THE APPLICATION OF MOLECULAR SIGNATURES TO CLASSIFICATION

THE APPLICATION OF MOLECULAR SIGNATURES AND PHYLOGENOMIC TECHNIQUES TO THE CLASSIFICATION AND IDENTIFICATION OF PROKARYOTIC ORGANISMS

By

Mobolaji Adeolu, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2016) McMaster University

(Biochemistry) Hamilton, Ontario

TITLE: The Application of Molecular Signatures and Phylogenomic Techniques to The Classification and Identification of Prokaryotic Organisms

AUTHOR: Mobolaji Adeolu, B.Sc. (McMaster University)

SUPERVISOR: Professor Radhey S. Gupta

NUMBER OF PAGES: xxiv, 213

ABSTRACT

The advent of large-scale genomic sequencing is providing researchers with an unparalleled wealth of information which can be used to elucidate the evolutionary relationships of living organisms. The newly available genome sequence data have enabled the use of comparative genomic techniques for the identification of novel molecular signatures, shared uniquely by evolutionarily related groups of organisms: conserved signature indels (CSIs) and conserved signature proteins (CSPs). These signatures allow for the unambiguous delineation of the prokaryotic taxa, independent of gene and genome based phylogenetic trees, and provide insights into novel aspects of their evolutionary relationships. The phylum Spirochaetes and the class *Betaproteobacteria* are large, diverse groups of bacteria, containing many important pathogenic and environmental organisms, which are classified primarily on the basis of 16S rRNA gene analysis. Here, I describe phylogenetic analyses of the phylum Spirochaetes based on genome derived molecular signatures. These analyses have yielded substantial evidence for differentiation between the three main sequenced groups of organisms within the phylum Spirochaetes and between the genus *Borrelia* from other closely related Spirochaetes. These findings have prompted a proposal to create three new orders and a new family within the phylum. These analyses have also supported the differentiation of two clinically distinct groups within the genus *Borrelia* and a proposal to divide the genus *Borrelia* into two genera. The use of molecular signatures and phylogenetic analysis of major groups within the class *Betaproteobacteria* are also described. The analyses of the order *Neisseriales* within this class resulted in a division of the order into two families, while the analyses of the genus *Burkholderia* supported the differentiation of the clinically relevant members of the genus *Burkholderia* from the plant-beneficial and environmental *Burkholderia* and a proposal to divide the genus into two genera. I also describe the use of phylogenomic techniques and molecular signatures to differentiate the seven main groups within the order *Enterobacteriales* and the integrated software pipeline used to produce the supermatrix based phylogenomic tree and genome distance calculations in the analysis of the order *Enterobacteriales*. The molecular signatures described in this thesis represent powerful new tools for evolutionary and systematic studies. Additionally, due to their taxon specificity, these molecular signatures are novel diagnostic markers for their specified group. Further analyses of these molecular signatures should lead to the discovery of novel functions and biological characteristics, mediated by CSIs and CSPs, which will provide important insights into the physiology, evolution, and adaptations of these groups.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my graduate supervisor, Dr. Radhey S. Gupta. Your endless passion for research has been infectious and has inspired me throughout my studies. You have given me both the guidance to build my knowledge and capabilities in my research area and the freedom to explore novel aspects of my research that I never would’ve expected when I joined the lab. You have had the patience to see me through my work, even on my worst days. Your mentorship has been invaluable and I owe every aspect of the body of work I have produced to your training.

I would also like to extend my gratitude to my supervisory committee, Dr. Herbert Schellhorn and Dr. Michael G. Surette. Their guidance and feedback during our numerous committee meetings over the course of my graduate research has helped guide my research and given me a more critical understanding of the field.

I also want to give my heartfelt thanks to my colleagues in the Gupta Lab who have worked by my side and helped me throughout my graduate research career. In particular, I want to acknowledge Sohail Naushad and Bijendra Khadka for their constant support and friendship in the Gupta Lab. Their presence in the lab brightened every day I experienced there.

I would like to express gratitude to my mother, Esther Adeolu, and my father, Bankole Adeolu. Your guidance, support, and tireless efforts have made me the person who I am today. I owe everything I am and have accomplished to you and love you both more than I can put into words.

Finally, I would like to give my thanks to my partner, Janelle Hinds, and my friend, Daniel Elbirt. Janelle has supported me and my research efforts throughout our time together. She has acted as a sounding board to bounce ideas off and has brought a critical perspective to my research efforts which I value. Her presence in my life has made me a better scientist and a better person. Daniel has made selfless efforts to aid me in my graduate work. Without his help and critical feedback, I would not have been able to complete this thesis. I deeply value the time he has taken to help me and his friendship.

PREFACE

The following work is a sandwich thesis. Chapters 2, 3, 4, and 5 are unaltered manuscripts, published in the years 2013 and 2014 while Chapter 7 is an unaltered manuscript, submitted for publication in June 2016. The preface section in each Chapter describes the details of the published and submitted articles, as well as my contribution to the multiple-authored work. Chapter 1, an introduction to the field of evolution and taxonomy research and the subjects of the manuscripts, provides context for the significance of the manuscripts included in this work. Chapter 6 describes an internally developed software pipeline for evolutionary genome analysis that has been utilized in the submitted manuscript included in Chapter 7. Chapter 8 reflects on the presented studies and describes the overall usefulness and future directions of the work. References for Chapters 1, 6, and 8 are provided at the end of this thesis. All chapters have been reproduced with the consent of all co-authors. Irrevocable, non-exclusive license has been granted to McMaster University and to the National Library of Canada from all publishers. Copies of permission and licenses have been submitted to the School of Graduate Studies.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 7

LIST OF ABBREVIATIONS

aa	Amino Acid
AAI	Average Amino Acid Identity
AF	Alignment Fraction
ANI	Average Nucleotide Identity
BCC	*Burkholderia cepacia* Complex
BLAST	Basic Local Alignment Search Tool
BLASTp	Protein vs Protein BLAST search
CSI	Conserved Signature Indel
CSP	Conserved Signature Protein
DDH	DNA-DNA Hybridization
del	Deletion
DNA	Deoxyribonucleic Acid
DnaK	Chaperone DnaK
E value	Expect value
GBDP	Genome BLAST Distance Phylogeny
GC or G+C	Guanine and Cysteine
GI	GenInfo Identifier
GLEAnS	Gupta Lab Evolutionary Analysis Software
GLIMPS	Gupta Lab Integrated Microbial Phylogeny and Supermatrix
GroEL	Chaperonin GroEL
GUI	Graphical User Interface
HMM	Hidden Markov Model
Indel	Insertion or Deletion
ins	Insertion
Kb	Kilobasepair (1000 base pairs)
MAFFT	Multiple Alignment based on Fast Fourier Transform
Mb	Megabasepair (1 000 000 basepairs)
MEGA	Molecular Evolutionary Genetics Analysis
MP	Maximum-Parsimony
ML	Maximum-Likelihood
MLSA	Multilocus Sequence Analysis
MLST	Multilocus Sequence Typing
MSA	Multiple Sequence Alignment
MUMi	Maximal Unique Matches Index
MUSCLE	Multiple Sequence Comparison by Log-Expectation
NCBI	National Center for Biotechnology Information
NUCMi	NUCmer Matches Index
NJ	Neighbour-Joining
ORF	Open Reading Frame
PATRIC	Pathosystems Resource Integration Center
PCR	Polymerase Chain Reaction
POCP	Percentage of Conserved Proteins

PROMi	PROmer Matches Index
PSI-BLAST	Position-Specific Iterative BLAST
RAxML	Randomized Axelerated Maximum Likelihood
RDP	Ribosomal Database Project
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
SH	Shimodaira and Hasegawa
SINA	SILVA Incremental Aligner
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
tRNA	Transfer RNA

GLOSSARY

16S ribosomal RNA or 16S rRNA: The small subunit of the 30S ribosomal complex. An integral part of protein production which is highly conserved and resistant to lateral transfer.

Alignment Trimming: Removal of spurious sequences or poorly aligned regions from a multiple sequence alignment.

Apomorphy: Specialized (derived) characters of an organism.

Archaea or Archaebacteria: One of the three domains of life, prokaryotic, differentiated from bacteria by genetic analysis, lacking peptidoglycan in their cell wall, and the presence of unique membrane lipids.

Average Amino Acid Identity: The average percentage of identical amino acids in alignments of proteins in two organisms.

Average Nucleotide Identity: The average percentage of identical nucleotides in alignments of genes in two organisms.

Bacteria or Eubacteria: One of the three domains of life, prokaryotic, differentiated from Archaea by genetic differences and the presence of peptidoglycan in their cell walls.

Bergey's Manual: The main resource for determining the identity of prokaryotic organisms, emphasizing bacterial species, using every characterizing aspect.

Bootstrap: A statistical procedure to assess the reliability of a result that involves resampling subsets of the data with replacement from the original data set. Jacknife is a similar procedure without replacement.

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

Comparative Genomics: A field of biological research which compares genomic features of different organisms such as sequence characteristics, genes, proteins, gene order, regulatory sequences, and other genetic or molecular characteristics in order to determine biological and evolutionary links between organisms.

Concatenation of Genes: Combining genetic data in a series and treating the combined data as a single gene for analysis.

Conserved Signature Indel (CSI): Insertions or deletions of a specific size uniquely present in a specific location in gene/protein sequences of organisms from the group of interest and absent in every other bacterial group. Flanked on both sides by conserved regions to ensure reliability.

Conserved Signature Protein (CSP): Lineage specific proteins found only in the group of interest with no homologs in any other bacterial group.

Convergent Evolution: The evolution of similar traits which occur due to similar adaptive benefits and not shared ancestry.

Core Genome: A term referring to the shared genes/proteins present in all members of a specified group.

Degenerate Oligonucleotide Primers: Primers to amplify the same region in related organisms. The sequence of the primers spans a range covering the different nucleotide sequences possible in region of amplification across different organisms.

DNA-DNA Hybridization: A technique used to determine the genetic distance between two organisms.

Effective Publication: A prokaryotic name which has been made generally available in published literature but has not met the requirements for valid publication.

Eukaryote: One of the three domains of life, differentiated from prokaryotes by the presence of membrane-bound organelles.

Genomic Distance: A measure of divergence between two genomes.

Graphical User Interface: The visual component of a computer application encompassing windows, icons, and menus.

Heuristic: Any approach that employs a practical method not guaranteed to be optimal, generally faster than optimal methods.

Hidden Markov Model: A statistical representation of a multiple sequence alignment.

Homologous genes/proteins: Sequences that are evolutionarily related by descent from a common ancestor.

International Code of Nomenclature of Bacteria or Bacteriological Code: The set of rules which govern the scientific names for Bacteria and Archaea.

Lateral Gene Transfer: Any movement of genetic material between organisms that does not occur during the transmission of DNA from a parent to a child.

Likelihood Ratio Test or SH-Like Test: A test comparing the likelihood of a null model (no specific relationship between organisms) to an alternative model (organisms X and Y are more related than organims X and Z) to determine the goodness of fit of the alternative model.

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

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

Maximum-Likelihood Tree: A phylogenetic tree built using the maximum-likelihood method which optimizes tree topology to maximize the likelihood of the tree being produced by the given alignment.

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

Multilocus Sequence Analysis: The analysis of multiple unlinked genes to determine phylogeny.

Multilocus Sequence Typing: The analysis of multiple unlinked genes to characterize and differentiate organisms.

Neighbour-Joining Tree: A phylogenetic tree built using the neighbor-joining method which clusters nodes based on a distance matrix.

Orthologous Gene/Protein or Ortholog: Sequences from different species that are evolutionarily related by descent from a common ancestral sequence and that diverged from one another as a result of speciation.

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

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

Paraphyletic: A group consisting of the group's last common ancestor and some, but not all, of the descendants of that ancestor.

Phenotype: An observable characteristic or trait of an organisms caused by an underlying genetic difference.

Phylogenetic Resolution: The ability to accurately elucidate the relationship between organisms.

Phylogenetic Tree: A branching “tree” diagram where bifurcations in the tree represent speciation events. Phylogenetic trees can contain additional information about branch reliability and divergence time.

Phylogenomic Tree: A phylogenetic tree based on the core genome of a group, can be produced using supertree or supermatrix methods.

Phylogenomics: Phylogenetic analysis using genome-scale data, encompasses phylogenetic trees and genomic distance measures.

Phylogeny: The evolutionary relationships between organisms.

Polyphasic Taxonomy: A methodology which includes disparate datatypes such as phenotypic, genotypic, molecular, and biochemical properties in taxonomy.

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

Prokaryotes: Organisms which lack a membrane bound nucleus and organelles. Prokaryotes can be divided into two main categories, Bacteria and Archaea.

SILVA: A curated 16s rRNA gene sequence database named after the Latin word *silva*, meaning forest.

Supermatrix: A concatenated set of all genes/proteins in a core genome.

Supertree: A consensus phylogenomic tree produced based on phylogenetic trees for all genes/proteins in a core genome.

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

Systematics: A field of biology dealing with the diversity of kinds. Systematics is usually divided into the two areas of phylogeny and taxonomy.

Taxonomic Framework: The structure of the nomenclatural classifications for a group of organisms.

Taxonomic Ranks: The levels within the taxonomic hierarchy (from most to least specific): species, genus, family, order, class, phylum, and domain.

Taxonomy: The science of naming and classifying organisms.

Tree topology: The arrangement of the various branches in a phylogenetic tree.

Valid Publication: A prokaryotic name is validly published if it is cited in the Approved Lists of Bacterial Names, published in the International Journal of Systematic and Evolutionary Microbiology or the International Journal of Systematic Bacteriology, or is published in a Validation List in one of the preceding journals.

CHAPTER 1

Background and Introduction

> *"Taxonomy is described sometimes as a science and sometimes as an art, but really it's a battleground. Even today there is more disorder in the system than most people realize."*
>
> *~Bill Bryson (A Short History of Nearly Everything, Chapter 23, 2003)*

An Early History of Prokaryotic Classification

The evolutionary history of living organisms on earth spans the most recent 3.5 billion years of the planet's 4.5 billion year history (Schopf, 1978; Woese et al., 1990). Unravelling the complex and circuitous history of life on earth constitutes one of the most fundamental and fascinating questions within the study of the life sciences (Schopf, 1978; Gupta & Griffiths, 2002). In particular, an understanding of the groupings of living organisms and the nature of their relationships to one another, codified as biological classifications and taxonomy, acts as the foundation which underlies and informs all modern fields of biology.

Carolus Linnaeus established the modern basis for rank-based taxonomic classification in the 18th century with the publication of the *Systema Naturae* (Linnaeus, 1758). However, it was not until Ferdinand Cohn began to classify bacteria into distinct genera in the 19th century, on the basis of their morphology, growth requirements, and pathogenic potential, that prokaryotes were given a meaningful standing in a modern Linnaean taxonomic classification system and were recognized as one of the earliest and most primitive divisions of life (Cohn, 1872, 1875). In the following decades, the bacterial classifications described by

Cohn were followed by an explosion of additional bacterial descriptions as the scientific community began to recognize the importance of prokaryotes as etiological agents of disease and to understand their role in food processing, agriculture, and ecology (Lehman & Neumann, 1896). As the range of diversity within the prokaryotes began to be appreciated, increased research attention, focussed on microorganisms, led to a number of novel insights regarding fundamental aspects of prokaryotic biochemistry and physiology. These studies yielded the first breakthroughs in understanding the diversity of metabolic pathways, the nature of oxygenic and anoxygenic photosynthesis, the carbon cycle, the extreme limits of life, symbiosis, and the mechanisms of information transfer in living organisms (Fred & Wilson, 1934; Waksman, 1934; Starkey & Waksman, 1943; Virtanen, 1947; Cohen, 1948; Gest & Kamen, 1948; Gest et al., 1950).

The increasingly diverse array of prokaryotes identified by microbiologists in the late 19th and early 20th centuries, exhibiting varied morphologies, physiologies, survival strategies, and life histories (Orla-Jensen, 1909; Pringsheim, 1923; Stanier & Van Niel, 1941), prompted the integration of increasing biochemical, physiological, and morphological properties in their descriptions and attempts at classification (Bergey et al., 1923; Stanier & Van Niel, 1941). This effort ultimately culminated in a universal Code of Bacteriological Nomenclature, approved at the 4th International Congress for Microbiology in 1947 (Huddleson, 1947; Stackebrandt, 2007). However, the number of readily determined phenotypic and biochemical properties in use to classify bacterial organisms in the first half of the 20th century were limited and were eventually found to exhibit high levels of convergence in unrelated organisms (Winogradsky, 1952; Stanier & Niel, 1962; Stanier et al., 1976). On this basis, many of the bacterial names described in the early 20th century were later found to be invalid or synonymous with other bacterial taxa.

In the late 20th century, advances in the determination of the nucleotide and amino acid sequences of DNA, RNA and protein molecules began to shine a light on the large number of poor and redundant taxa among the prokaryotes. In response, an effort was undertaken to purge bacterial taxonomy of all poorly defined, redundant, or ambiguous taxa (Lessel, 1971). The culmination of this effort was the concept of valid publication of bacterial nomenclature in a central repository (Lapage et al., 1973) and the Approved Lists of Bacterial Names (Skerman et al., 1980), a publication which contained all bacterial names deemed validly published and available for use by biologists. Of the 132 genera and 2703 species described in the 4th edition of Bergey's Manual of Determinative Bacteriology (Bergey et al., 1934), only 75 genera and 205 species were included in the Approved Lists of Bacterial Names (Skerman et al., 1980; Oren & Garrity, 2014).

16S rRNA and the Genetic Era of Prokaryotic Classification

The failure of bacterial classification systems based on phenotypic and biochemical properties in the middle of the 20th century, created an opportunity for alternative methods of phylogenetic inference to develop and gain prominence. In the 1950s, the discovery of the information transfer role and structure of deoxyribonucleic acid (DNA) (Hershey & Chase, 1952; Watson & Crick, 1953) provided researchers with a novel molecular target thought to encode all information underlying the phenotypic, physiological, and biochemical properties of an organism (Crick, 1970). Thus, one of the first methodologies developed to address the shortcomings of phenotype and biochemistry based classifications of prokaryotic organisms was the DNA-DNA hybridization (DDH) technique (Schildkraut et al., 1961; McCarthy & Bolton, 1963; Wayne et al., 1987). The DDH technique takes advantage of the weak bonds holding together the double strands of the DNA molecule. In the DDH technique, DNA molecules from two organisms are first heated and incubated, allowing the DNA strands to denature and dissociate, then then cooled, allowing the strands to reassociate. A subset of the reassociated DNA molecules is comprised of hybrids formed by the association of a strand from each of the two organisms. The strength of the association between the two strands of the hybridized DNA molecules is directly correlated with the similarity of the DNA sequences from those two organisms and can be calculated by determining the disassociation temperature (‘melting point’) of the hybridized DNA molecules. Thus, the DDH technique serves as a measure of the degree of genetic similarity between two organisms at a genome-wide level (McCarthy & Bolton, 1963; Wayne et al., 1987).

As classification based on phenotypic and biochemical properties fell out of favour, the DDH technique became widely used in prokaryotic systematics. The standardized definition of a species in prokaryotic systematics eventually became a group of organisms which share >70% DDH, correlated with a hybridized DNA melting point of <5ºC ΔT relative to the pure DNA molecules (Wayne et al., 1987; Tindall et al., 2010). However, the DDH technique has several important shortcomings. Notably, the determination of DDH values is a complicated, error-prone, time-consuming, and extremely laborious process, for which only a few laboratories are properly equipped (Rosselló-Mora, 2006). Additionally, several different methods for the measurement of DDH values exist which can produce different results (Grimont et al., 1980; Huss et al., 1983; Goris et al., 2007). Lastly, due to the comparative and experimental nature of the DDH technique, in which no sequence information is obtained, it is not possible to create incremental databases or scale the technique in any meaningful way (Goris et al., 2007; Schleifer, 2009). Due to these limitations, the DDH technique has proven unable to keep up with the growing rate of prokaryotic research and the growing diversity of described prokaryotic organisms.

In the late 1960s, the development of a method to partially characterize RNA sequences, referred to as oligonucleotide cataloguing (Sanger et al., 1965), and the development of the molecular clock concept, which allowed biological macromolecules to act as documents of evolutionary history (Zuckerkandl & Pauling, 1965), paved the way for the use of gene sequence analysis in evolutionary research. The 16S ribosomal RNA (rRNA) component of the 30S small ribosomal subunit quickly become the new gold standard in determining the evolutionary history of the prokaryotes (Fox et al., 1977b; Woese, 1987; Wilson, 1995; Garrity et al., 2001; Stackebrandt, 2006; Tindall et al., 2010). The 16S rRNA gene possessed a number of notable advantages that made it particularly suited to evolutionary inference. Firstly, the ribosome is essential for survival and directly comparable ribosomal genes are universally present in prokaryotes and eukaryotes, facilitating comparison between the multiple, disparate domains of life (Fox et al., 1980; Woese, 1987; Woese et al., 1990). Beyond its ubiquity, the 16S rRNA gene is easily isolated, and, as part of the large ribosomal complex, unlikely to undergo lateral gene transfer (Olsen et al., 1994; Patel, 2001; Janda & Abbott, 2007). Furthermore, the 16S rRNA gene contains both highly conserved and variable regions facilitating the classification of both closely related and highly divergent bacterial groups and the development of universal PCR primers that are able to amplify 16S rRNA genes readily from uncultured organisms (Greisen et al., 1994; Marchesi et al., 1998; Wang & Qian, 2009).

The use of 16S rRNA gene analysis was instrumental in one of the most significant advancements in modern taxonomy, the proposal of the three-domain model of life (Woese et al., 1990). Utilizing early oligonucleotide cataloguing techniques, Woese and colleagues compared the 16S rRNA genes of different prokaryotic organisms and the 18S rRNA genes of eukaryotic organisms (Fox et al., 1977a; Fox et al., 1977b; Olsen et al., 1985). These analyses shed new light on the genetic diversity among the prokaryotes and provided the first evidence that the Archaeabacteria were as distinct from Eubacteria as they were from the Eukaryotes (Fox et al., 1977b; Woese et al., 1990). Ultimately, these studies resulted in the proposal of the three-domain model of classification, in which Bacteria, Archaea, and Eukaryota are considered coequal and fundamental divisions of life on earth, which remains the dominant model for biological classification at the highest taxonomic levels (Woese et al., 1990).

The 16S rRNA gene has become the foundation of modern prokaryotic systematics. Analysis of the 16S rRNA gene sequence has been used to refine the classification of almost all described microbial groups (Garrity et al., 2005; Yarza et al., 2008; Kämpfer, 2012) and sequencing of the 16S rRNA gene has become an informal requirement for the description of all new prokaryotic species (Tindall et al., 2006; Tindall et al., 2010; Kämpfer & Glaeser, 2013). Bergey's Manual of Systematics of Archaea and Bacteria (Whitman, 2015a), the modern successor to Bergey’s Manual of Determinative Bacteriology, uses 16S rRNA gene sequence based phylogenies as its organizing basis and the All-Species Living Tree project, which has become the *de facto* tree of life for systematic purposes, is also based on alignments of the 16S rRNA gene sequence (Yarza et al., 2008; Yilmaz et al., 2013). Additionally, the research effort that has been focussed on the 16S rRNA gene sequence has led to the development of large, comprehensive databases of the 16S rRNA gene sequences, comprising nearly all described prokaryotic species and strains (Quast et al., 2013; Cole et al., 2014). 16S rRNA gene sequence similarity values have also superseded the use of DDH values for prokaryotic species demarcation (Stackebrandt & Goebel, 1994; Stackebrandt & Ebers, 2006; Tindall et al., 2006; Tindall et al., 2010). A 16S rRNA gene sequence similarity value of 97% is thought to correlate to the 70% DDH threshold for species demarcation (Stackebrandt & Goebel, 1994). However, the initial study that established that value was based on only 57 comparisons between 16S rRNA gene similarity values and DDH values (Stackebrandt & Goebel, 1994). Subsequent studies utilizing larger datasets have produced slightly different species thresholds, such as a 98.7% 16S rRNA gene sequence similarity threshold for species demarcation in a study using 380 comparisons (Stackebrandt & Ebers, 2006) and a 98.2% threshold in a study using 571 comparisons (Meier-Kolthoff et al., 2013). An additional threshold of 95% 16S rRNA gene sequence similarity for genus level demarcation has also been established in literature (Tindall et al., 2010). Until recently, there were no robust guidelines for the demarcation of taxonomic ranks above the genus level. However, a recent study examining the 16S rRNA gene sequences of 8602 type strains within the SILVA 16S rRNA database (Quast et al., 2013) established thresholds of 94.5%, 86.5%, 82%, 78.5%, and 75% 16S rRNA gene sequence similarity for the demarcation of prokaryotic taxa at the level of Genus, Family, Order, Class, and Phylum, respectively (Yarza et al., 2014), providing novel guidance for 16S rRNA gene based classifications. That said, it is important to note that all of the established thresholds are conservative guidelines and that their strict application can overlook important and distinct taxa that can be distinguished based on other means of analysis (Oren & Garrity, 2014; Yarza et al., 2014; Whitman, 2015b).

Despite the usefulness of the 16S rRNA gene for evolutionary studies, use of the 16S rRNA gene to elucidate evolutionary relationships among the prokaryotes, independent of other forms of evidence, has limitations. Firstly, the 16S rRNA gene has limited capacity to differentiate among very closely related and recently diverged species/strains of prokaryotes, due to the high sequence conservation and limited resolving power of the gene (Fox et al., 1992; Tang et al., 1998; Mignard & Flandrois, 2006; Janda & Abbott, 2007; Reller et al., 2007). The 16S rRNA gene also has limited capacity to resolve the relative branching orders of different prokaryotic phyla at the highest taxonomic levels (Garrity et al., 2001; Garrity et al., 2005; Yarza et al., 2008; Puigbo et al., 2009). Additionally, the GC content of 16S rRNA genes are correlated with the habitat and optimal growth temperatures of the prokaryote in which it is found; leading to convergent 16S rRNA gene GC content values in organisms with similar optimal growth temperatures (Stackebrandt et al., 2002; Stackebrandt et al., 2007; Gupta & Lali, 2013). Evolutionary inferences based on 16S rRNA gene sequence analysis can also be confounded by prokaryotic organisms possessing multiple copies of the 16S rRNA gene, which can differ by up to 2% or more of their sequence positions (Klappenbach et al., 2001; Boucher et al., 2004). Lastly, the structural elements of 16S rRNA gene are constrained and cannot freely change, leading these elements to change in sudden jumps rather than along a continuum, creating the potential for erroneous conclusions about the prokaryotic relationships which they support (Ludwig et al., 1998; Ludwig & Klenk, 2001). Hence the interest in the identification and use of other genes and proteins which have the potential to resolve evolutionary questions not sufficiently resolved by 16S rRNA gene sequence analysis.

The primary category of genes used as alternative evolutionary markers to the 16S rRNA gene are essential, single copy housekeeping genes such as the β-subunit of DNA gyrase (*gyrB*), the β-subunit of RNA polymerase (*rpoB*), the sigma 70 (sigma D) factor of RNA polymerase (*rpoD*), recombinase A (*recA*), the β-subunit of ATP synthase F0F1 (*atpD*), translation initiation factor IF-2 (*infB*), tRNA modification GTPase ThdF or TrmE (*thdF*), or the chaperonin GroEL (*groEL*) (Kämpfer, 2012; Glaeser & Kämpfer, 2015). These genes possess many of the same benefits as the 16S rRNA gene. They are ubiquitous among most organisms, essential for survival, large and slow evolving, and can be amplified and isolated using near universal degenerate PCR primer sets (Maiden et al., 1998; Gevers et al., 2005; Maiden, 2006). Additionally, the use of multiple genes for evolutionary inference limits the confounding effects of atypical evolutionary rates, genetic recombination, and lateral gene transfers at a single genetic locus (Rokas et al., 2003; Ciccarelli et al., 2006; Wu et al., 2009). The use of multiple (usually 5-10) housekeeping genes in genotypic characterization among prokaryotes is referred to as multilocus sequence typing (MLST) while the same methodology applied to the construction of prokaryotic phylogenetic trees is referred to as multilocus sequence analysis (MLSA) (Maiden et al., 1998; Gevers et al., 2005).

Unique sets of genetic loci have been identified and validated for the MLST-based characterization and differentiation of pathogenic prokaryotic and eukaryotic groups exhibiting significantly greater strain-level resolution than 16S rRNA based characterization (Jolley et al., 2004; Maiden, 2006; Jolley & Maiden, 2010; Maiden et al., 2013). Species and genus level MLST gene sequence similarity thresholds have been developed for specific groups to augment the universal 16S rRNA gene sequence similarity thresholds such as the genera *Burkholderia* (Vandamme & Peeters, 2014), *Streptomyces* (Rong et al., 2009), and *Chlamydia* (Sachse et al., 2015). These MLST gene sets have also been used for MLSA based phylogenetic analyses providing novel evolutionary and taxonomic insights for groups that are not clearly resolved based on the analysis of the 16S rRNA gene (Postic et al., 2007; Brady et al., 2013; Peeters et al., 2013; Glaeser & Kämpfer, 2015). Though universally conserved gene sets have been utilized for large-scale MLSA based phylogenetic analyses spanning the entire tree of life (Santos & Ochman, 2004; Jolley et al., 2012; Hug et al., 2016), these universally conserved gene sets cannot distinguish between many of the closely related organisms that group-specific MLST gene sets were designed to characterize and differentiate (Gevers et al., 2005; Glaeser & Kämpfer, 2015).

The Impact of Whole Genome Sequences on Prokaryotic Classification

The sequencing of the first microbial genome in 1995, belonging to the organism *Haemophilus influenzae* (Fleischmann et al., 1995), heralded the beginning of the genomic age of evolutionary biology. The 1.8 megabasepair (Mb) genome of *Haemophilus influenzae* used conventional Sanger sequencing techniques and cost hundreds of thousands of dollars to produce (Loman et al., 2012). The prohibitive cost of genome sequencing in the 1990s limited the use of sequenced genome data in evolution and taxonomy research. However, in 2005, the development of high-throughput next generation sequencing (NGS) technology massively reduced the cost of sequencing individual genomes (Metzker, 2005; Wetterstrand, 2016). With the advent of high-throughput NGS technologies, such as 454 parallel pyrosequencing, Sequencing by Oligonucleotide Ligation and Detection (SOLiD), ion semiconductor sequencing, and Illumina dye sequencing, the cost of genome sequencing has and continues to drop exponentially (Liu et al., 2012). Recently, the Illumina HiSeq X Ten, a genome sequencing platform which can generate up to 1 800 000 Mb of sequence data per run, has been able to sequence a human genome for less than $1000, a 99.999% reduction in cost from the first human genome sequence produced in 2001 (Venter et al., 2001; van Dijk et al., 2014; Wetterstrand, 2016). This massive decrease in the cost of genome sequencing has been associated with a commensurately massive increase in the number of available genome sequences.

To wit, here are currently over 75 000 genome sequences from over 16 000 organisms available in the NCBI genome database (NCBI, 2016).

This exponentially increasing wealth of genome sequence data has led to the development of several novel methods of understanding organismal relationships based on their genome sequences (Chun & Rainey, 2014). The most popular class of methods are overall genome relatedness indices. Overall genome relatedness indices are methods of measuring genome to genome distance, which serves as a proxy for the classic DDH value without its associated limitations. These indices include: average nucleotide identity (ANI), which measures the sequence identity of shared genes and has an established 95-96% identity threshold for species level demarcation (Konstantinidis & Tiedje, 2005; Richter & Rosselló-Móra, 2009; Kim et al., 2014; Varghese et al., 2015); average amino acid identity (AAI), which measures the sequence identity of shared proteins and provides greater stability for more distant comparisons than ANI (Konstantinidis & Tiedje, 2005; Rosselló-Mora, 2005; Thompson et al., 2013); percent of conserved proteins (POCP) and alignment fraction (AF), which measure the proportion of proteins/genes shared by two genomes (Qin et al., 2014; Varghese et al., 2015); genome BLAST distance phylogeny (GBDP) (Henz et al., 2005; Meier-Kolthoff et al., 2013), which uses a methodology similar to ANI but does not break the genome into artificial blocks and has a closer correlation to DDH values; and the maximal unique matches index (MUMi) (Deloger et al., 2009) and the related nucleotide matches (NUCMi) and protein matches (PROMi) indices

(Dias et al., 2011), which are based on the sequence similarity of shared genome segments identified during whole genome alignments. Each of these methods synthesizes large amounts of genome sequence data to determine evolutionary relationships. Their results generally correlate well with established phylogenies based on the 16S rRNA gene sequence while being robust against lateral gene transfer and other anomalous genetic information (Chun & Rainey, 2014; Zuo et al., 2015). In Chapter 3 of this thesis, AAI values are utilized to support the differentiation of two groups within the genus *Borrelia*, while POCP is utilized in Chapter 7 of this thesis to support the distinctiveness of the main groups within the order *Enterobacteriales*. An integrated software pipeline is described in Chapter 6 of this thesis, which can be utilized to produce both AAI and POCP values from genome sequence data.

Another class of methods for understanding organismal relationships based on their genomes is referred to as alignment independent genome to genome distance measures (Bonham-Carter et al., 2014; Chan et al., 2014). These methodologies utilize the nucleotide or amino acid composition of genomes to infer their overall relatedness. Alignment independent genome to genome distance measures can be broken down into four broad categories: factor frequencies (Liu et al., 2008), composition vectors (Lu et al., 2008; Chan et al., 2012), data compression (Otu & Sayood, 2003; Ulitsky et al., 2006), and common substrings (Ukkonen, 1985). Each alignment independent genome to genome distance measure determines genomic similarity, using statistical methodologies to compare the frequency of specific length sub-sections of the genome, referred to as words or *k*-mers, between pairs of genomes. Due to the alignment free nature of these methodologies, they can be computed extremely quickly and are often used as the first heuristic approach in sequence similarity search algorithms (Altschul et al., 1997; Kent, 2002; Edgar, 2010). However, alignment independent genome to genome distance measures only roughly correlate with 16S gene sequence analysis and are not regularly used in evolution and taxonomy research (Gao et al., 2007; Jun et al., 2010; Zuo et al., 2015). Moreover, none of the overall genome relatedness indices or alignment independent genome to genome distance measures can be used to produce phylogenetic trees which are significantly more robust than those already provided by analysis of the 16S rRNA gene (Verma et al., 2013; Chun & Rainey, 2014; Zuo et al., 2015). Thus, these methodologies are primarily limited to supplemental roles in polyphasic evolutionary analysis that already incorporates a robust phylogenetic methodology (Ramasamy et al., 2014; Vandamme & Peeters, 2014).

Genome-Scale Phylogenetic Tree Construction

Phylogenetic trees, which are hierarchal and bifurcating tree diagrams depicting the evolutionary history of a group of organisms, have formed the backbone of evolutionary and systematic research for the last 25 years (Woese et al., 1990; Stackebrandt & Goebel, 1994; Yilmaz et al., 2013; Oren & Garrity, 2014; Parte, 2014). The construction of phylogenetic trees is generally based on clustering similar organisms using measures of genetic or genomic distance, such as in the neighbour-joining approach (Saitou & Nei, 1987), or on the optimization of an overall tree score, such as in the maximum-parsimony (Fitch, 1971), maximum-likelihood (Felsenstein, 1981), and Bayesian inference (Rannala & Yang, 1996) approaches. Maximum-parsimony, maximum-likelihood, and Bayesian inference approaches attempt to optimize tree scores based on minimum number of changes required to reconcile the tree and the gene/protein alignment, the log-likelihood of the tree based on the gene/protein alignment, and the posterior probability of generating the tree from the gene/protein alignment, respectively, often using heuristic methodologies (Yang & Rannala, 2012). The strength (i.e. consistency) of the evolutionary relationships depicted in the phylogenetic tree are primarily determined by using statistical tests such as jackknife and bootstrap resampling (Quenouille, 1949; Efron, 1992) or likelihood ratio analysis (Shimodaira & Hasegawa, 1999; Anisimova & Gascuel, 2006).

The availability of genome sequence data allows for phylogenetic tree construction based on large amounts of genetic information—potentially consisting of the entire core genome—which has consistently been shown to have higher reliability and resolving power and to be more resistant to lateral gene transfer events than phylogenetic trees based on any single gene or protein (Rokas et al., 2003; Dutilh et al., 2004; Delsuc et al., 2005; Ciccarelli et al., 2006; Wu & Eisen, 2008; Puigbo et al., 2009; Wu et al., 2009). There are two main approaches to utilizing genomic sequence data in the construction of robust phylogenetic trees. The first approach involves the construction of individual phylogenetic trees, based on sequence alignments of each gene/protein in the shared core genome, which are later combined into a single consensus phylogenetic tree referred to as a supertree (Bininda-Emonds, 2004; Beiko et al., 2005; Puigbo et al., 2009; Lang et al., 2013). The supertree exhibits the dominant branching patterns present in the multiple individual phylogenetic trees, allowing their core trends to be readily visualized. This methodology has two main benefits. Firstly, due to the exponential increase in the difficulty of phylogenetic tree construction as the length of the analyzed gene sequence increases (Stamatakis, 2014), the supertree method is more computationally efficient than methods that attempt to analyze all of the genome at once. For example, reconstructing a phylogeny based on one alignment of size X takes more total computational power than reconstructing the phylogeny of ten alignments of size 0.1X. Secondly, the supertree method simultaneously produces individual gene trees as it produces the consensus supertree, providing additional gene based phylogenies which can be further analyzed and compared to the consensus supertree. The second approach to utilizing genomic sequence data in robust phylogenetic trees involves the individual alignment of either a limited number of genes/proteins or all genes/proteins in the shared genome, followed by the concatenation of these alignments into a single dataset referred to as a supermatrix (Brown et al., 2001; Snel et al., 2005; Ciccarelli et al., 2006; Lang et al., 2013; Segata et al., 2013; Hug et al., 2016). This supermatrix is then used to produce a highly robust phylogenetic tree. The supermatrix method has a few notable advantages over the supertree method including improved resolution of the relationships among organisms in the tree and compatibility with traditional statistical methods to determine the strength of the topological relationships within the tree, including bootstrap resampling and likelihood ratio analysis (Gadagkar et al., 2005; Ren et al., 2009; Lang et al., 2013). Chapter 6 of this thesis discusses an integrated software pipeline that can produce supermatrix based phylogenetic trees from genome sequence data.

The quality and reliability of supertrees and supermatrix based phylogenetic trees are dependent on the composition and size of the core genome of the examined organisms. In closely related organisms, where the core genome may consist of thousands of genes/proteins (Rasko et al., 2008; Bottacini et al., 2010; den Bakker et al., 2010; Valot et al., 2015), phylogenetic supertrees and phylogenetic trees based on concatenated sequences are particularly robust and reliable. However, the core genome for distantly related groups of organisms is limited in size, consisting largely of genes which are functionally interlinked (Ciccarelli et al., 2006; Dagan & Martin, 2006; Hug et al., 2016). Thus, supertrees and supermatrix based phylogenetic trees for diverse groups of organisms are limited in the numbers of genes they can include, and should be supplemented with additional forms of analysis.

The Utility of Molecular Signatures in Evolutionary and Taxonomic Studies

The wealth of available genomic sequence information also allows for the identification of conserved molecular signatures specific to related groups of prokaryotic organisms. The molecular signatures that are ideally suited for use in evolutionary studies as molecular signatures are homologous apomorphic characters that evolved only once (i.e. a synapomorphy) during the course of evolution (Stackebrandt & Schumann, 2006; Gupta, 2014). One such class of molecular signatures, that has been a focus of much recent evolutionary research, are Conserved Signature insertions and deletions, i.e. Indels, (CSIs) of defined lengths and locations in widely distributed proteins, which are specific for particular groups of organisms (Gupta, 2014; Gupta et al., 2015a; Gupta et al., 2015b; Gupta, 2016; Gupta et al., 2016). Indels of a defined size, flanked on both sides by conserved regions to ensure they constitute reliable characteristics which are not a result of alignment errors, provide extremely useful phylogenetic information (Gupta, 2014). The high conservation of their location in the genome suggests that they have high functional significance and are likely under significant selective pressure for retention (Gao & Gupta, 2012b; Gupta, 2014). Many of these conserved signature indels (CSIs), such as those found in the GroEL and DnaK proteins of many bacteria, are essential for bacterial growth and lead to cell death if removed or significantly altered (Singh & Gupta, 2009). Thus, CSIs in widely distributed proteins in a defined group of bacteria are extremely rare genetic changes and are highly specific molecular signatures which have functional significance and may be essential for bacterial growth (Rokas & Holland, 2000; Singh & Gupta, 2009; Zhi et al., 2012). The genetic changes which give rise to conserved indels are highly specific and extremely rare in occurrence, thus, such changes are unlikely to arise in different groups due to convergent evolution (Rokas & Holland, 2000; Naushad & Gupta, 2013; Gupta, 2014). Hence, the most parsimonious explanation for the unique presence of a CSI in a particular group of organisms is that the rare genetic change responsible for the CSI first occurred in a common ancestor of the group of species where the CSI was found and was then transferred vertically to its various descendants (Rivera & Lake, 1992; Rokas & Holland, 2000; Gupta, 2014). However, it is important to consider the possibility that the shared presence of a CSI could be due to cases of lateral gene transfers. Further, based upon the presence or absence of a particular CSI in various outgroup species, it is possible to infer whether the CSI under consideration is an insertion or a deletion in a given group, and which of the two character states of the protein is ancestral and which is derived (Rivera & Lake, 1992; Gupta, 1998; Gupta, 2014). Thus, by making use of CSIs that have been introduced at various stages of evolution, it is possible to derive a rooted evolutionary relationship among various groups or taxa under consideration independently of phylogenetic trees (Gupta, 2001; Gupta, 2014). The applications of CSI based evolutionary inference to the taxonomy of specific groups of bacteria are described in Chapters 2, 3, 4, 5, and 7 of this thesis.

In addition to conserved indels, comparative genomic analyses have been an essential resource in identifying another important class of molecular signatures useful to evolutionary studies. These markers consist of whole proteins found uniquely in monophyletic clades of bacteria (Lerat et al., 2005; Gao et al., 2006; Dutilh et al., 2008; Gupta, 2010; Gao & Gupta, 2012b). Many proteins of known and unknown functions, thought to be unique and distinctive, have been found to be characteristic of various species of bacteria from monophyletic clades of different phylogenetic depths (Snel et al., 2005; Dutilh et al., 2008; Gupta & Sharma, 2015; Gupta, 2016). Although the mechanisms responsible for the origin/evolution of genes for these proteins are unclear (Dutilh et al., 2008; Kuo & Ochman, 2009), their presence in a conserved state in all or most species/strains from a monophyletic clade, but nowhere else, suggests that the genes for these proteins first evolved in a common ancestor of these clades and were subsequently vertically passed down to its various descendants (Dutilh et al., 2008; Fang et al., 2008; Narra et al., 2008). Thus, like CSIs, these Conserved Signature Proteins (CSPs) provide valuable molecular signatures for evolutionary studies of different bacterial clades (Dutilh et al., 2008; Gupta & Gao, 2010; Gao & Gupta, 2012a; Gupta & Sharma, 2015). The identification of a number of CSPs which distinguish two closely related groups within the genus *Borrelia* are described in Chapter 3 of this thesis.

Research Objective and an Overview of The Phylum Spirocheates and The Class Betaproteobacteria

The overall objective of my graduate research has been the identification and analysis of molecular signatures, such as CSIs and CSPs, and the utilization of phylogenomic and comparative genomic techniques to elucidate the evolutionary history of the phylum Spirochaetes and the class *Betaproteobacteria* and their main constituent groups.

The phylum Spirochaetes consists of a large and diverse group of motile bacteria which are widespread in the environment and are highly prevalent disease causing agents (Seshadri et al., 2004; Paster, 2011). There are two particularly important genera within the phylum Spirochaetes whose species are the causative agents of many globally prevalent illnesses, *Treponema* and *Borrelia* (Bellgard et al., 2009). *Treponema pallidum* subspecies *pallidum* is the causative agent of syphilis, a sexually transmitted disease which affects at least twenty-five million adults worldwide (Gerbase et al., 1998). Members of the genus *Borrelia* are the causative agents of both Lyme disease, which is currently the most prevalent vector-borne disease in North America and temperate regions of Eurasia, and relapsing fever, which is a disease endemic to many disparate regions of the world (Lindgren & Jaenson, 2006; Cutler, 2010; Adams et al., 2013). However, despite the clinical importance and diverse characteristics of its members, the phylum Spirochaetes was, until recently, comprised of a single class, *Spirochaetia*, containing a single order, *Spirochaetales*, which was made up of four families (Paster, 2011).

Similarly, the class *Betaproteobacteria* is a large and diverse group within the phylum Proteobacteria, consisting of over 200 bacterial species divided into seven orders (Parte, 2014). Of the seven orders within the *Betaproteobacteria*, the orders *Neisseriales* and *Burkholderiales* are of particular interest due to their size and their pathogenic members. Namely, *Neisseria gonorrhoeae*, the causative agent of the increasingly drug resistant sexually transmitted infection gonorrhea, which affects approximately 88 million individuals a year worldwide (World Health Organization, 2011), *Neisseria meningitides*, the primary causative agent of infectious meningococcal meningitis (Stephens et al., 2007; Cohn et al., 2010), and the genus *Burkholderia*, a large group of soil bacteria which are ubiquitous in the environment and can act as opportunistic pathogens (White, 2003; Workowski et al., 2008; Lipuma, 2010). Despite the diversity within the order *Neisseriales* and the presence of important pathogens, until recently, all members of the order *Neisseriales* were placed within a single family, *Neisseriaceae*, and, until recently, all of the >70 diverse members of the genus *Burkholderia* were placed within one genus (Coenye & Vandamme, 2003; Palleroni, 2005).

Research Overview

The analyses completed in my research have been utilized to propose significant taxonomic revisions for the phylum Spirochaetes and major groups within the class *Betaproteobacteria*, reflecting the diversity present in these groups (Adeolu & Gupta, 2013; Gupta et al., 2013b; Adeolu & Gupta, 2014; Sawana et al., 2014). In Chapter 2 of this thesis, I describe the use of CSIs and phylogenetic trees to differentiate the three main sequenced groups of organisms within the phylum Spirochaetes and to differentiate the genus *Borrelia* from other closely related Spirochaetes. The chapter concludes with a proposal for a novel taxonomic framework for the phylum Spirochaetes including three new orders and a new family. Chapter 3 of this thesis details a corollary study focused on the genus *Borrelia*. In this chapter, I describe the use of CSIs and CSPs, phylogenetic trees, and average nucleotide identity analysis to differentiate two clinically distinct groups within the genus *Borrelia* and a proposal to divide the genus *Borrelia* into two genera.

In Chapter 4 of this thesis, I describe the use of CSIs and phylogenetic trees to differentiate the obligate host-associated members of the order *Neisseriales* from the other genera within the order and a proposal to recognize the distinctiveness of the host-associated members by limiting the family *Neisseriaceae* to only those members, while transferring the other genera within the order *Neisseriales* to a novel family. Chapter 5 of this thesis describes a subsequent study focused on the genus *Burkholderia*, in which CSIs and phylogenetic trees are utilized to differentiate the opportunistically pathogenic members of the genus *Burkholderia* from the plant-beneficial and environmental

Burkholderia and a division of the two groups within the genus into two distinct genera is proposed.

Chapter 7 of this thesis describes the use of CSIs, protein based phylogenetic trees, and shared protein content to differentiate the seven main groups within the order *Enterobacteriales* and proposes that each of the seven groups should be treated as family-level taxa. Chapter 6 of this thesis describes an integrated software pipeline that produces supermatrix based phylogenetic trees and calculates both shared protein content and average amino acid identity from genome sequences which is utilized in the study described in Chapter 7. Lastly, Chapter 8 reflects on the studies and phylogenomic tools presented herein,and describes the overall usefulness and future directions of the work.

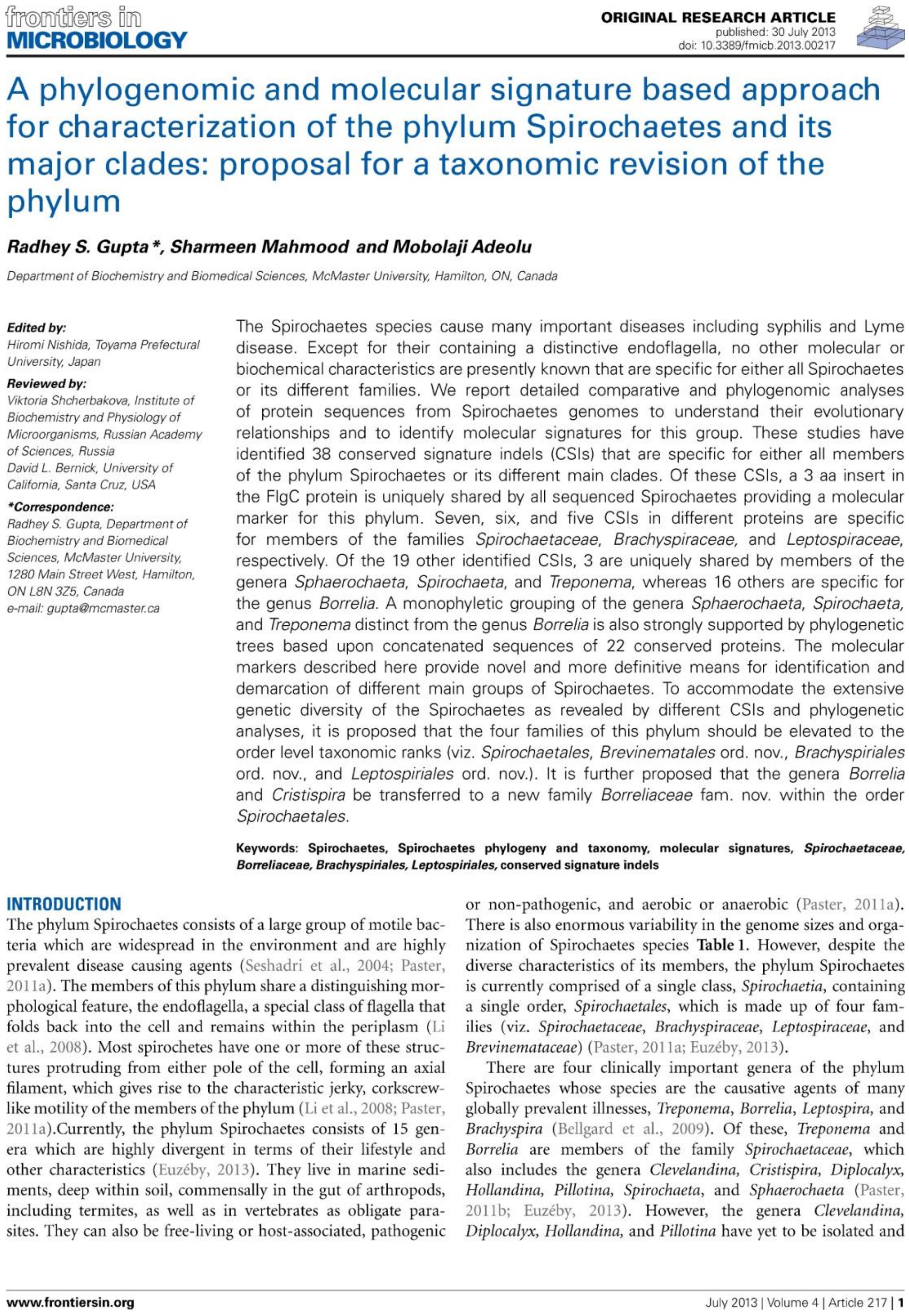
CHAPTER 2

A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum

This chapter describes the use of molecular signatures (CSIs) and phylogenetic trees to differentiate the three main sequenced groups of organisms within the phylum Spirochaetes. Additionally, this chapter describes the differentiation of the genus *Borrelia* from other closely related Spirochaetes genera (viz. *Treponema*, *Spirochaeta*, and *Sphaerochaeta*). The chapter concludes with a proposal for a novel taxonomic framework for the phylum Spirochaetes including three new orders and a new family. My contributions to the completion of this chapter include the construction of all phylogenetic trees shown, reexamination of the specificity of the identified of CSIs, the creation of the taxonomic proposals, the writing drafts and revisions of the manuscript, and the production of all main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials for this work are not included in the chapter but can be accessed along with the rest of the manuscript at:

Gupta, R. S., Mahmood, S., & Adeolu, M. (2013). *Frontiers in microbiology*, *4*, 217.



frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 30 July 2013
doi: 10.3389/fmicb.2013.00217

A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum

Radhey S. Gupta*, Sharmeen Mahmood and Mobolaji Adeolu

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

Edited by:
Hiromi Nishida, Toyama Prefectural University, Japan

Reviewed by:
Viktoria Shcherbakova, Institute of Biochemistry and Physiology of Microorganisms, Russian Academy of Sciences, Russia
David L. Bernick, University of California, Santa Cruz, USA

***Correspondence:**
Radhey S. Gupta, Department of Biochemistry and Biomedical Sciences, McMaster University, 1280 Main Street West, Hamilton, ON L8N 3Z5, Canada
e-mail: gupta@mcmaster.ca

The Spirochaetes species cause many important diseases including syphilis and Lyme disease. Except for their containing a distinctive endoflagella, no other molecular or biochemical characteristics are presently known that are specific for either all Spirochaetes or its different families. We report detailed comparative and phylogenomic analyses of protein sequences from Spirochaetes genomes to understand their evolutionary relationships and to identify molecular signatures for this group. These studies have identified 38 conserved signature indels (CSIs) that are specific for either all members of the phylum Spirochaetes or its different main clades. Of these CSIs, a 3 aa insert in the FlgC protein is uniquely shared by all sequenced Spirochaetes providing a molecular marker for this phylum. Seven, six, and five CSIs in different proteins are specific for members of the families *Spirochaetaceae*, *Brachyspiraceae,* and *Leptospiraceae*, respectively. Of the 19 other identified CSIs, 3 are uniquely shared by members of the genera *Sphaerochaeta*, *Spirochaeta*, and *Treponema*, whereas 16 others are specific for the genus *Borrelia*. A monophyletic grouping of the genera *Sphaerochaeta*, *Spirochaeta*, and *Treponema* distinct from the genus *Borrelia* is also strongly supported by phylogenetic trees based upon concatenated sequences of 22 conserved proteins. The molecular markers described here provide novel and more definitive means for identification and demarcation of different main groups of Spirochaetes. To accommodate the extensive genetic diversity of the Spirochaetes as revealed by different CSIs and phylogenetic analyses, it is proposed that the four families of this phylum should be elevated to the order level taxonomic ranks (viz. *Spirochaetales*, *Brevinematales* ord. nov., *Brachyspiriales* ord. nov., and *Leptospiriales* ord. nov.). It is further proposed that the genera *Borrelia* and *Cristispira* be transferred to a new family *Borreliaceae* fam. nov. within the order *Spirochaetales*.

Keywords: Spirochaetes, Spirochaetes phylogeny and taxonomy, molecular signatures, *Spirochaetaceae, Borreliaceae, Brachyspiriales, Leptospiriales,* conserved signature indels

INTRODUCTION

The phylum Spirochaetes consists of a large group of motile bacteria which are widespread in the environment and are highly prevalent disease causing agents (Seshadri et al., 2004; Paster, 2011a). The members of this phylum share a distinguishing morphological feature, the endoflagella, a special class of flagella that folds back into the cell and remains within the periplasm (Li et al., 2008). Most spirochetes have one or more of these structures protruding from either pole of the cell, forming an axial filament, which gives rise to the characteristic jerky, corkscrew-like motility of the members of the phylum (Li et al., 2008; Paster, 2011a).Currently, the phylum Spirochaetes consists of 15 genera which are highly divergent in terms of their lifestyle and other characteristics (Euzéby, 2013). They live in marine sediments, deep within soil, commensally in the gut of arthropods, including termites, as well as in vertebrates as obligate parasites. They can also be free-living or host-associated, pathogenic or non-pathogenic, and aerobic or anaerobic (Paster, 2011a). There is also enormous variability in the genome sizes and organization of Spirochaetes species **Table 1**. However, despite the diverse characteristics of its members, the phylum Spirochaetes is currently comprised of a single class, *Spirochaetia*, containing a single order, *Spirochaetales*, which is made up of four families (viz. *Spirochaetaceae*, *Brachyspiraceae*, *Leptospiraceae*, and *Brevinemataceae*) (Paster, 2011a; Euzéby, 2013).

There are four clinically important genera of the phylum Spirochaetes whose species are the causative agents of many globally prevalent illnesses, *Treponema*, *Borrelia*, *Leptospira,* and *Brachyspira* (Bellgard et al., 2009). Of these, *Treponema* and *Borrelia* are members of the family *Spirochaetaceae*, which also includes the genera *Clevelandina, Cristispira, Diplocalyx, Hollandina, Pillotina, Spirochaeta*, and *Sphaerochaeta* (Paster, 2011b; Euzéby, 2013). However, the genera *Clevelandina, Diplocalyx, Hollandina,* and *Pillotina* have yet to be isolated and

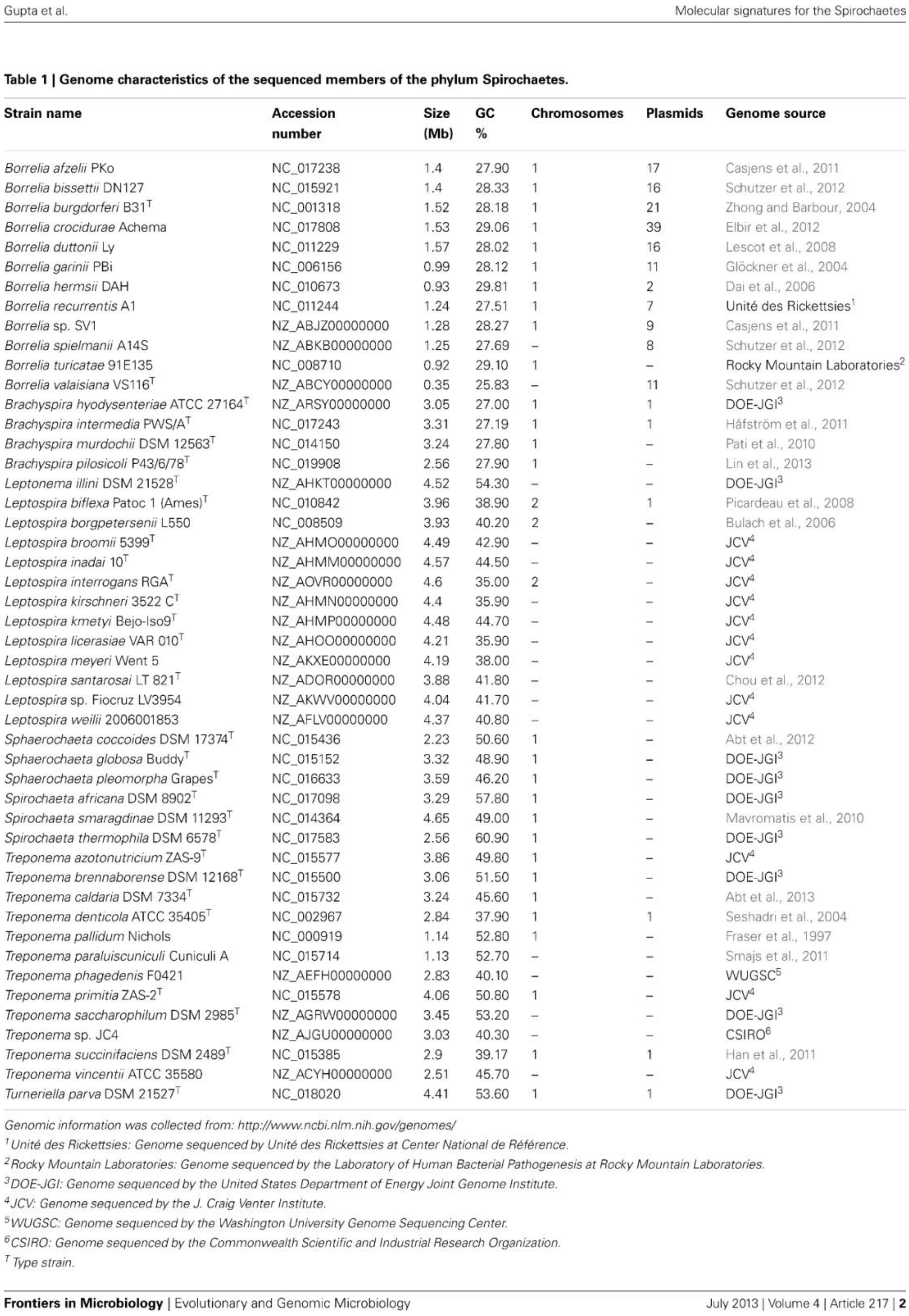


Table 1 | Genome characteristics of the sequenced members of the phylum Spirochaetes.

Strain name	Accession number	Size (Mb)	GC %	Chromosomes	Plasmids	Genome source
Borrelia afzelii PKo	NC_017238	1.4	27.90	1	17	Casjens et al., 2011
Borrelia bissettii DN127	NC_015921	1.4	28.33	1	16	Schutzer et al., 2012
Borrelia burgdorferi B31[T]	NC_001318	1.52	28.18	1	21	Zhong and Barbour, 2004
Borrelia crocidurae Achema	NC_017808	1.53	29.06	1	39	Elbir et al., 2012
Borrelia duttonii Ly	NC_011229	1.57	28.02	1	16	Lescot et al., 2008
Borrelia garinii PBi	NC_006156	0.99	28.12	1	11	Glöckner et al., 2004
Borrelia hermsii DAH	NC_010673	0.93	29.81	1	2	Dai et al., 2006
Borrelia recurrentis A1	NC_011244	1.24	27.51	1	7	Unité des Rickettsies[1]
Borrelia sp. SV1	NZ_ABJZ00000000	1.28	28.27	1	9	Casjens et al., 2011
Borrelia spielmanii A14S	NZ_ABKB00000000	1.25	27.69	–	8	Schutzer et al., 2012
Borrelia turicatae 91E135	NC_008710	0.92	29.10	1	–	Rocky Mountain Laboratories[2]
Borrelia valaisiana VS116[T]	NZ_ABCY00000000	0.35	25.83	–	11	Schutzer et al., 2012
Brachyspira hyodysenteriae ATCC 27164[T]	NZ_ARSY00000000	3.05	27.00	1	1	DOE-JGI[3]
Brachyspira intermedia PWS/A[T]	NC_017243	3.31	27.19	1	1	Håfström et al., 2011
Brachyspira murdochii DSM 12563[T]	NC_014150	3.24	27.80	1	–	Pati et al., 2010
Brachyspira pilosicoli P43/6/78[T]	NC_019908	2.56	27.90	1	–	Lin et al., 2013
Leptonema illini DSM 21528[T]	NZ_AHKT00000000	4.52	54.30	–	–	DOE-JGI[3]
Leptospira biflexa Patoc 1 (Ames)[T]	NC_010842	3.96	38.90	2	1	Picardeau et al., 2008
Leptospira borgpetersenii L550	NC_008509	3.93	40.20	2	–	Bulach et al., 2006
Leptospira broomii 5399[T]	NZ_AHMO00000000	4.49	42.90	–	–	JCV[4]
Leptospira inadai 10[T]	NZ_AHMM00000000	4.57	44.50	–	–	JCV[4]
Leptospira interrogans RGA[T]	NZ_AOVR00000000	4.6	35.00	2	–	JCV[4]
Leptospira kirschneri 3522 C[T]	NZ_AHMN00000000	4.4	35.90	–	–	JCV[4]
Leptospira kmetyi Bejo-Iso9[T]	NZ_AHMP00000000	4.48	44.70	–	–	JCV[4]
Leptospira licerasiae VAR 010[T]	NZ_AHOO00000000	4.21	35.90	–	–	JCV[4]
Leptospira meyeri Went 5	NZ_AKXE00000000	4.19	38.00	–	–	JCV[4]
Leptospira santarosai LT 821[T]	NZ_ADOR00000000	3.88	41.80	–	–	Chou et al., 2012
Leptospira sp. Fiocruz LV3954	NZ_AKWV00000000	4.04	41.70	–	–	JCV[4]
Leptospira weilii 2006001853	NZ_AFLV00000000	4.37	40.80	–	–	JCV[4]
Sphaerochaeta coccoides DSM 17374[T]	NC_015436	2.23	50.60	1	–	Abt et al., 2012
Sphaerochaeta globosa Buddy[T]	NC_015152	3.32	48.90	1	–	DOE-JGI[3]
Sphaerochaeta pleomorpha Grapes[T]	NC_016633	3.59	46.20	1	–	DOE-JGI[3]
Spirochaeta africana DSM 8902[T]	NC_017098	3.29	57.80	1	–	DOE-JGI[3]
Spirochaeta smaragdinae DSM 11293[T]	NC_014364	4.65	49.00	1	–	Mavromatis et al., 2010
Spirochaeta thermophila DSM 6578[T]	NC_017583	2.56	60.90	1	–	DOE-JGI[3]
Treponema azotonutricium ZAS-9[T]	NC_015577	3.86	49.80	1	–	JCV[4]
Treponema brennaborense DSM 12168[T]	NC_015500	3.06	51.50	1	–	DOE-JGI[3]
Treponema caldaria DSM 7334[T]	NC_015732	3.24	45.60	1	–	Abt et al., 2013
Treponema denticola ATCC 35405[T]	NC_002967	2.84	37.90	1	1	Seshadri et al., 2004
Treponema pallidum Nichols	NC_000919	1.14	52.80	1	–	Fraser et al., 1997
Treponema paraluiscuniculi Cuniculi A	NC_015714	1.13	52.70	–	–	Smajs et al., 2011
Treponema phagedenis F0421	NZ_AEFH00000000	2.83	40.10	–	–	WUGSC[5]
Treponema primitia ZAS-2[T]	NC_015578	4.06	50.80	1	–	JCV[4]
Treponema saccharophilum DSM 2985[T]	NZ_AGRW00000000	3.45	53.20	–	–	DOE-JGI[3]
Treponema sp. JC4	NZ_AJGU00000000	3.03	40.30	–	–	CSIRO[6]
Treponema succinifaciens DSM 2489[T]	NC_015385	2.9	39.17	1	1	Han et al., 2011
Treponema vincentii ATCC 35580	NZ_ACYH00000000	2.51	45.70	–	–	JCV[4]
Turneriella parva DSM 21527[T]	NC_018020	4.41	53.60	1	1	DOE-JGI[3]

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

[1] *Unité des Rickettsies: Genome sequenced by Unité des Rickettsies at Center National de Référence.*

[2] *Rocky Mountain Laboratories: Genome sequenced by the Laboratory of Human Bacterial Pathogenesis at Rocky Mountain Laboratories.*

[3] *DOE-JGI: Genome sequenced by the United States Department of Energy Joint Genome Institute.*

[4] *JCV: Genome sequenced by the J. Craig Venter Institute.*

[5] *WUGSC: Genome sequenced by the Washington University Genome Sequencing Center.*

[6] *CSIRO: Genome sequenced by the Commonwealth Scientific and Industrial Research Organization.*

[T] *Type strain.*

grown in pure or mixed culture and their phylogeny is based largely on analyses of morphological characteristics (Bermudes et al., 1988). *Treponema pallidum* subspecies *pallidum* is the causative agent of syphilis, a sexually transmitted disease which affects at least 25 million adults worldwide (Gerbase et al., 1998). Other members of the genus *Treponema* are responsible for diseases such bejel, yaws, and pinta and play important role in periodontal diseases (Ellen and Galimanas, 2005; Visser and Ellen, 2011; Smajs et al., 2012). Members of the genus *Borrelia*, namely *Borrelia burgdorferi s* and *Borrelia recurrentis*, are important human pathogens that cause Lyme disease and relapsing fever, respectively (Dworkin et al., 2008; Nau et al., 2009; Cutler, 2010). *Leptospira* and *Brachyspira*, are members of the families *Leptospiraceae* and *Brachyspiraceae*, and causative agents of the diseases leptospirosis and intestinal spirochaetosis, respectively (Adler and de la Peña Moctezuma, 2010; Anthony et al., 2013; Euzéby, 2013).

Despite the importance of species of the phylum Spirochaetes in causing many important human diseases, the evolutionary relationship of species within this phylum remains poorly understood and no distinguishing molecular features are known that are specific for all members of the different families (Olsen et al., 2000; Paster and Dewhirst, 2000; Paster, 2011a). The availability of genome sequences provides a valuable resource to identify/discover novel molecular markers that are helpful in these regards and to gain insights into their evolutionary relationships. Genomes from 48 species covering the three main families of the phylum Spirochaetes are now available in the NCBI database (**Table 1**) (NCBI, 2013). The availability of genome sequences allows for the use of comparative genomic approaches to identify molecular markers that are specific for different bacterial taxa at various taxonomic levels. Using genomic sequences, one useful approach pioneered by our lab involves the discovery of Conserved Signature insertions/deletions (i.e., Indels) or CSIs present in protein sequences that are specific for different groups of organisms. Due to the specificity of these CSIs for particular groups/taxa of species, they provide valuable molecular markers of common evolutionary descent (i.e., synapomorphies) for identification and demarcation of different phylogenetic/taxonomic clades of organisms in molecular terms. Additionally, based upon the presence or absence of these CSIs in outgroup species, it is possible to infer whether the observed genetic change is an insert or a deletion and a rooted phylogenetic relationship among different groups can be derived (Baldauf and Palmer, 1993; Gupta, 1998; Griffiths and Gupta, 2004; Gao and Gupta, 2012a).

In this work, we report the results of comparative analyses on protein sequences for the phylum Spirochaetes to identify molecular markers (CSIs) that are specific for the species from the phylum and its subgroups, or those that provide information regarding interrelationships among them. These studies have led to identification of 38 CSIs providing novel molecular markers for the species from the phylum and clarifying their evolutionary relationships. Additionally, we have also constructed a phylogenetic tree for all genome sequenced members of the phylum Spirochaetes based upon concatenated sequences for 22 conserved proteins. The inferences from different identified CSIs are strongly supported by the branching pattern of species in the phylogenetic tree indicating that the identified CSIs provide reliable molecular markers for the indicated groups of Spirochaetes.

METHODS

PHYLOGENETIC SEQUENCE ANALYSIS

Phylogenetic analysis was performed on a concatenated sequence alignment of 22 highly conserved proteins (viz. UvrD, GyrA, GyrB, RpoB, RpoC, EF-G, EF-Tu, RecA, ArgRS, IleRS, ThrRS, TrpRS, SecY, DnaK, and ribosomal proteins L2, L5, S2, S3, and S9) which have been widely used for phylogenetic analysis (Harris et al., 2003; Gao and Gupta, 2012a). Sequences for these proteins were obtained from the NCBI database for representative strains of all the sequenced Spirochaetes species (**Table 1**) and *Thermosynechococcus elongatus* and *Nostoc flagelliforme* which were used to root the tree. Multiple sequence alignments for these proteins were created using Clustal_X 1.83 (Jeanmougin et al., 1998) and concatenated into a single alignment file. Poorly aligned regions from this alignment file were removed using Gblocks 0.91 b (Castresana, 2000). The resulting alignment, which contained 7411 aligned amino acids, was used for phylogenetic analysis. The maximum likelihood (ML) and neighbor joining (NJ) trees based on 100 bootstrap replicates of this alignment were constructed using MEGA 5.1 (Tamura et al., 2011) employing the Whelan and Goldman (Whelan and Goldman, 2001) and Jones-Taylor-Thornton (Jones et al., 1992) substitution models, respectively.

A 16S rRNA gene sequence tree was also created for 107 sequences that included representative species for all 11 cultured Spirochaetes genera. 16S rRNA gene sequences larger than 1200 bp were obtained for all type species classified under the phylum Spirochaetes in release 114 of the SILVA database (Quast et al., 2013). Information for these sequences is provided in Supplemental Table 1. A ML tree based on these sequences was created using 100 bootstrap replicates of the 16S rRNA sequence alignments in MEGA 5.1 (Tamura et al., 2011) employing the General Time-Reversible (Tavaré, 1986) substitution model.

IDENTIFICATION OF MOLECULAR MARKERS (CSIs)

To identify CSIs that are commonly shared by different groups of Spirochaetes, BLASTp searches (Altschul et al., 1997) were performed on each protein in the genome of *Treponema pallidum* subspecies *pallidum* strain Nichols. These searches were performed using the default BLAST parameters against all available sequences in the GenBank non-redundant database. For those proteins for whom high scoring homologs (E-values $< 1e^{-20}$) were present in other species from the phylum Spirochaetes and some other bacterial groups multiple sequence alignments were created using the Clustal_X 1.83 program (Jeanmougin et al., 1998). These alignments were visually inspected for the presence of insertions or deletions that were flanked on both sides by at least 4–5 conserved amino acid residues in the neighboring 30–40 amino acids. Indels that were not flanked by conserved regions were not further considered, as they do not provide useful molecular markers (Gupta, 1998; Gao and Gupta, 2012a; Adeolu and Gupta, 2013). The specificity of potentially useful indels for members of the Spirochaetes was further evaluated by carrying out detailed Blastp searches on short sequence segments containing the indel and the flanking conserved regions (60–100 amino acids long). To ensure that the identified signatures are only present in the Spirochaetes homologs, a minimum of 250 blast hits with the highest similarity to the query sequence were examined for the presence or absence of these CSIs. In this work, we report the results of only those CSIs that are specific for different groups of Spirochaetes and where similar CSIs were not observed in any other bacteria in the top 250 blast hits. The sequence alignment files presented here contain sequence information for all sequenced genera within Spirochaetes. However, due to size restraints, different strains and/or species of the sequenced genera are not shown as they all exhibited similar patterns.

RESULTS

GENOMIC CHARACTERISTICS OF THE SEQUENCED SPIROCHAETES

There are currently 48 genome sequenced species of Spirochaetes. **Table 1** lists some characteristics of representative strains for all Spirochaetes species that have been completely sequenced. The genome sizes of these species of Spirochaetes showed a large amount of variation, ranging from 0.92 to 4.7 Mb in length. The G + C content of these species also showed a large amount of variation, ranging from 25.8 to 60.9%. The members of the phylum Spirochaetes also exhibited a large amount of variation in genome structure. The genome structure of members of genus *Borrelia* is one of the most unique among prokaryotes (Chaconas, 2005; Chaconas and Kobryn, 2010). The *Borrelia* genome consists of 6–24 DNA segments, including a linear chromosome about 900 kb in length which is accompanied by multiple essential linear and circular plasmids ranging from 5 to 220 kb in length (Chaconas and Kobryn, 2010). Linear chromosomes and plasmids terminated by covalently closed hairpin telomers are particularly uncommon genomic features among prokaryotes and are only found in the genomes of the *Borrelia* species and the species *Agrobacterium tumefaciens* (Goodner et al., 2001; Kobryn, 2007; Chaconas and Kobryn, 2010). Members of the genus *Leptospira* also have an unusual genome structure consisting of two circular chromosomes, a big chromosome about 3.6–4.2 Mb in length and a smaller chromosome about 300 kb in length (Ren et al., 2003; Picardeau et al., 2008).

PHYLOGENETIC ANALYSES OF THE SEQUENCED SPIROCHAETES

The branching order of species within the phylum Spirochaetes has primarily been determined using 16S rRNA sequence based phylogenetic trees (Paster and Dewhirst, 2000; Paster, 2011a). In these trees, the four families with the phylum branch into distinct monophyletic clades separated by long branches. However, the interrelationships of members of the family *Spirochaetaceae* are not reliably resolved (Paster, 2011b) (**Figure 2**). Phylogenetic trees derived from large numbers of conserved genes/proteins provide greater resolving power than those based on any single gene or protein (Rokas et al., 2003; Ciccarelli et al., 2006; Wu et al., 2009; Gao and Gupta, 2012a). In this study, we have constructed phylogenetic trees of the genome sequenced members of the phylum Spirochaetes listed in **Table 1** using 22 conserved housekeeping and ribosomal proteins. The trees were constructed using both the NJ and ML methodologies and branching patterns generated by both methodologies were highly similar (**Figure 1**).

In the concatenated protein trees, which are rooted using the species *T. elongatus* and *N. flagelliforme*, the members of the three sequenced families of Spirochaetes (viz. *Spirochaetaceae*, *Brachyspiraceae*, and *Leptospiraceae*) formed three distinct monophyletic clades (**Figure 1**). Additionally, the branching order of members of the family *Spirochaetaceae* is well-resolved in the concatenated protein trees. Within the *Spirochaetaceae* clade, the genera *Treponema*, *Spirochaeta*, and *Sphaerochaeta* formed a well-supported monophyletic clade separated from the members of the genus *Borrelia* by a long branch. The *Treponema*, *Spirochaeta*, and *Sphaerochaeta* clade exhibited a large amount of diversity and consisted of a number of strongly supported subclades. Members of each of the sequenced genera within Spirochaetes formed monophyletic clusters with the exception of the genus *Spirochaeta*, where *Spirochaeta smaragdinae* branched with the genus *Sphaerochaeta*. Another *Spirochaeta* species, *S. caldaria,* which branched within the *Treponema* has recently been reclassified as *Treponema caldaria* (Abt et al., 2013). The remaining *Spirochaeta* (viz. *S. thermophila* and *S. africana*) branched deeply within the *Treponema*, *Spirochaeta*, and *Sphaerochaeta* clade (**Figure 1**). The monophyletic clade containing all the members of the genus *Borrelia* consisted of two highly distinct subclades, one containing *Borrelia burgdorferi*, and related species of *Borrelia* and the other containing *Borrelia recurrentis* related species.

The 16S rRNA tree shown in **Figure 2** includes all of the members included in the concatenated protein tree as well as other cultured members of the phylum Spirochaetes which have yet to be genome sequenced. The branching patterns in the 16S rRNA phylogenetic tree were similar to those observed in the concatenated protein tree; all families within the phylum branched distinctly. Within the cluster consisting of members of the family *Spirochaetaceae* the genera *Treponema*, *Sphaerochaeta*, and most members of the genus *Spirochaeta* formed a monophyletic clade. The genera *Borrelia* and *Cristispira* also formed a well-supported monophyletic clade that was distinct from the genera *Treponema*, *Spirochaeta*, and *Sphaerochaeta* within the *Spirochaetaceae* clade. The different sequenced members of the genus *Borrelia* also formed two distinct clusters in the 16S rRNA tree (**Figure 2**).

CSI SPECIFIC FOR THE PHYLUM SPIROCHAETES

CSIs that are restricted to a group of related species are a novel class of molecular marker with high utility for evolutionary studies (Gupta, 1998; Rokas et al., 2003; Gupta, 2009; Gao and Gupta, 2012a). The co-occurrence of multiple CSIs in different species may be due to shared evolutionary history, convergent evolution, lateral gene transfer. However, the unique shared presence of multiple CSIs in a diverse range by a related group of species is most parsimoniously explained by the occurrence of the rare genetic changes that resulted in these CSIs in a common ancestor of the group, followed by vertical transmission of these CSIs to various descendant species (Gupta, 1998; Rokas and Holland, 2000; Gogarten et al., 2002; Gupta and Griffiths, 2002; Gao and Gupta, 2012a). Hence, these CSIs represent molecular synapomorphies of common evolutionary descent and they provide useful markers for identifying different groups of organisms in molecular terms and for understanding their interrelationships independently of

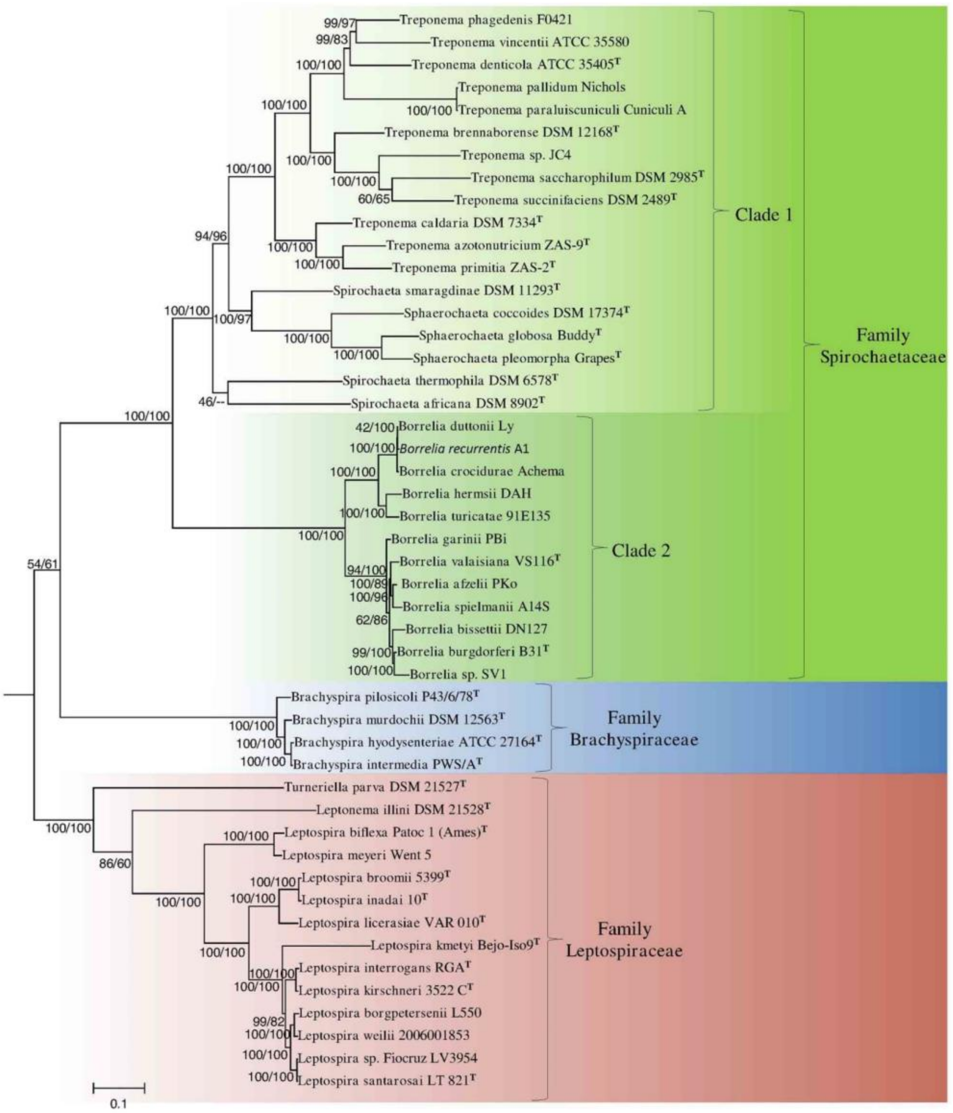

99/97
Treponema phagedenis F0421
99/83
Treponema vincentii ATCC 35580
100/100
Treponema denticola ATCC 35405^T
Treponema pallidum Nichols
100/100
100/100
Treponema paraluiscuniculi Cuniculi A
Treponema brennaborense DSM 12168^T
100/100
Treponema sp. JC4
100/100
100/100
Treponema saccharophilum DSM 2985^T
60/65
Treponema succinifaciens DSM 2489^T
Treponema caldaria DSM 7334^T
94/96
100/100
Treponema azotonutricium ZAS-9^T
100/100
Treponema primitia ZAS-2^T
Spirochaeta smaragdinae DSM 11293^T
100/100
100/97
Sphaerochaeta coccoides DSM 17374^T
100/100
Sphaerochaeta globosa Buddy^T
100/100
Sphaerochaeta pleomorpha Grapes^T
Spirochaeta thermophila DSM 6578^T
46/--
Spirochaeta africana DSM 8902^T
Clade 1
Family
Spirochaetaceae
100/100
42/100
Borrelia duttonii Ly
100/100
Borrelia recurrentis A1
100/100
Borrelia crocidurae Achema
Borrelia hermsii DAH
100/100
Borrelia turicatae 91E135
100/100
Borrelia garinii PBi
54/61
94/100
Borrelia valaisiana VS116^T
100/89
Borrelia afzelii PKo
100/96
Borrelia spielmanii A14S
62/86
Borrelia bissettii DN127
99/100
Borrelia burgdorferi B31^T
100/100
Borrelia sp. SV1
Clade 2
Brachyspira pilosicoli P43/6/78^T
100/100
Brachyspira murdochii DSM 12563^T
100/100
Brachyspira hyodysenteriae ATCC 27164^T
100/100
Brachyspira intermedia PWS/A^T
Family
Brachyspiraceae
Turneriella parva DSM 21527^T
100/100
Leptonema illini DSM 21528^T
100/100
Leptospira biflexa Patoc 1 (Ames)^T
86/60
Leptospira meyeri Went 5
100/100
Leptospira broomii 5399^T
100/100
100/100
Leptospira inadai 10^T
Leptospira licerasiae VAR 010^T
100/100
Leptospira kmetyi Bejo-Iso9^T
100/100
Leptospira interrogans RGA^T
100/100
Leptospira kirschneri 3522 C^T
99/82
Leptospira borgpetersenii L550
100/100
Leptospira weilii 2006001853
100/100
Leptospira sp. Fiocruz LV3954
100/100
Leptospira santarosai LT 821^T
Family
Leptospiraceae
0.1

FIGURE 1 | A phylogenetic tree of genome sequenced members of the phylum Spirochaetes based on the concatenated amino acid sequences of 22 conserved proteins. The tree shown is a maximum-likelihood (ML) distance tree. Bootstrap values are shown at branch nodes for both maximum-likelihood and neighbor-joining tree construction methods as ML/NJ. The different sequenced families and two main clades of the family *Spirochaetaceae* supported by the tree are marked. The letter ^{T} refers to the type strain of the species.

phylogenetic trees (Gupta, 1998; Gupta and Griffiths, 2002; Gao and Gupta, 2012a,b). The CSI-based approach has recently been used to propose important taxonomic changes for a number of groups of bacteria (viz. Chloroflexi, *Coriobacateriia*, *Neisseriales*, and *Bacillus*) at different taxonomic ranks (Gupta et al., 2012, 2013; Adeolu and Gupta, 2013; Bhandari et al., 2013). In the present work, we have completed comprehensive genomic analyses to identify CSIs that are primarily restricted to the phylum

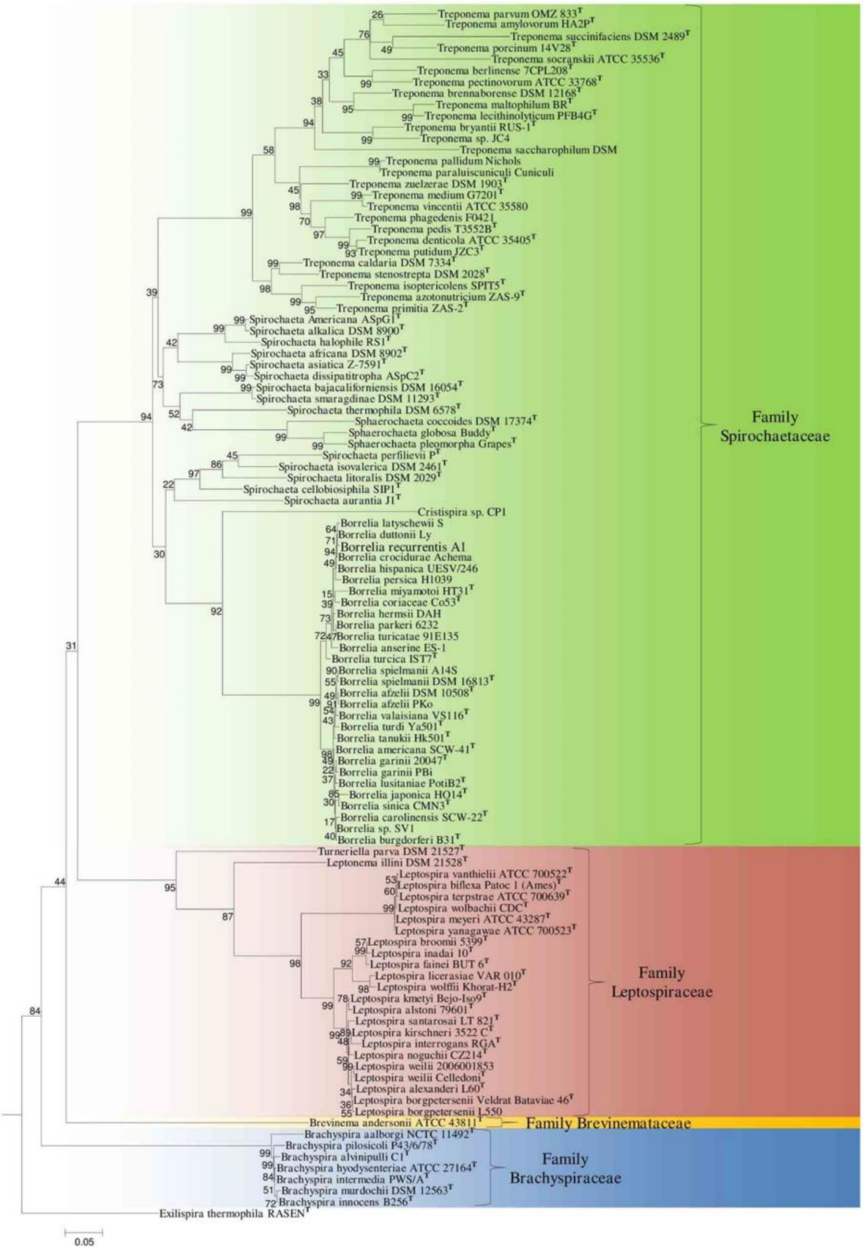
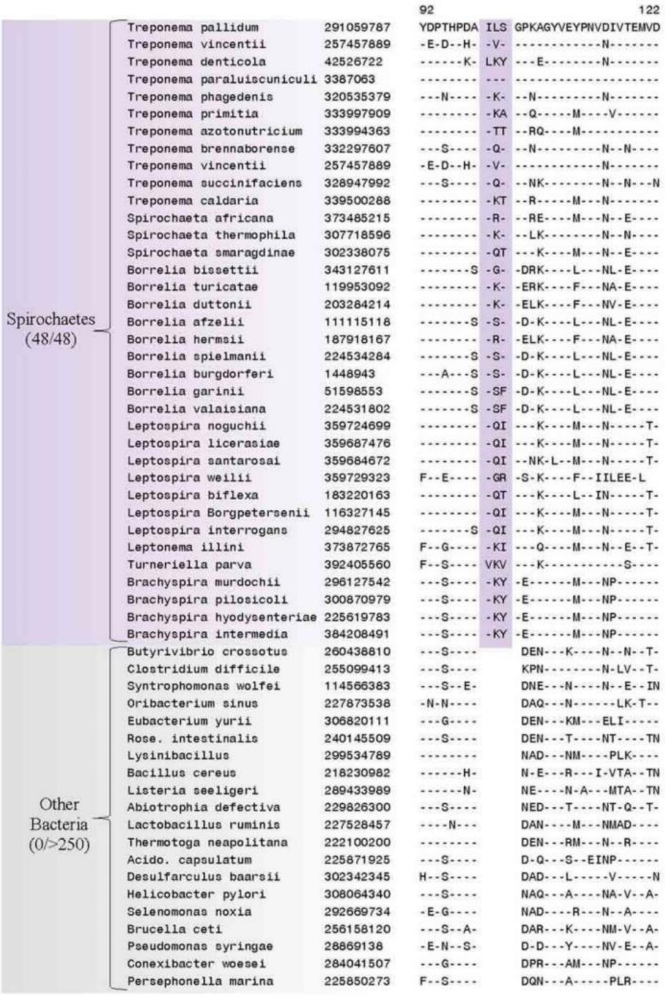

Treponema parvum OMZ 833^{T}
Treponema amylovorum HA2P^{T}
Treponema succinifaciens DSM 2489^{T}
Treponema porcinum 14V28^{T}
Treponema socranskii ATCC 35536^{T}
Treponema berlinense 7CPL208^{T}
Treponema pectinovorum ATCC 33768^{T}
Treponema brennaborense DSM 12168^{T}
Treponema maltophilum BR^{T}
Treponema lecithinolyticum PFB4G^{T}
Treponema bryantii RUS-1^{T}
Treponema sp. JC4
Treponema saccharophilum DSM
Treponema pallidum Nichols
Treponema paraluiscuniculi Cuniculi
Treponema zuelzerae DSM 1903^{T}
Treponema medium G7201^{T}
Treponema vincentii ATCC 35580
Treponema phagedenis F0421
Treponema pedis T3552B^{T}
Treponema denticola ATCC 35405^{T}
Treponema putidum JZC3^{T}
Treponema caldaria DSM 7334^{T}
Treponema stenostrepta DSM 2028^{T}
Treponema isoptericolens SPIT5^{T}
Treponema azotonutricium ZAS-9^{T}
Treponema primitia ZAS-2^{T}
Spirochaeta Americana ASpG1^{T}
Spirochaeta alkalica DSM 8900^{T}
Spirochaeta halophile RS1^{T}
Spirochaeta africana DSM 8902^{T}
Spirochaeta asiatica Z-7591^{T}
Spirochaeta dissipatitropha ASpC2^{T}
Spirochaeta bajacaliforniensis DSM 16054^{T}
Spirochaeta smaragdinae DSM 11293^{T}
Spirochaeta thermophila DSM 6578^{T}
Sphaerochaeta coccoides DSM 17374^{T}
Sphaerochaeta globosa Buddy^{T}
Sphaerochaeta pleomorpha Grapes^{T}
Spirochaeta perfilievii P^{T}
Spirochaeta isovalerica DSM 2461^{T}
Spirochaeta litoralis DSM 2029^{T}
Spirochaeta cellobiosiphila SIP1^{T}
Spirochaeta aurantia J1^{T}
Cristispira sp. CP1
Borrelia latyschewii S
Borrelia duttonii Ly
Borrelia recurrentis A1
Borrelia crocidurae Achema
Borrelia hispanica UESV/246
Borrelia persica H1039
Borrelia miyamotoi HT31^{T}
Borrelia coriaceae Co53^{T}
Borrelia hermsii DAH
Borrelia parkeri 6232
Borrelia turicatae 91E135
Borrelia anserine ES-1
Borrelia turcica IST7^{T}
Borrelia spielmanii A14S
Borrelia spielmanii DSM 16813^{T}
Borrelia afzelii DSM 10508^{T}
Borrelia afzelii PKo
Borrelia valaisiana VS116^{T}
Borrelia turdi Ya501^{T}
Borrelia tanukii Hk501^{T}
Borrelia americana SCW-41^{T}
Borrelia garinii 20047^{T}
Borrelia garinii PBi
Borrelia lusitaniae PotiB2^{T}
Borrelia japonica HO14^{T}
Borrelia sinica CMN3^{T}
Borrelia carolinensis SCW-22^{T}
Borrelia sp. SV1
Borrelia burgdorferi B31^{T}
Family Spirochaetaceae
Turneriella parva DSM 21527^{T}
Leptonema illini DSM 21528^{T}
Leptospira vanthielii ATCC 700522^{T}
Leptospira biflexa Patoc 1 (Ames)^{T}
Leptospira terpstrae ATCC 700639^{T}
Leptospira wolbachii CDC^{T}
Leptospira meyeri ATCC 43287^{T}
Leptospira yanagawae ATCC 700523^{T}
Leptospira broomii 5399^{T}
Leptospira inadai 10^{T}
Leptospira fainei BUT 6^{T}
Leptospira licerasiae VAR 010^{T}
Leptospira wolffii Khorat-H2^{T}
Leptospira kmetyi Bejo-Iso9^{T}
Leptospira alstoni 79601^{T}
Leptospira santarosai LT 821^{T}
Leptospira kirschneri 3522 C^{T}
Leptospira interrogans RGA^{T}
Leptospira noguchii CZ214^{T}
Leptospira weilii 2006001853
Leptospira weilii Celledoni^{T}
Leptospira alexanderi L60^{T}
Leptospira borgpetersenii Veldrat Bataviae 46^{T}
Leptospira borgpetersenii L550
Family Leptospiraceae
Brevinema andersonii ATCC 43811^{T}
Family Brevinemataceae
Brachyspira aalborgi NCTC 11492^{T}
Brachyspira pilosicoli P43/6/78^{T}
Brachyspira alvinipulli C1^{T}
Brachyspira hyodysenteriae ATCC 27164^{T}
Brachyspira intermedia PWS/A^{T}
Brachyspira murdochii DSM 12563^{T}
Brachyspira innocens B256^{T}
Family Brachyspiraceae
Exilispira thermophila RASEN^{T}
0.05

FIGURE 2 | A ML tree based on the 16S rRNA gene sequences of representative species from cultured genera within the phylum Spirochaetes. Bootstrap values are shown at branch nodes. The different families of the phylum Spirochaetes are marked. The letter^{T} refers to the type strain of the species. The accession numbers of the 16S rRNA gene sequences used in this analysis are provided in Supplemental Table 1.

Spirochaetes or its subgroups. Information regarding the species specificities of these CSIs and their evolutionary significances are discussed below.

Our analyses have identified 38 CSIs in diverse and important proteins that are specific for members of the Spirochaetes. One CSI has been identified that is specifically found in all of the sequenced members of the phylum Spirochaetes and not found in homologous proteins from any other bacterial species (in the top 250 Blast hits) (**Figure 3**). This CSI consists of a 3 amino acid (aa) insertion located in the flagellar basal-body rod protein FlgC, a component of the basal body which comprises a large portion of the flagella (Macnab, 2003). This CSI represents a unique molecular characteristic of the phylum Spirochaetes and may be related to


Spirochaetes
(48/48)
Other
Bacteria
(0/>250)

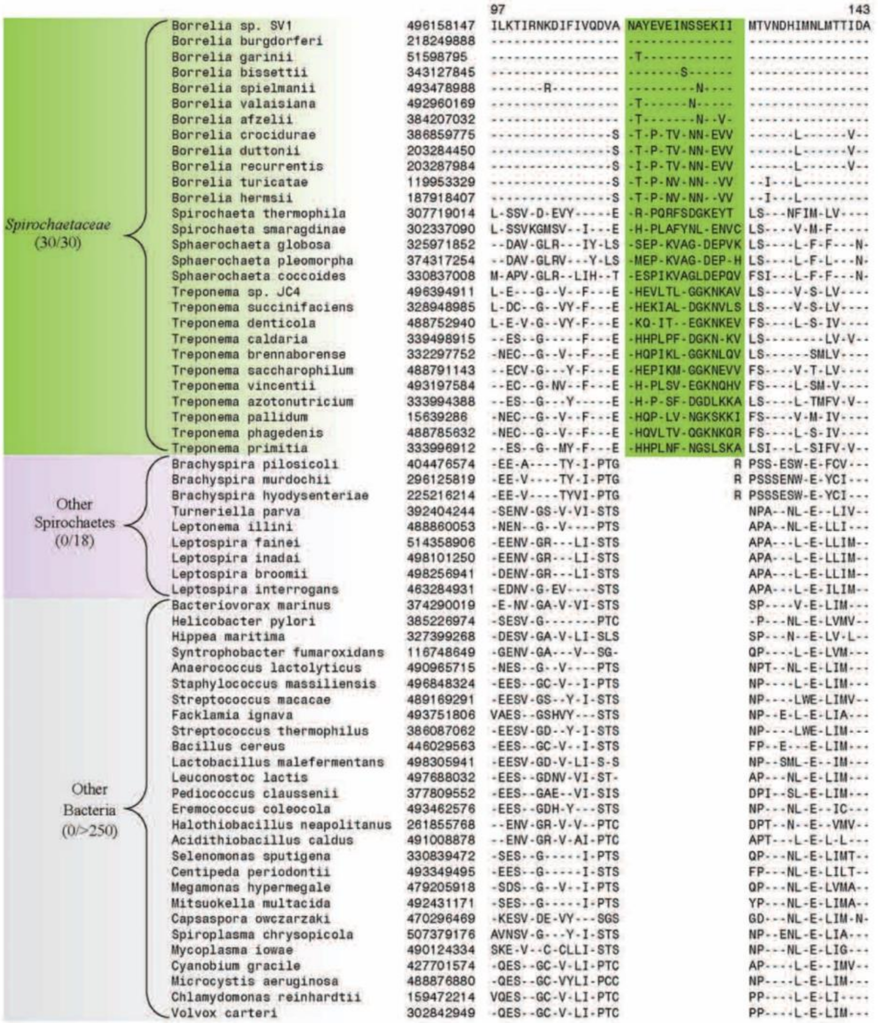
```
                                         92                       122
Treponema pallidum          291059787    YDPTHPDA ILS GPKAGYVEYPNVDIVTEMVD
Treponema vincentii         257457889    -E-D--H- -V- ------------N-------
Treponema denticola         42526722     ------K- LKY ---E--------N-------
Treponema paraluiscuniculi  3387063      -------- --- --------------------
Treponema phagedenis        320535379    ---N---- -K- --N---------N-------
Treponema primitia          333997909    -------- -KA --Q-----M----V------
Treponema azotonutricium    333994363    -------- -TT --RQ----M-----------
Treponema brennaborense     332297607    ---S---- -Q- --N---------N--N----
Treponema vincentii         257457889    -E-D--H- -V- ------------N-------
Treponema succinifaciens    328947992    ---S---- -Q- --NK--------N--N---N
Treponema caldaria          339500288    -------- -KT --R-----M---N-------
Spirochaeta africana        373485215    -------- -R- --RE----M---N--E----
Spirochaeta thermophila     307718596    -------- -K- --LK--------N--N----
Spirochaeta smaragdinae     302338075    -------- -QT ---K----M---N--E----
Borrelia bissettii          343127611    -------S -G- -DRK----L---NL-E----
Borrelia turicatae          119953092    -------- -K- -ERK----F---NA-E----
Borrelia duttonii           203284214    -------- -K- -ELK----F---NV-E----
Borrelia afzelii            111115118    -------S -S- -D-K----L---NL-E----
Borrelia hermsii            187918167    -------- -R- -ELK----F---NA-E----
Borrelia spielmanii         224534284    -------S -S- -D-K----L---NL-E----
Borrelia burgdorferi        1448943      ---A---S -S- -D-K----L---NL-E----
Borrelia garinii            51598553     -------S -SF -D-K----L---NL-E----
Borrelia valaisiana         224531802    -------S -SF -D-K----L---NL-E----
Leptospira noguchii         359724699    -------- -QI ---K----M---N-----T-
Leptospira licerasiae       359687476    -------- -QI ---K----M---N-----T-
Leptospira santarosai       359684672    -------- -QI --NK-L--M---N-----T-
Leptospira weilii           359729323    F--E---- -GR -S-K----F--IILEE-L
Leptospira biflexa          183220163    -------- -QT ---K----L--IN-----T-
Leptospira Borgpetersenii   116327145    -------- -QI ---K----M---N-----T-
Leptospira interrogans      294827625    -------S -QI ---K----M---N-----T-
Leptonema illini            373872765    F--G---- -KI ---Q----M---N--E--T-
Turneriella parva           392405560    F--S---- VKV ---K-----------S----
Brachyspira murdochii       296127542    ---S---- -KY -E------M---NP------
Brachyspira pilosicoli      300870979    ---S---- -KY -E------M---NP------
Brachyspira hyodysenteriae  225619783    ---S---- -KY -E------M---NP------
Brachyspira intermedia      384208491    ---S---- -KY -E------M---NP------
Butyrivibrio crossotus      260438810    ---S----      DEN---K----N--N--T-
Clostridium difficile       255099413    ---S----      KPN--------N-LV--T-
Syntrophomonas wolfei       114566383    ---S--E-      DNE---N----N--E--IN
Oribacterium sinus          227873538    -N-N----      DAQ---N------LK-T--
Eubacterium yurii           306820111    ---G----      DEN---KM---ELI-----
Rose. intestinalis          240145509    ---S----      DEN---T----NT----TN
Lysinibacillus              299534789    --------      NAD---NM----PLK----
Bacillus cereus             218230982    ------H-      N-E---R---I-VTA--TN
Listeria seeligeri          289433989    ------N-      NE----N-A---MTA--TN
Abiotrophia defectiva       229826300    ---S----      NED---T----NT-Q--T-
Lactobacillus ruminis       227528457    ----N---      DAN----M---NMAD----
Thermotoga neapolitana      222100200    --------      DEN---RM---N--R----
Acido. capsulatum           225871925    ---S----      D-Q---S--EINP------
Desulfarculus baarsii       302342345    H--S----      DAD---L-----V-----N
Helicobacter pylori         308064340    ---S----      NAQ---A----NA-V--A-
Selenomonas noxia           292669734    -E-G----      NAD----R---N--A----
Brucella ceti               256158120    ---S--A-      DAR---K----NM-V--A-
Pseudomonas syringae        28869138     -E-N--S-      D-D---Y----NV-E--A-
Conexibacter woesei         284041507    ---G----      DPR---AM---NP------
Persephonella marina        225850273    F--S----      DQN---A-----PLR----
```

FIGURE 3 | A partial sequence alignment of the flagellar basal-body rod protein FlgC, showing a CSI (boxed) that is uniquely present in all members of the phylum Spirochaetes. Sequence information for only a limited number of species from the Spirochaetes and other bacteria is shown here, but unless otherwise indicated similar CSIs were detected in all members of the indicated group and not detected in any other bacterial species in the top 250 Blastp hits. The dashes (–) in the alignments indicate identity with the residue in the top sequence. GenBank identification (GI) numbers for each sequence are indicated in the second column. Sequence homologs for this protein were not identified from members of the genus *Sphaerochaeta*.

the characteristic flagellar morphology shared by members of the phylum.

CSIs THAT ARE SPECIFIC FOR DIFFERENT FAMILIES OF SPIROCHAETES

Many of the CSIs identified by our analyses are specific for the different sequenced families within the phylum Spirochaetes (viz. *Spirochaetaceae*, *Brachyspiraceae*, and *Leptospiraceae*) allowing us to demarcate these families in clear molecular terms. Seven of the CSIs identified by our analyses are specific for the family *Spirochaetaceae*. One example of a CSI that is specific for the species from the family *Spirochaetaceae* is a 15 aa insertion in a highly conserved region of the protein phosphoribosylpyrophosphate synthetase, which is uniquely found in all members of the family *Spirochaetaceae* but not in any other sequenced bacterial groups (**Figure 4**). Sequence information for 6 other CSIs in diverse proteins (viz. Alanyl-tRNA synthetase,



```
                                          97                                                 143
Borrelia sp. SV1               496158147  ILKTIRNKDIFIVQDVA NAYEVEINSSEKII MTVNDHIMNLMTTIDA
Borrelia burgdorferi           218249888  ----------------- -------------- ----------------
Borrelia garinii               51598795   ----------------- -T------------ ----------------
Borrelia bissettii             343127845  ----------------- -------S------ ----------------
Borrelia spielmanii            493478988  -------R--------- ---------N---- ----------------
Borrelia valaisiana            492960169  ----------------- -T------N----- ----------------
Borrelia afzelii               384207032  ----------------- -T-------N--V- ----------------
Borrelia crocidurae            386859775  ----------------S -T-P-TV-NN-EVV ------L------V--
Borrelia duttonii              203284450  ----------------S -T-P-TV-NN-EVV ------L------V--
Borrelia recurrentis           203287984  ----------------S -I-P-TV-NN-EVV ------L------V--
Borrelia turicatae             119953329  ----------------S -T-P-NV-NN--VV --I---L---------
Borrelia hermsii               187918407  ----------------S -T-P-NV-NN--VV --I---L---------
Spirochaeta thermophila        307719014  L-SSV-D-EVY-----E -R-PQRFSDGKEYT LS---NFIM-LV----
Spirochaeta smaragdinae        302337090  L-SSVKGMSV--I---E -H-PLAFYNL-ENVC LS----V-M-F-----
Sphaerochaeta globosa          325971852  --DAV-GLR---IY-LS -SEP-KVAG-DEPVK LS----L-F-F---N-
Sphaerochaeta pleomorpha       374317254  --DAV-GLRV---Y-LS -MEP-KVAG-DEP-H LS----L-F-L---N-
Sphaerochaeta coccoides        330837008  M-APV-GLR--LIH--T -ESPIKVAGLDEPQV FSI---L-F-F---N-
Treponema sp. JC4              496394911  L-E---G--V--F---E -HEVLTL-GGKNKAV LS----V-S-LV----
Treponema succinifaciens       328948985  L-DC--G--VY-F---E -HEKIAL-DGKNVLS LS----V-S-LV----
Treponema denticola            488752940  L-E-V-G--VY-F---E -KQ-IT--EGKNKEV FS----L-S-IV----
Treponema caldaria             339498915  --ES--G-----F---E -HHPLPF-DGKN-KV LS--------LV-V--
Treponema brennaborense        332297752  -NEC--G--V--F---E -HQPIKL-GGKNLQV LS------SMLV----
Treponema saccharophilum       488791143  --ECV-G---Y-F---E -HEPIKM-GGKNEVV FS----V-T-LV----
Treponema vincentii            493197584  --EC--G-NV--F---E -H-PLSV-EGKNQHV FS----L-SM-V----
Treponema azotonutricium       333994388  --ES--G---Y-----E -H-P-SF-DGDLKKA LS----L-TMFV-V--
Treponema pallidum             15639286   -NEC--G--V--F---E -HQP-LV-NGKSKKI FS----V-M-IV----
Treponema phagedenis           488785632  -NEC--G--V--F---E -HQVLTV-QGKNKQR FS----L-S-IV----
Treponema primitia             333996912  --ES--G--MY-F---E -HHPLNF-NGSLSKA LSI---L-SIFV-V--
Brachyspira pilosicoli         404476574  -EE-A----TY-I-PTG               R PSS-ESW-E-FCV---
Brachyspira murdochii          296125819  -EE-V----TY-I-PTG               R PSSSENW-E-YCI---
Brachyspira hyodysenteriae     225216214  -EE-V----TYVI-PTG               R PSSSESW-E-YCI---
Turneriella parva              392404244  -SENV-GS-V-VI-STS                 NPA--NL-E--LIV--
Leptonema illini               488860053  -NEN--G--V----PTS                 APA--NL-E-LLI---
Leptospira fainei              514358906  -EENV-GR---LI-STS                 APA---L-E-LLIM--
Leptospira inadai              498101250  -EENV-GR---LI-STS                 APA---L-E-LLIM--
Leptospira broomii             498256941  -DENV-GR---LI-STS                 APA---L-E-LLIM--
Leptospira interrogans         463284931  -EDNV-G-EV----STS                 APA---L-E-ILIM--
Bacteriovorax marinus          374290019  -E-NV-GA-V-VI-STS                 SP----V-E-LIM---
Helicobacter pylori            385226974  -SESV-G-------PTC                 -P---NL-E-LVMV--
Hippea maritima                327399268  -DESV-GA-V-LI-SLS                 SP---N--E-LV-L--
Syntrophobacter fumaroxidans   116748649  -GENV-GA---V--SG-                 QP----L-E-LVM---
Anaerococcus lactolyticus      490965715  -NES--G--V----PTS                 NPT--NL-E-LIM---
Staphylococcus massiliensis    496848324  -EES--GC-V--I-PTS                 NP----L-E-LIM---
Streptococcus macacae          489169291  -EESV-GS--Y-I-STS                 NP---LWE-LIMV--
Facklamia ignava               493751806  VAES--GSHVY---STS                 NP--E-L-E-LIA---
Streptococcus thermophilus     386087062  -EESV-GD--Y-I-STS                 NP---LWE-LIM---
Bacillus cereus                446029563  -EES--GC-V--I-STS                 FP--E---E-LIM---
Lactobacillus malefermentans   498305941  -EESV-GD-V-LI-S-S                 NP--SML-E--IM---
Leuconostoc lactis             497688032  -EES--GDNV-VI-ST-                 AP---NL-E-LIM---
Pediococcus claussenii         377809552  -EES--GAE--VI-SIS                 DPI--SL-E-LIM---
Eremococcus coleocola          493462576  -EES--GDH-Y---STS                 NP---NL-E--IC---
Halothiobacillus neapolitanus  261855768  --ENV-GR-V-V--PTC                 DPT--N--E--VMV--
Acidithiobacillus caldus       491008878  --ENV-GR-V-AI-PTC                 APT---L-E-L-L---
Selenomonas sputigena          330839472  -SES--G-----I-PTS                 QP---NL-E-LIMT--
Centipeda periodontii          493349495  -EES--G-----I-STS                 FP---NL-E-LILT--
Megamonas hypermegale          479205918  -SDS--G--V--I-PTS                 QP---NL-E-LVMA--
Mitsuokella multacida          492431171  -SES--G-----I-PTS                 YP---NL-E-LIMA--
Capsaspora owczarzaki          470296469  -KESV-DE-VY---SGS                 GD---NL-E-LIM-N-
Spiroplasma chrysopicola       507379176  AVNSV-G---Y-I-STS                 NP--ENL-E-LIA---
Mycoplasma iowae               490124334  SKE-V--C-CLLI-STS                 NP---NL-E-LIG---
Cyanobium gracile              427701574  -QES--GC-V-LI-PTC                 AP----L-E--IMV--
Microcystis aeruginosa         488876880  -QES--GC-VYLI-PCC                 NP----L-E-LIM---
Chlamydomonas reinhardtii      159472214  VQES--GC-V-LI-PTC                 PP----L-E-LI----
Volvox carteri                 302842949  -QES--GC-V-LI-PTC                 PP----L-E-LIM---
```

FIGURE 4 | A partial sequence alignment of the protein alanyl-tRNA synthetase showing a two amino acid insertion (boxed) identified in homologs from the family *Spirochaetaceae*, but not found in the sequence homologs of any other sequenced bacteria. Sequence information for other *Spirochaetaceae* specific CSIs is presented in Supplemental Figures 3–6 and summarized in **Table 2**.

phosphoribosylpyrophosphate synthetase, preprotein translocase SecY, peptide chain release factor 2, DNA mismatch repair protein MutS, and DNA mismatch repair protein MutL) that are also specifically present in members of the family *Spirochaetaceae* is presented in Supplementary Figures 1–6 and some of their characteristics are summarized in **Table 2**.

Our analyses have also identified 6 CSIs in diverse proteins that are specifically found in members of the family *Brachyspiraceae* and absent in all other bacterial groups. One of these *Brachyspiraceae*-specific CSIs, a 1 aa insertion, is present in the flagellar hook-associated protein FlgK, a protein involved in flagellar hook morphogenesis (**Figure 5A**) (Homma et al., 1990). Another *Brachyspiraceae*-specific CSI, a 1 aa insertion, is found in a highly conserved region of DNA polymerase I (**Figure 5B**). These proteins represent highly conserved and essential components of members of the family *Brachyspiraceae* which contain conserved molecular changes not found in any other sequenced bacterial group. Sequence information for 4 other CSIs in three other proteins (viz. valyl-tRNA synthetase, ATP-dependent protease La, and glutamyl-tRNA amidotransferase subunit B) that are also specifically present in members of the family *Brachyspiraceae* is presented in Supplemental Figures 7–10 and some of their characteristics are summarized in **Table 3**.

We have also identified 5 CSIs that are uniquely present in members of the family *Leptospiraceae*. Two examples of such CSIs are shown in **Figure 6**. The first of these CSIs, an 8 aa insertion in the 50S ribosomal protein L14, is shown in **Figure 6A**, and the other CSI, a 4 aa insert in alanyl-tRNA synthetase, is shown in **Figure 6B**. Both of these CSIs are found in members of the the family *Leptospiraceae* and absent in every other sequenced bacterial group. Sequence information for 4 other CSIs in diverse proteins (viz. 30S Ribosomal protein S2, flagellar basal-body rod protein FlgG, and flagellar filament core protein FlaB) that are also specifically present in members of the family *Leptospiraceae* is presented in Supplemental Figures 11–14 and some of their characteristics are summarized in **Table 4**.

CSIs DISTINGUISHING TWO CLADES WITHIN THE FAMILY *Spirochaetaceae*

In addition to the numerous CSIs identified in our analyses for the sequenced families within the phylum Spirochaetes, we have also identified a number of CSIs that elucidate the relationship of the genera within the family *Spirochaetaceae*. Three of the identified CSIs are uniquely shared by the genera *Treponema*, *Spirochaeta*, and *Sphaerochaeta*. One example of a CSI specific to these three genera, a 1 aa deletion in the 30S ribosomal protein S13, a component of the protein translation complex, is shown in **Figure 7A**. Sequence information for 2 other CSIs specifically found in these three genera is provided in **Table 5** and Supplemental Figures 14, 15. An additional 16 CSIs were uniquely shared by members of the genus *Borrelia*. One example of a CSI consisting of a 6 aa insertion in the glycolysis related protein, phosphofructokinase, that is specific to the members of the genus *Borrelia* is shown in **Figure 7B**. Fifteen other CSIs were also specifically found in members of the genus *Borrelia* and information for them is presented in **Table 5** and Supplemental Figures 16–30.

DISCUSSION

The phylum Spirochaetes is currently distinguished from other bacteria on the basis of both branching in 16S rRNA sequence based phylogenies and the presence of the endoflagella that characterizes the phylum (Paster, 2011a; Euzéby, 2013). Apart from the presence of endoflagella, no reliable morphological, biochemical, or molecular characteristics are known that are specifically shared by all members of the phylum. Additionally, the phylum contains four divergent lineages, contained within a single class/order, that are demarcated largely on the basis of 16S rRNA sequence based phylogenies (Paster, 2011a). In this work, we have utilized comparative genomic techniques to identify large numbers of novel molecular signatures (CSIs) that are distinctive characteristics of either all members of the phylum Spirochaetes or for its different subgroups at multiple phylogenetic levels and which can be used to demarcate these groups in more definitive molecular terms. A summary diagram depicting the species distribution of the identified CSIs is shown in **Figure 8**.

The phylum Spirochaetes is rare in having a defining morphological characteristic, the endoflagella, which correlates to the clustering of the members of the phylum in 16S rRNA phylogenetic trees (Ludwig and Klenk, 2001; Cavalier-Smith, 2002; Paster, 2011a). The endoflagella is a unique feature of the phylum and is thought to responsible for the great pathogenic and ecological diversity of its many members (Ren et al., 2003). Of the 38 CSIs we have identified in this study, one was uniquely shared by all 48 members of the phylum Spirochaetes and absent in every other sequenced group of bacteria. The identified CSI is located in the flagellar basal-body rod protein FlgC, a core component of the motor complex of the flagella (Macnab, 2003). This CSI provides a novel means to distinguish the members of the phylum from all

Table 2 | Conserved signature Indels that are specific for members of the family *Spirochaetaceae*.

Protein name	Gene name	GI number	Figure number	Indel size	Indel position
Phosphoribosylpyrophosphate synthetase	prsA	496158147	**Figure 4**	15 aa ins	97–143
Alanyl-tRNA synthetase	alaS	386859446	Supplemental Figure 1	2 aa ins	277–306
Phosphoribosylpyrophosphate synthetase	prsA	387827445	Supplemental Figure 2	8 aa ins	256–297
Preprotein translocase	secY	15639201	Supplemental Figure 3	1 aa del	340–373
Peptide chain release factor 2	prfB	257457828	Supplemental Figure 4	1 aa del	137–176
DNA mismatch repair protein MutS	mutS	224532424	Supplemental Figure 5	2 aa del	720–751
DNA mismatch repair protein MutL	mutL	338706271	Supplemental Figure 6	4 aa del	494–520

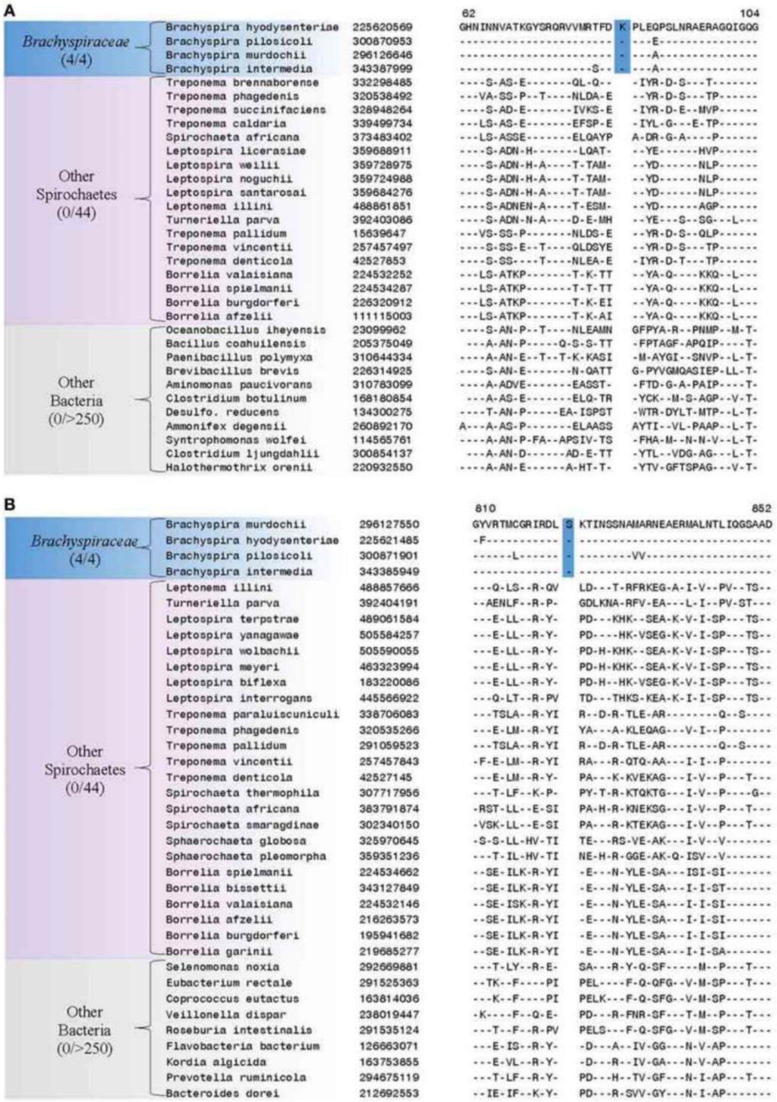

A
62 104
Brachyspiraceae (4/4)
Brachyspira hyodysenteriae 225620569 GHNINNVATKGYSRQRVVMRTFD K PLEQPSLNRAERAGQIGQG
Brachyspira pilosicoli 300870953 ---------------------- - ---E-------------
Brachyspira murdochii 296126646 ---------------------- - ---A-------------
Brachyspira intermedia 343387999 -------------------S-- - ---A-------------
Other Spirochaetes (0/44)
Treponema brennaborense 332298485 ----S-AS-E-------QL-Q-- -IYR-D-S---T------
Treponema phagedenis 320538492 ---VA-SS-P--T----NLDA-E --YR-D-S---TP-----
Treponema succinifaciens 328948264 ----S-AD-E-------IVKS-E -IYR-D-E--MVP-----
Treponema caldaria 339499734 ---LS-AS-E-------EFSP-E -IYL-G---E-TP-----
Spirochaeta africana 373483402 ---LS-ASSE-------ELQAYP A-DR-G-A----P-----
Leptospira licerasiae 359688911 ----S-ADN-H-------LQAT- --YE------HVP-----
Leptospira weilii 359728975 ----S-ADN-H-A----T-TAM- --YD------NLP-----
Leptospira noguchii 359724988 ----S-ADN-H-A----T-TAM- --YD------NLP-----
Leptospira santarosai 359684276 ----S-ADN-H-A----T-TAM- --YD------NLP-----
Leptonema illini 488861851 ----S-ADNEN-A----T-ESM- --YD------AGP-----
Turneriella parva 392403086 ----S-ADN-N-A----D-E-MH --YE---S--SG---L---
Treponema pallidum 15639647 ---VS-SS-P-------NLDS-E --YR-D-S--QLP-----
Treponema vincentii 257457497 ----S-SS-E--T----QLDSYE --YR-D-S---TP-----
Treponema denticola 42527853 ----S-SS----T----NLEA-E -IYR-D-T---TP-----
Borrelia valaisiana 224532252 ---LS-ATKP-------T-K-TT --YA-Q----KKQ--L---
Borrelia spielmanii 224534287 ---LS-ATKP-------T-T-TT --YA-Q----KKQ--L---
Borrelia burgdorferi 226320912 ---LS-ATKP-------T-K-EI --YA-Q----KKQ--L---
Borrelia afzelii 111115003 ---LS-ATKP-------T-K-AI --YA-Q----KKQ--L---
Other Bacteria (0/>250)
Oceanobacillus iheyensis 23099962 ----S-AN-P--T----NLEAMN GFPYA-R--PNMP--M-T-
Bacillus coahuilensis 205375049 ---A-AN-P-----Q-S-S-TT -FPTAGF-APQIP----T-
Paenibacillus polymyxa 310644334 ----A-AN-E--T--T-K-KASI -M-AYGI--SNVP--L-T-
Brevibacillus brevis 226314925 ----S-AN-E-------N-QATT G-PYVGMQASIEP-LL-T-
Aminomonas paucivorans 310783099 ----A-ADVE-------EASST- -FTD-G-A-PAIP----T-
Clostridium botulinum 168180854 ---A-AS-E-------ELQ-TR -YCK--M-S-AGP--V-T-
Desulfo. reducens 134300275 ----T-AN-P-----EA-ISPST -WTR-DYLT-MTP--L-T-
Ammonifex degensii 260892170 A---A-AS-P-------ELAASS AYTI--VL-PAAP--L-T-
Syntrophomonas wolfei 114565761 ----A-AN-P-FA--APSIV-TS -FHA-M--N-N-V--L-T-
Clostridium ljungdahlii 300854137 ---A-AN-D------AD-E-TT -YTL--VDG-AG---L-T-
Halothermothrix orenii 220932550 ----A-AN-E------A-HT-T- -YTV-GFTSPAG---V-T-
B
810 852
Brachyspiraceae (4/4)
Brachyspira murdochii 296127550 GYVRTMCGRIRDL S KTINSSNAMARNEAERMALNTLIQGSAAD
Brachyspira hyodysenteriae 225621485 -F----------- - -----------------------------
Brachyspira pilosicoli 300871901 ------L------ - --------VV-------------------
Brachyspira intermedia 343385949 ------------- - -----------------------------
Other Spirochaetes (0/44)
Leptonema illini 488857666 ---Q-LS--R-QV LD---T-RFRKEG-A-I-V--PV--TS--
Turneriella parva 392404191 --AENLF--R-P- GDLKNA-RFV-EA---L-I--PV-ST---
Leptospira terpstrae 489061584 ---E-LL--R-Y- PD---KHK--SEA-K-V-I-SP---TS--
Leptospira yanagawae 505584257 ---E-LL--R-Y- PD----HK-VSEG-K-V-I-SP---TS--
Leptospira wolbachii 505590055 ---E-LL--R-Y- PD-H-KHK--SEA-K-V-I-SP---TS--
Leptospira meyeri 463323994 ---E-LL--R-Y- PD-H-KHK--SEA-K-V-I-SP---TS--
Leptospira biflexa 183220086 ---E-LL--R-Y- PD-H--HK-VSEG-K-V-I-SP---TS--
Leptospira interrogans 445566922 ---Q-LT--R-PV TD---THKS-KEA-K-I-I-SP---TS--
Treponema paraluiscuniculi 338706083 ---TSLA--R-YI R--D-R-TLE-AR--------Q--S----
Treponema phagedenis 320535266 ---E-LM--R-YI YA---A-KLEQAG---V-I--P-------
Treponema pallidum 291059523 ---TSLA--R-YI R--D-R-TLE-AR--------Q--S----
Treponema vincentii 257457843 -F-E-LM--R-YI RA---R-QTQ-AA---I-I--P-------
Treponema denticola 42527145 ---E-LM--R-Y- PA---K-KVEKAG---I-V--P---T---
Spirochaeta thermophila 307717956 ---T-LF--K-P- PY-T-R-KTQKTG---I-V--P----G--
Spirochaeta africana 383791874 -RST-LL--E-SI PA-H-R-KNEKSG---I-V--P---T---
Spirochaeta smaragdinae 302340150 -VSK-LL--E-SI PA---R-KTEKAG---I-V--P---T---
Sphaerochaeta globosa 325970645 -S-S-LL-HV-TI TE---RS-VE-AK---I-V--V-------
Sphaerochaeta pleomorpha 359351236 ---T-IL-HV-TI NE-H-R-GGE-AK-Q-ISV--V-------
Borrelia spielmanii 224534662 --SE-ILK-R-YI -E---N-YLE-SA---ISI-SI-------
Borrelia bissettii 343127849 --SE-ILK-R-YI -E---N-YLE-SA---I-I-ST-------
Borrelia valaisiana 224532146 --SE-ISK-R-YI -E---N-YLE-SA---I-I-SI-------
Borrelia afzelii 216263573 --SE-ILK-R-YI -E---N-YLE-SA---I-I-SI-------
Borrelia burgdorferi 195941682 --SE-ILK-R-YI -E---N-YLE-SA---I-I-SI-------
Borrelia garinii 219685277 --SE-ILK-R-YI -E---N-YLE-SA---I-I-SA-------
Other Bacteria (0/>250)
Selenomonas noxia 292669881 ---T-LY--R-E- SA---R-Y-Q-SF-----M--P---T---
Eubacterium rectale 291525363 --TK--F----PI PEL----F-Q-QFG--V-M-SP---T---
Coprococcus eutactus 163814036 ---K--F----PI PELK---F-Q-SFG--V-M-SP-------
Veillonella dispar 238019447 -K----F--Q-E- PD---R-FNR-SF---T-M--P---T---
Roseburia intestinalis 291535124 ---T--F--R-PV PELS---F-Q-SFG--V-M-SP---T---
Flavobacteria bacterium 126663071 ---E-IS--R-Y- -D---A--IV-GG---N-V-AP-------
Kordia algicida 163753855 ---E-VL--R-Y- -D---R--IV-GA---N-V-AP-------
Prevotella ruminicola 294675119 ---T-LF--R-Y- PD---H--TV-GF---N-I-AP---T---
Bacteroides dorei 212692553 --IE-IF--K-Y- PD---R-SVV-GY---N-I-AP-------

FIGURE 5 | Partial sequence alignments of (A) Flagellar hook-associated protein FlgK and (B) DNA polymerase I, showing two CSIs that are specific for the family *Brachyspiraceae*, but not found in the sequence homologs of any other sequenced bacteria. Sequence homologs for flagellar hook-associated protein FlgK were not identified from members of the genus *Sphaerochaeta*. Sequence information for other *Brachyspiraceae* specific CSIs is presented in Supplemental Figures 7–10 and summarized in **Table 3**.

other bacteria in molecular terms and provides another delimiting marker for the group in addition to the endoflagella. While the role of this CSI in the function or morphology of the Spirochaetes flagella is currently unknown, the unique presence of this CSI in a flagellar protein in all members of the phylum Spirochaetes suggests that it may be related to the unique flagella ultrastructure of the phylum. Earlier work has established that the CSIs are primarily located on surface loops of proteins which are important

Table 3 | Conserved signature Indels that are specific for members of the family *Brachyspiraceae.*

Protein name	Gene name	GI number	Figure number	Indel size	Indel position
Flagellar hook-associated protein FlgK	flgK	225620569	**Figure 5A**	1 aa ins	62–104
DNA polymerase I	polA	296127550	**Figure 5B**	1 aa ins	810–852
Valyl-tRNA synthetase	valS	300871449	Supplemental Figure 7	1 aa ins	225–263
Valyl-tRNA synthetase	valS	300871449	Supplemental Figure 8	2 aa del	660–703
ATP-dependent protease La	lon	225620632	Supplemental Figure 9	1 aa ins	760–793
Glutamyl-tRNA amidotransferase subunit B	gatB	300871379	Supplemental Figure 10	1 aa ins	325–361

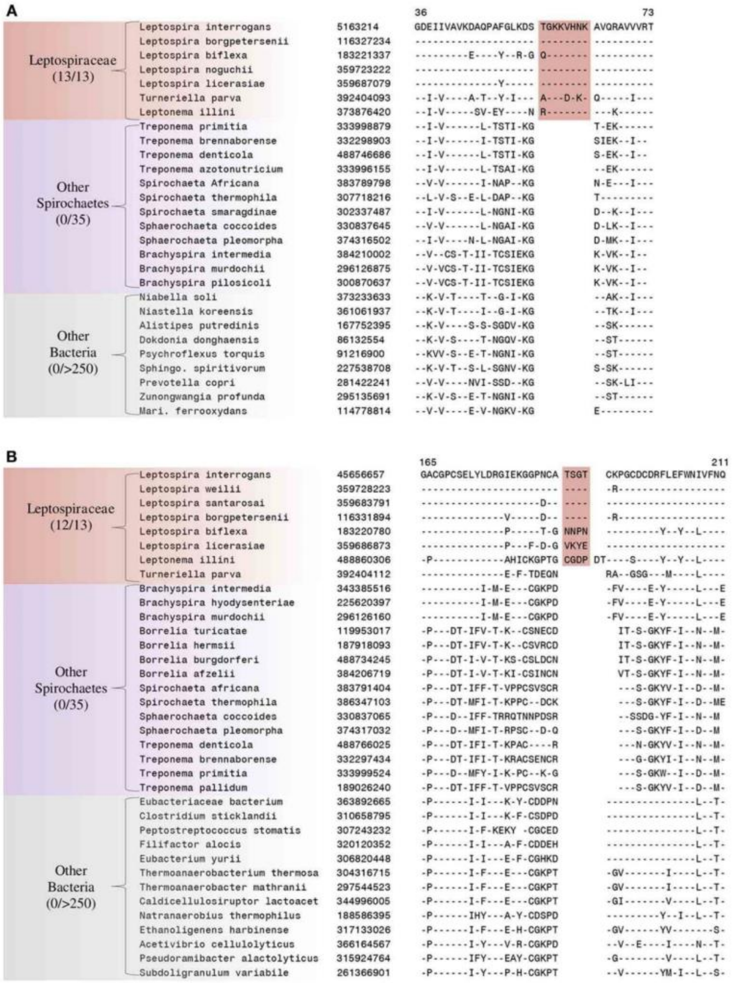


FIGURE 6 | Partial sequence alignments of (A) 50S Ribosomal protein L14 and (B) Alanyl-tRNA synthetase, showing two CSIs that are specific for the family *Leptospiraceae*, but not found in the sequence homologs of any other sequenced bacteria. Sequence information for other *Leptospiraceae* specific CSIs is presented in Supplemental Figures 11–13 and summarized in **Table 4**.

Table 4 | Conserved signature Indels that are specific for members of the family *Leptospiraceae*.

Protein name	Gene name	GI number	Figure Number	Indel Size	Indel Position
50S Ribosomal protein L14	rplN	5163214	**Figure 6A**	8 aa ins	36–73
Alanyl-tRNA synthetase	alaS	45656657	**Figure 6B**	4 aa ins	165–211
30S Ribosomal protein S2	rpsB	116330588	Supplemental Figure 11	2 aa ins	108–141
Flagellar filament core protein FlaB	flaB	12657818	Supplemental Figure 12	4 aa del	130–168
Flagellar basal-body rod protein FlgG	flgG	294828153	Supplemental Figure 13	1 aa ins	80–123

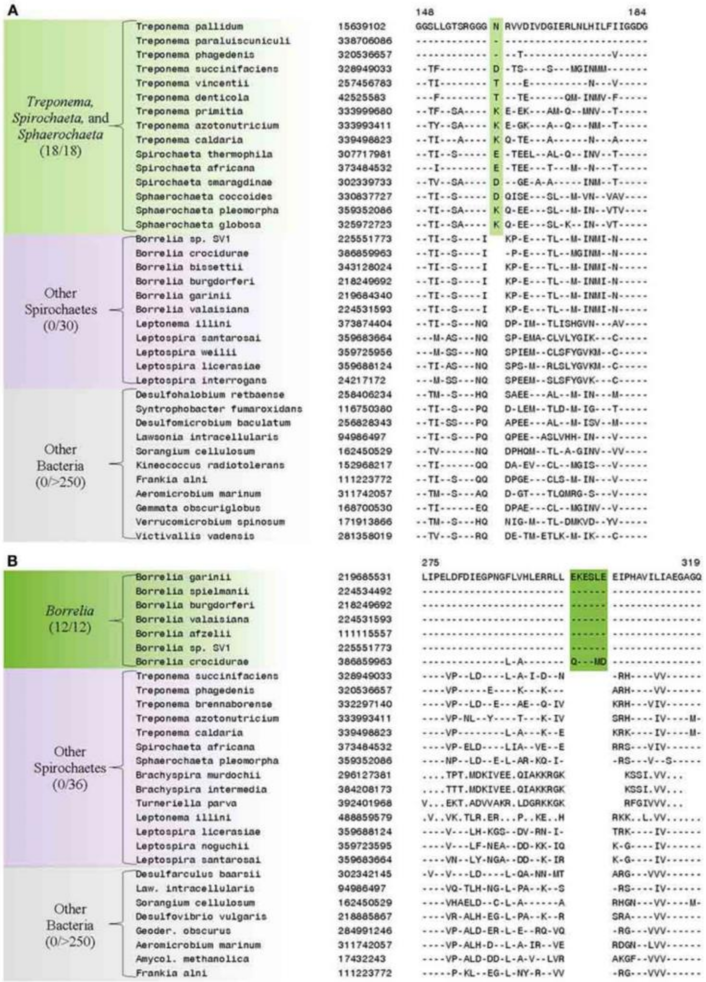

A
148
184
Treponema, Spirochaeta, and *Sphaerochaeta* (18/18)
Other Spirochaetes (0/30)
Other Bacteria (0/>250)
B
275
319
Borrelia (12/12)
Other Spirochaetes (0/36)
Other Bacteria (0/>250)

FIGURE 7 | (A) Partial sequence alignment of the protein 6-phosphofructokinase (pyrophosphate) containing a 1 amino acid insert in a conserved region that is specifically present in the species from the genera *Treponema, Spirochaeta, and Sphaerochaeta*, but not found in any other sequenced bacteria. **(B)** Partial sequence alignment of phosphofructokinase containing a 6 amino acid insert that is specific for the genera *Borrelia*. Sequence information for other CSIs showing similar specificities is provided in **Table 5** and in Supplemental Figures 14–30.

Table 5 | Conserved Signature Indels that are specific for groups within the family *Spirochaetaceae*.

Protein name	Gene name	GI Number	Figure Number	Specificity	Indel size	Indel position
6-phosphofructokinase (pyrophosphate)	pfp	15639102	**Figure 7A**	*Treponema, Spirochaeta* and *Sphaerochaeta*	1 aa ins	148–184
Bifunctional Hpr kinase/phosphatase	hprK	3322886	Supplemental Figure 14	*Treponema, Spirochaeta* and *Sphaerochaeta*	1 aa ins	183–221
30S ribosomal protein S13	rpsM	302337499	Supplemental Figure 15	*Treponema, Spirochaeta* and *Sphaerochaeta*	1 aa del	1–39
Phosphofructokinase	pfk	219685531	**Figure 7B**	*Borrelia*	6 aa ins	275–319
50S ribosomal protein L4	rplD	224534698	Supplemental Figure 16	*Borrelia*	1 aa ins	103–136
tRNA pseudouridine 55 synthase	truB	203284699	Supplemental Figure 17	*Borrelia*	2 aa ins	143–178
Translation elongation factor Tu	tuf	203284386	Supplemental Figure 18	*Borrelia*	1 aa del	330–369
Histidyl-tRNA synthetase	hisS	187918014	Supplemental Figure 19	*Borrelia*	1 aa del	273–301
Seryl-tRNA synthetase	serS	187918098	Supplemental Figure 20	*Borrelia*	1 aa del	231–264
Spoiiij-associtated protein	jag	219684344	Supplemental Figure 21	*Borrelia*	3 aa ins	114–154
Nicotinate phosphoribosyltransferase	pncB	187918492	Supplemental Figure 22	*Borrelia*	1 aa del	134–159
Ribose 5-phosphate isomerase	rpiA	119953435	Supplemental Figure 23	*Borrelia*	1 aa ins	86–110
Ribonuclease Z	rnz	195941574	Supplemental Figure 24	*Borrelia*	2 aa ins	64–94
Hypothetical protein BGAFAR04_0762	–	386859948	Supplemental Figure 25	*Borrelia*	1 aa ins	206–236
Signal recognition particle, subunit FFH/SRP54	–	119953471	Supplemental Figure 26	*Borrelia*	1 aa ins	374–412
Hypothetical protein BSV1_0075	–	15594416	Supplemental Figure 27	*Borrelia*	1 aa del	52–97
Aspartyl/glutamyl-tRNA amidotransferase subunit A	gatA	119953137	Supplemental Figure 28	*Borrelia*	1 aa ins	364–402
Ribosomal RNA methyltransferase	rlmE	203284234	Supplemental Figure 29	*Borrelia*	1 aa ins	15–48
LysM domain/M23/M37 peptidase domain protein	–	224534310	Supplemental Figure 30	*Borrelia*	1 aa ins	320–365

in protein-protein interactions (Akiva et al., 2008; Singh and Gupta, 2009; Gupta, 2010). Thus, the CSI identified in FlgC likely plays an important role in the cellular functions of the flagellar basal-body.

The phylum Spirochaetes contains 4 main lineages (viz. *Spirochaetaceae*, *Brachyspiraceae*, *Leptospiraceae*, and *Brevinemataceae*). These lineages have historically been distinguished from each other by their biochemical characteristics and their 16S rRNA gene sequences (Harwood and Canale-Parola, 1984; Paster et al., 1991; Paster, 2011a). In this study we have also identified 22 CSIs in a diverse range of proteins that are specific to each of the main sequenced lineages of the phylum Spirochaetes (viz. *Spirochaetaceae*, *Brachyspiraceae*, and *Leptospiraceae*), which serve to distinguish these lineages from themselves and all other bacteria. Seven of these identified CSIs were specific for the family *Spirochaetaceae*, 6 CSIs were identified that were specific for the family *Brachyspiraceae*, and 5 CSIs were identified that were specific to the family *Leptospiraceae*. Each of these lineages also branch distinctly and are separated by long branches in both 16S rRNA based and concatenated protein based phylogenetic trees (**Figures 1**, **2**). This molecular and phylogenetic evidence supports the current division of these lineages. However, the large number of CSIs discovered for each of these groups and their genetic distances suggests that these lineages may represent higher taxonomic divisions (viz. orders or classes) than currently recognized. It is noteworthy that two of the CSIs that are specific for the *Brachyspiraceae* family and one that is specific for the *Leptospiraceae* are again found in flagella-related proteins (viz. FlgK, FlgB, FlgG) indicating that there might be interesting differences in the structures and/or functions of flagella within the Spirochaete families.

The family *Spirochaetaceae*, which contains the genera *Borrelia, Clevelandina, Cristispira, Diplocalyx, Hollandina, Pillotina, Sphaerochaeta, Spirochaeta,* and *Treponema*, is the most diverse of the lineages within the phylum Spirochaetes (Paster, 2011b; Euzéby, 2013). The interrelationships between the genera within this family are not reliably resolved by 16S rRNA sequence analysis (Paster, 2011b) (**Figure 2**). In this study we have identified 19 CSIs which serve to delineate at least certain relationships within the family *Spirochaetaceae*. Three of the CSIs identified are specifically found in members of the genera *Sphaerochaeta*, *Spirochaeta*, and *Treponema* and 16 additional CSIs were identified that are specifically found in members of the genus *Borrelia*. These CSIs suggest that the genera *Sphaerochaeta*, *Spirochaeta*, and *Treponema* shared a common ancestor distinct from the members of the genus *Borrelia*. In our concatenated protein phylogenetic tree, the genera *Sphaerochaeta*, *Spirochaeta* and *Treponema* formed a well-supported monophyletic clade, which was separated from the members of the genus *Borrelia* by a long branch, supporting the relationship delineated by these CSIs. Both of these two clades also exhibit considerable phylogenetic diversity. The clade consisting of genera *Sphaerochaeta*, *Spirochaeta*, and *Treponema* contains a number of distinct smaller subclades while the members of the genus *Borrelia* form two highly distinct clades in the phylogenetic trees. However, further work to identify molecular markers will be required to determine the significance of the branching of these subclades. The genus *Cristispira* has not had its genome sequenced, but it branches

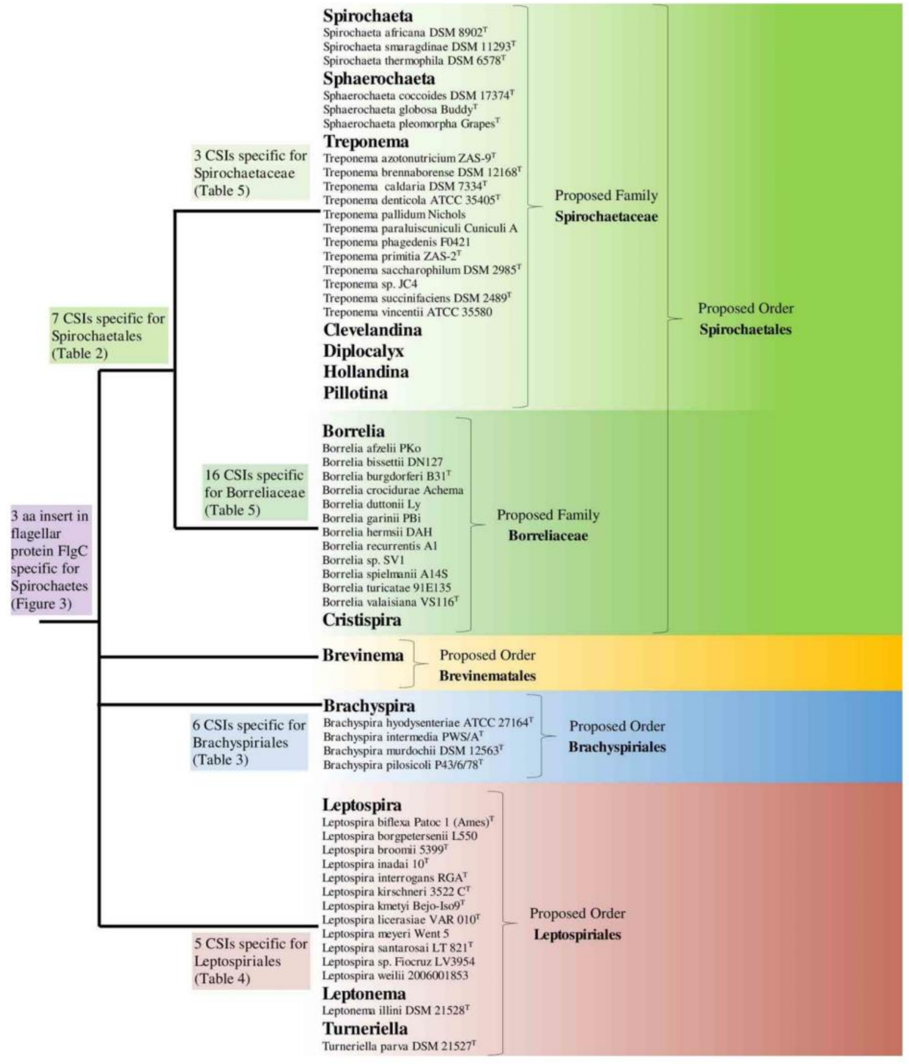

Spirochaeta
Spirochaeta africana DSM 8902^{T}
Spirochaeta smaragdinae DSM 11293^{T}
Spirochaeta thermophila DSM 6578^{T}
Sphaerochaeta
Sphaerochaeta coccoides DSM 17374^{T}
Sphaerochaeta globosa Buddy^{T}
Sphaerochaeta pleomorpha Grapes^{T}
Treponema
Treponema azotonutricium ZAS-9^{T}
Treponema brennaborense DSM 12168^{T}
Treponema caldaria DSM 7334^{T}
Treponema denticola ATCC 35405^{T}
Treponema pallidum Nichols
Treponema paraluiscuniculi Cuniculi A
Treponema phagedenis F0421
Treponema primitia ZAS-2^{T}
Treponema saccharophilum DSM 2985^{T}
Treponema sp. JC4
Treponema succinifaciens DSM 2489^{T}
Treponema vincentii ATCC 35580
Clevelandina
Diplocalyx
Hollandina
Pillotina
3 CSIs specific for Spirochaetaceae (Table 5)
Proposed Family Spirochaetaceae
7 CSIs specific for Spirochaetales (Table 2)
Proposed Order Spirochaetales
Borrelia
Borrelia afzelii PKo
Borrelia bissettii DN127
Borrelia burgdorferi B31^{T}
Borrelia crocidurae Achema
Borrelia duttonii Ly
Borrelia garinii PBi
Borrelia hermsii DAH
Borrelia recurrentis A1
Borrelia sp. SV1
Borrelia spielmanii A14S
Borrelia turicatae 91E135
Borrelia valaisiana VS116^{T}
Cristispira
16 CSIs specific for Borreliaceae (Table 5)
Proposed Family Borreliaceae
3 aa insert in flagellar protein FlgC specific for Spirochaetes (Figure 3)
Brevinema
Proposed Order Brevinematales
Brachyspira
Brachyspira hyodysenteriae ATCC 27164^{T}
Brachyspira intermedia PWS/A^{T}
Brachyspira murdochii DSM 12563^{T}
Brachyspira pilosicoli P43/6/78^{T}
6 CSIs specific for Brachyspiriales (Table 3)
Proposed Order Brachyspiriales
Leptospira
Leptospira biflexa Patoc 1 (Ames)^{T}
Leptospira borgpetersenii L550
Leptospira broomii 5399^{T}
Leptospira inadai 10^{T}
Leptospira interrogans RGA^{T}
Leptospira kirschneri 3522 C^{T}
Leptospira kmetyi Bejo-Iso9^{T}
Leptospira licerasiae VAR 010^{T}
Leptospira meyeri Went 5
Leptospira santarosai LT 821^{T}
Leptospira sp. Fiocruz LV3954
Leptospira weilii 2006001853
Leptonema
Leptonema illini DSM 21528^{T}
Turneriella
Turneriella parva DSM 21527^{T}
5 CSIs specific for Leptospiriales (Table 4)
Proposed Order Leptospiriales

FIGURE 8 | A summary diagram depicting the distribution of identified CSIs and the proposed reclassification of the groups within the phylum Spirochaetes. A representative strain is listed for each genome sequenced species. The letter ^{T} refers to the type strain of the species.

with the members of the genus *Borrelia* reliably in 16S rRNA based phylogenetic trees suggesting that some, if not all, of the *Borrelia* specific CSIs identified in this study may also be found in *Cristispira* (Paster, 2011b) (**Figure 2**). The remaining members of the family *Spirochaetaceae* (viz. *Clevelandina, Diplocalyx, Hollandina,* and *Pillotina*) have been identified in the hindguts of termite and cockroaches but have yet to be isolated and grown in pure or mixed culture. The current placement of the identified members of *Clevelandina, Diplocalyx, Hollandina,* and *Pillotina* in distinct genera within the family *Spirochaetaceae* is ambiguous and based largely on analyses of morphological characteristics (Bermudes et al., 1988). No genome or 16S rRNA sequences are currently available from these genera for phylogenetic analysis. However, the observations presented in this report suggest that the family *Spirochaetaceae* contains at least two distinct monophyletic groups: one consisting of the genera *Sphaerochaeta*, *Spirochaeta*, and *Treponema* and another consisting of the genera *Borrelia* and *Cristispira*.

TAXONOMIC IMPLICATIONS

The results presented here show that the main lineages of the phylum Spirochaetes are evolutionarily distinct. The families *Spirochaetaceae*, *Brachyspiraceae*, and *Leptospiraceae* are distinguished from each other and all other bacteria by large numbers of identified CSIs in widely distributed proteins. Additionally, these three families branch distinctly in both 16S rRNA based and concatenated protein based phylogenetic trees. The results presented here also show that the family *Spirochaetaceae* consists of two distinct monophyletic groups. The distinctiveness of these groups is supported by both molecular evidence, in the form of the large numbers of discovered CSIs, and phylogenetic analyses. Additionally, both of these distinct groups exhibit a large amount of phylogenetic diversity which is currently not reflected in their taxonomy. The current taxonomic organization of the phylum Spirochaetes places all of the main lineages (viz. *Spirochaetaceae*, *Brachyspiraceae*, *Leptospiraceae*, and *Brevinemataceae*) into a single order. However, to adequately recognize both distinctiveness of the main lineages within the phylum Spirochaetes and the distinctiveness and diversity of the two main groups within the family *Spirochaetaceae*, the main lineages of the phylum Spirochaetes would have to have their taxonomic rank increased. To recognize the distinctiveness of both the main lineages within the phylum Spirochaetes and the two main groups within the family *Spirochaetaceae* we are proposing a taxonomic rearrangement of the phylum as follows: We propose that the family *Leptospiraceae* be transferred to the novel order *Leptospiriales* ord. nov. within the class *Spirochaetia*, the family *Brachyspiraceae* be transferred to the novel order *Brachyspiriales* ord. nov. within the class *Spirochaetia*, the family *Brevinemataceae* be transferred to the novel order *Brevinematales* ord. nov. within the class *Spirochaetia*, and that the genera *Borrelia* and *Cristispira* be transferred to the novel family *Borreliaceae* fam. nov. within the order *Spirochaetales* (**Figure 8**). The emended descriptions of the order *Spirochaetales* and the family *Spirochaetaceae*, as well as a description of the new taxonomic groups *Leptospiriales* ord. nov., *Brachyspiriales* ord. nov., *Brevinematales* ord. nov., and *Borreliaceae* fam. nov. are provided below.

EMENDED DESCRIPTION OF THE ORDER *Spirochaetales* (BUCHANAN, 1917)

The order contains two families, *Spirochaetaceae* and *Borreliaceae*, of which *Spirochaetaceae* is the type family. Organisms are helical or coccoid, 0.1–75 μm in diameter and 3.5–250 μm in length. Cells do not have hooked ends. Cells may possess flagella. Periplasmic flagella overlap in the central region of the cell. The diamino acid component of the peptidoglycan is L-ornithine. Anaerobic, facultatively anaerobic, or microaerophilic. Organisms are Chemo-organotrophic and utilize carbohydrates or amino acids as carbon and energy sources. Both free living and host associated members. The G + C content of the DNA is 27–66 (mol%). The type genus is *Spirochaeta* (Ehrenberg, 1835).

Organisms from this order are distinguished from all other Bacteria by the conserved signature indels (CSIs) described in this report in the following proteins: Alanyl-tRNA synthetase, Phosphoribosylpyrophosphate synthetase, SecY preprotein translocase, peptide chain release factor 2, DNA mismatch repair protein MutS, and DNA mismatch repair protein MutL.

EMENDED DESCRIPTION OF THE FAMILY *Spirochaetaceae* (SWELLENGREBEL 1907 EMEND. ABT ET AL., 2012)

The family contains seven genera, *Clevelandina, Diplocalyx, Hollandina, Pillotina, Sphaerochaeta, Spirochaeta*, and *Treponema* of which *Spirochaeta* is the type genus. Organisms are helical or coccoid, 0.1–75 μm in diameter and 5–250 μm in length. Cells do not have hooked ends. Cells may possess flagella. Periplasmic flagella overlap in the central region of the cell. Cells can be anaerobic or facultatively anaerobic. The diamino acid component of the peptidoglycan is L-ornithine. Organisms are chemo-organotrophic and utilize carbohydrates or amino acids as carbon and energy sources. Both free living and host associated members. The G + C content of the DNA is 36–66 (mol%).

Organisms from this family are distinguished from all other bacteria by the CSIs described in this report in the following proteins: 6-phosphofructokinase (pyrophosphate), bifunctional Hpr kinase/phosphatase, and 30S ribosomal protein S13.

DESCRIPTION OF *Borreliaceae* fam. nov.

Borreliaceae (Bor.re'li.a'ce.ae. N.L. fem. n. *Borrelia* type genus of the family; -aceae ending to denote a family; M.L. fem. pl. n. *Borreliaceae* the *Borrelia* family).

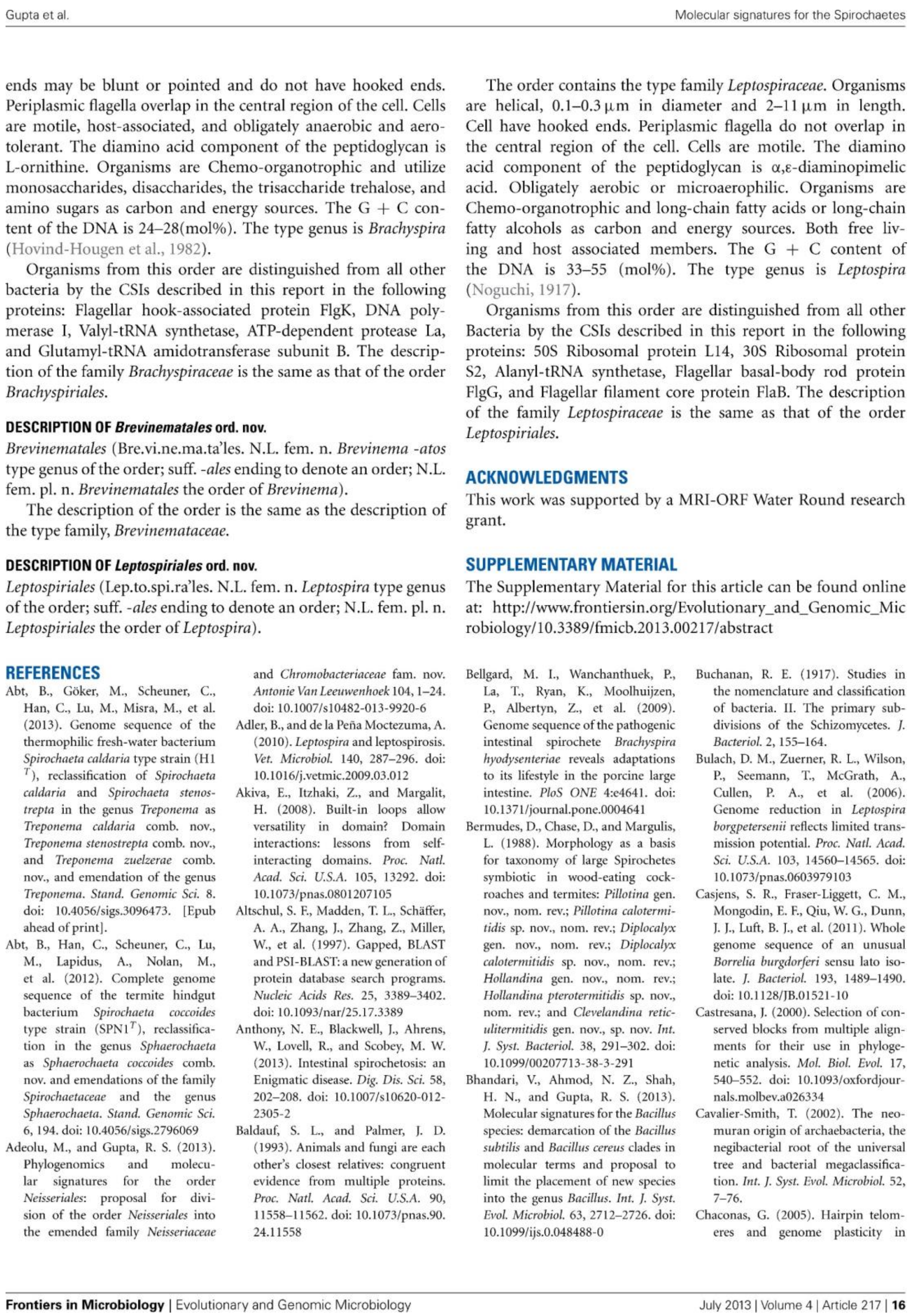
The family contains two genera, *Borrelia* and *Cristispira* of which *Borrelia* is the type genus. Organisms are helical, 0.2–3 μm in diameter and 3–180 μm in length. Cells do not have hooked ends. Periplasmic flagella overlap in the central region of the cell. Cells are motile, host-associated, and microaerophilic. The diamino acid component of the peptidoglycan is L-ornithine. Organisms are chemo-organotrophic and utilize carbohydrates or amino acids as carbon and energy sources. The G + C content of the DNA is 27–32 (mol%).

Organisms from this family are distinguished from all other Bacteria by the CSIs described in this report in the following proteins: Phosphofructokinase, 50S ribosomal protein L4, tRNA pseudouridine 55 synthase, Translation elongation factor-Tu, Histidyl-tRNA synthetase, Seryl-tRNA synthetase, Spoiiij-associtated protein, Nicotinate phosphoribosyltransferase, Ribose 5-phosphate isomerase, Ribonuclease Z, Hypothetical protein BGAFAR04_0762, Signal recognition particle subunit FFH/SRP54, Hypothetical protein BSV1_0075, Aspartyl/glutamyl-tRNA amidotransferase subunit A, Ribosomal RNA methyltransferase, and a LysM domain/M23/M37 peptidase domain protein.

DESCRIPTION OF *Brachyspiriales* ord. nov.

Brachyspiriales (Bra.chy.spi.ra'les. N.L. fem. n. Brachyspira type genus of the order; suff. *-ales* ending to denote an order; N.L. fem. pl. n. *Brachyspiriales* the order of *Brachyspira*).

The order contains the type family *Brachyspiraceae*. Organisms are helical, 0.2–0.4 μm in diameter and 2–11 μm in length. Cell



ends may be blunt or pointed and do not have hooked ends. Periplasmic flagella overlap in the central region of the cell. Cells are motile, host-associated, and obligately anaerobic and aerotolerant. The diamino acid component of the peptidoglycan is L-ornithine. Organisms are Chemo-organotrophic and utilize monosaccharides, disaccharides, the trisaccharide trehalose, and amino sugars as carbon and energy sources. The G + C content of the DNA is 24–28(mol%). The type genus is *Brachyspira* (Hovind-Hougen et al., 1982).

Organisms from this order are distinguished from all other bacteria by the CSIs described in this report in the following proteins: Flagellar hook-associated protein FlgK, DNA polymerase I, Valyl-tRNA synthetase, ATP-dependent protease La, and Glutamyl-tRNA amidotransferase subunit B. The description of the family *Brachyspiraceae* is the same as that of the order *Brachyspiriales*.

DESCRIPTION OF *Brevinematales* ord. nov.

Brevinematales (Bre.vi.ne.ma.ta'les. N.L. fem. n. *Brevinema -atos* type genus of the order; suff. *-ales* ending to denote an order; N.L. fem. pl. n. *Brevinematales* the order of *Brevinema*).

The description of the order is the same as the description of the type family, *Brevinemataceae*.

DESCRIPTION OF *Leptospiriales* ord. nov.

Leptospiriales (Lep.to.spi.ra'les. N.L. fem. n. *Leptospira* type genus of the order; suff. *-ales* ending to denote an order; N.L. fem. pl. n. *Leptospiriales* the order of *Leptospira*).

The order contains the type family *Leptospiraceae*. Organisms are helical, 0.1–0.3 μm in diameter and 2–11 μm in length. Cell have hooked ends. Periplasmic flagella do not overlap in the central region of the cell. Cells are motile. The diamino acid component of the peptidoglycan is α,ε-diaminopimelic acid. Obligately aerobic or microaerophilic. Organisms are Chemo-organotrophic and long-chain fatty acids or long-chain fatty alcohols as carbon and energy sources. Both free living and host associated members. The G + C content of the DNA is 33–55 (mol%). The type genus is *Leptospira* (Noguchi, 1917).

Organisms from this order are distinguished from all other Bacteria by the CSIs described in this report in the following proteins: 50S Ribosomal protein L14, 30S Ribosomal protein S2, Alanyl-tRNA synthetase, Flagellar basal-body rod protein FlgG, and Flagellar filament core protein FlaB. The description of the family *Leptospiraceae* is the same as that of the order *Leptospiriales*.

ACKNOWLEDGMENTS

This work was supported by a MRI-ORF Water Round research grant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/Evolutionary_and_Genomic_Microbiology/10.3389/fmicb.2013.00217/abstract

REFERENCES

Abt, B., Göker, M., Scheuner, C., Han, C., Lu, M., Misra, M., et al. (2013). Genome sequence of the thermophilic fresh-water bacterium *Spirochaeta caldaria* type strain (H1 ^{T}), reclassification of *Spirochaeta caldaria* and *Spirochaeta stenostrepta* in the genus *Treponema* as *Treponema caldaria* comb. nov., *Treponema stenostrepta* comb. nov., and *Treponema zuelzerae* comb. nov., and emendation of the genus *Treponema*. *Stand. Genomic Sci.* 8. doi: 10.4056/sigs.3096473. [Epub ahead of print].

Abt, B., Han, C., Scheuner, C., Lu, M., Lapidus, A., Nolan, M., et al. (2012). Complete genome sequence of the termite hindgut bacterium *Spirochaeta coccoides* type strain (SPN1^{T}), reclassification in the genus *Sphaerochaeta* as *Sphaerochaeta coccoides* comb. nov. and emendations of the family *Spirochaetaceae* and the genus *Sphaerochaeta*. *Stand. Genomic Sci.* 6, 194. doi: 10.4056/sigs.2796069

Adeolu, M., and Gupta, R. S. (2013). Phylogenomics and molecular signatures for the order *Neisseriales*: proposal for division of the order *Neisseriales* into the emended family *Neisseriaceae* and *Chromobacteriaceae* fam. nov. *Antonie Van Leeuwenhoek* 104, 1–24. doi: 10.1007/s10482-013-9920-6

Adler, B., and de la Peña Moctezuma, A. (2010). *Leptospira* and leptospirosis. *Vet. Microbiol.* 140, 287–296. doi: 10.1016/j.vetmic.2009.03.012

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. doi: 10.1073/pnas.0801207105

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped, BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Anthony, N. E., Blackwell, J., Ahrens, W., Lovell, R., and Scobey, M. W. (2013). Intestinal spirochetosis: an Enigmatic disease. *Dig. Dis. Sci.* 58, 202–208. doi: 10.1007/s10620-012-2305-2

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. doi: 10.1073/pnas.90.24.11558

Bellgard, M. I., Wanchanthuek, P., La, T., Ryan, K., Moolhuijzen, P., Albertyn, Z., et al. (2009). Genome sequence of the pathogenic intestinal spirochete *Brachyspira hyodysenteriae* reveals adaptations to its lifestyle in the porcine large intestine. *PloS ONE* 4:e4641. doi: 10.1371/journal.pone.0004641

Bermudes, D., Chase, D., and Margulis, L. (1988). Morphology as a basis for taxonomy of large Spirochetes symbiotic in wood-eating cockroaches and termites: *Pillotina* gen. nov., nom. rev.; *Pillotina calotermitidis* sp. nov., nom. rev.; *Diplocalyx* gen. nov., nom. rev.; *Diplocalyx calotermitidis* sp. nov., nom. rev.; *Hollandina* gen. nov., nom. rev.; *Hollandina pterotermitidis* sp. nov., nom. rev.; and *Clevelandina reticulitermitidis* gen. nov., sp. nov. *Int. J. Syst. Bacteriol.* 38, 291–302. doi: 10.1099/00207713-38-3-291

Bhandari, V., Ahmod, N. Z., Shah, H. N., and Gupta, R. S. (2013). Molecular signatures for the *Bacillus* species: demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. *Int. J. Syst. Evol. Microbiol.* 63, 2712–2726. doi: 10.1099/ijs.0.048488-0

Buchanan, R. E. (1917). Studies in the nomenclature and classification of bacteria. II. The primary subdivisions of the Schizomycetes. *J. Bacteriol.* 2, 155–164.

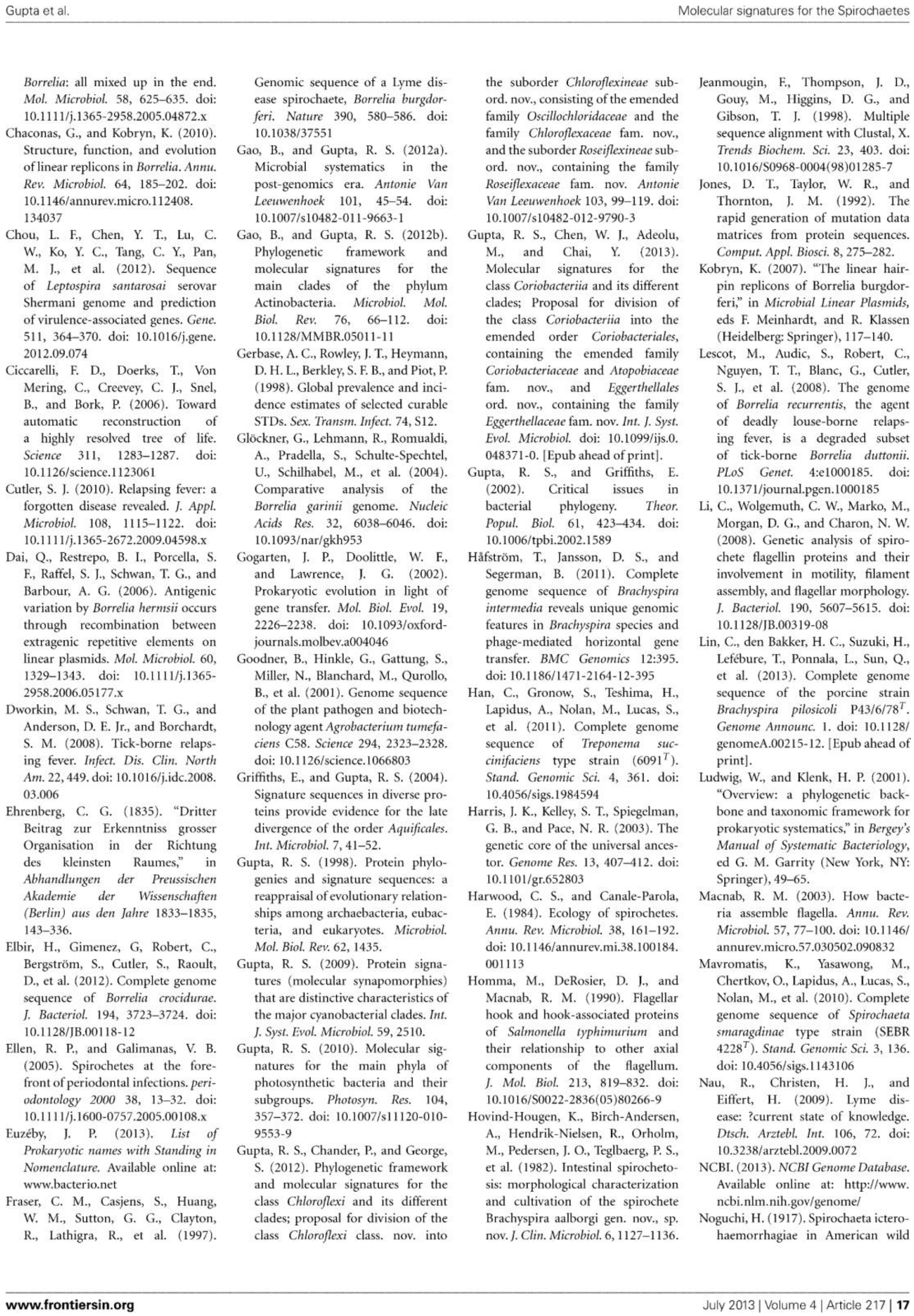
Bulach, D. M., Zuerner, R. L., Wilson, P., Seemann, T., McGrath, A., Cullen, P. A., et al. (2006). Genome reduction in *Leptospira borgpetersenii* reflects limited transmission potential. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14560–14565. doi: 10.1073/pnas.0603979103

Casjens, S. R., Fraser-Liggett, C. M., Mongodin, E. F., Qiu, W. G., Dunn, J. J., Luft, B. J., et al. (2011). Whole genome sequence of an unusual *Borrelia burgdorferi* sensu lato isolate. *J. Bacteriol.* 193, 1489–1490. doi: 10.1128/JB.01521-10

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* 52, 7–76.

Chaconas, G. (2005). Hairpin telomeres and genome plasticity in



Borrelia: all mixed up in the end. *Mol. Microbiol.* 58, 625–635. doi: 10.1111/j.1365-2958.2005.04872.x

Chaconas, G., and Kobryn, K. (2010). Structure, function, and evolution of linear replicons in *Borrelia*. *Annu. Rev. Microbiol.* 64, 185–202. doi: 10.1146/annurev.micro.112408.134037

Chou, L. F., Chen, Y. T., Lu, C. W., Ko, Y. C., Tang, C. Y., Pan, M. J., et al. (2012). Sequence of *Leptospira santarosai* serovar Shermani genome and prediction of virulence-associated genes. *Gene.* 511, 364–370. doi: 10.1016/j.gene.2012.09.074

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. doi: 10.1126/science.1123061

Cutler, S. J. (2010). Relapsing fever: a forgotten disease revealed. *J. Appl. Microbiol.* 108, 1115–1122. doi: 10.1111/j.1365-2672.2009.04598.x

Dai, Q., Restrepo, B. I., Porcella, S. F., Raffel, S. J., Schwan, T. G., and Barbour, A. G. (2006). Antigenic variation by *Borrelia hermsii* occurs through recombination between extragenic repetitive elements on linear plasmids. *Mol. Microbiol.* 60, 1329–1343. doi: 10.1111/j.1365-2958.2006.05177.x

Dworkin, M. S., Schwan, T. G., and Anderson, D. E. Jr., and Borchardt, S. M. (2008). Tick-borne relapsing fever. *Infect. Dis. Clin. North Am.* 22, 449. doi: 10.1016/j.idc.2008.03.006

Ehrenberg, C. G. (1835). "Dritter Beitrag zur Erkenntniss grosser Organisation in der Richtung des kleinsten Raumes," in *Abhandlungen der Preussischen Akademie der Wissenschaften (Berlin) aus den Jahre* 1833–1835, 143–336.

Elbir, H., Gimenez, G, Robert, C., Bergström, S., Cutler, S., Raoult, D., et al. (2012). Complete genome sequence of *Borrelia crocidurae*. *J. Bacteriol.* 194, 3723–3724. doi: 10.1128/JB.00118-12

Ellen, R. P., and Galimanas, V. B. (2005). Spirochetes at the forefront of periodontal infections. *periodontology 2000* 38, 13–32. doi: 10.1111/j.1600-0757.2005.00108.x

Euzéby, J. P. (2013). *List of Prokaryotic names with Standing in Nomenclature.* Available online at: www.bacterio.net

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., et al. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390, 580–586. doi: 10.1038/37551

Gao, B., and Gupta, R. S. (2012a). Microbial systematics in the post-genomics era. *Antonie Van Leeuwenhoek* 101, 45–54. doi: 10.1007/s10482-011-9663-1

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. doi: 10.1128/MMBR.05011-11

Gerbase, A. C., Rowley, J. T., Heymann, D. H. L., Berkley, S. F. B., and Piot, P. (1998). Global prevalence and incidence estimates of selected curable STDs. *Sex. Transm. Infect.* 74, S12.

Glöckner, G., Lehmann, R., Romualdi, A., Pradella, S., Schulte-Spechtel, U., Schilhabel, M., et al. (2004). Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res.* 32, 6038–6046. doi: 10.1093/nar/gkh953

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238. doi: 10.1093/oxfordjournals.molbev.a004046

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., et al. (2001). Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* 294, 2323–2328. doi: 10.1126/science.1066803

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

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.

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

Gupta, R. S. (2010). Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. *Photosyn. Res.* 104, 357–372. doi: 10.1007/s11120-010-9553-9

Gupta, R. S., Chander, P., and George, S. (2012). Phylogenetic framework and molecular signatures for the class *Chloroflexi* and its different clades; proposal for division of the class *Chloroflexi* class. nov. into the suborder *Chloroflexineae* subord. nov., consisting of the emended family *Oscillochloridaceae* and the family *Chloroflexaceae* fam. nov., and the suborder *Roseiflexineae* subord. nov., containing the family *Roseiflexaceae* fam. nov. *Antonie Van Leeuwenhoek* 103, 99–119. doi: 10.1007/s10482-012-9790-3

Gupta, R. S., Chen, W. J., Adeolu, M., and Chai, Y. (2013). Molecular signatures for the class *Coriobacteriia* and its different clades; Proposal for division of the class *Coriobacteriia* into the emended order *Coriobacteriales*, containing the emended family *Coriobacteriaceae* and *Atopobiaceae* fam. nov., and *Eggerthellales* ord. nov., containing the family *Eggerthellaceae* fam. nov. *Int. J. Syst. Evol. Microbiol.* doi: 10.1099/ijs.0.048371-0. [Epub ahead of print].

Gupta, R. S., and Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* 61, 423–434. doi: 10.1006/tpbi.2002.1589

Håfström, T., Jansson, D. S., and Segerman, B. (2011). Complete genome sequence of *Brachyspira intermedia* reveals unique genomic features in *Brachyspira* species and phage-mediated horizontal gene transfer. *BMC Genomics* 12:395. doi: 10.1186/1471-2164-12-395

Han, C., Gronow, S., Teshima, H., Lapidus, A., Nolan, M., Lucas, S., et al. (2011). Complete genome sequence of *Treponema succinifaciens* type strain (6091^T). *Stand. Genomic Sci.* 4, 361. doi: 10.4056/sigs.1984594

Harris, J. K., Kelley, S. T., Spiegelman, G. B., and Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Res.* 13, 407–412. doi: 10.1101/gr.652803

Harwood, C. S., and Canale-Parola, E. (1984). Ecology of spirochetes. *Annu. Rev. Microbiol.* 38, 161–192. doi: 10.1146/annurev.mi.38.100184.001113

Homma, M., DeRosier, D. J., and Macnab, R. M. (1990). Flagellar hook and hook-associated proteins of *Salmonella typhimurium* and their relationship to other axial components of the flagellum. *J. Mol. Biol.* 213, 819–832. doi: 10.1016/S0022-2836(05)80266-9

Hovind-Hougen, K., Birch-Andersen, A., Hendrik-Nielsen, R., Orholm, M., Pedersen, J. O., Teglbaerg, P. S., et al. (1982). Intestinal spirochetosis: morphological characterization and cultivation of the spirochete Brachyspira aalborgi gen. nov., sp. nov. *J. Clin. Microbiol.* 6, 1127–1136.

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. doi: 10.1016/S0968-0004(98)01285-7

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

Kobryn, K. (2007). "The linear hairpin replicons of Borrelia burgdorferi," in *Microbial Linear Plasmids,* eds F. Meinhardt, and R. Klassen (Heidelberg: Springer), 117–140.

Lescot, M., Audic, S., Robert, C., Nguyen, T. T., Blanc, G., Cutler, S. J., et al. (2008). The genome of *Borrelia recurrentis*, the agent of deadly louse-borne relapsing fever, is a degraded subset of tick-borne *Borrelia duttonii*. *PLoS Genet.* 4:e1000185. doi: 10.1371/journal.pgen.1000185

Li, C., Wolgemuth, C. W., Marko, M., Morgan, D. G., and Charon, N. W. (2008). Genetic analysis of spirochete flagellin proteins and their involvement in motility, filament assembly, and flagellar morphology. *J. Bacteriol.* 190, 5607–5615. doi: 10.1128/JB.00319-08

Lin, C., den Bakker, H. C., Suzuki, H., Lefébure, T., Ponnala, L., Sun, Q., et al. (2013). Complete genome sequence of the porcine strain *Brachyspira pilosicoli* P43/6/78^T. *Genome Announc.* 1. doi: 10.1128/genomeA.00215-12. [Epub ahead of print].

Ludwig, W., and Klenk, H. P. (2001). "Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics," in *Bergey's Manual of Systematic Bacteriology*, ed G. M. Garrity (New York, NY: Springer), 49–65.

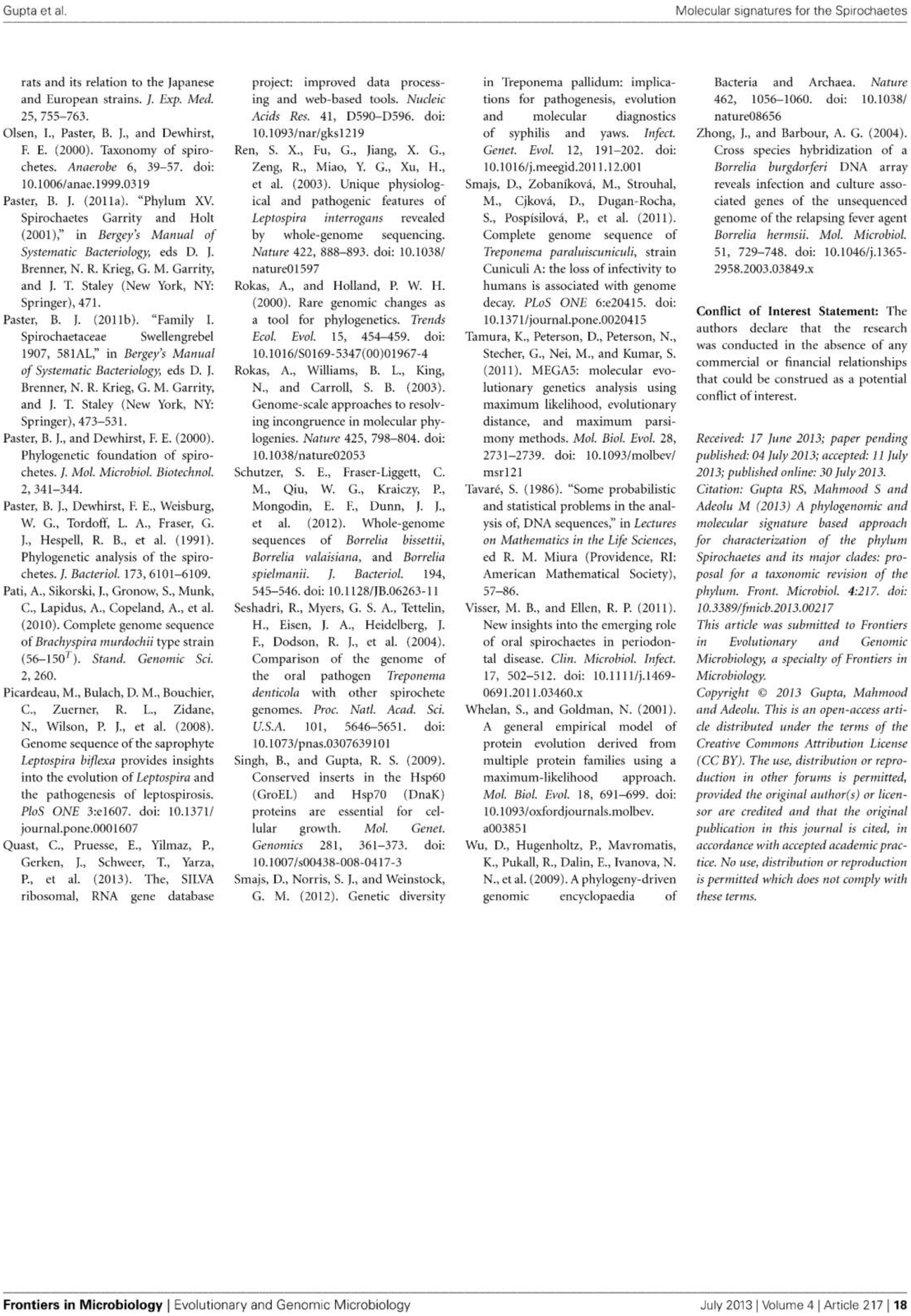
Macnab, R. M. (2003). How bacteria assemble flagella. *Annu. Rev. Microbiol.* 57, 77–100. doi: 10.1146/annurev.micro.57.030502.090832

Mavromatis, K., Yasawong, M., Chertkov, O., Lapidus, A., Lucas, S., Nolan, M., et al. (2010). Complete genome sequence of *Spirochaeta smaragdinae* type strain (SEBR 4228^T). *Stand. Genomic Sci.* 3, 136. doi: 10.4056/sigs.1143106

Nau, R., Christen, H. J., and Eiffert, H. (2009). Lyme disease: ?current state of knowledge. *Dtsch. Arztebl. Int.* 106, 72. doi: 10.3238/arztebl.2009.0072

NCBI. (2013). *NCBI Genome Database.* Available online at: http://www.ncbi.nlm.nih.gov/genome/

Noguchi, H. (1917). Spirochaeta icterohaemorrhagiae in American wild



rats and its relation to the Japanese and European strains. *J. Exp. Med.* 25, 755–763.

Olsen, I., Paster, B. J., and Dewhirst, F. E. (2000). Taxonomy of spirochetes. *Anaerobe* 6, 39–57. doi: 10.1006/anae.1999.0319

Paster, B. J. (2011a). "Phylum XV. Spirochaetes Garrity and Holt (2001)," in *Bergey's Manual of Systematic Bacteriology,* eds D. J. Brenner, N. R. Krieg, G. M. Garrity, and J. T. Staley (New York, NY: Springer), 471.

Paster, B. J. (2011b). "Family I. Spirochaetaceae Swellengrebel 1907, 581AL," in *Bergey's Manual of Systematic Bacteriology,* eds D. J. Brenner, N. R. Krieg, G. M. Garrity, and J. T. Staley (New York, NY: Springer), 473–531.

Paster, B. J., and Dewhirst, F. E. (2000). Phylogenetic foundation of spirochetes. *J. Mol. Microbiol. Biotechnol.* 2, 341–344.

Paster, B. J., Dewhirst, F. E., Weisburg, W. G., Tordoff, L. A., Fraser, G. J., Hespell, R. B., et al. (1991). Phylogenetic analysis of the spirochetes. *J. Bacteriol.* 173, 6101–6109.

Pati, A., Sikorski, J., Gronow, S., Munk, C., Lapidus, A., Copeland, A., et al. (2010). Complete genome sequence of *Brachyspira murdochii* type strain (56–150^T). *Stand. Genomic Sci.* 2, 260.

Picardeau, M., Bulach, D. M., Bouchier, C., Zuerner, R. L., Zidane, N., Wilson, P. J., et al. (2008). Genome sequence of the saprophyte *Leptospira biflexa* provides insights into the evolution of *Leptospira* and the pathogenesis of leptospirosis. *PloS ONE* 3:e1607. doi: 10.1371/journal.pone.0001607

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The, SILVA ribosomal, RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Ren, S. X., Fu, G., Jiang, X. G., Zeng, R., Miao, Y. G., Xu, H., et al. (2003). Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature* 422, 888–893. doi: 10.1038/nature01597

Rokas, A., and Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* 15, 454–459. doi: 10.1016/S0169-5347(00)01967-4

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. doi: 10.1038/nature02053

Schutzer, S. E., Fraser-Liggett, C. M., Qiu, W. G., Kraiczy, P., Mongodin, E. F., Dunn, J. J., et al. (2012). Whole-genome sequences of *Borrelia bissettii*, *Borrelia valaisiana*, and *Borrelia spielmanii*. *J. Bacteriol.* 194, 545–546. doi: 10.1128/JB.06263-11

Seshadri, R., Myers, G. S. A., Tettelin, H., Eisen, J. A., Heidelberg, J. F., Dodson, R. J., et al. (2004). Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 5646–5651. doi: 10.1073/pnas.0307639101

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. doi: 10.1007/s00438-008-0417-3

Smajs, D., Norris, S. J., and Weinstock, G. M. (2012). Genetic diversity in Treponema pallidum: implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. *Infect. Genet. Evol.* 12, 191–202. doi: 10.1016/j.meegid.2011.12.001

Smajs, D., Zobaníková, M., Strouhal, M., Cjková, D., Dugan-Rocha, S., Pospísilová, P., et al. (2011). Complete genome sequence of *Treponema paraluiscuniculi,* strain Cuniculi A: the loss of infectivity to humans is associated with genome decay. *PLoS ONE* 6:e20415. doi: 10.1371/journal.pone.0020415

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tavaré, S. (1986). "Some probabilistic and statistical problems in the analysis of, DNA sequences," in *Lectures on Mathematics in the Life Sciences*, ed R. M. Miura (Providence, RI: American Mathematical Society), 57–86.

Visser, M. B., and Ellen, R. P. (2011). New insights into the emerging role of oral spirochaetes in periodontal disease. *Clin. Microbiol. Infect.* 17, 502–512. doi: 10.1111/j.1469-0691.2011.03460.x

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699. doi: 10.1093/oxfordjournals.molbev.a003851

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056–1060. doi: 10.1038/nature08656

Zhong, J., and Barbour, A. G. (2004). Cross species hybridization of a *Borrelia burgdorferi* DNA array reveals infection and culture associated genes of the unsequenced genome of the relapsing fever agent *Borrelia hermsii*. *Mol. Microbiol.* 51, 729–748. doi: 10.1046/j.1365-2958.2003.03849.x

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: 17 June 2013; paper pending published: 04 July 2013; accepted: 11 July 2013; published online: 30 July 2013.

Citation: Gupta RS, Mahmood S and Adeolu M (2013) A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. Front. Microbiol. ***4****:217. doi: 10.3389/fmicb.2013.00217*

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

CHAPTER 3

A phylogenomic and molecular marker based proposal for the division of the genus *Borrelia* into two genera: the emended genus *Borrelia* containing only the members of the relapsing fever *Borrelia*, and the genus *Borreliella* gen. nov. containing the members of the Lyme disease *Borrelia* (*Borrelia burgdorferi sensu lato* complex).

This chapter describes the use of molecular signatures (CSIs and CSPs), phylogenetic trees, and genomic distance (average nucleotide identity) to differentiate two clinically distinct groups within the genus *Borrelia*. The chapter concludes with a proposal to divide the genus *Borrelia* into two genera, limiting the genus *Borrelia* to only the members of the relapsing fever *Borrelia* group, and transferring the members of the Lyme disease *Borrelia* group (also referred to as the *Borrelia burgdorferi sensu lato* complex) to the genus *Borreliella*. My contributions towards the completion of this chapter include the construction of all phylogenetic trees shown, identification of all CSIs and CSPs shown, the completion of the average nucleotide identity analysis, the creation of the taxonomic proposals, the writing of all drafts and revisions of the manuscript, and the production of all main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials for this work are not included in the chapter but can be accessed along with the rest of the manuscript at:

Adeolu, M., & Gupta, R. S. (2014). *Anton Leeuw Int J G, 105*(6), 1049-1072.

Antonie van Leeuwenhoek (2014) 105:1049–1072
DOI 10.1007/s10482-014-0164-x

ORIGINAL PAPER

A phylogenomic and molecular marker based proposal for the division of the genus *Borrelia* into two genera: the emended genus *Borrelia* containing only the members of the relapsing fever *Borrelia*, and the genus *Borreliella* gen. nov. containing the members of the Lyme disease *Borrelia* (*Borrelia burgdorferi* sensu lato complex)

Mobolaji Adeolu · Radhey S. Gupta

Received: 15 January 2014 / Accepted: 25 March 2014 / Published online: 18 April 2014

Abstract The genus *Borrelia* contains two groups of organisms: the causative agents of Lyme disease and their relatives and the causative agents of relapsing fever and their relatives. These two groups are morphologically indistinguishable and are difficult to distinguish biochemically. In this work, we have carried out detailed comparative genomic analyses on protein sequences from 38 *Borrelia* genomes to identify molecular markers in the forms of conserved signature inserts/deletions (CSIs) that are specifically found in the *Borrelia* homologues, and conserved signature proteins (CSPs) which are uniquely present in *Borrelia* species. Our analyses have identified 31 CSIs and 82 CSPs that are uniquely shared by all sequenced *Borrelia* species, providing molecular markers for this group of organisms. In addition, our work has identified 7 CSIs and 21 CSPs which are uniquely found in the Lyme disease *Borrelia* species and eight CSIs and four CSPs that are specific for members of the relapsing fever *Borrelia* group. Additionally, 38 other CSIs, in proteins which are uniquely found in *Borrelia* species, also distinguish these two groups of *Borrelia*. The identified CSIs and CSPs provide novel and highly specific molecular markers for identification and distinguishing between the Lyme disease *Borrelia* and the relapsing fever *Borrelia* species. We also report the results of average nucleotide identity (ANI) analysis on *Borrelia* genomes and phylogenetic analysis for these species based upon 16S rRNA sequences and concatenated sequences for 25 conserved proteins. These analyses also support the distinctness of the two *Borrelia* clades. On the basis of the identified molecular markers, the results from ANI and phylogenetic studies, and the distinct pathogenicity profiles and arthropod vectors used by different *Borrelia* spp. for their transmission, we are proposing a division of the genus *Borrelia* into two separate genera: an emended genus *Borrelia*, containing the causative agents of relapsing fever and a novel genus, *Borreliella* gen. nov., containing the causative agents of Lyme disease.

Keywords *Borrelia* · *Borreliella* · *Borrelia* taxonomy · Lyme disease · Phylogenetic trees · Average nucleotide identity · Conserved signature indels · Conserved signature proteins

Introduction

The genus *Borrelia* is an important pathogenic group of helical shaped, motile organisms that form a highly

Electronic supplementary material The online version of this article (doi:10.1007/s10482-014-0164-x) contains supplementary material, which is available to authorized users.

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

distinct, monophyletic lineage within the phylum Spirochaetes (Paster 2011; Wang and Schwartz 2011). Members of this genus are the causative agents of both Lyme disease, which is currently the most prevalent vector-borne disease in North America and temperate regions of Eurasia, and relapsing fever, which is a disease endemic to many disparate regions of the world (Lindgren and Jaenson 2006; Cutler 2010; Adams et al. 2013). Currently, the genus *Borrelia* contains 37 species which are carried by arthropod vectors and exhibit varying pathogenicity in mammalian and avian hosts (Margos et al. 2011; Wang and Schwartz 2011; Parte 2014). These species can be separated into two main groups based upon their pathogenicity profiles. The first group, containing the causative agents of Lyme disease, is commonly referred to as the *Borrelia burgdorferi* sensu lato complex, whereas the other group contains the causative agents of relapsing fever (Postic et al. 1990; Baranton et al. 1992; Wang et al. 1999; Margos et al. 2011; Wang and Schwartz 2011). Although, these two groups are morphologically indistinguishable from each other, their members can be distinguished from each other based on the arthropod vectors which transmit them and by a limited number of biochemical and genetic tests (Wang et al. 1999; Margos et al. 2011; Wang and Schwartz 2011). Our current understanding of the taxonomy and evolutionary relationships among the *Borrelia* species is based largely on DNA–DNA hybridization studies, 16S rRNA gene sequence analysis and multilocus sequence analysis (MLSA) (Margos et al. 2011; Wang and Schwartz 2011). Although these studies provide evidence suggesting separation of the members of the genus *Borrelia* into two distinguishable groups, due to lack of other reliable molecular, morphological, or biochemical characteristics that can distinguish these groups, no formal recognition of these two distinct groups of *Borrelia* has thus far been made (Wang and Schwartz 2011).

Whole genome sequences for members of the genus *Borrelia* are becoming increasingly available in public databases. There are currently 38 genomes from 18 species of *Borrelia* available in the NCBI database (NCBI 2014). These genomes provide a valuable resource to gain insight into the evolutionary history of this group of organisms and to identify novel shared molecular characteristics that are specific for this group of organisms. One useful comparative genomic approach, pioneered by our lab, involves the identification of conserved signature indels (CSIs), which are insertions/deletions uniquely present in protein sequences of organisms from the group of interest, and conserved signature proteins (CSPs), which are lineage specific proteins found only in the group of interest (Gupta and Griffiths 2006; Gupta 2010; Naushad et al. 2014). Due to the specificity of these markers (viz. CSIs and CSPs) for particular groups of bacteria, they represent molecular synapomorphies (markers of common evolutionary decent) which can be used to identify and demarcate specific bacterial groups in clear molecular terms. Additionally, whole genome sequences are also enabling the use of other computational algorithms to determine the overall genome similarity among different organisms (Richter and Rosselló-Móra 2009).

Our recent comparative analysis of Spirochaetes genomes has identified 38 CSIs that clearly delimit the major groups within the phylum and were used to revise the taxonomy of the phylum as a whole (Gupta et al. 2013b). In this work, we extend these studies by examining, in detail, the evolutionary relationships among the *Borrelia* species employing different phylogenetic and comparative genomic approaches. These analyses have identified 31 CSIs and 82 CSPs that are commonly shared by all sequenced *Borrelia* species. More importantly, these studies have identified of 53 CSIs and 25 CSPs, which serve to clearly distinguish the two main groups of *Borrelia* species and provide novel molecular markers to demarcate them in definitive terms. The distinctness of these two groups of *Borrelia* species is also supported by the results of an average nucleotide identity (ANI) analysis of *Borrelia* genomes and by phylogenetic trees constructed based upon 16S rRNA sequences and concatenated protein sequences. On the basis of the identified molecular markers, phylogenetic studies, and other evidence presented here, it is proposed that the genus *Borrelia* should be divided into two separate genera: an emended genus *Borrelia*, containing the causative agents of relapsing fever and a novel genus, *Borreliella* gen. nov., containing the causative agents of Lyme disease.

Methods

Phylogenetic sequence analysis

Phylogenetic analysis was performed on a concatenated sequence alignment of 25 highly conserved

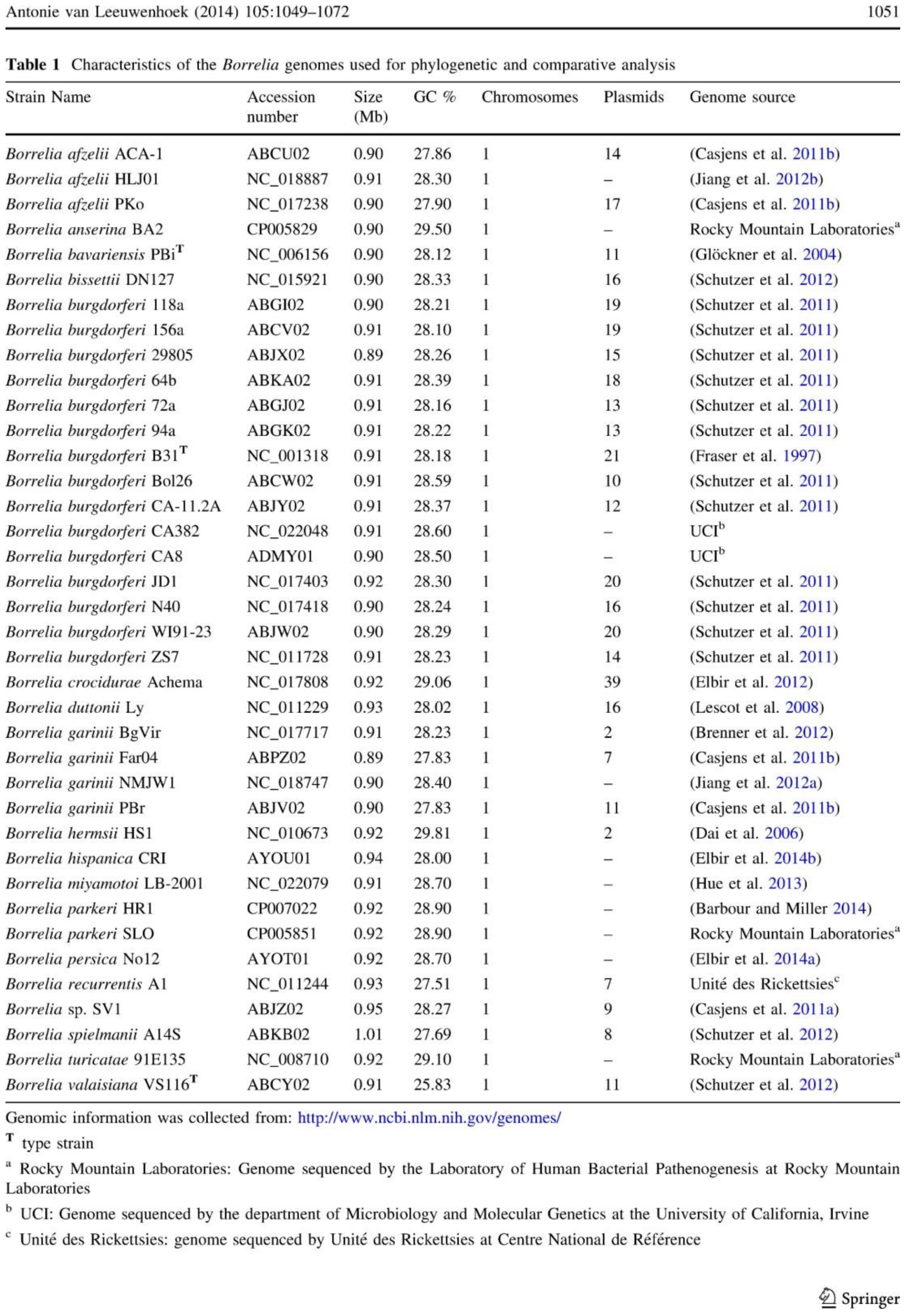


Table 1 Characteristics of the *Borrelia* genomes used for phylogenetic and comparative analysis

Strain Name	Accession number	Size (Mb)	GC %	Chromosomes	Plasmids	Genome source
Borrelia afzelii ACA-1	ABCU02	0.90	27.86	1	14	(Casjens et al. 2011b)
Borrelia afzelii HLJ01	NC_018887	0.91	28.30	1	–	(Jiang et al. 2012b)
Borrelia afzelii PKo	NC_017238	0.90	27.90	1	17	(Casjens et al. 2011b)
Borrelia anserina BA2	CP005829	0.90	29.50	1	–	Rocky Mountain Laboratories[a]
Borrelia bavariensis PBi[T]	NC_006156	0.90	28.12	1	11	(Glöckner et al. 2004)
Borrelia bissettii DN127	NC_015921	0.90	28.33	1	16	(Schutzer et al. 2012)
Borrelia burgdorferi 118a	ABGI02	0.90	28.21	1	19	(Schutzer et al. 2011)
Borrelia burgdorferi 156a	ABCV02	0.91	28.10	1	19	(Schutzer et al. 2011)
Borrelia burgdorferi 29805	ABJX02	0.89	28.26	1	15	(Schutzer et al. 2011)
Borrelia burgdorferi 64b	ABKA02	0.91	28.39	1	18	(Schutzer et al. 2011)
Borrelia burgdorferi 72a	ABGJ02	0.91	28.16	1	13	(Schutzer et al. 2011)
Borrelia burgdorferi 94a	ABGK02	0.91	28.22	1	13	(Schutzer et al. 2011)
Borrelia burgdorferi B31[T]	NC_001318	0.91	28.18	1	21	(Fraser et al. 1997)
Borrelia burgdorferi Bol26	ABCW02	0.91	28.59	1	10	(Schutzer et al. 2011)
Borrelia burgdorferi CA-11.2A	ABJY02	0.91	28.37	1	12	(Schutzer et al. 2011)
Borrelia burgdorferi CA382	NC_022048	0.91	28.60	1	–	UCI[b]
Borrelia burgdorferi CA8	ADMY01	0.90	28.50	1	–	UCI[b]
Borrelia burgdorferi JD1	NC_017403	0.92	28.30	1	20	(Schutzer et al. 2011)
Borrelia burgdorferi N40	NC_017418	0.90	28.24	1	16	(Schutzer et al. 2011)
Borrelia burgdorferi WI91-23	ABJW02	0.90	28.29	1	20	(Schutzer et al. 2011)
Borrelia burgdorferi ZS7	NC_011728	0.91	28.23	1	14	(Schutzer et al. 2011)
Borrelia crocidurae Achema	NC_017808	0.92	29.06	1	39	(Elbir et al. 2012)
Borrelia duttonii Ly	NC_011229	0.93	28.02	1	16	(Lescot et al. 2008)
Borrelia garinii BgVir	NC_017717	0.91	28.23	1	2	(Brenner et al. 2012)
Borrelia garinii Far04	ABPZ02	0.89	27.83	1	7	(Casjens et al. 2011b)
Borrelia garinii NMJW1	NC_018747	0.90	28.40	1	–	(Jiang et al. 2012a)
Borrelia garinii PBr	ABJV02	0.90	27.83	1	11	(Casjens et al. 2011b)
Borrelia hermsii HS1	NC_010673	0.92	29.81	1	2	(Dai et al. 2006)
Borrelia hispanica CRI	AYOU01	0.94	28.00	1	–	(Elbir et al. 2014b)
Borrelia miyamotoi LB-2001	NC_022079	0.91	28.70	1	–	(Hue et al. 2013)
Borrelia parkeri HR1	CP007022	0.92	28.90	1	–	(Barbour and Miller 2014)
Borrelia parkeri SLO	CP005851	0.92	28.90	1	–	Rocky Mountain Laboratories[a]
Borrelia persica No12	AYOT01	0.92	28.70	1	–	(Elbir et al. 2014a)
Borrelia recurrentis A1	NC_011244	0.93	27.51	1	7	Unité des Rickettsies[c]
Borrelia sp. SV1	ABJZ02	0.95	28.27	1	9	(Casjens et al. 2011a)
Borrelia spielmanii A14S	ABKB02	1.01	27.69	1	8	(Schutzer et al. 2012)
Borrelia turicatae 91E135	NC_008710	0.92	29.10	1	–	Rocky Mountain Laboratories[a]
Borrelia valaisiana VS116[T]	ABCY02	0.91	25.83	1	11	(Schutzer et al. 2012)

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

[T] type strain

[a] Rocky Mountain Laboratories: Genome sequenced by the Laboratory of Human Bacterial Pathenogenesis at Rocky Mountain Laboratories

[b] UCI: Genome sequenced by the department of Microbiology and Molecular Genetics at the University of California, Irvine

[c] Unité des Rickettsies: genome sequenced by Unité des Rickettsies at Centre National de Référence

proteins (viz. ArgRS, DnaK, EF-G, EF-Tu, GyrA, GyrB, Hsp60, Hsp70, IleRS, RecA, RpoB, RpoC, SecY, ThrRS, TrpRS, ValRS, and ribosomal proteins L1, L2, L5, L6, S3, S8, S9, S11, and S12) which represent a subset of the core proteins present in all bacteria that are widely used for phylogenetic analysis (Harris et al. 2003; Charlebois and Doolittle 2004; Ciccarelli et al. 2006; Vinuesa 2010; Gao and Gupta 2012b; Gupta et al. 2013b). Sequences for these proteins were obtained from the NCBI database for 38 sequenced *Borrelia* species (Table 1) and *Treponema pallidum* Nichols which was used to root the tree. Multiple sequence alignments for these proteins were created using Clustal_X 1.83 (Jeanmougin et al. 1998) and concatenated into a single alignment file. Poorly aligned regions from this alignment file were removed using Gblocks 0.91b (Castresana 2000). The resulting alignment, which contained 12,129 aligned amino acids, was used for phylogenetic analysis. The maximum likelihood tree based on 1,000 bootstrap replicates of this alignment was constructed using MEGA 6.0 (Tamura et al. 2013) employing the Le and Gascuel (Le and Gascuel 2008) substitution model.

A 16S rRNA gene sequence based phylogenetic tree was also created based on 53 sequences that included representative strains of all cultured *Borrelia* species (Supplemental Table 1). 16S rRNA gene sequences larger than 1,200 bp were obtained for all type strains classified under the genus *Borrelia* in release 115 of the SILVA database (Quast et al. 2013). 16S rRNA gene sequences were also obtained for representative strains from *Borrelia* species without a cultured type and for *T. pallidum* Nichols which was used to root the tree. A maximum likelihood tree based on these sequences was created using 1,000 bootstrap replicates of the 16S rRNA sequence alignments in MEGA 6.0 (Tamura et al. 2013) employing the General Time-Reversible (Tavaré 1986) substitution model.

Average nucleotide analysis

Average nucleotide identity values were calculated in order to assess the relatedness of the sequenced *Borrelia* genomes using the JSpecies v1.2.1 program (Richter and Rosselló-Móra 2009) which utilized an algorithm developed by Goris et al. (2007) to analyze the sequence identity of pairwise genome alignments created using the BLAST v2.2.26 program (Altschul et al. 1997).

Identification of conserved signature indels

To identify CSIs that are commonly shared by the different groups of *Borrelia*, BLAST searches (Altschul et al. 1997) were performed using each protein in the genome of *Borrelia recurrentis* A1 as queries. These searches were performed using the default BLAST parameters against all available sequences in the GenBank non-redundant database. For those proteins for whom high scoring homologues (E values $< 1e^{-20}$) were present in other *Borrelia* species, multiple sequence alignments were created using the Clustal_X 1.83 program (Jeanmougin et al. 1998). These alignments were visually inspected for the presence of insertions or deletions that were flanked on both sides by at least 5-6 conserved amino acid residues in the neighbouring 30–40 amino acids. Indels that were not flanked by conserved regions were not further considered, as they do not provide useful molecular markers (Gupta 2010; Naushad et al. 2014). The specificity of potentially useful indels for subgroups within of the genus *Borrelia* was further evaluated by carrying out detailed BLAST searches on short sequence segments containing the indel and the flanking conserved regions (60-100 amino acids long). To ensure that the identified signatures are only present in *Borrelia* homologues, 250 BLAST hits with the highest similarity to the query sequence were examined for the presence or absence of these CSIs. In this work, we report the results of CSIs that are specific for different groups within the *Borrelia* and where similar CSIs were not observed in any other bacteria in the top 250 BLAST hits. The sequence alignment files presented here contain sequence information for all sequenced species within the genus *Borrelia*. However, due to space constraints, different strains of the sequenced species are not shown, but they all displayed similar sequence characteristics.

Identification of conserved signature proteins

To identify proteins that are uniquely present in various groups of *Borrelia*, BLAST searches (Altschul et al. 1997) were performed using each protein in the genomes of *B. burgdorferi* B31 and *B. recurrentis* A1 as queries. These searches were performed using the default BLAST parameters against all available sequences in the GenBank non-redundant database. Proteins were considered CSPs if either all significant hits were from well-defined groups of *Borrelia* or which involved a large increase in E values from the last hit belonging to a particular group of *Borrelia* to the first hit from any other bacteria and the E values for the latter hits were $>1e^{-04}$, indicating weak similarity that could occur by chance (Gao and Gupta 2007; Naushad et al. 2014). In most cases, the lengths of various significant hits were very similar to those of the query proteins.

Results

Genomic characteristics of the sequenced *Borrelia*

Genome sequences for 38 *Borrelia* strains comprising 18 different species, which are currently available in the NCBI genome database, were used in these analyses. Some characteristics of these *Borrelia* genomes are summarized in Table 1. The genomes of most *Borrelia* species/strains, in addition to containing a linear chromosome, harboured large numbers of linear and circular plasmids, which is very unique among the prokaryotes (Chaconas 2005; Chaconas and Kobryn 2010). The chromosome sizes of the sequenced *Borrelia* fell within a narrow range between 0.89 and 1.01 Mb, with G+C content ranging between 25.83 and 29.81 %.

Phylogenetic sequence analysis

The current understanding of the phylogeny of the genus *Borrelia* is largely based on phylogenetic trees constructed using 16S rRNA, flagellin or housekeeping gene sequences (Fukunaga et al. 1996; Margos et al. 2009; Wang and Schwartz 2011). In this work, we have constructed a phylogenetic tree of the sequenced *Borrelia* species using concatenated sequences for 25 conserved housekeeping and ribosomal proteins (Fig. 1). Members of the genus *Borrelia* have shown some competence for the lateral transfer of tRNA synthetases (Ibba et al. 1997). However, phylogenetic trees based on concatenated sequences for a large number of unlinked and conserved loci minimize the effect of any instances of lateral gene transfer and provide greater resolving power than trees based on any single gene or protein (Rokas et al. 2003; Wu et al. 2009). In the concatenated protein tree, the sequenced *Borrelia* species clustered into two distinct monophyletic and strongly supported clades, which were separated by long branches. One of these clades consisted of the Lyme disease causing *B. burgdorferi* species (*B. burgdorferi* sensu stricto) and its relatives (*B. burgdorferi* sensu lato), while the other clade was comprised of the relapsing fever *Borrelia* (*B. recurrentis*) and its relatives (Fig. 1). These two clades of *Borrelia* are also clearly distinguished in a phylogenetic tree for 3,737 genome sequenced prokaryotes, which was constructed based upon >400 proteins (Segata et al. 2013).

A phylogenetic tree was also constructed based on the 16S rRNA gene sequences, which included representatives from all cultured *Borrelia* species (Fig. 2). Except for *Borrelia turcica*, all *Borrelia* species were grouped into two distinct clades similar to those seen in the concatenated protein tree. However, an earlier study showed that *B. turcica* clusters with several unnamed *Borrelia* isolates in a monophyletic clade related to the relapsing fever *Borrelia* (Takano et al. 2010). The members of the genus *Borrelia* have also been observed to branch into two distinct clades in a number of earlier phylogenetic studies based on 16S rRNA and other individual genes/protein sequences (Takano et al. 2010; Margos et al. 2011; Wang and Schwartz 2011).

Conserved signature indels that distinguish the two clades of Borrelia

CSIs and CSPs that are restricted to a given group of related species provide useful molecular characteristics for evolutionary studies (Gupta 1998; Rokas and Holland 2000; Gao and Gupta 2012a). Recently, CSIs have been used to define novel taxonomic groups and to propose important taxonomic changes for groups of bacteria (viz. Aquificae, *Bacillus*, Chloroflexi, *Neisseriales,* Spirochaetes, Synergistetes and Thermotoga) at different taxonomic ranks (Bhandari and Gupta 2012; Adeolu and Gupta 2013; Bhandari et al. 2013; Gupta et al. 2013a, b; Gupta and Lali 2013; Bhandari and Gupta 2014). In this work we have carried out comprehensive comparative analyses of *Borrelia* genomes in order to identify CSIs that clarify the relationship between the *Borrelia*. These studies have identified 31 CSIs that are specifically found in protein

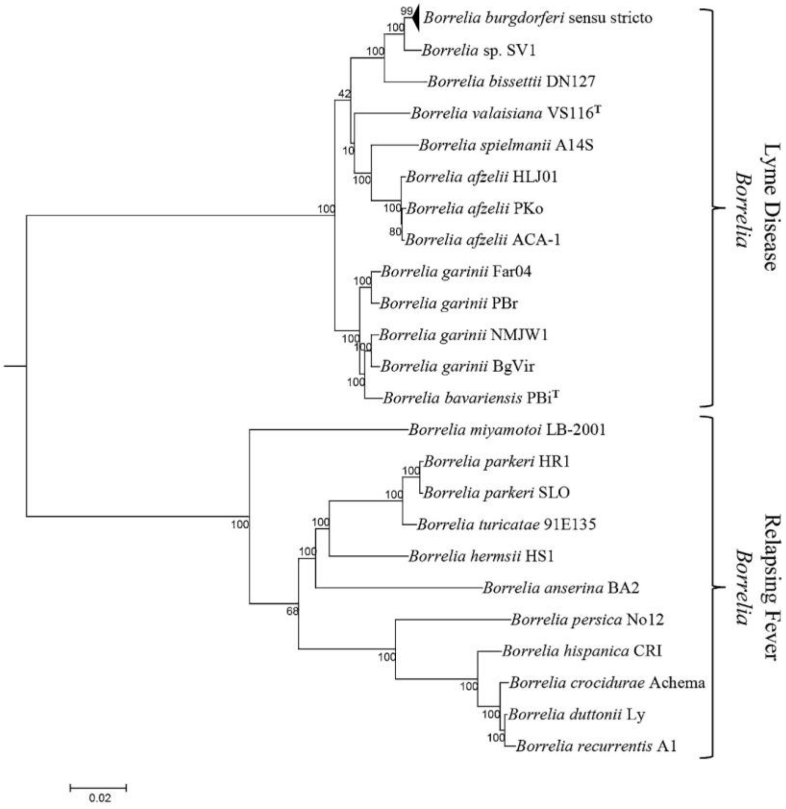


Fig. 1 A maximum likelihood phylogenetic tree of 38 sequenced members of the genus *Borrelia* based on the concatenated amino acid sequences of 25 conserved proteins. Bootstrap values are shown at *branch nodes*. The Lyme disease and relapsing fever clades of *Borrelia* are marked. The *letter*^{T} refers to the type strain of the species

homologues from members of the genus *Borrelia* as currently defined and absent in homologues from all other sequenced bacterial groups. Fifteen of these 31 CSIs are identified for the first time in this work, whereas the remaining 16 CSIs were identified in our earlier analysis of the phylum Spirochaetes (Gupta et al. 2013b). One example of a novel CSI that is uniquely found in all of the sequenced species from the genus *Borrelia* is shown in Fig. 3. In the example shown, a 3 aa insert in a conserved region of the bacterial rod-shaped determining protein MreB is uniquely present in all sequenced *Borrelia* species, but it is not found in sequences from any other Spirochaetes or other phyla of bacteria (Fig. 3). Sequence information for the 14 other novel CSIs that are also specific for the genus *Borrelia* is presented in Supp. Fig. 1–14 and a summary of all 31 *Borrelia* specific CSIs is presented in Table 2.

Our analyses have also identified 53 CSIs that are specific for or distinguish between the two main clades of *Borrelia* species, which are observed in the phylogenetic trees. Of these, seven CSIs are specific for the Lyme disease *Borrelia* clade, whereas another eight novel CSIs are uniquely found in the *Borrelia* species that are part of the relapsing fever clade. Examples of a CSI specific for the Lyme disease *Borrelia* clade and a CSI specific for the relapsing fever *Borrelia* clade are shown in Fig. 4. Figure 4a shows a 1 aa insert in a conserved region of Recombinase A that is uniquely found in all eight sequenced species from the Lyme disease *Borrelia* clade, whereas Fig. 4b shows a 1 aa deletion in the nicotinamide-nucleotide adenylyltransferase protein that is specific for members of the relapsing fever *Borrelia* clade. Sequence information for other CSIs that are specific for these two clades of *Borrelia*

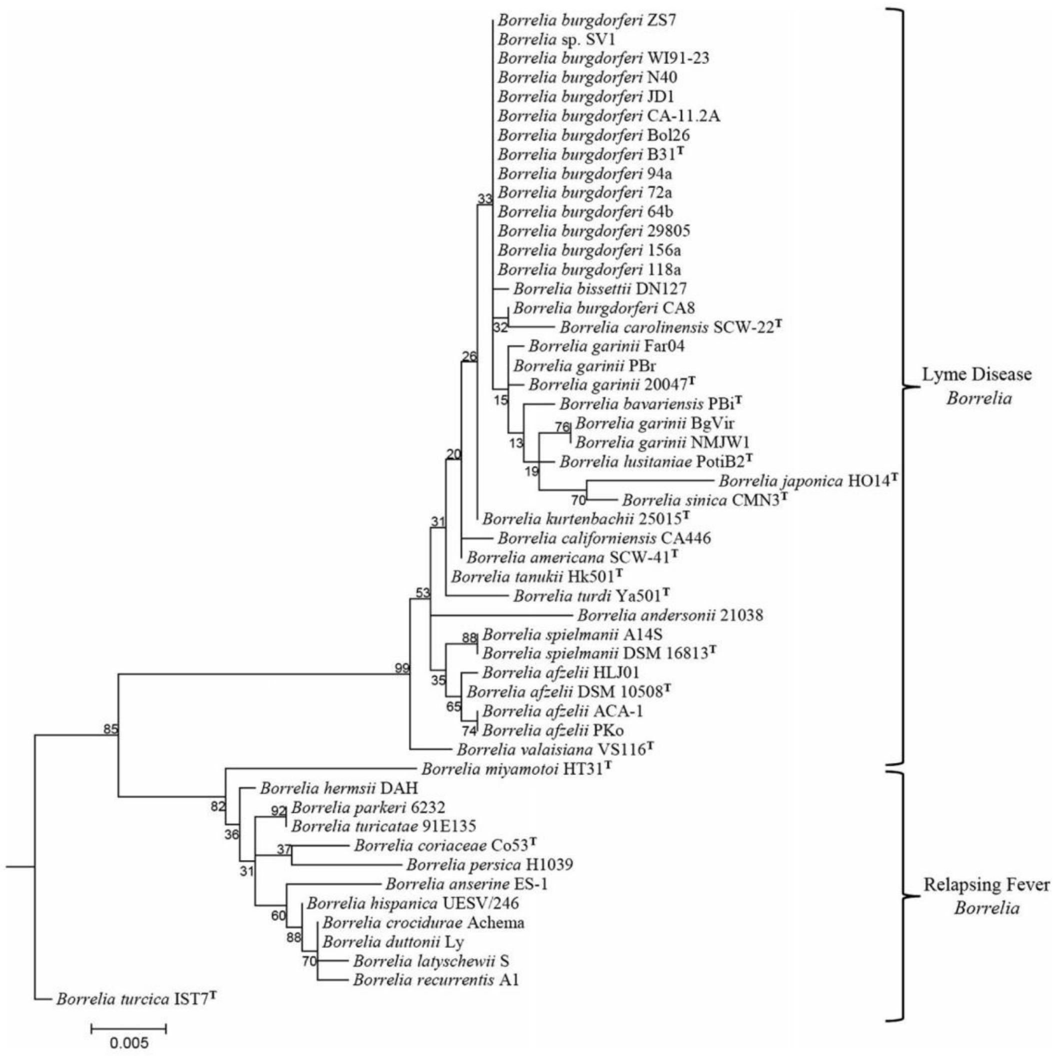
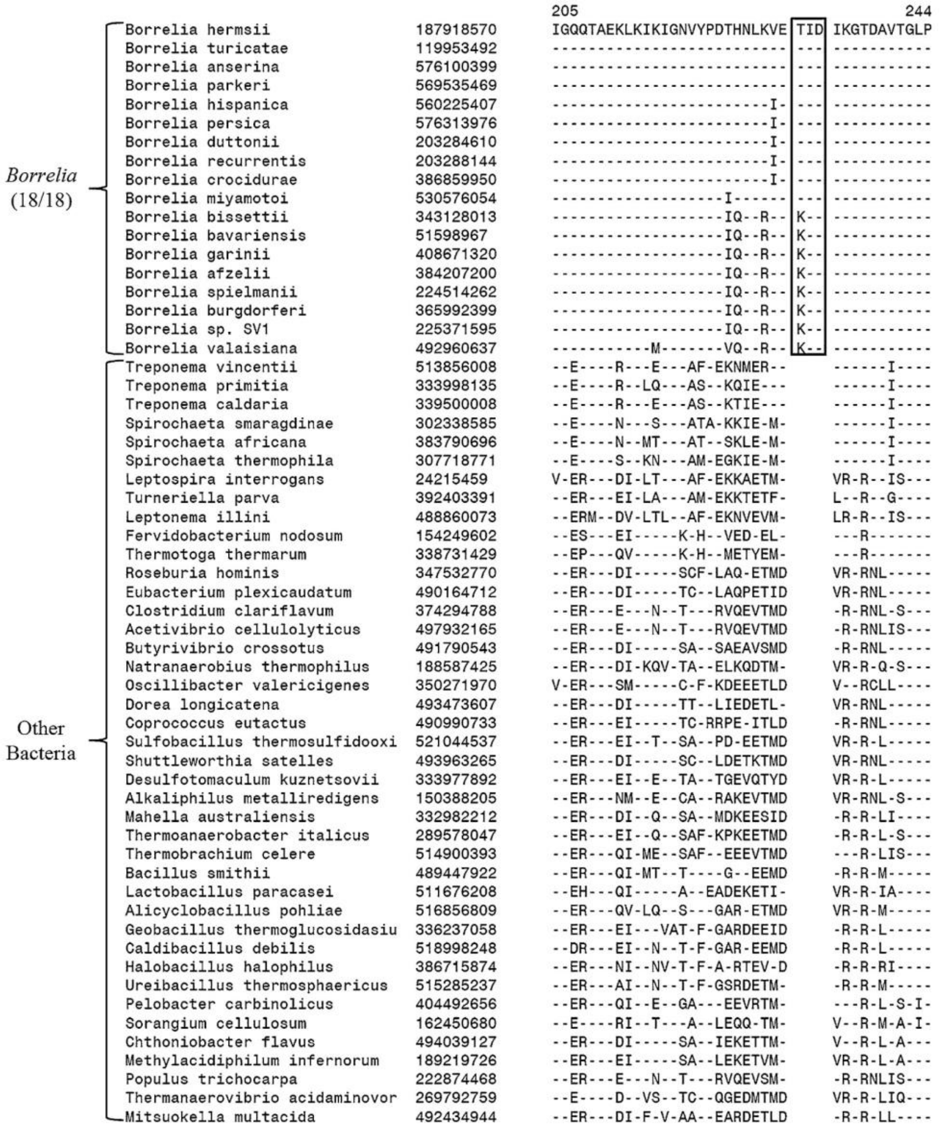


Fig. 2 A maximum likelihood tree based on the 16S rRNA gene sequences of representative strains of *Borrelia*. Bootstrap values are shown at *branch nodes*. The Lyme disease and relapsing fever clades of *Borrelia* are marked. The *letter* ^{T} refers to the type strain of the species. The accession numbers of the 16S rRNA gene sequences used in this analysis are provided in Supplemental Table 1

species are presented in Supp. Fig. 15–27 and Table 3. In addition to these 15 CSIs found in widely distributed proteins, 38 other CSIs in proteins that are mainly found in *Borrelia* species also serve to distinguish the Lyme disease *Borrelia* clade from the relapsing fever *Borrelia* clade. Because homologues for these proteins, or the conserved regions where these CSIs are present in these proteins, are not found in other bacteria, it is difficult to infer whether these CSIs represent insertions or deletions in the two groups. However, these CSIs still serve to distinguish between the two groups of *Borrelia*. One example of a 3 aa



```
                                             205                                     244
Borrelia hermsii               187918570    IGQQTAEKLKIKIGNVYPDTHNLKVE TID IKGTDAVTGLP
Borrelia turicatae             119953492    -------------------------- --- -----------
Borrelia anserina              576100399    -------------------------- --- -----------
Borrelia parkeri               569535469    -------------------------- --- -----------
Borrelia hispanica             560225407    ------------------------I- --- -----------
Borrelia persica               576313976    ------------------------I- --- -----------
Borrelia duttonii              203284610    ------------------------I- --- -----------
Borrelia recurrentis           203288144    ------------------------I- --- -----------
Borrelia crocidurae            386859950    ------------------------I- --- -----------
Borrelia miyamotoi             530576054    -------------------I------ --- -----------
Borrelia bissettii             343128013    -------------------IQ--R-- K-- -----------
Borrelia bavariensis           51598967     -------------------IQ--R-- K-- -----------
Borrelia garinii               408671320    -------------------IQ--R-- K-- -----------
Borrelia afzelii               384207200    -------------------IQ--R-- K-- -----------
Borrelia spielmanii            224514262    -------------------IQ--R-- K-- -----------
Borrelia burgdorferi           365992399    -------------------IQ--R-- K-- -----------
Borrelia sp. SV1               225371595    -------------------IQ--R-- K-- -----------
Borrelia valaisiana            492960637    -----------M-------VQ--R-- K-- -----------
Treponema vincentii            513856008    --E----R---E---AF-EKNMER--     ------I----
Treponema primitia             333998135    --E----R--LQ---AS--KQIE---     ------I----
Treponema caldaria             339500008    --E----R---E---AS--KTIE---     ------I----
Spirochaeta smaragdinae        302338585    --E----N---S---ATA-KKIE-M-     ------I----
Spirochaeta africana           383790696    --E----N--MT---AT--SKLE-M-     ------I----
Spirochaeta thermophila        307718771    --E----S--KN---AM-EGKIE-M-     ------I----
Leptospira interrogans         24215459     V-ER---DI-LT---AF-EKKAETM-     VR-R--IS---
Turneriella parva              392403391    --ER---EI-LA---AM-EKKTETF-     L--R--G----
Leptonema illini               488860073    --ERM--DV-LTL--AF-EKNVEVM-     LR-R--IS---
Fervidobacterium nodosum       154249602    --ES---EI-----K-H--VED-EL-     ---R-------
Thermotoga thermarum           338731429    --EP---QV-----K-H--METYEM-     ---R-------
Roseburia hominis              347532770    --ER---DI-----SCF-LAQ-ETMD     VR-RNL-----
Eubacterium plexicaudatum      490164712    --ER---DI-----TC--LAQPETID     VR-RNL-----
Clostridium clariflavum        374294788    --ER---E---N--T---RVQEVTMD     -R-RNL-S---
Acetivibrio cellulolyticus     497932165    --ER---E---N--T---RVQEVTMD     -R-RNLIS---
Butyrivibrio crossotus         491790543    --ER---DI-----SA--SAEAVSMD     -R-RNL-----
Natranaerobius thermophilus    188587425    --ER---DI-KQV-TA--ELKQDTM-     VR-R-Q-S---
Oscillibacter valericigenes    350271970    V-ER---SM-----C-F-KDEEETLD     V--RCLL----
Dorea longicatena              493473607    --ER---DI-----TT--LIEDETL-     VR-RNL-----
Coprococcus eutactus           490990733    --ER---EI-----TC-RRPE-ITLD     -R-RNL-----
Sulfobacillus thermosulfidooxi 521044537    --ER---EI--T--SA--PD-EETMD     VR-R-L-----
Shuttleworthia satelles        493963265    --ER---DI-----SC--LDETKTMD     VR-RNL-----
Desulfotomaculum kuznetsovii   333977892    --ER---EI--E--TA--TGEVQTYD     VR-R-L-----
Alkaliphilus metalliredigens   150388205    --ER---NM--E--CA--RAKEVTMD     VR-RNL-S---
Mahella australiensis          332982212    --ER---DI--Q--SA--MDKEESID     -R-R-LI----
Thermoanaerobacter italicus    289578047    --ER---EI--Q--SAF-KPKEETMD     -R-R-L-S---
Thermobrachium celere          514900393    --ER---QI-ME--SAF--EEEVTMD     ---R-LIS---
Bacillus smithii               489447922    --ER---QI-MT--T----G--EEMD     -R-R-M-----
Lactobacillus paracasei        511676208    --EH---QI-----A--EADEKETI-     VR-R-IA----
Alicyclobacillus pohliae       516856809    --ER---QV-LQ--S---GAR-ETMD     VR-R-M-----
Geobacillus thermoglucosidasiu 336237058    --ER---EI---VAT-F-GARDEEID     -R-R-L-----
Caldibacillus debilis          518998248    --DR---EI--N--T-F-GAR-EEMD     -R-R-L-----
Halobacillus halophilus        386715874    --ER---NI--NV-T-F-A-RTEV-D     -R-R-RI----
Ureibacillus thermosphaericus  515285237    --ER---AI--N--T-F-GSRDETM-     -R-R-M-----
Pelobacter carbinolicus        404492656    --ER---QI--E--GA---EEVRTM-     ---R-L-S-I-
Sorangium cellulosum           162450680    --E----RI--T---A--LEQQ-TM-     V--R-M-A-I-
Chthoniobacter flavus          494039127    --ER---DI-----SA--IEKETTM-     V--R-L-A---
Methylacidiphilum infernorum   189219726    --ER---EI-----SA--LEKETVM-     VR-R-L-A---
Populus trichocarpa            222874468    --ER---E---N--T---RVQEVSM-     -R-RNLIS---
Thermanaerovibrio acidaminovor 269792759    --E----D--VS--TC--QGEDMTMD     VR-R-LIQ---
Mitsuokella multacida          492434944    --ER---DI-F-V-AA--EARDETLD     -R-R-LL----
```

Fig. 3 A partial sequence alignment of the rod shape-determining protein *MreB*, showing a CSI (*boxed*) that is uniquely present in all members of the genus *Borrelia.* Sequence information for a single *Borrelia* strain from each of the 18 sequenced *Borrelia* species and a limited number other bacteria is shown here, but unless otherwise indicated similar CSIs were detected in all members of the indicated group and not detected in any other bacterial species in the top 250 BLAST hits. The *dashes* in the alignments indicate identity with the residue in the top sequence. GenBank identification (GI) numbers for each sequence are indicated in the *second column*. Sequence information for 30 other CSIs that are specific for all sequenced *Borrelia* species is provided in Supplemental figures 1–14 and Table 2

indel in a *Borrelia* specific protein of unknown function that distinguishes the Lyme disease *Borrelia* clade from the relapsing fever *Borrelia* clade is shown in Fig. 5. Sequence information for 37 other CSIs in different proteins that are of a similar kind is presented in Supp. Fig. 28–64 and Table 4.

Table 2 Conserved signature indels that are specific for all sequenced *Borrelia* species (both the Lyme disease *Borrelia* (*Borreliella*) and the relapsing fever *Borrelia*)

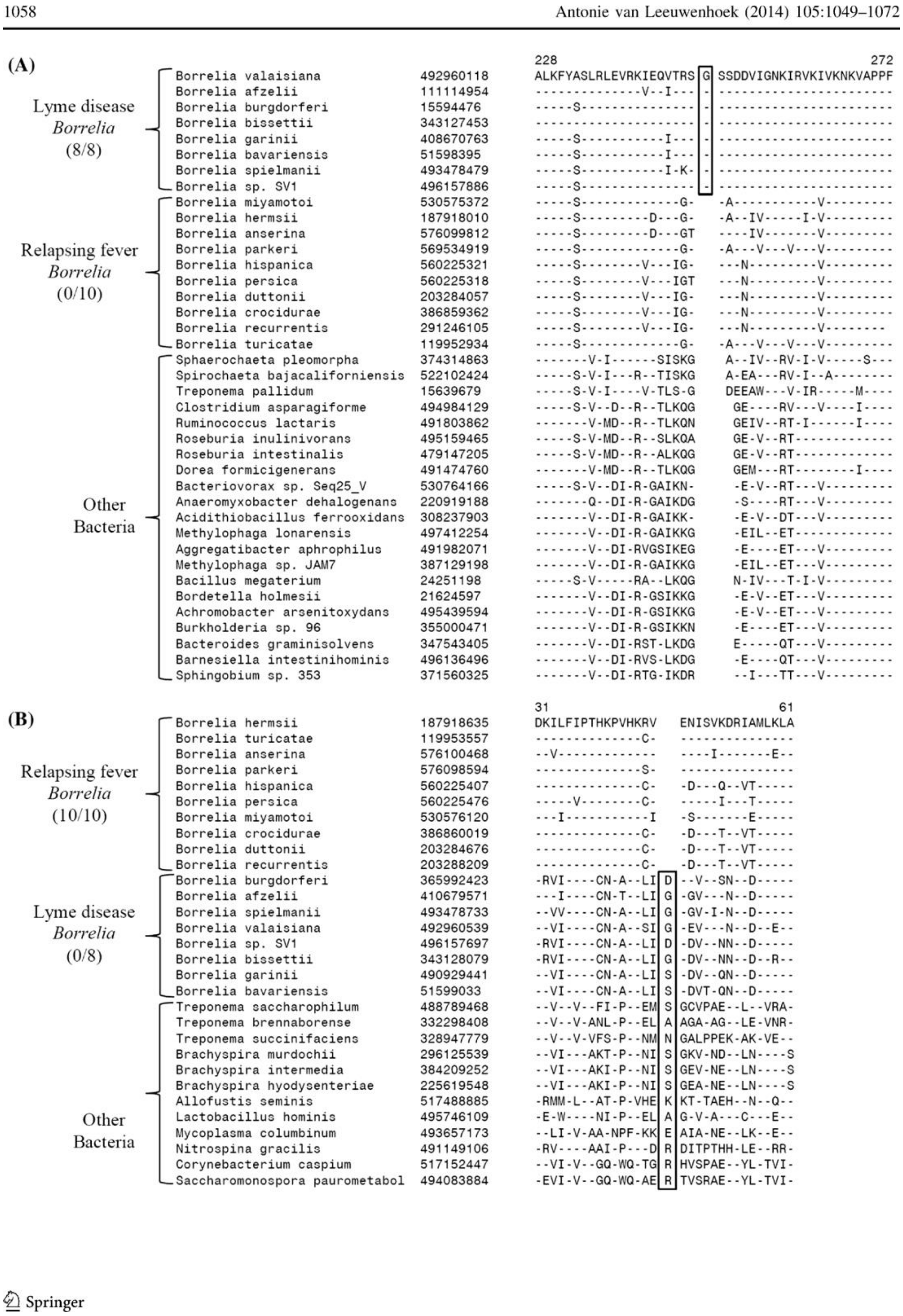
Protein Name	GI number	Figure number	Indel size	Indel position
Rod shape-determining protein MreB	187918570	Figure 3	3 aa ins	205–244
Flagellar motor switch protein FliM	119953077	Sup. Fig. 1	7 aa ins	8–65
ATP-dependent protease peptidase subunit	119953095	Sup. Fig. 2	3 aa ins	60–91
Mg^{2+} transporter MgtE	119953171	Sup. Fig. 3	1 aa ins	163–230
Mg^{2+} transporter MgtE	119953171	Sup. Fig. 4	4 aa ins	347–412
Cobyric acid synthase CobQ[a]	187918297	Sup. Fig. 5	1 aa del	147–184
Jag protein	119953232	Sup. Fig. 6	2 aa ins	121–180
CTP synthetase	119953361	Sup. Fig. 7	6 aa ins[b]	388–411
Chaperonin GroEL	187918505	Sup. Fig. 8	1 aa del	310–379
Ribose ABC transporter ATP-binding protein	15595022	Sup. Fig. 9	30 aa ins[b]	356–427
Phosphopantetheine adenylyltransferase	51598955	Sup. Fig. 10	2 aa ins	31–90
Asparaginyl-tRNA synthetase	365992288	Sup. Fig. 11	1 aa del	186–236
Chemotaxis protein CheY	119953336	Sup. Fig. 12	4 aa ins	66–121
Zn-ribbon protein	187918568	Sup. Fig. 13	1 aa ins	204–236
Chemotaxis protein CheW	15594910	Sup. Fig. 14	10 aa ins	51–119
Phosphofructokinase	219685531	(Gupta et al. 2013b)	6 aa ins	275–319
50S ribosomal protein L4	224534698	(Gupta et al. 2013b)	1 aa ins	103–136
tRNA pseudouridine 55 synthase	203284699	(Gupta et al. 2013b)	2 aa ins	143–178
Translation elongation factor Tu	203284386	(Gupta et al. 2013b)	1 aa del	330–369
Histidyl-tRNA synthetase	187918014	(Gupta et al. 2013b)	1 aa del	273–301
Seryl-tRNA synthetase	187918098	(Gupta et al. 2013b)	1 aa del	231–264
Spoiiij-associtated protein	219684344	(Gupta et al. 2013b)	3 aa ins	114–154
Nicotinate phosphoribosyltransferase	187918492	(Gupta et al. 2013b)	1 aa del	134–159
Ribose 5-phosphate isomerase	119953435	(Gupta et al. 2013b)	1 aa ins	86–110
Ribonuclease Z	195941574	(Gupta et al. 2013b)	2 aa ins	64–94
Hypothetical protein BGAFAR04_0762	386859948	(Gupta et al. 2013b)	1 aa ins	206–236
Signal recognition particle, subunit FFH/SRP54	119953471	(Gupta et al. 2013b)	1 aa ins	374–412
Hypothetical protein BSV1_0075	15594416	(Gupta et al. 2013b)	1 aa del	52–97
Aspartyl/glutamyl-tRNA amidotransferase subunit A	119953137	(Gupta et al. 2013b)	1 aa ins	364–402
Ribosomal RNA methyltransferase	203284234	(Gupta et al. 2013b)	1 aa ins	15–48
LysM domain/M23/M37 peptidase domain protein	224534310	(Gupta et al. 2013b)	1 aa ins	320–365

[a] Protein or indel containing region of the protein missing in two members of the *Borrelia*

[b] Indel was of different size in Lyme disease and Relapsing fever *Borrelia*

Conserved signature proteins which are specific for *Borrelia* or distinguish its two clades

Another useful category of molecular markers whose discovery has been enabled by comparative genomic analysis are conserved signature proteins (CSPs) that are uniquely present in different lineages of prokaryotes. Due to the specific presence of these genes/proteins in particular lineages of bacteria, they again provide useful molecular markers of common evolutionary decent for identifying and demarcating different bacterial groups in clear molecular terms. Our analyses of *Borrelia* genomes in this regard have led to identification of 107 proteins which are uniquely found either in all (or most) sequenced *Borrelia* species or are specific for only the Lyme disease *Borrelia* clade or the relapsing fever *Borrelia* clade. The results of BLAST searches for three CSPs that are specific to either all sequenced *Borrelia*, members of the Lyme disease *Borrelia*, or members of the



1058 Antonie van Leeuwenhoek (2014) 105:1049–1072

(A)

```
                                                                   228                                              272
                  Borrelia valaisiana              492960118       ALKFYASLRLEVRKIEQVTRS G SSDDVIGNKIRVKIVKNKVAPPF
                  Borrelia afzelii                 111114954       --------------V--I--- - -----------------------
Lyme disease      Borrelia burgdorferi             15594476        -----S--------------- - -----------------------
Borrelia          Borrelia bissettii               343127453       --------------------- - -----------------------
(8/8)             Borrelia garinii                 408670763       -----S-----------I--- - -----------------------
                  Borrelia bavariensis             51598395        -----S-----------I--- - -----------------------
                  Borrelia spielmanii              493478479       -----S-----------I-K- - -----------------------
                  Borrelia sp. SV1                 496157886       -----S--------------- - -----------------------
                  Borrelia miyamotoi               530575372       -----S-------------G-   -A-----------V---------
                  Borrelia hermsii                 187918010       -----S---------D---G-   -A--IV-----I-V---------
                  Borrelia anserina                576099812       -----S---------D---GT   ----IV-------V---------
Relapsing fever   Borrelia parkeri                 569534919       -----S-------------G-   -A---V---V---V---------
Borrelia          Borrelia hispanica               560225321       -----S--------V---IG-   ---N---------V---------
(0/10)            Borrelia persica                 560225318       -----S--------V---IGT   ---N---------V---------
                  Borrelia duttonii                203284057       -----S--------V---IG-   ---N---------V---------
                  Borrelia crocidurae              386859362       -----S--------V---IG-   ---N---------V---------
                  Borrelia recurrentis             291246105       -----S--------V---IG-   ---N---------V--------
                  Borrelia turicatae               119952934       -----S-------------G-   -A---V---V---V---------
                  Sphaerochaeta pleomorpha         374314863       -------V-I------SISKG    A--IV--RV-I-V-----S---
                  Spirochaeta bajacaliforniensis   522102424       -----S-V-I---R--TISKG    A-EA---RV-I--A--------
                  Treponema pallidum               15639679        -----S-V-I----V-TLS-G    DEEAW---V-IR-----M----
                  Clostridium asparagiforme        494984129       -----S-V--D--R--TLKQG     GE----RV---V----I----
                  Ruminococcus lactaris            491803862       -------V-MD--R--TLKQN     GEIV--RT-I------I----
                  Roseburia inulinivorans          495159465       -----S-V-MD--R--SLKQA     GE-V--RT-------------
                  Roseburia intestinalis           479147205       -----S-V-MD--R--ALKQG     GE-V--RT-------------
                  Dorea formicigenerans            491474760       -------V-MD--R--TLKQG     GEM---RT--------I----
                  Bacteriovorax sp. Seq25_V        530764166       -----S-V--DI-R-GAIKN-     -E-V--RT---V---------
Other             Anaeromyxobacter dehalogenans    220919188       -------Q--DI-R-GAIKDG     -S----RT---V---------
Bacteria          Acidithiobacillus ferrooxidans   308237903       -------V--DI-R-GAIKK-     -E-V--DT---V---------
                  Methylophaga lonarensis          497412254       -------V--DI-R-GAIKKG     -EIL--ET-------------
                  Aggregatibacter aphrophilus      491982071       -------V--DI-RVGSIKEG     -E----ET---V---------
                  Methylophaga sp. JAM7            387129198       -------V--DI-R-GAIKKG     -EIL--ET---V---------
                  Bacillus megaterium              24251198        -----S-V-----RA--LKQG     N-IV---T-I-V---------
                  Bordetella holmesii              21624597        -------V--DI-R-GSIKKG     -E-V--ET---V---------
                  Achromobacter arsenitoxydans     495439594       -------V--DI-R-GSIKKG     -E-V--ET---V---------
                  Burkholderia sp. 96              355000471       -------V--DI-R-GSIKKN     -E----ET---V---------
                  Bacteroides graminisolvens       347543405       -------V--DI-RST-LKDG     E-----QT---V---------
                  Barnesiella intestinihominis     496136496       -------V--DI-RVS-LKDG     -E----QT---V---------
                  Sphingobium sp. 353              371560325       -------V--DI-RTG-IKDR     --I---TT---V---------
```

(B)

```
                                                                   31                             61
                  Borrelia hermsii                 187918635       DKILFIPTHKPVHKRV   ENISVKDRIAMLKLA
                  Borrelia turicatae               119953557       --------------C-   ---------------
                  Borrelia anserina                576100468       --V-------------   ----I-------E--
Relapsing fever   Borrelia parkeri                 576098594       --------------S-   ---------------
Borrelia          Borrelia hispanica               560225407       --------------C-   -D---Q--VT-----
(10/10)           Borrelia persica                 560225476       -----V--------C-   -----I---T-----
                  Borrelia miyamotoi               530576120       ---I-----------I   -S-------E-----
                  Borrelia crocidurae              386860019       --------------C-   -D---T--VT-----
                  Borrelia duttonii                203284676       --------------C-   -D---T--VT-----
                  Borrelia recurrentis             203288209       --------------C-   -D---T--VT-----
                  Borrelia burgdorferi             365992423       -RVI----CN-A--LI D --V--SN--D-----
                  Borrelia afzelii                 410679571       ---I----CN-T--LI G -GV---N--D-----
Lyme disease      Borrelia spielmanii              493478733       --VV----CN-A--LI G -GV-I-N--D-----
Borrelia          Borrelia valaisiana              492960539       --VI----CN-A--SI G -EV---N--D--E--
(0/8)             Borrelia sp. SV1                 496157697       -RVI----CN-A--LI D -DV--NN--D-----
                  Borrelia bissettii               343128079       -RVI----CN-A--LI G -DV--NN--D--R--
                  Borrelia garinii                 490929441       --VI----CN-A--LI S -DV--QN--D-----
                  Borrelia bavariensis             51599033        --VI----CN-A--LI S -DVT-QN--D-----
                  Treponema saccharophilum         488789468       --V--V--FI-P--EM S GCVPAE--L--VRA-
                  Treponema brennaborense          332298408       --V--V-ANL-P--EL A AGA-AG--LE-VNR-
                  Treponema succinifaciens         328947779       --V--V-VFS-P--NM N GALPPEK-AK-VE--
                  Brachyspira murdochii            296125539       --VI---AKT-P--NI S GKV-ND--LN----S
                  Brachyspira intermedia           384209252       --VI---AKI-P--NI S GEV-NE--LN----S
                  Brachyspira hyodysenteriae       225619548       --VI---AKI-P--NI S GEA-NE--LN----S
                  Allofustis seminis               517488885       -RMM-L--AT-P-VHE K KT-TAEH--N--Q--
Other             Lactobacillus hominis            495746109       -E-W----NI-P--EL A G-V-A---C---E--
Bacteria          Mycoplasma columbinum            493657173       --LI-V-AA-NPF-KK E AIA-NE--LK--E--
                  Nitrospina gracilis              491149106       -RV----AAI-P---D R DITPTHH-LE--RR-
                  Corynebacterium caspium          517152447       --VI-V--GQ-WQ-TG R HVSPAE--YL-TVI-
                  Saccharomonospora paurometabol   494083884       -EVI-V--GQ-WQ-AE R TVSRAE--YL-TVI-
```

Springer

◀ **Fig. 4** Partial sequence alignments of (**a**) the protein Recombinase A showing a one amino acid insertion (*boxed*) identified in members of the Lyme disease *Borrelia* (**b**) the protein Nicotinamide-nucleotide adenylyltransferase showing a one amino acid deletion identified in members of the relapsing fever *Borrelia*. These CSIs were not found in the sequence homologues from any other sequenced bacteria. Sequence information for other Lyme disease or relapsing fever *Borrelia* specific CSIs is presented in Supplemental figures 15–27 and summarized in Table 3

relapsing fever *Borrelia* are shown in Table 5. As seen from this Table, high scoring homologues for these proteins are only found in different *Borrelia* species belonging to their specified clades, but not in any other bacterial organism. Thus, similar to the CSIs, these CSPs again are distinctive characteristics of the species from these clades and provide valuable molecular markers for their identification and demarcation. Of the CSPs that we have identified, 82 proteins are uniquely present in all or most of the sequenced *Borrelia* species and they are likely distinguishing characteristics of all members of the recently described family *Borreliaceae* (Table 6; Gupta et al. 2013b). In some cases, the homologues of these proteins were not detected in a few isolated strains of *Borrelia*. However, in every case, the proteins were not present in any other bacterial group, suggesting that the strains lacking these homologues have either undergone gene loss or that they are earlier branching lineages within these clades. In addition to the CSPs that are specific for all *Borrelia* (or the family *Borreliaceae*), we have also identified 21 CSPs whose homologues are only found in the Lyme disease *Borrelia* (Table 7) and four other CSPs, which are restricted to members of the relapsing fever *Borrelia* (Table 7). Some characteristics of the different CSPs are summarized in Tables 6 and 7. The cellular functions of most of these CSPs are unknown, but they may be related to some of the distinguishing properties exhibited by their specified clades.

Average nucleotide analysis

DNA–DNA hybridization is a commonly used method to determine the relatedness of different organisms and for assignment of species to either the same or different genera (Thompson et al. 2013). However, concerns have been raised about the scalability and reproducibility of these studies (Rosselló-Mora 2006). The availability of genome sequences have now made it possible to calculate pairwise ANI values between different genomes, which are analogous to DNA homology values (Richter and Rosselló-Móra 2009). We have compared the ANI values for all available genome sequenced *Borrelia* species (Fig. 6). The ANI values for different members within the genus *Borrelia* range between 73.03 and 99.34 % identity. However, based upon the comparisons of the ANI values, the *Borrelia* species can again be divided into two distinct clusters. One cluster, consisting of the members of the Lyme disease *Borrelia*, had intercluster ANI values which ranged between 91.33 and 98.06 % identity. The other cluster, which consisted of the members of the relapsing fever *Borrelia*, had intercluster ANI values which ranged between 82.51 and 99.34 % identity (Fig. 6). In contrast to the high ANI values for species within the two clusters, the ANI values of *Borrelia* species between the members of the two clusters were significantly lower, ranging between 73.03 and 74.85 % identity, indicating that the members of these clusters are distinct from each other.

Discussion

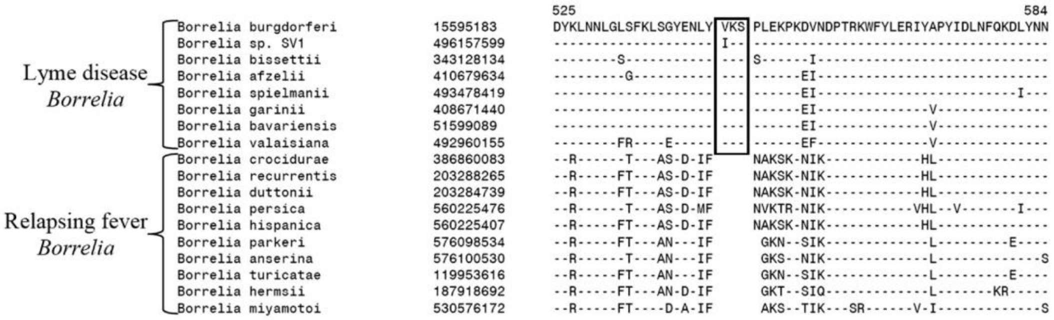
Genetic differences between the Lyme disease and relapsing fever *Borrelia* have been observed in a number of earlier studies (Postic et al. 1990; Fukunaga et al. 1996; Ras et al. 1996; Valsangiacomo et al. 1997; Margos et al. 2009). However, due to lack of distinct characteristics that can clearly distinguish the Lyme disease *Borrelia* from the relapsing fever *Borrelia*, it has proven difficult to reliably distinguish species from these two groups. This is responsible for the failure to diagnose or misdiagnosis of Lyme disease *Borrelia* in many individuals and also an underreporting of the overall incidence of this disease in the population (Wright et al. 2012; Ljostad and Mygland 2013). Detailed comparative analyses on genome sequences from *Borrelia* species that is reported here have identified numerous discrete molecular characteristics that are specifically shared by either members of the Lyme disease *Borrelia* clade or the relapsing fever *Borrelia* clade. The molecular markers described in this work provide novel and highly specific means for identification of members of the Lyme disease *Borrelia* group by either molecular sequence based (e.g. PCR, pyrosequencing, etc.) methods (Ahmod et al. 2011; Dunaj et al. 2013) or immunological methods (Wright et al. 2012; Ljostad and Mygland 2013).

Table 3 Conserved signature indels found in widely distributed proteins that are specific for either members of the Lyme disease *Borrelia* (*Borreliella*) or the relapsing fever *Borrelia*

Protein name	GI number	Figure number	Indel size	Indel position	Specificity
Recombinase A	492960118	Figure 4A	1 aa ins	228–272	Lyme disease *Borrelia*
Trigger factor Tig[a]	386854012	Sup. Fig. 15	2 aa ins	106–142	Lyme disease *Borrelia*
Chemotaxis protein CheY	15594760	Sup. Fig. 16	1 aa del	197–231	Lyme disease *Borrelia*
DNA polymerase III subunit beta	410679212	Sup. Fig. 17	1 aa del	135–176	Lyme disease *Borrelia*
Translation factor Sua5	15595079	Sup. Fig. 18	2 aa ins	149–182	Lyme disease *Borrelia*
Ferrous iron transporter A	51598605	Sup. Fig. 19	1 aa del	88–126	Lyme disease *Borrelia*
Glucose-6-phosphate isomerase	493478887	Sup. Fig. 20	1 aa ins	81–134	Lyme disease *Borrelia*
Nicotinamide-nucleotide adenylyltransferase	187918635	Figure 4B	1 aa del	31–61	Relapsing fever *Borrelia*
Hypothetical protein BRE16	203287484	Sup. Fig. 21	3 aa ins	64–98	Relapsing fever *Borrelia*
Hypothetical protein BDU327	203284245	Sup. Fig. 22	6 aa ins	866–907	Relapsing fever *Borrelia*
Hypothetical protein BT0471[b]	119953261	Sup. Fig. 23	1 aa del	216–261	Relapsing fever *Borrelia*
L-lactate permease	386859838	Sup. Fig. 24	1 aa ins	195–239	Relapsing fever *Borrelia*
1-phosphofructokinase	203288064	Sup. Fig. 25	1 aa del	101–139	Relapsing fever *Borrelia*
GTP-binding protein	203288075	Sup. Fig. 26	2 aa ins	42–87	Relapsing fever *Borrelia*
Sodium/panthothenate symporter	119953591	Sup. Fig. 27	1 aa ins	421–454	Relapsing fever *Borrelia*

[a] Indel also identified in one member of the relapsing fever *Borrelia*

[b] Protein or indel containing region of the protein missing in a member of the Lyme disease *Borrelia*



```
                                              525                                                            584
Borrelia burgdorferi        15595183          DYKLNNLGLSFKLSGYENLY VKS PLEKPKDVNDPTRKWFYLERIYAPYIDLNFQKDLYNN
Borrelia sp. SV1            496157599         -------------------- I-- -------------------------------------
Borrelia bissettii          343128134         --------S----------- --- S------I-----------------------------
Borrelia afzelii            410679634         ---------G---------- --- ------EI-----------------------------
Borrelia spielmanii         493478419         -------------------- --- ------EI-------------------------I---
Borrelia garinii            408671440         -------------------- --- ------EI--------------V--------------
Borrelia bavariensis        51599089          -------------------- --- ------EI--------------V--------------
Borrelia valaisiana         492960155         --------FR----E----- --- ------EF--------------V--------------
Borrelia crocidurae         386860083         --R------T---AS-D-IF     NAKSK-NIK------------HL--------------
Borrelia recurrentis        203288265         --R-----FT---AS-D-IF     NAKSK-NIK------------HL--------------
Borrelia duttonii           203284739         --R-----FT---AS-D-IF     NAKSK-NIK------------HL--------------
Borrelia persica            560225476         --R------T---AS-D-MF     NVKTR-NIK-----------VHL--V-------I---
Borrelia hispanica          560225407         --R-----FT---AS-D-IF     NAKSK-NIK------------HL--------------
Borrelia parkeri            576098534         --R-----FT---AN---IF      GKN--SIK-------------L---------E----
Borrelia anserina           576100530         --R------T---AN---IF      GKS--NIK-------------L-------------S
Borrelia turicatae          119953616         --R-----FT---AN---IF      GKN--SIK-------------L---------E----
Borrelia hermsii            187918692         --R-----FT---AN-D-IF      GKT--SIQ-------------L-------KR-----
Borrelia miyamotoi          530576172         --R-----FT----D-A-IF      AKS--TIK---SR------V-I-------------S
```

Fig. 5 A partial sequence alignment of a *Borrelia* lineage specific protein with currently unknown function (Hypothetical protein BB0838) showing a three amino acid insertion (*boxed*) which distinguishes the Lyme disease and relapsing fever *Borrelia*. Sequence information for other CSIs present in *Borrelia* lineage specific proteins is presented in Supplemental figures 28–64 and summarized in Table 4

The results reported here from multiple lines of investigations provide compelling evidence that the known *Borrelia* species are comprised of at least two evolutionary distinct groups of organisms corresponding to the Lyme disease *Borrelia* clade and the relapsing fever *Borrelia* clade. The different lines of investigation that support the distinctness of these two clades can be briefly summarized as follows:

1. In phylogenetic trees based on the 16S rRNA gene or concatenated sequences for 25 conserved proteins, the species from these two groups formed distinct and strongly supported clades that are separated from each other by long branches.
2. This work has identified 7 CSIs and 21 CSPs that are uniquely present in all of the genome

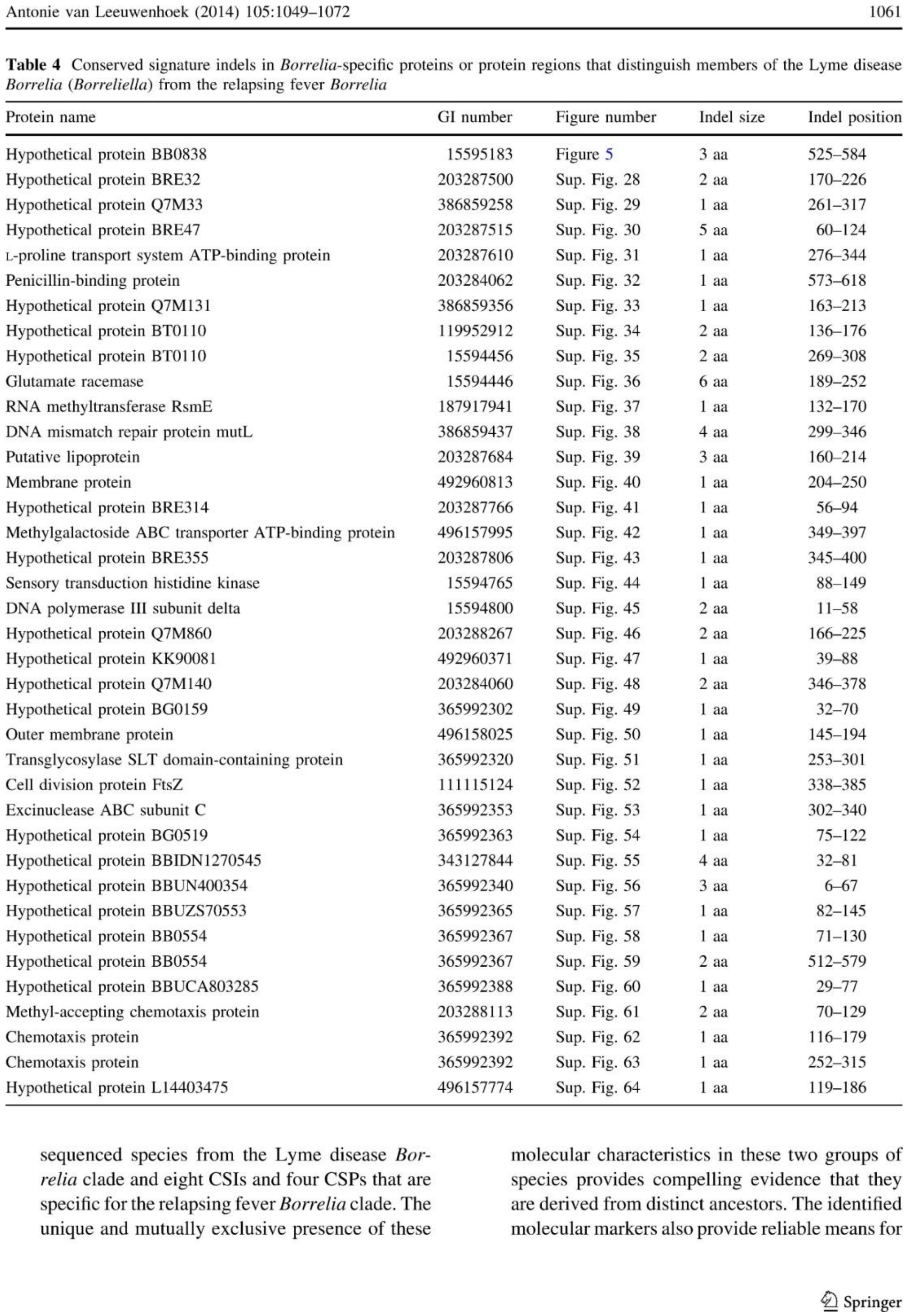


Table 4 Conserved signature indels in *Borrelia*-specific proteins or protein regions that distinguish members of the Lyme disease *Borrelia* (*Borreliella*) from the relapsing fever *Borrelia*

Protein name	GI number	Figure number	Indel size	Indel position
Hypothetical protein BB0838	15595183	Figure 5	3 aa	525–584
Hypothetical protein BRE32	203287500	Sup. Fig. 28	2 aa	170–226
Hypothetical protein Q7M33	386859258	Sup. Fig. 29	1 aa	261–317
Hypothetical protein BRE47	203287515	Sup. Fig. 30	5 aa	60–124
L-proline transport system ATP-binding protein	203287610	Sup. Fig. 31	1 aa	276–344
Penicillin-binding protein	203284062	Sup. Fig. 32	1 aa	573–618
Hypothetical protein Q7M131	386859356	Sup. Fig. 33	1 aa	163–213
Hypothetical protein BT0110	119952912	Sup. Fig. 34	2 aa	136–176
Hypothetical protein BT0110	15594456	Sup. Fig. 35	2 aa	269–308
Glutamate racemase	15594446	Sup. Fig. 36	6 aa	189–252
RNA methyltransferase RsmE	187917941	Sup. Fig. 37	1 aa	132–170
DNA mismatch repair protein mutL	386859437	Sup. Fig. 38	4 aa	299–346
Putative lipoprotein	203287684	Sup. Fig. 39	3 aa	160–214
Membrane protein	492960813	Sup. Fig. 40	1 aa	204–250
Hypothetical protein BRE314	203287766	Sup. Fig. 41	1 aa	56–94
Methylgalactoside ABC transporter ATP-binding protein	496157995	Sup. Fig. 42	1 aa	349–397
Hypothetical protein BRE355	203287806	Sup. Fig. 43	1 aa	345–400
Sensory transduction histidine kinase	15594765	Sup. Fig. 44	1 aa	88–149
DNA polymerase III subunit delta	15594800	Sup. Fig. 45	2 aa	11–58
Hypothetical protein Q7M860	203288267	Sup. Fig. 46	2 aa	166–225
Hypothetical protein KK90081	492960371	Sup. Fig. 47	1 aa	39–88
Hypothetical protein Q7M140	203284060	Sup. Fig. 48	2 aa	346–378
Hypothetical protein BG0159	365992302	Sup. Fig. 49	1 aa	32–70
Outer membrane protein	496158025	Sup. Fig. 50	1 aa	145–194
Transglycosylase SLT domain-containing protein	365992320	Sup. Fig. 51	1 aa	253–301
Cell division protein FtsZ	111115124	Sup. Fig. 52	1 aa	338–385
Excinuclease ABC subunit C	365992353	Sup. Fig. 53	1 aa	302–340
Hypothetical protein BG0519	365992363	Sup. Fig. 54	1 aa	75–122
Hypothetical protein BBIDN1270545	343127844	Sup. Fig. 55	4 aa	32–81
Hypothetical protein BBUN400354	365992340	Sup. Fig. 56	3 aa	6–67
Hypothetical protein BBUZS70553	365992365	Sup. Fig. 57	1 aa	82–145
Hypothetical protein BB0554	365992367	Sup. Fig. 58	1 aa	71–130
Hypothetical protein BB0554	365992367	Sup. Fig. 59	2 aa	512–579
Hypothetical protein BBUCA803285	365992388	Sup. Fig. 60	1 aa	29–77
Methyl-accepting chemotaxis protein	203288113	Sup. Fig. 61	2 aa	70–129
Chemotaxis protein	365992392	Sup. Fig. 62	1 aa	116–179
Chemotaxis protein	365992392	Sup. Fig. 63	1 aa	252–315
Hypothetical protein L14403475	496157774	Sup. Fig. 64	1 aa	119–186

sequenced species from the Lyme disease *Borrelia* clade and eight CSIs and four CSPs that are specific for the relapsing fever *Borrelia* clade. The unique and mutually exclusive presence of these molecular characteristics in these two groups of species provides compelling evidence that they are derived from distinct ancestors. The identified molecular markers also provide reliable means for

Table 5 Species specificity of selected conserved signature proteins

Protein specificity (GI number) function	All *Borrelia* (15594428) hypothetical		Lyme disease *Borrelia* (*Borreliella*) (365992370) hypothetical		Relapsing fever *Borrelia* (203288331) inclusion protein	
Organism	E value[a]	Length	E value[a]	Length	E value[a]	Length
Borrelia burgdorferi B31	0	432	$7.43e^{-111}$	174	–	–
Borrelia sp. SV1	0	432	$2.45e^{-103}$	174	–	–
Borrelia bissettii DN127	0	418	$4.98e^{-93}$	174	–	–
Borrelia spielmanii A14S	0	431	$1.95e^{-92}$	174	–	–
Borrelia garinii NMJW1	0	432	$4.27e^{-90}$	174	–	–
Borrelia afzelii HLJ01	0	431	$1.38e^{-88}$	174	–	–
Borrelia valaisiana VS116	0	431	$2.65e^{-94}$	174	–	–
Borrelia bavariensis PBi	0	432	$3.00e^{-89}$	174	–	–
Borrelia duttonii Ly	$9.83e^{-154}$	428	–	–	$3.00e^{-56}$	471
Borrelia crocidurae str. Achema	$1.39e^{-153}$	428	–	–	0	600
Borrelia recurrentis A1	$6.58e^{-153}$	428	–	–	0	622
Borrelia hispanica CRI	$1.00e^{-141}$	427	–	–	0	622
Borrelia persica No12	$4.00e^{-137}$	427	–	–	$1.00e^{-24}$	543
Borrelia turicatae 91E135	$1.63e^{-154}$	427	–	–	$4.00e^{-03}$	347
Borrelia hermsii HS1	$1.95e^{-145}$	427	–	–	$3.00e^{-11}$	575
Borrelia parkeri HR1	$4.00e^{-157}$	427	–	–	$3.00e^{-03}$	329
Borrelia anserina BA2	$1.00e^{-153}$	427	–	–	$2.00e^{-05}$	577
Borrelia miyamotoi LB-2001	$8.45e^{-145}$	428	–	–	$1.40e^{-02}$	146
Next Best BLAST Hit[b]	$7.09e^{00}$	1071	$4.00e^{00}$	1463	$1.07e^{02}$	1998

[a] E values smaller than $1.00e^{-180}$ are reported as 0

[b] Next best BLAST hits for protein 15594428, 365992370, and 203288331 are from *Leeuwenhoekiella blandensis*, *Trichomonas vaginalis*, and *Sulfolobus islandicus*, respectively

the demarcation of these two clades in molecular terms.

3. Whole genome ANI analyses of *Borrelia* genomes show that species from within either the Lyme disease *Borrelia* group or the relapsing fever *Borrelia* group had much higher ANI values when compared to other members of their group (range 82.51–99.34 %) than with members of the opposing *Borrelia* group (range 73.03–74.85 %).
4. The species from these two groups differ in terms of their pathogenicity profiles and the characteristics of the arthropod vectors which are involved in their transmission. The species which are part of the Lyme disease clade are transmitted via arthropod vectors that are hard tick species related to the *Ixodes ricinus* complex, while a majority of the members of the relapsing fever *Borrelia* clade are transferred by soft-bodied ticks within the family *Argasidae* (Table 8).

Taxonomic implications

The evidence obtained from different lines of investigations summarized above provides compelling evidence that the known *Borrelia* species are comprised of two main clades corresponding to the "Lyme disease *Borrelia* and its relatives" and the "relapsing fever *Borrelia* and its relatives". Of these two main groups, the Lyme disease *Borrelia* clade, based upon branching in the 16S rRNA gene tree and concatenated protein tree is comprised of the following 14 validly named species: *B. afzelii, B. americana, B. bavariensis, B. burgdorferi, B. carolinensis, B. garinii, B. japonica, B. kurtenbachii, B. lusitaniae, B. sinica, B. spielmanii, B. tanukii, B. turdi,* and *B. valaisiana*. All other currently validly named *Borrelia* species are part of the relapsing fever *Borrelia* clade. The observations presented in this work make a strong case for division of the existing genus *Borrelia* into two different

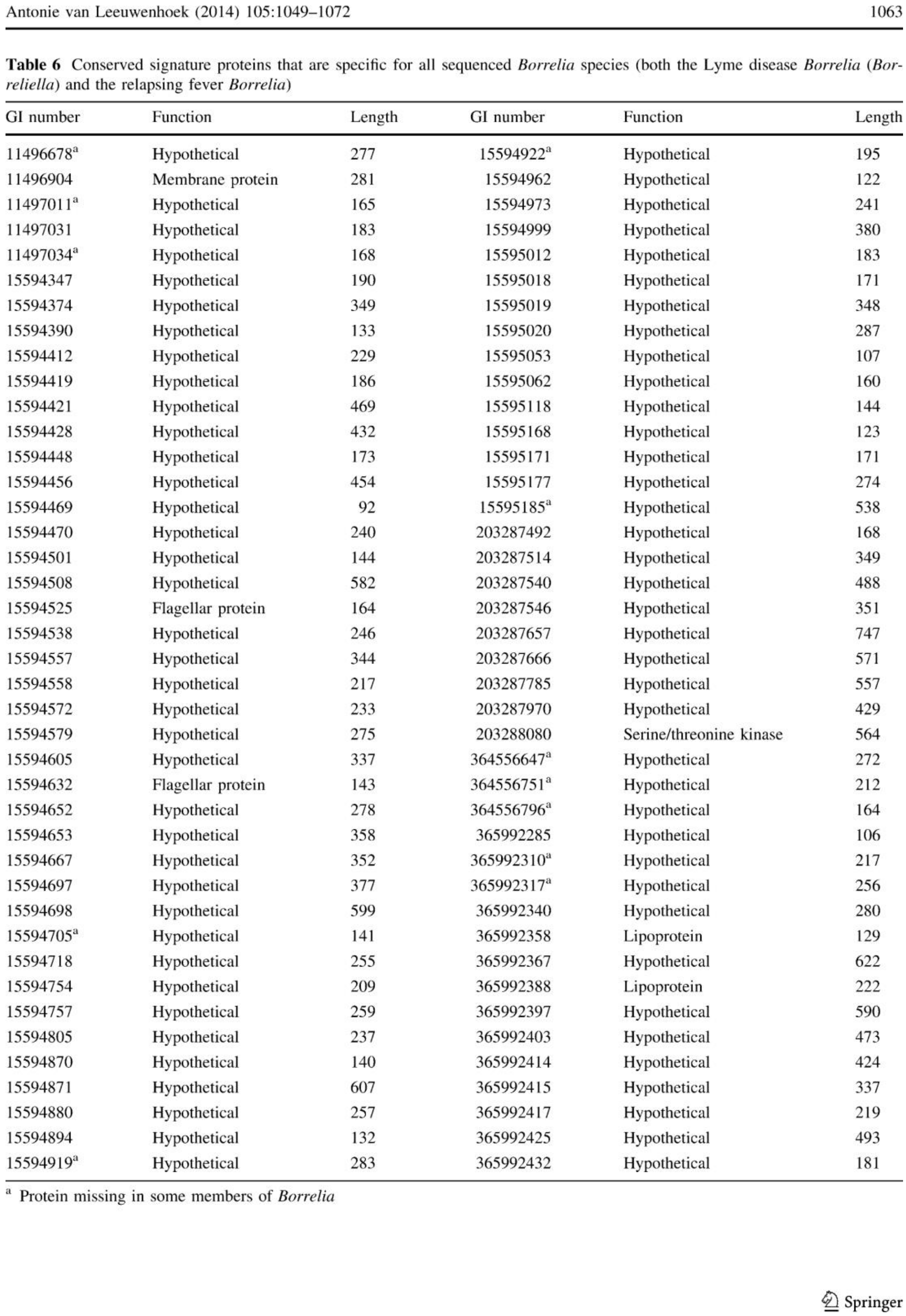


Table 6 Conserved signature proteins that are specific for all sequenced *Borrelia* species (both the Lyme disease *Borrelia* (*Borreliella*) and the relapsing fever *Borrelia*)

GI number	Function	Length	GI number	Function	Length
11496678[a]	Hypothetical	277	15594922[a]	Hypothetical	195
11496904	Membrane protein	281	15594962	Hypothetical	122
11497011[a]	Hypothetical	165	15594973	Hypothetical	241
11497031	Hypothetical	183	15594999	Hypothetical	380
11497034[a]	Hypothetical	168	15595012	Hypothetical	183
15594347	Hypothetical	190	15595018	Hypothetical	171
15594374	Hypothetical	349	15595019	Hypothetical	348
15594390	Hypothetical	133	15595020	Hypothetical	287
15594412	Hypothetical	229	15595053	Hypothetical	107
15594419	Hypothetical	186	15595062	Hypothetical	160
15594421	Hypothetical	469	15595118	Hypothetical	144
15594428	Hypothetical	432	15595168	Hypothetical	123
15594448	Hypothetical	173	15595171	Hypothetical	171
15594456	Hypothetical	454	15595177	Hypothetical	274
15594469	Hypothetical	92	15595185[a]	Hypothetical	538
15594470	Hypothetical	240	203287492	Hypothetical	168
15594501	Hypothetical	144	203287514	Hypothetical	349
15594508	Hypothetical	582	203287540	Hypothetical	488
15594525	Flagellar protein	164	203287546	Hypothetical	351
15594538	Hypothetical	246	203287657	Hypothetical	747
15594557	Hypothetical	344	203287666	Hypothetical	571
15594558	Hypothetical	217	203287785	Hypothetical	557
15594572	Hypothetical	233	203287970	Hypothetical	429
15594579	Hypothetical	275	203288080	Serine/threonine kinase	564
15594605	Hypothetical	337	364556647[a]	Hypothetical	272
15594632	Flagellar protein	143	364556751[a]	Hypothetical	212
15594652	Hypothetical	278	364556796[a]	Hypothetical	164
15594653	Hypothetical	358	365992285	Hypothetical	106
15594667	Hypothetical	352	365992310[a]	Hypothetical	217
15594697	Hypothetical	377	365992317[a]	Hypothetical	256
15594698	Hypothetical	599	365992340	Hypothetical	280
15594705[a]	Hypothetical	141	365992358	Lipoprotein	129
15594718	Hypothetical	255	365992367	Hypothetical	622
15594754	Hypothetical	209	365992388	Lipoprotein	222
15594757	Hypothetical	259	365992397	Hypothetical	590
15594805	Hypothetical	237	365992403	Hypothetical	473
15594870	Hypothetical	140	365992414	Hypothetical	424
15594871	Hypothetical	607	365992415	Hypothetical	337
15594880	Hypothetical	257	365992417	Hypothetical	219
15594894	Hypothetical	132	365992425	Hypothetical	493
15594919[a]	Hypothetical	283	365992432	Hypothetical	181

[a] Protein missing in some members of *Borrelia*

Table 7 Conserved signature proteins that are specific for either members of the Lyme disease *Borrelia* (*Borreliella*) or the relapsing fever *Borrelia*

GI number	Function	Length
CSPs that are specific for Lyme disease *Borrelia*		
11496594	Lipoprotein	192
11496595	Hypothetical	227
11496690[a]	Hypothetical	142
11496704	Hypothetical	155
11496896	S1 Antigen	417
11496905	Hypothetical	79
11496906	Lipoprotein	277
11496908	Lipoprotein	68
11496925	Membrane protein	257
11496937	Hypothetical	414
11496964	Lipoprotein	179
11496966	Hypothetical	201
11497026	Hypothetical	345
11497073[b]	Hemolysin	67
15594723	Hypothetical	220
15594749[a]	Hypothetical	138
15594801	Hypothetical	201
15594976[a]	Hypothetical	104
364556745[a]	Hypothetical	241
364556746[a]	Hypothetical	321
365992370	Hypothetical	174
CSPs that are specific for relapsing fever *Borrelia*		
203288331	Inclusion protein	622
203288332	Lipoprotein	619
203288333[a]	Lipoprotein	477
203288334[a]	Hypothetical	765

[a] Protein missing in some members of the specified clade

[b] Multiple copies of this CSP are present in the genome

genera corresponding to the Lyme disease *Borrelia* clade and the relapsing fever *Borrelia* clade. Ideally, the genus name *Borrelia* should be retained for the Lyme disease *Borrelia* clade, which includes the best known species from this genus, *B. burgdorferi*, the first identified causative agent of Lyme disease (Barbour 1984). However, the type species of the genus *Borrelia*, *Borrelia anserina*, is a part of the relapsing fever clade. Hence, the genus name *Borrelia* must be retained for the relapsing fever clade (Bergey 1925; Lapage et al. 1992; Wang and Schwartz 2011). Therefore, species from the Lyme disease clade must be transferred to a new genus indicating their distinctness from the relapsing fever clade (viz. the emended genus *Borrelia*). To minimize confusion among scientists and other health care professionals, we are proposing that the species that are part of the Lyme disease clade should be transferred to a new genus, *Borreliella* gen. nov. The proposed name retains much of the original name of the genus *Borrelia*, thus it is unlikely that the species with the new names (e.g. *B. burgdorferi*) could be confused with any other unrelated species. The emended description of the genus *Borrelia* and a description of the newly proposed genus, *Borreliella* gen. nov, containing 14 new combinations, are provided below.

Emended description of the genus *Borrelia* (Swellengrebel 1907) (approved lists 1980)

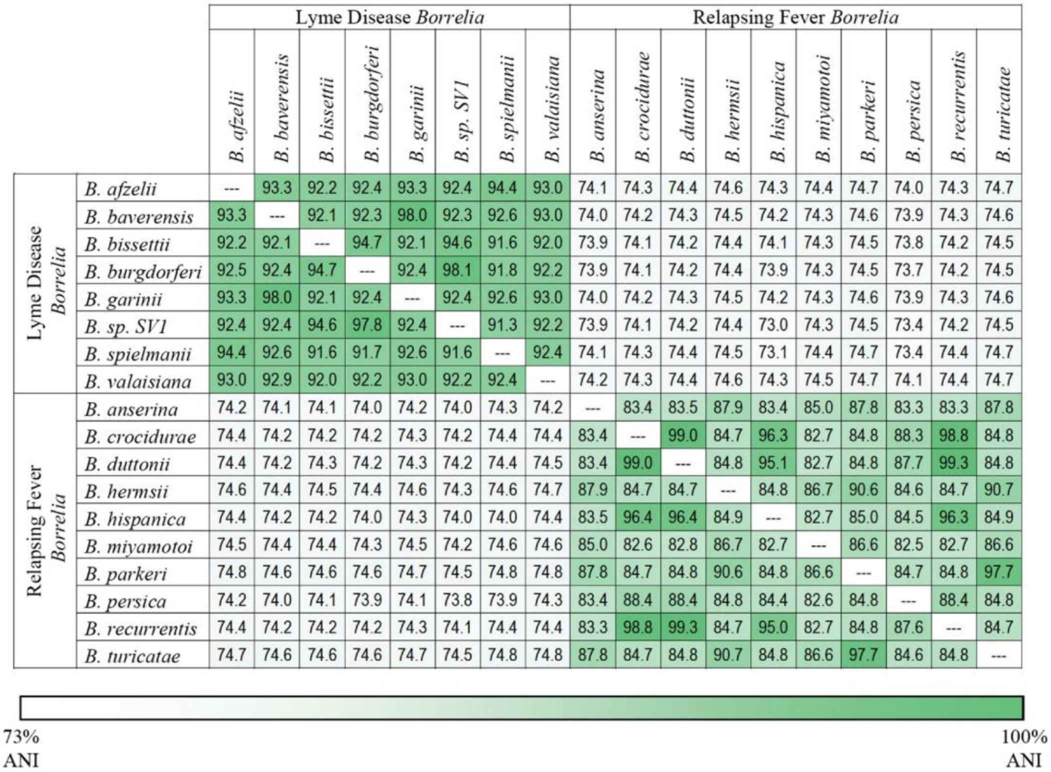
Organisms are helical, 0.2–3 μm in diameter and 3–180 μm in length. Cells do not have hooked ends. Periplasmic flagella overlap in the central region of the cell. Cells are motile, host-associated and microaerophilic. The diamino acid component of the peptidoglycan is L-ornithine. Organisms are chemoorganotrophic and utilize carbohydrates or amino acids as carbon and energy sources. Members of this genus are the causative agents of relapsing fever. The G+C content of the genomic DNA is 27–32 (mol%). The type species is *B. anserina* (Bergey 1925) (Approved Lists 1980) (Skerman et al. 1980).

Organisms from this genus are distinguished from all other bacteria examined to date by the CSIs and conserved signature proteins described in this report (Tables 3, 4, 7).

Description of *Borreliella* gen. nov.

Borreliella (Bor.re'li.el'la. N.L. fem. dim. n. *Borreliella*, named after Amédée Borrel, a French bacteriologist)

Organisms are helical, 0.2–0.3 μm in diameter and 20–30 μm in length. Cells do not have hooked ends. Periplasmic flagella overlap in the central region of the cell. Cells are motile, host-associated and microaerophilic. The diamino acid component of the peptidoglycan is L-ornithine. Organisms are chemoorganotrophic and utilize carbohydrates or amino acids as carbon and energy sources. Members of this genus are the causative agents of Lyme disease. The



		Lyme Disease *Borrelia*								Relapsing Fever *Borrelia*									
		B. afzelii	*B. baverensis*	*B. bissettii*	*B. burgdorferi*	*B. garinii*	*B. sp. SV1*	*B. spielmanii*	*B. valaisiana*	*B. anserina*	*B. crocidurae*	*B. duttonii*	*B. hermsii*	*B. hispanica*	*B. miyamotoi*	*B. parkeri*	*B. persica*	*B. recurrentis*	*B. turicatae*
Lyme Disease *Borrelia*	*B. afzelii*	---	93.3	92.2	92.4	93.3	92.4	94.4	93.0	74.1	74.3	74.4	74.6	74.3	74.4	74.7	74.0	74.3	74.7
	B. baverensis	93.3	---	92.1	92.3	98.0	92.3	92.6	93.0	74.0	74.2	74.3	74.5	74.2	74.3	74.6	73.9	74.3	74.6
	B. bissettii	92.2	92.1	---	94.7	92.1	94.6	91.6	92.0	73.9	74.1	74.2	74.4	74.1	74.3	74.5	73.8	74.2	74.5
	B. burgdorferi	92.5	92.4	94.7	---	92.4	98.1	91.8	92.2	73.9	74.1	74.2	74.4	73.9	74.3	74.5	73.7	74.2	74.5
	B. garinii	93.3	98.0	92.1	92.4	---	92.4	92.6	93.0	74.0	74.2	74.3	74.5	74.2	74.3	74.6	73.9	74.3	74.6
	B. sp. SV1	92.4	92.4	94.6	97.8	92.4	---	91.3	92.2	73.9	74.1	74.2	74.4	73.0	74.3	74.5	73.4	74.2	74.5
	B. spielmanii	94.4	92.6	91.6	91.7	92.6	91.6	---	92.4	74.1	74.3	74.4	74.5	73.1	74.4	74.7	73.4	74.4	74.7
	B. valaisiana	93.0	92.9	92.0	92.2	93.0	92.2	92.4	---	74.2	74.3	74.4	74.6	74.3	74.5	74.7	74.1	74.4	74.7
Relapsing Fever *Borrelia*	*B. anserina*	74.2	74.1	74.1	74.0	74.2	74.0	74.3	74.2	---	83.4	83.5	87.9	83.4	85.0	87.8	83.3	83.3	87.8
	B. crocidurae	74.4	74.2	74.2	74.2	74.3	74.2	74.4	74.4	83.4	---	99.0	84.7	96.3	82.7	84.8	88.3	98.8	84.8
	B. duttonii	74.4	74.2	74.3	74.2	74.3	74.2	74.4	74.5	83.4	99.0	---	84.8	95.1	82.7	84.8	87.7	99.3	84.8
	B. hermsii	74.6	74.4	74.5	74.4	74.6	74.3	74.6	74.7	87.9	84.7	84.7	---	84.8	86.7	90.6	84.6	84.7	90.7
	B. hispanica	74.4	74.2	74.2	74.0	74.3	74.0	74.0	74.4	83.5	96.4	96.4	84.9	---	82.7	85.0	84.5	96.3	84.9
	B. miyamotoi	74.5	74.4	74.4	74.3	74.5	74.2	74.6	74.6	85.0	82.6	82.8	86.7	82.7	---	86.6	82.5	82.7	86.6
	B. parkeri	74.8	74.6	74.6	74.6	74.7	74.5	74.8	74.8	87.8	84.7	84.8	90.6	84.8	86.6	---	84.7	84.8	97.7
	B. persica	74.2	74.0	74.1	73.9	74.1	73.8	73.9	74.3	83.4	88.4	88.4	84.8	84.4	82.6	84.8	---	88.4	84.8
	B. recurrentis	74.4	74.2	74.2	74.2	74.3	74.1	74.4	74.4	83.3	98.8	99.3	84.7	95.0	82.7	84.8	87.6	---	84.7
	B. turicatae	74.7	74.6	74.6	74.6	74.7	74.5	74.8	74.8	87.8	84.7	84.8	90.7	84.8	86.6	97.7	84.6	84.8	---

Fig. 6 A summary of the results of average nucleotide identity analysis performed on the sequenced members of the genus *Borrelia*. Cells with higher ANI values are *highlighted*. ANI results for multiple strains of the same species have been averaged

G+C content of the genomic DNA is 26–29 (mol%). The type species is *B. burgdorferi* comb. nov.

Organisms from this genus are distinguished from all other bacteria examined to date by the CSIs and conserved signature proteins described in this report (Tables 3, 4, 7).

Description of *Borreliella afzelii* comb. nov.

Basonym: *Borrelia afzelii* (Canica et al. 1994)

The description of the species is the same as the description given for *B. afzelii* by Canica et al. (1994). The species exhibits the genus properties and contains the CSIs and CSPs indicated in the description of *Borreliella*.

Type Strain: VS461^{T} (=ATCC 51567^{T} = CIP 103469^{T} = DSM 10508^{T})

Description of *Borreliella americana* comb. nov.

Basonym: *Borrelia americana* (Rudenko et al. 2010)

The description of the species is the same as the description given for *B. americana* by Rudenko et al. (2010). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: SCW-41^{T} (=ATCC BAA-1877^{T} = DSM 22541^{T})

Description of *Borreliella bavariensis* comb. nov.

Basonym: *Borrelia bavariensis* (Margos et al. 2013b)

The description of the species is the same as the description given for *B. bavariensis* by Margos et al. (2013a, b). The species exhibits the genus properties and contains the CSIs and CSPs indicated in the description of *Borreliella*.

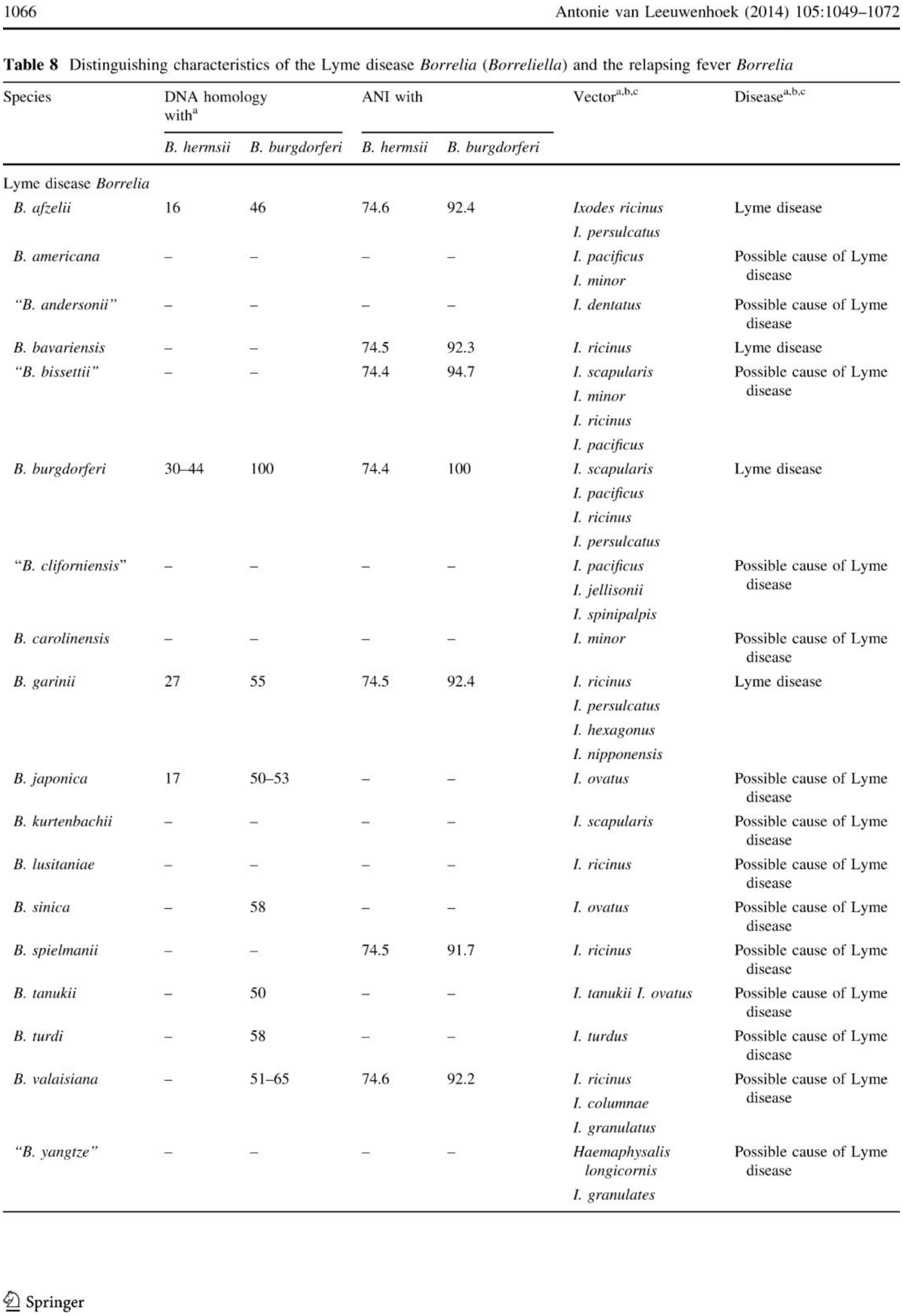


Table 8 Distinguishing characteristics of the Lyme disease *Borrelia* (*Borreliella*) and the relapsing fever *Borrelia*

Species	DNA homology with[a]		ANI with		Vector[a,b,c]	Disease[a,b,c]
	B. hermsii	*B. burgdorferi*	*B. hermsii*	*B. burgdorferi*		
Lyme disease *Borrelia*						
B. afzelii	16	46	74.6	92.4	*Ixodes ricinus* *I. persulcatus*	Lyme disease
B. americana	–	–	–	–	*I. pacificus* *I. minor*	Possible cause of Lyme disease
"B. andersonii"	–	–	–	–	*I. dentatus*	Possible cause of Lyme disease
B. bavariensis	–	–	74.5	92.3	*I. ricinus*	Lyme disease
"B. bissettii"	–	–	74.4	94.7	*I. scapularis* *I. minor* *I. ricinus* *I. pacificus*	Possible cause of Lyme disease
B. burgdorferi	30–44	100	74.4	100	*I. scapularis* *I. pacificus* *I. ricinus* *I. persulcatus*	Lyme disease
"*B. cliforniensis*"	–	–	–	–	*I. pacificus* *I. jellisonii* *I. spinipalpis*	Possible cause of Lyme disease
B. carolinensis	–	–	–	–	*I. minor*	Possible cause of Lyme disease
B. garinii	27	55	74.5	92.4	*I. ricinus* *I. persulcatus* *I. hexagonus* *I. nipponensis*	Lyme disease
B. japonica	17	50–53	–	–	*I. ovatus*	Possible cause of Lyme disease
B. kurtenbachii	–	–	–	–	*I. scapularis*	Possible cause of Lyme disease
B. lusitaniae	–	–	–	–	*I. ricinus*	Possible cause of Lyme disease
B. sinica	–	58	–	–	*I. ovatus*	Possible cause of Lyme disease
B. spielmanii	–	–	74.5	91.7	*I. ricinus*	Possible cause of Lyme disease
B. tanukii	–	50	–	–	*I. tanukii I. ovatus*	Possible cause of Lyme disease
B. turdi	–	58	–	–	*I. turdus*	Possible cause of Lyme disease
B. valaisiana	–	51–65	74.6	92.2	*I. ricinus* *I. columnae* *I. granulatus*	Possible cause of Lyme disease
"B. yangtze"	–	–	–	–	*Haemaphysalis longicornis* *I. granulates*	Possible cause of Lyme disease

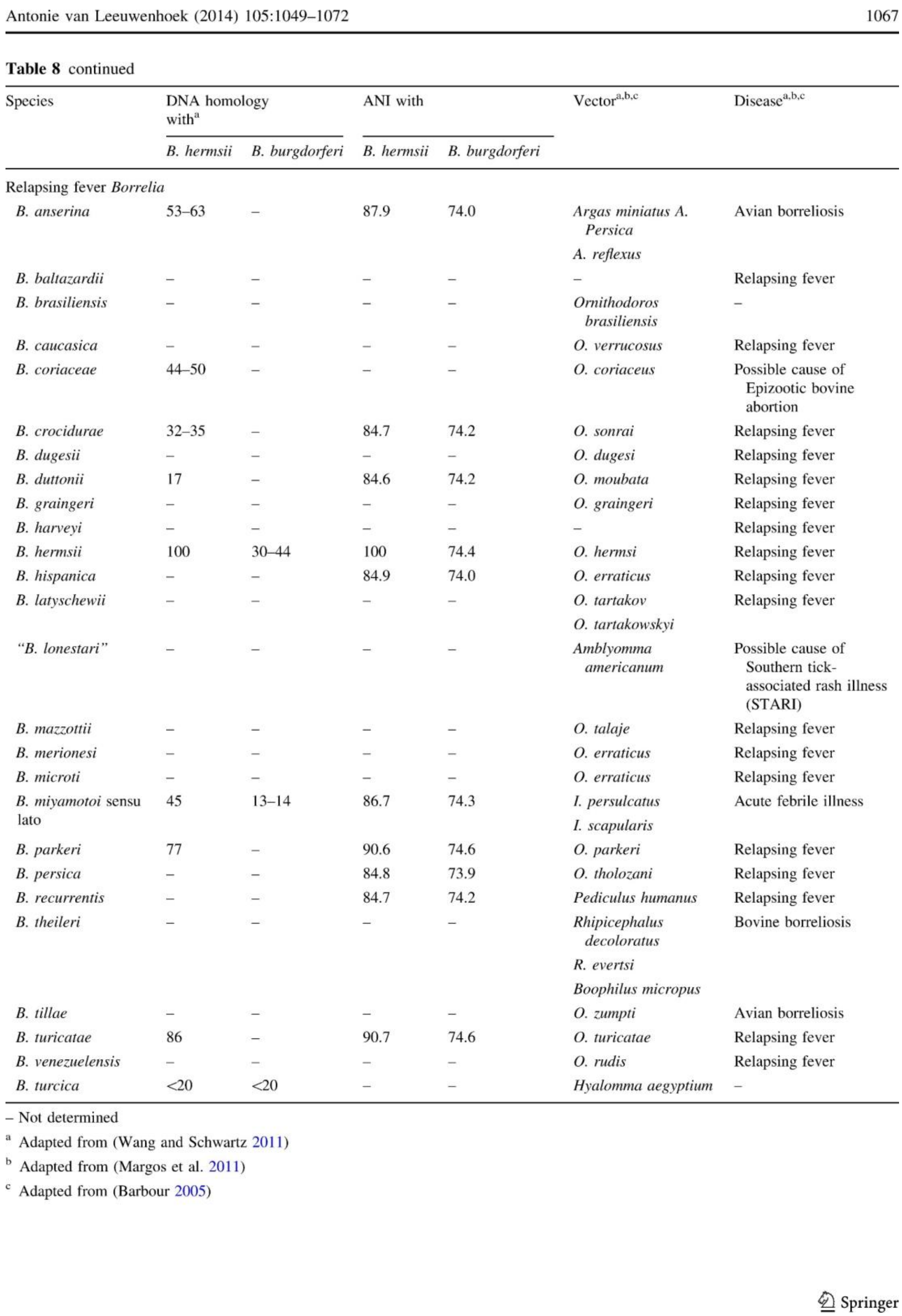


Table 8 continued

Species	DNA homology with[a]		ANI with		Vector[a,b,c]	Disease[a,b,c]
	B. hermsii	*B. burgdorferi*	*B. hermsii*	*B. burgdorferi*		
Relapsing fever *Borrelia*						
B. anserina	53–63	–	87.9	74.0	*Argas miniatus A. Persica* *A. reflexus*	Avian borreliosis
B. baltazardii	–	–	–	–	–	Relapsing fever
B. brasiliensis	–	–	–	–	*Ornithodoros brasiliensis*	–
B. caucasica	–	–	–	–	*O. verrucosus*	Relapsing fever
B. coriaceae	44–50	–	–	–	*O. coriaceus*	Possible cause of Epizootic bovine abortion
B. crocidurae	32–35	–	84.7	74.2	*O. sonrai*	Relapsing fever
B. dugesii	–	–	–	–	*O. dugesi*	Relapsing fever
B. duttonii	17	–	84.6	74.2	*O. moubata*	Relapsing fever
B. graingeri	–	–	–	–	*O. graingeri*	Relapsing fever
B. harveyi	–	–	–	–	–	Relapsing fever
B. hermsii	100	30–44	100	74.4	*O. hermsi*	Relapsing fever
B. hispanica	–	–	84.9	74.0	*O. erraticus*	Relapsing fever
B. latyschewii	–	–	–	–	*O. tartakov* *O. tartakowskyi*	Relapsing fever
"*B. lonestari*"	–	–	–	–	*Amblyomma americanum*	Possible cause of Southern tick-associated rash illness (STARI)
B. mazzottii	–	–	–	–	*O. talaje*	Relapsing fever
B. merionesi	–	–	–	–	*O. erraticus*	Relapsing fever
B. microti	–	–	–	–	*O. erraticus*	Relapsing fever
B. miyamotoi sensu lato	45	13–14	86.7	74.3	*I. persulcatus* *I. scapularis*	Acute febrile illness
B. parkeri	77	–	90.6	74.6	*O. parkeri*	Relapsing fever
B. persica	–	–	84.8	73.9	*O. tholozani*	Relapsing fever
B. recurrentis	–	–	84.7	74.2	*Pediculus humanus*	Relapsing fever
B. theileri	–	–	–	–	*Rhipicephalus decoloratus* *R. evertsi* *Boophilus micropus*	Bovine borreliosis
B. tillae	–	–	–	–	*O. zumpti*	Avian borreliosis
B. turicatae	86	–	90.7	74.6	*O. turicatae*	Relapsing fever
B. venezuelensis	–	–	–	–	*O. rudis*	Relapsing fever
B. turcica	<20	<20	–	–	*Hyalomma aegyptium*	–

– Not determined

[a] Adapted from (Wang and Schwartz 2011)

[b] Adapted from (Margos et al. 2011)

[c] Adapted from (Barbour 2005)

Type Strain: PBi^{T} (=DSM 23469^{T} = BAA-2496^{T})

Description of *Borreliella burgdorferi* comb. nov.

Basonym: *B*. *burgdorferi* (Johnson et al. 1984)

The description of the species is the same as the description given for *B*. *burgdorferi* by Johnson et al. (1984). The species exhibits the genus properties and contains the CSIs and CSPs indicated in the description of *Borreliella*.

Type Strain: B31^{T} (=ATCC 35210^{T} = CIP 102532^{T} = DSM 4680^{T})

Description of *Borreliella carolinensis* comb. nov.

Basonym: *Borrelia carolinensis* (Rudenko et al. 2011)

The description of the species is the same as the description given for *B*. *carolinensis* by Rudenko et al. (2011). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: SCW-22^{T} (=ATCC BAA-1773^{T} = DSM 22119^{T})

Description of *Borreliella garinii* comb. nov.

Basonym: *Borrelia garinii* (Baranton et al. 1992)

The description of the species is the same as the description given for *B*. *garinii* by Baranton et al. (1992). The species exhibits the genus properties and contains the CSIs and CSPs indicated in the description of *Borreliella*.

Type Strain: 20047^{T} (=ATCC 51383^{T} = CIP 103362^{T} = DSM 10534^{T})

Description of *Borreliella japonica* comb. nov.

Basonym: *Borrelia japonica* (Kawabata et al. 1994)

The description of the species is the same as the description given for *B*. *japonica* by Kawabata et al. (1994). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: HO14^{T} (=ATCC 51557^{T} = JCM 8951^{T})

Description of *Borreliella kurtenbachii* comb. nov.

Basonym: *Borrelia kurtenbachii* (Margos et al. 2013a)

The description of the species is the same as the description given for *B*. *kurtenbachii* by Margos et al. (2013a, b). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: 25015^{T} (=ATCC BAA-2495^{T} = DSM 26572^{T})

Description of *Borreliella lusitaniae* comb. nov.

Basonym: *Borrelia lusitaniae* (Le Fleche et al. 1997)

The description of the species is the same as the description given for *B*. *lusitaniae* by Le Fleche et al. (1997). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: PotiB2^{T} (=CIP 105366^{T})

Description of *Borreliella sinica* comb. nov.

Basonym: *Borrelia sinica* (Masuzawa et al. 2001)

The description of the species is the same as the description given for *B*. *sinica* by Masuzawa et al. (2001). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: CMN3^{T} (=DSM 23262^{T} = JCM 10505^{T})

Description of *Borreliella spielmanii* comb. nov.

Basonym: *Borrelia spielmanii* (Richter et al. 2006)

The description of the species is the same as the description given for *B*. *spielmanii* by Richter et al. (2006). The species exhibits the genus properties and contains the CSIs and CSPs indicated in the description of *Borreliella*.

Type Strain: PC-Eq17N5^{T} (=CIP 108855^{T} = DSM 16813^{T})

Description of *Borreliella tanukii* comb. nov.

Basonym: *Borrelia tanukii* (Fukunaga et al. 1997a)

The description of the species is the same as the description given for *B*. *tanukii* by Canica et al. (1994). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: Hk501^{T} (=ATCC BAA-127^{T} = JCM 9662^{T})

Description of *Borreliella turdi* comb. nov.

Basonym: *Borrelia turdi* (Fukunaga et al. 1997b)

The description of the species is the same as the description given for *B*. *turdi* by Fukunaga et al.

(1997). The species exhibits the genus properties indicated in the description of *Borreliella*.

Type Strain: Ya501^{T} (=ATCC BAA-126^{T} = JCM 9661^{T})

Description of *Borreliella valaisiana* comb. nov.

Basonym: *Borrelia valaisiana* (Wang et al. 1997)

The description of the species is the same as the description given for *B. valaisiana* by Wang et al. (1997). The species exhibits the genus properties and contains the CSIs and CSPs indicated in the description of *Borreliella*.

Type Strain: VS116^{T} (=CIP 105367^{T})

References

Adams DA, Gallagher KM, Jajosky RA, Kriseman J, Sharp P, Anderson WJ, Aranas AE, Mayes M, Wodajo MS, Onweh DH et al (2013) Summary of notifiable diseases—United States, 2011. MMWR Morb Mortal Wkly Rep 60(53): 1–117

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

Ahmod NZ, Gupta RS, Shah HN (2011) Identification of a *Bacillus anthracis* specific indel in the *yeaC* gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. J Microbiol Methods 87(3):278–285

Altschul SF, Madden TL, Schäffer 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(17):3389–3402

Baranton G, Postic D, Saint Girons I, Boerlin P, Piffaretti J-C, Assous M, Grimont PAD (1992) Delineation of *Borrelia burgdorferi* Sensu Stricto, *Borrelia garinii* sp. nov., and Group VS461 associated with Lyme borreliosis. Int J Syst Bacteriol 42(3):378–383

Barbour AG (1984) Isolation and cultivation of Lyme disease spirochetes. Yale J Biol Med 57(4):521

Barbour AG (2005) Relapsing fever. In: Goodman JL, Dennis DT, Sonenshine DE (eds) Tick-borne diseases of humans. ASM Press, Washington, pp 268–291

Barbour AG, Miller SC (2014) Genome sequence of *Borrelia parkeri*, an agent of enzootic relapsing fever in Western North America. Genome Announc 2(1):e00018

Bergey DH (1925) Bergey's manual of determinative bacteriology, 2nd edn. The Williams and Wilkins Co, Baltimore

Bhandari V, Gupta RS (2012) Molecular signatures for the phylum Synergistetes and some of its subclades. Antonie Van Leeuwenhoek 102(4):517–540

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

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

Brenner EV, Kurilshikov AM, Stronin OV, Fomenko NV (2012) Whole-genome sequencing of *Borrelia garinii* BgVir, isolated from Taiga ticks (*Ixodes persulcatus*). J Bacteriol 194(20):5713

Canica MM, Du Merle L, Mazie JC, Baranton G, Postic D (1994) *Borrelia afzelii* sp. nov. Validation of the publication of new names and new combinations previously effectively published outside the IJSB, list no 48. Int J Syst Bacteriol 44:182–183

Casjens SR, Fraser-Liggett CM, Mongodin EF, Qiu WG, Dunn JJ, Luft BJ, Schutzer SE (2011a) Whole genome sequence of an unusual *Borrelia burgdorferi* sensu lato isolate. J Bacteriol 193(6):1489–1490

Casjens SR, Mongodin EF, Qiu WG, Dunn JJ, Luft BJ, Fraser-Liggett CM, Schutzer SE (2011b) Whole-genome sequences of two *Borrelia afzelii* and two *Borrelia garinii* Lyme disease agent isolates. J Bacteriol 193(24): 6995–6996

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

Chaconas G (2005) Hairpin telomeres and genome plasticity in *Borrelia*: all mixed up in the end. Mol Microbiol 58(3): 625–635

Chaconas G, Kobryn K (2010) Structure, function, and evolution of linear replicons in *Borrelia*. Annu Rev Microbiol 64:185–202

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

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311(5765):1283–1287

Cutler SJ (2010) Relapsing fever: a forgotten disease revealed. J Appl Microbiol 108(4):1115–1122

Dai Q, Restrepo BI, Porcella SF, Raffel SJ, Schwan TG, Barbour AG (2006) Antigenic variation by *Borrelia hermsii* occurs through recombination between extragenic repetitive elements on linear plasmids. Mol Microbiol 60(6):1329–1343

Dunaj J, Moniuszko A, Zajkowska J, Pancewicz S (2013) The role of PCR in diagnostics of Lyme borreliosis. Przegl Epidemiol 67(1): 35–39, 119–123

Elbir H, Gimenez G, Robert C, Bergström S, Cutler S, Raoult D, Drancourt M (2012) Complete genome sequence of *Borrelia crocidurae*. J Bacteriol 194(14):3723–3724

Elbir H, Larsson P, Normark J, Upreti M, Korenberg E, Larsson C, Bergstrom S (2014a) Genome sequence of the Asiatic species *Borrelia persica*. Genome Announc 2(1):e01127

Elbir H, Larsson P, Upreti M, Normark J, Bergstrom S (2014b) Genome sequence of the relapsing fever borreliosis species *Borrelia hispanica*. Genome Announc 2(1):e01171

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK et al (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390(6660):580–586

Fukunaga M, Okada K, Nakao M, Konishi T, Sato Y (1996) Phylogenetic analysis of *Borrelia* species based on flagellin gene sequences and its application for molecular typing of Lyme disease borreliae. Int J Syst Bacteriol 46(4):898–905

Fukunaga M, Hamase A, Okada K, Nakao M (1997a) *Borrelia tanukii* sp. nov. Validation of the publication of new names and new combinations previously effectively published outside the IJSB, list no 63. Int J Syst Bacteriol 47:1274

Fukunaga M, Hamase A, Okada K, Nakao M (1997b) *Borrelia turdi* sp. nov. Validation of the publication of new names and new combinations previously effectively published outside the IJSB, list no 63. Int J Syst Bacteriol 47:1274

Gao B, Gupta R (2007) Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genom 8(1):86

Gao B, Gupta RS (2012a) Microbial systematics in the post-genomics era. Antonie Van Leeuwenhoek 101(1):45–54

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

Glöckner G, Lehmann R, Romualdi A, Pradella S, Schulte-Spechtel U, Schilhabel M, Wilske B, Sühnel J, Platzer M (2004) Comparative analysis of the *Borrelia garinii* genome. Nucleic Acids Res 32(20):6038–6046

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57(1):81–91

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

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

Gupta RS, Griffiths E (2006) *Chlamydiae*-specific proteins and indels: novel tools for studies. Trends Microbiol 14(12):527–535

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

Gupta RS, Chander P, George S (2013a) Phylogenetic framework and molecular signatures for the class *Chloroflexi* and its different clades; proposal for division of the class *Chloroflexia* class. nov. [corrected] into the suborder *Chloroflexineae* subord. nov., consisting of the emended family *Oscillochloridaceae* and the family *Chloroflexaceae* fam. nov., and the suborder *Roseiflexineae* subord. nov., containing the family *Roseiflexaceae* fam. nov. Antonie Van Leeuwenhoek 103(1):99–119

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

Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome Res 13(3): 407–412

Hue F, Ghalyanchi Langeroudi A, Barbour AG (2013) Chromosome sequence of *Borrelia miyamotoi*, an uncultivable tick-borne agent of human infection. Genome Announc 1(5):e00713

Ibba M, Bono JL, Rosa PA, Soll D (1997) Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. Proc Natl Acad Sci USA 94(26):14383–14388

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

Jiang B, Yao H, Tong Y, Yang X, Huang Y, Jiang J, Cao W (2012a) Genome sequence of *Borrelia garinii* strain NMJW1, isolated from China. J Bacteriol 194(23): 6660–6661

Jiang BG, Zheng YC, Tong YG, Jia N, Huo QB, Fan H, Ni XB, Ma L, Yang XF, Jiang JF et al (2012b) Genome sequence of *Borrelia afzelii* Strain HLJ01, isolated from a patient in China. J Bacteriol 194(24):7014–7015

Johnson RC, Schmid GP, Hyde FW, Steigerwalt AG, Brenner DJ (1984) *Borrelia burgdorferi* sp. nov.: etiologic agent of Lyme disease. Int J Syst Bacteriol 34(4):496–497

Kawabata H, Masuzawa T, Yanagihara Y (1994) *Borrelia japonica* sp. nov. Validation of the publication of new names and new combinations previously effectively published outside the IJSB, list no 50. Int J Syst Bacteriol 44:595

Lapage SP, Sneath PHA, Lessel EF, Skerman VBD, Seeliger HPR, Clark WA (1992) International code of nomenclature of bacteria: bacteriological code, 1990 revision. ASM Press International Union of Microbiological Societies, Washington

Le Fleche A, Postic D, Girardet K, Peter O, Baranton G (1997) Characterization of *Borrelia lusitaniae* sp. nov. by 16S ribosomal DNA sequence analysis. Int J Syst Bacteriol 47(4):921–925

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25(7):1307–1320

Lescot M, Audic S, Robert C, Nguyen TT, Blanc G, Cutler SJ, Wincker P, Couloux A, Claverie JM, Raoult D (2008) The genome of *Borrelia recurrentis,* the agent of deadly louse-borne relapsing fever, is a degraded subset of tick-borne *Borrelia duttonii*. PLoS Genet 4(9):e1000185

Lindgren E, Jaenson TG (2006) Lyme borreliosis in Europe: influences of climate and climate change, epidemiology, ecology and adaptation measures. WHO Regional Office for Europe, Copenhagen

Ljostad U, Mygland A (2013) Chronic Lyme; diagnostic and therapeutic challenges. Acta Neurol Scand 127 Suppl(196): 38–47

Margos G, Vollmer SA, Cornet M, Garnier M, Fingerle V, Wilske B, Bormane A, Vitorino L, Collares-Pereira M, Drancourt M et al (2009) A new *Borrelia* species defined by multilocus sequence analysis of housekeeping genes. Appl Environ Microbiol 75(16):5410–5416

Margos G, Vollmer SA, Ogden NH, Fish D (2011) Population genetics, taxonomy, phylogeny and evolution of *Borrelia burgdorferi* sensu lato. Infect Genet Evol 11(7):1545–1563

Margos G, Piesman J, Lane RS, Ogden NH, Sing A, Straubinger RK, Fingerle V (2013a). *Borrelia kurtenbachii* sp. nov.: a widely distributed member of the *Borrelia burgdorferi* sensu lato species complex in North America. Int J Syst Evol Microbiol 64(Pt 1):128–30. doi:10.1099/ijs.0.054593-0

Margos G, Wilske B, Sing A, Hizo-Teufel C, Cao WC, Chu C, Scholz H, Straubinger RK, Fingerle V (2013b) *Borrelia bavariensis* sp. nov. is widely distributed in Europe and Asia. Int J Syst Evol Microbiol 63(Pt 11):4284–4288

Masuzawa T, Takada N, Kudeken M, Fukui T, Yano Y, Ishiguro F, Kawamura Y, Imai Y, Ezaki T (2001) *Borrelia sinica* sp. nov., a lyme disease-related *Borrelia* species isolated in China. Int J Syst Evol Microbiol 51(Pt 5):1817–1824

Naushad HS, Lee B, Gupta RS (2014) Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*. Int J Syst Evol Microbiol 64(2):366–383

NCBI (2014) NCBI genome database. http://www.ncbi.nlm.nih.gov/genome/

Parte AC (2014) LPSN—list of prokaryotic names with standing in nomenclature. Nucleic Acids Res 42(D1):D613–D616

Paster BJ (2011) Phylum XV. Spirochaetes Garrity and Holt 2001. In Brenner DJ, Krieg NR, Garrity GM, Staley JT (eds) Bergey's Manual of Systematic Bacteriology, 2nd edn, vol 3. Springer, New York, pp 471–471 (reprinted from: not in File)

Postic D, Edlinger C, Richaud C, Grimont F, Dufresne Y, Perolat P, Baranton G, Grimont PAD (1990) Two genomic species in *Borrelia burgdorferi*. Res Microbiol 141(4): 465–475

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(D1):D590–D596

Ras NM, Lascola B, Postic D, Cutler SJ, Rodhain F, Baranton G, Raoult D (1996) Phylogenesis of relapsing fever *Borrelia* spp. Int J Syst Bacteriol 46(4):859–865

Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci 106(45):19126–19131

Richter D, Postic D, Sertour N, Livey I, Matuschka FR, Baranton G (2006) Delineation of *Borrelia burgdorferi* sensu lato species by multilocus sequence analysis and confirmation of the delineation of *Borrelia spielmanii* sp. nov. Int J Syst Evol Microbiol 56(Pt 4):873–881

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

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

Rosselló-Mora R (2006) DNA–DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In: Stackebrandt E (ed) Molecular identification, systematics, and population structure of prokaryotes. Springer, Berlin, pp 23–50

Rudenko N, Golovchenko M, Lin T, Gao L, Grubhoffer L, Oliver JH Jr (2010) *Borrelia americana* sp. nov. List of new names and new combinations previously effectively, but not validly, published, list no 135. Int J Syst Evol Microbiol 60:1985–1986

Rudenko N, Golovchenko M, Grubhoffer L, Oliver JH Jr (2011) *Borrelia carolinensis* sp. nov., a novel species of the *Borrelia burgdorferi* sensu lato complex isolated from rodents and a tick from the south-eastern USA. Int J Syst Evol Microbiol 61(Pt 2):381–383

Schutzer SE, Fraser-Liggett CM, Casjens SR, Qiu WG, Dunn JJ, Mongodin EF, Luft BJ (2011) Whole-genome sequences of thirteen isolates of *Borrelia burgdorferi*. J Bacteriol 193(4):1018–1020

Schutzer SE, Fraser-Liggett CM, Qiu WG, Kraiczy P, Mongodin EF, Dunn JJ, Luft BJ, Casjens SR (2012) Whole-genome sequences of *Borrelia bissettii, Borrelia valaisiana,* and *Borrelia spielmanii*. J Bacteriol 194(2):545–546

Segata N, Bornigen D, Morgan XC, Huttenhower C (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun 4:2304

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

Swellengrebel NH (1907) Sur la cytologie comparée des spirochètes et des spirilles. Ann Inst Pasteur (Paris) 21:562–586

Takano A, Goka K, Une Y, Shimada Y, Fujita H, Shiino T, Watabane H, Kawabata H (2010) Isolation and characterization of a novel *Borrelia* group of tick-borne borreliae from imported reptiles and their associated ticks. Environ Microbiol 12(1):134–146

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30(12):2725–2729

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In Miura RM (ed) Lectures on mathematics in the life sciences, 17th edn. American Mathematical Society, Providence, pp 57–86 (reprinted from: not in file)

Thompson CC, Chimetto L, Edwards RA, Swings J, Stackebrandt E, Thompson FL (2013) Microbial genomic taxonomy. BMC Genom 14(1):913

Valsangiacomo C, Balmelli T, Piffaretti JC (1997) A phylogenetic analysis of *Borrelia burgdorferi* sensu lato based on sequence information from the hbb gene, coding for a histone-like protein. Int J Syst Bacteriol 47(1):1–10

Vinuesa P (2010) Multilocus sequence analysis and bacterial species phylogeny estimation. In: Oren A, Papke RT (eds) Molecular phylogeny of microorganisms. Caister Academic Press, Norfolk, pp 41–64

Wang G, Schwartz I (2011) Genus II. *Borrelia Swellengrebel* 1907, 582^{AL}. In: Brenner DJ, Krieg NR, Garrity GM, Staley JT (eds) Bergey's manual of systematic bacteriology, vol 3, 2nd edn. Springer, New York, pp 484–498

Wang G, van Dam AP, Le Fleche A, Postic D, Peter O, Baranton G, de Boer R, Spanjaard L, Dankert J (1997) Genetic and phenotypic analysis of *Borrelia valaisiana* sp. nov. (*Borrelia* genomic groups VS116 and M19). Int J Syst Bacteriol 47(4):926–932

Wang G, van Dam AP, Schwartz I, Dankert J (1999) Molecular typing of *Borrelia burgdorferi* sensu lato: taxonomic, epidemiological, and clinical implications. Clin Microbiol Rev 12(4):633–653

Wright WF, Riedel DJ, Talwani R, Gilliam BL (2012) Diagnosis and management of Lyme disease. Am Fam Physician 85(11):1086–1093

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

CHAPTER 4

Phylogenomics and molecular signatures for the order *Neisseriales*: proposal for division of the order *Neisseriales* into the emended family *Neisseriaceae* and *Chromobacteriaceae* fam. nov.

This chapter describes the use of molecular signatures (CSIs) and phylogenetic trees to differentiate the obligate host-associated members of the order *Neisseriales* from the other genera within the order. The chapter also includes a brief discussion of the evolutionary history of the host-associated members of the order *Neisseriales* based on the phylogenetic trees and identified CSIs. The chapter concludes with a proposal to limit the family *Neisseriaceae* to the obligate host-associated members of the order *Neisseriales*, and to transfer the other genera within the order *Neisseriales* to the novel family *Chromobacteriaceae*. My contributions towards the completion of this chapter include the construction of all phylogenetic trees, identification of all CSIs, the creation of the taxonomic proposals, the writing of all drafts and revisions of the manuscript, and the production of all main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials for this work are not included in the chapter but can be accessed along with the rest of the manuscript at:

Adeolu, M., & Gupta, R. S. (2013). *Anton Leeuw Int J G, 104*(1), 1-24.

Antonie van Leeuwenhoek (2013) 104:1–24
DOI 10.1007/s10482-013-9920-6

ORIGINAL PAPER

Phylogenomics and molecular signatures for the order *Neisseriales*: proposal for division of the order *Neisseriales* into the emended family *Neisseriaceae* and *Chromobacteriaceae* fam. nov.

Mobolaji Adeolu · Radhey S. Gupta

Received: 1 February 2013 / Accepted: 1 April 2013 / Published online: 11 April 2013

Abstract The species from the order *Neisseriales* are currently distinguished from other bacteria on the basis of branching in 16S rRNA gene trees. For this order containing a single family, *Neisseriaceae*, no distinctive molecular, biochemical, or phenotypic characters are presently known. We report here detailed phylogenetic and comparative analyses on the 27 genome sequenced species of the order *Neisseriales*. Our comparative genomic analyses have identified 54 conserved signature indels (CSIs) in widely distributed proteins that are specific for either all of the sequenced *Neisseriales* species or a number of clades within this order that are also supported by phylogenetic analyses. Of these CSIs, 11 are specifically present in all of the sequenced species from this order, but are not found in homologous proteins from any other bacteria. These CSIs provide novel molecular markers specific for, and delimiting, this order. Twenty-one CSIs in diverse proteins are specific for a group comprised of the genera *Neisseria*, *Eikenella*, *Kingella*, and *Simonsiella* (Clade I), which are obligate host-associated organisms, lacking flagella and exhibiting varied morphology. The species from these genera also formed a strongly supported clade in phylogenetic trees based upon concatenated protein sequences; a monophyletic grouping of these genera and other genera displaying similar morphological characteristics was also observed in the 16S rRNA gene tree. A second clade (Clade II), supported by seven of the identified CSIs and phylogenetic trees based upon concatenated protein sequences, grouped together species from the genera *Chromobacterium*, *Laribacter*, and *Pseudogulbenkiania* that are rod-shaped bacteria, which display flagella-based motility and are capable of free living. The remainder of the CSIs were uniquely shared by smaller groups within these two main clades. Our analyses also provide novel insights into the evolutionary history of the *Neisseriales* and suggest that the CSIs that are specific for the Clade I species may play an important role in the evolution of obligate host-association within this order. On the basis of phylogenetic analysis, the identified CSIs, and conserved phenotypic characteristics of different *Neisseriales* genera, we propose a division of this order into two families: an emended family *Neisseriaceae* (corresponding to Clade I) containing the genera *Alysiella*, *Bergeriella*, *Conchiformibius*, *Eikenella*, *Kingella*, *Neisseria*, *Simonsiella*, *Stenoxybacter*, *Uruburuella* and *Vitreoscilla* and a new family, *Chromobacteriaceae* fam. nov., harboring the remainder of the genera from this order (viz. *Andreprevotia*, *Aquaspirillum*, *Aquitalea*, *Chitinibacter*, *Chitinilyticum*, *Chitiniphilus*, *Chromobacterium*, *Deefgea*, *Formivibrio*, *Gulbenkiania*, *Iodobacter*, *Jeongeupia*, *Laribacter*, *Leeia*,

Electronic supplementary material The online version of this article (doi:10.1007/s10482-013-9920-6) contains supplementary material, which is available to authorized users.

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

Microvirgula, *Paludibacterium*, *Pseudogulbenkiania*, *Silvimonas*, and *Vogesella*).

Keywords *Neisseriales* · *Neisseriales* taxonomy · Betaproteobacteria · Phylogenetic trees · *Neisseriaceae* · *Chromobacteriaceae* · Conserved signature indels · Molecular signatures

Introduction

The order *Neisseriales* is described as a group of gram-negative, non-spore forming, aerobic, and mesophilic bacteria (Tønjum 2005b). However, none of these traits are unique characteristics of the order. Currently, 32 genera of *Neisseriales* have been described, spanning a wide range of morphologies, habitats, and growth requirements, including important pathogens such as *Neisseria gonorrhoeae* and *N. meningitidis* (Tønjum 2005b; Euzeby 2012). *Neisseria gonorrhoeae*, the causative agent of the sexually transmitted infection gonorrhea, is an extremely prevalent pathogen that infects approximately 88 million individuals a year worldwide (World Health Organization 2011). *Neisseria meningitidis* is the primary causative agent of infectious meningococcal meningitis which has a mortality rate of over 70 % without treatment and has prevalence that ranges from less than 1 to over 1,000 cases per 100,000 individuals worldwide (Stephens et al. 2007; Cohn et al. 2010). However, despite the diversity within the order and the presence of important pathogens, all 32 genera within the order *Neisseriales* are currently placed within a single family, *Neisseriaceae*.

The current taxonomy of the order *Neisseriales* is primarily based on 16S rRNA sequence identity studies and phylogenetic trees (Bøvre 1984; Harmsen et al. 2001; Hedlund and Staley 2002; Tønjum 2005b; Yarza et al. 2008). In these trees, species from the order *Neisseriales* form a distinct clade, which provides the primary means for distinguishing them from all other bacteria. However, in these trees, the interrelationships among different genera that are part of this order are not resolved, leading to placement of all of them into a single family. Except for their branching in the 16S rRNA trees, currently, no other combination of morphological or biochemical properties are known that can reliably identify, or delimit, the species from this order, or can form the basis for its division into distinct subgroups (Bøvre 1984; Harmsen et al. 2001; Hedlund and Staley 2002; Tønjum 2005b; Hedlund and Kuhn 2006).

In recent years genome sequences from a large numbers of species from the order *Neisseriales* have become available in the public domain. These genome sequences should enable determination of the phylogeny of these bacteria based upon larger data sets of sequences, which provide a more reliable indication of their true phylogenetic affinities, than a single gene or protein (Rokas et al. 2003; Ciccarelli et al. 2006; Gupta and Mok 2007; Wu et al. 2009; Gao and Gupta 2012a). Genomic sequence data has already been used to more reliably elucidate the interrelationships of species within the genus *Neisseria* (Bennett et al. 2012). Additionally, comparative analyses of these genome sequences allow for the discovery of novel molecular markers (or signatures) that are capable of more reliably distinguishing these bacteria from all others. These comparative genomic studies should also provide important insights into the evolutionary relationships amongst the different taxa that are part of this order, independently of phylogenetic trees (Gupta 1998; Gupta et al. 2012; Gao and Gupta 2012a).

Currently, genome sequences for over 140 strains representing 27 species from the order *Neisseriales* are publicly available (Table 1) (NCBI 2012). We have used genomic information to construct a robust phylogenetic tree for the sequenced species based upon the concatenated sequences of 20 conserved proteins (Rokas et al. 2003; Ciccarelli et al. 2006; Gao et al. 2009; Gao and Gupta 2012a). Additionally, we have performed comprehensive comparative analyses on protein sequences from these genomes to identify molecular signatures comprising of conserved signature inserts or deletions (i.e. indels) (CSIs) in protein sequences that are uniquely shared by different species within this order. These studies have led to identification of >50 CSIs in different proteins involved in a broad range of functions that are specific for either all sequenced *Neisseriales* or a number of well-supported clades within this order at multiple phylogenetic levels. In particular, large numbers of the identified signatures are specific for a subclade of the *Neisseriales*, which is strongly supported in both the concatenated protein tree and the 16S rRNA tree. This clade is mainly comprised of species that are obligatory host-associated organisms and which lack flagella. To recognize the distinctness of this clade from all other *Neisseriales*, it is proposed that the order *Neisseriales*

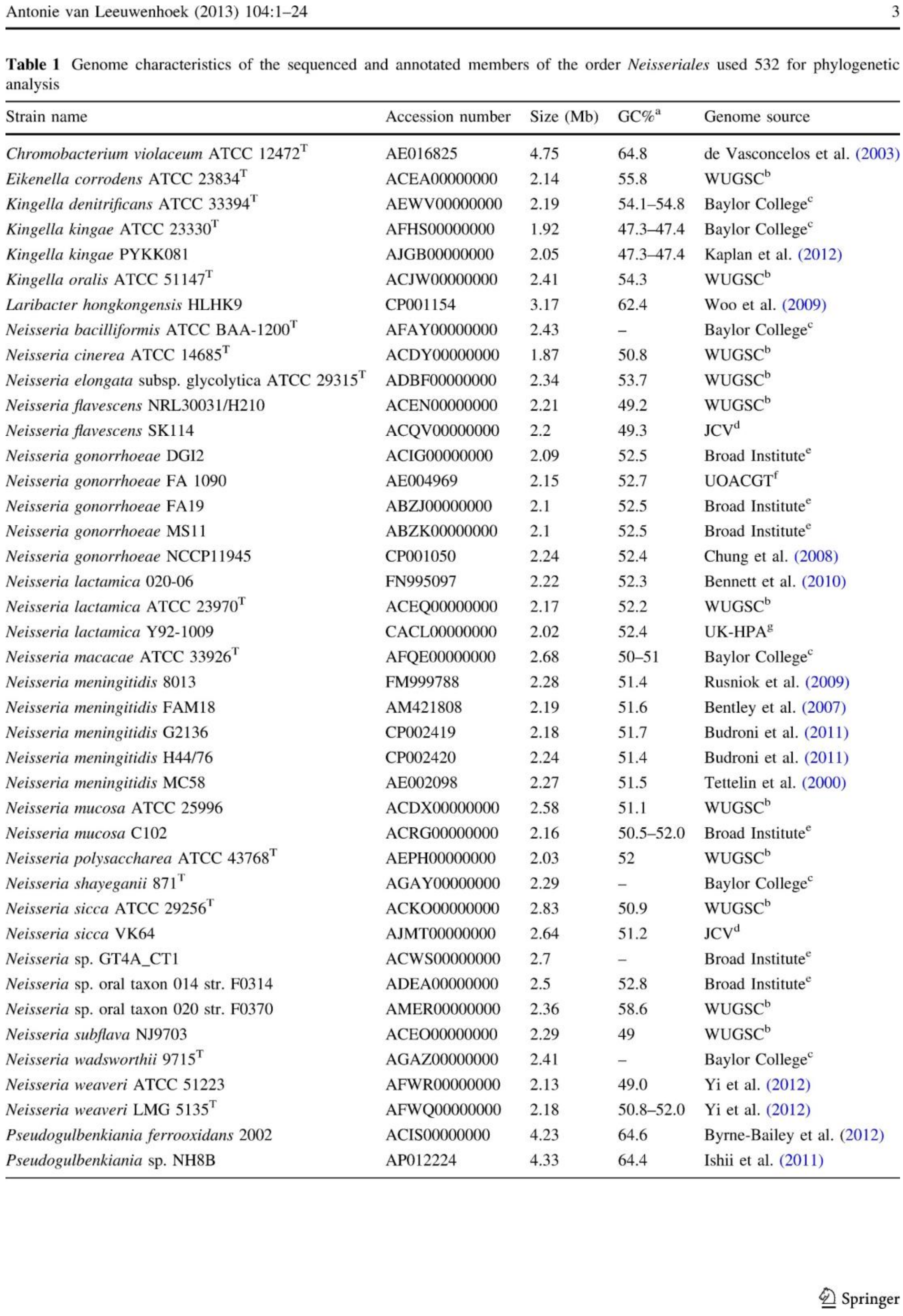


Table 1 Genome characteristics of the sequenced and annotated members of the order *Neisseriales* used 532 for phylogenetic analysis

Strain name	Accession number	Size (Mb)	GC%[a]	Genome source
Chromobacterium violaceum ATCC 12472[T]	AE016825	4.75	64.8	de Vasconcelos et al. (2003)
Eikenella corrodens ATCC 23834[T]	ACEA00000000	2.14	55.8	WUGSC[b]
Kingella denitrificans ATCC 33394[T]	AEWV00000000	2.19	54.1–54.8	Baylor College[c]
Kingella kingae ATCC 23330[T]	AFHS00000000	1.92	47.3–47.4	Baylor College[c]
Kingella kingae PYKK081	AJGB00000000	2.05	47.3–47.4	Kaplan et al. (2012)
Kingella oralis ATCC 51147[T]	ACJW00000000	2.41	54.3	WUGSC[b]
Laribacter hongkongensis HLHK9	CP001154	3.17	62.4	Woo et al. (2009)
Neisseria bacilliformis ATCC BAA-1200[T]	AFAY00000000	2.43	–	Baylor College[c]
Neisseria cinerea ATCC 14685[T]	ACDY00000000	1.87	50.8	WUGSC[b]
Neisseria elongata subsp. glycolytica ATCC 29315[T]	ADBF00000000	2.34	53.7	WUGSC[b]
Neisseria flavescens NRL30031/H210	ACEN00000000	2.21	49.2	WUGSC[b]
Neisseria flavescens SK114	ACQV00000000	2.2	49.3	JCV[d]
Neisseria gonorrhoeae DGI2	ACIG00000000	2.09	52.5	Broad Institute[e]
Neisseria gonorrhoeae FA 1090	AE004969	2.15	52.7	UOACGT[f]
Neisseria gonorrhoeae FA19	ABZJ00000000	2.1	52.5	Broad Institute[e]
Neisseria gonorrhoeae MS11	ABZK00000000	2.1	52.5	Broad Institute[e]
Neisseria gonorrhoeae NCCP11945	CP001050	2.24	52.4	Chung et al. (2008)
Neisseria lactamica 020-06	FN995097	2.22	52.3	Bennett et al. (2010)
Neisseria lactamica ATCC 23970[T]	ACEQ00000000	2.17	52.2	WUGSC[b]
Neisseria lactamica Y92-1009	CACL00000000	2.02	52.4	UK-HPA[g]
Neisseria macacae ATCC 33926[T]	AFQE00000000	2.68	50–51	Baylor College[c]
Neisseria meningitidis 8013	FM999788	2.28	51.4	Rusniok et al. (2009)
Neisseria meningitidis FAM18	AM421808	2.19	51.6	Bentley et al. (2007)
Neisseria meningitidis G2136	CP002419	2.18	51.7	Budroni et al. (2011)
Neisseria meningitidis H44/76	CP002420	2.24	51.4	Budroni et al. (2011)
Neisseria meningitidis MC58	AE002098	2.27	51.5	Tettelin et al. (2000)
Neisseria mucosa ATCC 25996	ACDX00000000	2.58	51.1	WUGSC[b]
Neisseria mucosa C102	ACRG00000000	2.16	50.5–52.0	Broad Institute[e]
Neisseria polysaccharea ATCC 43768[T]	AEPH00000000	2.03	52	WUGSC[b]
Neisseria shayeganii 871[T]	AGAY00000000	2.29	–	Baylor College[c]
Neisseria sicca ATCC 29256[T]	ACKO00000000	2.83	50.9	WUGSC[b]
Neisseria sicca VK64	AJMT00000000	2.64	51.2	JCV[d]
Neisseria sp. GT4A_CT1	ACWS00000000	2.7	–	Broad Institute[e]
Neisseria sp. oral taxon 014 str. F0314	ADEA00000000	2.5	52.8	Broad Institute[e]
Neisseria sp. oral taxon 020 str. F0370	AMER00000000	2.36	58.6	WUGSC[b]
Neisseria subflava NJ9703	ACEO00000000	2.29	49	WUGSC[b]
Neisseria wadsworthii 9715[T]	AGAZ00000000	2.41	–	Baylor College[c]
Neisseria weaveri ATCC 51223	AFWR00000000	2.13	49.0	Yi et al. (2012)
Neisseria weaveri LMG 5135[T]	AFWQ00000000	2.18	50.8–52.0	Yi et al. (2012)
Pseudogulbenkiania ferrooxidans 2002	ACIS00000000	4.23	64.6	Byrne-Bailey et al. (2012)
Pseudogulbenkiania sp. NH8B	AP012224	4.33	64.4	Ishii et al. (2011)

Table 1 continued

Strain name	Accession number	Size (Mb)	GC%[a]	Genome source
Simonsiella muelleri ATCC 29453[T]	ADCY00000000	2.39	41.3	Broad Institute[e]

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

T Type strain

[a] Genomic GC% of some species obtained from (Brenner et al. 2005)

[b] WUGSC: genome sequenced by Washington University Genome Sequencing Center

[c] Baylor College: genome sequenced by Baylor College of Medicine

[d] JCV: genome sequenced by J. Craig Venter Institute

[e] Broad Institute: genome sequenced by The Broad Institute Genome Sequencing Platform

[f] UOACGT: University of Oklahoma Advanced Center for Genome Technology

[g] UK-HPA: genome sequenced by UK Health Protection Agency

be divided into two families, an emended family *Neisseriaceae* comprising of this well-supported clade and a new family, *Chromobacteriaceae* fam. nov., harboring the other genera from this order.

Methodology

Phylogenetic sequence analysis

Phylogenetic analysis was performed on a concatenated sequence alignment of 20 highly conserved proteins (viz. UvrD, GyrA, GyrB, RpoB, RpoC, EF-G, EF-Tu, RecA, ArgRS, IleRS, ThrRS, TrpRS, SecY, DnaK and ribosomal proteins L2, L5, S2, S3, and S9) which are present in most bacteria and have been extensively used in phylogenetic studies (Harris et al. 2003; Gao and Gupta 2012a). The trees were constructed for 44 strains from the order *Neisseriales* that are listed in Table 1. Except for *N. gonorrhoeae* and *N. meningitides*, for which only a number of representative strains were included, this includes all of the species/strains whose genomes are now available. The amino acid sequences for the above mentioned 20 proteins were obtained from NCBI for all of the species/strains listed in Table 1 as well as from *Bordetella pertussis* and *Burkholderia ambifaria*, which served as outgroups. Multiple sequence alignments for these proteins were created using Clustal_X 1.83 (Jeanmougin et al. 1998) and concatenated into a single alignment file. Poorly aligned regions from this alignment file were removed using Gblocks 0.91b (Castresana 2000). The resulting alignment, which contained 11,954 aligned positions, was used for phylogenetic analysis. The maximum likelihood (ML) and neighbour joining (NJ) trees based on 100 bootstrap replicates of this alignment were constructed using MEGA 5.05 (Tamura et al. 2011) employing the Whelan and Goldman (2001) and Jones et al. (1992) substitution models, respectively.

A 16S rRNA gene sequence tree was also created for 94 sequences that included representative species for 32 genera that are part the order *Neisseriales* (Supplemental Table 1). 16S rRNA gene sequences larger than 1,300 bp were obtained for all type species classified under the order *Neisseriales* in release 114 of the SILVA database (Quast et al. 2013). 16S rRNA genes were also obtained for *Amantichitinum ursilacus*, which has yet to be added to the SILVA database, and every genome sequenced strain included in the concatenated protein tree (excluding *Neisseria lactamica* 020-06, *Neisseria sicca* VK64, and *Neisseria* sp. oral taxon 020 str. F0370 which have annotated 16S rRNA genes smaller than 1,300 bp). The accession numbers of different 16S rRNA gene sequences used in this work are provided in Supplemental Table 1. A maximum likelihood tree based on these sequences was created using 100 bootstrap replicates of the 16S rRNA sequence alignments in MEGA 5.05 (Tamura et al. 2011) employing the general time-reversible (Tavaré 1986) substitution model.

Identification of molecular markers (CSIs)

To identify CSIs that are commonly shared by different *Neisseriales* species, Blastp searches were performed on each protein in the genomes of *N. meningitides* FAM18 and *Pseudogulbenkiania* sp. NH8B. For those proteins for whom high scoring homologs (E values $<1e^{-20}$) were present in other species from the order *Neisseriales* and some other bacterial groups multiple sequence alignments were created using the Clustal_X 1.83 program (Jeanmougin et al. 1998). These alignments were visually inspected for the presence of insertions or deletions that were flanked on both sides by at least 5–6 identical/conserved amino acid residues in the neighbouring 30–40 amino acids. Indels that were not flanked by conserved regions were not further considered, as they do not provide useful molecular markers (Gupta 1998, 2001, 2009). The species specificity of each indel thus identified was then further evaluated by conducting Blastp searches on short sequence segments containing the indels and their flanking conserved regions (60–100 amino acids long). The searches were carried out against the NCBI non-redundant (nr) database and in all cases, a minimum of 250 blast hits were examined for the presence or absence of these CSIs to ascertain their specificities. All CSIs that were primarily restricted to members of the order *Neisseriales* were independently evaluated to determine their species specificity and the relationships among the members of this order that they support. In this work, we report the results of only those CSIs that are specific for the species from the order *Neisseriales* and where similar CSIs were not observed in any other bacteria in the top 250 blast hits. The CSIs present in only a single species, or those that were specific for the larger clades of *Betaproteobacteria*, and those which were shared with some other bacteria are not shown here, as they are of limited utility for the present work. The sequence alignment files presented here contain information for all detected *Neisseriales* homologs, but only a limited number of species from other bacterial groups. Sequence information for different strains of various species is also not shown as they all exhibited similar pattern. Additionally, unless otherwise indicated, all of these CSIs are specific for the indicated group of species and similar CSIs were not detected in any other bacteria at least in the top 250 blast hits.

Results

Phylogenetic analysis of the order *Neisseriales*

The genome sequences for >140 strains from the order *Neisseriales* are now available in the NCBI database. Although large numbers of these genomes are for two important pathogenic species *N. gonorrhoeae* and *N. meningitides*, the sequenced genomes include information for 27 different species from the order *Neisseriales*. Some characteristics of the genome sequences for different *Neisseriales* species/strains is shown in Table 1. For *N. gonorrhoeae* and *N. meningitides*, for whom large numbers of genomes have been sequenced, information for only a limited number of strains is presented in this table. The genome sizes of the species from the order *Neisseriales* vary from 1.87 to 4.75 Mb and their G+C contents are in the range of 42–65 %. Generally, the genomes of free-living species from this order (viz. *Chromobacterium*, *Laribacter*, and *Pseudogulbenkiania*) were much larger than those of the host-associated organisms (viz. *Neisseria*, *Eikenella*, *Kingella*, and *Simonsiella*).

The current taxonomy of the *Neisseriales* is primarily based on 16S rRNA sequence based studies (Bøvre 1984; Tønjum 2005b; Yarza et al. 2008). In the present work, we have constructed phylogenetic trees of all genome sequenced species from the order *Neisseriales* as well as a number of their strains based on concatenated sequences of 20 conserved housekeeping and ribosomal proteins (Table 1). The trees were constructed using both the maximum likelihood method, which is shown in Fig. 1, and the neighbor joining algorithm, which is shown in Supplemental Fig. 1. The branching patterns of the trees created by both these algorithms were very similar. These trees provide a phylogenetic framework for interpreting the significance of various molecular signatures (i.e. CSIs) that are identified by our comparative genomic analyses. In these trees, which were rooted using sequences from *Burkholderia* spp. (not shown), the genera from the order *Neisseriales* formed two distinct, well-supported subclades, which were separated by a long branch. One of these clades, marked Clade I, consisted of the genera *Neisseria*, *Eikenella*, *Kingella*, and *Simonsiella*, whereas a second deeper branching clade (Clade II) grouped together species from the genera *Chromobacterium*, *Laribacter*, and *Pseudogulbenkiania* (Fig. 1).

Within Clade I, a number of distinct subclades were also observed. One of these subclades consisted of all of the species from the genus *Neisseria*, except *N. shayeganii*. This latter *Neisseria* species instead grouped with the species *Eikenella corrodens*.

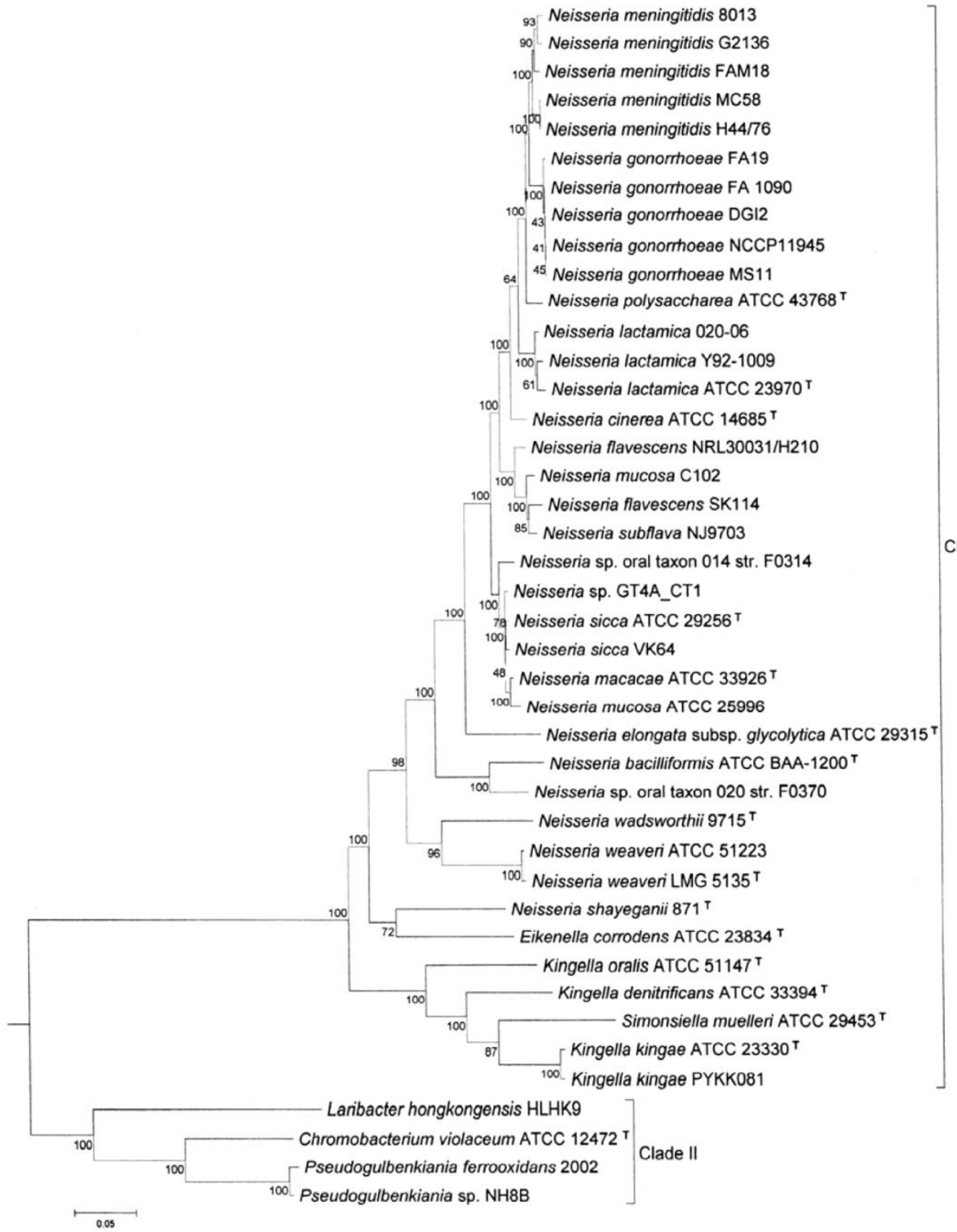

93
Neisseria meningitidis 8013
90
Neisseria meningitidis G2136
100
Neisseria meningitidis FAM18
Neisseria meningitidis MC58
100
100
Neisseria meningitidis H44/76
Neisseria gonorrhoeae FA19
100
Neisseria gonorrhoeae FA 1090
100
43
Neisseria gonorrhoeae DGI2
41
Neisseria gonorrhoeae NCCP11945
45
Neisseria gonorrhoeae MS11
64
Neisseria polysaccharea ATCC 43768 T
Neisseria lactamica 020-06
100
100
Neisseria lactamica Y92-1009
61
Neisseria lactamica ATCC 23970 T
100
Neisseria cinerea ATCC 14685 T
Neisseria flavescens NRL30031/H210
100
Neisseria mucosa C102
100
100
Neisseria flavescens SK114
85
Neisseria subflava NJ9703
Neisseria sp. oral taxon 014 str. F0314
Neisseria sp. GT4A_CT1
100
100
78
Neisseria sicca ATCC 29256 T
100
Neisseria sicca VK64
48
Neisseria macacae ATCC 33926 T
100
100
Neisseria mucosa ATCC 25996
Neisseria elongata subsp. glycolytica ATCC 29315 T
98
Neisseria bacilliformis ATCC BAA-1200 T
100
Neisseria sp. oral taxon 020 str. F0370
Neisseria wadsworthii 9715 T
100
96
Neisseria weaveri ATCC 51223
100
Neisseria weaveri LMG 5135 T
100
Neisseria shayeganii 871 T
72
Eikenella corrodens ATCC 23834 T
Kingella oralis ATCC 51147 T
Kingella denitrificans ATCC 33394 T
100
Simonsiella muelleri ATCC 29453 T
100
87
Kingella kingae ATCC 23330 T
100
Kingella kingae PYKK081
Clade I
Laribacter hongkongensis HLHK9
100
Chromobacterium violaceum ATCC 12472 T
100
Pseudogulbenkiania ferrooxidans 2002
100
Pseudogulbenkiania sp. NH8B
Clade II
0.05

Fig. 1 A Phylogenetic tree of genome sequenced members of the order *Neisseriales* based on the concatenated amino acid sequences of 20 conserved proteins. The tree shown is a maximum-likelihood (ML) distance tree. *Bootstrap values* are shown at *branch nodes*. The two main clades of *Neisseriales* supported by the tree are marked. The *letter* ^{T} refers to the type strain of the species

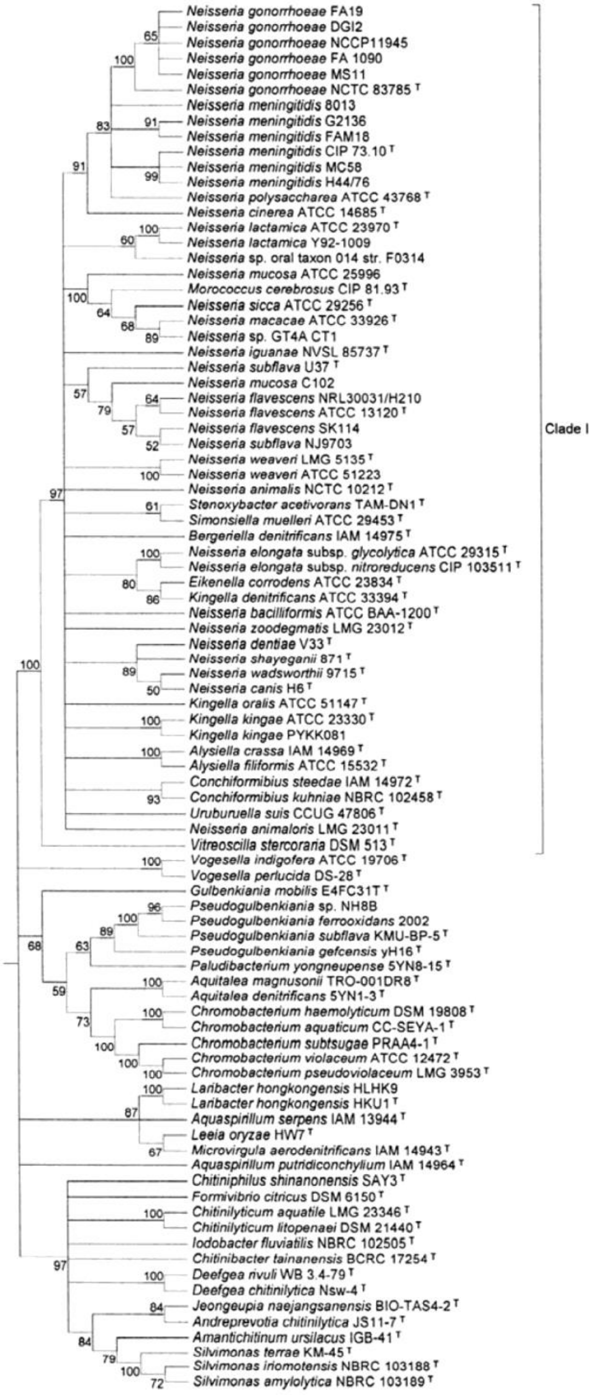

Neisseria gonorrhoeae FA19
Neisseria gonorrhoeae DGI2
65
Neisseria gonorrhoeae NCCP11945
100
Neisseria gonorrhoeae FA 1090
Neisseria gonorrhoeae MS11
Neisseria gonorrhoeae NCTC 83785 ^{T}
Neisseria meningitidis 8013
91
Neisseria meningitidis G2136
83
Neisseria meningitidis FAM18
Neisseria meningitidis CIP 73.10 ^{T}
91
Neisseria meningitidis MC58
99
Neisseria meningitidis H44/76
Neisseria polysaccharea ATCC 43768 ^{T}
Neisseria cinerea ATCC 14685 ^{T}
100
Neisseria lactamica ATCC 23970 ^{T}
60
Neisseria lactamica Y92-1009
Neisseria sp. oral taxon 014 str. F0314
Neisseria mucosa ATCC 25996
Morococcus cerebrosus CIP 81.93 ^{T}
100
Neisseria sicca ATCC 29256 ^{T}
64
Neisseria macacae ATCC 33926 ^{T}
68
89
Neisseria sp. GT4A CT1
Neisseria iguanae NVSL 85737 ^{T}
Neisseria subflava U37 ^{T}
Neisseria mucosa C102
57
64
Neisseria flavescens NRL30031/H210
79
Neisseria flavescens ATCC 13120 ^{T}
57
Neisseria flavescens SK114
Clade I
52
Neisseria subflava NJ9703
Neisseria weaveri LMG 5135 ^{T}
100
Neisseria weaveri ATCC 51223
97
Neisseria animalis NCTC 10212 ^{T}
61
Stenoxybacter acetivorans TAM-DN1 ^{T}
Simonsiella muelleri ATCC 29453 ^{T}
Bergeriella denitrificans IAM 14975 ^{T}
100
Neisseria elongata subsp. glycolytica ATCC 29315 ^{T}
Neisseria elongata subsp. nitroreducens CIP 103511 ^{T}
80
Eikenella corrodens ATCC 23834 ^{T}
86
Kingella denitrificans ATCC 33394 ^{T}
Neisseria bacilliformis ATCC BAA-1200 ^{T}
Neisseria zoodegmatis LMG 23012 ^{T}
Neisseria dentiae V33 ^{T}
100
Neisseria shayeganii 871 ^{T}
89
Neisseria wadsworthii 9715 ^{T}
50
Neisseria canis H6 ^{T}
Kingella oralis ATCC 51147 ^{T}
100
Kingella kingae ATCC 23330 ^{T}
Kingella kingae PYKK081
100
Alysiella crassa IAM 14969 ^{T}
Alysiella filiformis ATCC 15532 ^{T}
Conchiformibius steedae IAM 14972 ^{T}
93
Conchiformibius kuhniae NBRC 102458 ^{T}
Uruburuella suis CCUG 47806 ^{T}
Neisseria animaloris LMG 23011 ^{T}
Vitreoscilla stercoraria DSM 513 ^{T}
100
Vogesella indigofera ATCC 19706 ^{T}
Vogesella perlucida DS-28 ^{T}
Gulbenkiania mobilis E4FC31T ^{T}
96
Pseudogulbenkiania sp. NH8B
100
Pseudogulbenkiania ferrooxidans 2002
89
Pseudogulbenkiania subflava KMU-BP-5 ^{T}
68
63
Pseudogulbenkiania gefcensis yH16 ^{T}
Paludibacterium yongneupense 5YN8-15 ^{T}
100
Aquitalea magnusonii TRO-001DR8 ^{T}
59
Aquitalea denitrificans 5YN1-3 ^{T}
100
Chromobacterium haemolyticum DSM 19808 ^{T}
73
Chromobacterium aquaticum CC-SEYA-1 ^{T}
100
Chromobacterium subtsugae PRAA4-1 ^{T}
100
Chromobacterium violaceum ATCC 12472 ^{T}
100
Chromobacterium pseudoviolaceum LMG 3953 ^{T}
100
Laribacter hongkongensis HLHK9
Laribacter hongkongensis HKU1 ^{T}
87
Aquaspirillum serpens IAM 13944 ^{T}
Leeia oryzae HW7 ^{T}
67
Microvirgula aerodenitrificans IAM 14943 ^{T}
Aquaspirillum putridiconchylium IAM 14964 ^{T}
Chitiniphilus shinanonensis SAY3 ^{T}
Formivibrio citricus DSM 6150 ^{T}
100
Chitinilyticum aquatile LMG 23346 ^{T}
Chitinilyticum litopenaei DSM 21440 ^{T}
Iodobacter fluviatilis NBRC 102505 ^{T}
Chitinibacter tainanensis BCRC 17254 ^{T}
97
100
Deefgea rivuli WB 3.4-79 ^{T}
Deefgea chitinilytica Nsw-4 ^{T}
84
Jeongeupia naejangsanensis BIO-TAS4-2 ^{T}
Andreprevotia chitinilytica JS11-7 ^{T}
84
Amantichitinum ursilacus IGB-41 ^{T}
Silvimonas terrae KM-45 ^{T}
79
100
Silvimonas iriomotensis NBRC 103188 ^{T}
72
Silvimonas amylolytica NBRC 103189 ^{T}

Fig. 2 A ML tree based on the 16S rRNA gene sequences of representative species from different genera within the order *Neisseriales*. *Bootstrap values* based upon 100 replicates are shown at the *nodes* and *nodes* with *bootstrap values* below 50 % are condensed. The *letter* ^{T} refers to the type strain of the species. The accession numbers of the 16S rRNA gene sequences used in this analysis are provided in the Supplementary Table 1

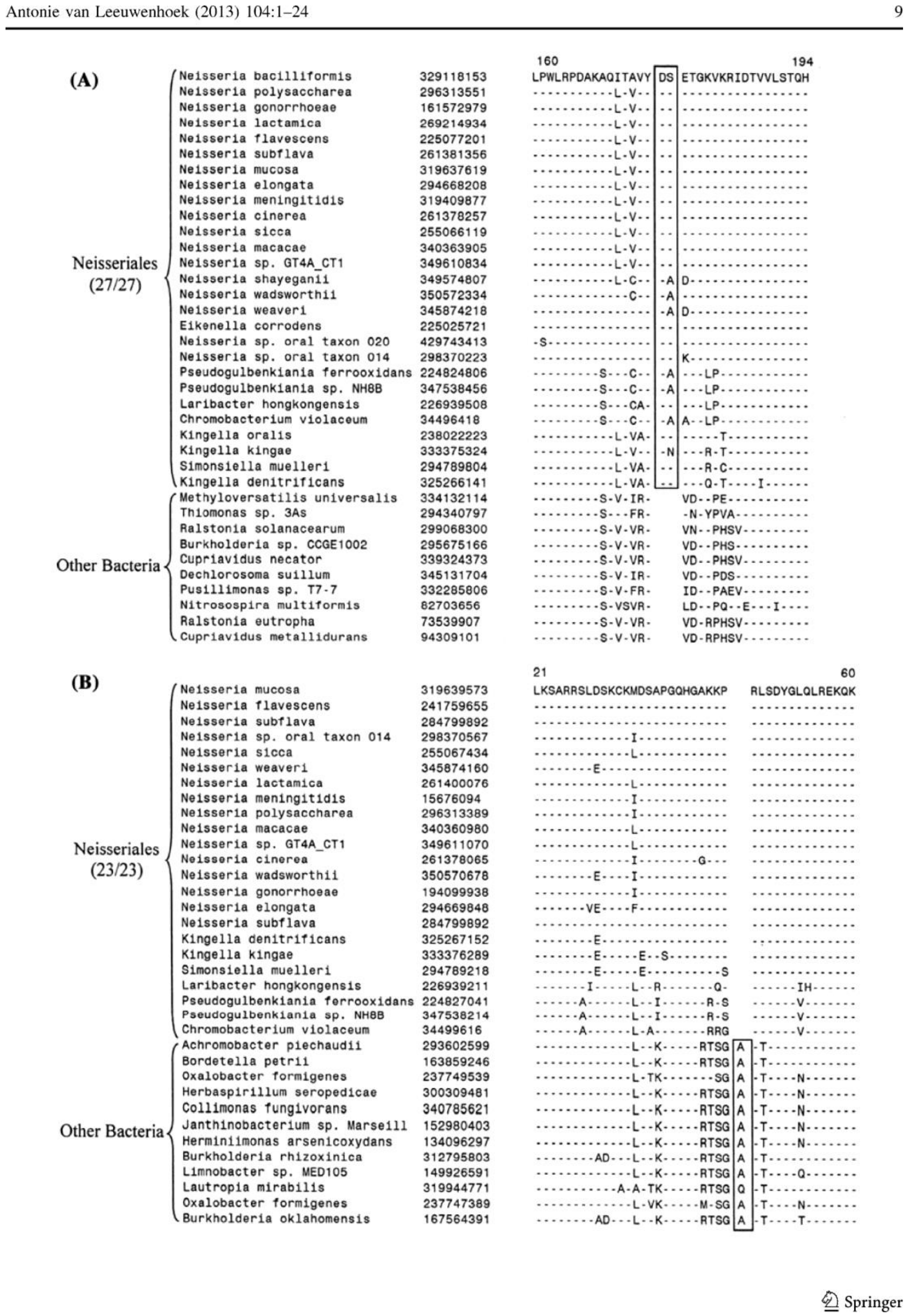
Another distinct subclade within the Clade I was comprised of the genera *Kingella* and *Simonsiella*. However, within this subclade, *S. muelleri* was found to branch between two *Kingella* species, thereby making the genus *Kingella* polyphyletic. It should also be noted that within the genus *Neisseria*, the two strains of *Neisseria mucosa* did not branch together and were part of different clusters. The strain *N. mucosa* ATCC 25996 branched with the closely related genera *N. macacae* and *N. sicca* in a well-supported cluster (Tønjum 2005a; Tanner et al. 2007; Bennett et al. 2012). The other strain, *N. mucosa* C102, was found to consistently branch within a cluster consisting of *N. flavescens* and *N. subflava*. Bennett et al. (2012) have recently reported a detailed phylogenetic analysis of the members of the genus *Neisseria* based on 246 conserved genes. While the overall branching pattern and clustering of the *Neisseria* species and that of *N. mucosa* ATCC 25996 was very similar in their study as observed in the present work, the strain *N. mucosa* C102, which is showing anomalous branching pattern, was not included in their work. Thus, it is likely that the strain *N. mucosa* C102 is presently miscategorised and it is more closely related to *N. flavescens* and *N. subflava* species.

In parallel, we also constructed a phylogenetic tree based upon 16S rRNA gene sequences (Fig. 2). This tree also included representative species from other genera for whom no genome sequences are currently available. The 16S rRNA tree also revealed the existence of a strongly supported monophyletic clade containing the genera *Neisseria*, *Eikenella*, *Kingella*, and *Simonsiella* (Clade I); additionally, this clade also grouped together species from a number of other genera, whose genomes have not yet been sequenced (Fig. 2). Within this clade, *Neisseria mucosa* C102 once again clustered with *Neisseria flavescens* and *Neisseria subflava* in a clade distinct from *N. mucosa* ATCC 25996, *N. sicca* ATCC 29256, and *Neisseria macacae* ATCC 33926. Additionally, in the 16S rRNA tree, the species *Morococcus cerebrosus* was observed to branch within a well-supported grouping of *N. mucosa* ATCC 25996, *N. sicca* ATCC 29256, *N. macacae* ATCC 33926, and *Neisseria* sp. GT4A_CT1 suggesting a close association of the genera *Morococcus* and *Neisseria*. The branching of *M. cerebrosus* CIP 81.93, which is the only reported isolate of this genus, within a strongly supported clade of *Neisseria* species, strongly suggests that this species should not be part of a separate genus, but it should be reclassified as a *Neisseria* species (Sly 2005). It should also be noted that the 16S rRNA sequence for the only isolated species of *Prolinoborus* branches within the class *Gammaproteobacteria* with the genus *Acinetobacter* (not shown) and is likely wrongly assigned to the order *Neisseriales* within the class *Betaproteobacteria*. However, apart from Clade I, the relationship among many of the other genera within the *Neisseriales* was largely unresolved in this tree, and no clade that grouped together the genera *Chromobacterium*, *Laribacter*, and *Pseudogulbenkiania* (Clade II) was observed in the 16S rRNA tree.

Fig. 3 Partial sequence alignments of the proteins. **a** Methionine adenosyltransferase (MetK) and **b** 30S ribosomal protein S4 (RpsD), showing two CSIs (*boxed*) that are uniquely present in various members of the order *Neisseriales*. Sequence information for only a limited number of species from other bacteria is shown here, but unless otherwise indicated similar CSIs were not detected in any other species in the top 250 blast hits. The *dashes* in the alignments indicate identity with the residue in the top sequence. *GI numbers* are indicated for each sequence and are provided for the type strain when available. Information for other CSIs that are specific for the order *Neisseriales* is presented in Table 2 and Supplementary Figs. 2–10

Importance of the CSIs for evolutionary studies and identification of CSIs specific for the order *Neisseriales*

The CSIs in genes/proteins that are restricted to a given group of related species provide very useful molecular markers for evolutionary studies (Gupta 1998, 2009; Rokas and Holland 2000; Gao and Gupta 2012b). The unique shared presence of these highly specific molecular markers in a related group of species is most parsimoniously explained by the occurrence of the rare genetic changes that resulted in these CSIs in a common ancestor of the group, followed by vertical transmission of these CSIs to various descendant species (Gupta 1998; Rokas and Holland 2000; Gao and Gupta 2012a). Hence, these CSIs represent molecular synapomorphies of common evolutionary descent and they provide useful markers for identifying different groups of organisms in molecular terms and for understanding their interrelationships (Gao et al. 2009; Gupta and Bhandari 2011; Gupta et al. 2012; Gao and Gupta 2012b). The CSI-based approach has recently been used to propose



Antonie van Leeuwenhoek (2013) 104:1–24 9

(A)

			160		194
Neisseriales (27/27)	Neisseria bacilliformis	329118153	LPWLRPDAKAQITAVY	DS	ETGKVKRIDTVVLSTQH
	Neisseria polysaccharea	296313551	-----------L-V--	--	-----------------
	Neisseria gonorrhoeae	161572979	-----------L-V--	--	-----------------
	Neisseria lactamica	269214934	-----------L-V--	--	-----------------
	Neisseria flavescens	225077201	-----------L-V--	--	-----------------
	Neisseria subflava	261381356	-----------L-V--	--	-----------------
	Neisseria mucosa	319637619	-----------L-V--	--	-----------------
	Neisseria elongata	294668208	-----------L-V--	--	-----------------
	Neisseria meningitidis	319409877	-----------L-V--	--	-----------------
	Neisseria cinerea	261378257	-----------L-V--	--	-----------------
	Neisseria sicca	255066119	-----------L-V--	--	-----------------
	Neisseria macacae	340363905	-----------L-V--	--	-----------------
	Neisseria sp. GT4A_CT1	349610834	-----------L-V--	--	-----------------
	Neisseria shayeganii	349574807	-----------L-C--	-A	D----------------
	Neisseria wadsworthii	350572334	-------------C--	-A	-----------------
	Neisseria weaveri	345874218	----------------	-A	D----------------
	Eikenella corrodens	225025721	----------------	--	-----------------
	Neisseria sp. oral taxon 020	429743413	-S--------------	--	-----------------
	Neisseria sp. oral taxon 014	298370223	----------------	--	K----------------
	Pseudogulbenkiania ferrooxidans	224824806	---------S---C--	-A	---LP------------
	Pseudogulbenkiania sp. NH8B	347538456	---------S---C--	-A	---LP------------
	Laribacter hongkongensis	226939508	---------S---CA-	--	---LP------------
	Chromobacterium violaceum	34496418	---------S---C--	-A	A--LP------------
	Kingella oralis	238022223	-----------L-VA-	--	-----T-----------
	Kingella kingae	333375324	-----------L-V--	-N	---R-T-----------
	Simonsiella muelleri	294789804	-----------L-VA-	--	---R-C-----------
	Kingella denitrificans	325266141	-----------L-VA-	--	---Q-T----I------
Other Bacteria	Methyloversatilis universalis	334132114	---------S-V-IR-		VD--PE-----------
	Thiomonas sp. 3As	294340797	---------S---FR-		-N-YPVA----------
	Ralstonia solanacearum	299068300	---------S-V-VR-		VN--PHSV---------
	Burkholderia sp. CCGE1002	295675166	---------S-V-VR-		VD--PHS----------
	Cupriavidus necator	339324373	---------S-V-VR-		VD--PHSV---------
	Dechlorosoma suillum	345131704	---------S-V-IR-		VD--PDS----------
	Pusillimonas sp. T7-7	332285806	---------S-V-FR-		ID--PAEV---------
	Nitrosospira multiformis	82703656	---------S-VSVR-		LD--PQ--E---I----
	Ralstonia eutropha	73539907	---------S-V-VR-		VD-RPHSV---------
	Cupriavidus metallidurans	94309101	---------S-V-VR-		VD-RPHSV---------

(B)

			21		60
Neisseriales (23/23)	Neisseria mucosa	319639573	LKSARRSLDSKCKMDSAPGQHGAKKP		RLSDYGLQLREKQK
	Neisseria flavescens	241759655	--------------------------		--------------
	Neisseria subflava	284799892	--------------------------		--------------
	Neisseria sp. oral taxon 014	298370567	-------------I------------		--------------
	Neisseria sicca	255067434	-------------L------------		--------------
	Neisseria weaveri	345874160	--------E-----------------		--------------
	Neisseria lactamica	261400076	-------------L------------		--------------
	Neisseria meningitidis	15676094	-------------I------------		--------------
	Neisseria polysaccharea	296313389	-------------I------------		--------------
	Neisseria macacae	340360980	-------------L------------		--------------
	Neisseria sp. GT4A_CT1	349611070	-------------L------------		--------------
	Neisseria cinerea	261378065	-------------I--------G---		--------------
	Neisseria wadsworthii	350570678	--------E----I------------		--------------
	Neisseria gonorrhoeae	194099938	-------------I------------		--------------
	Neisseria elongata	294669848	-------VE----F------------		--------------
	Neisseria subflava	284799892	--------------------------		--------------
	Kingella denitrificans	325267152	--------E-----------------		--------------
	Kingella kingae	333376289	--------E-----E--S--------		--------------
	Simonsiella muelleri	294789218	--------E-----E----------S		--------------
	Laribacter hongkongensis	226939211	-------I-----L--R-------Q-		------IH------
	Pseudogulbenkiania ferrooxidans	224827041	------A------L--I------R-S		------V-------
	Pseudogulbenkiania sp. NH8B	347538214	------A------L--I------R-S		------V-------
	Chromobacterium violaceum	34499616	------A------L-A-------RRG		------V-------
Other Bacteria	Achromobacter piechaudii	293602599	-------------L--K-----RTSG	A	-T------------
	Bordetella petrii	163859246	-------------L--K-----RTSG	A	-T------------
	Oxalobacter formigenes	237749539	-------------L-TK-------SG	A	-T---N--------
	Herbaspirillum seropedicae	300309481	-------------L--K-----RTSG	A	-T---N--------
	Collimonas fungivorans	340785621	-------------L--K-----RTSG	A	-T---N--------
	Janthinobacterium sp. Marseill	152980403	-------------L--K-----RTSG	A	-T---N--------
	Herminiimonas arsenicoxydans	134096297	-------------L--K-----RTSG	A	-T---N--------
	Burkholderia rhizoxinica	312795803	--------AD---L--K-----RTSG	A	-T------------
	Limnobacter sp. MED105	149926591	-------------L--K-----RTSG	A	-T---Q--------
	Lautropia mirabilis	319944771	-----------A-A-TK-----RTSG	Q	-T------------
	Oxalobacter formigenes	237747389	-------------L-VK-----M-SG	A	-T---N--------
	Burkholderia oklahomensis	167564391	--------AD---L--K-----RTSG	A	-T---T--------

Springer

important taxonomic changes for a number of groups of bacteria (viz. *Chloroflexi*, *Bacillus*, and *Coriobacateriia*) at different taxonomic ranks (Gupta et al. 2012, 2013; Bhandari et al. 2013). In the present work, a comprehensive study was carried out to identify CSIs that are commonly shared by different sequenced species from the order *Neisseriales*. These studies have identified 54 CSIs in diverse and important proteins that are specific for the order *Neisseriales* or a number of its subclades. Brief descriptions of the species specificities of these CSIs and their evolutionary significances are discussed below.

Of the 54 CSIs identified in this work, 11 are specifically found in all of the sequenced species from the order *Neisseriales* and they are not found in homologous proteins from any other bacterial species (in the top 250 Blast hits) (Table 1). The sequence information for two of the CSIs that are specific for the order *Neisseriales* is presented in Fig. 3. In the first example, a 2 amino acid insertion in a highly conserved region of the protein methionine adenosyltransferase (MetK) (Fig. 3a) is uniquely present in all of the sequenced members of the order *Neisseriales*, but not found in any other bacteria. In the second example, a 1 amino acid deletion is present in a highly conserved region of the ribosomal protein S4 (RpsD) that is specific for all detected homologs from the order *Neisseriales* (Fig. 3b). Sequence information for 8 other CSIs in diverse proteins (viz. Porphobilinogen synthase HemB, Single-Stranded-DNA-specific exonuclease RecJ, transcription-repair coupling factor Mfd, ATP phosphoribosyltransferase HisG, Glycine cleavage system aminomethyltransferase GcvT, Hypothetical Protein CV_3579 and NAD(P)+ transhydrogenase (AB-specific) PntA) that are also specifically present in different sequenced species from the order *Neisseriales* is presented in Supplemental Figs. 2–10 and some of their characteristics are summarized in Table 2.

CSIs that are specific for the Clade I species of the order *Neisseriales*

The order *Neisseriales* is currently comprised of a single family, *Neisseriaceae*, containing all 32 genera from this order (Euzeby, 2012). Our analysis has identified 21 CSIs in different proteins that are uniquely present in all sequenced members of the genera *Neisseria*, *Eikenella*, *Kingella* and *Simonsiella* (referred to as Clade I), and absent in any other sequenced *Neisseriales* or in any other bacteria (Table 3). The distinctiveness of the Clade I species within the order *Neisseriales* is independently and strongly supported by their monophyletic grouping in the phylogenetic tree based on concatenated protein sequences and in the 16S rRNA tree. Two examples of the CSIs that are specific for the Clade I species are shown in Fig. 4. In the two examples shown, a 2 amino acid insert in a conserved region of malate dehydrogenase (Mdh) (Fig. 4a), and a 1 amino acid insertion in

Table 2 Conserved signature indels that are specific for members of the order *Neisseriales*

Protein name	Gene name	GenInfo identifier (GI)[a]	Figure number	Indel size	Indel position[b]
Methionine adenosyltransferase	MetK	329118153	Fig. 3a	2 aa ins	160–194
30S ribosomal protein S4	RpsD	319639573	Fig. 3b	1 aa del	21–60
Single-stranded-DNA-specific exonuclease	RecJ	329118639	Supp. Figure 2	1 aa del	159–179
Transcription-repair coupling factor	Mfd	325204285	Supp. Figure 3	2 aa del	401–458
Porphobilinogen synthase	HemB	284799728	Supp. Figure 4	1 aa del	504–553
ATP phosphoribosyltransferase	HisG	329119453	Supp. Figure 5	1 aa ins	78–122
Glycine Cleavage System Aminomethyltransferase T	GcvT	296314946	Supp. Figure 6	3 aa ins	266–303
Hypothetical Protein CV_3579	–	34499034	Supp. Figure 7	1 aa ins	138–171
NAD(P)+ transhydrogenase (AB-specific)	PntA	347539809	Supp. Figure 8	1 aa ins	278–333
NAD(P)+ transhydrogenase (AB-specific)	PntA	347539809	Supp. Figure 9	2 aa del	201–249
Guanine deaminase	GuaD	347541590	Supp. Figure 10	2 aa ins	293–332

[a] GI number provided for the protein used in Blastp query to determine indel specificity within the top 250 hits

[b] The indel-containing region of the protein indicated here corresponds to the amino acid sequence of the protein indicated by the GI number on the same line

Table 3 Conserved signature Indels that are specific for species from Clade I of the *Neisseriales* (*Nesisseria*, *Eikenella*, *Kingella*, and *Simonsiella*)

Protein name	Gene name	GenInfo identifier (GI)[a]	Figure number	Indel size	Indel position[b]
Malate dehydrogenase	Mdh	297250895	Fig. 4a	2 aa ins	209–256
Ribosomal RNA large subunit methyltransferase J	RrmJ	329120491	Fig. 4b	1 aa ins	90–125
4-Hydroxy-3-methylbut-2-enyl diphosphate reductase	IspH	329118302	Supp. Figure 11	1 aa ins	129–173
Deoxyuridine 5′-triphosphate nucleotidohydrolase	Dut	329120658	Supp. Figure 12	1 aa del	20–59
Dihydroorotase	PyrC	329120699	Supp. Figure 13	1 aa ins	217–251
DNA polymerase III subunit alpha	DnaE	241759501	Supp. Figure 14	1 aa ins	145–203
Fructose-1,6-bisphosphatase	Fbp	329120261	Supp. Figure 15	1 aa del	151–200
Fructose-1,6-bisphosphatase	Fbp	329120261	Supp. Figure 16	2 aa del	107–160
GMP synthase	GuaA	325266943	Supp. Figure 17	5 aa del	126–162
Histidine–tRNA ligase	HisS	329120375	Supp. Figure 18	7 aa ins	8–59
Hypothetical protein EIKCOROL_00874	–	225024007	Supp. Figure 19	1 aa ins	426–467
Hypothetical protein EIKCOROL_00974	–	225024106	Supp. Figure 20	1 aa del	298–348
Hypothetical protein NEIFLAOT_00147	–	225075147	Supp. Figure 21	2 aa ins	84–125
Hypothetical protein NEIFLAOT_01683	–	225076633	Supp. Figure 22	1 aa ins	8–36
Hypothetical protein NG_O0349	–	254493262	Supp. Figure 23	2 aa del	126–172
Methionine–tRNA ligase	MetG	309379858	Supp. Figure 24	1 aa ins	619–661
S-adenosylmethionine:tRNA ribosyltransferase-isomerase	QueA	329118190	Supp. Figure 25	1 aa ins	241–272
YhgF-like protein	TexN	329118647	Supp. Figure 26	4 aa del	351–406
Amidophosphoribosyltransferase	PurF	241759844	Supp. Figure 27	1 aa ins	127–170
Anthranilate phosphoribosyltransferase	TrpD	329119146	Supp. Figure 28	1 aa ins	301–336
Succinyldiaminopimelate transaminase	ArgD	329120406	Supp. Figure 29	1 aa del	124–182

[a] GI number provided for the protein used in Blastp query to determine indel specificity within the top 250 hits

[b] The indel-containing region of the protein indicated here corresponds to the amino acid sequence of the protein indicated by the GI number on the same line

the 50S rRNA methyltransferase J (RrmJ) protein (Fig. 4b), are specifically present in all sequenced Clade I species. Both of these CSIs are present in highly conserved regions of these important and widely distributed proteins and except for the sequenced members of the Clade I species, these CSIs are not present in other detected *Neisseriales* homologs or any other bacteria (in the top 250 blast hits). Sequence information for other CSIs that are specific for the Clade I species is presented in Supplementary Figs. 11–29 and a summary of them is provided in Table 3.

Molecular markers that are specific for other clades within the order *Neisseriales*

In addition to the CSIs that are specific for all sequenced *Neisseriales* or the Clade I species, our analyses have also identified many other CSIs that are specific for other *Neisseriales* species. Eight of these CSIs are specifically present in all sequenced species from the three genera of *Neisseriales* that make up Clade II (*Chromobacterium*, *Laribacter*, and *Pseudogulbenkiania*) that are not part of the Clade I. One example of a CSI that is specific for the species from these genera is shown in Fig. 5a. In this case, a 1 aa insert in the Glycine cleavage system aminomethyltransferase T (GcvT) protein is specifically present in the homologs of all four sequenced species from these genera, but not in any other bacteria. In the another example of a CSI that is specific for the Clade II *Neisseriales*, which is shown in Fig. 5b, a 3 aa deletion in the protein propionyl CoA-carboxylase protein (PccB) is specifically present in all four sequenced species from these genera, but not in any other bacteria. Five other identified CSIs in different proteins are also specific for the species from these three genera and sequence information for them is present in Supplementary Figs. 30–34 and information for them is summarized in Table 4.

Our analyses have also identified a number of CSIs that are specific for other smaller clades of *Neisseriales*, which are observed in the concatenated protein trees. Eight of the identified CSIs are specific for a clade comprising of the sequenced members from the genera *Kingella* and *Simonsiella*, 6 others CSIs are uniquely present in the three sequenced members from the genera *Chromobacterium* and *Pseudogulbenkiania*, and a single CSI is largely specific for the genera *Neisseria* and *Eikenella* (Table 5). Two examples of such CSIs, one consisting of a 3 amino acid insert in the protein dTDP-Glucose 4,6-dehydratase (RfbB) that is specific for the genera *Kingella* and *Simonsiella*, and the other consisting of a 6 amino acid insert in a tellurium resistance protein (TerC) that is unique to the genera *Chromobacterium* and *Pseudogulbenkiania* are shown in Fig. 6a, b, respectively. The sequence information for other CSIs that are specific for these clades is presented in Supplementary Figs. 36–47 and a summary of them is provided in Table 5.

Discussion

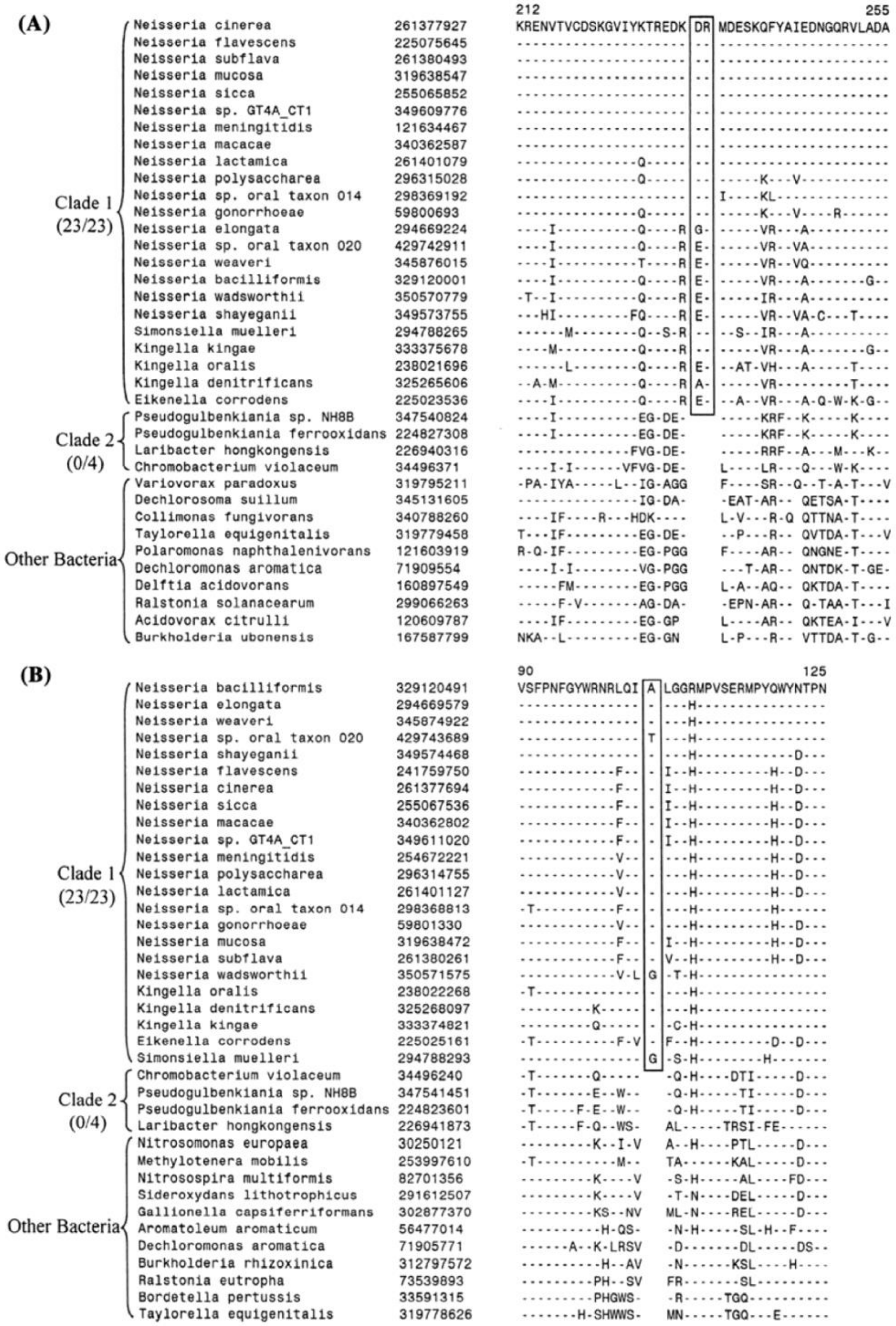
The order *Neisseriales* presently contains 32 genera spanning a wide range of morphologies, habitats, and growth requirements (Tønjum 2005b; Euzeby 2012). The species from this important order are presently distinguished from other bacteria primarily on the basis of their branching in the 16S rRNA gene trees, and no reliable biochemical, molecular or morphological characteristic that is specific for this group is known. Further, all 32 genera from this order are presently grouped into a single family and it has proven difficult to reliably distinguish any distinct subgroup within this order based upon their branching in the 16S rRNA tree or other known characteristics (Bøvre 1984; Harmsen et al. 2001; Hedlund and Staley 2002; Tønjum 2005b; Yarza et al. 2008). The results presented in this study are significant in this regard (Fig. 7).

In phylogenetic trees based upon concatenated sequences for 20 conserved proteins that were constructed in this work, a clade consisting of the genera *Neisseria*, *Eikenella*, *Kingella* and *Simonsiella* (Clade I), was clearly distinguished from all other sequenced genera from this order. A clade encompassing these genera, as well as several other genera for which genome sequences are not available, was also strongly supported in the 16S rRNA gene tree. Importantly, all 12 genera that are part of the Clade 1 in the 16S rRNA tree (viz. *Alysiella*, *Bergeriella*, *Conchiformibius*, *Eikenella*, *Kingella*, *Morococcus*, *Neisseria*, *Simonsiella*, *Stenoxybacter*, *Uruburuella* and *Vitreoscilla*) are obligatory host-associated organisms (except *Vitreoscilla*, which is found in multiple habitats) and they lack flagella (Table 6) (Dewhirst et al. 1989; Xie and Yokota 2005; Tønjum 2005a; Wertz and Breznak 2007). The remaining 19 genera from the order *Neisseriales* are all rod-shaped organisms (with the sole exception of the *Aquaspirillum*, which is a spirillum), which display flagella-based motility, and all are capable of free living (Table 6) (Patureau et al. 1998; Gillis and Logan 2005; Stackebrandt et al. 2007; Yoon et al. 2010). Thus, the Clade I species are also distinct from the other *Neisseriales* genera in terms of their biochemical and morphological characteristics.

Fig. 4 Partial sequence alignments of **a** malate dehydrogenase▸ (Mdh), and **b** ribosomal RNA large subunit methyltransferase protein (RrmJ), showing two CSIs that are specific for the Clade I *Neisseriales*, but not found in the sequence homologs of any other bacteria. Sequence information for other Clade I specific CSIs is presented in Supplementary Figs. 11–29 and summarized in Table 3

In this work we have also identified a large number of molecular markers, consisting of CSIs, which are specific for either all sequenced *Neisseriales* species or for distinct subgroups of them. As indicated earlier, these CSIs, due to their unique presence in specific groups of species, provide valuable markers for evolutionary and taxonomic studies. The work on identification of these CSIs was carried out independently of the phylogenetic trees. In the present study, 11 CSIs in divergent proteins were identified that are specifically present in all sequenced species from the order *Neisseriales*. Based upon earlier work on CSIs from other bacterial phyla/taxa, it is expected that many of these CSIs, if not all, will also be found in other *Neisseriales* species for which no sequence information is available at present (Gupta 2009; Gao and Gupta 2012b). Thus, these CSIs provide us, for the first time, multiple markers for identification of the *Neisseriales* species and the demarcation of this order in molecular terms.

This study also identified 21 CSIs that are specific for the Clade I species (*Nesisseria*, *Eikenella*, *Kingella*, and *Simonsiella*). Based upon the fact that the


(A)
212
KRENVTVCDSKGVIYKTREDK DR MDESKQFYAIEDNGQRVLADA
255
Clade 1 (23/23)
Neisseria cinerea 261377927
Neisseria flavescens 225075645
Neisseria subflava 261380493
Neisseria mucosa 319638547
Neisseria sicca 255065852
Neisseria sp. GT4A_CT1 349609776
Neisseria meningitidis 121634467
Neisseria macacae 340362587
Neisseria lactamica 261401079
Neisseria polysaccharea 296315028
Neisseria sp. oral taxon 014 298369192
Neisseria gonorrhoeae 59800693
Neisseria elongata 294669224
Neisseria sp. oral taxon 020 429742911
Neisseria weaveri 345876015
Neisseria bacilliformis 329120001
Neisseria wadsworthii 350570779
Neisseria shayeganii 349573755
Simonsiella muelleri 294788265
Kingella kingae 333375678
Kingella oralis 238021696
Kingella denitrificans 325265606
Eikenella corrodens 225023536
Clade 2 (0/4)
Pseudogulbenkiania sp. NH8B 347540824
Pseudogulbenkiania ferrooxidans 224827308
Laribacter hongkongensis 226940316
Chromobacterium violaceum 34496371
Other Bacteria
Variovorax paradoxus 319795211
Dechlorosoma suillum 345131605
Collimonas fungivorans 340788260
Taylorella equigenitalis 319779458
Polaromonas naphthalenivorans 121603919
Dechloromonas aromatica 71909554
Delftia acidovorans 160897549
Ralstonia solanacearum 299066263
Acidovorax citrulli 120609787
Burkholderia ubonensis 167587799
(B)
90
VSFPNFGYWRNRLQI A LGGRMPVSERMPYQWYNTPN
125
Clade 1 (23/23)
Neisseria bacilliformis 329120491
Neisseria elongata 294669579
Neisseria weaveri 345874922
Neisseria sp. oral taxon 020 429743689
Neisseria shayeganii 349574468
Neisseria flavescens 241759750
Neisseria cinerea 261377694
Neisseria sicca 255067536
Neisseria macacae 340362802
Neisseria sp. GT4A_CT1 349611020
Neisseria meningitidis 254672221
Neisseria polysaccharea 296314755
Neisseria lactamica 261401127
Neisseria sp. oral taxon 014 298368813
Neisseria gonorrhoeae 59801330
Neisseria mucosa 319638472
Neisseria subflava 261380261
Neisseria wadsworthii 350571575
Kingella oralis 238022268
Kingella denitrificans 325268097
Kingella kingae 333374821
Eikenella corrodens 225025161
Simonsiella muelleri 294788293
Clade 2 (0/4)
Chromobacterium violaceum 34496240
Pseudogulbenkiania sp. NH8B 347541451
Pseudogulbenkiania ferrooxidans 224823601
Laribacter hongkongensis 226941873
Other Bacteria
Nitrosomonas europaea 30250121
Methylotenera mobilis 253997610
Nitrosospira multiformis 82701356
Sideroxydans lithotrophicus 291612507
Gallionella capsiferriformans 302877370
Aromatoleum aromaticum 56477014
Dechloromonas aromatica 71905771
Burkholderia rhizoxinica 312797572
Ralstonia eutropha 73539893
Bordetella pertussis 33591315
Taylorella equigenitalis 319778626

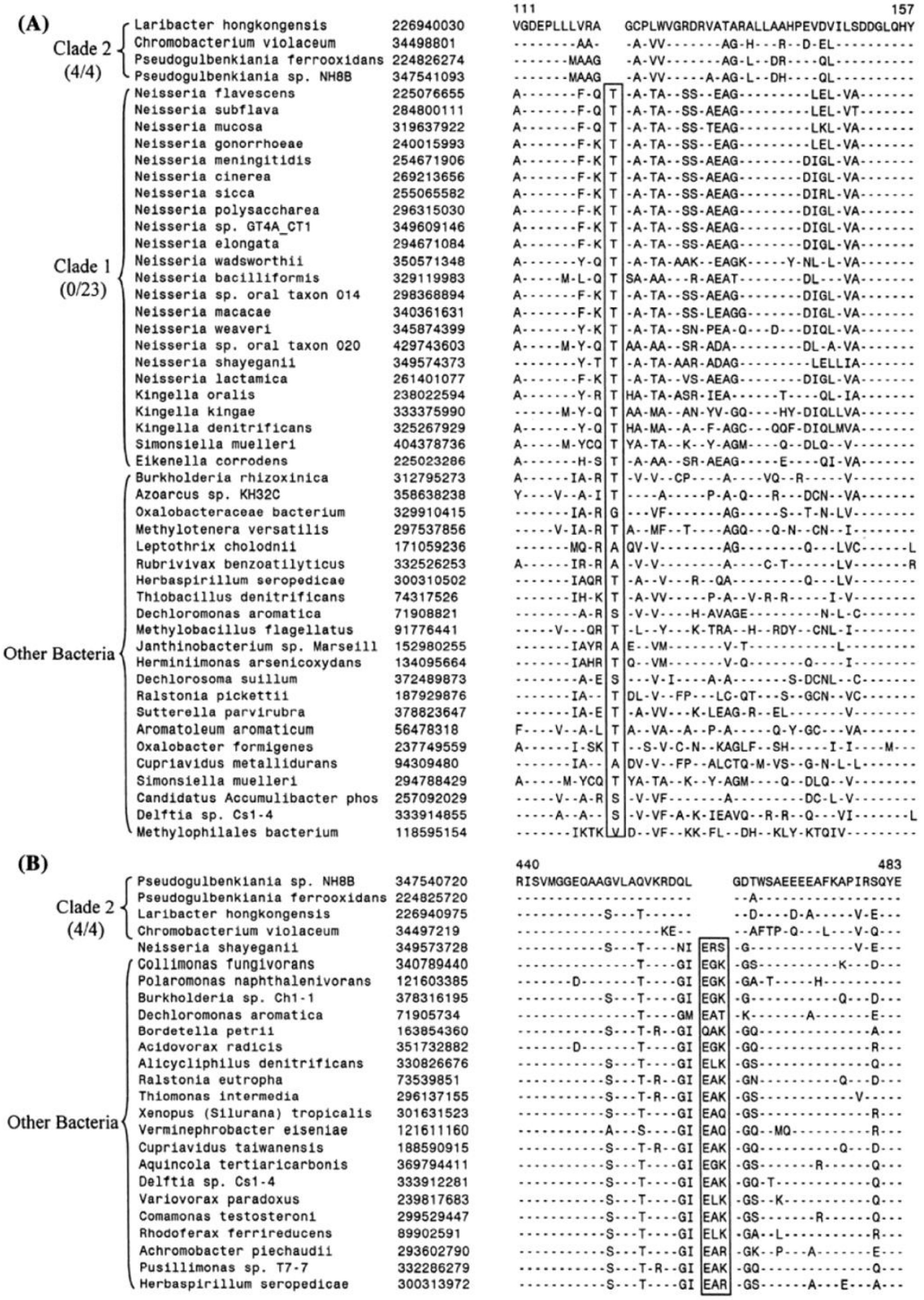
four genera for which sequences are available are dispersed within Clade I, it is highly likely that these CSIs are distinctive characteristics of the Clade I. Hence, it is expected that many of the CSIs, if not all, will also be present in other species/genera from Clade I, for whom sequence information is not available at present. The discovery of these large number of synapomorphic molecular markers for the Clade I species provides compelling evidence that this group of species represents a distinct subclade within the order *Neisseriales*.

This work also identified 7 CSIs that were specifically present in the other sequenced *Neisseriales* that are not part of the Clade I. Although these genera formed a well-supported clade in the protein tree (Clade II), in the 16S rRNA tree, where sequence information was included for additional genera within the order *Neisseriales*, no specific grouping of the Clade II genera was observed. Due to the divergent branching of these genera in the16S rRNA tree, and the paucity of genome sequence information for them, the evolutionary significance of these latter CSIs is unclear at present. Additional sequence information from genera that are not part of the Clade I is needed to determine their evolutionary significance.

Multiple CSIs were also identified in the present study that are specific for one of two smaller clades of *Neisseriales*. One of these clades consisted of species from the genera *Kingella* and *Simonsiella* and the other indicated a close affinity of species from the genera *Chromobacterium* and *Pseudogulbenkiania* (Table 5). The latter two genera are part of a distinct clade in the 16S rRNA tree that also includes some additional genera (Fig. 1b). Thus, it is possible that these CSIs could prove useful in demarcating some additional distinct clades within the order *Neisseriales*. However, sequence information from additional genera within the order *Neisseriales* will be needed to reliably determine the evolutionary significance of these CSIs. It should be noted that most of the identified CSIs are present in genes/proteins that contain many conserved regions. Hence, degenerate PCR primers based on conserved regions flanking these CSIs can be designed to specifically amplify the intervening regions to determine the presence or absence of these CSIs in species for which genome sequences are not available (Gao and Gupta 2005; Griffiths et al. 2005).

Fig. 5 Partial sequence alignments of the proteins. **a** Tetra-acyldisaccharide 4′-kinase (LpxA) and **b** propionyl-CoA carboxylase beta subunit (PecB) showing two CSIs that are specific for the other *Neisseriales* genera, except those from Clade I. The sequence homologs for the Clade I species were not detected in the BLAST search for propionyl-CoA carboxylase beta subunit (PecB). Sequence information for other CSIs exhibiting similar specificities is presented in Table 4 and Supplementary Figs. 30–34

Our results also provide novel insight into the evolutionary history of the order *Neisseriales*. The phylogenetic trees and CSIs identified in this study suggest that all obligate host-associated *Neisseriales* (Clade I) are members of a distinct monophyletic lineage, which differs from the rest of the *Neisseriales*, that are capable of free-living, in many important characteristics. Of these two major groups within the order, the free-living *Neisseriales* exhibit deeper branching in the phylogenetic trees and they are separated from the host-associated organisms by a long branch, which is indicative of rapid sequence divergence. Obligate host-associated organisms have been found to exhibit faster sequence evolution than their free living relatives (Moran 1996; Wernegreen and Moran 1999; Wernegreen 2011). These observations suggest that the common ancestor of all obligate host-associated *Neisseriales* was a bacterium capable of free-living that underwent rapid divergence and lost the ability to live independently of a host. It is also of interest to note, in this regard, that the deepest branching member of Clade I, *Vitreoscilla*, is the only member within the lineage that has been isolated from both host-associated and environmental samples. The evolution of this lineage from a free-living to obligate host associated is further evidenced by the apparent reduction in the genome sizes of the Clade I species. Reductive genome evolution is a characteristic of adaptation to obligate host-associated environments. Host-associated environments are relatively stable and provide host-associated bacteria with a number of metabolites and biosynthetic intermediaries which they no longer need to produce themselves allowing for significant genome reduction (Moran 2002; McCutcheon and Moran 2011). The genomes of the Clade I *Neisseriales* range in size from 1.87 to 2.83 Mb while the Clade II species have genomes that range in size from 3.17 to 4.75 Mb (Table 1). The present work has also identified large numbers of CSIs in important proteins that are uniquely shared by all of the Clade I species. The shared presence of these CSIs



```
(A)                                                                   111                                               157
                  Laribacter hongkongensis            226940030      VGDEPLLLVRA   GCPLWVGRDRVATARALLAAHPEVDVILSDDGLQHY
Clade 2           Chromobacterium violaceum           34498801       --------AA-   -A-VV-------AG-H---R--D-EL---------
(4/4)             Pseudogulbenkiania ferrooxidans     224826274      -------MAAG   -A-VV-------AG-L--DR----QL---------
                  Pseudogulbenkiania sp. NH8B         347541093      -------MAAG   -A-VV-----A-AG-L--DH----QL---------
                  Neisseria flavescens                225076655      A-------F-Q | T | -A-TA--SS--EAG---------LEL-VA------
                  Neisseria subflava                  284800111      A-------F-Q | T | -A-TA--SS-AEAG---------LEL-VT------
                  Neisseria mucosa                    319637922      A-------F-Q | T | -A-TA--SS-TEAG---------LKL-VA------
                  Neisseria gonorrhoeae               240015993      A-------F-K | T | -A-TA--SS--EAG---------LEL-VA------
                  Neisseria meningitidis              254671906      A-------F-K | T | -A-TA--SS-AEAG--------DIGL-VA------
                  Neisseria cinerea                   269213656      A-------F-K | T | -A-TA--SS-AEAG--------DIGL-VA------
                  Neisseria sicca                     255065582      A-------F-K | T | -A-TA--SS-AEAG--------DIRL-VA------
                  Neisseria polysaccharea             296315030      A-------F-K | T | -A-TA--SS-AEAG--------DIGL-VA------
                  Neisseria sp. GT4A_CT1              349609146      A-------F-K | T | -A-TA--SS-AEAG--------DIGL-VA------
                  Neisseria elongata                  294671084      A-------F-K | T | -A-TA--SS-AEAG--------DIGL-VA------
Clade 1           Neisseria wadsworthii               350571348      A-------Y-Q | T | -A-TA-AAK--EAGK-----Y-NL-L-VA------
(0/23)            Neisseria bacilliformis             329119983      A-----M-L-Q | T | SA-AA---R-AEAT--------DL---VA------
                  Neisseria sp. oral taxon 014        298368894      A-------F-K | T | -A-TA--SS-AEAG--------DIGL-VA------
                  Neisseria macacae                   340361631      A-------F-K | T | -A-TA--SS-LEAGG-------DIGL-VA------
                  Neisseria weaveri                   345874399      A-------Y-K | T | -A-TA--SN-PEA-Q---D---DIQL-VA------
                  Neisseria sp. oral taxon 020        429743603      A-----M-Y-Q | T | AA-AA--SR-ADA---------DL-A-VA------
                  Neisseria shayeganii                349574373      --------Y-T | T | -A-TA-AAR-ADAG---------LELLIA------
                  Neisseria lactamica                 261401077      A-------F-K | T | -A-TA--VS-AEAG--------DIGL-VA------
                  Kingella oralis                     238022594      A-------Y-R | T | HA-TA-ASR-IEA------T----QL-IA------
                  Kingella kingae                     333375990      ------M-Y-Q | T | AA-MA--AN-YV-GQ----HY-DIQLLVA------
                  Kingella denitrificans              325267929      A-------Y-Q | T | HA-MA--A--F-AGC---QQF-DIQLMVA------
                  Simonsiella muelleri                404378736      A-----M-YCQ | T | YA-TA--K--Y-AGM----Q--DLQ--V-------
                  Eikenella corrodens                 225023286      A-------H-S | T | -A-AA--SR-AEAG-----E----QI-VA------
                  Burkholderia rhizoxinica            312795273      A------IA-R | T | -V-V--CP----A----VQ--R-----V-------
                  Azoarcus sp. KH32C                  358638238      Y----V--A-I | T | ----A-----P-A-Q---R---DCN--VA------
                  Oxalobacteraceae bacterium          329910415      -------IA-R | G | ---VF-------AG-----S--T-N-LV-------
                  Methylotenera versatilis            297537856      -----V-IA-R | T | A--MF--T----AGQ---Q-N--CN--I-------
                  Leptothrix cholodnii                171059236      -------MQ-R | A | QV-V--------AG--------Q---LVC-----L
                  Rubrivivax benzoatilyticus          332526253      A------IR-R | A | -V-V--------A----C-T------LV------R
                  Herbaspirillum seropedicae          300310502      -------IAQR | T | -A--V---R--QA---------Q---LV-------
                  Thiobacillus denitrificans          74317526       -------IH-K | T | -A-VV-----P-A--V-R-R-----I-V-------
                  Dechloromonas aromatica             71908821       --------A-R | S | -V-V----H-AVAGE---------N-L-C------
                  Methylobacillus flagellatus         91776441       -----V---QR | T | -L--Y---K-TRA--H--RDY--CNL-I-------
Other Bacteria    Janthinobacterium sp. Marseill      152980255      -------IAYR | A | E--VM-------V-T-----------L--------
                  Herminiimonas arsenicoxydans        134095664      -------IAHR | T | Q--VM-------V-Q-------Q----I-------
                  Dechlorosoma suillum                372489873      --------A-E | S | ---V-I----A-A-------S-DCNL--C------
                  Ralstonia pickettii                 187929876      -------IA-- | T | DL-V--FP---LC-QT---S--GCN--VC------
                  Sutterella parvirubra               378823647      -------IA-E | T | -A-VV---K-LEAG-R--EL-------V-------
                  Aromatoleum aromaticum              56478318       F----V--A-L | T | A--VA--A--P-A-----Q-Y-GC---VA------
                  Oxalobacter formigenes              237749559      A------I-SK | T | --S-V-C-N--KAGLF--SH-----I-I----M--
                  Cupriavidus metallidurans           94309480       -------IA-- | A | DV-V--FP--ALCTQ-M-VS--G-N-L-L------
                  Simonsiella muelleri                294788429      A-----M-YCQ | T | YA-TA--K--Y-AGM----Q--DLQ--V-------
                  Candidatus Accumulibacter phos      257092029      -----V--A-R | S | -V-VF-------A---------DC-L-V-------
                  Delftia sp. Cs1-4                   333914855      -----A--A-- | S | -V-VF-A-K-IEAVQ--R-R--Q---VI------L
                  Methylophilales bacterium           118595154      -------IKTK | V | D--VF--KK-FL--DH--KLY-KTQIV--------

(B)                                                                   440                                     483
                  Pseudogulbenkiania sp. NH8B         347540720      RISVMGGEQAAGVLAQVKRDQL         GDTWSAEEEEAFKAPIRSQYE
Clade 2           Pseudogulbenkiania ferrooxidans     224825720      ----------------------         --A------------------
(4/4)             Laribacter hongkongensis            226940975      -----------S---T------         --D----D-A-----V-E---
                  Chromobacterium violaceum           34497219       -----------------KE--          --AFTP-Q---L---V-Q---
                  Neisseria shayeganii                349573728      -----------S---T----NI | ERS | -G-------------V-E---
                  Collimonas fungivorans              340789440      ---------------T----GI | EGK | -GS----------K---D---
                  Polaromonas naphthalenivorans       121603385      -------D-------T----GI | EGK | -GA-T-----H----------
                  Burkholderia sp. Ch1-1              378316195      -----------S---T----GI | EGK | -G-----------Q---D---
                  Dechloromonas aromatica             71905734       ---------------T----GM | EAT | -K-------A-------E---
                  Bordetella petrii                   163854360      -----------S---T-R--GI | QAK | -GQ--------------A---
                  Acidovorax radicis                  351732882      -------D-------T----GI | EGK | -GQ--------------R---
                  Alicycliphilus denitrificans        330826676      -----------S---T----GI | ELK | -GS--------------Q---
                  Ralstonia eutropha                  73539851       -----------S---T-R--GI | EAK | -GN----------Q---D---
                  Thiomonas intermedia                296137155      -----------S---T-R--GI | EAK | -GS------------V-----
Other Bacteria    Xenopus (Silurana) tropicalis       301631523      -----------S---T----GI | EAQ | -GS--------------R---
                  Verminephrobacter eiseniae          121611160      -----------A---S----GI | EAQ | -GQ--MQ----------R---
                  Cupriavidus taiwanensis             188590915      -----------S---T-R--GI | EAK | -GQ----------Q---D---
                  Aquincola tertiaricarbonis          369794411      -----------S---T----GI | EGK | -GS-------R------Q---
                  Delftia sp. Cs1-4                   333912281      -----------S---T----GI | EAK | -GQ-T------------Q---
                  Variovorax paradoxus                239817683      -----------S---T----GI | ELK | -GS--K-----------Q---
                  Comamonas testosteroni              299529447      -----------S---T----GI | EAK | -GS-------R------Q---
                  Rhodoferax ferrireducens            89902591       -----------S---T----GI | ELK | -GA--L-----------R---
                  Achromobacter piechaudii            293602790      -----------S---T----GI | EAR | -GK--P---A-------E---
                  Pusillimonas sp. T7-7               332286279      -----------S---T-R--GI | EAK | -GQ--------------Q---
                  Herbaspirillum seropedicae          300313972      -----------S---T----GI | EAR | -GS------A---E---A---
```

Table 4 Conserved signature indels that are specific for species from Clade II of the *Neisseriales* (*Chromobacterium*, *Laribacter*, and *Pseudogulbenkiania*)

Protein name	Gene name	GenInfo identifier (GI)[a]	Figure number	indel size	Indel position[b]
Tetraacyldisaccharide 4′-kinase	LpxK	226940030	Figure 5a	1 aa del	111–157
Propionyl-CoA carboxylase subunit beta	PccB	347540720	Figure 5b	3 aa del	440–486
Glycine cleavage system aminomethyltransferase T	GcvT	226941640	Supp. Figure 30	1 aa ins	287–325
Hypothetical protein CV_3451	–	34498906	Supp. Figure 31	1 aa ins	271–305
Methylmalonate-semialdehyde dehydrogenase	IolA	347539569	Supp. Figure 32	1 aa del	292–340
Ribonucleoside-diphosphate reductase subunit alpha	NrdE	226940518	Supp. Figure 33	1 aa del	205–263
Succinyl-CoA synthetase subunit alpha	SucC	226941328	Supp. Figure 34	2 aa del	96–143

[a] GI number provided for the protein used in Blastp query to determine indel specificity within the top 250 hits

[b] The indel-containing region of the protein indicated here corresponds to the amino acid sequence of the protein indicated by the GI number on the same line

in all of the Clade I species indicate that the genetic changes responsible for them were introduced in a common ancestor of the Clade I species, presumably at the stage when obligate host-association was initially established, prior to the differentiation of this lineage into its various decedent organisms. Additionally, the presence of these CSIs exclusively in all sequenced obligate host-associated *Neisseriales* suggests that they may play an essential, functional role in the adaptation of these organisms to an obligate host-associated lifestyle (Singh and Gupta 2009). The data reported here thus provides the first clear insights into the evolutionary history of the obligate host-associated *Neisseriales* and provides a framework for further evolutionary studies on the remaining lineages within this order.

Taxonomic implications

The results presented here show that the order *Neisseriales* is comprised of at least two distinct higher order clades. One large clade consisting of the genera *Alysiella*, *Bergeriella*, *Conchiformibius*, *Eikenella*, *Kingella*, *Morococcus*, *Neisseria*, *Simonsiella*, *Stenoxybacter*, *Uruburuella* and *Vitreoscilla* is strongly supported by the 16S rRNA gene tree. The distinctness of the genera that are part of this clade is also strongly supported by the tree based upon concatenated protein sequences and by the large numbers of discovered CSIs that are specific for the species from this clade. Members of this clade are also distinguished from other *Neisseriales* due to being comprised of obligatory host-associated organisms that lack flagella and show varied morphology (Dewhirst et al. 1989; Xie and Yokota 2005; Tønjum 2005a; Wertz and Breznak 2007). In contrast, the remainder of the genera within the order *Neisseriales* are free living rod-shaped organisms which exhibit flagella-based motility (with the sole exception of the *Aquaspirillum*, which is a spirillum) (Patureau et al. 1998; Gillis and Logan 2005; Stackebrandt et al. 2007; Yoon et al. 2010). Although some CSIs were identified for the sequenced members of the remaining *Neisseriales*, the species from these genera do not form a coherent grouping in the 16S rRNA gene tree. Branching in the 16S rRNA tree suggests that these other genera within the order *Neisseriales* would likely form more than one distinct higher taxonomic grouping within this order. However, reliable grouping of these genera into distinct taxonomic groups requires additional sequence information from genera within the order *Neisseriales*. Nevertheless, Clade I species/genera are indicated to be distinct from all other *Neisseriales* by different lines of evidences. To recognize the distinctiveness of the Clade I species we are proposing division of the order *Neisseriales* into two families. In this proposal the existing family *Neisseriaceae* will be emended to retain only the genera *Alysiella*, *Bergeriella*, *Conchiformibius*, *Eikenella*, *Kingella*, *Morococcus*, *Neisseria*, *Simonsiella*, *Stenoxybacter*, *Uruburuella* and *Vitreoscilla* that correspond to Clade I. The remainder of the genera from the order *Neisseriales (*viz. *Andreprevotia*, *Aquaspirillum*, *Aquitalea*, *Chitinibacter*, *Chitinilyticum*, *Chitiniphilus*, *Chromobacterium*, *Deefgea*, *Formivibrio*, *Gulbenkiania*, *Iodobacter*, *Jeongeupia*, *Laribacter*, *Leeia*, *Microvirgula*, *Paludibacterium*, *Prolinoborus*, *Pseudogulbenkiania*, *Silvimonas*, and *Vogesella*) will

Table 5 Conserved signature Indels that are specific for smaller clades within the order *Neisseriales*

Protein name	Gene name	GenInfo identifier (GI)[a]	Figure number	Specificity	Indel size	Indel position[b]
Glycine cleavage system aminomethyltransferase T	GcvT	226941640	Supp. Figure 35	*Neisseria* and *Eikenella*	2 aa ins	287–325
dTDP-glucose 4,6-dehydratase	RfbB	333376110	Figure 6A	*Kingella* and *Simonsiella*	3 aa ins	240–294
Anthranilate phosphoribosyltransferase	TrpD	333374977	Supp. Figure 36	*Kingella* and *Simonsiella*	1 aa ins	248–307
Multifunctional CCA protein	Cca	333375100	Supp. Figure 37	*Kingella* and *Simonsiella*	1 aa ins	281–340
NADH-quinone oxidoreductase subunit D	NqrD	325267317	Supp. Figure 38	*Kingella* and *Simonsiella*	1 aa ins	288–343
Hypothetical protein GCWU000324_00882	–	238020985	Supp. Figure 39	*Kingella* and *Simonsiella*	8 aa ins	386–433
Hypothetical protein GCWU000324_02377	–	238022469	Supp. Figure 40	*Kingella* and *Simonsiella*	1 aa ins	104–128
Hypothetical protein GCWU000324_02583	–	238022674	Supp. Figure 41	*Kingella* and *Simonsiella*	1 aa ins	11–58
Pyruvate kinase	Pyk	238022702	Supp. Figure 42	*Kingella* and *Simonsiella*	1 aa ins	201–243
Tellurium resistance protein	TerC	224825610	Figure 6B	*Chromobacterium* and *Pseudogulbenkiania*	6 aa ins	267–309
Helicase C2	DinG	224824996	Supp. Figure 43	*Chromobacterium* and *Pseudogulbenkiania*	1 aa del	30–74
Electron-transferring-flavoprotein dehydrogenase	EtfD	34499371	Supp. Figure 44	*Chromobacterium* and *Pseudogulbenkiania*	2 aa del	141–182
Acetate permease	ActP	224825666	Supp. Figure 45	*Chromobacterium* and *Pseudogulbenkiania*	2 aa del	241–285
Hypothetical protein CV_2031	–	347540101	Supp. Figure 46	*Chromobacterium* and *Pseudogulbenkiania*	2 aa del	200–249
UDP-N-acetylenolpyruvoylglucosamine reductase	MurB	34497047	Supp. Figure 47	*Chromobacterium* and *Pseudogulbenkiania*	1 aa ins	126–172

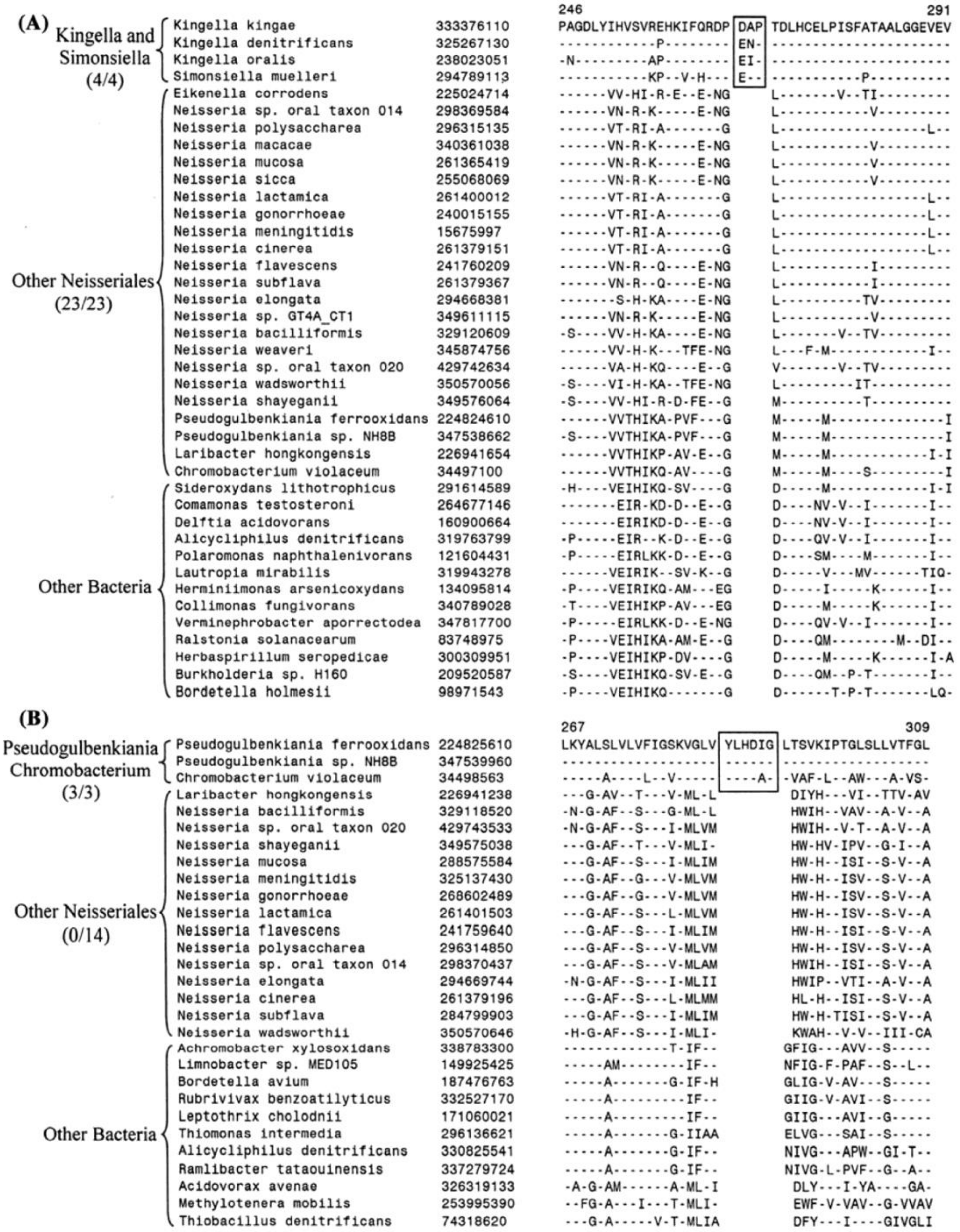
[a] GI number provided for the protein used in Blastp query to determine indel specificity within the top 250 hits

[b] The indel-containing region of the protein indicated here corresponds to the amino acid sequence of the protein indicated by the GI number on the same line

be transferred to a new family, *Chromobacteriaceae* fam. nov. The emended descriptions of the order *Neisseriales* and the family *Neisseriaceae*, as well as a description of the new family *Chromobacteriaceae* fam. nov. are provided below.

Emended description of the order *Neisseriales* (Tønjum 2006)

The order contains two families, *Neisseriaceae* and *Chromobacteriaceae*, of which *Neisseriaceae* is the type family. Organisms are coccal, coccoid, or distinctly rod-shaped occurring singly, in pairs, in masses, or in short chains. Endospores are not formed. The cells are Gram-negative, but there may be a tendency to resist decolouration. Flagella and swimming motility are present in some genera. Surface-bound motility ("twitching motility") is frequently observed. Fimbriae (pili) are often present. All species grow aerobically with optimal temperature of approximately 32–36 °C. Capsules may be present. Colonies are not pigmented except those of *Chromobacterium* and strains of *Vogesella*. The mol% G+C of the DNA is 41–70. The type genus is *Neisseria* (Trevisan 1885).



```
(A)
                                           246                                          291
Kingella and Simonsiella (4/4)
Kingella kingae                 333376110  PAGDLYIHVSVREHKIFQRDP DAP TDLHCELPISFATAALGGEVEV
Kingella denitrificans          325267130  ------------P-------- EN- ----------------------
Kingella oralis                 238023051  -N---------AP-------- EI- ----------------------
Simonsiella muelleri            294789113  -----------KP--V-H--- E-- -----------P----------
Other Neisseriales (23/23)
Eikenella corrodens             225024714  ------VV-HI-R-E--E-NG     L-------V--TI---------
Neisseria sp. oral taxon 014    298369584  ------VN-R-K-----E-NG     L-----------V---------
Neisseria polysaccharea         296315135  ------VT-RI-A-------G     L------------------L--
Neisseria macacae               340361038  ------VN-R-K-----E-NG     L-----------V---------
Neisseria mucosa                261365419  ------VN-R-K-----E-NG     L-----------V---------
Neisseria sicca                 255068069  ------VN-R-K-----E-NG     L-----------V---------
Neisseria lactamica             261400012  ------VT-RI-A-------G     L------------------L--
Neisseria gonorrhoeae           240015155  ------VT-RI-A-------G     L------------------L--
Neisseria meningitidis          15675997   ------VT-RI-A-------G     L------------------L--
Neisseria cinerea               261379151  ------VT-RI-A-------G     L------------------L--
Neisseria flavescens            241760209  ------VN-R--Q----E-NG     L-----------I---------
Neisseria subflava              261379367  ------VN-R--Q----E-NG     L-----------I---------
Neisseria elongata              294668381  -------S-H-KA----E-NG     L----------TV---------
Neisseria sp. GT4A_CT1          349611115  ------VN-R-K-----E-NG     L-----------V---------
Neisseria bacilliformis         329120609  -S----VV-H-KA----E-NG     L-------V--TV---------
Neisseria weaveri               345874756  ------VV-H-K---TFE-NG     L---F-M------------I--
Neisseria sp. oral taxon 020    429742634  ------VA-H-KQ----E--G     V-------V--TV---------
Neisseria wadsworthii           350570056  -S----VI-H-KA--TFE-NG     L---------IT----------
Neisseria shayeganii            349576064  -S----VV-HI-R-D-FE--G     M----------T----------
Pseudogulbenkiania ferrooxidans 224824610  ------VVTHIKA-PVF---G     M-----M--------------I
Pseudogulbenkiania sp. NH8B     347538662  -S----VVTHIKA-PVF---G     M-----M--------------I
Laribacter hongkongensis        226941654  ------VVTHIKP-AV-E--G     M-----M------------I-I
Chromobacterium violaceum       34497100   ------VVTHIKQ-AV----G     M-----M----S---------I
Other Bacteria
Sideroxydans lithotrophicus     291614589  -H----VEIHIKQ-SV----G     D-----M------------I-I
Comamonas testosteroni          264677146  -------EIR-KD-D--E--G     D----NV-V--I-------I--
Delftia acidovorans             160900664  -------EIRIKD-D--E--G     D----NV-V--I-------I--
Alicycliphilus denitrificans    319763799  -P-----EIR--K-D--E--G     D----QV-V--I-------I--
Polaromonas naphthalenivorans   121604431  -P-----EIRLKK-D--E--G     D----SM----M-------I--
Lautropia mirabilis             319943278  ------VEIRIK--SV-K--G     D-----V---MV------TIQ-
Herminiimonas arsenicoxydans    134095814  -P----VEIRIKQ-AM---EG     D-----I-----K------I--
Collimonas fungivorans          340789028  -T----VEIHIKP-AV---EG     D-----M-----K------I--
Verminephrobacter aporrectodea  347817700  -P-----EIRLKK-D--E-NG     D----QV-V--I-------I--
Ralstonia solanacearum          83748975   -P----VEIHIKA-AM-E--G     D----QM--------M--DI--
Herbaspirillum seropedicae      300309951  -P----VEIHIKP-DV----G     D-----M-----K------I-A
Burkholderia sp. H160           209520587  -S----VEIHIKQ-SV-E--G     D----QM--P-T-------I--
Bordetella holmesii             98971543   -P----VEIHIKQ-------G     D------T-P-T-------LQ-

(B)
                                           267                                       309
Pseudogulbenkiania Chromobacterium (3/3)
Pseudogulbenkiania ferrooxidans 224825610  LKYALSLVLVFIGSKVGLV YLHDIG LTSVKIPTGLSLLVTFGL
Pseudogulbenkiania sp. NH8B     347539960  ------------------- ------ ------------------
Chromobacterium violaceum       34498563   -----A----L--V----- ----A- -VAF-L--AW---A-VS-
Other Neisseriales (0/14)
Laribacter hongkongensis        226941238  ---G-AV--T---V-ML-L          DIYH---VI--TTV-AV
Neisseria bacilliformis         329118520  -N-G-AF--S---G-ML-L          HWIH--VAV--A-V--A
Neisseria sp. oral taxon 020    429743533  -N-G-AF--S---I-MLVM          HWIH--V-T--A-V--A
Neisseria shayeganii            349575038  ---G-AF--T---V-MLI-          HW-HV-IPV--G-I--A
Neisseria mucosa                288575584  ---G-AF--S---I-MLIM          HW-H--ISI--S-V--A
Neisseria meningitidis          325137430  ---G-AF--G---V-MLVM          HW-H--ISV--S-V--A
Neisseria gonorrhoeae           268602489  ---G-AF--G---V-MLVM          HW-H--ISV--S-V--A
Neisseria lactamica             261401503  ---G-AF--S---L-MLVM          HW-H--ISV--S-V--A
Neisseria flavescens            241759640  ---G-AF--S---I-MLIM          HW-H--ISI--S-V--A
Neisseria polysaccharea         296314850  ---G-AF--S---V-MLVM          HW-H--ISV--S-V--A
Neisseria sp. oral taxon 014    298370437  ---G-AF--S---V-MLAM          HWIH--ISI--S-V--A
Neisseria elongata              294669744  -N-G-AF--S---I-MLII          HWIP--VTI--A-V--A
Neisseria cinerea               261379196  ---G-AF--S---L-MLMM          HL-H--ISI--S-V--A
Neisseria subflava              284799903  ---G-AF--S---I-MLIM          HW-H-TISI--S-V--A
Neisseria wadsworthii           350570646  -H-G-AF--S---I-MLI-          KWAH--V-V--III-CA
Other Bacteria
Achromobacter xylosoxidans      338783300  -------------T-IF--         GFIG---AVV--S-----
Limnobacter sp. MED105          149925425  -----AM--------IF--         NFIG-F-PAF--S--L--
Bordetella avium                187476763  -----A-------G-IF-H         GLIG-V-AV---S-----
Rubrivivax benzoatilyticus      332527170  -----A---------IF--         GIIG-V-AVI--S-----
Leptothrix cholodnii            171060021  -----A---------IF--         GIIG---AVI--G-----
Thiomonas intermedia            296136621  -----A-------G-IIAA         ELVG---SAI--S-----
Alicycliphilus denitrificans    330825541  -----A-------G-IF--         NIVG---APW--GI-T--
Ramlibacter tataouinensis       337279724  -----A-------G-IF--         NIVG-L-PVF--G--A--
Acidovorax avenae               326319133  -A-G-AM------A-ML-I          DLY---I-YA----GA-
Methylotenera mobilis           253995390  --FG-A---I---T-MLI-          EWF-V-VAV--G-VVAV
Thiobacillus denitrificans      74318620   ---G-A-----V-T-MLIA          DFY---I----GIVGLI
```

◂**Fig. 6** **a** Partial sequence alignment of the protein dTDP-glucose 4,6-dehydratase (RfbB) containing a three amino acid insert in a conserved region that is specifically present in the species from the genera *Kingella* and *Simonsiella*, but not found in any other bacteria. **b** Partial sequence alignment of a tellurium resistance protein (TerC) containing a six amino acid insert that is specific for the genera *Chromobacterium* and *Pseudogulbenkiania*. The homologs for some *Neisseriales* species were not detected in the BLAST searches for TerC protein. Sequence information for other CSIs showing similar specificities is provided in Table 5 and in Supplementary Figs. 35–47

Organisms from this order are distinguished from all other *Betaproteobacteria* by the conserved signature indels described in this report in the following proteins: 30S ribosomal protein S4 (RpsD), ATP phosphoribosyltransferase (HisG), glycine cleavage system aminomethyltransferase (GcvT), hypothetical protein CV_3579, methionine adenosyltransferase (MetK), NAD(P)+ transhydrogenase (AB-specific) (PntA), porphobilinogen synthase (HemB), single-stranded-DNA-specific exonuclease (RecJ), guanine deaminase (GuaD), and in the transcription-repair coupling factor Mfd.

Emended description of the family *Neisseriaceae* (Prévot 1933 emend. Dewhirst et al. 1989)

The genera included in this family are *Alysiella*, *Bergeriella*, *Conchiformibius*, *Eikenella*, *Kingella*, (*Morococcus*), *Neisseria*, *Simonsiella*, *Stenoxybacter*, *Uruburuella* and *Vitreoscilla*. The type genus of this family is *Neisseria* (Trevisan 1885). Organisms are coccal, coccoid, or distinctly rod-shaped occurring singly, in pairs, in masses, or in short chains. Cells of *Simonsiella* and *Alysiella* may exhibit a characteristic multicellular micromorphology. Flagella and swimming motility are absent. The cells are largely nonmotile in liquid media, however, gliding motility is observed in certain strains. Surface-bound motility ("twitching motility") is frequently observed. Fimbriae (pili) are often present. All species are obligate host-associated organisms; one genus, *Vitreoscilla*, is an exception and has been found in multiple habitats. Colonies are not pigmented. Several species have complex growth factor requirements, while some species grow readily on simple defined media. It

Fig. 7 A summary diagram depicting the different clades of the order *Neisseriales* that are distinguished based upon different identified CSIs. Only those genera from whom genome sequences are available are listed here

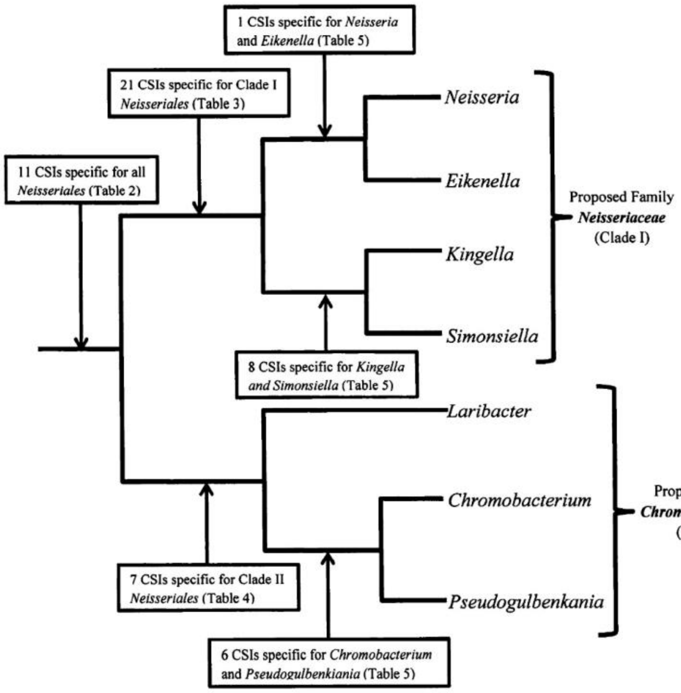

1 CSIs specific for *Neisseria* and *Eikenella* (Table 5)
21 CSIs specific for Clade I *Neisseriales* (Table 3)
11 CSIs specific for all *Neisseriales* (Table 2)
Neisseria
Eikenella
Kingella
Simonsiella
Proposed Family *Neisseriaceae* (Clade I)
8 CSIs specific for *Kingella* and *Simonsiella* (Table 5)
Laribacter
Chromobacterium
Pseudogulbenkania
Proposed Family *Chromobacteriaceae* (Clade II)
7 CSIs specific for Clade II *Neisseriales* (Table 4)
6 CSIs specific for *Chromobacterium* and *Pseudogulbenkiania* (Table 5)

Table 6 Phenotypic characteristics of the genera within the order *Neisseriales*

Genus	Proposed family	Cell morphology	GC%	Motility	Habitat	Reference
Alysiella	*Neisseriaceae*	Rod	44–48	Gliding motility	H	Xie and Yokota (2005)
Bergeriella	*Neisseriaceae*	Cocci	56	Non-motile	H	Xie and Yokota (2005)
Conchiformibius	*Neisseriaceae*	Rod	50–55	Gliding motility	H	Xie and Yokota (2005)
Eikenella	*Neisseriaceae*	Rod	56–58	Non-motile	H	Jackson and Goodman (1972)
Kingella	*Neisseriaceae*	Rod	47–58	Twitching motility[a]	H	Dewhirst et al. (1993)
Morococcus	*Neisseriaceae*	Cocci	52	Non-motile	H	Long et al. (1981)
Neisseria	*Neisseriaceae*	Primarily cocci	48–56	Non-motile	H	Tønjum (2005a)
Simonsiella	*Neisseriaceae*	Rod	41–55	Gliding motility	H	Dewhirst et al. (1989)
Stenoxybacter	*Neisseriaceae*	Rod	54	Non-motile	H	Wertz and Breznak (2007)
Uruburuella	*Neisseriaceae*	Coccobacilli	55	Non-motile	H	Vela et al. (2005)
Vitreoscilla	*Neisseriaceae*	Rod	42–63	Gliding motility	M	Strohl et al. (1986)
Amantichitinum	*Chromobacteriaceae*	Rod	61.5	Motile (flagella)	T	Moß et al. (2012)
Andreprevotia	*Chromobacteriaceae*	Rod	62–63	Motile (flagella)	T	Weon et al. (2007)
Aquaspirillum	*Chromobacteriaceae*	Spirilla	49–66	Motile (flagella)	A	Kumar et al. (1974)
Aquitalea	*Chromobacteriaceae*	Rod	59	Motile (flagella)	A	Lau et al. (2006)
Chitinibacter	*Chromobacteriaceae*	Rod	56–58	Motile (flagella)	T	Chern et al. (2004)
Chitinilyticum	*Chromobacteriaceae*	Rod	62–70	Motile (flagella)	A	Chang et al. (2007)
Chitiniphilus	*Chromobacteriaceae*	Rod	68	Motile (flagella)	A	Sato et al. (2009)
Chromobacterium	*Chromobacteriaceae*	Rod	65–68	Motile (flagella)	M	Gillis and Logan (2005)
Deefgea	*Chromobacteriaceae*	Rod	49–54	Motile (flagella)	A	Stackebrandt et al. (2007)
Formivibrio	*Chromobacteriaceae*	Rod	59–61	Motile (flagella)	A	Tanaka et al. (1991)
Gulbenkiania	*Chromobacteriaceae*	Rod	63	Motile (flagella)	A	Vaz-Moreira et al. (2007)
Iodobacter	*Chromobacteriaceae*	Rod	50–52	Motile (flagella)	A	Logan and Logan (1989)
Jeongeupia	*Chromobacteriaceae*	Rod	64	Motile (flagella)	T	Yoon et al. (2010)
Laribacter	*Chromobacteriaceae*	Rod	68	Motile (flagella)	M	Yuen et al. (2001)
Leeia	*Chromobacteriaceae*	Rod	56	Motile (flagella)	T	Lim et al. (2007)
Microvirgula	*Chromobacteriaceae*	Rod	65	Motile (flagella)	A	Patureau et al. (1998)
Paludibacterium	*Chromobacteriaceae*	Rod	63	Motile (flagella)	T	Kwon et al. (2008)
Prolinoborus	*Chromobacteriaceae*	Rod	62–65	Motile (flagella)	A	Pot et al. (1992)
Pseudogulbenkiania	*Chromobacteriaceae*	Rod	63	Motile (flagella)	A	Lin et al. (2008)
Silvimonas	*Chromobacteriaceae*	Rod	58–60	Motile (flagella)	T	Yang et al. (2005)
Vogesella	*Chromobacteriaceae*	Rod	65–69	Motile (flagella)	T	Grimes et al. (1997)

A aquatic, *T* terrestrial, *H* host-associated, *M* multiple

should be noted that, while *Morococcus* is currently placed in this family, the only isolated *Morococcus* strain is likely a member of the genus *Neisseria*, thus the placement of genus *Morococcus* is uncertain at present. The mol% G+C of the DNA is 41–56.

Organisms from this order are distinguished from all other *Neisseriales* by the conserved signature indels described in this report in the following proteins: 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (IspH), amidophosphoribosyltransferase (PurF), anthranilate phosphoribosyltransferase (TrpD), deoxyuridine 5′-triphosphate nucleotidohydrolase (Dut), dihydroorotase (PryC), DNA polymerase III subunit alpha (DnaE), fructose-1,6-bisphosphatase (Fbp), GMP synthase (GuaA), histidine–tRNA Ligase (HisS), hypothetical protein EIKCOROL_00874, hypothetical protein EIKCOROL_00974, hypothetical protein NEIFLAOT_00147, hypothetical protein NEIFLAOT_01683, hypothetical protein NG_O0349, malate dehydrogenase (Mdh), methionine–tRNA ligase (MetG), ribosomal RNA large subunit methyltransferase J (RrmJ), S-adenosylmethionine:tRNA ribosyltransferase-isomerase (QueA), succinyldiaminopimelate transaminase (ArgD), and a YhgF-like protein (TexN).

Description of *Chromobacteriaceae* fam. nov.

Chromobacteriaceae (Chro.mo.bac.teri.a'ce.ae. M.L. neut. n. *Chromobacterium* type genus of the family; -aceae ending to denote a family; M.L. fem. pl. n. *Chromobacteriaceae* the *Chromobacterium* family)

The genera that are part of this family include *Andreprevotia*, *Aquaspirillum*, *Aquitalea*, *Chitinibacter*, *Chitinilyticum*, *Chitiniphilus*, *Chromobacterium*, *Deefgea*, *Formivibrio*, *Gulbenkiania*, *Iodobacter*, *Jeongeupia*, *Laribacter*, *Leeia*, *Microvirgula*, *Paludibacterium*, (*Prolinoborus*), *Pseudogulbenkiania*, *Silvimonas*, and *Vogesella*. Of these, *Chromobacterium* is the type genus of this family (Bergonzini 1881). Cells are rod-shaped, occurring singly, in pairs, or in short chains; one genus, *Aquaspirillum*, is an exception and contains cells with helical morphology. Flagella and swimming motility are present in all genera with the exception of *Formivibrio* which exhibit surface-bound motility ("twitching motility"). Colonies of most genera are not pigmented except those of *Chromobacterium* and strains of *Vogesella*. All species are capable of free living; however some species may be facultative pathogens. Several species have complex growth factor requirements, while some species grow readily on simple defined media. It should be noted that the 16S rRNA sequence for the only isolated species of *Prolinoborus* branches within the class *Gammaproteobacteria* with the genus *Acinetobacter* and is unlikely to remain a member of this family. The mol% G+C of the DNA is 49–70.

Organisms from this family are distinguished by the presence of the CSIs that are specific for the order *Neisseriales* and the absence of the CSIs that are specific for the family *Neisseriaceae*. In addition, conserved signature indels described in this report in the following proteins may be present in some or all of the species from this family. Glycine cleavage system aminomethyltransferase T (GcvT), hypothetical protein CV_3451, methylmalonate-semialdehyde dehydrogenase (IolA), ribonucleoside-diphosphate reductase subunit alpha (NrdE), propionyl-CoA carboxylase subunit beta (PccB), succinyl-CoA synthetase subunit alpha (SucC), and tetraacyldisaccharide 4′-kinase (LpxK).

References

Bennett JS, Bentley SD, Vernikos GS, Quail MA, Cherevach I, White B, Parkhill J, Maiden MCJ (2010) Independent evolution of the core and accessory gene sets in the genus *Neisseria*: insights gained from the genome of Neisseria lactamica isolate 020–06. BMC Genomics 11:652

Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, Maiden MCJ (2012) A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. Microbiology 158:1570–1580

Bentley SD, Vernikos GS, Snyder LAS, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K (2007) Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet 3:e23

Bergonzini C (1881) Sopra un nuovo bacterio colorato. Annuar Soc Nat Modena 2:149–158

Bhandari V, Ahmod NZ, Shah HN, and Gupta RS (2013). Molecular signatures for the *Bacillus* species: demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. Int J Syst Evol Microbiol. doi:10.1099/ijs.0.051805-0

Bøvre K (1984) Family VIII *Neisseriaceae* Prévot 1933; 119. In: Holt JG (ed) Bergey's manual of systematic bacteriology. Springer, Berlin, pp 288–296

Brenner DJ, Krieg NR, Garrity GM, Staley JT (2005) Bergey's manual of systematic bacteriology: the proteobacteria. Springer, New York

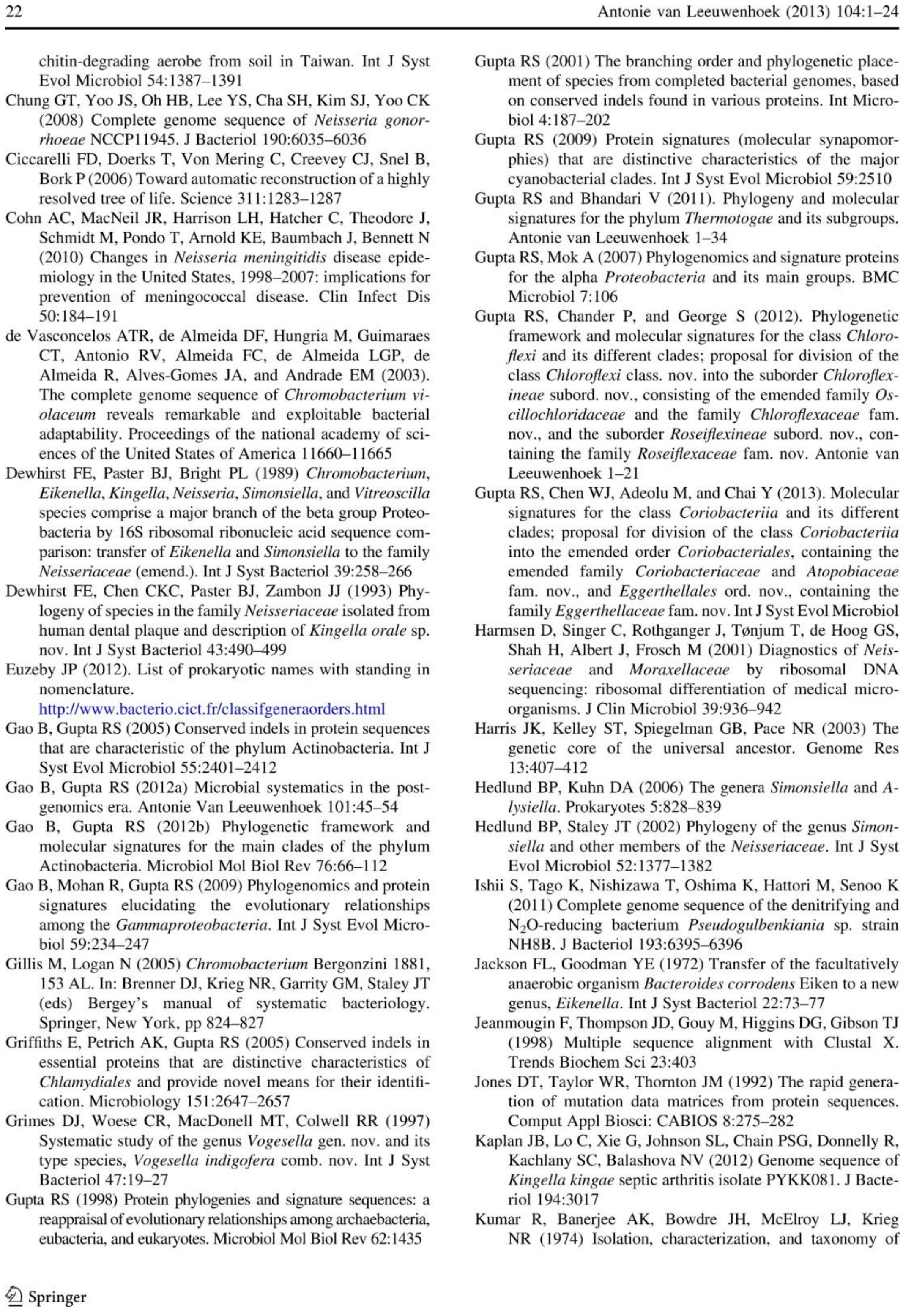
Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci 108:4494–4499

Byrne-Bailey KG, Weber KA, Coates JD (2012) Draft genome sequence of the anaerobic, nitrate-dependent, Fe(II)-oxidizing bacterium *Pseudogulbenkiania ferrooxidans* strain 2002. J Bacteriol 194:2400–2401

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

Chang SC, Chen WM, Wang JT, Wu MC (2007) *Chitinilyticum aquatile* gen. nov., sp. nov., a chitinolytic bacterium isolated from a freshwater pond used for Pacific white shrimp culture. Int J Syst Evol Microbiol 57:2854–2860

Chern LL, Stackebrandt E, Lee SF, Lee FL, Chen JK, Fu HM (2004) *Chitinibacter tainanensis* gen. nov., sp. nov., a



chitin-degrading aerobe from soil in Taiwan. Int J Syst Evol Microbiol 54:1387–1391

Chung GT, Yoo JS, Oh HB, Lee YS, Cha SH, Kim SJ, Yoo CK (2008) Complete genome sequence of *Neisseria gonorrhoeae* NCCP11945. J Bacteriol 190:6035–6036

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

Cohn AC, MacNeil JR, Harrison LH, Hatcher C, Theodore J, Schmidt M, Pondo T, Arnold KE, Baumbach J, Bennett N (2010) Changes in *Neisseria meningitidis* disease epidemiology in the United States, 1998–2007: implications for prevention of meningococcal disease. Clin Infect Dis 50:184–191

de Vasconcelos ATR, de Almeida DF, Hungria M, Guimaraes CT, Antonio RV, Almeida FC, de Almeida LGP, de Almeida R, Alves-Gomes JA, and Andrade EM (2003). The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. Proceedings of the national academy of sciences of the United States of America 11660–11665

Dewhirst FE, Paster BJ, Bright PL (1989) *Chromobacterium*, *Eikenella*, *Kingella*, *Neisseria*, *Simonsiella*, and *Vitreoscilla* species comprise a major branch of the beta group Proteobacteria by 16S ribosomal ribonucleic acid sequence comparison: transfer of *Eikenella* and *Simonsiella* to the family *Neisseriaceae* (emend.). Int J Syst Bacteriol 39:258–266

Dewhirst FE, Chen CKC, Paster BJ, Zambon JJ (1993) Phylogeny of species in the family *Neisseriaceae* isolated from human dental plaque and description of *Kingella orale* sp. nov. Int J Syst Bacteriol 43:490–499

Euzeby JP (2012). List of prokaryotic names with standing in nomenclature.
http://www.bacterio.cict.fr/classifgeneraorders.html

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, Gupta RS (2012a) Microbial systematics in the post-genomics era. Antonie Van Leeuwenhoek 101:45–54

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

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

Gillis M, Logan N (2005) *Chromobacterium* Bergonzini 1881, 153 AL. In: Brenner DJ, Krieg NR, Garrity GM, Staley JT (eds) Bergey's manual of systematic bacteriology. Springer, New York, pp 824–827

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

Grimes DJ, Woese CR, MacDonell MT, Colwell RR (1997) Systematic study of the genus *Vogesella* gen. nov. and its type species, *Vogesella indigofera* comb. nov. Int J Syst Bacteriol 47:19–27

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

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

Gupta RS and Bhandari V (2011). Phylogeny and molecular signatures for the phylum *Thermotogae* and its subgroups. Antonie van Leeuwenhoek 1–34

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

Gupta RS, Chander P, and George S (2012). Phylogenetic framework and molecular signatures for the class *Chloroflexi* and its different clades; proposal for division of the class *Chloroflexi* class. nov. into the suborder *Chloroflexineae* subord. nov., consisting of the emended family *Oscillochloridaceae* and the family *Chloroflexaceae* fam. nov., and the suborder *Roseiflexineae* subord. nov., containing the family *Roseiflexaceae* fam. nov. Antonie van Leeuwenhoek 1–21

Gupta RS, Chen WJ, Adeolu M, and Chai Y (2013). Molecular signatures for the class *Coriobacteriia* and its different clades; proposal for division of the class *Coriobacteriia* into the emended order *Coriobacteriales*, containing the emended family *Coriobacteriaceae* and *Atopobiaceae* fam. nov., and *Eggerthellales* ord. nov., containing the family *Eggerthellaceae* fam. nov. Int J Syst Evol Microbiol

Harmsen D, Singer C, Rothganger J, Tønjum T, de Hoog GS, Shah H, Albert J, Frosch M (2001) Diagnostics of *Neisseriaceae* and *Moraxellaceae* by ribosomal DNA sequencing: ribosomal differentiation of medical microorganisms. J Clin Microbiol 39:936–942

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

Hedlund BP, Kuhn DA (2006) The genera *Simonsiella* and *Alysiella*. Prokaryotes 5:828–839

Hedlund BP, Staley JT (2002) Phylogeny of the genus *Simonsiella* and other members of the *Neisseriaceae*. Int J Syst Evol Microbiol 52:1377–1382

Ishii S, Tago K, Nishizawa T, Oshima K, Hattori M, Senoo K (2011) Complete genome sequence of the denitrifying and N_2O-reducing bacterium *Pseudogulbenkiania* sp. strain NH8B. J Bacteriol 193:6395–6396

Jackson FL, Goodman YE (1972) Transfer of the facultatively anaerobic organism *Bacteroides corrodens* Eiken to a new genus, *Eikenella*. Int J Syst Bacteriol 22:73–77

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

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci: CABIOS 8:275–282

Kaplan JB, Lo C, Xie G, Johnson SL, Chain PSG, Donnelly R, Kachlany SC, Balashova NV (2012) Genome sequence of *Kingella kingae* septic arthritis isolate PYKK081. J Bacteriol 194:3017

Kumar R, Banerjee AK, Bowdre JH, McElroy LJ, Krieg NR (1974) Isolation, characterization, and taxonomy of

Aquaspirillum bengal sp. nov. Int J Syst Bacteriol 24: 453–458

Kwon SW, Kim BY, Kim WG, Yoo KH, Yoo SH, Son JA, Weon HY (2008) *Paludibacterium yongneupense* gen. nov., sp. nov., isolated from a wetland, Yongneup, in Korea. Int J Syst Evol Microbiol 58:190–194

Lau HT, Faryna J, Triplett EW (2006) *Aquitalea magnusonii* gen. nov., sp. nov., a novel Gram-negative bacterium isolated from a humic lake. Int J Syst Evol Microbiol 56:867–871

Lim JM, Jeon CO, Lee GS, Park DJ, Kang UG, Park CY, Kim CJ (2007) *Leeia oryzae* gen. nov., sp. nov., isolated from a rice field in Korea. Int J Syst Evol Microbiol 57:1204–1208

Lin MC, Chou JH, Arun AB, Young CC, Chen WM (2008) *Pseudogulbenkiania subflava* gen. nov., sp. nov., isolated from a cold spring. Int J Syst Evol Microbiol 58:2384–2388

Logan NA, Logan NA (1989) Numerical taxonomy of violet-pigmented, gram-negative bacteria and description of *Iodobacter fluviatile* gen. nov., comb. nov. Int J Syst Bacteriol 39:450–456

Long PA, Sly LI, Pham AV, Davis GHG (1981) Characterization of *Morococcus cerebrosus* gen. nov., sp. nov. and comparison with *Neisseria mucosa*. Int J Syst Bacteriol 31:294–301

McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26

Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci 93:2873–2878

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586

Moß KS, Hartmann SC, Müller I, Fritz C, Krügener S, Zibek S, Hirth T, Rupp S (2012) *Amantichitinum ursilacus* gen. nov., sp. nov., a chitin-degrading bacterium found at the Bärensee, Stuttgart, Germany. Int J Syst Evol Microbiol 63:98–103

NCBI (2012) NCBI genome database. http://www.ncbi.nlm.nih.gov/genome/

Patureau D, Godon JJ, Dabert P, Bouchez T, Bernet N, Delgenes JP, Moletta R (1998) *Microvirgula aerodenitrificans* gen. nov., sp. nov., a new Gram-negative bacterium exhibiting co-respiration of oxygen and nitrogen oxides up to oxygen-saturated conditions. Int J Syst Bacteriol 48:775–782

Pot B, Willems A, Gillis M, De Ley J (1992) Intra-and intergeneric relationships of the genus *Aquaspirillum*: *prolinoborus*, a new genus for *Aquaspirillum fasciculus*, with the species *Prolinoborus fasciculus* comb. nov. Int J Syst Bacteriol 42:44–57

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596

Rokas A, Holland PWH (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

Rusniok C, Vallenet D, Floquet S, Ewles H, Mouze-Soulama C, Brown D, Lajus A, Buchrieser C, Medigue C, Glaser P (2009) NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen *Neisseria meningitidis*. Genome Biol 10:R110

Sato K, Kato Y, Taguchi G, Nogawa M, Yokota A, Shimosaka M (2009) *Chitiniphilus shinanonensis* gen. nov., sp. nov., a novel chitin-degrading bacterium belonging to *Betaproteobacteria*. J Gen Appl Microbiol 55:147–153

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

Sly LI (2005). Genus incertae sedis XV. *Morococcus* Long, Sly, Pham and Davis 1981, 300^{VP}. In: Brenner DJ, Krieg NR, Garrity GM, and Staley JT (eds) Bergey's manual of systematic bacteriology. Springer, New York, pp 861–863

Stackebrandt E, Lang E, Cousin S, Päuker O, Brambilla E, Kroppenstedt R, Lünsdorf H (2007) *Deefgea rivuli* gen. nov., sp. nov., a member of the class *Betaproteobacteria*. Int J Syst Evol Microbiol 57:639–645

Stephens DS, Greenwood B, Brandtzaeg P (2007) Epidemic meningitis, meningococcaemia, and *Neisseria meningitidis*. Lancet 369:2196–2210

Strohl WR, Schmidt TM, Lawry NH, Mezzino MJ, Larkin JM (1986) Characterization of *Vitreoscilla beggiatoides* and *Vitreoscilla filiformis* sp. nov., nom. rev., and comparison with *Vitreoscilla stercoraria* and *Beggiatoa alba*. Int J Syst Bacteriol 36:302–313

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

Tanaka K, Nakamura K, Mikami E (1991) Fermentation of S-citramalate, citrate, mesaconate, and pyruvate by a gram-negative strictly anaerobic non-spore-former, *Formivibrio citricus* gen. nov., sp. nov. Arch Microbiol 155:491–495

Tanner A, Maiden MF, Paster BJ, Dewhirst FE (2007) The impact of 16S ribosomal RNA-based phylogeny on the taxonomy of oral bacteria. Periodontology 2000(5):26–51

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

Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 287:1809–1815

Tønjum T (2005a). Genus I. *Neisseria* Trevisan 1885, 105^{AL}. In: Brenner DJ, Krieg NR, Garrity GM and Staley JT (eds) Bergey's manual of systematic bacteriology. Springer, New York, pp 777–798

Tønjum T (2005b). Order IV. *Neisseriales* ord. nov. In: Brenner DJ, Krieg NR, Garrity GM and Staley JT (eds) Bergey's manual of systematic bacteriology. Springer, New York, p 774

Trevisan V (1885) Caratteri di alcuni nuovi generi di Batteriacee. Atti della Accademia Fisio-Medico-Statistica in Milano. Series 4:92–107

Vaz-Moreira I, Nobre MF, Nunes OC, Manaia CM (2007) *Gulbenkiania mobilis* gen. nov., sp. nov., isolated from treated municipal wastewater. Int J Syst Evol Microbiol 57:1108–1112

Vela AI, Collins MD, Lawson PA, García N, Domínguez L, Fernández-Garayzábal JF (2005) *Uruburuella suis* gen.

nov., sp. nov., isolated from clinical specimens of pigs. Int J Syst Evol Microbiol 55:643–647

Weon HY, Kim BY, Yoo SH, Joa JH, Kwon SW, Kim WG (2007) *Andreprevotia chitinilytica* gen. nov., sp. nov., isolated from forest soil from Halla Mountain, Jeju Island, Korea. Int J Syst Evol Microbiol 57:1572–1575

Wernegreen JJ (2011) Reduced selective constraint in endosymbionts: elevation in radical amino acid replacements occurs genome-wide. PLoS One 6:e28905

Wernegreen JJ, Moran NA (1999) Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. Mol Biol Evol 16:83–97

Wertz JT, Breznak JA (2007) *Stenoxybacter acetivorans* gen. nov., sp. nov., an acetate-oxidizing obligate microaerophile among diverse O_2-consuming bacteria from termite guts. Appl Environ Microbiol 73:6819–6828

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

Woo PCY, Lau SKP, Tse H, Teng JLL, Curreem SOT, Tsang AKL, Fan RYY, Wong GKM, Huang Y, Loman NJ, Snyder LAS, Cai JJ, Huang JD, Mak W, Pallen MJ, Lok S, Yuen KY (2009) The complete genome and proteome of *Laribacter hongkongensis* reveal potential mechanisms for adaptations to different temperatures and habitats. PLoS Genet 5:e1000416

World Health Organization (2011). Prevalence and incidence of selected sexually transmitted infections, *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, syphilis and *Trichomonas vaginalis*: methods and results used by WHO to generate 2005 estimates. Geneva: the Organization; 2011. World Health Organization, Geneva. ISBN 9789241563598

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

Xie CH, Yokota A (2005) Phylogenetic analysis of *Alysiella* and related genera of *Neisseriaceae*: proposal of *Alysiella crassa* comb. nov., *Conchiformibium steedae* gen. nov., comb. nov., *Conchiformibium kuhniae* sp. nov. and *Bergeriella denitrificans* gen. nov., comb. nov. J Gene Appl Microbiol 51:1–10

Yang HC, Im WT, An DS, Park W, Kim IS, Lee ST (2005) *Silvimonas terrae* gen. nov., sp. nov., a novel chitin-degrading facultative anaerobe belonging to the *Betaproteobacteria*. Int J Syst Evol Microbiol 55:2329–2332

Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glöckner FO, Rosselló-Móra R (2008) The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 31:241–250

Yi H, Cho YJ, Yoon SH, Park SC, Chun J (2012) Comparative genomics of *Neisseria weaveri* clarifies the taxonomy of this species and identifies genetic determinants that may be associated with virulence. FEMS Microbiol Lett 328(2):100–105

Yoon JH, Choi JH, Kang SJ, Choi NS, Lee JS, Song JJ (2010) *Jeongeupia naejangsanensis* gen. nov., sp. nov., a cellulose-degrading bacterium isolated from forest soil from Naejang Mountain in Korea. Int J Syst Evol Microbiol 60:615–619

Yuen KY, Woo PCY, Teng JLL, Leung KW, Wong MKM, Lau SKP (2001) *Laribacter hongkongensis* gen. nov., sp. nov., a novel gram-negative bacterium isolated from a cirrhotic patient with bacteremia and empyema. J Clin Microbiol 39:4227–4232

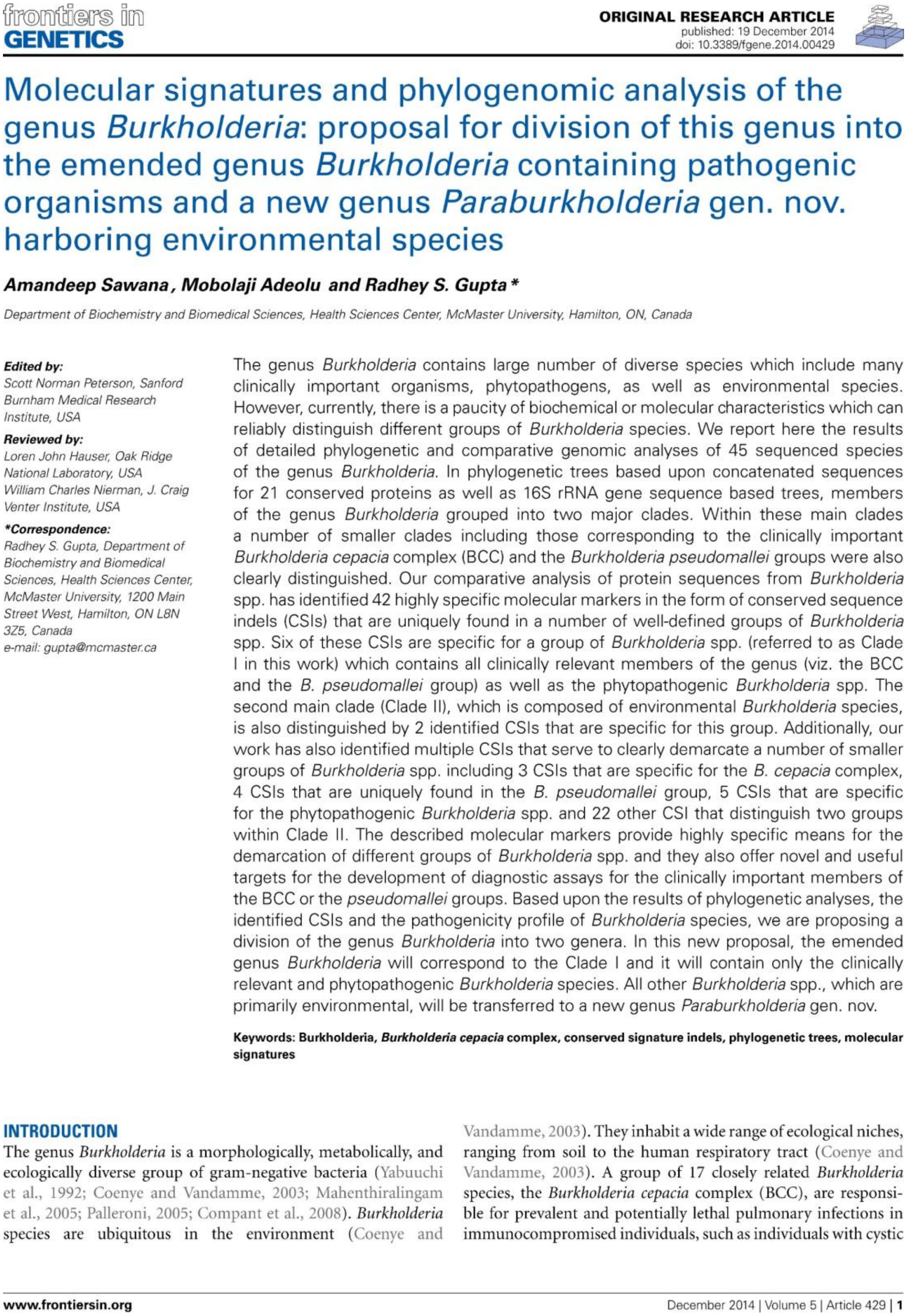
CHAPTER 5

Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring environmental species.

This chapter describes the use of molecular signatures (CSIs) and phylogenetic trees to differentiate the opportunistically pathogenic members of the genus *Burkholderia* from the plant-beneficial and environmental *Burkholderia*. The chapter also describes unique CSIs which distinguish the clinically relevant *Burkholderia cepacia* complex, the pathogenic *Burkholderia pseudomallei* group, or the phytopathogenic *Burkholderia* group, and includes a brief discussion of their diagnostic potential. The chapter concludes with a proposal to limit the genus *Burkholderia* to opportunistically pathogenic members of the genus, and to transfer the plant-beneficial and environmental *Burkholderia* to the novel genus *Paraburkholderia*. My contributions towards the completion of this chapter include the construction of the 16S rRNA based phylogenetic tree, the initial identification of some CSIs, the creation of the taxonomic proposals, the writing of drafts and revisions of the manuscript, and involvement in the production of main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials for this work are not included in the chapter but can be accessed along with the rest of the manuscript at:

Sawana, A., Adeolu, M., & Gupta, R. S. (2014). *Frontiers in genetics, 5*, 429.



frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 19 December 2014
doi: 10.3389/fgene.2014.00429

Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring environmental species

Amandeep Sawana, Mobolaji Adeolu and Radhey S. Gupta*

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

Edited by:
Scott Norman Peterson, Sanford Burnham Medical Research Institute, USA

Reviewed by:
Loren John Hauser, Oak Ridge National Laboratory, USA
William Charles Nierman, J. Craig Venter Institute, USA

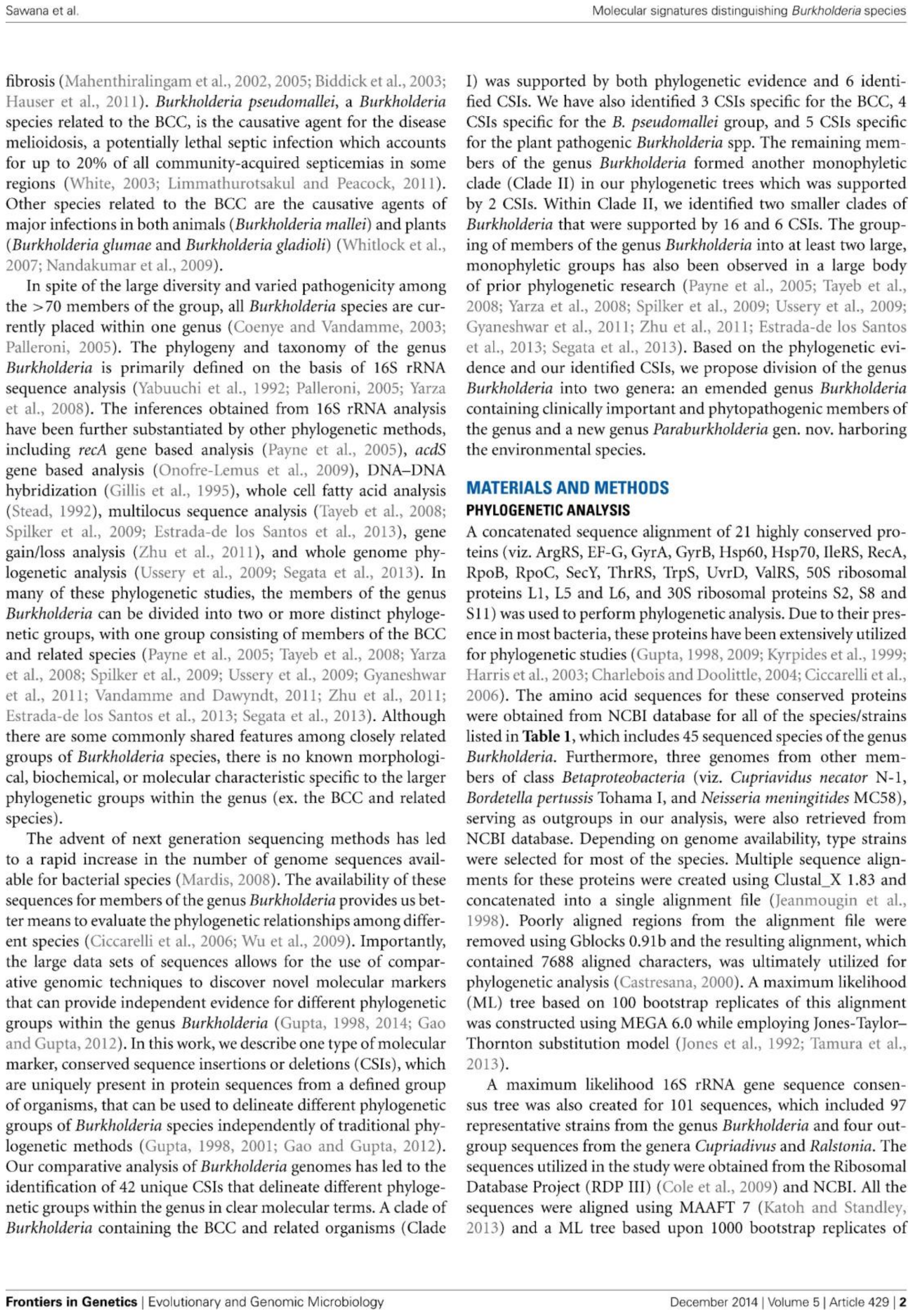
****Correspondence:***
Radhey S. Gupta, Department of Biochemistry and Biomedical Sciences, Health Sciences Center, McMaster University, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada
e-mail: gupta@mcmaster.ca

The genus *Burkholderia* contains large number of diverse species which include many clinically important organisms, phytopathogens, as well as environmental species. However, currently, there is a paucity of biochemical or molecular characteristics which can reliably distinguish different groups of *Burkholderia* species. We report here the results of detailed phylogenetic and comparative genomic analyses of 45 sequenced species of the genus *Burkholderia*. In phylogenetic trees based upon concatenated sequences for 21 conserved proteins as well as 16S rRNA gene sequence based trees, members of the genus *Burkholderia* grouped into two major clades. Within these main clades a number of smaller clades including those corresponding to the clinically important *Burkholderia cepacia* complex (BCC) and the *Burkholderia pseudomallei* groups were also clearly distinguished. Our comparative analysis of protein sequences from *Burkholderia* spp. has identified 42 highly specific molecular markers in the form of conserved sequence indels (CSIs) that are uniquely found in a number of well-defined groups of *Burkholderia* spp. Six of these CSIs are specific for a group of *Burkholderia* spp. (referred to as Clade I in this work) which contains all clinically relevant members of the genus (viz. the BCC and the *B. pseudomallei* group) as well as the phytopathogenic *Burkholderia* spp. The second main clade (Clade II), which is composed of environmental *Burkholderia* species, is also distinguished by 2 identified CSIs that are specific for this group. Additionally, our work has also identified multiple CSIs that serve to clearly demarcate a number of smaller groups of *Burkholderia* spp. including 3 CSIs that are specific for the *B. cepacia* complex, 4 CSIs that are uniquely found in the *B. pseudomallei* group, 5 CSIs that are specific for the phytopathogenic *Burkholderia* spp. and 22 other CSI that distinguish two groups within Clade II. The described molecular markers provide highly specific means for the demarcation of different groups of *Burkholderia* spp. and they also offer novel and useful targets for the development of diagnostic assays for the clinically important members of the BCC or the *pseudomallei* groups. Based upon the results of phylogenetic analyses, the identified CSIs and the pathogenicity profile of *Burkholderia* species, we are proposing a division of the genus *Burkholderia* into two genera. In this new proposal, the emended genus *Burkholderia* will correspond to the Clade I and it will contain only the clinically relevant and phytopathogenic *Burkholderia* species. All other *Burkholderia* spp., which are primarily environmental, will be transferred to a new genus *Paraburkholderia* gen. nov.

Keywords: Burkholderia, *Burkholderia cepacia* complex, conserved signature indels, phylogenetic trees, molecular signatures

INTRODUCTION

The genus *Burkholderia* is a morphologically, metabolically, and ecologically diverse group of gram-negative bacteria (Yabuuchi et al., 1992; Coenye and Vandamme, 2003; Mahenthiralingam et al., 2005; Palleroni, 2005; Compant et al., 2008). *Burkholderia* species are ubiquitous in the environment (Coenye and Vandamme, 2003). They inhabit a wide range of ecological niches, ranging from soil to the human respiratory tract (Coenye and Vandamme, 2003). A group of 17 closely related *Burkholderia* species, the *Burkholderia cepacia* complex (BCC), are responsible for prevalent and potentially lethal pulmonary infections in immunocompromised individuals, such as individuals with cystic



fibrosis (Mahenthiralingam et al., 2002, 2005; Biddick et al., 2003; Hauser et al., 2011). *Burkholderia pseudomallei*, a *Burkholderia* species related to the BCC, is the causative agent for the disease melioidosis, a potentially lethal septic infection which accounts for up to 20% of all community-acquired septicemias in some regions (White, 2003; Limmathurotsakul and Peacock, 2011). Other species related to the BCC are the causative agents of major infections in both animals (*Burkholderia mallei*) and plants (*Burkholderia glumae* and *Burkholderia gladioli*) (Whitlock et al., 2007; Nandakumar et al., 2009).

In spite of the large diversity and varied pathogenicity among the >70 members of the group, all *Burkholderia* species are currently placed within one genus (Coenye and Vandamme, 2003; Palleroni, 2005). The phylogeny and taxonomy of the genus *Burkholderia* is primarily defined on the basis of 16S rRNA sequence analysis (Yabuuchi et al., 1992; Palleroni, 2005; Yarza et al., 2008). The inferences obtained from 16S rRNA analysis have been further substantiated by other phylogenetic methods, including *recA* gene based analysis (Payne et al., 2005), *acdS* gene based analysis (Onofre-Lemus et al., 2009), DNA–DNA hybridization (Gillis et al., 1995), whole cell fatty acid analysis (Stead, 1992), multilocus sequence analysis (Tayeb et al., 2008; Spilker et al., 2009; Estrada-de los Santos et al., 2013), gene gain/loss analysis (Zhu et al., 2011), and whole genome phylogenetic analysis (Ussery et al., 2009; Segata et al., 2013). In many of these phylogenetic studies, the members of the genus *Burkholderia* can be divided into two or more distinct phylogenetic groups, with one group consisting of members of the BCC and related species (Payne et al., 2005; Tayeb et al., 2008; Yarza et al., 2008; Spilker et al., 2009; Ussery et al., 2009; Gyaneshwar et al., 2011; Vandamme and Dawyndt, 2011; Zhu et al., 2011; Estrada-de los Santos et al., 2013; Segata et al., 2013). Although there are some commonly shared features among closely related groups of *Burkholderia* species, there is no known morphological, biochemical, or molecular characteristic specific to the larger phylogenetic groups within the genus (ex. the BCC and related species).

The advent of next generation sequencing methods has led to a rapid increase in the number of genome sequences available for bacterial species (Mardis, 2008). The availability of these sequences for members of the genus *Burkholderia* provides us better means to evaluate the phylogenetic relationships among different species (Ciccarelli et al., 2006; Wu et al., 2009). Importantly, the large data sets of sequences allows for the use of comparative genomic techniques to discover novel molecular markers that can provide independent evidence for different phylogenetic groups within the genus *Burkholderia* (Gupta, 1998, 2014; Gao and Gupta, 2012). In this work, we describe one type of molecular marker, conserved sequence insertions or deletions (CSIs), which are uniquely present in protein sequences from a defined group of organisms, that can be used to delineate different phylogenetic groups of *Burkholderia* species independently of traditional phylogenetic methods (Gupta, 1998, 2001; Gao and Gupta, 2012). Our comparative analysis of *Burkholderia* genomes has led to the identification of 42 unique CSIs that delineate different phylogenetic groups within the genus in clear molecular terms. A clade of *Burkholderia* containing the BCC and related organisms (Clade I) was supported by both phylogenetic evidence and 6 identified CSIs. We have also identified 3 CSIs specific for the BCC, 4 CSIs specific for the *B. pseudomallei* group, and 5 CSIs specific for the plant pathogenic *Burkholderia* spp. The remaining members of the genus *Burkholderia* formed another monophyletic clade (Clade II) in our phylogenetic trees which was supported by 2 CSIs. Within Clade II, we identified two smaller clades of *Burkholderia* that were supported by 16 and 6 CSIs. The grouping of members of the genus *Burkholderia* into at least two large, monophyletic groups has also been observed in a large body of prior phylogenetic research (Payne et al., 2005; Tayeb et al., 2008; Yarza et al., 2008; Spilker et al., 2009; Ussery et al., 2009; Gyaneshwar et al., 2011; Zhu et al., 2011; Estrada-de los Santos et al., 2013; Segata et al., 2013). Based on the phylogenetic evidence and our identified CSIs, we propose division of the genus *Burkholderia* into two genera: an emended genus *Burkholderia* containing clinically important and phytopathogenic members of the genus and a new genus *Paraburkholderia* gen. nov. harboring the environmental species.

MATERIALS AND METHODS

PHYLOGENETIC ANALYSIS

A concatenated sequence alignment of 21 highly conserved proteins (viz. ArgRS, EF-G, GyrA, GyrB, Hsp60, Hsp70, IleRS, RecA, RpoB, RpoC, SecY, ThrRS, TrpS, UvrD, ValRS, 50S ribosomal proteins L1, L5 and L6, and 30S ribosomal proteins S2, S8 and S11) was used to perform phylogenetic analysis. Due to their presence in most bacteria, these proteins have been extensively utilized for phylogenetic studies (Gupta, 1998, 2009; Kyrpides et al., 1999; Harris et al., 2003; Charlebois and Doolittle, 2004; Ciccarelli et al., 2006). The amino acid sequences for these conserved proteins were obtained from NCBI database for all of the species/strains listed in **Table 1**, which includes 45 sequenced species of the genus *Burkholderia*. Furthermore, three genomes from other members of class *Betaproteobacteria* (viz. *Cupriavidus necator* N-1, *Bordetella pertussis* Tohama I, and *Neisseria meningitides* MC58), serving as outgroups in our analysis, were also retrieved from NCBI database. Depending on genome availability, type strains were selected for most of the species. Multiple sequence alignments for these proteins were created using Clustal_X 1.83 and concatenated into a single alignment file (Jeanmougin et al., 1998). Poorly aligned regions from the alignment file were removed using Gblocks 0.91b and the resulting alignment, which contained 7688 aligned characters, was ultimately utilized for phylogenetic analysis (Castresana, 2000). A maximum likelihood (ML) tree based on 100 bootstrap replicates of this alignment was constructed using MEGA 6.0 while employing Jones-Taylor–Thornton substitution model (Jones et al., 1992; Tamura et al., 2013).

A maximum likelihood 16S rRNA gene sequence consensus tree was also created for 101 sequences, which included 97 representative strains from the genus *Burkholderia* and four outgroup sequences from the genera *Cupriadivus* and *Ralstonia*. The sequences utilized in the study were obtained from the Ribosomal Database Project (RDP III) (Cole et al., 2009) and NCBI. All the sequences were aligned using MAAFT 7 (Katoh and Standley, 2013) and a ML tree based upon 1000 bootstrap replicates of

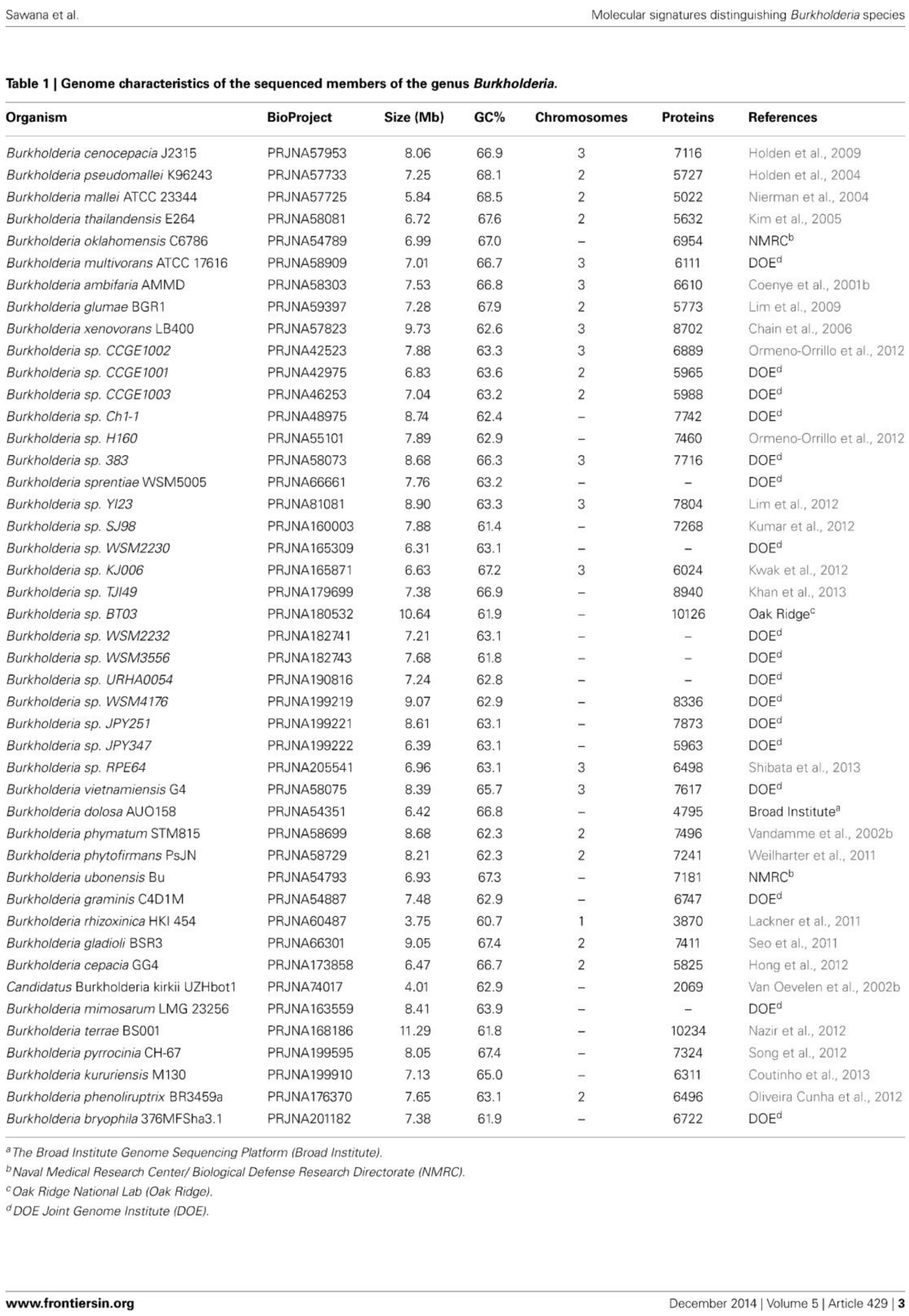


Table 1 | Genome characteristics of the sequenced members of the genus *Burkholderia*.

Organism	BioProject	Size (Mb)	GC%	Chromosomes	Proteins	References
Burkholderia cenocepacia J2315	PRJNA57953	8.06	66.9	3	7116	Holden et al., 2009
Burkholderia pseudomallei K96243	PRJNA57733	7.25	68.1	2	5727	Holden et al., 2004
Burkholderia mallei ATCC 23344	PRJNA57725	5.84	68.5	2	5022	Nierman et al., 2004
Burkholderia thailandensis E264	PRJNA58081	6.72	67.6	2	5632	Kim et al., 2005
Burkholderia oklahomensis C6786	PRJNA54789	6.99	67.0	–	6954	NMRC[b]
Burkholderia multivorans ATCC 17616	PRJNA58909	7.01	66.7	3	6111	DOE[d]
Burkholderia ambifaria AMMD	PRJNA58303	7.53	66.8	3	6610	Coenye et al., 2001b
Burkholderia glumae BGR1	PRJNA59397	7.28	67.9	2	5773	Lim et al., 2009
Burkholderia xenovorans LB400	PRJNA57823	9.73	62.6	3	8702	Chain et al., 2006
Burkholderia sp. CCGE1002	PRJNA42523	7.88	63.3	3	6889	Ormeno-Orrillo et al., 2012
Burkholderia sp. CCGE1001	PRJNA42975	6.83	63.6	2	5965	DOE[d]
Burkholderia sp. CCGE1003	PRJNA46253	7.04	63.2	2	5988	DOE[d]
Burkholderia sp. Ch1-1	PRJNA48975	8.74	62.4	–	7742	DOE[d]
Burkholderia sp. H160	PRJNA55101	7.89	62.9	–	7460	Ormeno-Orrillo et al., 2012
Burkholderia sp. 383	PRJNA58073	8.68	66.3	3	7716	DOE[d]
Burkholderia sprentiae WSM5005	PRJNA66661	7.76	63.2	–	–	DOE[d]
Burkholderia sp. YI23	PRJNA81081	8.90	63.3	3	7804	Lim et al., 2012
Burkholderia sp. SJ98	PRJNA160003	7.88	61.4	–	7268	Kumar et al., 2012
Burkholderia sp. WSM2230	PRJNA165309	6.31	63.1	–	–	DOE[d]
Burkholderia sp. KJ006	PRJNA165871	6.63	67.2	3	6024	Kwak et al., 2012
Burkholderia sp. TJI49	PRJNA179699	7.38	66.9	–	8940	Khan et al., 2013
Burkholderia sp. BT03	PRJNA180532	10.64	61.9	–	10126	Oak Ridge[c]
Burkholderia sp. WSM2232	PRJNA182741	7.21	63.1	–	–	DOE[d]
Burkholderia sp. WSM3556	PRJNA182743	7.68	61.8	–	–	DOE[d]
Burkholderia sp. URHA0054	PRJNA190816	7.24	62.8	–	–	DOE[d]
Burkholderia sp. WSM4176	PRJNA199219	9.07	62.9	–	8336	DOE[d]
Burkholderia sp. JPY251	PRJNA199221	8.61	63.1	–	7873	DOE[d]
Burkholderia sp. JPY347	PRJNA199222	6.39	63.1	–	5963	DOE[d]
Burkholderia sp. RPE64	PRJNA205541	6.96	63.1	3	6498	Shibata et al., 2013
Burkholderia vietnamiensis G4	PRJNA58075	8.39	65.7	3	7617	DOE[d]
Burkholderia dolosa AUO158	PRJNA54351	6.42	66.8	–	4795	Broad Institute[a]
Burkholderia phymatum STM815	PRJNA58699	8.68	62.3	2	7496	Vandamme et al., 2002b
Burkholderia phytofirmans PsJN	PRJNA58729	8.21	62.3	2	7241	Weilharter et al., 2011
Burkholderia ubonensis Bu	PRJNA54793	6.93	67.3	–	7181	NMRC[b]
Burkholderia graminis C4D1M	PRJNA54887	7.48	62.9	–	6747	DOE[d]
Burkholderia rhizoxinica HKI 454	PRJNA60487	3.75	60.7	1	3870	Lackner et al., 2011
Burkholderia gladioli BSR3	PRJNA66301	9.05	67.4	2	7411	Seo et al., 2011
Burkholderia cepacia GG4	PRJNA173858	6.47	66.7	2	5825	Hong et al., 2012
Candidatus Burkholderia kirkii UZHbot1	PRJNA74017	4.01	62.9	–	2069	Van Oevelen et al., 2002b
Burkholderia mimosarum LMG 23256	PRJNA163559	8.41	63.9	–	–	DOE[d]
Burkholderia terrae BS001	PRJNA168186	11.29	61.8	–	10234	Nazir et al., 2012
Burkholderia pyrrocinia CH-67	PRJNA199595	8.05	67.4	–	7324	Song et al., 2012
Burkholderia kururiensis M130	PRJNA199910	7.13	65.0	–	6311	Coutinho et al., 2013
Burkholderia phenoliruptrix BR3459a	PRJNA176370	7.65	63.1	2	6496	Oliveira Cunha et al., 2012
Burkholderia bryophila 376MFSha3.1	PRJNA201182	7.38	61.9	–	6722	DOE[d]

[a] *The Broad Institute Genome Sequencing Platform (Broad Institute).*

[b] *Naval Medical Research Center/ Biological Defense Research Directorate (NMRC).*

[c] *Oak Ridge National Lab (Oak Ridge).*

[d] *DOE Joint Genome Institute (DOE).*

this alignment was constructed using the General Time Reversible Model (Tavaré, 1986) in MEGA 6.0 (Tamura et al., 2013).

IDENTIFICATION OF MOLECULAR MARKERS (CSIs)

BLASTp searches were conducted for all proteins from chromosomes 2 and 3 (accession numbers NC_008061 and NC_008061) of *Burkholderia cenocepacia* J2315 (Holden et al., 2009) to identify CSIs that are shared by different members of the genus *Burkholderia*. Species that appeared as top hits with high scoring homologs (E values $< 1e^{-20}$) from the genus *Burkholderia* and other outgroups were selected. Multiple sequence alignments were created using the Clustal_X 1.83 (Jeanmougin et al., 1998). These alignments were visually inspected for the presence of insertions or deletions (indels) restricted to either some or all members of the genus *Burkholderia* and flanked by at least 5–6 conserved amino acid residues on both sides in the neighboring 30–40 amino acids. Indel queries that were not flanked by conserved regions were not further evaluated. The species specificity of the indel queries meeting the above criterion was further evaluated by performing BLASTp searches on short sequence segments containing the insertions or deletions, and their flanking conserved regions (60–100 amino acids long). The searches were conducted against the NCBI non-redundant (nr) database and a minimum of 250 BLAST hits were examined for the presence or absence of CSIs. The results of these analyses were evaluated as described in detail in our recent work (Gupta, 2014). Signature files for the CSIs that were specific for members of the genus *Burkholderia* were created and formatted using the programs SIG_CREATE and SIG_STYLE (accessible from Gleans.net) as described by Gupta (2014). The sequence alignment files presented here contain information for all detected insertions or deletions from the *Burkholderia* group of interest, but only a limited number from species that are serving as outgroups. Sequence information for different strains of various species is not shown, but they all exhibited similar pattern. Lastly, unless otherwise indicated, the CSIs shown here are specifically found in the indicated groups and similar CSIs were not detected in the 250 Blast hits with the query sequences.

RESULTS

BRANCHING PATTERN OF *BURKHOLDERIA* SPECIES IN CONCATENATED PROTEIN AND 16S rRNA TREES

Genome sequences of 45 species of *Burkholderia* were available from the NCBI genome database at the time of this work (NCBI, 2014). Some characteristics of these genomes are listed in **Table 1**. The genome sizes of the sequenced *Burkholderia* species show large variation (from 3.75–11.29 Mb) and the numbers of proteins in them also varied in a similar proportion. In this work we have produced a ML phylogenetic tree based on the concatenated amino acid sequences of 21 conserved housekeeping and ribosomal proteins obtained from 45 sequenced *Burkholderia* species (**Figure 1**). The *Burkholderia* species formed two large clades in the protein based ML tree: One consisting of the BCC and related organisms (Clade I) and another comprised mainly of environmental or poorly characterized *Burkholderia* species (Clade II). Within Clade I, three smaller, distinct clades are also observed. The first of these clades (Clade Ia) is wholly comprised of the sequenced BCC species, the second clade (Clade Ib) groups *B. pseudomallei* and closely related species, and the third clade (Clade Ic) consists of the plant pathogenic species, *B. glumae* and *B. gladioli*. Clade II could also be divided into two smaller clades, Clade IIa and Clade IIb. Clade IIa is separated from Clade IIb by a long branch, suggesting that a large amount of genetic divergence has occurred between the two groups. In addition to the two main clades of *Burkholderia*, two species, *Burkholderia* sp. JPY347 and *Burkholderia rhizoxinica*, branched early in the tree and did not associate with either Clade I or II.

We have also constructed a 16S rRNA based ML phylogenetic tree for 97 *Burkholderia* strains and candidate species (**Figure 2**). In this 16S rRNA based phylogenetic tree we observed broadly similar patterns to our protein based phylogeny. A clade consisting of the BCC and related organisms (Clade I) was clearly resolved. The three subclades within Clade I, the BCC (Clade Ia), the *B. pseudomallei* group (Clade Ib), and the plant pathogenic species (Clade Ic) were well resolved, though some species exhibited aberrant branching (ex. *B. oklahomensis* and *B. pseudomultivorans*). A large assemblage of the remaining *Burkholderia* species, roughly corresponding to Clade II in our concatenated protein based phylogenetic tree, was also observed in the 16S rRNA tree. However, due to significant number of unsequenced *Burkholderia* species which are present in the 16S rRNA database it is difficult to accurately identify the groups within Clade II of the 16S rRNA tree which correspond to Clades IIa and IIb in our concatenated protein based phylogenetic tree. Bootstrap support for branches in the 16S rRNA based tree were also significantly lower than they were in the concatenated protein tree indicating that some of the observed branching patterns may not be reliable. However, the clade consisting of the BCC and related organisms (Clade I) has strong bootstrap support and has been identified in a large number of previous 16S rRNA based phylogenetic studies (Yabuuchi et al., 1992; Palleroni, 2005; Yarza et al., 2008; Suarez-Moreno et al., 2012).

MOLECULAR SIGNATURES DISTINGUISHING THE CLADE I AND CLADE II *BURKHOLDERIA*

Rare genetic changes, such as insertions and deletions in essential genes/proteins, which occur in a common ancestor can be inherited by the various decedent species related to this common ancestor (Gupta, 1998; Rokas and Holland, 2000; Gogarten et al., 2002; Gupta and Griffiths, 2002). Due to the rarity and the specific presence of these rare genetic changes to a related group of organisms, they can serve as important molecular markers and provide a novel means to understand the evolutionary interrelationships between different closely related species (Gupta, 1998; Gupta and Griffiths, 2002; Gao and Gupta, 2012).

The comparative analysis of protein sequences from *Burkholderia* species that was carried out in the present work has identified a number of CSIs that serve to clearly distinguish a number of different clades within the genus *Burkholderia*. These studies have led to identification of 6 CSIs that are specific for the Clade I *Burkholderia*, consisting of the BCC and related organisms, enabling clear distinction of this group from all other *Burkholderia*. This clade, which contains all well characterized pathogens within the genus, represents the most clinically

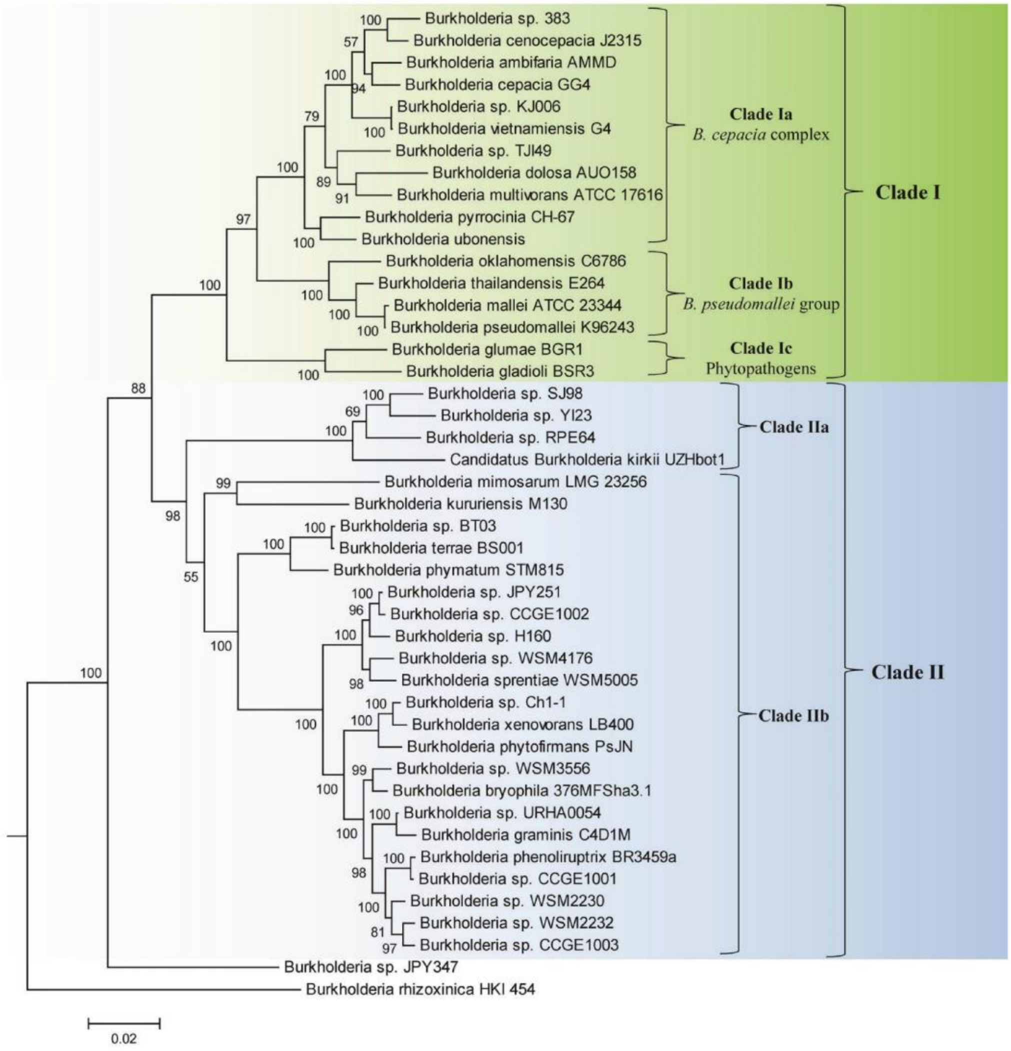

Burkholderia sp. 383
Burkholderia cenocepacia J2315
Burkholderia ambifaria AMMD
Burkholderia cepacia GG4
Burkholderia sp. KJ006
Burkholderia vietnamiensis G4
Burkholderia sp. TJI49
Burkholderia dolosa AUO158
Burkholderia multivorans ATCC 17616
Burkholderia pyrrocinia CH-67
Burkholderia ubonensis
Burkholderia oklahomensis C6786
Burkholderia thailandensis E264
Burkholderia mallei ATCC 23344
Burkholderia pseudomallei K96243
Burkholderia glumae BGR1
Burkholderia gladioli BSR3
Burkholderia sp. SJ98
Burkholderia sp. YI23
Burkholderia sp. RPE64
Candidatus Burkholderia kirkii UZHbot1
Burkholderia mimosarum LMG 23256
Burkholderia kururiensis M130
Burkholderia sp. BT03
Burkholderia terrae BS001
Burkholderia phymatum STM815
Burkholderia sp. JPY251
Burkholderia sp. CCGE1002
Burkholderia sp. H160
Burkholderia sp. WSM4176
Burkholderia sprentiae WSM5005
Burkholderia sp. Ch1-1
Burkholderia xenovorans LB400
Burkholderia phytofirmans PsJN
Burkholderia sp. WSM3556
Burkholderia bryophila 376MFSha3.1
Burkholderia sp. URHA0054
Burkholderia graminis C4D1M
Burkholderia phenoliruptrix BR3459a
Burkholderia sp. CCGE1001
Burkholderia sp. WSM2230
Burkholderia sp. WSM2232
Burkholderia sp. CCGE1003
Burkholderia sp. JPY347
Burkholderia rhizoxinica HKI 454
Clade Ia
B. cepacia complex
Clade Ib
B. pseudomallei group
Clade Ic
Phytopathogens
Clade I
Clade IIa
Clade IIb
Clade II
100
57
100
94
79
100
100
89
91
97
100
100
100
100
100
100
88
100
69
100
99
98
100
100
55
100
96
100
100
98
100
100
100
99
100
100
100
100
98
100
81
97
100
0.02

FIGURE 1 | A maximum likelihood phylogenetic tree of the genome sequenced members of the genus *Burkholderia* based upon concatenated sequences of 21 conserved proteins. The tree was rooted using *Cupriavidus necator* N-1, *Bordetella pertussis* Tohama I, and *Neisseria meningitides* MC58. Bootstrap analysis scores are indicated for each node. The major *Burkholderia* clades (Clades I and II) and their main sub-clades are indicated by brackets.

relevant group within the *Burkholderia*. All species within this clade are potentially pathogenic to human, animals, or plants and most have been isolated from clinical human samples (Simpson et al., 1994; Mahenthiralingam et al., 2002, 2005; Biddick et al., 2003; O'Carroll et al., 2003). One example of a CSI that is specific to the Clade I *Burkholderia* is shown in **Figure 3A**. In this case, a one amino acid deletion is present in a highly conserved region of a periplasmic amino acid-binding protein. The indel is flanked on both sides by highly conserved regions indicating that it is not the result of alignment artifacts and that it is a reliable genetic characteristic. This CSI is present in all of the sequenced members of the Clade I *Burkholderia*, but absent in all other bacterial homologs of this protein. Our work has identified 5 additional CSIs in other widely distributed proteins that are

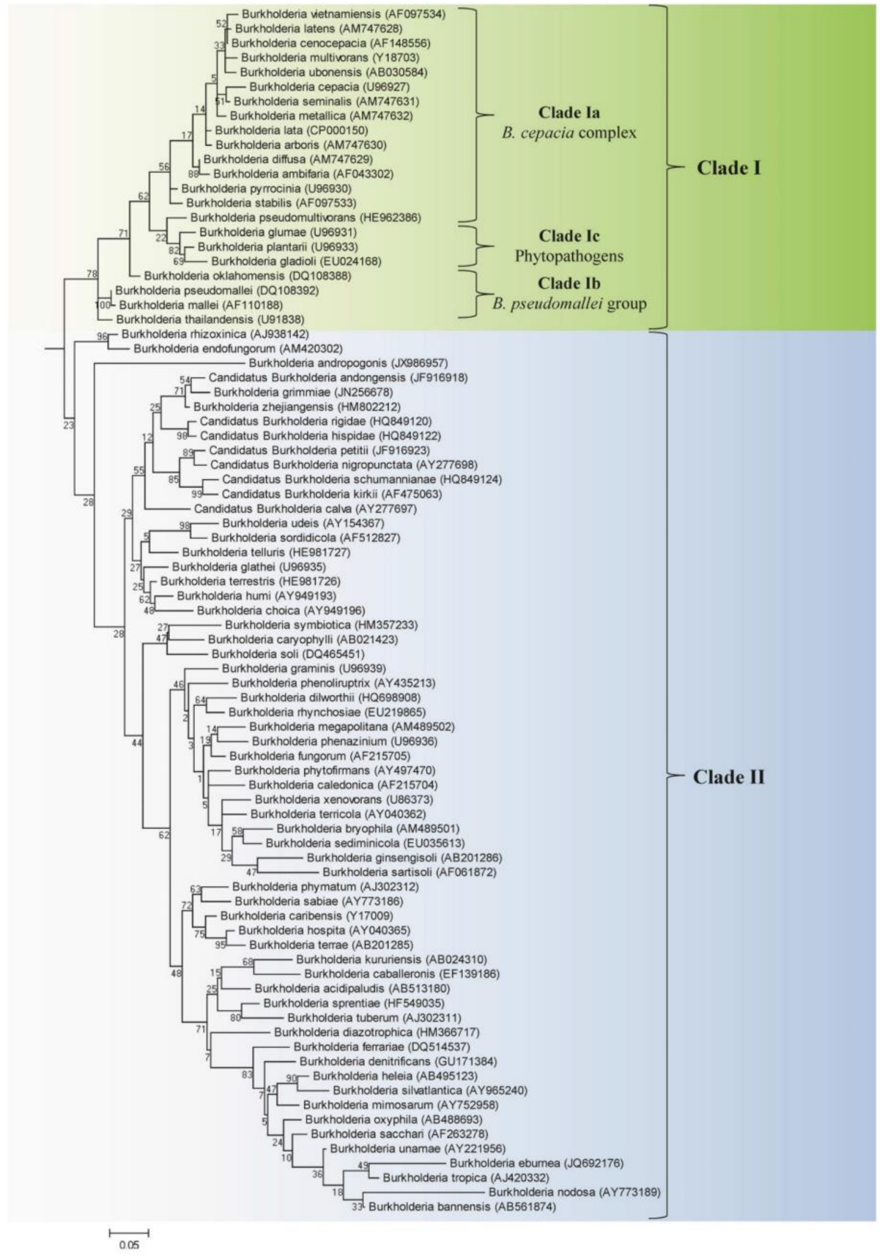

Burkholderia vietnamiensis (AF097534)
52 Burkholderia latens (AM747628)
33 Burkholderia cenocepacia (AF148556)
Burkholderia multivorans (Y18703)
Burkholderia ubonensis (AB030584)
5 Burkholderia cepacia (U96927)
51 Burkholderia seminalis (AM747631)
14 Burkholderia metallica (AM747632)
Burkholderia lata (CP000150)
17 Burkholderia arboris (AM747630)
Burkholderia diffusa (AM747629)
56 88 Burkholderia ambifaria (AF043302)
Burkholderia pyrrocinia (U96930)
62 Burkholderia stabilis (AF097533)
Burkholderia pseudomultivorans (HE962386)
71 22 Burkholderia glumae (U96931)
82 Burkholderia plantarii (U96933)
69 Burkholderia gladioli (EU024168)
78 Burkholderia oklahomensis (DQ108388)
Burkholderia pseudomallei (DQ108392)
100 Burkholderia mallei (AF110188)
Burkholderia thailandensis (U91838)
96 Burkholderia rhizoxinica (AJ938142)
Burkholderia endofungorum (AM420302)
Burkholderia andropogonis (JX986957)
54 Candidatus Burkholderia andongensis (JF916918)
71 Burkholderia grimmiae (JN256678)
25 Burkholderia zhejiangensis (HM802212)
Candidatus Burkholderia rigidae (HQ849120)
23 12 98 Candidatus Burkholderia hispidae (HQ849122)
89 Candidatus Burkholderia petitii (JF916923)
Candidatus Burkholderia nigropunctata (AY277698)
55 85 Candidatus Burkholderia schumannianae (HQ849124)
99 Candidatus Burkholderia kirkii (AF475063)
28 Candidatus Burkholderia calva (AY277697)
29 98 Burkholderia udeis (AY154367)
5 Burkholderia sordidicola (AF512827)
Burkholderia telluris (HE981727)
27 Burkholderia glathei (U96935)
25 Burkholderia terrestris (HE981726)
62 Burkholderia humi (AY949193)
48 Burkholderia choica (AY949196)
27 Burkholderia symbiotica (HM357233)
28 47 Burkholderia caryophylli (AB021423)
Burkholderia soli (DQ465451)
Burkholderia graminis (U96939)
46 Burkholderia phenoliruptrix (AY435213)
64 Burkholderia dilworthii (HQ698908)
2 Burkholderia rhynchosiae (EU219865)
14 Burkholderia megapolitana (AM489502)
3 19 Burkholderia phenazinium (U96936)
44 Burkholderia fungorum (AF215705)
Burkholderia phytofirmans (AY497470)
1 Burkholderia caledonica (AF215704)
Burkholderia xenovorans (U86373)
5 Burkholderia terricola (AY040362)
17 58 Burkholderia bryophila (AM489501)
62 Burkholderia sediminicola (EU035613)
29 Burkholderia ginsengisoli (AB201286)
47 Burkholderia sartisoli (AF061872)
63 Burkholderia phymatum (AJ302312)
Burkholderia sabiae (AY773186)
72 Burkholderia caribensis (Y17009)
75 Burkholderia hospita (AY040365)
95 Burkholderia terrae (AB201285)
68 Burkholderia kururiensis (AB024310)
48 15 Burkholderia caballeronis (EF139186)
25 Burkholderia acidipaludis (AB513180)
Burkholderia sprentiae (HF549035)
80 Burkholderia tuberum (AJ302311)
71 Burkholderia diazotrophica (HM366717)
Burkholderia ferrariae (DQ514537)
7 Burkholderia denitrificans (GU171384)
83 90 Burkholderia heleia (AB495123)
7 47 Burkholderia silvatlantica (AY965240)
Burkholderia mimosarum (AY752958)
5 Burkholderia oxyphila (AB488693)
Burkholderia sacchari (AF263278)
24 Burkholderia unamae (AY221956)
10 49 Burkholderia eburnea (JQ692176)
36 Burkholderia tropica (AJ420332)
18 Burkholderia nodosa (AY773189)
33 Burkholderia bannensis (AB561874)
Clade Ia
B. cepacia complex
Clade Ic
Phytopathogens
Clade Ib
B. pseudomallei group
Clade I
Clade II
0.05

FIGURE 2 | A maximum likelihood tree based on the 16S rRNA gene sequences of 97 members of the genus *Burkholderia*. Accession numbers for the 16S rRNA sequenced used for each organism are provided in the brackets following the name of the organism. The tree was rooted using four species from the genera *Cupriadivus* and *Ralstonia*. Bootstrap analysis scores are indicated for each node. The major *Burkholderia* clades (Clades I and II) and the subclades within Clade I are indicated by brackets.

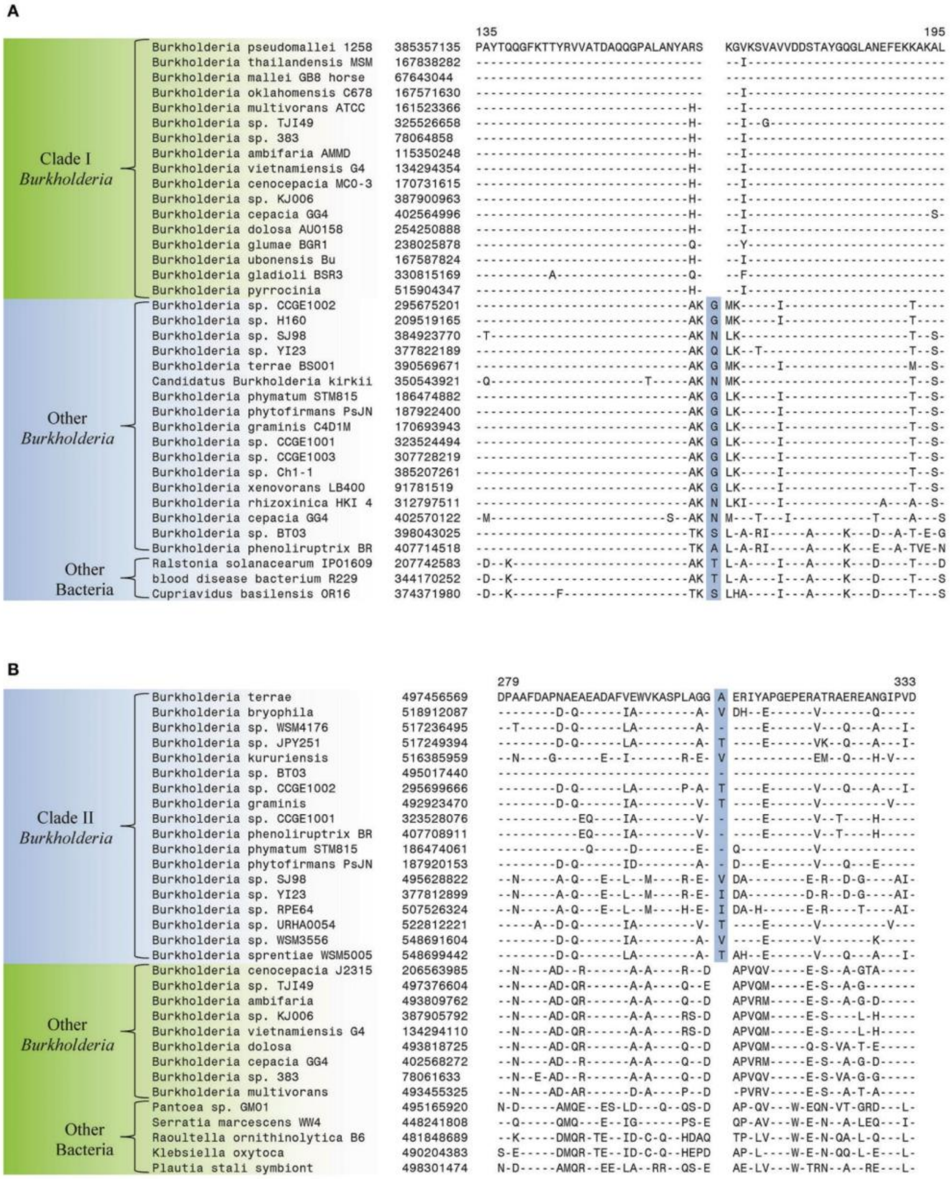

A
135
195
PAYTQQGFKTTYRVVATDAQQGPALANYARS KGVKSVAVVDDSTAYGQGLANEFEKKAKAL
Clade I Burkholderia
Burkholderia pseudomallei 1258 385357135
Burkholderia thailandensis MSM 167838282
Burkholderia mallei GB8 horse 67643044
Burkholderia oklahomensis C678 167571630
Burkholderia multivorans ATCC 161523366
Burkholderia sp. TJI49 325526658
Burkholderia sp. 383 78064858
Burkholderia ambifaria AMMD 115350248
Burkholderia vietnamiensis G4 134294354
Burkholderia cenocepacia MCO-3 170731615
Burkholderia sp. KJ006 387900963
Burkholderia cepacia GG4 402564996
Burkholderia dolosa AU0158 254250888
Burkholderia glumae BGR1 238025878
Burkholderia ubonensis Bu 167587824
Burkholderia gladioli BSR3 330815169
Burkholderia pyrrocinia 515904347
Other Burkholderia
Burkholderia sp. CCGE1002 295675201
Burkholderia sp. H160 209519165
Burkholderia sp. SJ98 384923770
Burkholderia sp. YI23 377822189
Burkholderia terrae BS001 390569671
Candidatus Burkholderia kirkii 350543921
Burkholderia phymatum STM815 186474882
Burkholderia phytofirmans PsJN 187922400
Burkholderia graminis C4D1M 170693943
Burkholderia sp. CCGE1001 323524494
Burkholderia sp. CCGE1003 307728219
Burkholderia sp. Ch1-1 385207261
Burkholderia xenovorans LB400 91781519
Burkholderia rhizoxinica HKI 4 312797511
Burkholderia cepacia GG4 402570122
Burkholderia sp. BT03 398043025
Burkholderia phenoliruptrix BR 407714518
Other Bacteria
Ralstonia solanacearum IPO1609 207742583
blood disease bacterium R229 344170252
Cupriavidus basilensis OR16 374371980
B
279
333
DPAAFDAPNAEAEADAFVEWVKASPLAGG A ERIYAPGEPERATRAEREANGIPVD
Clade II Burkholderia
Burkholderia terrae 497456569
Burkholderia bryophila 518912087
Burkholderia sp. WSM4176 517236495
Burkholderia sp. JPY251 517249394
Burkholderia kururiensis 516385959
Burkholderia sp. BT03 495017440
Burkholderia sp. CCGE1002 295699666
Burkholderia graminis 492923470
Burkholderia sp. CCGE1001 323528076
Burkholderia phenoliruptrix BR 407708911
Burkholderia phymatum STM815 186474061
Burkholderia phytofirmans PsJN 187920153
Burkholderia sp. SJ98 495628822
Burkholderia sp. YI23 377812899
Burkholderia sp. RPE64 507526324
Burkholderia sp. URHA0054 522812221
Burkholderia sp. WSM3556 548691604
Burkholderia sprentiae WSM5005 548699442
Other Burkholderia
Burkholderia cenocepacia J2315 206563985
Burkholderia sp. TJI49 497376604
Burkholderia ambifaria 493809762
Burkholderia sp. KJ006 387905792
Burkholderia vietnamiensis G4 134294110
Burkholderia dolosa 493818725
Burkholderia cepacia GG4 402568272
Burkholderia sp. 383 78061633
Burkholderia multivorans 493455325
Other Bacteria
Pantoea sp. GM01 495165920
Serratia marcescens WW4 448241808
Raoultella ornithinolytica B6 481848689
Klebsiella oxytoca 490204383
Plautia stali symbiont 498301474

FIGURE 3 | Partial sequence alignments of (A) a periplasmic amino acid-binding protein showing a 1 amino acid deletion identified in all members of Clade I of the genus *Burkholderia* (B) a dehydrogenase showing a 1 amino acid insertion (boxed) identified only in members of Clade II of the genus *Burkholderia*. These CSIs were not found in the sequence homologs of these proteins from any other sequenced bacteria. In each case, sequence information for a *Burkholderia* species and a limited number other bacteria are shown, but unless otherwise indicated, similar CSIs were detected in all members of the indicated group and not detected in any other bacterial species in the top 250 BLAST hits. The dashes (–) in the alignments indicate identity with the residue in the top sequence. GenBank identification (GI) numbers for each sequence are presented in the second column. Sequence information for other CSIs specific to the members of Clade I and Clade II of the genus *Burkholderia* are presented in Supplemental Figures 1–5 and Supplemental Figure 6, respectively, and their characteristics are summarized in **Table 2**.

specific for the Clade I *Burkholderia* and sequence alignments for these CSIs are shown in Supplemental Figures 1–5 and a summary of their characteristics is provided in **Table 2.**

Two additional CSIs identified in this work are specific for the Clade II *Burkholderia* species which is made up of mainly environmental organisms. One of these CSIs, shown in **Figure 3B**, consists of a one amino acid insertion in a dehydrogenase protein that is uniquely found in members of the Clade II *Burkholderia* and absent in all other *Burkholderia* species as well all other bacterial groups. A sequence alignment for another CSI that is specific for the Clade II *Burkholderia* (a 2 aa deletion in a LysR family of transcription regulator protein) is shown in Supplemental Figure 6 and its characteristics are summarized in **Table 2.**

CSIs DISTINGUISHING DIFFERENT MAIN GROUPS WITHIN THE CLADE I *BURKHOLDERIA*

The species within Clade I of the genus *Burkholderia* are responsible for a range of human, animal, and plant diseases (Biddick et al., 2003; Mahenthiralingam et al., 2005). The members of Clade I (i.e., the BCC and related *Burkholderia*) are commonly separated into 3 main groups which correspond to clades identified in our phylogenetic trees. The first group, the members of the BCC (Clade 1a), are prevalent pathogens in cystic fibrosis patients, the second group, the *B. pseudomallei* group (Clade Ib), contains the causative agents of melioidosis and glanders, while the third group contains the plant pathogenic *Burkholderia* species (Clade Ic) (White, 2003; Mahenthiralingam et al., 2005; Whitlock et al., 2007; Nandakumar et al., 2009). Our analysis has identified 3 CSIs that are specific for all members of the BCC clade (Clade 1a). One example of a BCC clade specific CSI is shown in **Figure 4A.** This CSI consists of a 2 amino acid insertion in a conserved region of a histidine utilization repressor which is only found in members of the BCC. Sequence alignments for two other BCC clade specific CSIs are shown in Supplemental Figures 7, 8 and their characteristics are summarized in **Table 3.**

Our work has also identified 4 CSIs that are specific for the *B. pseudomallei* group (Clade Ib) which contains the most prevalent human pathogen within the genus, *B. pseudomallei* (Wiersinga et al., 2006). One example of a CSI specific to the *B. pseudomallei* group, which consists of a 1 amino acid insertion in a conserved region of a periplasmic oligopeptide-binding protein, is shown in **Figure 4B.** Sequence alignments for three other CSIs in three different proteins that are specific for the *B. pseudomallei* group are shown in Supplemental Figures 9–11 and their characteristics are summarized in **Table 3.**

We have also identified 5 CSIs that are specific for the major plant pathogenic group within the genus *Burkholderia* (Clade 1c) which contains the species *B. glumae* and *B. gladioli*. An example of a CSI representing this group is shown in **Figure 4C.** This CSI consists of a 1 amino acid insertion in a conserved region of a SMP-30/gluconolaconase/LRE-like region-containing protein that is found in the members of Clade 1c of the genus *Burkholderia* but absent in all other *Burkholderia* and all other bacterial groups. Sequence alignments for the other 4 CSIs are shown in Supplemental Figures 12–15 and their key features are highlighted in **Table 3.**

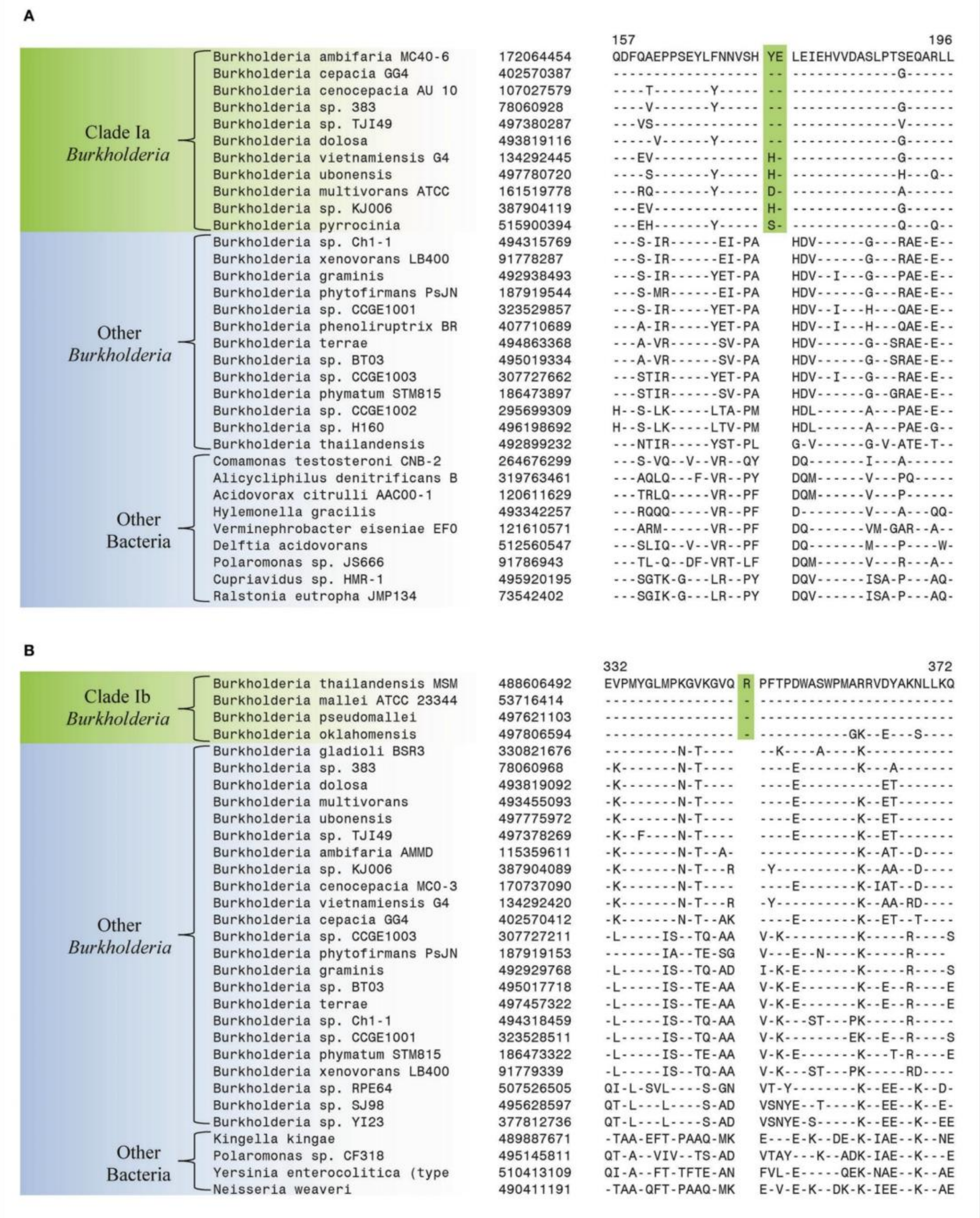
CSIs THAT ARE SPECIFIC FOR TWO GROUPS WITHIN THE CLADE II *BURKHOLDERIA*

The species within Clade II of the genus *Burkholderia* inhabit a variety of environmental niches, but there is little evidence of their colonization of healthy or immunocompromised human patients (Coenye and Vandamme, 2003). The branching of different groups within Clade II is not well resolved in 16S rRNA trees and there is currently a lack of sequence data that can be used to generate trees based on concatenated gene sets that reliably resolve the interrelationships of the clade while sufficiently reflecting the total diversity of species within the clade (**Figures 1, 2**) (Cole et al., 2009; NCBI, 2014). Despite the limited sequence data, we have been able to identify two robust groups within Clade II that are supported by a number of CSIs. The first Clade, Clade IIa, primarily consists of unclassified members of the genus and candidatus *Burkholderia* species (**Figure 1**). Clade IIa is supported by 16 CSIs identified in this work. One example of a CSI specific for Clade IIa, consisting of a 1 amino acid insertion in 3-phosphoglycerate dehydrogenase, is shown in **Figure 5A.** This insertion is present in a highly conserved region of this protein in all sequenced members of Clade IIa and absent in all other *Burkholderia* and all other bacterial groups. Sequence alignments for the other 15 CSIs that are specific for Clade IIa *Burkholderia* spp. are shown in Supplemental Figures 16–30 and their characteristics are summarized in **Table 3.**

Table 2 | Conserved signature indels specific for the two major clades within the genus *Burkholderia*.

Protein Name	GI Number	Figures	Indel size	Indel position[a]	Specificity
Periplasmic amino acid-binding protein	385357135	**Figure 3A**	1 aa del	135–195	Clade I
Putative lyase	167724527	Supplemental Figure 1	1 aa del	70–121	Clade I
4-hydroxybenzoate 3-monooxygenase	238023559	Supplemental Figure 2	1 aa ins	101–171	Clade I
6-phosphogluconate dehydrogenase, decarboxylating	330820932	Supplemental Figure 3	1 aa ins	137–202	Clade I
Putative lipoprotein	121598811	Supplemental Figure 4	1 aa del	363–393	Clade I
Sarcosine oxidase subunit alpha	493818877	Supplemental Figure 5	3 aa ins	904–965	Clade I
Dehydrogenase	497456569	**Figure 3B**	1 aa ins	279–333	Clade II
LysR family transcriptional regulator	187919777	Supplemental Figure 6	2 aa del	260–294	Clade II

[a] *The region of the specified protein that contains the indel.*



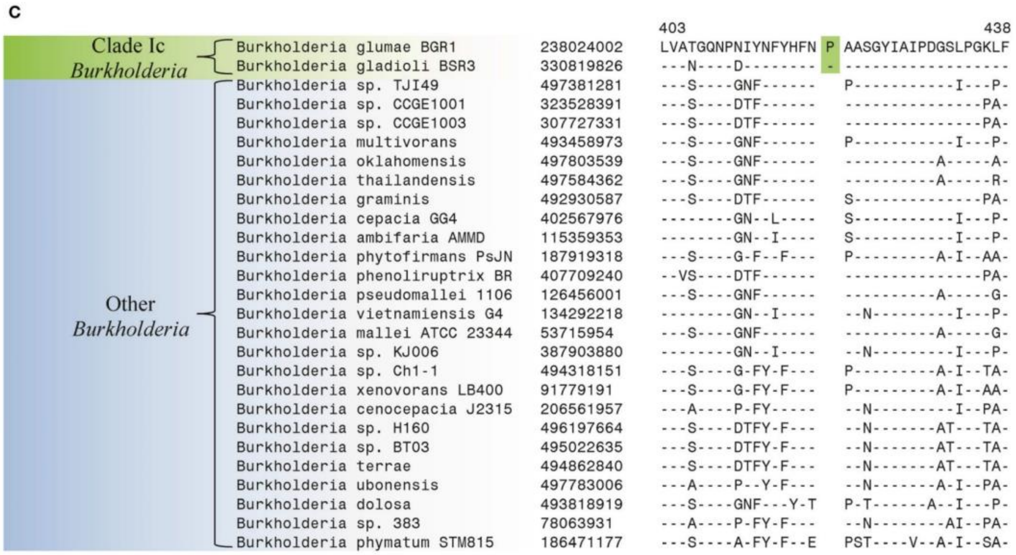
A

			157		196
Clade Ia *Burkholderia*	Burkholderia ambifaria MC40-6	172064454	QDFQAEPPSEYLFNNVSH	YE	LEIEHVVDASLPTSEQARLL
	Burkholderia cepacia GG4	402570387	------------------	--	-------------G------
	Burkholderia cenocepacia AU 10	107027579	----T-------Y-----	--	--------------------
	Burkholderia sp. 383	78060928	----V-------Y-----	--	-------------G------
	Burkholderia sp. TJI49	497380287	---VS-------------	--	-------------V------
	Burkholderia dolosa	493819116	-----V------Y-----	--	-------------G------
	Burkholderia vietnamiensis G4	134292445	---EV-------------	H-	-------------G------
	Burkholderia ubonensis	497780720	----S-------Y-----	H-	-------------H---Q--
	Burkholderia multivorans ATCC	161519778	---RQ-------Y-----	D-	-------------A------
	Burkholderia sp. KJ006	387904119	---EV-------------	H-	-------------G------
	Burkholderia pyrrocinia	515900394	---EH-------Y-----	S-	-------------Q---Q--
Other *Burkholderia*	Burkholderia sp. Ch1-1	494315769	---S-IR------EI-PA		HDV------G---RAE-E--
	Burkholderia xenovorans LB400	91778287	---S-IR------EI-PA		HDV------G---RAE-E--
	Burkholderia graminis	492938493	---S-IR-----YET-PA		HDV--I---G---PAE-E--
	Burkholderia phytofirmans PsJN	187919544	---S-MR------EI-PA		HDV------G---RAE-E--
	Burkholderia sp. CCGE1001	323529857	---S-IR-----YET-PA		HDV--I---H---QAE-E--
	Burkholderia phenoliruptrix BR	407710689	---A-IR-----YET-PA		HDV--I---H---QAE-E--
	Burkholderia terrae	494863368	---A-VR------SV-PA		HDV------G--SRAE-E--
	Burkholderia sp. BT03	495019334	---A-VR------SV-PA		HDV------G--SRAE-E--
	Burkholderia sp. CCGE1003	307727662	---STIR-----YET-PA		HDV--I---G---RAE-E--
	Burkholderia phymatum STM815	186473897	---STIR------SV-PA		HDV------G--GRAE-E--
	Burkholderia sp. CCGE1002	295699309	H--S-LK-----LTA-PM		HDL------A---PAE-E--
	Burkholderia sp. H160	496198692	H--S-LK-----LTV-PM		HDL------A---PAE-G--
	Burkholderia thailandensis	492899232	---NTIR-----YST-PL		G-V------G-V-ATE-T--
Other Bacteria	Comamonas testosteroni CNB-2	264676299	---S-VQ--V--VR--QY		DQ-------I---A------
	Alicycliphilus denitrificans B	319763461	---AQLQ---F-VR--PY		DQM------V---PQ-----
	Acidovorax citrulli AAC00-1	120611629	---TRLQ-----VR--PF		DQM------V---P------
	Hylemonella gracilis	493342257	---RQQQ-----VR--PF		D--------V---A---QQ-
	Verminephrobacter eiseniae EF0	121610571	---ARM------VR--PF		DQ-------VM-GAR--A--
	Delftia acidovorans	512560547	---SLIQ--V--VR--PF		DQ-------M---P----W-
	Polaromonas sp. JS666	91786943	---TL-Q--DF-VRT-LF		DQM------V---R---A--
	Cupriavidus sp. HMR-1	495920195	---SGTK-G---LR--PY		DQV------ISA-P---AQ-
	Ralstonia eutropha JMP134	73542402	---SGIK-G---LR--PY		DQV------ISA-P---AQ-

B

			332		372
Clade Ib *Burkholderia*	Burkholderia thailandensis MSM	488606492	EVPMYGLMPKGVKGVQ	R	PFTPDWASWPMARRVDYAKNLLKQ
	Burkholderia mallei ATCC 23344	53716414	----------------	-	------------------------
	Burkholderia pseudomallei	497621103	----------------	-	------------------------
	Burkholderia oklahomensis	497806594	----------------	-	-----------GK--E---S----
Other *Burkholderia*	Burkholderia gladioli BSR3	330821676	---------N-T----		--K----A----K-----------
	Burkholderia sp. 383	78060968	-K-------N-T----		----E-------K---A-------
	Burkholderia dolosa	493819092	-K-------N-T----		----E----------ET-------
	Burkholderia multivorans	493455093	-K-------N-T----		----E-------K--ET-------
	Burkholderia ubonensis	497775972	-K-------N-T----		----E-------K--ET-------
	Burkholderia sp. TJI49	497378269	-K--F----N-T----		----E-------K--ET-------
	Burkholderia ambifaria AMMD	115359611	-K-------N-T--A-		------------K--AT--D----
	Burkholderia sp. KJ006	387904089	-K-------N-T---R		-Y----------K--AA--D----
	Burkholderia cenocepacia MCO-3	170737090	-K-------N-T----		----E-------K-IAT--D----
	Burkholderia vietnamiensis G4	134292420	-K-------N-T---R		-Y----------K--AA-RD----
	Burkholderia cepacia GG4	402570412	-K-------N-T--AK		----E-------K--ET--T----
	Burkholderia sp. CCGE1003	307727211	-L-----IS--TQ-AA		V-K---------K-----R----S
	Burkholderia phytofirmans PsJN	187919153	-------IA--TE-SG		V---E--N----K-----R----
	Burkholderia graminis	492929768	-L-----IS--TQ-AD		I-K-E-------K-----R----S
	Burkholderia sp. BT03	495017718	-L-----IS--TE-AA		V-K-E-------K--E--R----E
	Burkholderia terrae	497457322	-L-----IS--TE-AA		V-K-E-------K--E--R----E
	Burkholderia sp. Ch1-1	494318459	-L-----IS--TQ-AA		V-K---ST---PK-----R-----
	Burkholderia sp. CCGE1001	323528511	-L-----IS--TQ-AA		V-K--------EK--E--R----S
	Burkholderia phymatum STM815	186473322	-L-----IS--TE-AA		V-K-E-------K---T-R----E
	Burkholderia xenovorans LB400	91779339	-L-----IS--TQ-AA		V-K---ST---PK-----RD----
	Burkholderia sp. RPE64	507526505	QI-L-SVL----S-GN		VT-Y--------K--EE--K--D-
	Burkholderia sp. SJ98	495628597	QT-L---L----S-AD		VSNYE--T----K--EE--K--E-
	Burkholderia sp. YI23	377812736	QT-L---L----S-AD		VSNYE-S-----K--EE--K--EE
Other Bacteria	Kingella kingae	489887671	-TAA-EFT-PAAQ-MK		E---E-K--DE-K-IAE--K--NE
	Polaromonas sp. CF318	495145811	QT-A--VIV--TS-AD		VTAY---K--ADK-IAE--K---E
	Yersinia enterocolitica (type	510413109	QI-A--FT-TFTE-AN		FVL-E-----QEK-NAE--K--AE
	Neisseria weaveri	490411191	-TAA-QFT-PAAQ-MK		E-V-E-K--DK-K-IEE--K--AE

FIGURE 4 | Continued


C
Clade Ic
Burkholderia
Other
Burkholderia

```
                                                         403                                   438
Burkholderia glumae BGR1              238024002   LVATGQNPNIYNFYHFN P AASGYIAIPDGSLPGKLF
Burkholderia gladioli BSR3            330819826   ---N----D-------- - -----------------
Burkholderia sp. TJI49                497381281   ---S----GNF------   P-----------I---P-
Burkholderia sp. CCGE1001             323528391   ---S----DTF------   ---------------PA-
Burkholderia sp. CCGE1003             307727331   ---S----DTF------   ---------------PA-
Burkholderia multivorans              493458973   ---S----GNF------   P-----------I---P-
Burkholderia oklahomensis             497803539   ---S----GNF------   ----------A-----A-
Burkholderia thailandensis            497584362   ---S----GNF------   ----------A-----R-
Burkholderia graminis                 492930587   ---S----DTF------   S--------------PA-
Burkholderia cepacia GG4              402567976   --------GN--L----   S-----------I---P-
Burkholderia ambifaria AMMD           115359353   --------GN--I----   S-----------I---P-
Burkholderia phytofirmans PsJN        187919318   ---S----G-F--F---   P---------A-I--AA-
Burkholderia phenoliruptrix BR        407709240   --VS----DTF------   ---------------PA-
Burkholderia pseudomallei 1106        126456001   ---S----GNF------   ----------A-----G-
Burkholderia vietnamiensis G4         134292218   --------GN--I----   --N---------I---P-
Burkholderia mallei ATCC 23344        53715954    ---S----GNF------   ----------A-----G-
Burkholderia sp. KJ006                387903880   --------GN--I----   --N---------I---P-
Burkholderia sp. Ch1-1                494318151   ---S----G-FY-F---   P---------A-I--TA-
Burkholderia xenovorans LB400         91779191    ---S----G-FY-F---   P---------A-I--AA-
Burkholderia cenocepacia J2315        206561957   ---A----P-FY-----   --N----------I--PA-
Burkholderia sp. H160                 496197664   ---S----DTFY-F---   --N-------AT---TA-
Burkholderia sp. BT03                 495022635   ---S----DTFY-F---   --N-------AT---TA-
Burkholderia terrae                   494862840   ---S----DTFY-F---   --N-------AT---TA-
Burkholderia ubonensis                497783006   ---A----P--Y-F---   --N-------A-I--PA-
Burkholderia dolosa                   493818919   ---S----GNF---Y-T   P-T------A--I---P-
Burkholderia sp. 383                  78063931    ---A----P-FY-F---   --N---------AI--PA-
Burkholderia phymatum STM815          186471177   ---S----A-FY-F--E   PST----V--A-I--SA-
```

FIGURE 4 | Partial sequence alignments of (A) a histidine utilization repressor showing a 2 amino acid insertion (boxed) identified in all members of the *Burkholderia cepacia* complex (Clade Ia) within the genus *Burkholderia* (B) a periplasmic oligopeptide-binding protein showing a 1 amino acid insertion (boxed) identified in all members of the *Burkholderia pseudomallei* group (Clade Ib) within the genus *Burkholderia* (C) a SMP-30/gluconolaconase/LRE-like region-containing protein showing a 1 amino acid insertion (boxed) identified in all members of the phytopathogenic *Burkholderia* clade (Clade Ic). These CSIs were not found in the sequence homologs of these proteins from any other sequenced bacteria in the top 250 BLAST hits. Sequence information for other CSIs specific to subclades within Clade I of the genus *Burkholderia* are presented in Supplemental Figures 7–15 and their characteristics are summarized in **Table 3**.

The second group within Clade II of the *Burkholderia* (Clade IIb), is comprised of a large variety of environmental *Burkholderia* species (Coenye and Vandamme, 2003; Suarez-Moreno et al., 2012). Our analysis has identified 6 CSIs that are specific to this large group of *Burkholderia* species. One example of a CSI specific to the members of Clade IIb of the genus *Burkholderia* is shown in **Figure 5B**. The CSI consists of a one amino acid insertion in 4-hydroxyacetophenone monooxygenase, which is only present in members of Clade IIb of the genus *Burkholderia* and not in protein homologs from any other sequenced bacterial group. Information for other 5 CSIs which are specific to members of Clade IIb of the genus *Burkholderia* are shown in Supplemental Figures 31–35 and their characteristics are summarized in **Table 3**.

DISCUSSION

The genus *Burkholderia* is one of the largest groups of species within the class *Betaproteobacteria* (Palleroni, 2005; Parte, 2013). The genus contains a variety of bacteria that inhabit a wide range of ecological niches including a number of bacteria that have pathogenic potential (Yabuuchi et al., 1992; Coenye and Vandamme, 2003; Mahenthiralingam et al., 2005; Palleroni, 2005; Compant et al., 2008). The phylogeny of the genus *Burkholderia* has been studied using a wide array of methodologies based on phenotypic, biochemical, genetic, and genomic characteristics (Stead, 1992; Gillis et al., 1995; Payne et al., 2005; Tayeb et al., 2008; Onofre-Lemus et al., 2009; Spilker et al., 2009; Ussery et al., 2009; Gyaneshwar et al., 2011; Vandamme and Dawyndt, 2011; Zhu et al., 2011; Estrada-de los Santos et al., 2013). These studies have provided novel insights into the evolutionary relationship of the species within the genus *Burkholderia*. However, no taxonomic changes have been made to date due to a lack of discrete, distinguishing characteristics identified for the different phylogenetic lineages within the genus (Estrada-de los Santos et al., 2013).

In the present work, we have outlined two major groups of species within the genus *Burkholderia*: Clade I, which contains all pathogenic members of the genus, and Clade II, which contains a large variety of environmental species. These two groups were found to branch distinctly in a highly resolved phylogenetic tree based on a large number of concatenated protein sequences produced in this work (**Figure 1**). Evidence for the distinctness of Clade I organisms from other *Burkholderia* species has been observed in a wide range of previous phylogenetic studies (Payne et al., 2005; Tayeb et al., 2008; Yarza et al., 2008; Spilker et al., 2009; Ussery et al., 2009; Gyaneshwar et al., 2011; Vandamme and Dawyndt, 2011; Zhu et al., 2011; Suarez-Moreno et al., 2012; Estrada-de los Santos et al., 2013; Segata et al., 2013). Importantly, we have also identified 6 and 2 CSIs that serve as discrete molecular characteristics of Clade I and Clade II, respectively (**Figure 6** and **Table 2**). These CSIs are the

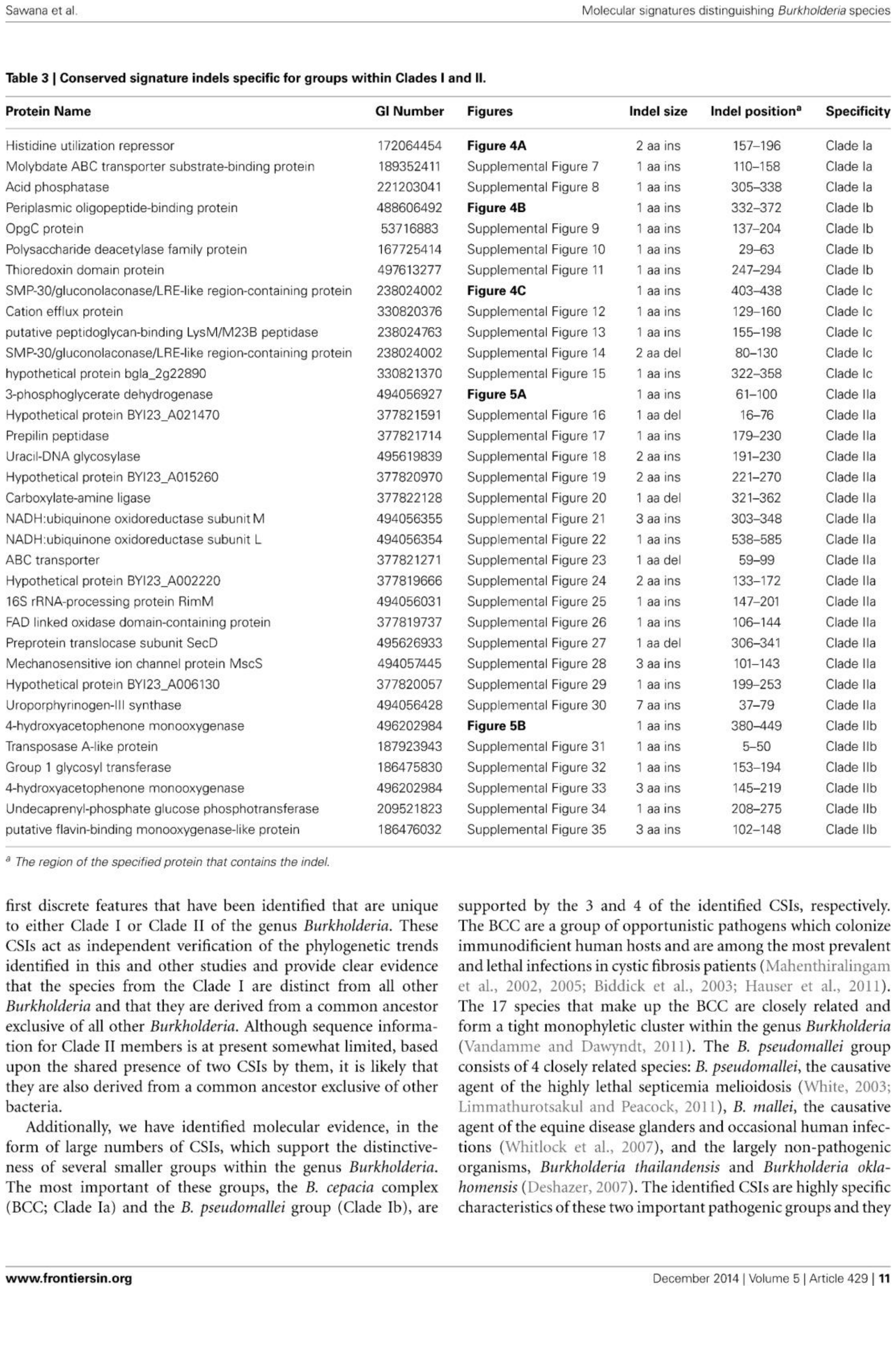


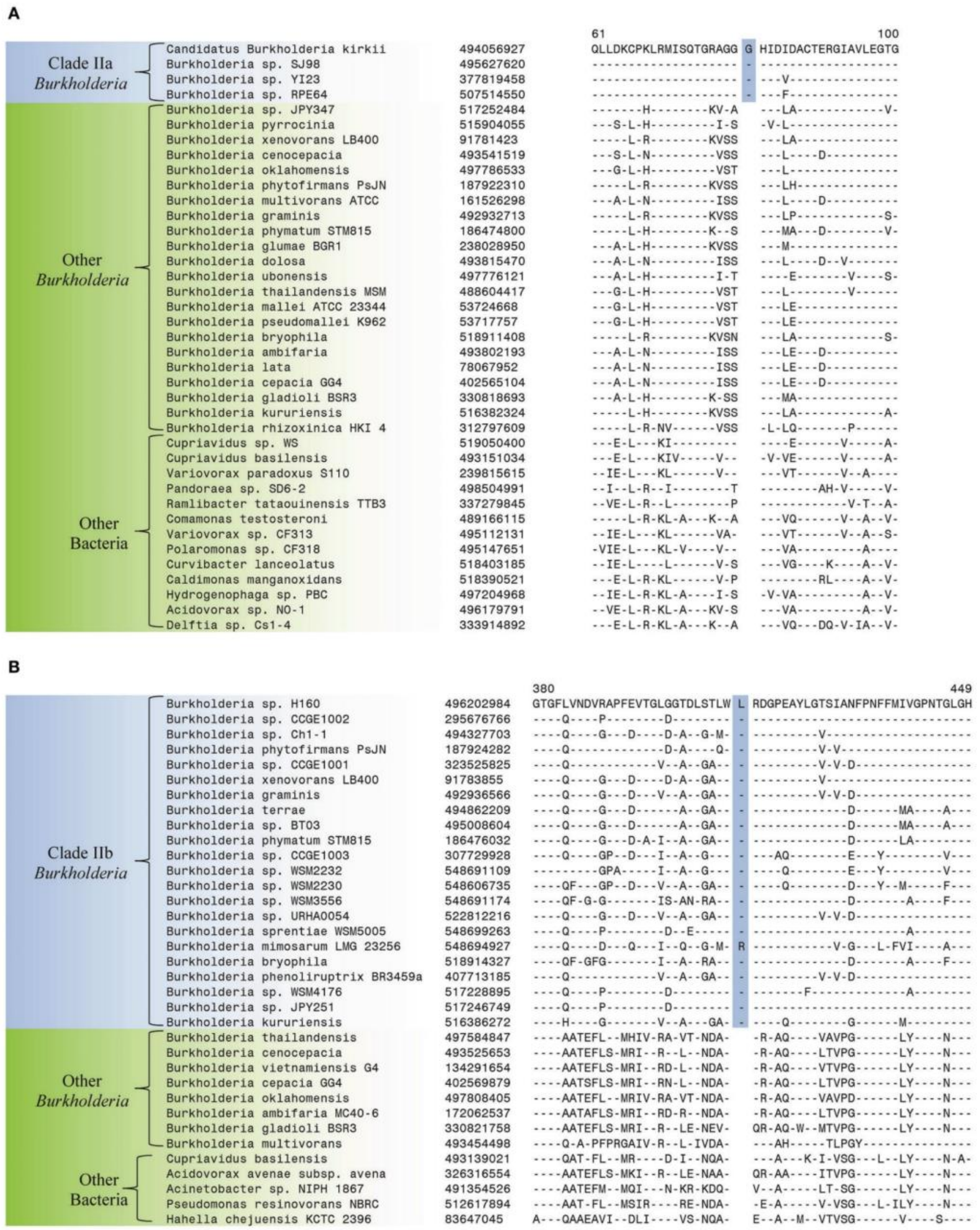
Table 3 | Conserved signature indels specific for groups within Clades I and II.

Protein Name	GI Number	Figures	Indel size	Indel position[a]	Specificity
Histidine utilization repressor	172064454	**Figure 4A**	2 aa ins	157–196	Clade Ia
Molybdate ABC transporter substrate-binding protein	189352411	Supplemental Figure 7	1 aa ins	110–158	Clade Ia
Acid phosphatase	221203041	Supplemental Figure 8	1 aa ins	305–338	Clade Ia
Periplasmic oligopeptide-binding protein	488606492	**Figure 4B**	1 aa ins	332–372	Clade Ib
OpgC protein	53716883	Supplemental Figure 9	1 aa ins	137–204	Clade Ib
Polysaccharide deacetylase family protein	167725414	Supplemental Figure 10	1 aa ins	29–63	Clade Ib
Thioredoxin domain protein	497613277	Supplemental Figure 11	1 aa ins	247–294	Clade Ib
SMP-30/gluconolaconase/LRE-like region-containing protein	238024002	**Figure 4C**	1 aa ins	403–438	Clade Ic
Cation efflux protein	330820376	Supplemental Figure 12	1 aa ins	129–160	Clade Ic
putative peptidoglycan-binding LysM/M23B peptidase	238024763	Supplemental Figure 13	1 aa ins	155–198	Clade Ic
SMP-30/gluconolaconase/LRE-like region-containing protein	238024002	Supplemental Figure 14	2 aa del	80–130	Clade Ic
hypothetical protein bgla_2g22890	330821370	Supplemental Figure 15	1 aa ins	322–358	Clade Ic
3-phosphoglycerate dehydrogenase	494056927	**Figure 5A**	1 aa ins	61–100	Clade IIa
Hypothetical protein BYI23_A021470	377821591	Supplemental Figure 16	1 aa del	16–76	Clade IIa
Prepilin peptidase	377821714	Supplemental Figure 17	1 aa ins	179–230	Clade IIa
Uracil-DNA glycosylase	495619839	Supplemental Figure 18	2 aa ins	191–230	Clade IIa
Hypothetical protein BYI23_A015260	377820970	Supplemental Figure 19	2 aa ins	221–270	Clade IIa
Carboxylate-amine ligase	377822128	Supplemental Figure 20	1 aa del	321–362	Clade IIa
NADH:ubiquinone oxidoreductase subunit M	494056355	Supplemental Figure 21	3 aa ins	303–348	Clade IIa
NADH:ubiquinone oxidoreductase subunit L	494056354	Supplemental Figure 22	1 aa ins	538–585	Clade IIa
ABC transporter	377821271	Supplemental Figure 23	1 aa del	59–99	Clade IIa
Hypothetical protein BYI23_A002220	377819666	Supplemental Figure 24	2 aa ins	133–172	Clade IIa
16S rRNA-processing protein RimM	494056031	Supplemental Figure 25	1 aa ins	147–201	Clade IIa
FAD linked oxidase domain-containing protein	377819737	Supplemental Figure 26	1 aa ins	106–144	Clade IIa
Preprotein translocase subunit SecD	495626933	Supplemental Figure 27	1 aa del	306–341	Clade IIa
Mechanosensitive ion channel protein MscS	494057445	Supplemental Figure 28	3 aa ins	101–143	Clade IIa
Hypothetical protein BYI23_A006130	377820057	Supplemental Figure 29	1 aa ins	199–253	Clade IIa
Uroporphyrinogen-III synthase	494056428	Supplemental Figure 30	7 aa ins	37–79	Clade IIa
4-hydroxyacetophenone monooxygenase	496202984	**Figure 5B**	1 aa ins	380–449	Clade IIb
Transposase A-like protein	187923943	Supplemental Figure 31	1 aa ins	5–50	Clade IIb
Group 1 glycosyl transferase	186475830	Supplemental Figure 32	1 aa ins	153–194	Clade IIb
4-hydroxyacetophenone monooxygenase	496202984	Supplemental Figure 33	3 aa ins	145–219	Clade IIb
Undecaprenyl-phosphate glucose phosphotransferase	209521823	Supplemental Figure 34	1 aa ins	208–275	Clade IIb
putative flavin-binding monooxygenase-like protein	186476032	Supplemental Figure 35	3 aa ins	102–148	Clade IIb

[a] *The region of the specified protein that contains the indel.*

first discrete features that have been identified that are unique to either Clade I or Clade II of the genus *Burkholderia*. These CSIs act as independent verification of the phylogenetic trends identified in this and other studies and provide clear evidence that the species from the Clade I are distinct from all other *Burkholderia* and that they are derived from a common ancestor exclusive of all other *Burkholderia*. Although sequence information for Clade II members is at present somewhat limited, based upon the shared presence of two CSIs by them, it is likely that they are also derived from a common ancestor exclusive of other bacteria.

Additionally, we have identified molecular evidence, in the form of large numbers of CSIs, which support the distinctiveness of several smaller groups within the genus *Burkholderia*. The most important of these groups, the *B. cepacia* complex (BCC; Clade Ia) and the *B. pseudomallei* group (Clade Ib), are supported by the 3 and 4 of the identified CSIs, respectively. The BCC are a group of opportunistic pathogens which colonize immunodificient human hosts and are among the most prevalent and lethal infections in cystic fibrosis patients (Mahenthiralingam et al., 2002, 2005; Biddick et al., 2003; Hauser et al., 2011). The 17 species that make up the BCC are closely related and form a tight monophyletic cluster within the genus *Burkholderia* (Vandamme and Dawyndt, 2011). The *B. pseudomallei* group consists of 4 closely related species: *B. pseudomallei*, the causative agent of the highly lethal septicemia melioidosis (White, 2003; Limmathurotsakul and Peacock, 2011), *B. mallei*, the causative agent of the equine disease glanders and occasional human infections (Whitlock et al., 2007), and the largely non-pathogenic organisms, *Burkholderia thailandensis* and *Burkholderia oklahomensis* (Deshazer, 2007). The identified CSIs are highly specific characteristics of these two important pathogenic groups and they



```
A
                                                   61                                  100
Candidatus Burkholderia kirkii       494056927     QLLDKCPKLRMISQTGRAGG G HIDIDACTERGIAVLEGTG
Burkholderia sp. SJ98                495627620     -------------------- - -------------------
Burkholderia sp. YI23                377819458     -------------------- - ---V---------------
Burkholderia sp. RPE64               507514550     -------------------- - ---F---------------
Burkholderia sp. JPY347              517252484     -------H--------KV-A   ---LA-----------V-
Burkholderia pyrrocinia              515904055     ---S-L-H---------I-S   -V-L---------------
Burkholderia xenovorans LB400        91781423      -----L-R--------KVSS   ---LA--------------
Burkholderia cenocepacia             493541519     ---S-L-N---------VSS   ---L----D----------
Burkholderia oklahomensis            497786533     ---G-L-H---------VST   ---L---------------
Burkholderia phytofirmans PsJN       187922310     -----L-R--------KVSS   ---LH--------------
Burkholderia multivorans ATCC        161526298     ---A-L-N---------ISS   ---L----D----------
Burkholderia graminis                492932713     -----L-R--------KVSS   ---LP-----------S-
Burkholderia phymatum STM815         186474800     -----L-H--------K--S   ---MA---D--------V-
Burkholderia glumae BGR1             238028950     ---A-L-H--------KVSS   ---M---------------
Burkholderia dolosa                  493815470     ---A-L-N---------ISS   ---L----D--V-------
Burkholderia ubonensis               497776121     ---A-L-H---------I-T   ----E-------V----S-
Burkholderia thailandensis MSM       488604417     ---G-L-H---------VST   ---L--------V------
Burkholderia mallei ATCC 23344       53724668      ---G-L-H---------VST   ---LE--------------
Burkholderia pseudomallei K962       53717757      ---G-L-H---------VST   ---LE--------------
Burkholderia bryophila               518911408     -----L-R--------KVSN   ---LA-----------S-
Burkholderia ambifaria               493802193     ---A-L-N---------ISS   ---LE---D----------
Burkholderia lata                    78067952      ---A-L-N---------ISS   ---LE---D----------
Burkholderia cepacia GG4             402565104     ---A-L-N---------ISS   ---LE---D----------
Burkholderia gladioli BSR3           330818693     ---A-L-H--------K-SS   ---MA--------------
Burkholderia kururiensis             516382324     -----L-H--------KVSS   ---LA-----------A-
Burkholderia rhizoxinica HKI 4       312797609     -----L-R-NV------VSS   -L-LQ-------P------
Cupriavidus sp. WS                   519050400     ---E-L---KI---------   ----E------V-----A-
Cupriavidus basilensis               493151034     ---E-L---KIV-----V--   -V-VE------V-----A-
Variovorax paradoxus S110            239815615     --IE-L---KL------V--   ---VT------V--A----
Pandoraea sp. SD6-2                  498504991     --I--L-R--I--------T   --------AH-V-----V-
Ramlibacter tataouinensis TTB3       337279845     --VE-L-R--L--------P   ------------V-T--A-
Comamonas testosteroni               489166115     -----L-R-KL-A---K--A   ---VQ------V--A--V-
Variovorax sp. CF313                 495112131     --IE-L---KL------VA-   ---VT------V--A--S-
Polaromonas sp. CF318                495147651     -VIE-L---KL-V----V--   ---VA---------A----
Curvibacter lanceolatus              518403185     --IE-L----L------V-S   ---VG----K----A--V-
Caldimonas manganoxidans             518390521     ---E-L-R-KL------V-P   --------RL----A--V-
Hydrogenophaga sp. PBC               497204968     --IE-L-R-KL-A----I-S   -V-VA---------A--V-
Acidovorax sp. NO-1                  496179791     --VE-L-R-KL-A---KV-S   ---VA---------A--V-
Delftia sp. Cs1-4                    333914892     ---E-L-R-KL-A---K--A   ---VQ---DQ-V-IA--V-

B
                                                380                                                                449
Burkholderia sp. H160                 496202984 GTGFLVNDVRAPFEVTGLGGTDLSTLW L RDGPEAYLGTSIANFPNFFMIVGPNTGLGH
Burkholderia sp. CCGE1002             295676766 ----Q----P--------D-------- - -----------------------------
Burkholderia sp. Ch1-1                494327703 ----Q----G---D----D-A--G-M- - ---------V-------------------
Burkholderia phytofirmans PsJN        187924282 ----Q-------------D-A----Q- - ---------V-V-----------------
Burkholderia sp. CCGE1001             323525825 ----Q------------V--A--GA-- - ---------V-V-D---------------
Burkholderia xenovorans LB400         91783855  ----Q----G---D----D-A--GA-- - ---------V-------------------
Burkholderia graminis                 492936566 ----Q----G---D---V--A--GA-- - ---------V-V-D---------------
Burkholderia terrae                   494862209 ----Q----G---D------A--GA-- - -------------D------MA----A---
Burkholderia sp. BT03                 495008604 ----Q----G---D------A--GA-- - -------------D------MA----A---
Burkholderia phymatum STM815          186476032 ----Q----G---D-A-I--A--GA-- - -------------D------LA--------
Burkholderia sp. CCGE1003             307729928 ----Q----GP--D---I--A--G--- - ---AQ--------E---Y--------V---
Burkholderia sp. WSM2232              548691109 ---------GPA-----I--A--G--- - ----Q--------E---Y--------V---
Burkholderia sp. WSM2230              548606735 ----QF---GP--D---V--A--GA-- - ----Q--------D---Y--M-----F---
Burkholderia sp. WSM3556              548691174 ----QF-G-G-------IS-AN-RA-- - -------------D-------A----F---
Burkholderia sp. URHA0054             522812216 ----Q----G---D---V--A--GA-- - ---------V-V-D---------------
Burkholderia sprentiae WSM5005        548699263 ----Q----P--------D--E----- - ---------------------A--------
Burkholderia mimosarum LMG 23256      548694927 ----Q----D---Q---I--Q--G-M- R -----------V-G---L-FVI----A---
Burkholderia bryophila                518914327 ----QF-GFG-------I--A--RA-- - -------------D-------A----F---
Burkholderia phenoliruptrix BR3459a   407713185 ----Q------------V--A--GA-- - ---------V-V-D---------------
Burkholderia sp. WSM4176              517228895 ----Q----P--------D-------- - -------F-------------A--------
Burkholderia sp. JPY251               517246749 ----Q----P--------D-------- - -----------------------------
Burkholderia kururiensis              516386272 ----H----G-------V--A---GA- - ----Q--------G------M---------
Burkholderia thailandensis            497584847 ----AATEFL--MHIV-RA-VT-NDA-   -R-AQ----VAVPG------LY----N---
Burkholderia cenocepacia              493525653 ----AATEFLS-MRI--R--L--NDA-   -R-AQ----LTVPG------LY----N---
Burkholderia vietnamiensis G4         134291654 ----AATEFLS-MRI--RD-L--NDA-   -R-AQ----VTVPG------LY----N---
Burkholderia cepacia GG4              402569879 ----AATSFLS-MRI--RN-L--NDA-   -R-AQ----LTVPG------LY----N---
Burkholderia oklahomensis             497808405 ----AATEFL--MRIV-RA-VT-NDA-   -R-AQ----VAVPD------LY----N---
Burkholderia ambifaria MC40-6         172062537 ----AATAFLS-MRI--RD-R--NDA-   -R-AQ----LTVPG------LY----N---
Burkholderia gladioli BSR3            330821758 ----AATEFLS-MRI--R--LE-NEV-   QR-AQ-W--MTVPG------LY----N---
Burkholderia multivorans              493454498 ----Q-A-PFPRGAIV-R--L-IVDA-   ---AH-----TLPGY--------------
Cupriavidus basilensis                493139021 ----QAT-FL--MR----D-I--NQA-   ---A---K-I-VSG---L--LY----N-A-
Acidovorax avenae subsp. avena        326316554 ----AATEFLS-MKI--R--LE-NAA-   QR-AA----ITVPG------LY----N---
Acinetobacter sp. NIPH 1867           491354526 ----AATEFM--MQI---N-KR-KDQ-   V--A-----LT-SG------LY----N---
Pseudomonas resinovorans NBRC         512617894 ----AAT-FL--MSIR----RE-NDA-   -E-A-----V-VSG---L-ILY----N---
Hahella chejuensis KCTC 2396          83647045  A---QAAEAVI--DLI----VS-NQA-   E--A--M--VTVSG------V----S----
```

FIGURE 5 | Partial sequence alignments of (A) 3-phosphoglycerate dehydrogenase showing a 1 amino acid insertion (boxed) identified in all members of Clade IIa of the genus *Burkholderia* (B) 4-hydroxyacetophenone monooxygenase showing a 1 amino acid insertion (boxed) identified only in members of Clade IIb of the genus *Burkholderia*. These CSIs were not found in the sequence homologs of these proteins from any other sequenced bacteria in the top 250 BLAST hits. Sequence information for other CSIs specific to subclades within Clade II of the genus *Burkholderia* are presented in Supplemental Figures 16–35 and their characteristics are summarized in **Table 3**.

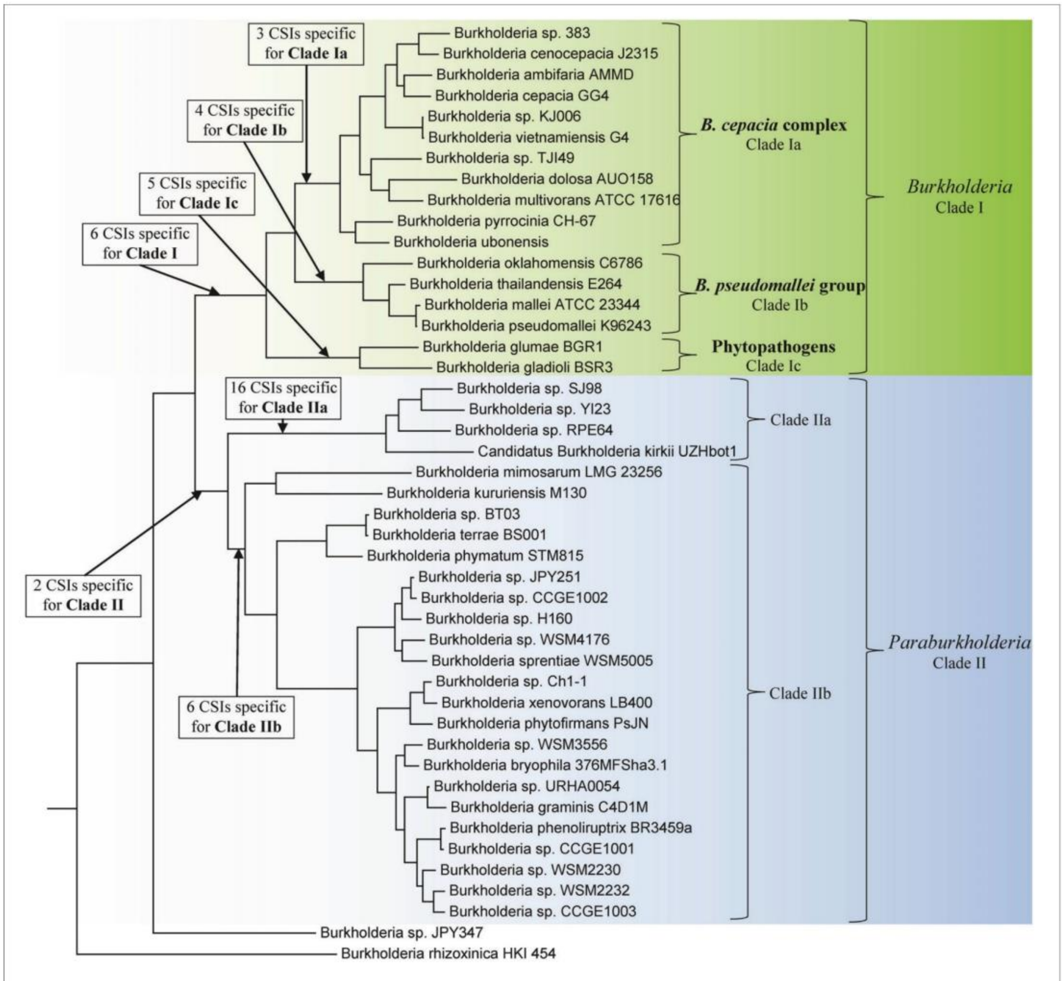

3 CSIs specific for Clade Ia
4 CSIs specific for Clade Ib
5 CSIs specific for Clade Ic
6 CSIs specific for Clade I
16 CSIs specific for Clade IIa
2 CSIs specific for Clade II
6 CSIs specific for Clade IIb
Burkholderia sp. 383
Burkholderia cenocepacia J2315
Burkholderia ambifaria AMMD
Burkholderia cepacia GG4
Burkholderia sp. KJ006
Burkholderia vietnamiensis G4
Burkholderia sp. TJI49
Burkholderia dolosa AUO158
Burkholderia multivorans ATCC 17616
Burkholderia pyrrocinia CH-67
Burkholderia ubonensis
Burkholderia oklahomensis C6786
Burkholderia thailandensis E264
Burkholderia mallei ATCC 23344
Burkholderia pseudomallei K96243
Burkholderia glumae BGR1
Burkholderia gladioli BSR3
Burkholderia sp. SJ98
Burkholderia sp. YI23
Burkholderia sp. RPE64
Candidatus Burkholderia kirkii UZHbot1
Burkholderia mimosarum LMG 23256
Burkholderia kururiensis M130
Burkholderia sp. BT03
Burkholderia terrae BS001
Burkholderia phymatum STM815
Burkholderia sp. JPY251
Burkholderia sp. CCGE1002
Burkholderia sp. H160
Burkholderia sp. WSM4176
Burkholderia sprentiae WSM5005
Burkholderia sp. Ch1-1
Burkholderia xenovorans LB400
Burkholderia phytofirmans PsJN
Burkholderia sp. WSM3556
Burkholderia bryophila 376MFSha3.1
Burkholderia sp. URHA0054
Burkholderia graminis C4D1M
Burkholderia phenoliruptrix BR3459a
Burkholderia sp. CCGE1001
Burkholderia sp. WSM2230
Burkholderia sp. WSM2232
Burkholderia sp. CCGE1003
Burkholderia sp. JPY347
Burkholderia rhizoxinica HKI 454
B. cepacia complex
Clade Ia
B. pseudomallei group
Clade Ib
Phytopathogens
Clade Ic
Burkholderia
Clade I
Clade IIa
Clade IIb
Paraburkholderia
Clade II

FIGURE 6 | A summary diagram depicting the distribution of identified CSIs and the proposed names of the two major groups (Clade I and II) within *Burkholderia*. The major *Burkholderia* clades are indicated by brackets and highlighting.

provide novel and useful targets for the development of diagnostic assays for either the BCC or the *B. pseudomallei* group (Ahmod et al., 2011; Wong et al., 2014). We have identified CSIs for three other groups within the genus *Burkholderia*: A group of plant pathogenic *Burkholderia* related to the BCC and *B. pseudomallei* group (Clade Ic), a group containing unnamed and candidate *Burkholderia* species (Clade IIa), and a group consisting of environmental *Burkholderia* (Clade IIb). We have identified 6, 16, and 6 CSIs for these three groups, respectively. These CSIs provide important differentiating characteristics for these groups, particularly for Clades IIa and IIb which are related groups that have no other identified differentiating characteristics (Suarez-Moreno et al., 2012).

The phylogenetic analyses, identified CSIs, and the pathogenic characteristics of the different *Burkholderia* species presented in this work strongly suggest that the genus *Burkholderia* is made up of at least two distinct lineages. One lineage consisting of the BCC and related organisms (Clade I) and another consisting of a wide range of environmental organisms (Clade II). This latter clade is phylogenetically highly diverse and there is a paucity of sequence information available for its members. Thus, it is possible that in future this latter clade may be found to consist of more than one distinct bacterial lineage, however, it is currently clear that Clade I and Clade II represent distinct lineages. Evidence for the distinctness of the Clade I members from other *Burkholderia* species has been identified in

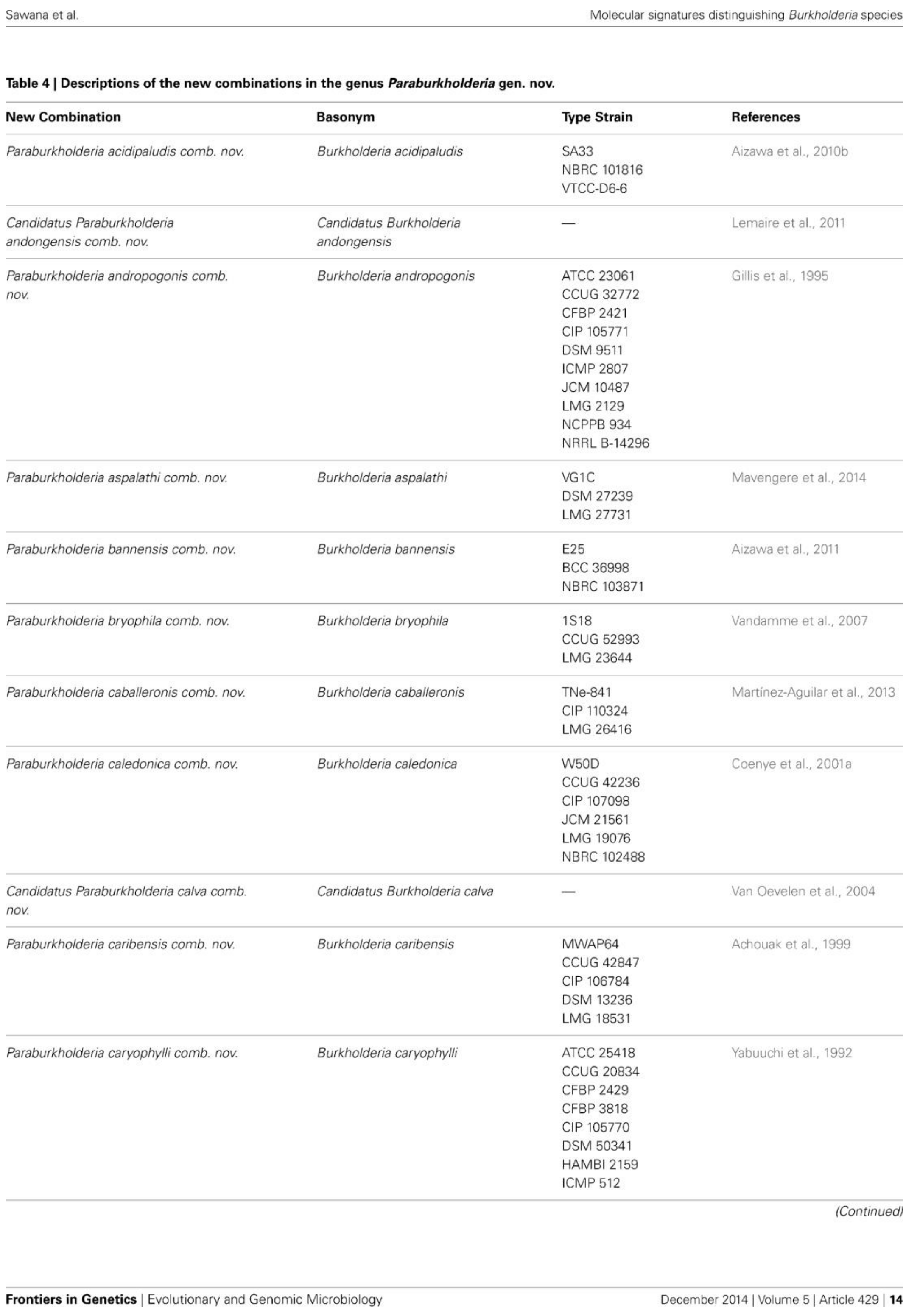


Table 4 | Descriptions of the new combinations in the genus *Paraburkholderia* gen. nov.

New Combination	Basonym	Type Strain	References
Paraburkholderia acidipaludis comb. nov.	*Burkholderia acidipaludis*	SA33 NBRC 101816 VTCC-D6-6	Aizawa et al., 2010b
Candidatus Paraburkholderia andongensis comb. nov.	*Candidatus Burkholderia andongensis*	—	Lemaire et al., 2011
Paraburkholderia andropogonis comb. nov.	*Burkholderia andropogonis*	ATCC 23061 CCUG 32772 CFBP 2421 CIP 105771 DSM 9511 ICMP 2807 JCM 10487 LMG 2129 NCPPB 934 NRRL B-14296	Gillis et al., 1995
Paraburkholderia aspalathi comb. nov.	*Burkholderia aspalathi*	VG1C DSM 27239 LMG 27731	Mavengere et al., 2014
Paraburkholderia bannensis comb. nov.	*Burkholderia bannensis*	E25 BCC 36998 NBRC 103871	Aizawa et al., 2011
Paraburkholderia bryophila comb. nov.	*Burkholderia bryophila*	1S18 CCUG 52993 LMG 23644	Vandamme et al., 2007
Paraburkholderia caballeronis comb. nov.	*Burkholderia caballeronis*	TNe-841 CIP 110324 LMG 26416	Martínez-Aguilar et al., 2013
Paraburkholderia caledonica comb. nov.	*Burkholderia caledonica*	W50D CCUG 42236 CIP 107098 JCM 21561 LMG 19076 NBRC 102488	Coenye et al., 2001a
Candidatus Paraburkholderia calva comb. nov.	*Candidatus Burkholderia calva*	—	Van Oevelen et al., 2004
Paraburkholderia caribensis comb. nov.	*Burkholderia caribensis*	MWAP64 CCUG 42847 CIP 106784 DSM 13236 LMG 18531	Achouak et al., 1999
Paraburkholderia caryophylli comb. nov.	*Burkholderia caryophylli*	ATCC 25418 CCUG 20834 CFBP 2429 CFBP 3818 CIP 105770 DSM 50341 HAMBI 2159 ICMP 512	Yabuuchi et al., 1992

(Continued)

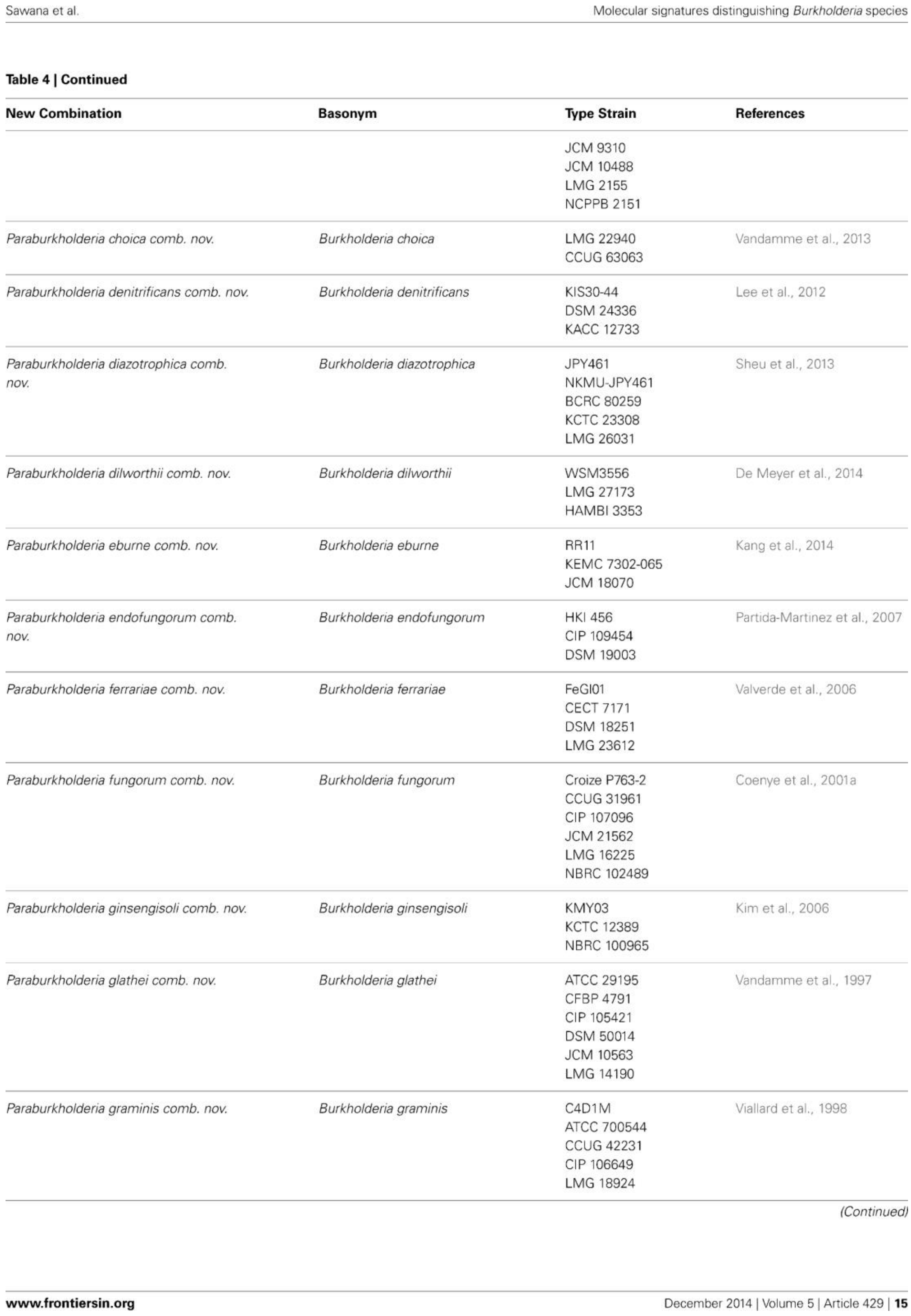


Table 4 | Continued

New Combination	Basonym	Type Strain	References
		JCM 9310 JCM 10488 LMG 2155 NCPPB 2151	
Paraburkholderia choica comb. nov.	*Burkholderia choica*	LMG 22940 CCUG 63063	Vandamme et al., 2013
Paraburkholderia denitrificans comb. nov.	*Burkholderia denitrificans*	KIS30-44 DSM 24336 KACC 12733	Lee et al., 2012
Paraburkholderia diazotrophica comb. nov.	*Burkholderia diazotrophica*	JPY461 NKMU-JPY461 BCRC 80259 KCTC 23308 LMG 26031	Sheu et al., 2013
Paraburkholderia dilworthii comb. nov.	*Burkholderia dilworthii*	WSM3556 LMG 27173 HAMBI 3353	De Meyer et al., 2014
Paraburkholderia eburne comb. nov.	*Burkholderia eburne*	RR11 KEMC 7302-065 JCM 18070	Kang et al., 2014
Paraburkholderia endofungorum comb. nov.	*Burkholderia endofungorum*	HKI 456 CIP 109454 DSM 19003	Partida-Martinez et al., 2007
Paraburkholderia ferrariae comb. nov.	*Burkholderia ferrariae*	FeGI01 CECT 7171 DSM 18251 LMG 23612	Valverde et al., 2006
Paraburkholderia fungorum comb. nov.	*Burkholderia fungorum*	Croize P763-2 CCUG 31961 CIP 107096 JCM 21562 LMG 16225 NBRC 102489	Coenye et al., 2001a
Paraburkholderia ginsengisoli comb. nov.	*Burkholderia ginsengisoli*	KMY03 KCTC 12389 NBRC 100965	Kim et al., 2006
Paraburkholderia glathei comb. nov.	*Burkholderia glathei*	ATCC 29195 CFBP 4791 CIP 105421 DSM 50014 JCM 10563 LMG 14190	Vandamme et al., 1997
Paraburkholderia graminis comb. nov.	*Burkholderia graminis*	C4D1M ATCC 700544 CCUG 42231 CIP 106649 LMG 18924	Viallard et al., 1998

(Continued)

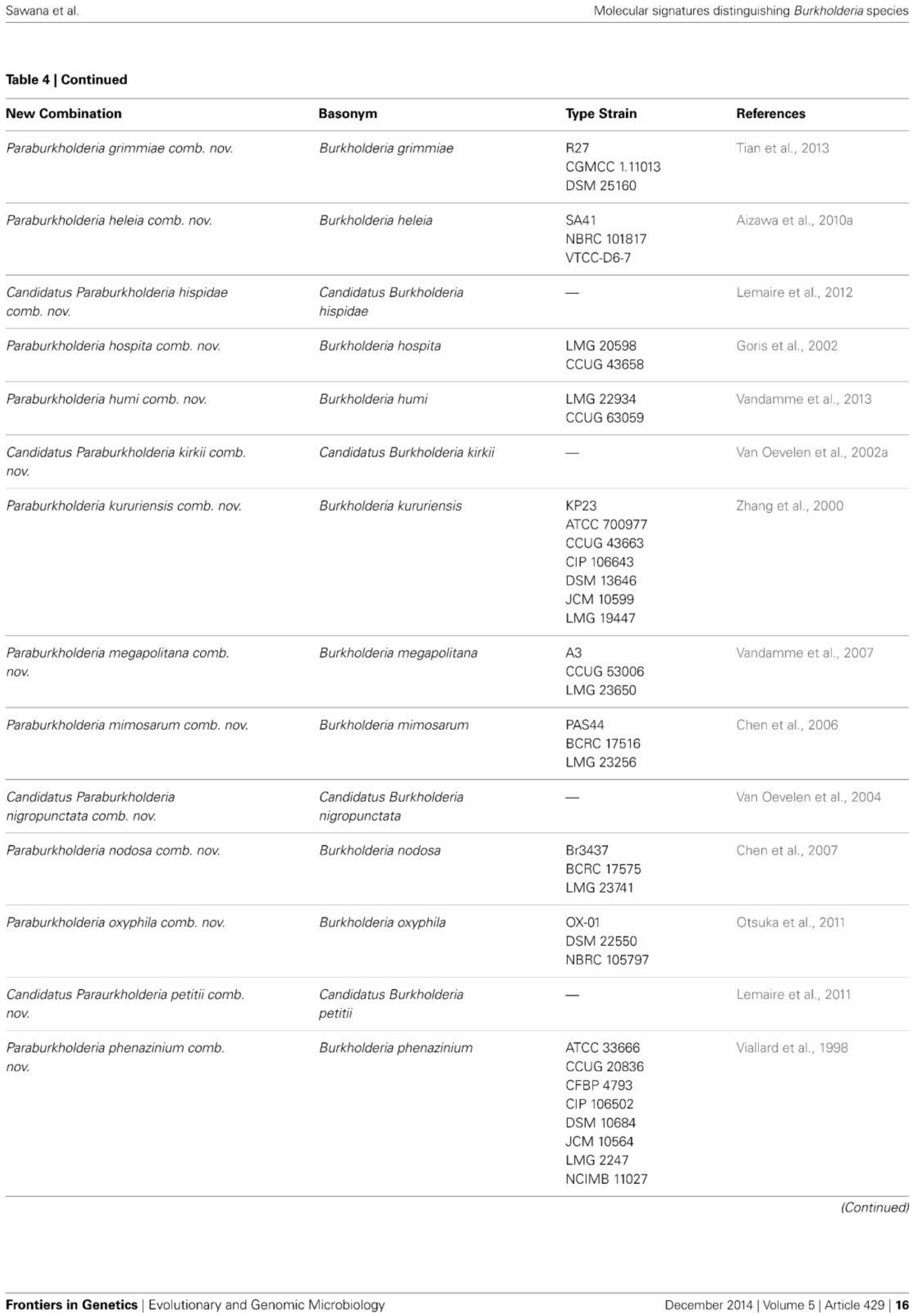


Table 4 | Continued

New Combination	Basonym	Type Strain	References
Paraburkholderia grimmiae comb. nov.	*Burkholderia grimmiae*	R27 CGMCC 1.11013 DSM 25160	Tian et al., 2013
Paraburkholderia heleia comb. nov.	*Burkholderia heleia*	SA41 NBRC 101817 VTCC-D6-7	Aizawa et al., 2010a
Candidatus Paraburkholderia hispidae comb. nov.	*Candidatus Burkholderia hispidae*	—	Lemaire et al., 2012
Paraburkholderia hospita comb. nov.	*Burkholderia hospita*	LMG 20598 CCUG 43658	Goris et al., 2002
Paraburkholderia humi comb. nov.	*Burkholderia humi*	LMG 22934 CCUG 63059	Vandamme et al., 2013
Candidatus Paraburkholderia kirkii comb. nov.	*Candidatus Burkholderia kirkii*	—	Van Oevelen et al., 2002a
Paraburkholderia kururiensis comb. nov.	*Burkholderia kururiensis*	KP23 ATCC 700977 CCUG 43663 CIP 106643 DSM 13646 JCM 10599 LMG 19447	Zhang et al., 2000
Paraburkholderia megapolitana comb. nov.	*Burkholderia megapolitana*	A3 CCUG 53006 LMG 23650	Vandamme et al., 2007
Paraburkholderia mimosarum comb. nov.	*Burkholderia mimosarum*	PAS44 BCRC 17516 LMG 23256	Chen et al., 2006
Candidatus Paraburkholderia nigropunctata comb. nov.	*Candidatus Burkholderia nigropunctata*	—	Van Oevelen et al., 2004
Paraburkholderia nodosa comb. nov.	*Burkholderia nodosa*	Br3437 BCRC 17575 LMG 23741	Chen et al., 2007
Paraburkholderia oxyphila comb. nov.	*Burkholderia oxyphila*	OX-01 DSM 22550 NBRC 105797	Otsuka et al., 2011
Candidatus Paraurkholderia petitii comb. nov.	*Candidatus Burkholderia petitii*	—	Lemaire et al., 2011
Paraburkholderia phenazinium comb. nov.	*Burkholderia phenazinium*	ATCC 33666 CCUG 20836 CFBP 4793 CIP 106502 DSM 10684 JCM 10564 LMG 2247 NCIMB 11027	Viallard et al., 1998

(Continued)

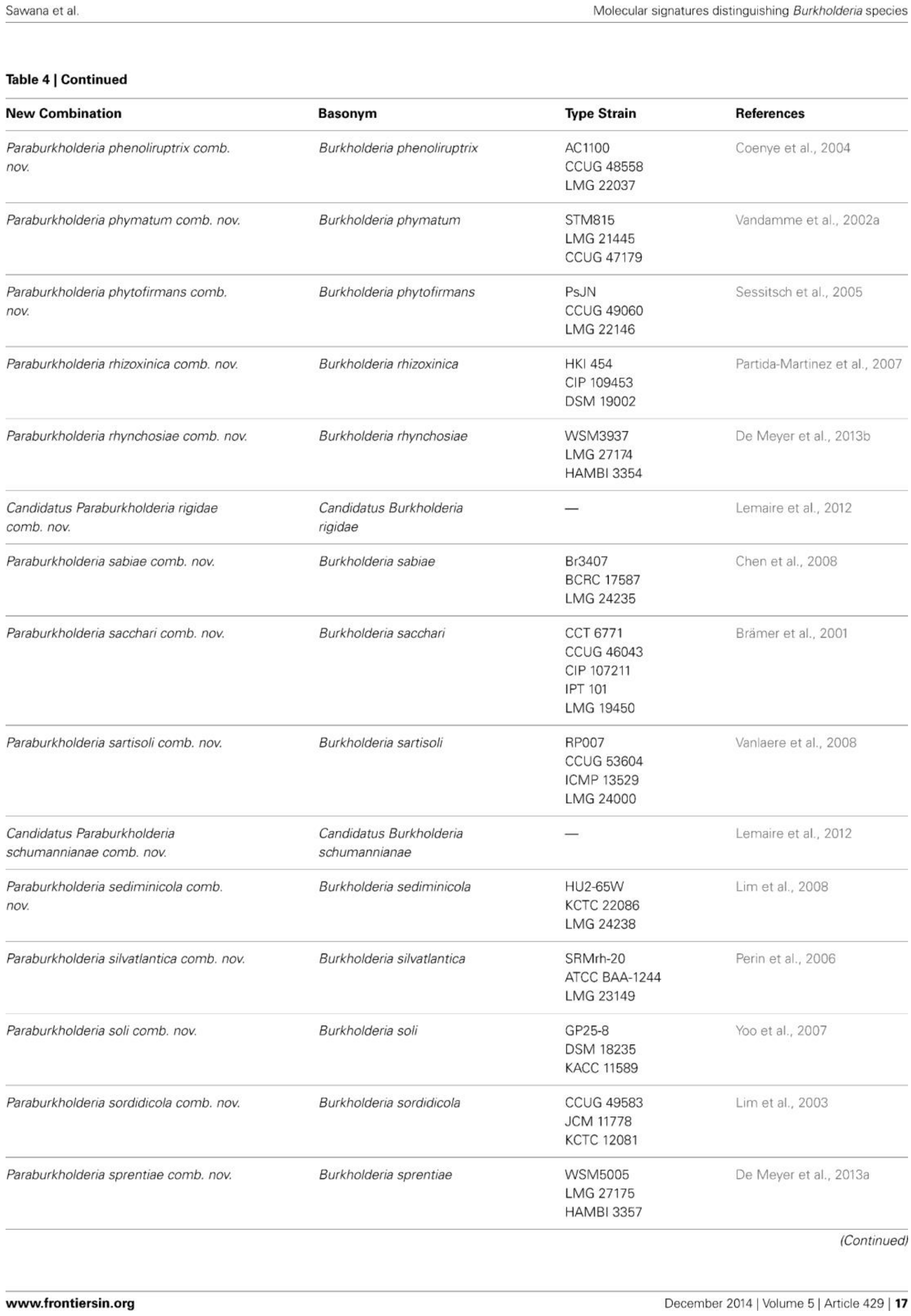


Table 4 | Continued

New Combination	Basonym	Type Strain	References
Paraburkholderia phenoliruptrix comb. nov.	*Burkholderia phenoliruptrix*	AC1100 CCUG 48558 LMG 22037	Coenye et al., 2004
Paraburkholderia phymatum comb. nov.	*Burkholderia phymatum*	STM815 LMG 21445 CCUG 47179	Vandamme et al., 2002a
Paraburkholderia phytofirmans comb. nov.	*Burkholderia phytofirmans*	PsJN CCUG 49060 LMG 22146	Sessitsch et al., 2005
Paraburkholderia rhizoxinica comb. nov.	*Burkholderia rhizoxinica*	HKI 454 CIP 109453 DSM 19002	Partida-Martinez et al., 2007
Paraburkholderia rhynchosiae comb. nov.	*Burkholderia rhynchosiae*	WSM3937 LMG 27174 HAMBI 3354	De Meyer et al., 2013b
Candidatus Paraburkholderia rigidae comb. nov.	*Candidatus Burkholderia rigidae*	—	Lemaire et al., 2012
Paraburkholderia sabiae comb. nov.	*Burkholderia sabiae*	Br3407 BCRC 17587 LMG 24235	Chen et al., 2008
Paraburkholderia sacchari comb. nov.	*Burkholderia sacchari*	CCT 6771 CCUG 46043 CIP 107211 IPT 101 LMG 19450	Brämer et al., 2001
Paraburkholderia sartisoli comb. nov.	*Burkholderia sartisoli*	RP007 CCUG 53604 ICMP 13529 LMG 24000	Vanlaere et al., 2008
Candidatus Paraburkholderia schumannianae comb. nov.	*Candidatus Burkholderia schumannianae*	—	Lemaire et al., 2012
Paraburkholderia sediminicola comb. nov.	*Burkholderia sediminicola*	HU2-65W KCTC 22086 LMG 24238	Lim et al., 2008
Paraburkholderia silvatlantica comb. nov.	*Burkholderia silvatlantica*	SRMrh-20 ATCC BAA-1244 LMG 23149	Perin et al., 2006
Paraburkholderia soli comb. nov.	*Burkholderia soli*	GP25-8 DSM 18235 KACC 11589	Yoo et al., 2007
Paraburkholderia sordidicola comb. nov.	*Burkholderia sordidicola*	CCUG 49583 JCM 11778 KCTC 12081	Lim et al., 2003
Paraburkholderia sprentiae comb. nov.	*Burkholderia sprentiae*	WSM5005 LMG 27175 HAMBI 3357	De Meyer et al., 2013a

(Continued)

Table 4 | Continued

New Combination	Basonym	Type Strain	References
Paraburkholderia symbiotica comb. nov.	*Burkholderia symbiotica*	JPY-345 NKMU-JPY-345 BCRC 80258 KCTC 23309 LMG 26032	Sheu et al., 2012
Paraburkholderia telluris comb. nov.	*Burkholderia telluris*	LMG 22936 CCUG 63060	Vandamme et al., 2013
Paraburkholderia terrae comb. nov.	*Burkholderia terrae*	KMY02 KCTC 12388 NBRC 100964	Yang et al., 2006
Paraburkholderia terrestris comb. nov.	*Burkholderia terrestris*	LMG 22937 CCUG 63062	Vandamme et al., 2013
Paraburkholderia terricola comb. nov.	*Burkholderia terricola*	CCUG 44527 LMG 20594	Goris et al., 2002
Paraburkholderia tropica comb. nov.	*Burkholderia tropica*	Ppe8 ATCC BAA-831 DSM 15359 LMG 22274	Reis et al., 2004
Paraburkholderia tuberum comb. nov.	*Burkholderia tuberum*	STM678 CCUG 47178 LMG 21444	Vandamme et al., 2002a
Paraburkholderia udeis comb. nov.	*Burkholderia udeis*	LMG 27134 CCUG 63061	Vandamme et al., 2013
Paraburkholderia unamae comb. nov.	*Burkholderia unamae*	MTI-641 ATCC BAA-744 CIP 107921	Caballero-Mellado et al., 2004
Paraburkholderia xenovorans comb. nov.	*Burkholderia xenovorans*	LB400 CCUG 46959 LMG 21463 NRRL B-18064	Goris et al., 2004
Paraburkholderia zhejiangensis comb. nov.	*Burkholderia zhejiangensis*	OP-1 KCTC 23300	Lu et al., 2012

a number of previous phylogenetic studies (Payne et al., 2005; Tayeb et al., 2008; Yarza et al., 2008; Spilker et al., 2009; Ussery et al., 2009; Gyaneshwar et al., 2011; Vandamme and Dawyndt, 2011; Zhu et al., 2011; Suarez-Moreno et al., 2012; Estrada-de los Santos et al., 2013; Segata et al., 2013). Estrada-de los Santos et al. (2013) recently completed a phylogenetic analysis of the genus *Burkholderia* utilizing the multilocus sequence analysis of *atpD*, *gltB*, *lepA*, and *recA* genes in combination with the 16S rRNA gene, which provides compelling evidence for the presence of two distinct evolutionary lineages within the genus *Burkholderia*. However, these authors have refrained from formally proposing a division of the genus into two genera due to a paucity of differentiating characteristics for the two groups. Our comparative analysis of *Burkholderia* genomes has identified a set of distinctive molecular characteristics that clearly differentiate the two evolutionary lineages within the genus *Burkholderia* in addition the phylogenetic evidence. In light of the abundance of phylogenetic and molecular evidence for the presence of two distinct evolutionary lineages within the genus *Burkholderia,* and the distinct pathogenicity profiles of the members of these two groups, we are proposing that genus *Burkholderia* should be divided into two separate genera. The first of these monophyletic genera, which comprises of all the clinically relevant species and clearly distinguished from all other *Burkholderia* species, will retain the name *Burkholderia* (Clade I). For the remainder of the *Burkholderia* species (Clade II), which include a wide range of environmental species, we propose the name *Paraburkholderia* gen. nov. An emended description of the genus *Burkholderia* and a description of *Paraburkholderia* gen. nov. are provided below. Brief descriptions of the new species combinations within *Paraburkholderia* gen. nov. are presented in **Table 4**.

EMENDED DESCRIPTION OF THE GENUS *BURKHOLDERIA* (Yabuuchi et al., 1993 EMEND. Gillis et al., 1995)

The genus contains the type species *B. cepacia* (Yabuuchi et al., 1993). The species from this genus are gram-negative, straight or slightly curved rods, which exhibit motility mediated by one or more polar flagella. Only, *B. mallei* lacks flagella and is non-motile. The species do not produce sheaths or prosthecae and do not go through any resting stages. Most species are able to accumulate and utilize poly-β-hydroxybutyrate (PHB) for growth. The species are mostly aerobic chemoorganotrophs, but some species are capable of anaerobic respiration using nitrate as the terminal electron acceptor. The G+C content for the members of the genus ranges from 65.7 to 68.5%. The members of the genus form a distinct monophyletic clade in phylogenetic trees, and they are distinguished from all other bacteria by the conserved sequence indels reported in this work in the following proteins: Periplasmic amino acid-binding protein, 4-hydroxybenzoate 3-monooxygenase, 6-phosphogluconate dehydrogenase, Sarcosine oxidase subunit alpha, a putative lipoprotein, and a putative lyase (**Table 2**).

DESCRIPTION OF THE GENUS *PARABURKHOLDERIA* GEN. NOV.

The genus contains the type species *Paraburkholderia graminis* comb. nov. (Basonym: *Burkholderia graminis*, Viallard et al., 1998) The species from this genus are gram-negative straight or slightly curved rods with one or more polar flagella. Other morphological and metabolic characteristics are similar to genus *Burkholderia.* The G+C content for the members of the genus ranges from 61.4 to 65.0%. The species are not associated with humans. The members of this genus generally form a distinct clade in the neighborhood of genus *Burkholderia* in phylogenetic trees, and they lack the molecular signatures which are specific for *Burkholderia*. Most of the sequenced members from this genus contain the conserved sequence indels reported in this work in the protein sequences of an unnamed dehydrogenase and a LysR family transcriptional regulator (**Table 2**).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00429/abstract

REFERENCES

Achouak, W., Christen, R., Barakat, M., Martel, M.-H., and Heulin, T. (1999). *Burkholderia caribensis* sp. nov., an exopolysaccharide-producing bacterium isolated from vertisol microaggregates in Martinique. *Int. J. Syst. Bacteriol.* 49, 787–794. doi: 10.1099/00207713-49-2-787

Ahmod, N. Z., Gupta, R. S., and Shah, H. N. (2011). Identification of a *Bacillus anthracis* specific indel in the *yeaC* gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. *J. Microbiol. Methods* 87, 278–285. doi: 10.1016/j.mimet.2011.08.015

Aizawa, T., Ve, N. B., Nakajima, M., and Sunairi, M. (2010a). *Burkholderia heleia* sp. nov., a nitrogen-fixing bacterium isolated from an aquatic plant, *Eleocharis dulcis*, that grows in highly acidic swamps in actual acid sulfate soil areas of Vietnam. *Int. J. Syst. Evol. Microbiol.* 60, 1152–1157. doi: 10.1099/ijs.0.015198-0

Aizawa, T., Ve, N. B., Vijarnsorn, P., Nakajima, M., and Sunairi, M. (2010b). *Burkholderia acidipaludis* sp. nov., aluminium-tolerant bacteria isolated from Chinese water chestnut (*Eleocharis dulcis*) growing in highly acidic swamps in South-East Asia. *Int. J. Syst. Evol. Microbiol.* 60, 2036–2041. doi: 10.1099/ijs.0.018283-0

Aizawa, T., Vijarnsorn, P., Nakajima, M., and Sunairi, M. (2011). *Burkholderia bannensis* sp. nov., an acid-neutralizing bacterium isolated from torpedo grass (*Panicum repens*) growing in highly acidic swamps. *Int. J. Syst. Evol. Microbiol.* 61, 1645–1650. doi: 10.1099/ijs.0.026278-0

Biddick, R., Spilker, T., Martin, A., and LiPuma, J. J. (2003). Evidence of transmission of *Burkholderia cepacia*, *Burkholderia multivorans* and *Burkholderia dolosa* among persons with cystic fibrosis. *FEMS Microbiol. Lett.* 228, 57–62. doi: 10.1016/S0378-1097(03)00724-9

Brämer, C. O., Vandamme, P., da Silva, L. F., Gomez, J., and Steinbüchel, A. (2001). Polyhydroxyalkanoate-accumulating bacterium isolated from soil of a sugar-cane plantation in Brazil. *Int. J. Syst. Evol. Microbiol.* 51, 1709–1713. doi: 10.1099/00207713-51-5-1709

Caballero-Mellado, J., Martínez-Aguilar, L., Paredes-Valdez, G., and Estrada-de los Santos, P. (2004). *Burkholderia unamae* sp. nov., an N2-fixing rhizospheric and endophytic species. *Int. J. Syst. Evol. Microbiol.* 54, 1165–1172. doi: 10.1099/ijs.0.02951-0

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chain, P. S., Denef, V. J., Konstantinidis, K. T., Vergez, L. M., Agullo, L., Reyes, V. L., et al. (2006). *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15280–15287. doi: 10.1073/pnas.0606924103

Charlebois, R. L., and Doolittle, W. F. (2004). Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 14, 2469–2477. doi: 10.1101/gr.3024704

Chen, W.-M., de Faria, S. M., Chou, J.-H., James, E. K., Elliott, G. N., Sprent, J. I., et al. (2008). *Burkholderia sabiae* sp. nov., isolated from root nodules of *Mimosa caesalpiniifolia. Int. J. Syst. Evol. Microbiol.* 58, 2174–2179. doi: 10.1099/ijs.0.65816-0

Chen, W.-M., De Faria, S. M., James, E. K., Elliott, G. N., Lin, K.-Y., Chou, J.-H., et al. (2007). *Burkholderia nodosa* sp. nov., isolated from root nodules of the woody Brazilian legumes *Mimosa bimucronata* and *Mimosa scabrella. Int. J. Syst. Evol. Microbiol.* 57, 1055–1059. doi: 10.1099/ijs.0.64873-0

Chen, W.-M., James, E. K., Coenye, T., Chou, J.-H., Barrios, E., De Faria, S. M., et al. (2006). *Burkholderia mimosarum* sp. nov., isolated from root nodules of Mimosa spp. from Taiwan and South America. *Int. J. Syst. Evol. Microbiol.* 56, 1847–1851. doi: 10.1099/ijs.0.64325-0

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. doi: 10.1126/science.1123061

Coenye, T., Henry, D., Speert, D. P., and Vandamme, P. (2004). *Burkholderia phenoliruptrix* sp. nov., to accommodate the 2, 4, 5-trichlorophenoxyacetic acid and halophenol-degrading strain AC1100. *Syst. Appl. Microbiol.* 27, 623–627. doi: 10.1078/0723202042369992

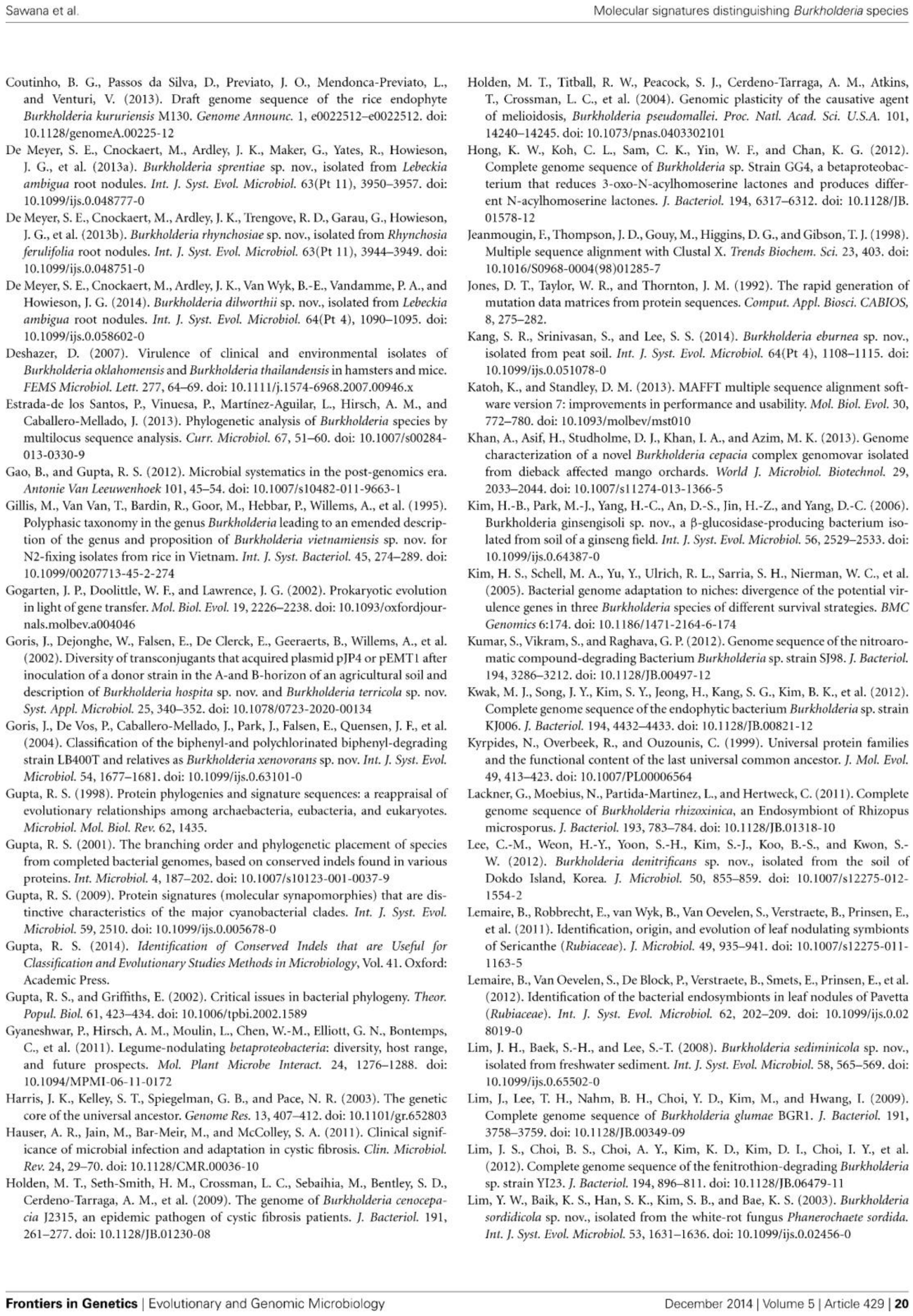
Coenye, T., Laevens, S., Willems, A., Ohlén, M., Hannant, W., Govan, J., et al. (2001a). *Burkholderia fungorum* sp. nov. and *Burkholderia caledonica* sp. nov., *two* new species isolated from the environment, animals and human clinical samples. *Int. J. Syst. Evol. Microbiol.* 51, 1099–1107. doi: 10.1099/00207713-51-3-1099

Coenye, T., Mahenthiralingam, E., Henry, D., LiPuma, J. J., Laevens, S., Gillis, M., et al. (2001b). *Burkholderia ambifaria* sp. nov., a novel member of the *Burkholderia cepacia* complex including biocontrol and cystic fibrosis-related isolates. *Int. J. Syst. Evol. Microbiol.* 51, 1481–1490. doi: 10.1099/00207713-51-4-1481

Coenye, T., and Vandamme, P. (2003). Diversity and significance of Burkholderia species occupying diverse ecological niches. *Environ. Microbiol.* 5, 719–729. doi: 10.1046/j.1462-2920.2003.00471.x

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–D145. doi: 10.1093/nar/gkn879

Compant, S., Nowak, J., Coenye, T., Clement, C., and Ait Barka, E. (2008). Diversity and occurrence of *Burkholderia* spp. in the natural environment. *FEMS Microbiol. Rev.* 32, 607–626. doi: 10.1111/j.1574-6976.2008.00113.x



Coutinho, B. G., Passos da Silva, D., Previato, J. O., Mendonca-Previato, L., and Venturi, V. (2013). Draft genome sequence of the rice endophyte *Burkholderia kururiensis* M130. *Genome Announc.* 1, e0022512–e0022512. doi: 10.1128/genomeA.00225-12

De Meyer, S. E., Cnockaert, M., Ardley, J. K., Maker, G., Yates, R., Howieson, J. G., et al. (2013a). *Burkholderia sprentiae* sp. nov., isolated from *Lebeckia ambigua* root nodules. *Int. J. Syst. Evol. Microbiol.* 63(Pt 11), 3950–3957. doi: 10.1099/ijs.0.048777-0

De Meyer, S. E., Cnockaert, M., Ardley, J. K., Trengove, R. D., Garau, G., Howieson, J. G., et al. (2013b). *Burkholderia rhynchosiae* sp. nov., isolated from *Rhynchosia ferulifolia* root nodules. *Int. J. Syst. Evol. Microbiol.* 63(Pt 11), 3944–3949. doi: 10.1099/ijs.0.048751-0

De Meyer, S. E., Cnockaert, M., Ardley, J. K., Van Wyk, B.-E., Vandamme, P. A., and Howieson, J. G. (2014). *Burkholderia dilworthii* sp. nov., isolated from *Lebeckia ambigua* root nodules. *Int. J. Syst. Evol. Microbiol.* 64(Pt 4), 1090–1095. doi: 10.1099/ijs.0.058602-0

Deshazer, D. (2007). Virulence of clinical and environmental isolates of *Burkholderia oklahomensis* and *Burkholderia thailandensis* in hamsters and mice. *FEMS Microbiol. Lett.* 277, 64–69. doi: 10.1111/j.1574-6968.2007.00946.x

Estrada-de los Santos, P., Vinuesa, P., Martínez-Aguilar, L., Hirsch, A. M., and Caballero-Mellado, J. (2013). Phylogenetic analysis of *Burkholderia* species by multilocus sequence analysis. *Curr. Microbiol.* 67, 51–60. doi: 10.1007/s00284-013-0330-9

Gao, B., and Gupta, R. S. (2012). Microbial systematics in the post-genomics era. *Antonie Van Leeuwenhoek* 101, 45–54. doi: 10.1007/s10482-011-9663-1

Gillis, M., Van Van, T., Bardin, R., Goor, M., Hebbar, P., Willems, A., et al. (1995). Polyphasic taxonomy in the genus *Burkholderia* leading to an emended description of the genus and proposition of *Burkholderia vietnamiensis* sp. nov. for N2-fixing isolates from rice in Vietnam. *Int. J. Syst. Bacteriol.* 45, 274–289. doi: 10.1099/00207713-45-2-274

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238. doi: 10.1093/oxfordjournals.molbev.a004046

Goris, J., Dejonghe, W., Falsen, E., De Clerck, E., Geeraerts, B., Willems, A., et al. (2002). Diversity of transconjugants that acquired plasmid pJP4 or pEMT1 after inoculation of a donor strain in the A-and B-horizon of an agricultural soil and description of *Burkholderia hospita* sp. nov. and *Burkholderia terricola* sp. nov. *Syst. Appl. Microbiol.* 25, 340–352. doi: 10.1078/0723-2020-00134

Goris, J., De Vos, P., Caballero-Mellado, J., Park, J., Falsen, E., Quensen, J. F., et al. (2004). Classification of the biphenyl-and polychlorinated biphenyl-degrading strain LB400T and relatives as *Burkholderia xenovorans* sp. nov. *Int. J. Syst. Evol. Microbiol.* 54, 1677–1681. doi: 10.1099/ijs.0.63101-0

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.

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. doi: 10.1007/s10123-001-0037-9

Gupta, R. S. (2009). Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. *Int. J. Syst. Evol. Microbiol.* 59, 2510. doi: 10.1099/ijs.0.005678-0

Gupta, R. S. (2014). *Identification of Conserved Indels that are Useful for Classification and Evolutionary Studies Methods in Microbiology*, Vol. 41. Oxford: Academic Press.

Gupta, R. S., and Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* 61, 423–434. doi: 10.1006/tpbi.2002.1589

Gyaneshwar, P., Hirsch, A. M., Moulin, L., Chen, W.-M., Elliott, G. N., Bontemps, C., et al. (2011). Legume-nodulating *betaproteobacteria*: diversity, host range, and future prospects. *Mol. Plant Microbe Interact.* 24, 1276–1288. doi: 10.1094/MPMI-06-11-0172

Harris, J. K., Kelley, S. T., Spiegelman, G. B., and Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Res.* 13, 407–412. doi: 10.1101/gr.652803

Hauser, A. R., Jain, M., Bar-Meir, M., and McColley, S. A. (2011). Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev.* 24, 29–70. doi: 10.1128/CMR.00036-10

Holden, M. T., Seth-Smith, H. M., Crossman, L. C., Sebaihia, M., Bentley, S. D., Cerdeno-Tarraga, A. M., et al. (2009). The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *J. Bacteriol.* 191, 261–277. doi: 10.1128/JB.01230-08

Holden, M. T., Titball, R. W., Peacock, S. J., Cerdeno-Tarraga, A. M., Atkins, T., Crossman, L. C., et al. (2004). Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14240–14245. doi: 10.1073/pnas.0403302101

Hong, K. W., Koh, C. L., Sam, C. K., Yin, W. F., and Chan, K. G. (2012). Complete genome sequence of *Burkholderia* sp. Strain GG4, a betaproteobacterium that reduces 3-oxo-N-acylhomoserine lactones and produces different N-acylhomoserine lactones. *J. Bacteriol.* 194, 6317–6312. doi: 10.1128/JB.01578-12

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. doi: 10.1016/S0968-0004(98)01285-7

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

Kang, S. R., Srinivasan, S., and Lee, S. S. (2014). *Burkholderia eburnea* sp. nov., isolated from peat soil. *Int. J. Syst. Evol. Microbiol.* 64(Pt 4), 1108–1115. doi: 10.1099/ijs.0.051078-0

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Khan, A., Asif, H., Studholme, D. J., Khan, I. A., and Azim, M. K. (2013). Genome characterization of a novel *Burkholderia cepacia* complex genomovar isolated from dieback affected mango orchards. *World J. Microbiol. Biotechnol.* 29, 2033–2044. doi: 10.1007/s11274-013-1366-5

Kim, H.-B., Park, M.-J., Yang, H.-C., An, D.-S., Jin, H.-Z., and Yang, D.-C. (2006). Burkholderia ginsengisoli sp. nov., a β-glucosidase-producing bacterium isolated from soil of a ginseng field. *Int. J. Syst. Evol. Microbiol.* 56, 2529–2533. doi: 10.1099/ijs.0.64387-0

Kim, H. S., Schell, M. A., Yu, Y., Ulrich, R. L., Sarria, S. H., Nierman, W. C., et al. (2005). Bacterial genome adaptation to niches: divergence of the potential virulence genes in three *Burkholderia* species of different survival strategies. *BMC Genomics* 6:174. doi: 10.1186/1471-2164-6-174

Kumar, S., Vikram, S., and Raghava, G. P. (2012). Genome sequence of the nitroaromatic compound-degrading Bacterium *Burkholderia* sp. strain SJ98. *J. Bacteriol.* 194, 3286–3212. doi: 10.1128/JB.00497-12

Kwak, M. J., Song, J. Y., Kim, S. Y., Jeong, H., Kang, S. G., Kim, B. K., et al. (2012). Complete genome sequence of the endophytic bacterium *Burkholderia* sp. strain KJ006. *J. Bacteriol.* 194, 4432–4433. doi: 10.1128/JB.00821-12

Kyrpides, N., Overbeek, R., and Ouzounis, C. (1999). Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* 49, 413–423. doi: 10.1007/PL00006564

Lackner, G., Moebius, N., Partida-Martinez, L., and Hertweck, C. (2011). Complete genome sequence of *Burkholderia rhizoxinica*, an Endosymbiont of Rhizopus microsporus. *J. Bacteriol.* 193, 783–784. doi: 10.1128/JB.01318-10

Lee, C.-M., Weon, H.-Y., Yoon, S.-H., Kim, S.-J., Koo, B.-S., and Kwon, S.-W. (2012). *Burkholderia denitrificans* sp. nov., isolated from the soil of Dokdo Island, Korea. *J. Microbiol.* 50, 855–859. doi: 10.1007/s12275-012-1554-2

Lemaire, B., Robbrecht, E., van Wyk, B., Van Oevelen, S., Verstraete, B., Prinsen, E., et al. (2011). Identification, origin, and evolution of leaf nodulating symbionts of Sericanthe (*Rubiaceae*). *J. Microbiol.* 49, 935–941. doi: 10.1007/s12275-011-1163-5

Lemaire, B., Van Oevelen, S., De Block, P., Verstraete, B., Smets, E., Prinsen, E., et al. (2012). Identification of the bacterial endosymbionts in leaf nodules of Pavetta (*Rubiaceae*). *Int. J. Syst. Evol. Microbiol.* 62, 202–209. doi: 10.1099/ijs.0.028019-0

Lim, J. H., Baek, S.-H., and Lee, S.-T. (2008). *Burkholderia sediminicola* sp. nov., isolated from freshwater sediment. *Int. J. Syst. Evol. Microbiol.* 58, 565–569. doi: 10.1099/ijs.0.65502-0

Lim, J., Lee, T. H., Nahm, B. H., Choi, Y. D., Kim, M., and Hwang, I. (2009). Complete genome sequence of *Burkholderia glumae* BGR1. *J. Bacteriol.* 191, 3758–3759. doi: 10.1128/JB.00349-09

Lim, J. S., Choi, B. S., Choi, A. Y., Kim, K. D., Kim, D. I., Choi, I. Y., et al. (2012). Complete genome sequence of the fenitrothion-degrading *Burkholderia* sp. strain YI23. *J. Bacteriol.* 194, 896–811. doi: 10.1128/JB.06479-11

Lim, Y. W., Baik, K. S., Han, S. K., Kim, S. B., and Bae, K. S. (2003). *Burkholderia sordidicola* sp. nov., isolated from the white-rot fungus *Phanerochaete sordida*. *Int. J. Syst. Evol. Microbiol.* 53, 1631–1636. doi: 10.1099/ijs.0.02456-0

Limmathurotsakul, D., and Peacock, S. J. (2011). Melioidosis: a clinical overview. *Br. Med. Bull.* 99, 125–139. doi: 10.1093/bmb/ldr007

Lu, P., Zheng, L.-Q., Sun, J.-J., Liu, H.-M., Li, S.-P., Hong, Q., et al. (2012). *Burkholderia zhejiangensis* sp. nov., a methyl-parathion-degrading bacterium isolated from a wastewater-treatment system. *Int. J. Syst. Evol. Microbiol.* 62(Pt 6), 1337–1341. doi: 10.1099/ijs.0.035428-0

Mahenthiralingam, E., Baldwin, A., and Vandamme, P. (2002). *Burkholderia cepacia* complex infection in patients with cystic fibrosis. *J. Med. Microbiol.* 51, 533–538.

Mahenthiralingam, E., Urban, T. A., and Goldberg, J. B. (2005). The multifarious, multireplicon Burkholderia cepacia complex. *Nat. Rev. Microbiol.* 3, 144–156. doi: 10.1038/nrmicro1085

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141. doi: 10.1016/j.tig.2007.12.007

Martínez-Aguilar, L., Salazar-Salazar, C., Méndez, R. D., Caballero-Mellado, J., Hirsch, A. M., Vásquez-Murrieta, M. S., et al. (2013). *Burkholderia caballeronis* sp. nov., a nitrogen fixing species isolated from tomato (*Lycopersicon esculentum*) with the ability to effectively nodulate *Phaseolus vulgaris. Antonie Van Leeuwenhoek* 104, 1063–1071. doi: 10.1007/s10482-013-0028-9

Mavengere, N. R., Ellis, A. G., and Le Roux, J. J. (2014). *Burkholderia aspalathi* sp. nov., isolated from root nodules of the South African legume *Aspalathus abietina* Thunb. *Int. J. Syst. Evol. Microbiol.* 64, 1906–1912. doi: 10.1099/ijs.0.057067-0

Nandakumar, R., Shahjahan, A., Yuan, X., Dickstein, E., Groth, D., Clark, C., et al. (2009). *Burkholderia glumae* and *B. gladioli* cause bacterial panicle blight in rice in the southern United States. *Plant Dis.* 93, 896–905. doi: 10.1094/PDIS-93-9-0896

Nazir, R., Hansen, M. A., Sorensen, S., and van Elsas, J. D. (2012). Draft genome sequence of the soil bacterium *Burkholderia terrae* strain BS001, which interacts with fungal surface structures. *J. Bacteriol.* 194, 4480–4481. doi: 10.1128/J.B.00725-12

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

Nierman, W. C., DeShazer, D., Kim, H. S., Tettelin, H., Nelson, K. E., Feldblyum, T., et al. (2004). Structural flexibility in the *Burkholderia mallei* genome. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14246–14251. doi: 10.1073/pnas.0403 306101

O'Carroll, M. R., Kidd, T. J., Coulter, C., Smith, H. V., Rose, B. R., Harbour, C., et al. (2003). *Burkholderia pseudomallei*: another emerging pathogen in cystic fibrosis. *Thorax* 58, 1087–1091. doi: 10.1136/thorax.58.12.1087

Oliveira Cunha, C., Goda Zuleta, L. F., Paula de Almeida, L. G., Prioli Ciapina, L., Lustrino Borges, W., Pitard, R. M., et al. (2012). Complete genome sequence of *Burkholderia phenoliruptrix* BR3459a (CLA1), a heat-tolerant, nitrogen-fixing symbiont of *Mimosa flocculosa. J. Bacteriol.* 194, 6675–6676. doi: 10.1128/JB.01821-12

Onofre-Lemus, J., Hernández-Lucas, I., Girard, L., and Caballero-Mellado, J. (2009). ACC (1-aminocyclopropane-1-carboxylate) deaminase activity, a widespread trait in *Burkholderia* species, and its growth-promoting effect on tomato plants. *Appl. Environ. Microbiol.* 75, 6581–6590. doi: 10.1128/AEM.01240-09

Ormeno-Orrillo, E., Rogel, M. A., Chueire, L. M., Tiedje, J. M., Martinez-Romero, E., and Hungria, M. (2012). Genome sequences of *Burkholderia* sp. strains CCGE1002 and H160, isolated from legume nodules in Mexico and Brazil. *J. Bacteriol.* 194, 6927–6912. doi: 10.1128/JB.01756-12

Otsuka, Y., Muramatsu, Y., Nakagawa, Y., Matsuda, M., Nakamura, M., and Murata, H. (2011). Burkholderia oxyphila sp. nov., a bacterium isolated from acidic forest soil that catabolizes (+)-catechin and its putative aromatic derivatives. *Int. J. Syst. Evol. Microbiol.* 61, 249–254. doi: 10.1099/ijs.0.017368-0

Palleroni, N. J. (2005). "Genus I. *Burkholderia* Yabuuchi et al. 1993, 398^{VP} (Effective publication: Yabuuchi et al. 1992, 1268) emend. Gillis et al. 1995, 286*," in *Bergey's Manual of Systematic Bacteriology, 2 Edn*, Vol. 2, eds D. J. Brenner, N. R. Krieg, G. M. Garrity, and J. T. Staley (New York, NY: Springer), 575–600.

Parte, A. C. (2013). LPSN–list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res.* 42, D613–D616. doi: 10.1093/nar/gkt1111

Partida-Martinez, L. P., Groth, I., Schmitt, I., Richter, W., Roth, M., and Hertweck, C. (2007). *Burkholderia rhizoxinica* sp. nov. and Burkholderia endofungorum sp. nov., bacterial endosymbionts of the plant-pathogenic fungus *Rhizopus microsporus. Int. J. Syst. Evol. Microbiol.* 57, 2583–2590. doi: 10.1099/ijs.0.64660-0

Payne, G. W., Vandamme, P., Morgan, S. H., LiPuma, J. J., Coenye, T., Weightman, A. J., et al. (2005). Development of a recA gene-based identification approach for the entire *Burkholderia* genus. *Appl. Environ. Microbiol.* 71, 3917–3927. doi: 10.1128/AEM.71.7.3917-3927.2005

Perin, L., Martínez-Aguilar, L., Paredes-Valdez, G., Baldani, J., Estrada-de Los Santos, P., Reis, V., et al. (2006). *Burkholderia silvatlantica* sp. nov., a diazotrophic bacterium associated with sugar cane and maize. *Int. J. Syst. Evol. Microbiol.* 56, 1931–1937. doi: 10.1099/ijs.0.64362-0

Reis, V., Estrada-De los Santos, P., Tenorio-Salgado, S., Vogel, J., Stoffels, M., Guyon, S., et al. (2004). *Burkholderia tropica* sp. nov., a novel nitrogen-fixing, plant-associated bacterium. *Int. J. Syst. Evol. Microbiol.* 54, 2155–2162. doi: 10.1099/ijs.0.02879-0

Rokas, A., and Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* 15, 454–459. doi: 10.1016/S0169-5347(00)01967-4

Segata, N., Bornigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* 4, 2304. doi: 10.1038/ncomms3304

Seo, Y. S., Lim, J., Choi, B. S., Kim, H., Goo, E., Lee, B., et al. (2011). Complete genome sequence of *Burkholderia gladioli* BSR3. *J. Bacteriol.* 193, 3149–3111. doi: 10.1128/JB.00420-11

Sessitsch, A., Coenye, T., Sturz, A., Vandamme, P., Barka, E. A., Salles, J., et al. (2005). *Burkholderia phytofirmans* sp. nov., a novel plant-associated bacterium with plant-beneficial properties. *Int. J. Syst. Evol. Microbiol.* 55, 1187–1192. doi: 10.1099/ijs.0.63149-0

Sheu, S.-Y., Chou, J.-H., Bontemps, C., Elliott, G. N., Gross, E., James, E. K., et al. (2012). *Burkholderia symbiotica* sp. nov., isolated from root nodules of *Mimosa* spp. native to north-east Brazil. *Int. J. Syst. Evol. Microbiol.* 62(Pt 9), 2272–2278. doi: 10.1099/ijs.0.037408-0

Sheu, S.-Y., Chou, J.-H., Bontemps, C., Elliott, G. N., Gross, E., dos Reis Junior, F. B., et al. (2013). *Burkholderia diazotrophica* sp. nov., isolated from root nodules of *Mimosa* spp. *Int. J. Syst. Evol. Microbiol.* 63(Pt 2), 435–441. doi: 10.1099/ijs.0.039859-0

Shibata, T. F., Maeda, T., Nikoh, N., Yamaguchi, K., Oshima, K., Hattori, M., et al. (2013). Complete genome sequence of *Burkholderia* sp. Strain RPE64, bacterial symbiont of the bean bug *Riptortus pedestris*. *Genome Announc.* 1, 10–13. doi: 10.1128/genomeA.00441-13

Simpson, I. N., Finlay, J., Winstanley, D. J., Dewhurst, N., Nelson, J. W., Butler, S. L., et al. (1994). Multi-resistance isolates possessing characteristics of both *Burkholderia* (*Pseudomonas*) *cepacia* and *Burkholderia gladioli* from patients with cystic fibrosis. *J. Antimicrob. Chemother.* 34, 353–361. doi: 10.1093/jac/34.3.353

Song, J. Y., Kwak, M. J., Lee, K. Y., Kong, H. G., Kim, B. K., Kwon, S. K., et al. (2012). Draft genome sequence of the antifungal-producing plant-benefiting bacterium *Burkholderia pyrrocinia* CH-67. *J. Bacteriol.* 194, 6649–6650. doi: 10.1128/JB.01779-12

Spilker, T., Baldwin, A., Bumford, A., Dowson, C. G., Mahenthiralingam, E., and LiPuma, J. J. (2009). Expanded multilocus sequence typing for *Burkholderia* species. *J. Clin. Microbiol.* 47, 2607–2610. doi: 10.1128/JCM.00770-09

Stead, D. (1992). Grouping of plant-pathogenic and some other *Pseudomonas* spp. by using cellular fatty acid profiles. *Int. J. Syst. Bacteriol.* 42, 281–295. doi: 10.1099/00207713-42-2-281

Suarez-Moreno, Z. R., Caballero-Mellado, J., Coutinho, B. G., Mendonca-Previato, L., James, E. K., and Venturi, V. (2012). Common features of environmental and potentially beneficial plant-associated *Burkholderia*. *Microb. Ecol.* 63, 249–266. doi: 10.1007/s00248-011-9929-1

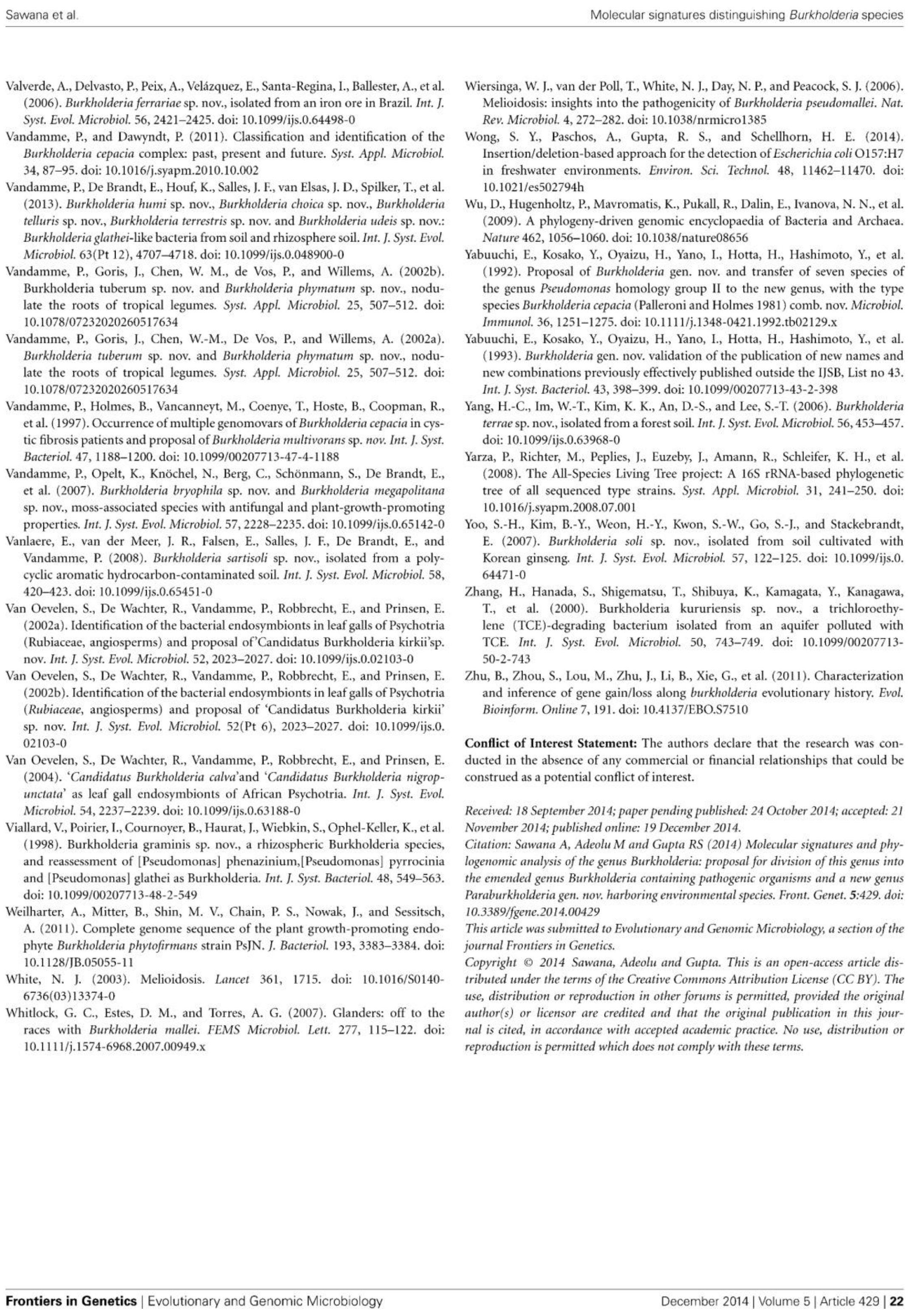
Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tavaré, S. (1986). "Some probabilistic and statistical problems in the analysis of DNA sequences," in *Lectures on Mathematics in the Life Sciences, 17 Edn*, ed R. M. Miura (Providence (RI): American Mathematical Society), 57–86.

Tayeb, L. A., Lefevre, M., Passet, V., Diancourt, L., Brisse, S., and Grimont, P. A. (2008). Comparative phylogenies of *Burkholderia*, *Ralstonia*, *Comamonas*, *Brevundimonas* and related organisms derived from *rpoB*, *gyrB* and *rrs* gene sequences. *Res. Microbiol.* 159, 169–177. doi: 10.1016/j.resmic.2007.12.005

Tian, Y., Kong, B. H., Liu, S. L., Li, C. L., Yu, R., Liu, L., et al. (2013). *Burkholderia grimmiae* sp. nov., isolated from a xerophilous moss (*Grimmia montana*). *Int. J. Syst. Evol. Microbiol.* 63(Pt 6), 2108–2113. doi: 10.1099/ijs.0.045492-0

Ussery, D. W., Kiil, K., Lagesen, K., Sicheritz-Ponten, T., Bohlin, J., and Wassenaar, T. M. (2009). The genus *Burkholderia*: analysis of 56 genomic sequences. *Genome Dyn.* 6, 140–457. doi: 10.1159/000235768



Valverde, A., Delvasto, P., Peix, A., Velázquez, E., Santa-Regina, I., Ballester, A., et al. (2006). *Burkholderia ferrariae* sp. nov., isolated from an iron ore in Brazil. *Int. J. Syst. Evol. Microbiol.* 56, 2421–2425. doi: 10.1099/ijs.0.64498-0

Vandamme, P., and Dawyndt, P. (2011). Classification and identification of the *Burkholderia cepacia* complex: past, present and future. *Syst. Appl. Microbiol.* 34, 87–95. doi: 10.1016/j.syapm.2010.10.002

Vandamme, P., De Brandt, E., Houf, K., Salles, J. F., van Elsas, J. D., Spilker, T., et al. (2013). *Burkholderia humi* sp. nov., *Burkholderia choica* sp. nov., *Burkholderia telluris* sp. nov., *Burkholderia terrestris* sp. nov. and *Burkholderia udeis* sp. nov.: *Burkholderia glathei*-like bacteria from soil and rhizosphere soil. *Int. J. Syst. Evol. Microbiol.* 63(Pt 12), 4707–4718. doi: 10.1099/ijs.0.048900-0

Vandamme, P., Goris, J., Chen, W. M., de Vos, P., and Willems, A. (2002b). Burkholderia tuberum sp. nov. and *Burkholderia phymatum* sp. nov., nodulate the roots of tropical legumes. *Syst. Appl. Microbiol.* 25, 507–512. doi: 10.1078/07232020260517634

Vandamme, P., Goris, J., Chen, W.-M., De Vos, P., and Willems, A. (2002a). *Burkholderia tuberum* sp. nov. and *Burkholderia phymatum* sp. nov., nodulate the roots of tropical legumes. *Syst. Appl. Microbiol.* 25, 507–512. doi: 10.1078/07232020260517634

Vandamme, P., Holmes, B., Vancanneyt, M., Coenye, T., Hoste, B., Coopman, R., et al. (1997). Occurrence of multiple genomovars of *Burkholderia cepacia* in cystic fibrosis patients and proposal of *Burkholderia multivorans* sp. *nov. Int. J. Syst. Bacteriol.* 47, 1188–1200. doi: 10.1099/00207713-47-4-1188

Vandamme, P., Opelt, K., Knöchel, N., Berg, C., Schönmann, S., De Brandt, E., et al. (2007). *Burkholderia bryophila* sp. nov. and *Burkholderia megapolitana* sp. nov., moss-associated species with antifungal and plant-growth-promoting properties. *Int. J. Syst. Evol. Microbiol.* 57, 2228–2235. doi: 10.1099/ijs.0.65142-0

Vanlaere, E., van der Meer, J. R., Falsen, E., Salles, J. F., De Brandt, E., and Vandamme, P. (2008). *Burkholderia sartisoli* sp. nov., isolated from a polycyclic aromatic hydrocarbon-contaminated soil. *Int. J. Syst. Evol. Microbiol.* 58, 420–423. doi: 10.1099/ijs.0.65451-0

Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. (2002a). Identification of the bacterial endosymbionts in leaf galls of Psychotria (Rubiaceae, angiosperms) and proposal of'Candidatus Burkholderia kirkii'sp. nov. *Int. J. Syst. Evol. Microbiol.* 52, 2023–2027. doi: 10.1099/ijs.0.02103-0

Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. (2002b). Identification of the bacterial endosymbionts in leaf galls of Psychotria (*Rubiaceae*, angiosperms) and proposal of 'Candidatus Burkholderia kirkii' sp. nov. *Int. J. Syst. Evol. Microbiol.* 52(Pt 6), 2023–2027. doi: 10.1099/ijs.0.02103-0

Van Oevelen, S., De Wachter, R., Vandamme, P., Robbrecht, E., and Prinsen, E. (2004). '*Candidatus Burkholderia calva*'and '*Candidatus Burkholderia nigropunctata*' as leaf gall endosymbionts of African Psychotria. *Int. J. Syst. Evol. Microbiol.* 54, 2237–2239. doi: 10.1099/ijs.0.63188-0

Viallard, V., Poirier, I., Cournoyer, B., Haurat, J., Wiebkin, S., Ophel-Keller, K., et al. (1998). Burkholderia graminis sp. nov., a rhizospheric Burkholderia species, and reassessment of [Pseudomonas] phenazinium,[Pseudomonas] pyrrocinia and [Pseudomonas] glathei as Burkholderia. *Int. J. Syst. Bacteriol.* 48, 549–563. doi: 10.1099/00207713-48-2-549

Weilharter, A., Mitter, B., Shin, M. V., Chain, P. S., Nowak, J., and Sessitsch, A. (2011). Complete genome sequence of the plant growth-promoting endophyte *Burkholderia phytofirmans* strain PsJN. *J. Bacteriol.* 193, 3383–3384. doi: 10.1128/JB.05055-11

White, N. J. (2003). Melioidosis. *Lancet* 361, 1715. doi: 10.1016/S0140-6736(03)13374-0

Whitlock, G. C., Estes, D. M., and Torres, A. G. (2007). Glanders: off to the races with *Burkholderia mallei*. *FEMS Microbiol. Lett.* 277, 115–122. doi: 10.1111/j.1574-6968.2007.00949.x

Wiersinga, W. J., van der Poll, T., White, N. J., Day, N. P., and Peacock, S. J. (2006). Melioidosis: insights into the pathogenicity of *Burkholderia pseudomallei*. *Nat. Rev. Microbiol.* 4, 272–282. doi: 10.1038/nrmicro1385

Wong, S. Y., Paschos, A., Gupta, R. S., and Schellhorn, H. E. (2014). Insertion/deletion-based approach for the detection of *Escherichia coli* O157:H7 in freshwater environments. *Environ. Sci. Technol.* 48, 11462–11470. doi: 10.1021/es502794h

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056–1060. doi: 10.1038/nature08656

Yabuuchi, E., Kosako, Y., Oyaizu, H., Yano, I., Hotta, H., Hashimoto, Y., et al. (1992). Proposal of *Burkholderia* gen. nov. and transfer of seven species of the genus *Pseudomonas* homology group II to the new genus, with the type species *Burkholderia cepacia* (Palleroni and Holmes 1981) comb. nov. *Microbiol. Immunol.* 36, 1251–1275. doi: 10.1111/j.1348-0421.1992.tb02129.x

Yabuuchi, E., Kosako, Y., Oyaizu, H., Yano, I., Hotta, H., Hashimoto, Y., et al. (1993). *Burkholderia* gen. nov. validation of the publication of new names and new combinations previously effectively published outside the IJSB, List no 43. *Int. J. Syst. Bacteriol.* 43, 398–399. doi: 10.1099/00207713-43-2-398

Yang, H.-C., Im, W.-T., Kim, K. K., An, D.-S., and Lee, S.-T. (2006). *Burkholderia terrae* sp. nov., isolated from a forest soil. *Int. J. Syst. Evol. Microbiol.* 56, 453–457. doi: 10.1099/ijs.0.63968-0

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K. H., et al. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* 31, 241–250. doi: 10.1016/j.syapm.2008.07.001

Yoo, S.-H., Kim, B.-Y., Weon, H.-Y., Kwon, S.-W., Go, S.-J., and Stackebrandt, E. (2007). *Burkholderia soli* sp. nov., isolated from soil cultivated with Korean ginseng. *Int. J. Syst. Evol. Microbiol.* 57, 122–125. doi: 10.1099/ijs.0.64471-0

Zhang, H., Hanada, S., Shigematsu, T., Shibuya, K., Kamagata, Y., Kanagawa, T., et al. (2000). Burkholderia kururiensis sp. nov., a trichloroethylene (TCE)-degrading bacterium isolated from an aquifer polluted with TCE. *Int. J. Syst. Evol. Microbiol.* 50, 743–749. doi: 10.1099/00207713-50-2-743

Zhu, B., Zhou, S., Lou, M., Zhu, J., Li, B., Xie, G., et al. (2011). Characterization and inference of gene gain/loss along *burkholderia* evolutionary history. *Evol. Bioinform. Online* 7, 191. doi: 10.4137/EBO.S7510

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: 18 September 2014; paper pending published: 24 October 2014; accepted: 21 November 2014; published online: 19 December 2014.

Citation: Sawana A, Adeolu M and Gupta RS (2014) Molecular signatures and phylogenomic analysis of the genus Burkholderia: proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus Paraburkholderia gen. nov. harboring environmental species. Front. Genet. ***5****:429. doi: 10.3389/fgene.2014.00429*

This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics.

CHAPTER 6

GLIMPS: A User-Friendly Pipeline for the production of Core Genome and Concatenated Protein based Phylogenetic Trees and Protein based Comparative Genomic Analyses

Background

The construction and analysis of accurate phylogenetic trees has come to form the backbone of modern evolutionary biology and systematics research (Woese et al., 1990; Stackebrandt & Goebel, 1994; Yilmaz et al., 2013; Oren & Garrity, 2014; Parte, 2014). The growing availability of whole genome sequences for a large number of microbial organisms provides researchers with a powerful tool for the production of large, robust, and accurate multi-gene phylogenetic trees. Phylogenetic trees, when based on the entire shared core genome of the analysed group, are referred to as phylogenomic trees. Phylogenomic trees provide a number of advantages over single gene trees, including increased phylogenetic signal, improved resolution of relationships among organisms in the tree, and resistance to phylogenetic artifacts caused by lateral gene transfers and other anomalous genetic events (Rokas et al., 2003; Dutilh et al., 2004; Delsuc et al., 2005; Ciccarelli et al., 2006; Wu & Eisen, 2008; Puigbo et al., 2009; Wu et al., 2009). However, the production of phylogenomic trees is computationally intensive and presents three main challenges: identification of orthologous protein families, multiple sequence alignment (MSA), and the construction of the phylogenomic tree.

A number of phylogenomic tree building pipelines have been previously described in published literature (Wu & Eisen, 2008; Robbertse et al., 2011; Rodriguez-R et al., 2012; Wu & Scott, 2012; Dunn et al., 2013; Pearse & Purvis, 2013; Segata et al., 2013; Grant & Katz, 2014; Kumar et al., 2015). However, these pipelines are primarily command-line tools (Wu & Eisen, 2008; Robbertse et al., 2011; Rodriguez-R et al., 2012; Wu & Scott, 2012; Dunn et al., 2013; Pearse & Purvis, 2013; Segata et al., 2013; Grant & Katz, 2014), are generally designed and validated for use on eukaryotic transcriptome data (Robbertse et al., 2011; Dunn et al., 2013; Pearse & Purvis, 2013; Grant & Katz, 2014; Kumar et al., 2015), are often limited in their use of heuristics or computational acceleration methods (Wu & Eisen, 2008; Robbertse et al., 2011; Wu & Scott, 2012; Dunn et al., 2013; Pearse & Purvis, 2013; Grant & Katz, 2014), and, in a few cases, are designed to use preselected sets of near universal genes instead of the shared core genome of the organisms to be analyzed (Wu & Eisen, 2008; Wu & Scott, 2012; Segata et al., 2013).

Here I describe an integrated software pipeline for the production of phylogenomic trees called the Gupta Lab Integrated Microbial Phylogeny and Supermatrix (GLIMPS) pipeline (Figure 6.1). The GLIMPS pipeline uses heuristic tools to accelerate the orthologous protein family identification and phylogenomic tree construction stages of the phylogenomic tree building process. Additionally, the GLIMPS pipeline uses thread-aware multicore processing strategies to accelerate the sequence search and MSA stages of the phylogenomic tree building process. As well as producing phylogenomic trees, the GLIMPS pipeline is capable of producing presence-absence matrices of the shared protein families in the analyzed genomes, and calculating matrices for both the proportion of shared protein content and average amino acid identity of the analyzed genome sequences. Lastly, the GLIMPS pipeline includes a simple, user-friendly graphical user interface (GUI) which will allow researchers, who are outside of the field of bioinformatics or who may not be comfortable with command line based tools, to generate robust and reliable *de novo* phylogenomic trees.

Description of the GLIMPS pipeline and Graphical User Interface

Identification of Orthologous Protein Families

Orthologous proteins, hereafter referred to as orthologs, are defined as members of homologous protein families which have been separated by speciation events (Fitch, 1970). For example, the DNA gyrase proteins in *Escherichia coli* and *Bacillus subtilis* are orthologs (i.e. they are the “same” protein in different organisms, separated by speciation). The identification of orthologs is the crucial first step in phylogenomic analysis. The optimal methodology for identification of orthologs is the tree reconciliation method, which involves the comparison of a known species tree to the phylogenetic tree generated by individual genes/proteins (Zmasek & Eddy, 2001; Kristensen et al., 2011; Trachana et al., 2011). However, this methodology requires the computationally intensive task of creating accurate gene/protein based phylogenetic trees, and necessitates the presence of a known species tree, which is generally unavailable in microbial organisms. In contrast, the fastest current methods of ortholog identification involve threshold based protein clustering (Li & Godzik, 2006; Edgar, 2010; Fu et al., 2012). However, these methods are limited to proteins sharing 50% or greater sequence identity, and they cannot distinguish between orthologs and other types of homologous proteins. Consequently, the most common methodology for the identification of orthologs is the reciprocal best hit methodology. The reciprocal best hit method involves the use of all-vs-all sequence similarity comparisons of each gene/protein in a pair of genomes. Proteins which share the highest similarity to each other in different genomes are identified as orthologs (Remm et al., 2001). However, due to the exponential rate of increase in the number of comparisons required, this methodology does not scale well beyond about 50 genomes (Lechner et al., 2011).

In the GLIMPS pipeline, we utilize CD-Hit (Fu et al., 2012), a threshold based protein clustering program to generate initial protein families which share 50% or greater sequence similarity with each other. We then use Clustal Omega (Sievers et al., 2011), a fast and accurate MSA program, to generate alignments of these protein families. These MSAs are converted into profile Hidden Markov Models (HMMs) (Eddy, 1998), which are statistical representations of the MSA, using HMMer (Eddy, 2011). The profile HMMs are then used to search for other members of the protein families in the input genomes. This process is similar to the highly sensitive PSI-BLAST algorithm (Altschul et al., 1997) and the phylogenomic clustering methodology utilized by the PATRIC database (Wattam et al., 2014). Overall, this methodology has the benefit of combining two extremely fast ortholog detection procedures, protein clustering and non-reciprocal sequence similarity searches, to create a fast ortholog detection process that manages to retain a significant amount of sensitivity. Once the identification of orthologous protein families steps are complete, the GLIMPS pipeline is able to generate a presence-absence table and calculate the percentage of shared proteins (Qin et al., 2014) for each pair of genomes in the analysis (Figure 6.1).

Multiple Sequence Alignment

The quality and accuracy of an MSA has significant impacts on the accuracy of the resultant phylogenetic tree, a property often referred to as the “garbage in, garbage out” principle (Ogden & Rosenberg, 2006; Talavera & Castresana, 2007; Liu et al., 2010a; Wang et al., 2011). The production of accurate MSAs is a computationally difficult task and there is a strong inverse relationship between alignment accuracy and alignment speed (Notredame et al., 2000; Katoh & Toh, 2007; Liu et al., 2010b; Sievers et al., 2011). In the GLIMPS pipeline, we utilize the program Clustal Omega for our MSAs (Sievers et al., 2011) which is a preferable MSA program for phylogenomic analyses for two main reasons. Firstly, Clustal Omega scores within 5-10% of the alignment quality of the most accurate MSA programs on benchmarks, similar to popular alignment programs such as MAFFT (Katoh & Standley, 2013) and MUSCLE (Edgar, 2004), while being up to two orders of magnitude faster than the most accurate MSA programs, and faster than other similarly accurate alignment programs (Sievers et al., 2011). Secondly, unlike the popular and accurate MAFFT (Katoh & Standley, 2013) L-INS-i setting, Clustal Omega does not presuppose that the sequence being aligned is from a single domain, globular protein. This allows Clustal Omega to be more adaptable to the varied protein types encountered in whole genomes (Sievers et al., 2011). In the GLIMPS pipeline, we have accelerated the MSA stage of phylogenomic tree construction by utilizing the Python multiprocessing module to assign individual instances of Clustal Omega to each available thread of execution on the host computer. Once each of the protein families has been aligned, the GLIMPS pipeline is able to calculate the average amino acid identity (Thompson et al., 2013) of the shared proteins in each pair of genomes in the analysis (Figure 6.1).

The quality of MSAs can be improved by removing poorly aligned regions in a process known as alignment trimming. Alignment trimming is thought to increase the signal to noise ratio of the MSA (Talavera & Castresana, 2007; Capella-Gutierrez et al., 2009; Wu et al., 2012) and has been shown to generally improve the power of phylogenetic inference (Talavera & Castresana, 2007; Löytynoja & Goldman, 2008; Cummins & McInerney, 2011). In the GLIMPS pipeline, we have utilized the alignment trimming program TrimAl (Capella-Gutierrez et al., 2009) to trim our alignments before concatenation into a supermatrix. As a sequence based alignment trimming program, TrimAl processes alignments multiple orders of magnitude faster than confidence based alignment trimming programs (Chang et al., 2014) and, unlike the widely used sequence based alignment trimming program GBlocks (Castresana, 2000), TrimAl is capable of automatically optimizing the parameters used to trim each sequence alignment in the core genome based on the sequence characteristics of each input MSA (Capella-Gutierrez et al., 2009). This quality facilitates the use of TrimAl for alignment trimming in large phylogenomic datasets.

Phylogenomic Tree Construction

The difficulty of phylogenetic tree construction increases exponentially with the length of analyzed alignment (Stamatakis, 2014), making phylogenetic tree construction based on supermatrices computationally intensive. Constructing a phylogenetic tree based on a genome-scale supermatrix can take weeks on a consumer grade desktop computer using the fastest maximum-likelihood based phylogeny programs currently available, PhyML (Guindon et al., 2010) and RAxML (Stamatakis, 2014). FastTree, a program developed specifically to create large-scale phylogenies, uses heuristic methodology to approximate maximum-likelihood phylogenies, which are nearly as accurate as the maximum-likelihood phylogenies produced by PhyML or RAxML, and is at least two orders of magnitude faster in its execution (Price et al., 2010; Liu et al., 2011). In the GLIMPS pipeline, we utilize FastTree to construct an approximately maximum-likelihood phylogenetic tree which we then pass as input to RAxML, instead of the default maximum-parsimony tree. This greatly reduces the time RAxML requires to optimize individual branch lengths and perform local rearrangements in order to identify the optimal maximum-likelihood topology.

Graphical User Interface

The GLIMPS pipeline includes a simple GUI written using the Python Tk interface module (Figure 6.2). The GUI for the GLIMPS pipeline consists of three main components: the main input screen, the settings screen, and the activity log. The main input screen, shown in Figure 6.2A, allows the user to select the directory containing the translated protein files, in the fasta format, for each genome to be included in the phylogenomic analysis, and to select the save location for the GLIMPS output files. The main input screen also allows the user to optionally provide a user created file, in the fasta format, containing a set of curated protein sequences, such as multilocus sequence analysis proteins or ribosomal proteins, which can be used to generate the phylogenomic tree instead of the proteins in the core genome. The settings screen, shown in Figure 6.2B, allows the user to select which output files will be generated during the phylogenomic analysis, including the presence-absence table, the percentage of shared proteins matrix, and the average amino acid identity matrix. The user may also select the minimum percentage of the input genomes in which a protein must be found to be utilized in the phylogenomic analysis, and may select or modify the local paths for the software utilized by the GLIMPS pipeline. Lastly, the GLIMPS pipeline provides the user with a real-time log of the current status of the phylogenomic analysis, allowing the user to easily monitor the state and performance of the pipeline.

Discussion

The GLIMPS phlyogenomic analysis pipeline is a simple, integrated tool capable of quickly producing accurate and robust phylogenomic trees for use in complete comparative genomic analyses. The GLIMPS pipeline uses well-established and validated tools and several heuristic steps to rapidly generate publication quality phylogenomic trees. Early versions of the GLIMPS pipeline have already been utilized to produce phylogenomic trees in a number of published evolutionary microbiology and systematic studies (Campbell et al., 2015; Gupta et al., 2015b; Naushad et al., 2015a; Gupta et al., 2016; Zhang et al., 2016). The GLIMPS pipeline has also been utilized to generate the protein based phylogenetic trees and the percentage of conserved proteins matrix shown in the submitted manuscript presented in Chapter 7 of this thesis. The binary executables for Windows, macOS, and Linux for the GLIMPS pipeline will be available on the Gupta Lab Evolutionary Analysis Software website (GLEAnS.net) once completed and the source code for the pipeline will be hosted on GitHub. The GLIMPS pipeline represents a step forward in providing bioinformatics tools to the wider research community and will allow researchers to generate robust and reliable *de novo* phylogenomic trees without the requirement of extensive bioinformatics or computing skills.

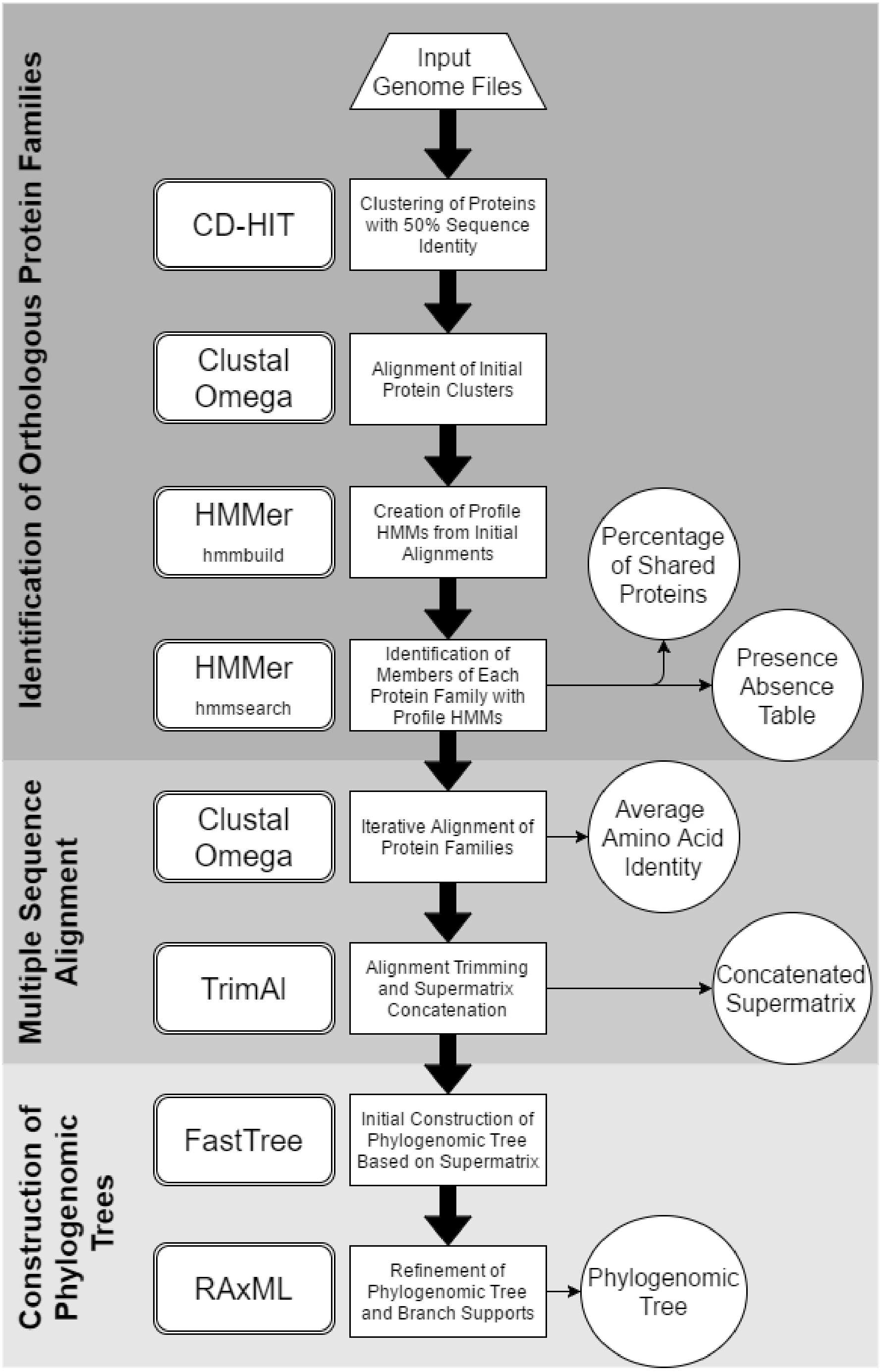
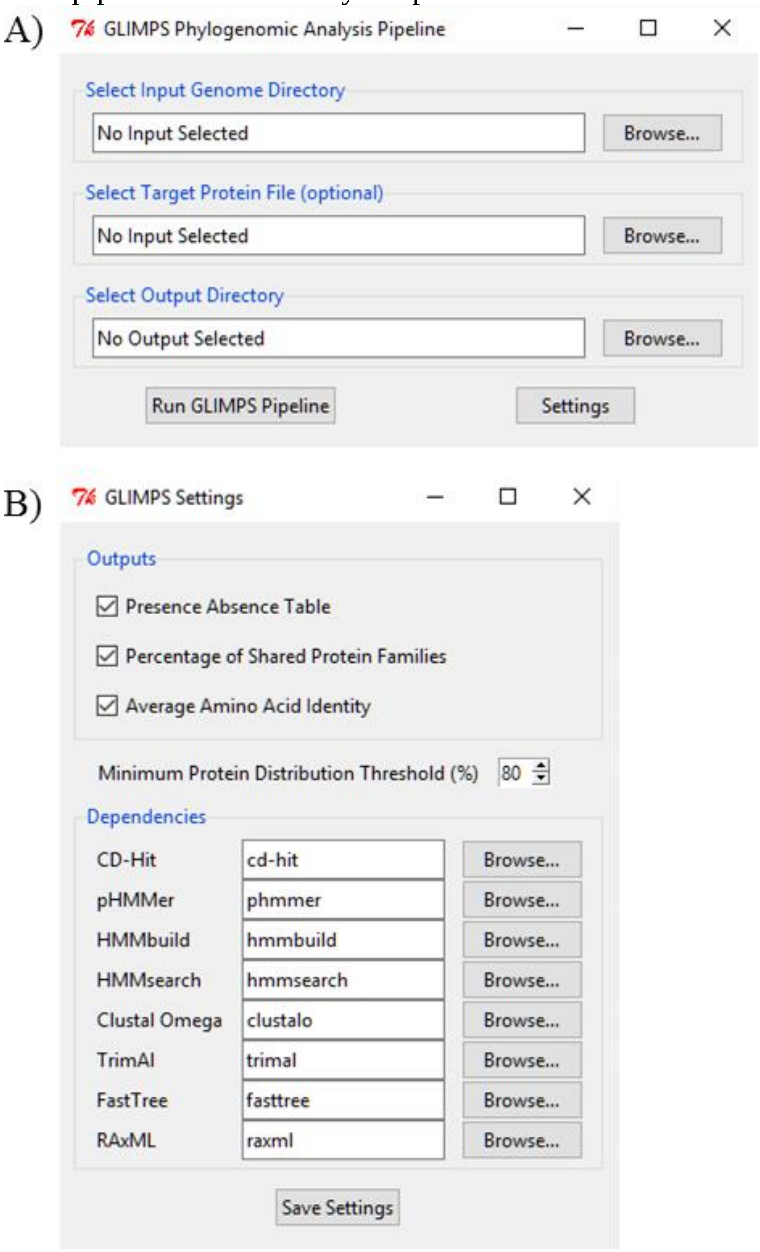

Identification of Orthologous Protein Families
Multiple Sequence Alignment
Construction of Phylogenomic Trees
Input Genome Files
CD-HIT
Clustering of Proteins with 50% Sequence Identity
Clustal Omega
Alignment of Initial Protein Clusters
HMMer hmmbuild
Creation of Profile HMMs from Initial Alignments
HMMer hmmsearch
Identification of Members of Each Protein Family with Profile HMMs
Percentage of Shared Proteins
Presence Absence Table
Clustal Omega
Iterative Alignment of Protein Families
Average Amino Acid Identity
TrimAl
Alignment Trimming and Supermatrix Concatenation
Concatenated Supermatrix
FastTree
Initial Construction of Phylogenomic Tree Based on Supermatrix
RAxML
Refinement of Phylogenomic Tree and Branch Supports
Phylogenomic Tree

Figure 6.1 A flowchart depicting the program logic of the GLIMPS pipeline. The three main phases of the pipeline are highlighted in different shades of grey. Each step of the pipeline is described in white rectangles. The names of the programs used in each step of the pipeline are in rounded rectangles beside the description

of the step. The five main outputs of the pipeline are shown in circles connected to the step in the pipeline in which they are produced.



Select Input Genome Directory: No Input Selected

Select Target Protein File (optional): No Input Selected

Select Output Directory: No Output Selected

Outputs

☑ Presence Absence Table

☑ Percentage of Shared Protein Families

☑ Average Amino Acid Identity

Minimum Protein Distribution Threshold (%) 80

Dependencies

CD-Hit	cd-hit
pHMMer	phmmer
HMMbuild	hmmbuild
HMMsearch	hmmsearch
Clustal Omega	clustalo
TrimAl	trimal
FastTree	fasttree
RAxML	raxml

Figure 6.2 Examples of the user-friendly graphical user interface for the GLIMPS pipeline showing (A) the main input interface and (B) the settings screen. The user interface for the GLIMPS pipeline also provides the user with a real-time log of the status of the current phylogenomic analysis (not shown).

CHAPTER 7

A molecular and genomic examination of the phylogeny and taxonomy of the order *Enterobacteriales*: proposal to divide the order *Enterobacteriales* into seven families (*Enterobacteriaceae, Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov.)

This chapter describes the use of molecular signatures (CSIs), protein based phylogenetic trees, and genomic distance (shared protein content) to differentiate the seven main groups within the order *Enterobacteriales*. A version of the tool described in Chapter 6 is utilized to produce the phylogenetic trees and to calculate shared protein content of the genomes examined in this chapter. The chapter concludes with a proposal to divide the order *Enterobacteriales* into seven families (*Enterobacteriaceae*, *Erwiniaceae*, *Pectobacteriaceae*, *Yersiniaceae*, *Hafniaceae*, *Morganellaceae*, and *Budviciaceae*). My contributions towards the completion of this chapter include the construction of phylogenetic trees based on the core genome, ribosomal proteins, and multi-locus sequence analysis proteins, the production of the shared protein content matrix and the presence absence matrix, the creation of the taxonomic proposals, the writing of all drafts and revisions of the manuscript, and involvement in the production of main and supplemental figures and tables in the manuscript.

Due to limited space, supplementary materials for this work are not included in the chapter but will become available with the rest of the manuscript upon publication.

Genome based phylogeny and taxonomy of the Enterobacteriales: proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov.

Mobolaji Adeolu[†], Seema Alnajar[†], Sohail Naushad, and Radhey S. Gupta*
Department of Biochemistry and Biomedical Sciences,
McMaster University, Hamilton, Ontario, Canada L8N 3Z5

Short Title: Phylogeny and taxonomy of the order Enterobacteriales

Content Category; Evolution, Phylogeny and Biodiversity

Keywords: Enterobacteriales, *Enterobacterales*, *Enterobacteriaceae*, phylogeny, taxonomy, conserved signature indels

[†] Mobolaji Adeolu and Seema Alnajar contributed equally to this work.

*Contact information for the Corresponding Author:
Phone: (905) 525-9140 ext. 22639
Fax: (905) 522-9033
Email: gupta@mcmaster.ca

Abstract

Understanding of the phylogeny and interrelationships of the genera within the order "Enterobacteriales" has proven difficult using the 16S rRNA gene and other single-gene or limited multi-gene approaches. In this work, we have completed comprehensive comparative genomic analyses of the members of the order "Enterobacteriales" which includes phylogenetic reconstructions based on 1548 core proteins, 53 ribosomal proteins, 4 multilocus sequence analysis proteins, as well as examining the overall genome similarity amongst the members of this order. The results of these analyses all support the existence of 7 distinct monophyletic groups of genera within the order "Enterobacteriales". In parallel, our analyses of protein sequences from the "Enterobacteriales" genomes have identified numerous molecular characteristics in the forms of Conserved Signature Insertions/deletions, which are specifically shared by the members of the identified clades and independently support their monophyly and distinctness. Many of these groupings, either in part or in whole, have been recognized in previous evolutionary studies, but have not been consistently resolved as monophyletic entities in 16S rRNA trees. The work presented here represents the first comprehensive, genome-scale taxonomic analysis of the entirety of the order "Enterobacteriales". On the basis of phylogenetic analyses and the numerous identified conserved molecular characteristics, which clearly distinguish members of the order "Enterobacteriales" and the seven reported clades within this order, a proposal is made here for the order *Enterobacterales* ord. nov. which consists of 7 families: *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov.

Introduction

The order “Enterobacteriales” is a large and diverse group of Gram-negative, facultatively anaerobic, non-spore-forming rod-shaped bacteria within the class *Gammaproteobacteria*. The members of the group inhabit a number of different ecological niches and have been found in soil, water, and in association with living organisms including plants, insects, animals and humans (Brenner & Farmer III, 2005). Many members of the order “Enterobacteriales” have been implicated as pathogens in humans and animals, such as the species *Escherichia coli*, *Salmonella enterica*, and *Yersinia pestis*, and as economically devastating phytopathogens, such as members of the genera *Dickeya*, *Pectobacterium*, *Brenneria*, *Erwinia*, and *Pantoea* (Hauben et al., 1998; Bonn & van der Zwet, 2000; Tyler & Triplett, 2008; Coutinho & Venter, 2009; Croxen & Finlay, 2010; Livermore, 2012). The order “Enterobacteriales” currently contains 60 validly named genera (Parte, 2014; NamesforLife, 2016) and two additional genera which have been recently described but not yet validly published (viz. “Atlantibacter” and “Chania”) (Ee et al., 2016; Hata et al., 2016). Most genera within the order “Enterobacteriales”, encompassing over 250 species, are placed within the sole, validly described family within the order, *Enterobacteriaceae*; making the family *Enterobacteriaceae* one of the most taxonomically diverse bacterial families currently recognized (Parte, 2014; NamesforLife, 2016). A number of distinct groupings of genera within the family *Enterobacteriaceae* are well known (viz. the groupings of the genera *Salmonella*, *Citrobacter*, and *Escherichia*/*Shigella*, and the genera *Dickeya*, *Pectobacterium*, and *Brenneria*, the close associations between the genera *Xenorhabdus* and *Photorhabdus*, the genera *Erwinia* and *Pantoea*, and the genera *Obesumbacterium* and *Hafnia*) (Hauben et al., 1998; Samuel et al., 2004; Goodrich-Blair & Clarke, 2007; Naushad et al., 2014; Octavia & Lan, 2014; Zhang & Qiu, 2015; Zhang et al., 2016), but these groupings are not recognized as unique taxonomic units.

The biochemical diversity and the large number of organisms within the order “Enterobacteriales” has made biochemical descriptions of the order and its constituent subgroups difficult (Brenner & Farmer III, 2005; Octavia & Lan, 2014). Our current understanding of the phylogeny and interrelationships of the members of the order “Enterobacteriales” is primarily based on the 16S rRNA gene (Hauben et al., 1998; Spröer et al., 1999; Francino et al., 2006; Naum et al., 2008). However, the 16S rRNA gene has low discriminatory power and interrelationships of the members of the order “Enterobacteriales” are poorly resolved in 16S rRNA gene based phylogenetic trees (Hauben et al., 1998; Naum et al., 2008; Octavia & Lan, 2014). Additionally, the branching of the genera and species within "Enterobacteriales" in 16S rRNA gene based phylogenies shows considerable stochasticity depending on the algorithms used and the organisms analyzed (Naum et al., 2008; Octavia & Lan, 2014). Most concerning, comprehensive 16S rRNA phylogenetic trees for the order "Enterobacteriales" and other members of the class *Gammaproteobacteria* suggest that the order "Enterobacteriales" exhibits polyphyletic branching and does not form a coherent monophyletic grouping (Brenner & Farmer III, 2005; Yarza et al., 2008; Yilmaz et al., 2013; Octavia & Lan, 2014). A number of alternative genes have been employed in phylogenetic analysis of the order "Enterobacteriales" in order to gain additional insight into the interrelationships of the members of the order such as *gyrB* (Dauga, 2002; Fukushima et al., 2002), *dnaJ* (Nhung et al., 2007), *oriC* (Roggenkamp, 2007), and recA (Tailliez et al., 2010). More recently, multiple gene/protein based multilocus sequence analysis (MLSA) studies have been conducted to further elucidate the phylogeny of the order "Enterobacteriales" including studies based on the genes *tuf* and *atpD* (Paradis et al., 2005), the genes *atpD*, *carA*, and *recA* (Young & Park, 2007), the genes *gapA*, *gyrA* and *ompA* (Naum et al., 2011), the genes *rpoB*, *gyrB*, *dnaJ*, and *recA* (Hata et al., 2016), the genes *fusA*, *pyrG*, *rplB*, *rpoB* and sucA (Ee et al., 2016), and, most commonly, the genes *gyrB*, *rpoB*, *atpD* and *infB* (Brady et al., 2008; Brady et al., 2013; Brady et al., 2014b; Glaeser & Kämpfer, 2015; Zhang & Qiu, 2015). These studies have led to a significant number of reclassifications within the order "Enterobacteriales" and have alleviated many of the issues related to polyphyletic genera within the order. However, no family-level divisions within the order "Enterobacteriales" have thus far been proposed.

The increasing prevalence and ubiquity of genome sequencing technology has led to an increasing wealth of publically available genome sequence data. Currently, there are over 14 000 genomes from 54 validly named genera within the order "Enterobacteriales" available in the NCBI genome database (NCBI, 2016). These genome sequences are enabling the increasing use of robust and reliable core genome based phylogenetic reconstructions in "Enterobacteriales" research (Husnik et al., 2011; Wattam et al., 2014; Zhang & Qiu, 2015; Zhang et al., 2016), which have been shown to mitigate the effects of recombination or lateral gene transfer and provide greater resolving power than phylogenetic trees based on single genes/proteins (Rokas et al., 2003; Ciccarelli et al., 2006; Gao et al., 2009; Wu et al., 2009). Genome sequence data is also enabling the detection of conserved molecular characteristics shared by evolutionarily related groups of organisms. One particular class of conserved molecular characteristics, which have recently been utilized to great effect in prokaryotic taxonomy are conserved signature insertions/deletions (CSIs) present in widely distributed proteins (Gupta, 2014; Naushad et al., 2014; Gupta, 2016). CSIs are insertions or deletions (indels) that are uniquely present in a related group of organisms. The most parsimonious explanation of the presence of the CSI in a related group of organisms is the existence of a common ancestor in which the genetic change leading to the CSI occurred, and which was subsequently inherited by all of its various decedents. Thus, CSIs represent synapomorphic characteristics and they provide reliable evidence, independent of phylogenetic trees, that the species from the groups in which they are found are specifically related to each other due to common ancestry. Recently, on the basis of CSIs and other molecular characteristics, the taxonomy of a number of important prokaryotic groups, ranging from genus to phylum level taxa, has been revised (Naushad et al., 2014; Sawana et al., 2014; Campbell et al., 2015; Gupta et al., 2015a; Gupta et al., 2015b; Naushad et al., 2015b; Gupta, 2016; Gupta et al., 2016).

In our earlier work, a limited number of CSIs and unique proteins, referred to as conserved signature proteins, were identified that were distinctive characteristics of either all *Gammaproteobacteria* or were commonly shared by members from certain orders of *Gammaproteobacteria* which reliably grouped together in phylogenetic trees constructed in this work (Gupta, 2000; Gao et al., 2009). We have also previously completed comprehensive studies in order to identify large numbers of CSIs utilized to reclassify members within the gammaproteobacterial orders *Pasteurellales* and *Xanthomonadales* (Naushad & Gupta, 2012, 2013; Naushad et al., 2015a; Naushad et al., 2015b). In the present study, we have extended our earlier work on *Gammaproteobacteria* by carrying out comprehensive phylogenetic and comparative genomic studies on members of the order “Enterobacteriales” to examine their evolutionary relationships and taxonomy. Using whole genome sequences of 179 representative genome sequenced members of the order “Enterobacteriales”, we have constructed a highly robust phylogenetic tree based on 1548 shared core proteins, as well as phylogenetic trees based on 53 ribosomal proteins and 4 MLSA proteins, and to identify conserved molecular characteristics that can be used to determine the interrelationships within the order “Enterobacteriales”. Here we present 5 CSIs which are unique characteristics of all

“Enterobacteriales” and an additional 64 CSIs which are specific for 7 main groups of genera within the order “Enterobacteriales” identified in our phylogenetic trees. The 69 CSIs identified in this work, when combined with previously discovered CSIs (Naushad et al., 2014) and the highly robust phylogenetic trees constructed here, provide for a comprehensive understanding of interrelationships within the order “Enterobacteriales” and form the basis for a novel taxonomic framework. On the basis of the phylogenetic analyses and the identified conserved molecular characteristics presented here, we propose a division of the order “Enterobacteriales” (now renamed as the order *Enterobacterales* ord. nov.) into 7 families: *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov.

Methods

Phylogenetic and Genomic Analyses of the order Enterobacteriales

Three phylogenetic trees were produced in this work utilizing 179 representative genome sequenced members of the order “Enterobacteriales” (Supplemental Table 1) and 4 members of the families *Pasteurellaceae* and *Vibrionaceae* as outgroups. Representative genomes for the genus *Plesiomonas* and the endosymbiotic genera *Buchnera* and *Wigglesworthia* were not included in the phylogenetic trees shown in the main figures due to the potential for phylogenetic artifacts caused by long branch attraction effects (Bergsten, 2005; Philippe et al., 2005), but are shown in the respective supplemental figures for each phylogenetic tree. A core genome phylogeny was produced based on the concatenated sequences of 1548 core proteins. The core protein families used in the core genome phylogeny were identified using the UCLUST algorithm (Edgar, 2010) to identify protein families which shared at least 50% sequence identity and 50% sequence length. 1548 identified proteins families which were present in at least 80% of the input genomes were used in the phylogenetic analysis. The 53 ribosomal proteins were identified using HMMer 3.1 (Eddy, 2011) based on profile hidden Markov models (Supplemental Table 2) obtained from the Pfam database (Finn et al., 2016). The 4 MLSA proteins (viz. GyrB, RpoB, AtpD and InfB) were identified using HMMer 3.1 (Eddy, 2011) based on amino acid sequences from *Escherichia coli* K12 (Blattner et al., 1997) (Supplemental Table 2) obtained from the UniProt database (UniProt, 2015). In each case, each identified protein family was individually aligned using Clustal Omega (Sievers et al., 2011), trimmed using Gblocks 0.91b (Castresana, 2000) with relaxed parameters (Talavera & Castresana, 2007), and concatenated with the other proteins in its dataset. The concatenated alignments were 458,971, 5930, and 3535 aligned amino acids long for the core protein, ribosomal protein, and MLSA protein datasets, respectively. Maximum-likelihood trees based on these concatenated alignments were constructed using FastTree 2 (Price et al., 2010) employing the Whelan and Goldman model of protein sequence evolution (Whelan & Goldman, 2001) and RAxML 8 (Stamatakis, 2014) using the Le and Gascuel model of protein sequence evolution (Le & Gascuel, 2008). SH-like statistical support values (Guindon et al., 2010) for each branch node in the final phylogenetic trees were calculated using RAxML 8 (Stamatakis, 2014). The resultant phylogenetic trees were drawn using MEGA 6 (Tamura et al., 2013). This process was completed using an internally developed software pipeline. A manuscript for this pipeline is currently under preparation and the pipeline will be available for public use on Gleans.net once released. We have also utilized the protein families identified by the USearch algorithm (Edgar, 2010) for our core protein based phylogenetic tree to calculate the proportion of shared protein families in each pair of genomes in our dataset.

Identification of Conserved Signature Indels

Conserved signature indels were identified as detailed by Gupta (2014) using protein sequences found in the genomes of *Shimwellia Blattae* DSM 4481 (Brzuszkiewicz et al., 2012), *Providencia stuartii* MRSN 2154 (Clifford et al., 2012), *Pragia fontium* 24613 (Snopková et al., 2015) and *Dickeya zeae* Ech586 (Pritchard et al., 2013) as the starting points. BLASTp (Altschul et al., 1997) searches were conducted on each of the protein sequences in these genomes that were >75 amino acids in length against the NCBI non-redundant database. From the results of the BLASTp searches, 15-20 homologues belonging different genera of "Enterobacteriales" and 6-8 species from other orders/classes of proteobacteria were selected. The selected sequences were aligned using Clustal_X 2.1 (Jeanmougin et al., 1998). The alignments were then visually inspected for the presence of insertions or deletions that were flanked on both sides by at least 5-6 conserved amino acid residues in the neighboring 30–40 amino acids. Gaps that were of a variable length or that were not flanked by conserved residues were not further investigated. Detailed BLASTp searches were then carried out on short sequence segments containing the indel and the flanking conserved regions (60-100 amino acids long) and compared against the top 500 BLAST hits to determine the specificity of the indels. In some cases, an additional BLASTp search was conducted to include a more diverse representation of the "Enterobacteriales" species involving 1000 alignments, or excluding overrepresented species. SIG_CREATE and SIG_STYLE (available on Gleans.net) were then used to create Signature files for identified CSIs that were specific to the order "Enterobacteriales" or one of its subgroups as described by Gupta (2014). Due to the large number of genome sequences available for the order "Enterobacteriales", the sequence alignment files presented here contain sequence information for only a limited number of species. However, unless otherwise indicated, homologs of all members of the specified groups displayed similar sequence characteristics.

Results

Phylogenetic and Genomic Analyses of the order Enterobacteriales

Phylogenetic analyses of the order Enterobacteriales

In this work, we have produced 3 phylogenetic trees for 179 representative members of the order "Enterobacteriales", encompassing 49 validly named genera within the order: one tree based on 1548 core proteins, another based on 53 ribosomal proteins, and a third based on 4 MLSA proteins (Figure 1A-1C and Supplemental Figures 1-3). The 1548 core protein based phylogeny produced for this work, covering a majority of the diversity present within the order, represents one of the most comprehensive genome based phylogenetic trees for the order "Enterobacteriales" produced to date. Additionally, a 16S rRNA gene based phylogenetic tree of the "Enterobacteriales", produced as part of the All-Species Living Tree project release 123 (Yarza et al., 2008; Yilmaz et al., 2013), is shown in Figure 1D and Supplemental Figure 4.

The branching pattern of the main groups within the order "Enterobacteriales" in the genome based tree, the ribosomal protein tree, and the MLSA based phylogenetic tree are highly consistent. In each of the phylogenetic trees, the members of the order "Enterobacteriales" form 7 main groups/clades which are labelled in the phylogenetic tree figures. The first group, referred to as the *Enterobacter-Escherichia* clade, is the largest group within the order "Enterobacteriales" and consists of the genera "Atlantibacter"*, Buttiauxella, Cedecea, Citrobacter, Cronobacter, Enterobacter, Escherichia, Franconibacter, Klebsiella, Kluyvera,* *Kosakonia, Leclercia, Lelliottia, Mangrovibacter, Pluralibacter, Raoultella, Salmonella, Shigella, Shimwellia, Siccibacter, Trabulsiella,* and *Yokenella*. The *Erwinia-Pantoea* clade, which is present in a monophyletic grouping with the *Enterobacter-Escherichia* clade, consists of the genera *Erwinia, Pantoea, Phaseolibacter,* and *Tatumella*. The *Pectobacterium-Dickeya* clade consists of the genera *Brenneria, Dickeya, Lonsdalea, Pectobacterium* and *Sodalis*, the *Yersinia-Serratia* clade consists of the genera "Chania", *Ewingella, Rahnella, Rouxiella, Serratia,* and *Yersinia*, the *Hafnia-Edwardsiella* clade consists of the genera *Edwardsiella, Hafnia,* and *Obesumbacterium*, the *Proteus-Xenorhabdus* clade consists of the genera *Arsenophonus, Moellerella, Morganella, Photorhabdus, Proteus, Providencia,* and *Xenorhabdus*, and, lastly, the *Budvicia* clade consists of the genera *Budvicia, Leminorella,* and *Pragia*. Apart from one exception, the genera within the order "Enterobacteriales" consistently branch together within the clades described above as distinct monophyletic groupings in the phylogenetic trees. The sole exception to these groupings is observed in the ribosomal protein based phylogenetic tree. In the ribosomal protein based phylogenetic tree, the two representative members of the genus *Sodalis*, which are early branching members of the *Pectobacterium-Dickeya* clade in other phylogenetic trees, branch outside of the *Pectobacterium-Dickeya* clade, exhibiting no branching affinity for any of the main clades within the order "Enterobacteriales" in the ribosomal protein based phylogenetic tree. The early branching of the genus *Sodalis* from other members of the *Pectobacterium-Dickeya* clade in the genome and MLSA based phylogenetic trees and the lack of branching affinity of the genus *Sodalis* to any main clade within the order "Enterobacteriales" in the ribosomal protein based phylogenetic tree may be a result of the endosymbiotic adaptations of the genus *Sodalis* which have led to significant genome degradation and genetic divergence from its closest relatives (Toh et al., 2006).

The genera *Buchnera*, *Plesiomonas*, and *Wigglesworthia* exhibit atypical branching characteristics and are not included in the main figures, but the results for them are presented in the Supplemental Figures 1B, 2B, and 3B. The endosymbiotic genera *Buchnera* and *Wigglesworthia* possess extremely long branches and form a monophyletic cluster. However, the monophyletic clustering of *Buchnera* and *Wigglesworthia* is potentially a consequence of long branch attraction artefacts, compositional bias due to their small A+T-rich genomes, and rooting (Bergsten, 2005; Herbeck et al., 2005; Philippe et al., 2005; Williams et al., 2010; Husník et al., 2011). The genera *Buchnera* and *Wigglesworthia* branch between the *Enterobacter-Escherichia* and the *Erwinia-Pantoea* clades in both the genome and ribosomal protein based phylogenetic trees (Supplemental Figures 1B and 2B), but branch earlier, after the *Budvicia* clade, in the MLSA based phylogenetic tree. In contrast to these two genera, genus *Plesiomonas* forms an early diverging outgroup of the order "Enterobacteriales" in the genome and MLSA based phylogenetic trees (Supplemental Figures 1B and 3B), and branches between the *Vibrionaceae* and *Pasteurellaceae* members in the ribosomal protein based phylogenetic tree (Supplemental Figure 2B). It is of interest to note that *Plesiomonas* has historically been difficult to place in a specific taxonomic group due to its atypical phenotypic characteristics and highly recombinant genome (Salerno et al., 2007; Janda et al., 2016). The genus *Plesiomonas* was originally placed within the family *Vibrionaceae* before transfer to the family *Enterobacteriaceae* (Ruimy et al., 1994; Janda, 2005).

The genera within the "Enterobacteriales" in the 16S rRNA based phylogenetic tree (Figure 1D and Supplemental Figure 4) exhibit extensive polyphyly and many of the clades identified in the genome, ribosomal protein, and MLSA based phylogenetic trees are poorly resolved or unsupported in the 16S rRNA based phylogenetic tree. Similar to the genome, ribosomal protein, and MLSA based phylogenetic trees, a monophyletic grouping of the genera within the *Enterobacter-Escherichia* clade and the *Erwinia-Pantoea* clade is observed in the 16S rRNA gene based phylogenetic tree. However, the members of the *Erwinia-Pantoea* clade branch within the *Enterobacter-Escherichia* clade in the 16S rRNA gene based phylogeny instead of branching as two distinct, but related groups. In the 16S rRNA gene based phylogenetic tree, the *Yersinia-Serratia* clade and the *Hafnia-Edwardsiella* clade, as well as the genus *Budvicia* from the *Budvicia* clade, form a highly intermixed, paraphyletic outgroup of the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades (simply labelled as the *Yersinia-Serratia* clade in Figure 1D). The *Pectobacterium-Dickeya* clade forms a distinct, monophyletic grouping in the 16S rRNA based phylogenetic tree that is largely consistent with the branching seen in the genome, ribosomal protein, and MLSA based phylogenetic trees. The members of the *Proteus-Xenorhabdus* clade cluster together in a paraphyletic grouping. Notably, the earliest branching members of the order "Enterobacteriales" in the genome, ribosomal protein, and MLSA based phylogenetic trees (viz. the *Proteus-Xenorhabdus* and *Budvicia* clades) and the members of the *Pectobacterium-Dickeya* exhibit closer affinity to other families within the class *Gammaproteobacteria* (viz. *Pasteurellaceae*, *Orbaceae*, and *Thorselliaceae*) than to the other members of the *Enterobacteriaceae*, making the order "Enterobacteriales" polyphyletic in the 16S rRNA based phylogenetic tree.

Genome relatedness of the members of the order Enterobacteriales

The gold standard technique in microbial classification is the DNA-DNA hybridization methodology (Gevers et al., 2005; Goris et al., 2007). Recently, *in silico* measures of genome to genome relatedness have been used in classification as replacements for the DNA-DNA hybridization procedure (Konstantinidis & Tiedje, 2005; Rosselló-Mora, 2006; Auch et al., 2010). Here we utilize a measure of genome to genome relatedness with applications for phylogeny and classification, the proportion of shared protein families in a pair of genomes, that has alternately been referred to as Percentage of Conserved Proteins (Qin et al., 2014) and Alignment Fraction (Varghese et al., 2015) in prior studies (Figure 2). This measure of genome to genome relatedness is particularly useful at higher taxonomic ranks because of its large dynamic range which extends from > 60% for closely related organisms (Qin et al., 2014; Varghese et al., 2015) to <1% for distantly related organisms (Ciccarelli et al., 2006; Dagan & Martin, 2006). The 7 main groups of genera observed in our phylogenetic trees (Figure 1) exhibit distinctly higher genome to genome relatedness to each other than to other groups of genera in our analysis of shared protein families (Figure 2). Additionally, the proportion of shared protein content also supports the general branching order observed in the phylogenetic trees with the *Enterobacter-Escherichia*, *Erwinia-Pantoea*, *Yersinia-Serratia*, *Hafnia-Edwardsiella*, and *Pectobacterium-Dickeya* clades exhibiting more a higher proportion of shared protein families with each other than to the early branching *Proteus-Xenorhabdus* and *Budvicia* clades.

Identification of Conserved Signature Indels

Molecular characteristics which are unique to the order Enterobacteriales

In this work, we have completed a comprehensive comparative analysis of the publically available genomes from members of the order "Enterobacteriales" in order to identify discrete markers of common evolutionary ancestry in the form of CSIs. We have identified 69 CSIs which are distinctive characteristics of the order "Enterobacteriales" and its main constituent clades. Five of these CSIs are a shared, distinguishing characteristic of the members of the order "Enterobacteriales" in its entirety. An example of one such CSI, consisting of a single amino acid (aa) insertion in the L-arabinose isomerase protein, is shown in Figure 3. This insertion is present in homologs from all sequenced members (>150) from the order "Enterobacteriales" and is absent in homologs from all other bacteria (top 1000 Blastp hits examined). More detailed information for this CSI is shown in Supplemental Figure 5. Four additional CSIs, which are distinguishing characteristics of the members of the order "Enterobacteriales"**,** were identified in elongation factor P-like protein YeiP, peptide ABC transporter permease, pyrophosphatase, and a hypothetical protein and sequence alignments for these CSIs are shown in Supplemental Figures 6-9 and some properties of these CSIs are briefly summarized in Table 1. The unique shared presence of these CSIs in all of the "Enterobacteriales", but in no other bacteria, except for 1-2 isolated exceptions provides evidence, independent of the phylogenetic trees, that the order "Enterobacteriales" is monophyletic in nature and these CSIs are distinguishing characteristics of this large group of bacteria. Homologs from the genera *Buchnera* and *Wigglesworthia* were not identified in any of the 5 proteins containing CSIs shared by all "Enterobacteriales", while homologs from the genus *Plesiomonas* were only identified for the peptide ABC transporter permease (Supplemental Figure 8) and pyrophosphatase (Supplemental Figure 9). In both cases, the genus *Plesiomonas* did not share the CSI shared by all other "Enterobacteriales".

Molecular characteristics distinguishing the main clades within the order Enterobacteriales

The main focus of this study is on the identification of unique shared characteristics, which can be used to distinguish the main groups within the order "Enterobacteriales". We have identified a total of 66 CSIs which distinguish the 7 main groups of genera within the order "Enterobacteriales", observed in the phylogenetic trees, from each other and from all other bacteria. A number of additional CSIs distinguishing the *Pectobacterium-Dickeya* clade were identified in a previous study (Naushad et al., 2014) whose specificities were re-examined in this work. The identified CSIs which distinguish each of the 7 main clades of the order "Enterobacteriales" are described below.

Clade 1: The *Enterobacter-Escherichia* clade

The members of the genera *Salmonella*, *Citrobacter*, *Escherichia*, and *Shigella* are a well- recognized and highly researched grouping of genera within the order "Enterobacteriales" (Fukushima et al., 2002; Samuel et al., 2004; Nataro et al., 2011; Gordienko et al., 2013).

Escherichia coli, in particular, is one of the most important model organisms in microbiology and has been highly studied and sequenced (Blattner et al., 1997; Chaudhuri & Henderson, 2012; Gordienko et al., 2013; NCBI, 2016). These genera and their closest relatives (viz. *Enterobacter*, *Cronobacter, Klebsiella*, etc.) are the largest grouping of genera within the order "Enterobacteriales". This grouping of genera, labelled the *Enterobacter-Escherichia* clade, is clearly observed in our genome, ribosomal protein, and MLSA based phylogenetic trees and an association between these genera is also seen in 16S rRNA based phylogenies (Figure 1 and Supplemental Figures 1-4). We have identified 21 CSIs which are shared, distinguishing characteristics of the members of the *Enterobacter-Escherichia* clade in our phylogenetic trees, providing evidence that the members of the *Enterobacter-Escherichia* clade form a coherent phylogenetic grouping. An example of a unique, characterizing CSI which is shared by the members of the *Enterobacter-Escherichia* clade is depicted in Figure 4A. The CSI consists of a 3 aa insert in the protein NADH:ubiquinone-oxidoreductase (subunit M), which is present in all of the sequenced species/homologs belonging to this group, and absent in other homologs from the "Enterobacteriales". More detailed information for this signature is shown in Supplemental Figure 10 and the sequence alignments for the 20 other signatures depicting the different identified CSIs which are also distinguishing characteristics of the *Enterobacter-Escherichia* clade are shown in Supplemental Figures 11-30 and their properties are briefly summarized in Table 2.

Clade 2: The *Erwinia-Pantoea* clade

The genera *Erwinia* and *Pantoea* are a well-studied grouping of bacteria containing a number of insect and plant pathogens (Coutinho & Venter, 2009; Zhang & Qiu, 2015). These genera and their closest relatives, labelled the *Erwinia-Pantoea* clade in our phylogenetic trees, exhibit a close association with the members of the *Enterobacter-Escherichia* clade. In our genome, ribosomal protein, and MLSA-based phylogenetic trees the members of the *Erwinia-Pantoea* clade branch as a distinct subgroup in a monophyletic grouping with the *Enterobacter-Escherichia* clade and branch within the *Enterobacter-Escherichia* clade in 16S rRNA based phylogenetic trees (Figure 1 and Supplemental Figures 1-4). We have identified 12 CSIs that are unique distinguishing characteristics of the *Erwinia-Pantoea* clade and an additional 6 CSIs that are shared characteristics of both the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades. An example of each type of CSI is shown here. The first CSI consists of a 1 aa insertion in the protein glutamate-cysteine ligase that is uniquely present in all sequenced members (>20) of the *Erwinia-Pantoea* clade (Figure 4B), while the second CSI consists of a 1 aa insertion in the protein cysteine synthase A that is uniquely present in homologs from members of both the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades (Figure 5A). In both cases similar insertions were not identified in any other related protein homologs from other organisms. More detailed information for these two CSIs as well sequence alignments for the 16 other CSIs, which are specific for either the *Erwinia-Pantoea* clade or supporting a grouping of the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades are shown in Supplemental Figures 31-48 and their properties are briefly summarized in Table 3.

It is of much interest that of the 12 CSI-containing proteins that are distinguishing characteristics of the *Erwinia-Pantoea* clade, homologs for 3 of them were detected in *Buchnera aphidicola* (Figure 4B and Supplementary Figures 31, 36, 41). In each case, *Buchnera aphidicola* shared the characteristic CSIs identified in the CSI-containing proteins with the members of the *Erwinia-Pantoea* clade. Additionally, *Buchnera aphidicola* homologs were identified for 2 proteins containing CSIs shared by both the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades (Figure 5A and Supplementary Figures 43 and 45). These results provide reliable evidence that support previous assertions that *Buchnera aphidicola* is specifically related to the members of the *Erwinia-Pantoea* clade (Husník et al., 2011). Homologs for most of the CSIs-containing proteins shared by the *Erwinia-Pantoea* clade or the *Enterobacter-Escherichia* clade were not found in *Wigglesworthia glossinidia* and, in the few cases where they were found (Supplementary Figures 24 and 36), *Wigglesworthia glossinidia* did not share the CSI with either of the two clades. However, *Wigglesworthia glossinidia* was found to specifically share a CSI in a ribonucleotide reductase stimulatory protein which is a distinguishing characteristic of both the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades (Supplementary Figure 46). This CSI supports the view that *Wigglesworthia glossinidia* is also specifically related to either the *Erwinia-Pantoea* clade or the *Enterobacter-Escherichia* clade, though it is likely a more distant relative of either clade than *Buchnera aphidicola*.

Clade 3: The *Pectobacterium-Dickeya* clade

The members of the genera *Dickeya*, *Pectobacterium*, and *Brenneria* are important plant-pathogenic bacteria (Hauben et al., 1998; Ma et al., 2007; Young & Park, 2007; Zhang et al., 2016). *Dickeya*, *Pectobacterium*, and *Brenneria* branch with the genera *Lonsdalea* and *Sodalis* in our genome and MLSA based phylogenetic trees (Figure 1A and 1C), in a grouping referred to as the *Pectobacterium-Dickeya* clade. However, *Sodalis* does not branch with the other members of this clade in our ribosomal protein based phylogenetic tree (Figure 1B). Here we describe 4 CSIs which are shared by *Brenneria, Dickeya, Lonsdalea, Pectobacterium* and *Sodalis* providing independent evidence of the unique shared ancestry of this group of species. An example of one of these CSIs, consisting of a 2 aa insertion in a hypothetical protein that is uniquely present in homologs from *Brenneria, Dickeya, Lonsdalea, Pectobacterium* and *Sodalis* and absent in all other bacterial groups is shown in Figure 5B. More detailed information for this CSI is shown in Supplemental Figure 49. In earlier work, we have reported 10 CSIs which, at that time, were indicated to be specific for the *Dickeya*, *Pectobacterium*, and *Brenneria* genera (Naushad et al., 2014). A re-examination of these CSIs has shown that 2 of these previously identified CSIs (in a two-component sensor histidine kinase protein and flagellar motor protein MotB) were found in all members of the *Pectobacterium-Dickeya* clade. However, the remaining 8 CSIs identified in our earlier work (not described here) (Naushad et al., 2014) were either not found in homologs from *Sodalis* or the homologs of these proteins were not detected in members of the genus *Sodalis*, and thus they are specific for a subclade of the enlarged *Pectobacterium-Dickeya* clade described here. Sequence alignments for the 3 other CSIs which are distinguishing characteristics of the *Pectobacterium-Dickeya* clade are shown in Supplemental Figures 50-52 and their properties are briefly summarized in Table 4.

Clade 4: The *Yersinia-Serratia* clade

The genus *Yersinia* contains the causative agent of the plague, a disease which led to one of the most devastating pandemics in human history. Consequently, the members of the genus *Yersinia* are the subjects of significant research interest (Perry & Fetherston, 1997; Parkhill et al., 2001; Eppinger et al., 2010; Morelli et al., 2010). In our genome, ribosomal protein, and MLSA based phylogenetic trees (Figure 1A-1C) the members of the genus *Yersinia* are part of a distinct group which contains the genera "Chania", *Ewingella, Rahnella, Rouxiella,* and *Serratia*, referred to as the *Yersinia-Serratia* clade. We have identified 3 CSIs which are shared, distinguishing characteristics of the members of the *Yersinia-Serratia* clade, providing independent evidence that the members of these genera shared a unique common ancestor. One example of such a CSI, shown in Figure 6A, consists of a single aa insertion in the TetR family transcriptional regulator protein found in homologs from the members of the *Yersinia-Serratia* clade. More detailed information for this signature as well as sequence alignments for the 2 other identified CSIs which are distinguishing characteristics of the *Yersinia-Serratia* clade are shown in Supplemental Figures 53-55 and their properties are briefly summarized in Table 4.

Clade 5: The *Hafnia-Edwardsiella* clade

The genera *Edwardsiella, Hafnia,* and *Obesumbacterium* are minor pathogens of animals and humans (Michael & Abbott, 1993; Janda & Abbott, 2006; Koivula et al., 2006; Huys et al., 2010). An association between the genera *Hafnia,* and *Obesumbacterium* has been observed in a number of previous studies (Paradis et al., 2005; Priest & Barker, 2010; Octavia & Lan, 2014), however, the genus *Edwardsiella* shows limited association with the genera *Hafnia* and *Obesumbacterium* in 16S rRNA based phylogenetic trees (Supplemental Figure 4). The genera *Edwardsiella, Hafnia,* and *Obesumbacterium* form a distinct phylogenetic grouping, referred to as the *Hafnia-Edwardsiella* clade, in our genome, ribosomal protein, and MLSA based phylogenetic trees (Figure 1A-1C). We have also identified 4 CSIs which are shared by *Edwardsiella, Hafnia,* and *Obesumbacterium*. An example of one CSI that is uniquely shared by the members of the *Hafnia-Edwardsiella* clade is shown in Figure 6B. This CSI consists of a 4 aa insertion in the two-component system response regulator protein GIrR, which is uniquely found in homologs from *Edwardsiella, Hafnia* and *Obesumbacterium.* More detailed information for this CSI and the sequence alignments for the 3 other CSIs which are distinguishing characteristics of the *Hafnia-Edwardsiella* clade are shown in Supplemental Figures 56-59 and their properties are briefly summarized in Table 4.

Clade 6: The *Proteus-Xenorhabdus* clade

The genera *Xenorhabdus* and *Photorhabdus* are a closely related group of symbiotic bacteria associated with nematode hosts with which they have synergistic entomopathogenic effects against insects (Forst et al., 1997; Nielsen-LeRoux et al., 2012). Previous research has suggested that the closest evolutionary neighbours of *Xenorhabdus* and *Photorhabdus* are the genera *Arsenophonus, Proteus,* and *Providencia* (Boemare & Akhurst, 2006; Trowbridge et al., 2006; Tailliez et al., 2010). However, *Xenorhabdus*, *Photorhabdus*, *Arsenophonus, Proteus,* and *Providencia* do not form a monophyletic clade in 16S rRNA based phylogenetic trees (Figure 1D). In our genome, ribosomal protein, and MLSA based phylogenetic trees (Figure 1A-1C) the genera *Arsenophonus, Moellerella, Morganella, Photorhabdus, Proteus, Providencia,* and *Xenorhabdus* form a distinct, monophyletic grouping, referred to as the *Proteus-Xenorhabdus* clade. We have identified 7 CSIs which are uniquely shared characteristics of the members of the *Proteus-Xenorhabdus* clade. One of these CSIs, a 1 aa deletion in the protein dihydrolipoamide succinyltransferase, in homologs from the *Proteus-Xenorhabdus* clade, is shown in Figure 7A. More detailed information for this CSI as well as the sequence information for the 6 other identified CSIs which are distinguishing characteristics of the *Proteus-Xenorhabdus* clade are shown in Supplemental Figures 60-66 and their properties are briefly summarized in Table 4. These CSIs provide independent evidence in support of the inference from core genome, ribosomal protein, and MLSA-based phylogenetic trees, that the members of the *Proteus-Xenorhabdus* clade form a monophyletic clade derived from a unique common ancestor.

Clade 7: The *Budvicia* clade

The members of the genera *Budvicia, Leminorella,* and *Pragia* are characterized by their H_2S-positive phenotypes and have long been thought to be related (Schindler et al., 1991; Spröer et al., 1999; Paradis et al., 2005; Janda, 2006). A grouping of these three genera, referred to as the *Budvicia* clade, is observed in our genome, ribosomal protein, and MLSA based phylogenetic trees (Figure 1A-1C). A previously reported CSI in the *atpD* gene also supports a specific relationship of the genera *Budvicia, Leminorella,* and *Pragia* (Paradis et al., 2005). Here, we have identified 9 additional CSIs which are shared by these three genera. One example of a CSI shared by the genera *Budvicia, Leminorella,* and *Pragia* is shown in Figure 7B. The CSI consists of a 4 aa insertion in the protein bifunctional Bifunctional protein-disulfide isomerise/oxidoreductase DsbC in homologs from *Budvicia, Leminorella,* and *Pragia* which is absent in homologs from all other species. Detailed information for this signature is shown in Supplemental Figure 67. Sequence alignments for the 8 additional CSIs which are also distinguishing characteristics of the *Budvicia* clade are shown in Supplemental Figures 68-75 and their properties are briefly summarized in Table 4.

Discussion

Understanding the phylogeny and interrelationships of the genera within the order "Enterobacteriales" has proven difficult using the 16S rRNA gene and other single-gene based approaches (Hauben et al., 1998; Spröer et al., 1999; Dauga, 2002; Fukushima et al., 2002; Francino et al., 2006; Nhung et al., 2007; Roggenkamp, 2007; Naum et al., 2008; Tailliez et al., 2010). The advent of ubiquitous genome sequencing technology now presents us with a wealth of genomic sequence data from a broad range of organisms, spanning a majority of the diversity within the order "Enterobacteriales" (NCBI, 2016), from which novel and reliable inferences regarding the evolutionary relationships of the genera within the order "Enterobacteriales" can be drawn. The analyses of the members of the order "Enterobacteriales" presented here, consisting of phylogenetic reconstructions based on 1548 core proteins, 53 ribosomal proteins, and 4 MLSA proteins (Figure 1A-1C), analyses of overall genome similarity (Figure 2), and the identification of shared distinguishing molecular characteristics (Figure 8 and Tables 1-4), represent the first comprehensive, genome-scale taxonomic analysis of the entirety of the order "Enterobacteriales".

The phylogenetic trees produced in this study, utilizing 1548 core proteins, 53 ribosomal proteins, and 4 MLSA proteins from 179 representative genomes from the order "Enterobacteriales", consistently support the existence of the 7 main groups of genera within the order. Additionally, an independently created genome based phylogenetic tree produced by the curators of the PATRIC database (Wattam et al., 2014) utilizing over 1000 genome sequences from members of the order "Enterobacteriales" exhibits highly similar inter-genus level branching to the phylogenetic trees produced in this work and supports the same groupings. The 7 main groupings of genera were also supported by a measure of genomic similarity known as Percentage of Conserved Proteins (Qin et al., 2014) or Alignment Fraction (Varghese et al., 2015) (Figure 2) which is based on the shared gene/protein families present in the genomes. Conversely, phylogenetic trees produced based on the 16S rRNA gene sequence (Figure 1D) exhibit limited ability to resolve the clades identified in the genome, ribosomal protein, and MLSA based phylogenetic trees (Hauben et al., 1998; Naum et al., 2008; Octavia & Lan, 2014).

Additionally, the branching of the genera and species within the order "Enterobacteriales" in 16S rRNA gene based phylogenies shows considerable stochasticity depending on the algorithms used and the organisms analyzed (Naum et al., 2008; Octavia & Lan, 2014). Overall, the results obtained here substantiate previous suggestions that the 16S rRNA gene possesses limited utility in accurate phylogenetic reconstruction of inter-genus level relationships within the order "Enterobacteriales" (Naum et al., 2008; Naum et al., 2011; Octavia & Lan, 2014).

The CSIs identified in this work provide a novel means of elucidating the common evolutionary ancestry of different groups within the order "Enterobacteriales" independently of phylogenetic trees. The most parsimonious explanation of the unique presence of multiple CSIs in a related group of organisms is the existence of a unique shared ancestor in which the genetic changes leading to these CSIs occurred which were then inherited by the descendent species. Thus, CSIs which are restricted to well-defined groups of organisms can be treated synapomorphic traits and used as independent support of monophyletic phylogenetic relationships (Rokas & Holland, 2000; Jones, 2012; Gupta, 2014). Here we describe 71 CSIs which are distinctive characteristics of the order "Enterobacteriales" and its main constituent clades. 5 of the identified CSIs are shared by the entire order "Enterobacteriales", 21 CSIs are shared by the *Enterobacter-Escherichia* clade, 12 CSIs are shared by the *Erwinia-Pantoea* clade, 4 CSIs are shared by the *Pectobacterium-Dickeya* clade, 3 CSIs are shared by the *Yersinia-Serratia* clade, 4 CSIs are shared by the *Hafnia-Edwardsiella* clade, 7 CSIs are shared by the *Proteus-Xenorhabdus* clade, and 9 CSIs are shared by the *Budvicia* clade. Each of these CSIs provide independent support for the branching and the groupings of genera seen in the genome, ribosomal protein, and MLSA based phylogenetic trees produced in this work. Additionally, it is now possible to differentiate these groups of genera from each other and all other bacteria on the basis of the presence or absence of these unique CSIs either *in silico* or utilizing PCR-based assays (Ahmod et al., 2011; Wong et al., 2014).

The single constituent family within the order "Enterobacteriales" contains over 60 genera and 250 species, making the family *Enterobacteriaceae* one of the most taxonomically diverse bacterial families (Parte, 2014; NamesforLife, 2016). The analyses completed in this study, including phylogenetic reconstructions based on 1548 core proteins, 53 ribosomal proteins, and 4 multilocus sequence analysis (MLSA) proteins, analysis of overall genome similarity, and the identification of shared CSIs, strongly support the existence of at least 7 main groups within the order "Enterobacteriales". A division of the family *Enterobacteriaceae* into additional family-level taxa would provide a more coherent taxonomic framework for the order "Enterobacteriales" that more accurately reflects the interrelationships of the various groups of genera within the order. Additionally, a new taxonomic framework for the order "Enterobacteriales" would guide future taxonomic revisions and play a significant role in reducing the prevalence of polyphyletic genera within the order (Brenner & Farmer III, 2005; Brady et al., 2013; Octavia & Lan, 2014). Thus, on the basis of the phylogenetic analyses and utilizing numerous identified conserved molecular characteristics described here, we propose a division of the order "Enterobacteriales" into 7 families: an emended family *Enterobacteriaceae* (The *Enterobacter-Escherichia* clade), *Erwiniaceae* fam. nov. (The *Erwinia-Pantoea* clade), *Pectobacteriaceae* fam. nov. (The *Pectobacterium-Dickeya* clade), *Yersiniaceae* fam. nov. (The *Yersinia-Serratia* clade), *Hafniaceae* fam. nov. (The *Hafnia-Edwardsiella* clade), *Morganellaceae* fam. nov. (The *Proteus-Xenorhabdus* clade), and *Budviciaceae* fam. nov. (The *Budvicia* clade). Genera which are not sequenced (viz. *Biostraticola, Cosenzaea, Enterobacillus, Gibbsiella, Pseudocitrobacter, Rosenbergiella, Saccharobacter*, and *Samsonia*) are placed into one of the families based on 16S rRNA gene sequence identity (Supplemental Table 5). The branching affinity of the genera *Buchnera* and *Wigglesworthia* within the order "Enterobacteriales" has been difficult to resolve in past studies (Lerat et al., 2003; Herbeck et al., 2005; Williams et al., 2010; Husník et al., 2011). Here, we have observed that the genera *Buchnera* and *Wigglesworthia* branch between the *Enterobacter-Escherichia* and the *Erwinia-Pantoea* clades in both the genome and ribosomal protein based phylogenetic trees. Furthermore, the genus *Buchnera* shares 5 CSIs with either the *Erwinia-Pantoea* clade or both the *Enterobacter-Escherichia* and the *Erwinia-Pantoea* clades, while the genus *Wigglesworthia* shares a single CSI with both the *Enterobacter-Escherichia* and the *Erwinia-Pantoea* clades. These findings provide strong suggestive evidence of a specific relationship between the genus *Buchnera* and the *Erwinia-Pantoea* clade and evidence for an association between the genus *Wigglesworthia* and both the *Enterobacter-Escherichia* and the *Erwinia-Pantoea* clades. Thus, at present, the genera *Buchnera* and *Wigglesworthia* will be assigned to the *Erwinia-Pantoea* clade (*Erwiniaceae* fam. nov.). The genus *Plesiomonas* is difficult to place in any of the described families based on phylogeny, CSIs, and 16S rRNA gene sequence identity. Additionally, the homologs of the CSI-containing proteins, specific for all "Enterobacteriales", which were found in the genus *Plesiomonas* did not contain the CSIs shared by all other members of the order "Enterobacteriales". Further, the genus *Plesiomonas* was found to consistently branch either earlier than all other members of the "Enterobacteriales" or with greater affinity to other orders within *Gammaproteobacteria* in phylogenetic trees. These results suggest that the genus *Plesiomonas* has limited association with other members of the order "Enterobacteriales" and it may not belong in the order at all. Thus, the genus *Plesiomonas* will not assigned to any family within the order "Enterobacteriales", at present, and will be considered family *incertae sedis*. A summary of the taxonomic revisions proposed here is available in Figure 8 and descriptions of the new and emended taxa are provided below.

Nomenclature of the order Enterobacteriales

The order "Enterobacteriales" has never been validly published in accordance to the rules of the *International Code of Nomenclature of Bacteria* (Lapage et al., 1992). The latest edition of *Bergey's Manual of Systematic Bacteriology* lists the type genus of the order "Enterobacteriales" as *Escherichia*, which is the same as the type genus of the family *Enterobacteriaceae* (Imhoff, 2005). However, the name *Enterobacteriaceae* predates the *International Code of Nomenclature of Bacteria* and its original derivation is uncertain (Judicial Commission of the International Committee on Systematic Bacteriology, 1981). The name *Enterobacteriaceae* was validated by the Judicial Commission of the International Committee on Systematic Bacteriology with the type genus *Escherichia* for historical reasons, despite this nomenclature not being in accordance to the rules of the *International Code of Nomenclature of Bacteria* (Wayne, 1982; Brenner, 1983). Thus, an order with the type genus *Escherichia* should be named "Escherichiales", not "Enterobacteriales", according to the rules of the *International Code of Nomenclature of Bacteria* (Lapage et al., 1992). Furthermore, an order with the type genus *Enterobacter* should be named "Enterobacterales" not "Enterobacteriales". To limit the confusion regarding the nomenclature of the "Enterobacteriales" which could arise if the name "Escherichiales" were to be used to describe the order, we have chosen to utilize the name *Enterobacterales* ord. nov. with the type genus *Enterobacter* to describe the order containing the family *Enterobacteriaceae*. A description of the order *Enterobacterales* is provided below.

Description of the order *Enterobacterales* ord. nov.

Enterobacterales (En.te.ro.bac.te.r.a'les. N.L. n. *Enterobacter* the type genus of the order; *-ales* ending to denote an order; N.L. fem. pl. n. *Enterobacterales* the order whose nomenclatural type is the genus *Enterobacter*)

The *Enterobacterales* are an order of gram negative, non-spore forming, rod shaped facultative anaerobes. The order contains the type genus *Enterobacter* (Rahn, 1937) as well as the families *Enterobacteriaceae* (Rahn, 1937), *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. The description of the order is the same as that of the family *Enterobacteriaceae* given by Brenner and Farmer III (2005) with the following modifications: the members of the order *Enterobacterales* can be distinguished from all other bacteria by the 5 conserved signature indels in the proteins peptide ABC transporter permease, elongation factor P-like protein YeiP, L-arabinose isomerase, pyrophosphatase, and a hypothetical protein (Table 1).

Emended Description of the family *Enterobacteriaceae* Rahn 1937 (Approved Lists 1980)

The family *Enterobacteriaceae* contains the type genus *Escherichia* (Castellani & Chambers, 1919; Lapage et al., 1992) and the genera "Atlantibacter" (Hata et al., 2016), *Biostraticola* (Verbarg et al., 2008), *Buttiauxella* (Ferragut et al., 1981), *Cedecea* (Grimont et al., 1981), *Citrobacter* (Werkman & Gillen, 1932), *Cronobacter* (Iversen et al., 2008), *Enterobacillus* (Patil et al., 2015), *Enterobacter* (Rahn, 1937), *Franconibacter* (Stephan et al., 2014), *Gibbsiella* (Brady et al., 2010a), *Izhakiella* (Aizenberg-Gershtein et al., 2016), *Klebsiella* (Drancourt et al., 2001), *Kluyvera* (Farmer et al., 1981), *Kosakonia* (Brady et al., 2013), *Leclercia* (Tamura et al., 1986), *Lelliottia* (Brady et al., 2013), *Mangrovibacter* (Rameshkumar et al., 2010), *Pluralibacter* (Brady et al., 2013), *Pseudocitrobacter* (Kämpfer et al., 2014), *Raoultella* (Drancourt et al., 2001), *Rosenbergiella* (Halpern et al., 2013b), *Saccharobacter* (Yaping et al., 1990), *Salmonella* (Lignieres, 1900), *Shigella* (Castellani & Chambers, 1919), *Shimwellia* (Priest & Barker, 2010), *Siccibacter* (Stephan et al., 2014), *Trabulsiella* (McWhorter et al., 1991), and *Yokenella* (Kosako et al., 1984). All genera belonging to this group are catalase positive and oxidase negative. Members of the family *Enterobacteriaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from all other members of the order *Enterobacterales* by 21 conserved signature indels in the proteins NADH:ubiquinone oxidoreductase (subunit M), Twitching motility protein PilT, 2,3-dihydroxybenzoate-AMP ligase, ATP/GTP-binding protein, Multifunctional fatty acid oxidation complex (subunit alpha), S-formylglutathione hydrolase, Aspartate-semialdehyde dehydrogenase, Epimerase, Membrane protein, Formate dehydrogenylase (subunit 7), Glutathione S-transferase, Major facilitator superfamily transporter, Phosphoglucosamine mutase, Glycosyl hydrolase 1 family protein, 23S rrna (uracil(1939)-C(5))-methyltransferase, Co-chaperone HscB, N-acetylmuramoyl-L-alanine amidase, Sulfate ABC transporter ATP-binding protein CysA, and LPS assembly protein LptD (Table 2).

Description of *Erwiniaceae* fam. nov.

Erwiniaceae (Er.wi.ni.a.ce'ae. N.L. fem. n. *Erwinia* type genus of the family; *-aceae* ending to denote a family; N.L. fem. pl. n. *Erwiniaceae* the family whose nomenclatural type is the genus *Erwinia*)

The family *Erwiniaceae* contains the type genus *Erwinia* (Hauben et al., 1998) and the genera *Buchnera* (Munson et al., 1991), *Pantoea* (Brady et al., 2010b), *Phaseolibacter* (Halpern et al., 2013a), *Tatumella* (Hollis et al., 1981) and *Wigglesworthia* (Aksoy, 1995). These bacteria are catalase positive, oxidase negative, and do not produce indole or hydrogen disulfide. Most *Erwiniaceae* species are positive for Voges-Proskauer, with the exception of *Erwinia toletena, E. typographi* and some strains of *E. olae.* Members of the family *Erwiniaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from the all other bacteria by 12 conserved signature indels in the proteins Glutamate--cysteine ligase, DNA gyrase (subunit B), LPS assembly protein LptD, Thiol:disulfide interchange protein DsbA precursor, Two-component sensor histidine kinase, RNA helicase, tRNA pseudouridine(13) synthase TruD, Glycine/betaine ABC transporter ATP-binding protein, Superoxide dismutase, and Stationary phase inducible protein CsiE (Table 3).

Description of *Pectobacteriaceae* fam. nov.

Pectobacteriaceae (Pec.to.bac.te.ri.a.ce'ae N.L. neut. n. *Pectobacterium* type genus of the family; *-aceae* ending to denote a family; N.L. fem. pl. n. *Pectobacteriaceae* the family whose nomenclatural type is the genus *Pectobacterium*)

The family *Pectobacteriaceae* contains the type genus *Pectobacterium* (Hauben et al., 1998) and the genera *Brenneria* (Brady et al., 2014a), *Dickeya* (Gardan, 2005), *Lonsdalea* (Brady et al., 2012), and *Sodalis* (Dale & Maudlin, 1999). *Pectobacteriaceae* species produce acid from N-acetylglucosamine and are negative for arginine dihydrolase, orthinine decarboxylase and lysine decarboxylase. These bacteria are catalase positive, oxidase negative, and do not produce hydrogen disulfide. Members of the family *Pectobacteriaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from the all other bacteria by 4 conserved signature indels in the proteins Transcriptional activator RhaS, Flagellar motor protein MotB, a Two-component sensor histidine kinase protein and a Hypothetical protein (Table 4).

Description of *Yersiniaceae* fam. nov.

Yersiniaceae (Yer.si.ni.a.ce'ae. N.L. fem. n. *Yersinia* type genus of the family; *-aceae* ending to denote a family; N.L. fem. pl. n. *Yersiniaceae* the family whose nomenclatural type is the genus *Yersinia*)

The family *Yersiniaceae* contains the type genus *Yersinia* (Van Loghem, 1944) and the genera "Chania" (Ee et al., 2016), *Ewingella* (Grimont et al., 1983), *Rahnella* (Izard et al., 1978), *Rouxiella* (Le Fleche-Mateos et al., 2015), *Samsonia* (Sutra et al., 2001), and *Serratia* (Bizio, 1823). These bacteria are motile, catalase positive, and do not produce hydrogen disulfide. Members of the family *Yersiniaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from the all other bacteria by 3 conserved signature indels in the protein TetR family transcriptional regulator and a Hypothetical protein (Table 4).

Description of *Hafniaceae* fam. nov.

Hafniaceae (Haf.ni.a.ce'ae. N.L. fem. n. *Hafnia* type genus of the family; *-aceae* ending to denote a family; N.L. fem. pl. n. *Hafniaceae* the family whose nomenclatural type is the genus *Hafnia*)

The family *Hafniaceae* contains the type genus *Hafnia* (Møller, 1954) and the genera *Edwardsiella* (Ewing et al., 1965) and *Obesumbacterium* (Shimwell, 1963). *Hafniaceae* species are catalase positive, oxidase negative, and are negative for lysine decarboxylase. These bacteria are also able to grow on MacConkey media, and are capable of reducing nitrate. Members of the family *Hafniaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from the all other bacteria by 4 conserved signature indels in the proteins Two-component system response regulator GIrR, Glucose-1-phosphate adenylyltransferase, Transcriptional activator NhaR, and the Hybrid sensor histidine kinase/response regulator (Table 4).

Description of *Morganellaceae* fam. nov.

Morganellaceae (Mor.ga.nel.la.ce'ae. N.L. fem. n. *Morganella* the type genus of the family; -*aceae* ending to denote a family; N.L. fem. pl. n. *Morganellaceae* the family whose nomenclatural type is the genus *Morganella*)

The family *Morganellaceae* contains the type genus *Morganella* (Fulton, 1943) and the genera *Arsenophonus* (Gherna et al., 1991), *Cosenzaea* (Giammanco et al., 2011), *Moellerella* (Hickman-Brenner et al., 1984), *Photorhabdus* (Boemare et al., 1993), *Proteus* (Hauser, 1885), *Providencia* (Ewing, 1962), and *Xenorhabdus* (Thomas & Poinar Jr, 1979). These bacteria are oxidase negative, and negative for arginine decarboxylase and Voges-Proskauer. Members of the family *Morganellaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from the all other bacteria by 7 conserved signature indels in the proteins Dihydrolipoamide succinyltransferase, Xaa-Pro dipeptidase, Bifunctional UDP-sugar hydrolase (5'-nucleotidase), Transcriptional repair coupling factor, Phosphate acetyltransferase, Histidine—tRNA ligase, and N-acetylmuramoyl-L-alanine amidase (Table 4).

Description of *Budviciaceae* fam. nov.

Budviciaceae (Bud.vi.ci.a.ce'ae. L. fem. n. *Budvicia* type genus of the family; -*aceae* ending to denote a family; N.L. fem. pl. n. *Budviciaceae* the family whose nomenclatural type is the genus *Budvicia*)

The family *Budviciaceae* contains the type genus *Budvicia* (Lang et al., 2013) and the genera *Leminorella* (Hickman-Brenner et al., 1985) and *Pragia* (Aldová et al., 1988). *Budviciaceae* species are catalase positive, oxidase negative, and negative for indole, arginine dihydrolase, orthinine decarboxylase, and lysing decarboxylase. These bacteria are capable of producing hydrogen disulfide and reducing nitrate, but are incapable of growing on KCN media.

Members of the family *Budviciaceae* form a distinct monophyletic cluster in genome and multi-gene based phylogenetic trees and can be distinguished from the all other bacteria by 9 conserved signature indels in the proteins Bifunctional protein-disulfide isomerise/oxidoreductase DsbC, L-methionine/branched chain amino acid transporter, D-alanine—D-alanine ligase, and Hypothetical proteins (Table 4).

Acknowledgements

This work was supported by the research grant No. 249924 from the Natural Science and Engineering Research Council of Canada to RSG. We thank Sanam Gurm and Bo "Jim" Zeng for their expert technical assistance in identifying some of the CSIs described in this work.

Works Cited

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

Aizenberg-Gershtein, Y., Laviad, S., Samuni-Blank, M., & Halpern, M. (2016). *Izhakiella capsodis* gen. nov. sp. nov., in the family *Enterobacteriaceae*, isolated from the mirid bug Capsodes infuscatus. *Int J Syst Evol Microbiol*.

Aksoy, S. (1995). *Wigglesworthia* gen. nov. and *Wigglesworthia glossinidia* sp. nov., taxa consisting of the mycetocyte-associated, primary endosymbionts of tsetse flies. *Int J Syst Evol Microbiol, 45*(4), 848-851.

Aldová, E., Hausner, O., Brenner, D. J., Kocmoud, Z., Schindler, J., Potužníková, B., & Petráš, P. (1988). *Pragia fontium* gen. nov., sp. nov. of the family *Enterobacteriaceae*, isolated from water. *Int J Syst Evol Microbiol, 38*(2), 183-189.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Auch, A. F., von Jan, M., Klenk, H.-P., & Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci, 2*(1), 117-134.

Bergsten, J. (2005). A review of long - branch attraction. *Cladistics, 21*(2), 163-193.

Bizio, B. (1823). Lettera di Bartolomeo Bizio al chiarissimo canonico Angelo Bellani sopra il fenomeno della polenta porporina. *Biblioteca Italiana o sia Giornale di Letteratura, Scienze e Arti, 30*, 275-295.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science, 277*(5331), 1453-1462.

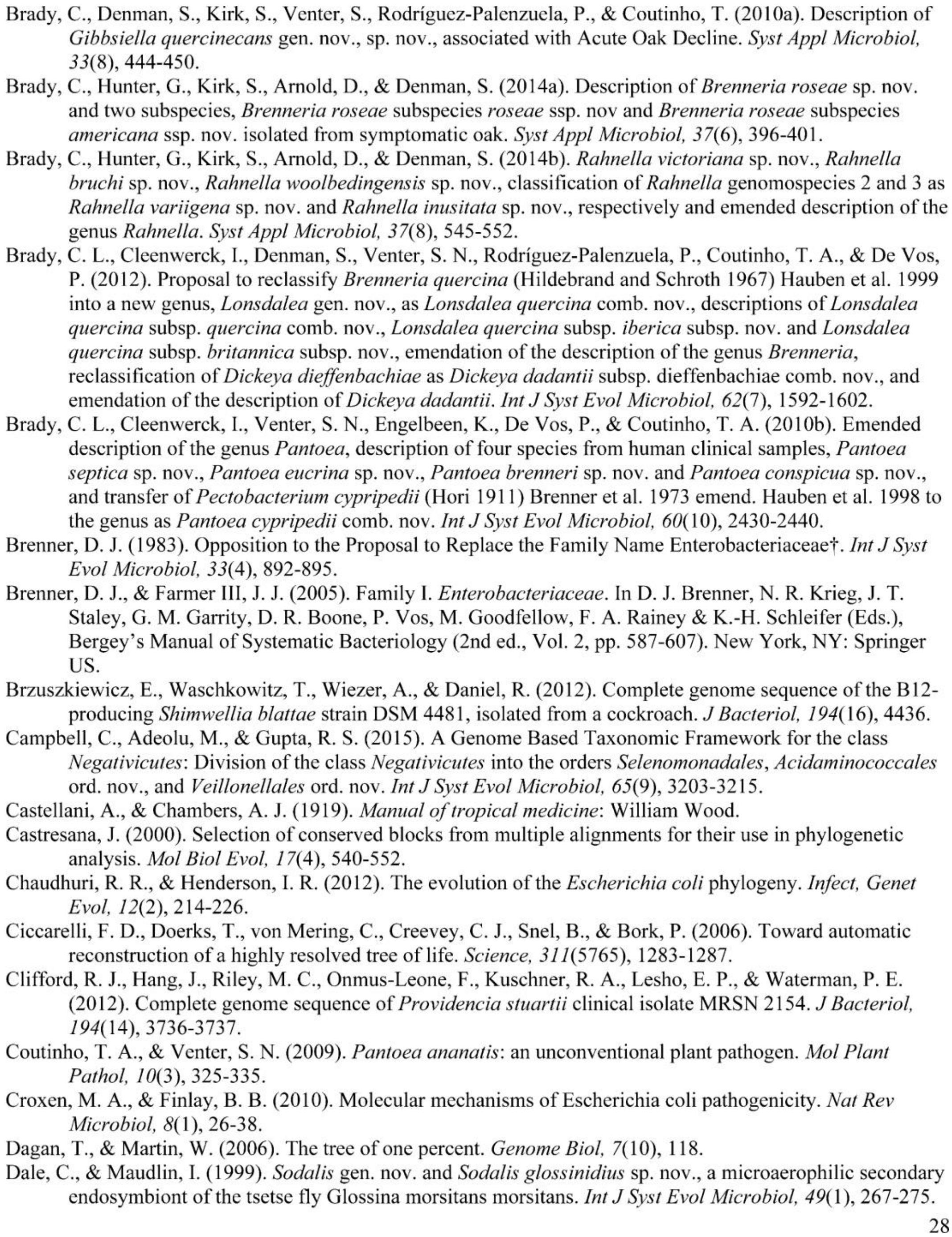
Boemare, N., & Akhurst, R. (2006). The genera *Photorhabdus* and *Xenorhabdus* The prokaryotes (pp. 451-494): Springer.

Boemare, N., Akhurst, R., & Mourant, R. (1993). DNA relatedness between *Xenorhabdus* spp.(*Enterobacteriaceae*), symbiotic bacteria of entomopathogenic nematodes, and a proposal to transfer *Xenorhabdus luminescens* to a new genus, *Photorhabdus* gen. nov. *Int J Syst Evol Microbiol, 43*(2), 249-255.

Bonn, W. G., & van der Zwet, T. (2000). Distribution and economic importance of fire blight. *Fire blight: the disease and its causative agent, Erwinia amylovora*, 37-53.

Brady, C., Cleenwerck, I., Venter, S., Coutinho, T., & De Vos, P. (2013). Taxonomic evaluation of the genus *Enterobacter* based on multilocus sequence analysis (MLSA): proposal to reclassify *E. nimipressuralis* and *E. amnigenus* into *Lelliottia* gen. nov. as *Lelliottia nimipressuralis* comb. nov. and *Lelliottia amnigena* comb. nov., respectively, *E. gergoviae* and *E. pyrinus* into *Pluralibacter* gen. nov. as *Pluralibacter gergoviae* comb. nov. and *Pluralibacter pyrinus* comb. nov., respectively, *E. cowanii*, *E. radicincitans*, *E. oryzae* and *E. arachidis* into *Kosakonia* gen. nov. as *Kosakonia cowanii* comb. nov., *Kosakonia radicincitans* comb. nov., *Kosakonia oryzae* comb. nov. and *Kosakonia arachidis* comb. nov., respectively, and *E. turicensis*, *E. helveticus* and *E. pulveris* into *Cronobacter* as *Cronobacter zurichensis* nom. nov., *Cronobacter helveticus* comb. nov. and *Cronobacter pulveris* comb. nov., respectively, and emended description of the genera *Enterobacter* and *Cronobacter*. *Syst Appl Microbiol, 36*(5), 309-319.

Brady, C., Cleenwerck, I., Venter, S., Vancanneyt, M., Swings, J., & Coutinho, T. (2008). Phylogeny and identification of *Pantoea* species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). *Syst Appl Microbiol, 31*(6-8), 447-460.



Brady, C., Denman, S., Kirk, S., Venter, S., Rodríguez-Palenzuela, P., & Coutinho, T. (2010a). Description of *Gibbsiella quercinecans* gen. nov., sp. nov., associated with Acute Oak Decline. *Syst Appl Microbiol, 33*(8), 444-450.

Brady, C., Hunter, G., Kirk, S., Arnold, D., & Denman, S. (2014a). Description of *Brenneria roseae* sp. nov. and two subspecies, *Brenneria roseae* subspecies *roseae* ssp. nov and *Brenneria roseae* subspecies *americana* ssp. nov. isolated from symptomatic oak. *Syst Appl Microbiol, 37*(6), 396-401.

Brady, C., Hunter, G., Kirk, S., Arnold, D., & Denman, S. (2014b). *Rahnella victoriana* sp. nov., *Rahnella bruchi* sp. nov., *Rahnella woolbedingensis* sp. nov., classification of *Rahnella* genomospecies 2 and 3 as *Rahnella variigena* sp. nov. and *Rahnella inusitata* sp. nov., respectively and emended description of the genus *Rahnella*. *Syst Appl Microbiol, 37*(8), 545-552.

Brady, C. L., Cleenwerck, I., Denman, S., Venter, S. N., Rodríguez-Palenzuela, P., Coutinho, T. A., & De Vos, P. (2012). Proposal to reclassify *Brenneria quercina* (Hildebrand and Schroth 1967) Hauben et al. 1999 into a new genus, *Lonsdalea* gen. nov., as *Lonsdalea quercina* comb. nov., descriptions of *Lonsdalea quercina* subsp. *quercina* comb. nov., *Lonsdalea quercina* subsp. *iberica* subsp. nov. and *Lonsdalea quercina* subsp. *britannica* subsp. nov., emendation of the description of the genus *Brenneria*, reclassification of *Dickeya dieffenbachiae* as *Dickeya dadantii* subsp. dieffenbachiae comb. nov., and emendation of the description of *Dickeya dadantii*. *Int J Syst Evol Microbiol, 62*(7), 1592-1602.

Brady, C. L., Cleenwerck, I., Venter, S. N., Engelbeen, K., De Vos, P., & Coutinho, T. A. (2010b). Emended description of the genus *Pantoea*, description of four species from human clinical samples, *Pantoea septica* sp. nov., *Pantoea eucrina* sp. nov., *Pantoea brenneri* sp. nov. and *Pantoea conspicua* sp. nov., and transfer of *Pectobacterium cypripedii* (Hori 1911) Brenner et al. 1973 emend. Hauben et al. 1998 to the genus as *Pantoea cypripedii* comb. nov. *Int J Syst Evol Microbiol, 60*(10), 2430-2440.

Brenner, D. J. (1983). Opposition to the Proposal to Replace the Family Name Enterobacteriaceae†. *Int J Syst Evol Microbiol, 33*(4), 892-895.

Brenner, D. J., & Farmer III, J. J. (2005). Family I. *Enterobacteriaceae*. In D. J. Brenner, N. R. Krieg, J. T. Staley, G. M. Garrity, D. R. Boone, P. Vos, M. Goodfellow, F. A. Rainey & K.-H. Schleifer (Eds.), Bergey's Manual of Systematic Bacteriology (2nd ed., Vol. 2, pp. 587-607). New York, NY: Springer US.

Brzuszkiewicz, E., Waschkowitz, T., Wiezer, A., & Daniel, R. (2012). Complete genome sequence of the B12-producing *Shimwellia blattae* strain DSM 4481, isolated from a cockroach. *J Bacteriol, 194*(16), 4436.

Campbell, C., Adeolu, M., & Gupta, R. S. (2015). A Genome Based Taxonomic Framework for the class *Negativicutes*: Division of the class *Negativicutes* into the orders *Selenomonadales*, *Acidaminococcales* ord. nov., and *Veillonellales* ord. nov. *Int J Syst Evol Microbiol, 65*(9), 3203-3215.

Castellani, A., & Chambers, A. J. (1919). *Manual of tropical medicine*: William Wood.

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

Chaudhuri, R. R., & Henderson, I. R. (2012). The evolution of the *Escherichia coli* phylogeny. *Infect, Genet Evol, 12*(2), 214-226.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science, 311*(5765), 1283-1287.

Clifford, R. J., Hang, J., Riley, M. C., Onmus-Leone, F., Kuschner, R. A., Lesho, E. P., & Waterman, P. E. (2012). Complete genome sequence of *Providencia stuartii* clinical isolate MRSN 2154. *J Bacteriol, 194*(14), 3736-3737.

Coutinho, T. A., & Venter, S. N. (2009). *Pantoea ananatis*: an unconventional plant pathogen. *Mol Plant Pathol, 10*(3), 325-335.

Croxen, M. A., & Finlay, B. B. (2010). Molecular mechanisms of Escherichia coli pathogenicity. *Nat Rev Microbiol, 8*(1), 26-38.

Dagan, T., & Martin, W. (2006). The tree of one percent. *Genome Biol, 7*(10), 118.

Dale, C., & Maudlin, I. (1999). *Sodalis* gen. nov. and *Sodalis glossinidius* sp. nov., a microaerophilic secondary endosymbiont of the tsetse fly Glossina morsitans morsitans. *Int J Syst Evol Microbiol, 49*(1), 267-275.

28

Dauga, C. (2002). Evolution of the gyrB gene and the molecular phylogeny of *Enterobacteriaceae*: a model molecule for molecular systematic studies. *Int J Syst Evol Microbiol, 52*(2), 531-547.

Drancourt, M., Bollet, C., Carta, A., & Rousselier, P. (2001). Phylogenetic analyses of *Klebsiella* species delineate *Klebsiella* and *Raoultella* gen. nov., with description of *Raoultella ornithinolytica* comb. nov., *Raoultella terrigena* comb. nov. and *Raoultella planticola* comb. nov. *Int J Syst Evol Microbiol, 51*(3), 925-932.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol, 7*(10), e1002195.

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

Ee, R., Madhaiyan, M., Ji, L., Lim, Y.-L., Nor, N. M., Tee, K.-K., Chen, J.-W., & Yin, W.-F. (2016). *Chania multitudinisentens* gen. nov., sp. nov., a N-acyl-homoserine lactone-producing bacterium in the family of *Enterobacteriaceae* isolated from a former municipal landfill site soil. *Int J Syst Evol Microbiol*.

Eppinger, M., Worsham, P. L., Nikolich, M. P., Riley, D. R., Sebastian, Y., Mou, S., Achtman, M., Lindler, L. E., & Ravel, J. (2010). Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. *J Bacteriol, 192*(6), 1685-1699.

Ewing, W. (1962). The tribe Proteae: its nomenclature and taxonomy. *Int J Syst Evol Microbiol, 12*(3), 93-102.

Ewing, W., McWhorter, A., Escobar, M., & Lubin, A. (1965). *Edwardsiella*, a new genus of *Enterobacteriaceae* based on a new species, *E. tarda*. *Int J Syst Evol Microbiol, 15*(1), 33-38.

Farmer, J., Fanning, G., Huntley-Carter, G., Holmes, B., Hickman, F., Richard, C., & Brenner, D. (1981). *Kluyvera*, a new (redefined) genus in the family *Enterobacteriaceae*: identification of *Kluyvera ascorbata* sp. nov. and *Kluyvera cryocrescens* sp. nov. in clinical specimens. *J Clin Microbiol, 13*(5), 919-933.

Ferragut, C., Izard, D., Gavini, F., Lefebvre, B., & Leclerc, H. (1981). *Buttiauxella*, a new genus of the family *Enterobacteraceae*. *Zentralblatt für Bakteriologie Mikrobiologie und Hygiene: I Abt Originale C: Allgemeine, angewandte und ökologische Mikrobiologie, 2*(1), 33-44.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res, 44*(D1), D279-285.

Forst, S., Dowds, B., Boemare, N., & Stackebrandt, E. (1997). *Xenorhabdus* and *Photorhabdus* spp.: bugs that kill bugs. *Annual Reviews in Microbiology, 51*(1), 47-72.

Francino, M. P., Santos, S. R., & Ochman, H. (2006). Phylogenetic relationships of bacteria with special reference to endosymbionts and enteric species The Prokaryotes (pp. 41-59): Springer.

Fukushima, M., Kakinuma, K., & Kawaguchi, R. (2002). Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. *J Clin Microbiol, 40*(8), 2779-2785.

Fulton, M. (1943). The identity of Bacterium columbensis Castellani. *J Bacteriol, 46*(1), 79.

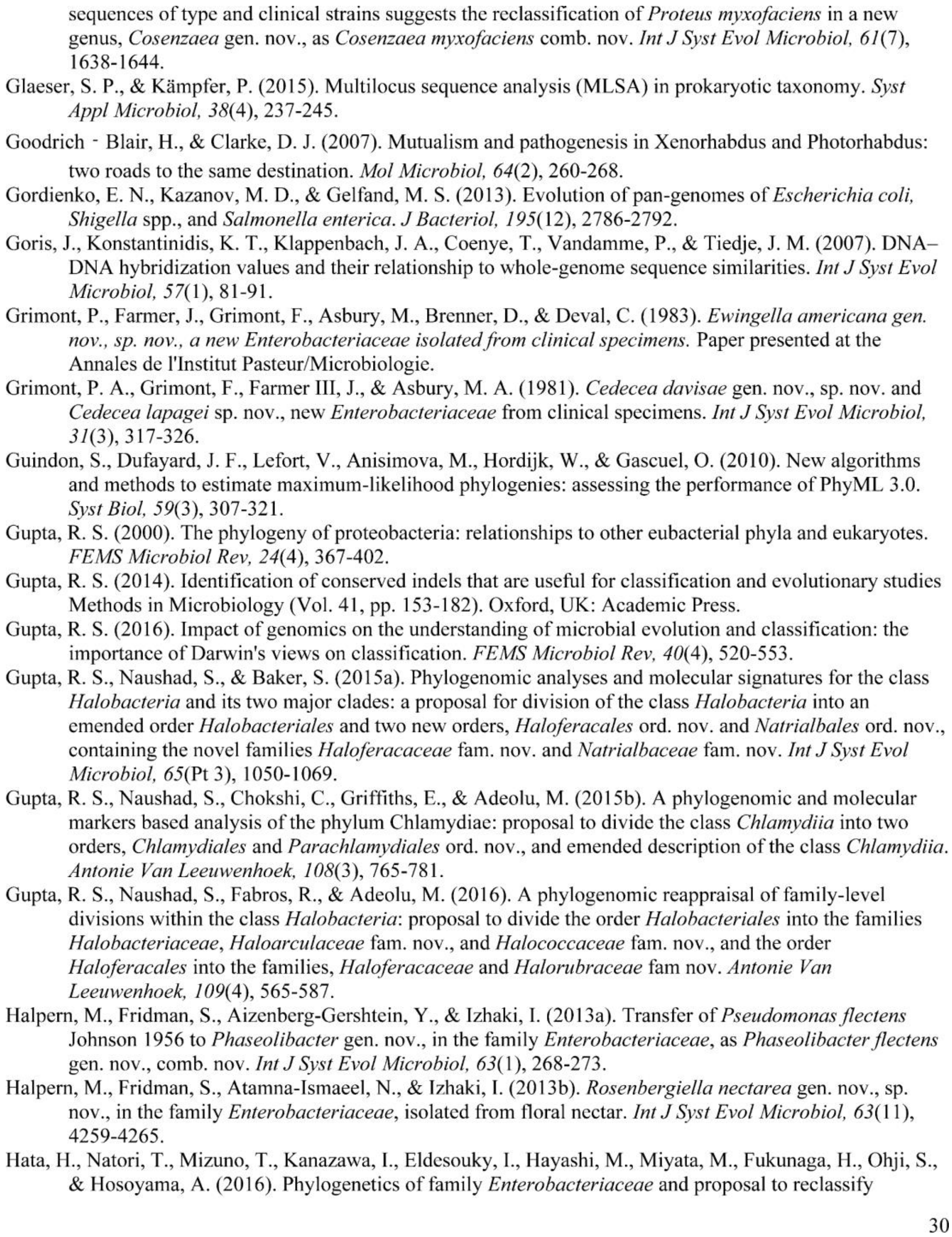
Gao, B., Mohan, R., & Gupta, R. S. (2009). Phylogenomics and protein signatures elucidating the evolutionary relationships among the *Gammaproteobacteria*. *Int J Syst Evol Microbiol, 59*(2), 234-247.

Gardan, L. (2005). Transfer of *Pectobacterium chrysanthemi* (Burkholder et al. 1953) Brenner et al. 1973 and Brenneria paradisiaca to the genus *Dickeya* gen. nov. as *Dickeya chrysanthemi* comb. nov. and *Dickeya paradisiaca* comb. nov. and delineation of four novel species, *Dickeya dadantii* sp. nov., *Dickeya dianthicola* sp. nov., *Dickeya dieffenbachiae* sp. nov. and *Dickeya zeae* sp. nov. *Int J Syst Evol Microbiol, 55*(14151427), 0.02791-02790.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., & Thompson, F. L. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology, 3*(9), 733-739.

Gherna, R., Werren, J., Weisburg, W., Cote, R., Woeste, C., Mandelco, L., & Brenner, D. (1991). *Arsenophonus nasoniae* gen. nov., sp. nov., the causative agent of the son-killer trait in the parasitic wasp Nasonia vitripennis. *International journal of systematic bacteriology (USA)*.

Giammanco, G. M., Grimont, P. A., Grimont, F., Lefevre, M., Giammanco, G., & Pignato, S. (2011). Phylogenetic analysis of the genera *Proteus*, *Morganella* and *Providencia* by comparison of *rpoB* gene



sequences of type and clinical strains suggests the reclassification of *Proteus myxofaciens* in a new genus, *Cosenzaea* gen. nov., as *Cosenzaea myxofaciens* comb. nov. *Int J Syst Evol Microbiol, 61*(7), 1638-1644.

Glaeser, S. P., & Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst Appl Microbiol, 38*(4), 237-245.

Goodrich ‐ Blair, H., & Clarke, D. J. (2007). Mutualism and pathogenesis in Xenorhabdus and Photorhabdus: two roads to the same destination. *Mol Microbiol, 64*(2), 260-268.

Gordienko, E. N., Kazanov, M. D., & Gelfand, M. S. (2013). Evolution of pan-genomes of *Escherichia coli, Shigella* spp., and *Salmonella enterica*. *J Bacteriol, 195*(12), 2786-2792.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol, 57*(1), 81-91.

Grimont, P., Farmer, J., Grimont, F., Asbury, M., Brenner, D., & Deval, C. (1983). *Ewingella americana gen. nov., sp. nov., a new Enterobacteriaceae isolated from clinical specimens.* Paper presented at the Annales de l'Institut Pasteur/Microbiologie.

Grimont, P. A., Grimont, F., Farmer III, J., & Asbury, M. A. (1981). *Cedecea davisae* gen. nov., sp. nov. and *Cedecea lapagei* sp. nov., new *Enterobacteriaceae* from clinical specimens. *Int J Syst Evol Microbiol, 31*(3), 317-326.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol, 59*(3), 307-321.

Gupta, R. S. (2000). The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev, 24*(4), 367-402.

Gupta, R. S. (2014). Identification of conserved indels that are useful for classification and evolutionary studies Methods in Microbiology (Vol. 41, pp. 153-182). Oxford, UK: Academic Press.

Gupta, R. S. (2016). Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. *FEMS Microbiol Rev, 40*(4), 520-553.

Gupta, R. S., Naushad, S., & Baker, S. (2015a). Phylogenomic analyses and molecular signatures for the class *Halobacteria* and its two major clades: a proposal for division of the class *Halobacteria* into an emended order *Halobacteriales* and two new orders, *Haloferacales* ord. nov. and *Natrialbales* ord. nov., containing the novel families *Haloferacaceae* fam. nov. and *Natrialbaceae* fam. nov. *Int J Syst Evol Microbiol, 65*(Pt 3), 1050-1069.

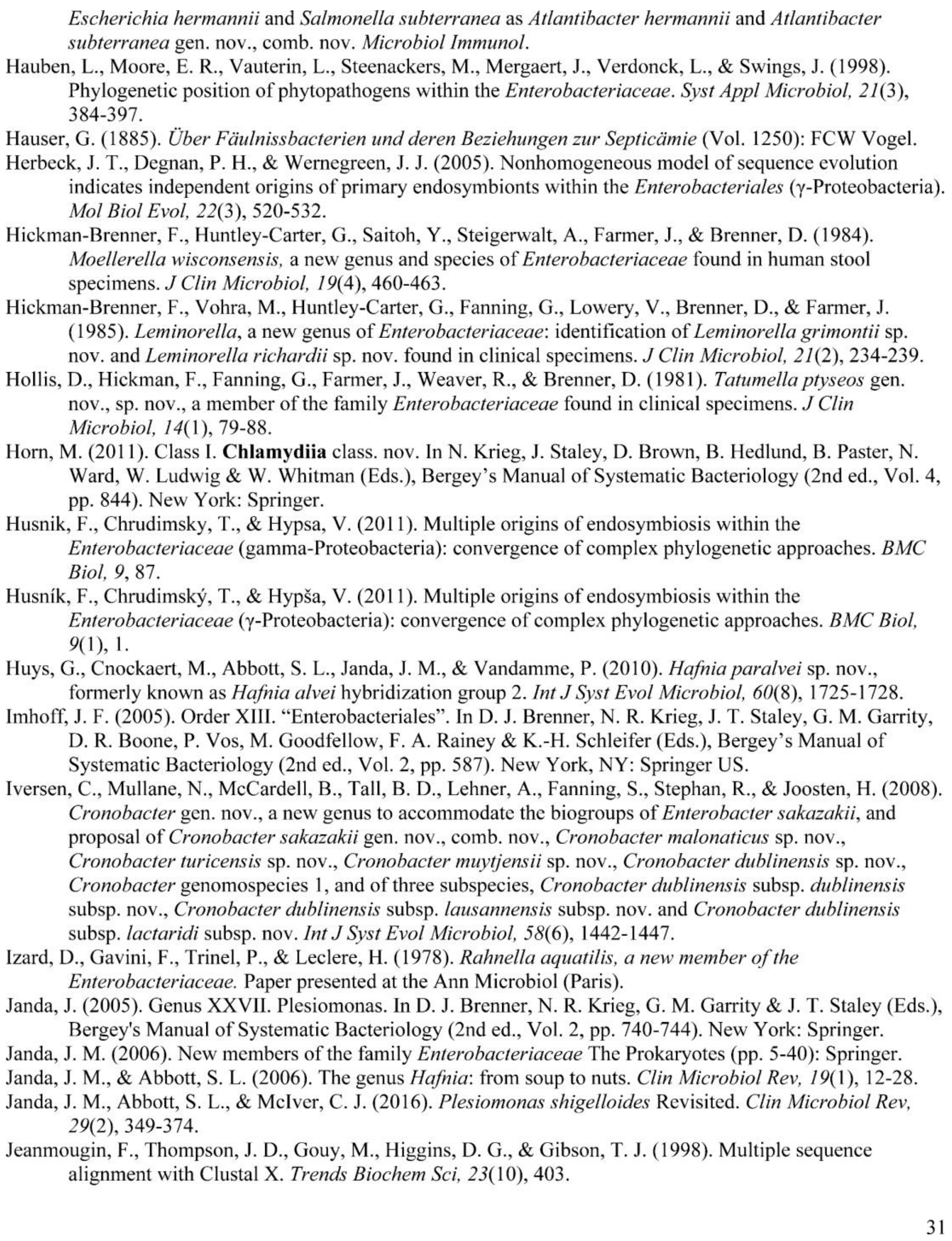
Gupta, R. S., Naushad, S., Chokshi, C., Griffiths, E., & Adeolu, M. (2015b). A phylogenomic and molecular markers based analysis of the phylum Chlamydiae: proposal to divide the class *Chlamydiia* into two orders, *Chlamydiales* and *Parachlamydiales* ord. nov., and emended description of the class *Chlamydiia*. *Antonie Van Leeuwenhoek, 108*(3), 765-781.

Gupta, R. S., Naushad, S., Fabros, R., & Adeolu, M. (2016). A phylogenomic reappraisal of family-level divisions within the class *Halobacteria*: proposal to divide the order *Halobacteriales* into the families *Halobacteriaceae*, *Haloarculaceae* fam. nov., and *Halococcaceae* fam. nov., and the order *Haloferacales* into the families, *Haloferacaceae* and *Halorubraceae* fam nov. *Antonie Van Leeuwenhoek, 109*(4), 565-587.

Halpern, M., Fridman, S., Aizenberg-Gershtein, Y., & Izhaki, I. (2013a). Transfer of *Pseudomonas flectens* Johnson 1956 to *Phaseolibacter* gen. nov., in the family *Enterobacteriaceae*, as *Phaseolibacter flectens* gen. nov., comb. nov. *Int J Syst Evol Microbiol, 63*(1), 268-273.

Halpern, M., Fridman, S., Atamna-Ismaeel, N., & Izhaki, I. (2013b). *Rosenbergiella nectarea* gen. nov., sp. nov., in the family *Enterobacteriaceae*, isolated from floral nectar. *Int J Syst Evol Microbiol, 63*(11), 4259-4265.

Hata, H., Natori, T., Mizuno, T., Kanazawa, I., Eldesouky, I., Hayashi, M., Miyata, M., Fukunaga, H., Ohji, S., & Hosoyama, A. (2016). Phylogenetics of family *Enterobacteriaceae* and proposal to reclassify



Escherichia hermannii and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter subterranea* gen. nov., comb. nov. *Microbiol Immunol*.

Hauben, L., Moore, E. R., Vauterin, L., Steenackers, M., Mergaert, J., Verdonck, L., & Swings, J. (1998). Phylogenetic position of phytopathogens within the *Enterobacteriaceae*. *Syst Appl Microbiol, 21*(3), 384-397.

Hauser, G. (1885). *Über Fäulnissbacterien und deren Beziehungen zur Septicämie* (Vol. 1250): FCW Vogel.

Herbeck, J. T., Degnan, P. H., & Wernegreen, J. J. (2005). Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the *Enterobacteriales* (γ-Proteobacteria). *Mol Biol Evol, 22*(3), 520-532.

Hickman-Brenner, F., Huntley-Carter, G., Saitoh, Y., Steigerwalt, A., Farmer, J., & Brenner, D. (1984). *Moellerella wisconsensis,* a new genus and species of *Enterobacteriaceae* found in human stool specimens. *J Clin Microbiol, 19*(4), 460-463.

Hickman-Brenner, F., Vohra, M., Huntley-Carter, G., Fanning, G., Lowery, V., Brenner, D., & Farmer, J. (1985). *Leminorella*, a new genus of *Enterobacteriaceae*: identification of *Leminorella grimontii* sp. nov. and *Leminorella richardii* sp. nov. found in clinical specimens. *J Clin Microbiol, 21*(2), 234-239.

Hollis, D., Hickman, F., Fanning, G., Farmer, J., Weaver, R., & Brenner, D. (1981). *Tatumella ptyseos* gen. nov., sp. nov., a member of the family *Enterobacteriaceae* found in clinical specimens. *J Clin Microbiol, 14*(1), 79-88.

Horn, M. (2011). Class I. **Chlamydiia** class. nov. In N. Krieg, J. Staley, D. Brown, B. Hedlund, B. Paster, N. Ward, W. Ludwig & W. Whitman (Eds.), Bergey's Manual of Systematic Bacteriology (2nd ed., Vol. 4, pp. 844). New York: Springer.

Husnik, F., Chrudimsky, T., & Hypsa, V. (2011). Multiple origins of endosymbiosis within the *Enterobacteriaceae* (gamma-Proteobacteria): convergence of complex phylogenetic approaches. *BMC Biol, 9*, 87.

Husník, F., Chrudimský, T., & Hypša, V. (2011). Multiple origins of endosymbiosis within the *Enterobacteriaceae* (γ-Proteobacteria): convergence of complex phylogenetic approaches. *BMC Biol, 9*(1), 1.

Huys, G., Cnockaert, M., Abbott, S. L., Janda, J. M., & Vandamme, P. (2010). *Hafnia paralvei* sp. nov., formerly known as *Hafnia alvei* hybridization group 2. *Int J Syst Evol Microbiol, 60*(8), 1725-1728.

Imhoff, J. F. (2005). Order XIII. "Enterobacteriales". In D. J. Brenner, N. R. Krieg, J. T. Staley, G. M. Garrity, D. R. Boone, P. Vos, M. Goodfellow, F. A. Rainey & K.-H. Schleifer (Eds.), Bergey's Manual of Systematic Bacteriology (2nd ed., Vol. 2, pp. 587). New York, NY: Springer US.

Iversen, C., Mullane, N., McCardell, B., Tall, B. D., Lehner, A., Fanning, S., Stephan, R., & Joosten, H. (2008). *Cronobacter* gen. nov., a new genus to accommodate the biogroups of *Enterobacter sakazakii*, and proposal of *Cronobacter sakazakii* gen. nov., comb. nov., *Cronobacter malonaticus* sp. nov., *Cronobacter turicensis* sp. nov., *Cronobacter muytjensii* sp. nov., *Cronobacter dublinensis* sp. nov., *Cronobacter* genomospecies 1, and of three subspecies, *Cronobacter dublinensis* subsp. *dublinensis* subsp. nov., *Cronobacter dublinensis* subsp. *lausannensis* subsp. nov. and *Cronobacter dublinensis* subsp. *lactaridi* subsp. nov. *Int J Syst Evol Microbiol, 58*(6), 1442-1447.

Izard, D., Gavini, F., Trinel, P., & Leclere, H. (1978). *Rahnella aquatilis, a new member of the Enterobacteriaceae.* Paper presented at the Ann Microbiol (Paris).

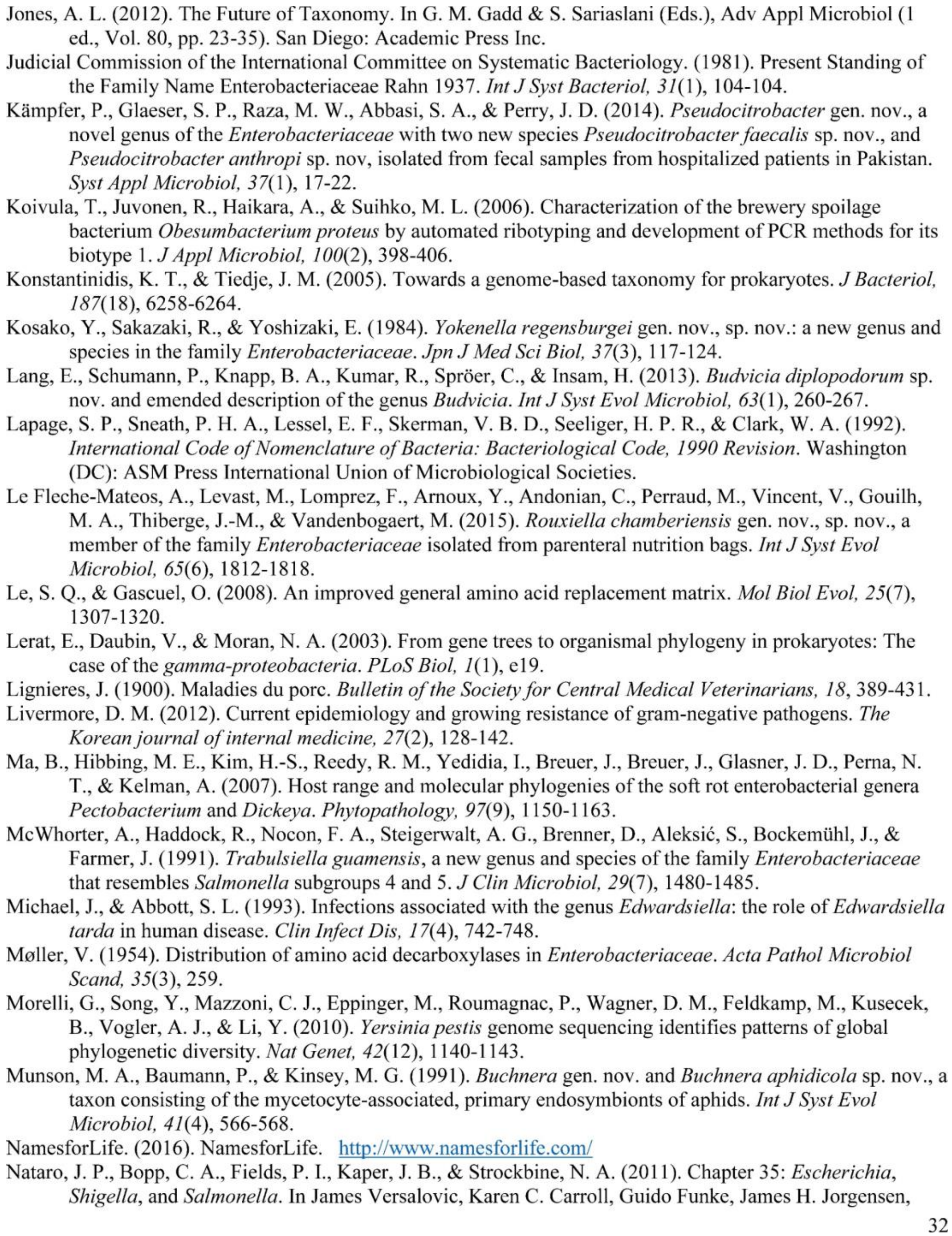
Janda, J. (2005). Genus XXVII. Plesiomonas. In D. J. Brenner, N. R. Krieg, G. M. Garrity & J. T. Staley (Eds.), Bergey's Manual of Systematic Bacteriology (2nd ed., Vol. 2, pp. 740-744). New York: Springer.

Janda, J. M. (2006). New members of the family *Enterobacteriaceae* The Prokaryotes (pp. 5-40): Springer.

Janda, J. M., & Abbott, S. L. (2006). The genus *Hafnia*: from soup to nuts. *Clin Microbiol Rev, 19*(1), 12-28.

Janda, J. M., Abbott, S. L., & McIver, C. J. (2016). *Plesiomonas shigelloides* Revisited. *Clin Microbiol Rev, 29*(2), 349-374.

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



Jones, A. L. (2012). The Future of Taxonomy. In G. M. Gadd & S. Sariaslani (Eds.), Adv Appl Microbiol (1 ed., Vol. 80, pp. 23-35). San Diego: Academic Press Inc.

Judicial Commission of the International Committee on Systematic Bacteriology. (1981). Present Standing of the Family Name Enterobacteriaceae Rahn 1937. *Int J Syst Bacteriol, 31*(1), 104-104.

Kämpfer, P., Glaeser, S. P., Raza, M. W., Abbasi, S. A., & Perry, J. D. (2014). *Pseudocitrobacter* gen. nov., a novel genus of the *Enterobacteriaceae* with two new species *Pseudocitrobacter faecalis* sp. nov., and *Pseudocitrobacter anthropi* sp. nov, isolated from fecal samples from hospitalized patients in Pakistan. *Syst Appl Microbiol, 37*(1), 17-22.

Koivula, T., Juvonen, R., Haikara, A., & Suihko, M. L. (2006). Characterization of the brewery spoilage bacterium *Obesumbacterium proteus* by automated ribotyping and development of PCR methods for its biotype 1. *J Appl Microbiol, 100*(2), 398-406.

Konstantinidis, K. T., & Tiedje, J. M. (2005). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol, 187*(18), 6258-6264.

Kosako, Y., Sakazaki, R., & Yoshizaki, E. (1984). *Yokenella regensburgei* gen. nov., sp. nov.: a new genus and species in the family *Enterobacteriaceae*. *Jpn J Med Sci Biol, 37*(3), 117-124.

Lang, E., Schumann, P., Knapp, B. A., Kumar, R., Spröer, C., & Insam, H. (2013). *Budvicia diplopodorum* sp. nov. and emended description of the genus *Budvicia*. *Int J Syst Evol Microbiol, 63*(1), 260-267.

Lapage, S. P., Sneath, P. H. A., Lessel, E. F., Skerman, V. B. D., Seeliger, H. P. R., & Clark, W. A. (1992). *International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision*. Washington (DC): ASM Press International Union of Microbiological Societies.

Le Fleche-Mateos, A., Levast, M., Lomprez, F., Arnoux, Y., Andonian, C., Perraud, M., Vincent, V., Gouilh, M. A., Thiberge, J.-M., & Vandenbogaert, M. (2015). *Rouxiella chamberiensis* gen. nov., sp. nov., a member of the family *Enterobacteriaceae* isolated from parenteral nutrition bags. *Int J Syst Evol Microbiol, 65*(6), 1812-1818.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol, 25*(7), 1307-1320.

Lerat, E., Daubin, V., & Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: The case of the *gamma-proteobacteria*. *PLoS Biol, 1*(1), e19.

Lignieres, J. (1900). Maladies du porc. *Bulletin of the Society for Central Medical Veterinarians, 18*, 389-431.

Livermore, D. M. (2012). Current epidemiology and growing resistance of gram-negative pathogens. *The Korean journal of internal medicine, 27*(2), 128-142.

Ma, B., Hibbing, M. E., Kim, H.-S., Reedy, R. M., Yedidia, I., Breuer, J., Breuer, J., Glasner, J. D., Perna, N. T., & Kelman, A. (2007). Host range and molecular phylogenies of the soft rot enterobacterial genera *Pectobacterium* and *Dickeya*. *Phytopathology, 97*(9), 1150-1163.

McWhorter, A., Haddock, R., Nocon, F. A., Steigerwalt, A. G., Brenner, D., Aleksić, S., Bockemühl, J., & Farmer, J. (1991). *Trabulsiella guamensis*, a new genus and species of the family *Enterobacteriaceae* that resembles *Salmonella* subgroups 4 and 5. *J Clin Microbiol, 29*(7), 1480-1485.

Michael, J., & Abbott, S. L. (1993). Infections associated with the genus *Edwardsiella*: the role of *Edwardsiella tarda* in human disease. *Clin Infect Dis, 17*(4), 742-748.

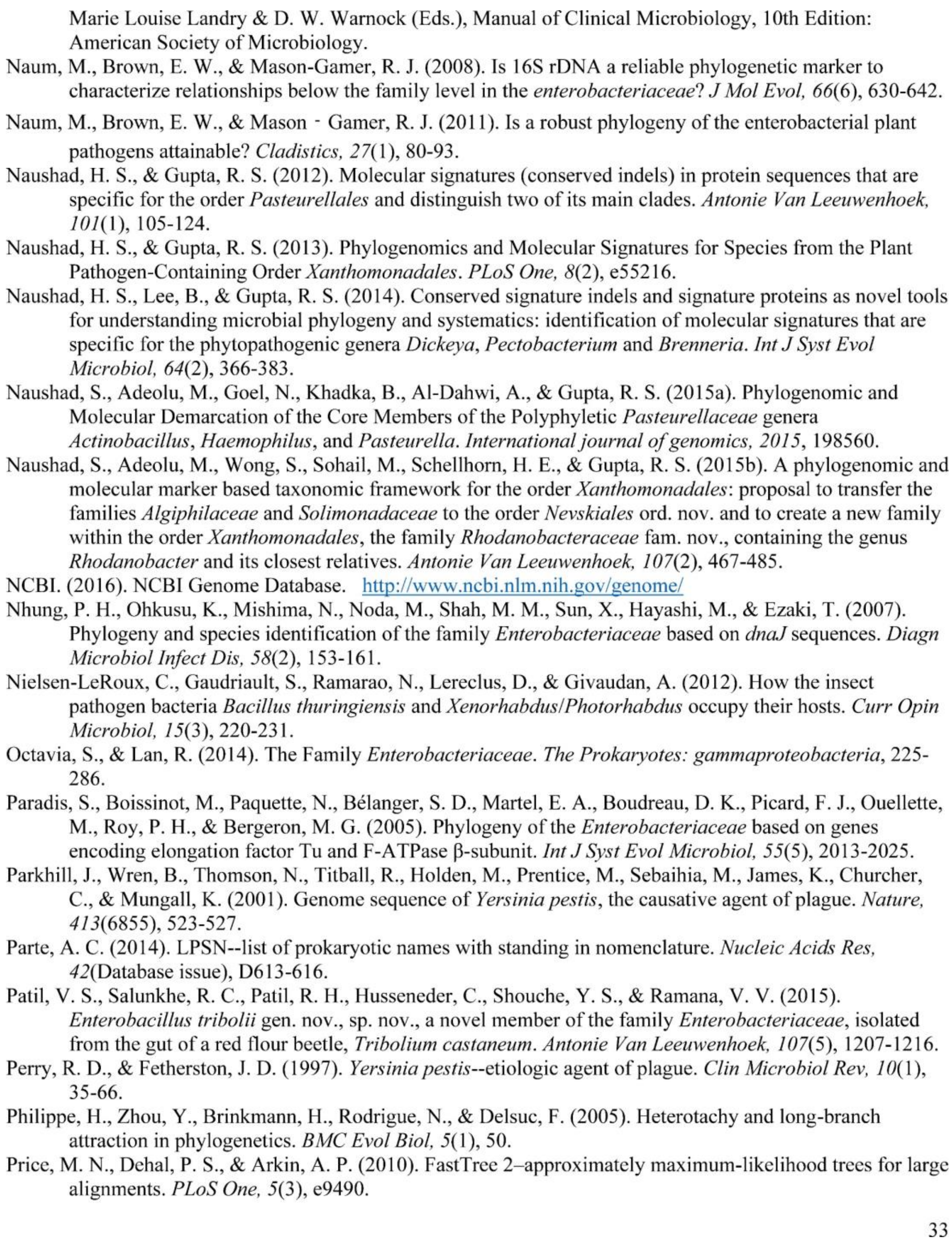
Møller, V. (1954). Distribution of amino acid decarboxylases in *Enterobacteriaceae*. *Acta Pathol Microbiol Scand, 35*(3), 259.

Morelli, G., Song, Y., Mazzoni, C. J., Eppinger, M., Roumagnac, P., Wagner, D. M., Feldkamp, M., Kusecek, B., Vogler, A. J., & Li, Y. (2010). *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet, 42*(12), 1140-1143.

Munson, M. A., Baumann, P., & Kinsey, M. G. (1991). *Buchnera* gen. nov. and *Buchnera aphidicola* sp. nov., a taxon consisting of the mycetocyte-associated, primary endosymbionts of aphids. *Int J Syst Evol Microbiol, 41*(4), 566-568.

NamesforLife. (2016). NamesforLife. http://www.namesforlife.com/

Nataro, J. P., Bopp, C. A., Fields, P. I., Kaper, J. B., & Strockbine, N. A. (2011). Chapter 35: *Escherichia*, *Shigella*, and *Salmonella*. In James Versalovic, Karen C. Carroll, Guido Funke, James H. Jorgensen,



Marie Louise Landry & D. W. Warnock (Eds.), Manual of Clinical Microbiology, 10th Edition: American Society of Microbiology.

Naum, M., Brown, E. W., & Mason-Gamer, R. J. (2008). Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the *enterobacteriaceae*? *J Mol Evol, 66*(6), 630-642.

Naum, M., Brown, E. W., & Mason - Gamer, R. J. (2011). Is a robust phylogeny of the enterobacterial plant pathogens attainable? *Cladistics, 27*(1), 80-93.

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

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

Naushad, H. S., Lee, B., & Gupta, R. S. (2014). Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*. *Int J Syst Evol Microbiol, 64*(2), 366-383.

Naushad, S., Adeolu, M., Goel, N., Khadka, B., Al-Dahwi, A., & Gupta, R. S. (2015a). Phylogenomic and Molecular Demarcation of the Core Members of the Polyphyletic *Pasteurellaceae* genera *Actinobacillus*, *Haemophilus*, and *Pasteurella*. *International journal of genomics, 2015*, 198560.

Naushad, S., Adeolu, M., Wong, S., Sohail, M., Schellhorn, H. E., & Gupta, R. S. (2015b). A phylogenomic and molecular marker based taxonomic framework for the order *Xanthomonadales*: proposal to transfer the families *Algiphilaceae* and *Solimonadaceae* to the order *Nevskiales* ord. nov. and to create a new family within the order *Xanthomonadales*, the family *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. *Antonie Van Leeuwenhoek, 107*(2), 467-485.

NCBI. (2016). NCBI Genome Database. http://www.ncbi.nlm.nih.gov/genome/

Nhung, P. H., Ohkusu, K., Mishima, N., Noda, M., Shah, M. M., Sun, X., Hayashi, M., & Ezaki, T. (2007). Phylogeny and species identification of the family *Enterobacteriaceae* based on *dnaJ* sequences. *Diagn Microbiol Infect Dis, 58*(2), 153-161.

Nielsen-LeRoux, C., Gaudriault, S., Ramarao, N., Lereclus, D., & Givaudan, A. (2012). How the insect pathogen bacteria *Bacillus thuringiensis* and *Xenorhabdus/Photorhabdus* occupy their hosts. *Curr Opin Microbiol, 15*(3), 220-231.

Octavia, S., & Lan, R. (2014). The Family *Enterobacteriaceae*. *The Prokaryotes: gammaproteobacteria*, 225-286.

Paradis, S., Boissinot, M., Paquette, N., Bélanger, S. D., Martel, E. A., Boudreau, D. K., Picard, F. J., Ouellette, M., Roy, P. H., & Bergeron, M. G. (2005). Phylogeny of the *Enterobacteriaceae* based on genes encoding elongation factor Tu and F-ATPase β-subunit. *Int J Syst Evol Microbiol, 55*(5), 2013-2025.

Parkhill, J., Wren, B., Thomson, N., Titball, R., Holden, M., Prentice, M., Sebaihia, M., James, K., Churcher, C., & Mungall, K. (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature, 413*(6855), 523-527.

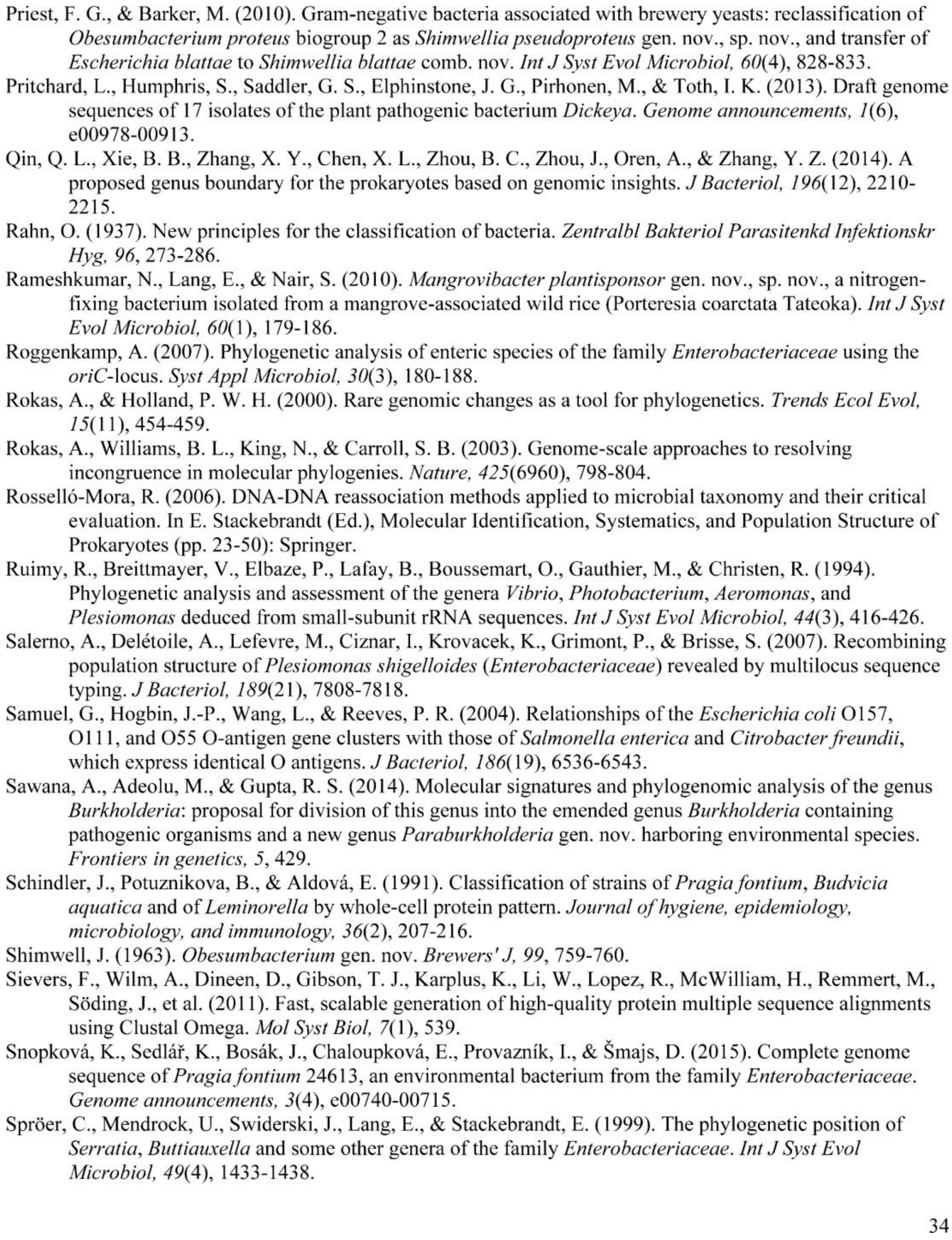
Parte, A. C. (2014). LPSN--list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res, 42*(Database issue), D613-616.

Patil, V. S., Salunkhe, R. C., Patil, R. H., Husseneder, C., Shouche, Y. S., & Ramana, V. V. (2015). *Enterobacillus tribolii* gen. nov., sp. nov., a novel member of the family *Enterobacteriaceae*, isolated from the gut of a red flour beetle, *Tribolium castaneum*. *Antonie Van Leeuwenhoek, 107*(5), 1207-1216.

Perry, R. D., & Fetherston, J. D. (1997). *Yersinia pestis*--etiologic agent of plague. *Clin Microbiol Rev, 10*(1), 35-66.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., & Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol, 5*(1), 50.

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



Priest, F. G., & Barker, M. (2010). Gram-negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. *Int J Syst Evol Microbiol, 60*(4), 828-833.

Pritchard, L., Humphris, S., Saddler, G. S., Elphinstone, J. G., Pirhonen, M., & Toth, I. K. (2013). Draft genome sequences of 17 isolates of the plant pathogenic bacterium *Dickeya*. *Genome announcements, 1*(6), e00978-00913.

Qin, Q. L., Xie, B. B., Zhang, X. Y., Chen, X. L., Zhou, B. C., Zhou, J., Oren, A., & Zhang, Y. Z. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol, 196*(12), 2210-2215.

Rahn, O. (1937). New principles for the classification of bacteria. *Zentralbl Bakteriol Parasitenkd Infektionskr Hyg, 96*, 273-286.

Rameshkumar, N., Lang, E., & Nair, S. (2010). *Mangrovibacter plantisponsor* gen. nov., sp. nov., a nitrogen-fixing bacterium isolated from a mangrove-associated wild rice (Porteresia coarctata Tateoka). *Int J Syst Evol Microbiol, 60*(1), 179-186.

Roggenkamp, A. (2007). Phylogenetic analysis of enteric species of the family *Enterobacteriaceae* using the *oriC*-locus. *Syst Appl Microbiol, 30*(3), 180-188.

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

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

Rosselló-Mora, R. (2006). DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In E. Stackebrandt (Ed.), Molecular Identification, Systematics, and Population Structure of Prokaryotes (pp. 23-50): Springer.

Ruimy, R., Breittmayer, V., Elbaze, P., Lafay, B., Boussemart, O., Gauthier, M., & Christen, R. (1994). Phylogenetic analysis and assessment of the genera *Vibrio*, *Photobacterium*, *Aeromonas*, and *Plesiomonas* deduced from small-subunit rRNA sequences. *Int J Syst Evol Microbiol, 44*(3), 416-426.

Salerno, A., Delétoile, A., Lefevre, M., Ciznar, I., Krovacek, K., Grimont, P., & Brisse, S. (2007). Recombining population structure of *Plesiomonas shigelloides* (*Enterobacteriaceae*) revealed by multilocus sequence typing. *J Bacteriol, 189*(21), 7808-7818.

Samuel, G., Hogbin, J.-P., Wang, L., & Reeves, P. R. (2004). Relationships of the *Escherichia coli* O157, O111, and O55 O-antigen gene clusters with those of *Salmonella enterica* and *Citrobacter freundii*, which express identical O antigens. *J Bacteriol, 186*(19), 6536-6543.

Sawana, A., Adeolu, M., & Gupta, R. S. (2014). Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring environmental species. *Frontiers in genetics, 5*, 429.

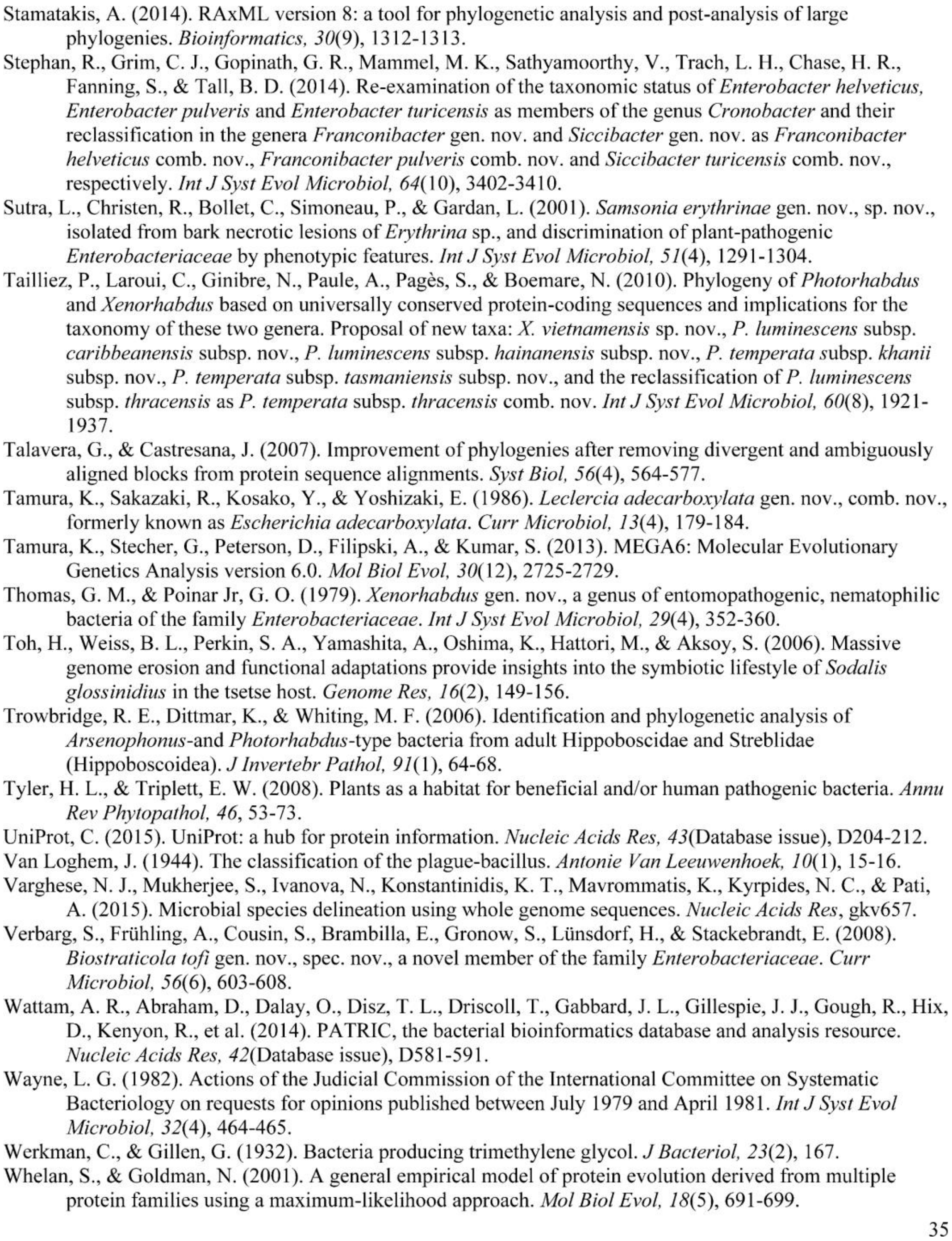
Schindler, J., Potuznikova, B., & Aldová, E. (1991). Classification of strains of *Pragia fontium*, *Budvicia aquatica* and of *Leminorella* by whole-cell protein pattern. *Journal of hygiene, epidemiology, microbiology, and immunology, 36*(2), 207-216.

Shimwell, J. (1963). *Obesumbacterium* gen. nov. *Brewers' J, 99*, 759-760.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol, 7*(1), 539.

Snopková, K., Sedlář, K., Bosák, J., Chaloupková, E., Provazník, I., & Šmajs, D. (2015). Complete genome sequence of *Pragia fontium* 24613, an environmental bacterium from the family *Enterobacteriaceae*. *Genome announcements, 3*(4), e00740-00715.

Spröer, C., Mendrock, U., Swiderski, J., Lang, E., & Stackebrandt, E. (1999). The phylogenetic position of *Serratia*, *Buttiauxella* and some other genera of the family *Enterobacteriaceae*. *Int J Syst Evol Microbiol, 49*(4), 1433-1438.



Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313.

Stephan, R., Grim, C. J., Gopinath, G. R., Mammel, M. K., Sathyamoorthy, V., Trach, L. H., Chase, H. R., Fanning, S., & Tall, B. D. (2014). Re-examination of the taxonomic status of *Enterobacter helveticus, Enterobacter pulveris* and *Enterobacter turicensis* as members of the genus *Cronobacter* and their reclassification in the genera *Franconibacter* gen. nov. and *Siccibacter* gen. nov. as *Franconibacter helveticus* comb. nov., *Franconibacter pulveris* comb. nov. and *Siccibacter turicensis* comb. nov., respectively. *Int J Syst Evol Microbiol, 64*(10), 3402-3410.

Sutra, L., Christen, R., Bollet, C., Simoneau, P., & Gardan, L. (2001). *Samsonia erythrinae* gen. nov., sp. nov., isolated from bark necrotic lesions of *Erythrina* sp., and discrimination of plant-pathogenic *Enterobacteriaceae* by phenotypic features. *Int J Syst Evol Microbiol, 51*(4), 1291-1304.

Tailliez, P., Laroui, C., Ginibre, N., Paule, A., Pagès, S., & Boemare, N. (2010). Phylogeny of *Photorhabdus* and *Xenorhabdus* based on universally conserved protein-coding sequences and implications for the taxonomy of these two genera. Proposal of new taxa: *X. vietnamensis* sp. nov., *P. luminescens* subsp. *caribbeanensis* subsp. nov., *P. luminescens* subsp. *hainanensis* subsp. nov., *P. temperata s*ubsp. *khanii* subsp. nov., *P. temperata* subsp. *tasmaniensis* subsp. nov., and the reclassification of *P. luminescens* subsp. *thracensis* as *P. temperata* subsp. *thracensis* comb. nov. *Int J Syst Evol Microbiol, 60*(8), 1921-1937.

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol, 56*(4), 564-577.

Tamura, K., Sakazaki, R., Kosako, Y., & Yoshizaki, E. (1986). *Leclercia adecarboxylata* gen. nov., comb. nov., formerly known as *Escherichia adecarboxylata*. *Curr Microbiol, 13*(4), 179-184.

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

Thomas, G. M., & Poinar Jr, G. O. (1979). *Xenorhabdus* gen. nov., a genus of entomopathogenic, nematophilic bacteria of the family *Enterobacteriaceae*. *Int J Syst Evol Microbiol, 29*(4), 352-360.

Toh, H., Weiss, B. L., Perkin, S. A., Yamashita, A., Oshima, K., Hattori, M., & Aksoy, S. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res, 16*(2), 149-156.

Trowbridge, R. E., Dittmar, K., & Whiting, M. F. (2006). Identification and phylogenetic analysis of *Arsenophonus*-and *Photorhabdus*-type bacteria from adult Hippoboscidae and Streblidae (Hippoboscoidea). *J Invertebr Pathol, 91*(1), 64-68.

Tyler, H. L., & Triplett, E. W. (2008). Plants as a habitat for beneficial and/or human pathogenic bacteria. *Annu Rev Phytopathol, 46*, 53-73.

UniProt, C. (2015). UniProt: a hub for protein information. *Nucleic Acids Res, 43*(Database issue), D204-212.

Van Loghem, J. (1944). The classification of the plague-bacillus. *Antonie Van Leeuwenhoek, 10*(1), 15-16.

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res*, gkv657.

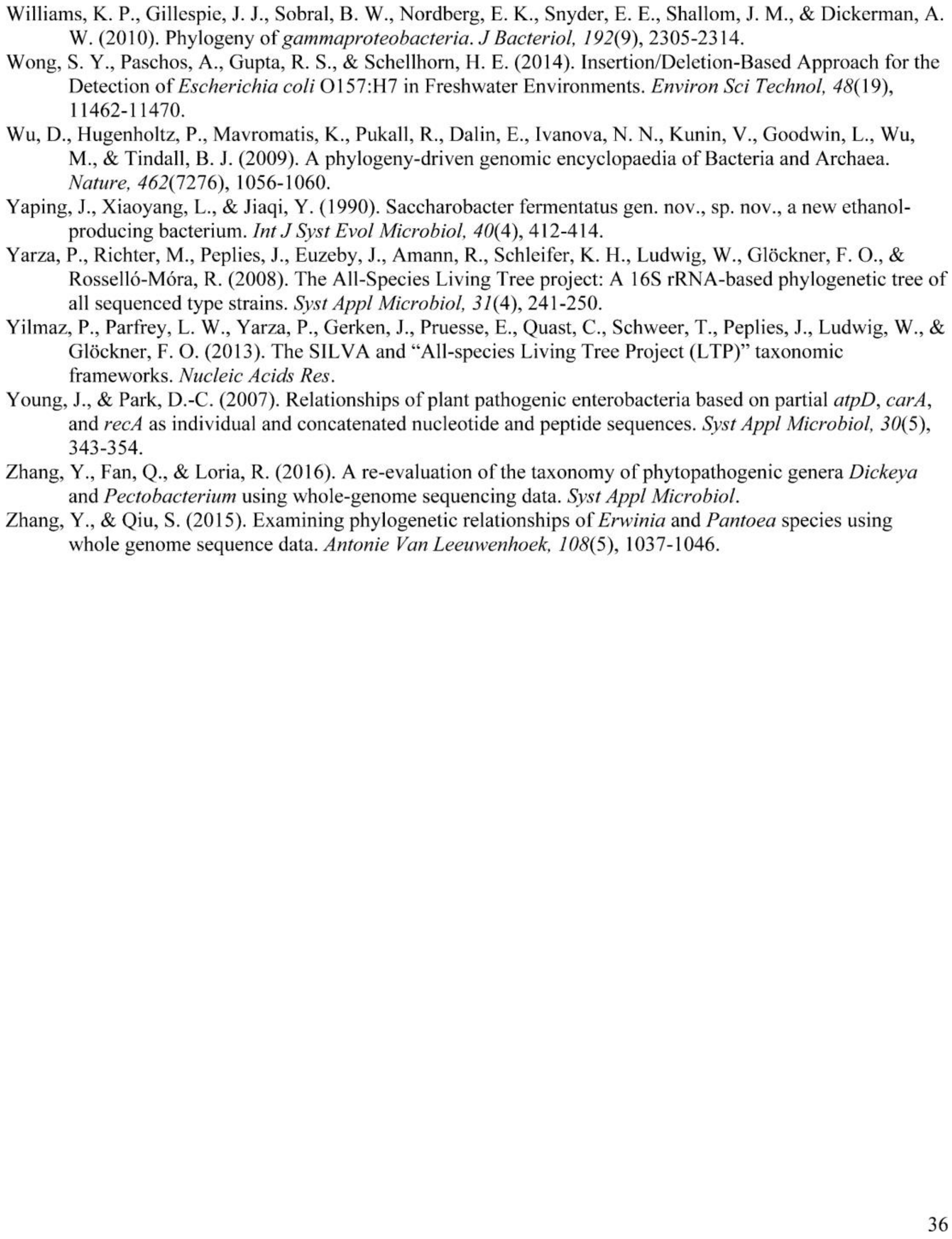
Verbarg, S., Frühling, A., Cousin, S., Brambilla, E., Gronow, S., Lünsdorf, H., & Stackebrandt, E. (2008). *Biostraticola tofi* gen. nov., spec. nov., a novel member of the family *Enterobacteriaceae*. *Curr Microbiol, 56*(6), 603-608.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res, 42*(Database issue), D581-591.

Wayne, L. G. (1982). Actions of the Judicial Commission of the International Committee on Systematic Bacteriology on requests for opinions published between July 1979 and April 1981. *Int J Syst Evol Microbiol, 32*(4), 464-465.

Werkman, C., & Gillen, G. (1932). Bacteria producing trimethylene glycol. *J Bacteriol, 23*(2), 167.

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



Williams, K. P., Gillespie, J. J., Sobral, B. W., Nordberg, E. K., Snyder, E. E., Shallom, J. M., & Dickerman, A. W. (2010). Phylogeny of *gammaproteobacteria*. *J Bacteriol, 192*(9), 2305-2314.

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

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

Yaping, J., Xiaoyang, L., & Jiaqi, Y. (1990). Saccharobacter fermentatus gen. nov., sp. nov., a new ethanol-producing bacterium. *Int J Syst Evol Microbiol, 40*(4), 412-414.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K. H., Ludwig, W., Glöckner, F. O., & Rosselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol, 31*(4), 241-250.

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

Young, J., & Park, D.-C. (2007). Relationships of plant pathogenic enterobacteria based on partial *atpD*, *carA*, and *recA* as individual and concatenated nucleotide and peptide sequences. *Syst Appl Microbiol, 30*(5), 343-354.

Zhang, Y., Fan, Q., & Loria, R. (2016). A re-evaluation of the taxonomy of phytopathogenic genera *Dickeya* and *Pectobacterium* using whole-genome sequencing data. *Syst Appl Microbiol*.

Zhang, Y., & Qiu, S. (2015). Examining phylogenetic relationships of *Erwinia* and *Pantoea* species using whole genome sequence data. *Antonie Van Leeuwenhoek, 108*(5), 1037-1046.

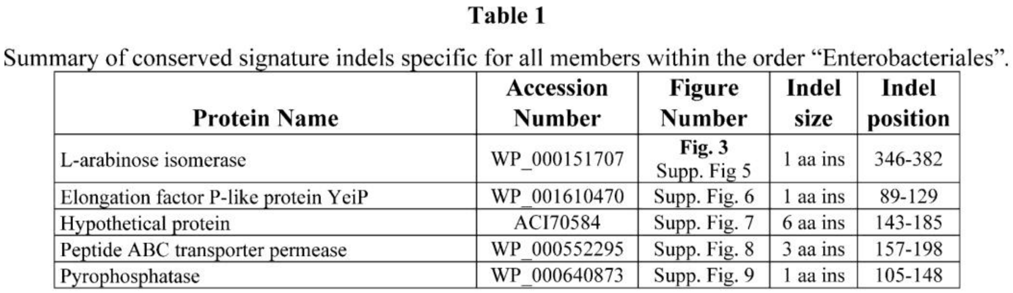


Table 1

Summary of conserved signature indels specific for all members within the order “Enterobacteriales”.

Protein Name	Accession Number	Figure Number	Indel size	Indel position
L-arabinose isomerase	WP_000151707	**Fig. 3** Supp. Fig 5	1 aa ins	346-382
Elongation factor P-like protein YeiP	WP_001610470	Supp. Fig. 6	1 aa ins	89-129
Hypothetical protein	ACI70584	Supp. Fig. 7	6 aa ins	143-185
Peptide ABC transporter permease	WP_000552295	Supp. Fig. 8	3 aa ins	157-198
Pyrophosphatase	WP_000640873	Supp. Fig. 9	1 aa ins	105-148

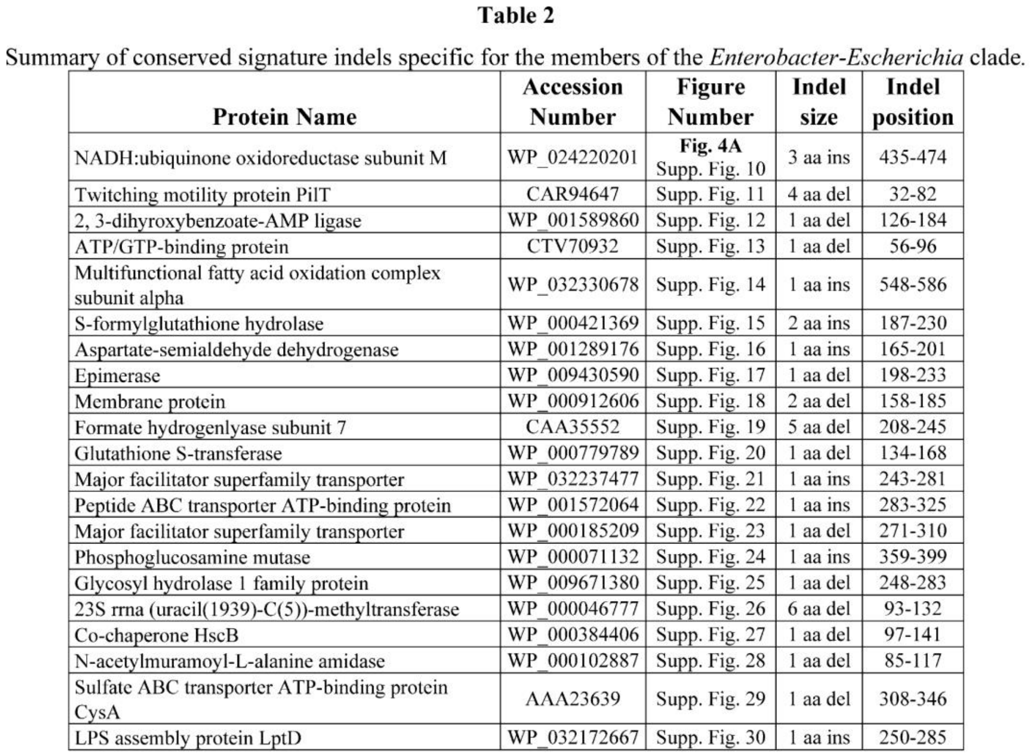


Table 2

Summary of conserved signature indels specific for the members of the *Enterobacter-Escherichia* clade.

Protein Name	Accession Number	Figure Number	Indel size	Indel position
NADH:ubiquinone oxidoreductase subunit M	WP_024220201	**Fig. 4A** Supp. Fig. 10	3 aa ins	435-474
Twitching motility protein PilT	CAR94647	Supp. Fig. 11	4 aa del	32-82
2, 3-dihyroxybenzoate-AMP ligase	WP_001589860	Supp. Fig. 12	1 aa del	126-184
ATP/GTP-binding protein	CTV70932	Supp. Fig. 13	1 aa del	56-96
Multifunctional fatty acid oxidation complex subunit alpha	WP_032330678	Supp. Fig. 14	1 aa ins	548-586
S-formylglutathione hydrolase	WP_000421369	Supp. Fig. 15	2 aa ins	187-230
Aspartate-semialdehyde dehydrogenase	WP_001289176	Supp. Fig. 16	1 aa ins	165-201
Epimerase	WP_009430590	Supp. Fig. 17	1 aa del	198-233
Membrane protein	WP_000912606	Supp. Fig. 18	2 aa del	158-185
Formate hydrogenlyase subunit 7	CAA35552	Supp. Fig. 19	5 aa del	208-245
Glutathione S-transferase	WP_000779789	Supp. Fig. 20	1 aa del	134-168
Major facilitator superfamily transporter	WP_032237477	Supp. Fig. 21	1 aa ins	243-281
Peptide ABC transporter ATP-binding protein	WP_001572064	Supp. Fig. 22	1 aa ins	283-325
Major facilitator superfamily transporter	WP_000185209	Supp. Fig. 23	1 aa del	271-310
Phosphoglucosamine mutase	WP_000071132	Supp. Fig. 24	1 aa ins	359-399
Glycosyl hydrolase 1 family protein	WP_009671380	Supp. Fig. 25	1 aa del	248-283
23S rrna (uracil(1939)-C(5))-methyltransferase	WP_000046777	Supp. Fig. 26	6 aa del	93-132
Co-chaperone HscB	WP_000384406	Supp. Fig. 27	1 aa del	97-141
N-acetylmuramoyl-L-alanine amidase	WP_000102887	Supp. Fig. 28	1 aa del	85-117
Sulfate ABC transporter ATP-binding protein CysA	AAA23639	Supp. Fig. 29	1 aa del	308-346
LPS assembly protein LptD	WP_032172667	Supp. Fig. 30	1 aa ins	250-285

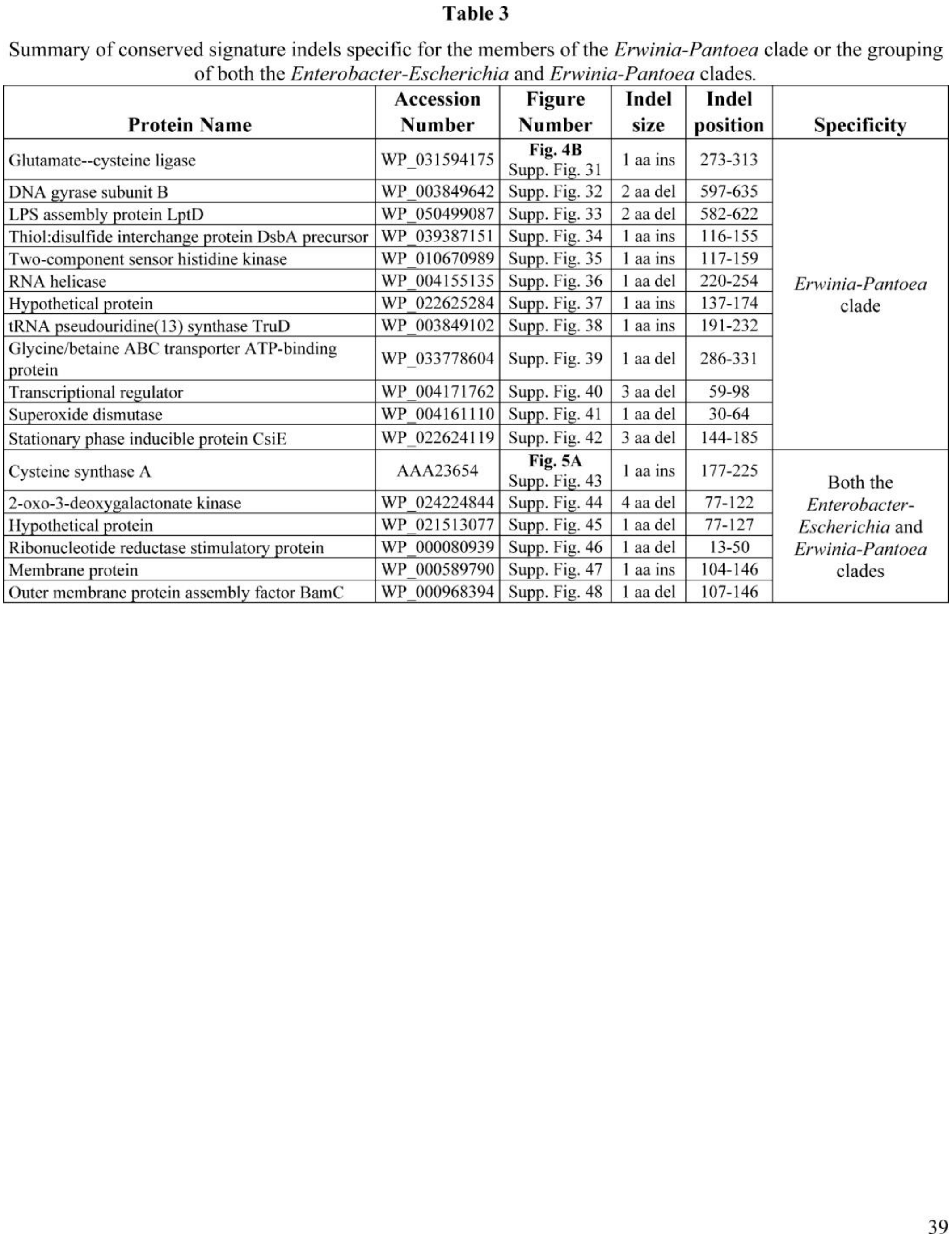


Table 3

Summary of conserved signature indels specific for the members of the *Erwinia-Pantoea* clade or the grouping of both the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades.

Protein Name	Accession Number	Figure Number	Indel size	Indel position	Specificity
Glutamate--cysteine ligase	WP_031594175	**Fig. 4B** Supp. Fig. 31	1 aa ins	273-313	*Erwinia-Pantoea* clade
DNA gyrase subunit B	WP_003849642	Supp. Fig. 32	2 aa del	597-635	
LPS assembly protein LptD	WP_050499087	Supp. Fig. 33	2 aa del	582-622	
Thiol:disulfide interchange protein DsbA precursor	WP_039387151	Supp. Fig. 34	1 aa ins	116-155	
Two-component sensor histidine kinase	WP_010670989	Supp. Fig. 35	1 aa ins	117-159	
RNA helicase	WP_004155135	Supp. Fig. 36	1 aa del	220-254	
Hypothetical protein	WP_022625284	Supp. Fig. 37	1 aa ins	137-174	
tRNA pseudouridine(13) synthase TruD	WP_003849102	Supp. Fig. 38	1 aa ins	191-232	
Glycine/betaine ABC transporter ATP-binding protein	WP_033778604	Supp. Fig. 39	1 aa del	286-331	
Transcriptional regulator	WP_004171762	Supp. Fig. 40	3 aa del	59-98	
Superoxide dismutase	WP_004161110	Supp. Fig. 41	1 aa del	30-64	
Stationary phase inducible protein CsiE	WP_022624119	Supp. Fig. 42	3 aa del	144-185	
Cysteine synthase A	AAA23654	**Fig. 5A** Supp. Fig. 43	1 aa ins	177-225	Both the *Enterobacter-Escherichia* and *Erwinia-Pantoea* clades
2-oxo-3-deoxygalactonate kinase	WP_024224844	Supp. Fig. 44	4 aa del	77-122	
Hypothetical protein	WP_021513077	Supp. Fig. 45	1 aa del	77-127	
Ribonucleotide reductase stimulatory protein	WP_000080939	Supp. Fig. 46	1 aa del	13-50	
Membrane protein	WP_000589790	Supp. Fig. 47	1 aa ins	104-146	
Outer membrane protein assembly factor BamC	WP_000968394	Supp. Fig. 48	1 aa del	107-146	

39

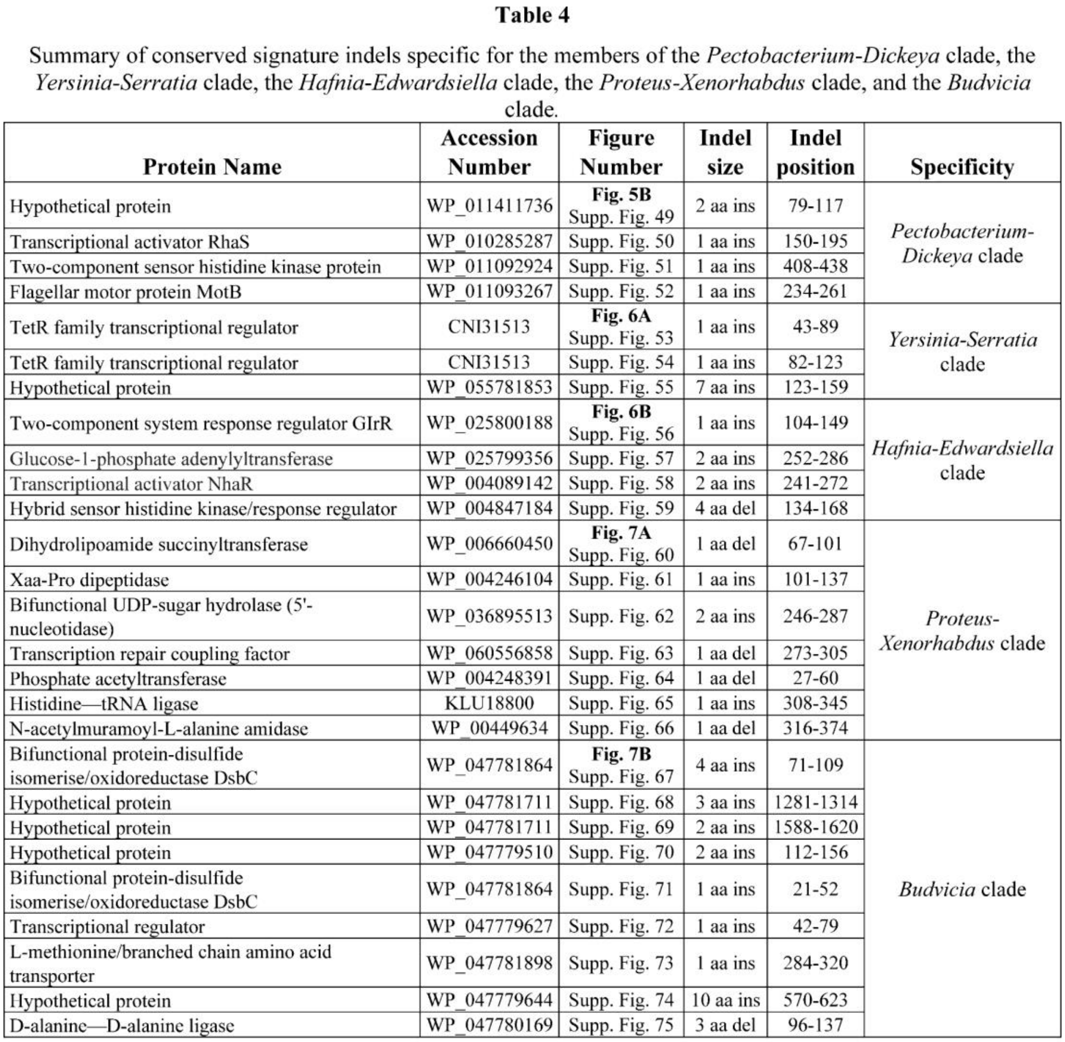


Table 4

Summary of conserved signature indels specific for the members of the *Pectobacterium-Dickeya* clade, the *Yersinia-Serratia* clade, the *Hafnia-Edwardsiella* clade, the *Proteus-Xenorhabdus* clade, and the *Budvicia* clade.

Protein Name	Accession Number	Figure Number	Indel size	Indel position	Specificity
Hypothetical protein	WP_011411736	**Fig. 5B** Supp. Fig. 49	2 aa ins	79-117	*Pectobacterium-Dickeya* clade
Transcriptional activator RhaS	WP_010285287	Supp. Fig. 50	1 aa ins	150-195	
Two-component sensor histidine kinase protein	WP_011092924	Supp. Fig. 51	1 aa ins	408-438	
Flagellar motor protein MotB	WP_011093267	Supp. Fig. 52	1 aa ins	234-261	
TetR family transcriptional regulator	CNI31513	**Fig. 6A** Supp. Fig. 53	1 aa ins	43-89	*Yersinia-Serratia* clade
TetR family transcriptional regulator	CNI31513	Supp. Fig. 54	1 aa ins	82-123	
Hypothetical protein	WP_055781853	Supp. Fig. 55	7 aa ins	123-159	
Two-component system response regulator GIrR	WP_025800188	**Fig. 6B** Supp. Fig. 56	1 aa ins	104-149	*Hafnia-Edwardsiella* clade
Glucose-1-phosphate adenylyltransferase	WP_025799356	Supp. Fig. 57	2 aa ins	252-286	
Transcriptional activator NhaR	WP_004089142	Supp. Fig. 58	2 aa ins	241-272	
Hybrid sensor histidine kinase/response regulator	WP_004847184	Supp. Fig. 59	4 aa del	134-168	
Dihydrolipoamide succinyltransferase	WP_006660450	**Fig. 7A** Supp. Fig. 60	1 aa del	67-101	*Proteus-Xenorhabdus* clade
Xaa-Pro dipeptidase	WP_004246104	Supp. Fig. 61	1 aa ins	101-137	
Bifunctional UDP-sugar hydrolase (5'-nucleotidase)	WP_036895513	Supp. Fig. 62	2 aa ins	246-287	
Transcription repair coupling factor	WP_060556858	Supp. Fig. 63	1 aa del	273-305	
Phosphate acetyltransferase	WP_004248391	Supp. Fig. 64	1 aa del	27-60	
Histidine—tRNA ligase	KLU18800	Supp. Fig. 65	1 aa ins	308-345	
N-acetylmuramoyl-L-alanine amidase	WP_00449634	Supp. Fig. 66	1 aa del	316-374	
Bifunctional protein-disulfide isomerise/oxidoreductase DsbC	WP_047781864	**Fig. 7B** Supp. Fig. 67	4 aa ins	71-109	*Budvicia* clade
Hypothetical protein	WP_047781711	Supp. Fig. 68	3 aa ins	1281-1314	
Hypothetical protein	WP_047781711	Supp. Fig. 69	2 aa ins	1588-1620	
Hypothetical protein	WP_047779510	Supp. Fig. 70	2 aa ins	112-156	
Bifunctional protein-disulfide isomerise/oxidoreductase DsbC	WP_047781864	Supp. Fig. 71	1 aa ins	21-52	
Transcriptional regulator	WP_047779627	Supp. Fig. 72	1 aa ins	42-79	
L-methionine/branched chain amino acid transporter	WP_047781898	Supp. Fig. 73	1 aa ins	284-320	
Hypothetical protein	WP_047779644	Supp. Fig. 74	10 aa ins	570-623	
D-alanine—D-alanine ligase	WP_047780169	Supp. Fig. 75	3 aa del	96-137	

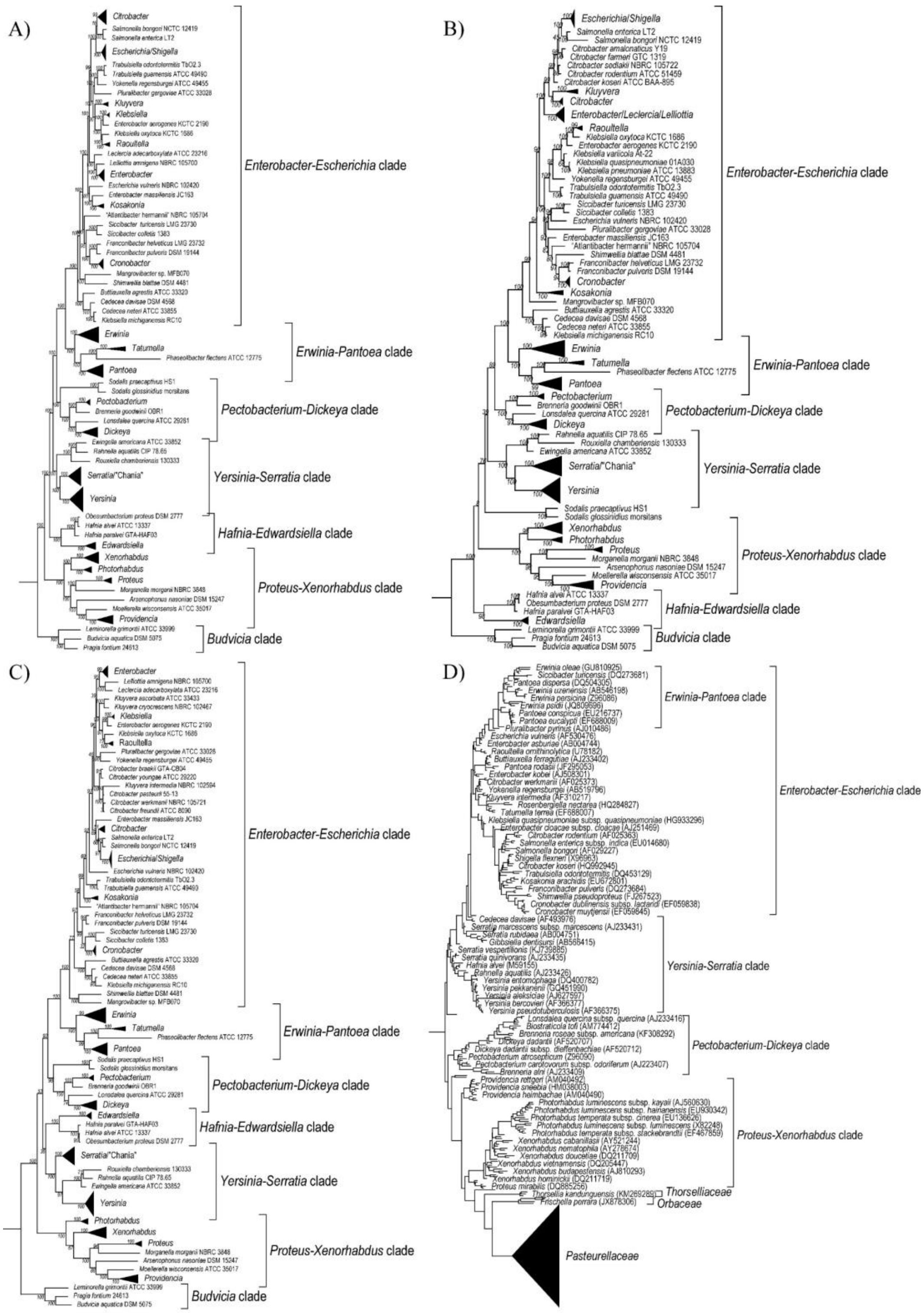

A)
Citrobacter
Salmonella bongori NCTC 12419
Salmonella enterica LT2
Escherichia/Shigella
Trabulsiella odontotermitis TbO2.3
Trabulsiella guamensis ATCC 49490
Yokenella regensburgei ATCC 49455
Pluralibacter gergoviae ATCC 33028
Kluyvera
Klebsiella
Enterobacter aerogenes KCTC 2190
Klebsiella oxytoca KCTC 1686
Raoultella
Leclercia adecarboxylata ATCC 23216
Lelliottia amnigena NBRC 105700
Enterobacter
Escherichia vulneris NBRC 102420
Enterobacter massiliensis JC163
Kosakonia
"Atlantibacter hermannii" NBRC 105704
Siccibacter turicensis LMG 23730
Siccibacter colletis 1383
Franconibacter helveticus LMG 23732
Franconibacter pulveris DSM 19144
Cronobacter
Mangrovibacter sp. MFB070
Shimwellia blattae DSM 4481
Buttiauxella agrestis ATCC 33320
Cedecea davisae DSM 4568
Cedecea neteri ATCC 33855
Klebsiella michiganensis RC10
Enterobacter-Escherichia clade
Erwinia
Tatumella
Phaseolibacter flectens ATCC 12775
Pantoea
Erwinia-Pantoea clade
Sodalis praecaptivus HS1
Sodalis glossinidius morsitans
Pectobacterium
Brenneria goodwinii OBR1
Lonsdalea quercina ATCC 29281
Dickeya
Pectobacterium-Dickeya clade
Ewingella americana ATCC 33852
Rahnella aquatilis CIP 78.65
Rouxiella chamberiensis 130333
Serratia/"Chania"
Yersinia
Yersinia-Serratia clade
Obesumbacterium proteus DSM 2777
Hafnia alvei ATCC 13337
Hafnia paralvei GTA-HAF03
Edwardsiella
Hafnia-Edwardsiella clade
Xenorhabdus
Photorhabdus
Proteus
Morganella morganii NBRC 3848
Arsenophonus nasoniae DSM 15247
Moellerella wisconsensis ATCC 35017
Providencia
Proteus-Xenorhabdus clade
Leminorella grimontii ATCC 33999
Budvicia aquatica DSM 5075
Pragia fontium 24613
Budvicia clade
B)
Escherichia/Shigella
Salmonella enterica LT2
Salmonella bongori NCTC 12419
Citrobacter amalonaticus Y19
Citrobacter farmeri GTC 1319
Citrobacter sedlakii NBRC 105722
Citrobacter rodentium ATCC 51459
Citrobacter koseri ATCC BAA-895
Kluyvera
Citrobacter
Enterobacter/Leclercia/Lelliottia
Raoultella
Klebsiella oxytoca KCTC 1686
Enterobacter aerogenes KCTC 2190
Klebsiella variicola At-22
Klebsiella quasipneumoniae 01A030
Klebsiella pneumoniae ATCC 13883
Yokenella regensburgei ATCC 49455
Trabulsiella odontotermitis TbO2.3
Trabulsiella guamensis ATCC 49490
Siccibacter turicensis LMG 23730
Siccibacter colletis 1383
Escherichia vulneris NBRC 102420
Pluralibacter gergoviae ATCC 33028
Enterobacter massiliensis JC163
"Atlantibacter hermannii" NBRC 105704
Shimwellia blattae DSM 4481
Franconibacter helveticus LMG 23732
Franconibacter pulveris DSM 19144
Cronobacter
Kosakonia
Mangrovibacter sp. MFB070
Buttiauxella agrestis ATCC 33320
Cedecea davisae DSM 4568
Cedecea neteri ATCC 33855
Klebsiella michiganensis RC10
Enterobacter-Escherichia clade
Erwinia
Tatumella
Phaseolibacter flectens ATCC 12775
Pantoea
Erwinia-Pantoea clade
Pectobacterium
Brenneria goodwinii OBR1
Lonsdalea quercina ATCC 29281
Dickeya
Pectobacterium-Dickeya clade
Rahnella aquatilis CIP 78.65
Rouxiella chamberiensis 130333
Ewingella americana ATCC 33852
Serratia/"Chania"
Yersinia
Yersinia-Serratia clade
Sodalis praecaptivus HS1
Sodalis glossinidius morsitans
Xenorhabdus
Photorhabdus
Proteus
Morganella morganii NBRC 3848
Arsenophonus nasoniae DSM 15247
Moellerella wisconsensis ATCC 35017
Providencia
Proteus-Xenorhabdus clade
Hafnia alvei ATCC 13337
Obesumbacterium proteus DSM 2777
Hafnia paralvei GTA-HAF03
Edwardsiella
Hafnia-Edwardsiella clade
Leminorella grimontii ATCC 33999
Pragia fontium 24613
Budvicia aquatica DSM 5075
Budvicia clade
C)
Enterobacter
Lelliottia amnigena NBRC 105700
Leclercia adecarboxylata ATCC 23216
Kluyvera ascorbata ATCC 33433
Kluyvera cryocrescens NBRC 102467
Klebsiella
Enterobacter aerogenes KCTC 2190
Klebsiella oxytoca KCTC 1686
Raoultella
Pluralibacter gergoviae ATCC 33028
Yokenella regensburgei ATCC 49455
Citrobacter braakii GTA-CB04
Citrobacter youngae ATCC 29220
Kluyvera intermedia NBRC 102594
Citrobacter pasteurii 55-13
Citrobacter werkmanii NBRC 105721
Citrobacter freundii ATCC 8090
Enterobacter massiliensis JC163
Citrobacter
Salmonella enterica LT2
Salmonella bongori NCTC 12419
Escherichia/Shigella
Escherichia vulneris NBRC 102420
Trabulsiella odontotermitis TbO2.3
Trabulsiella guamensis ATCC 49490
Kosakonia
"Atlantibacter hermannii" NBRC 105704
Franconibacter helveticus LMG 23732
Franconibacter pulveris DSM 19144
Siccibacter turicensis LMG 23730
Siccibacter colletis 1383
Cronobacter
Buttiauxella agrestis ATCC 33320
Cedecea davisae DSM 4568
Cedecea neteri ATCC 33855
Klebsiella michiganensis RC10
Shimwellia blattae DSM 4481
Mangrovibacter sp. MFB070
Enterobacter-Escherichia clade
Erwinia
Tatumella
Phaseolibacter flectens ATCC 12775
Pantoea
Erwinia-Pantoea clade
Sodalis praecaptivus HS1
Sodalis glossinidius morsitans
Pectobacterium
Brenneria goodwinii OBR1
Lonsdalea quercina ATCC 29281
Dickeya
Pectobacterium-Dickeya clade
Edwardsiella
Hafnia paralvei GTA-HAF03
Hafnia alvei ATCC 13337
Obesumbacterium proteus DSM 2777
Hafnia-Edwardsiella clade
Serratia/"Chania"
Rouxiella chamberiensis 130333
Rahnella aquatilis CIP 78.65
Ewingella americana ATCC 33852
Yersinia
Yersinia-Serratia clade
Photorhabdus
Xenorhabdus
Proteus
Morganella morganii NBRC 3848
Arsenophonus nasoniae DSM 15247
Moellerella wisconsensis ATCC 35017
Providencia
Proteus-Xenorhabdus clade
Leminorella grimontii ATCC 33999
Pragia fontium 24613
Budvicia aquatica DSM 5075
Budvicia clade
D)
Erwinia oleae (GU810925)
Siccibacter turicensis (DQ273681)
Pantoea dispersa (DQ504305)
Erwinia uzenensis (AB546198)
Erwinia persicina (Z96086)
Erwinia psidii (JQ809696)
Pantoea conspicua (EU216737)
Pantoea eucalypti (EF688009)
Pluralibacter pyrinus (AJ010486)
Erwinia-Pantoea clade
Escherichia vulneris (AF530476)
Enterobacter asburiae (AB004744)
Raoultella ornithinolytica (U78182)
Buttiauxella ferragutiae (AJ233402)
Pantoea rodasii (JF295053)
Enterobacter kobei (AJ508301)
Citrobacter werkmanii (AF025373)
Yokenella regensburgei (AB519796)
Kluyvera intermedia (AF310217)
Rosenbergiella nectarea (HQ284827)
Tatumella terrea (EF688007)
Klebsiella quasipneumoniae subsp. quasipneumoniae (HG933296)
Enterobacter cloacae subsp. cloacae (AJ251469)
Citrobacter rodentium (AF025363)
Salmonella enterica subsp. indica (EU014680)
Salmonella bongori (AF029227)
Shigella flexneri (X96963)
Citrobacter koseri (HQ992945)
Trabulsiella odontotermitis (DQ453129)
Kosakonia arachidis (EU672801)
Franconibacter pulveris (DQ273684)
Shimwellia pseudoproteus (FJ267523)
Cronobacter dublinensis subsp. lactaridi (EF059838)
Cronobacter muytjensii (EF059845)
Enterobacter-Escherichia clade
Cedecea davisae (AF493976)
Serratia marcescens subsp. marcescens (AJ233431)
Serratia rubidaea (AB004751)
Gibbsiella dentisursi (AB566415)
Serratia vespertilionis (KJ739885)
Serratia quinivorans (AJ233435)
Hafnia alvei (M59155)
Rahnella aquatilis (AJ233426)
Yersinia entomophaga (DQ400782)
Yersinia pekkanenii (GQ451990)
Yersinia aleksiciae (AJ627597)
Yersinia bercovieri (AF366377)
Yersinia pseudotuberculosis (AF366375)
Yersinia-Serratia clade
Lonsdalea quercina subsp. quercina (AJ233416)
Biostraticola tofi (AM774412)
Brenneria roseae subsp. americana (KF308292)
Dickeya dadantii (AF520707)
Dickeya dadantii subsp. dieffenbachiae (AF520712)
Pectobacterium atrosepticum (Z96090)
Pectobacterium carotovorum subsp. odoriferum (AJ223407)
Brenneria alni (AJ233409)
Pectobacterium-Dickeya clade
Providencia rettgeri (AM040492)
Providencia sneebia (HM038003)
Providencia heimbachae (AM040490)
Photorhabdus luminescens subsp. kayaii (AJ560630)
Photorhabdus luminescens subsp. hainanensis (EU930342)
Photorhabdus temperata subsp. cinerea (EU136626)
Photorhabdus luminescens subsp. luminescens (X82248)
Photorhabdus temperata subsp. stackebrandtii (EF467859)
Xenorhabdus cabanillasii (AY521244)
Xenorhabdus nematophila (AY278674)
Xenorhabdus doucetiae (DQ211709)
Xenorhabdus vietnamensis (DQ205447)
Xenorhabdus budapestensis (AJ810293)
Xenorhabdus hominickii (DQ211719)
Proteus mirabilis (DQ885256)
Proteus-Xenorhabdus clade
Thorsellia kandunguensis (KM269289)
Thorselliaceae
Frischella perrara (JX878306)
Orbaceae
Pasteurellaceae

Figure 1

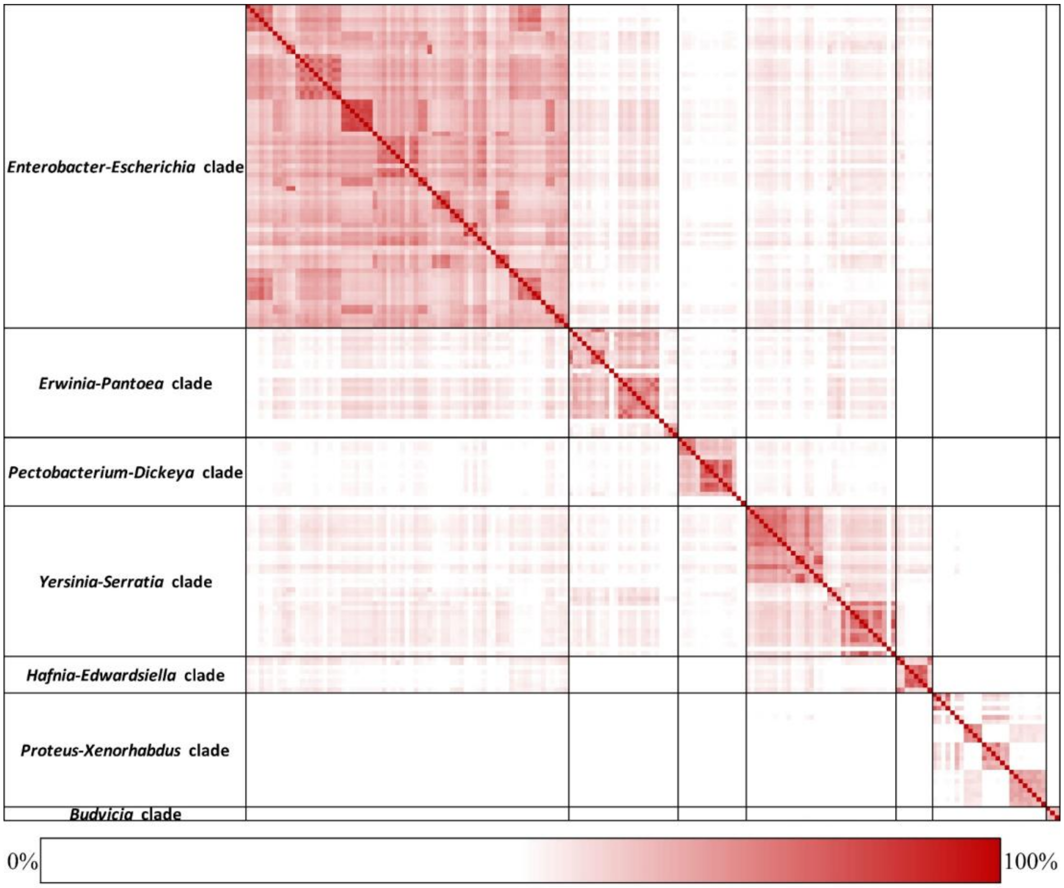
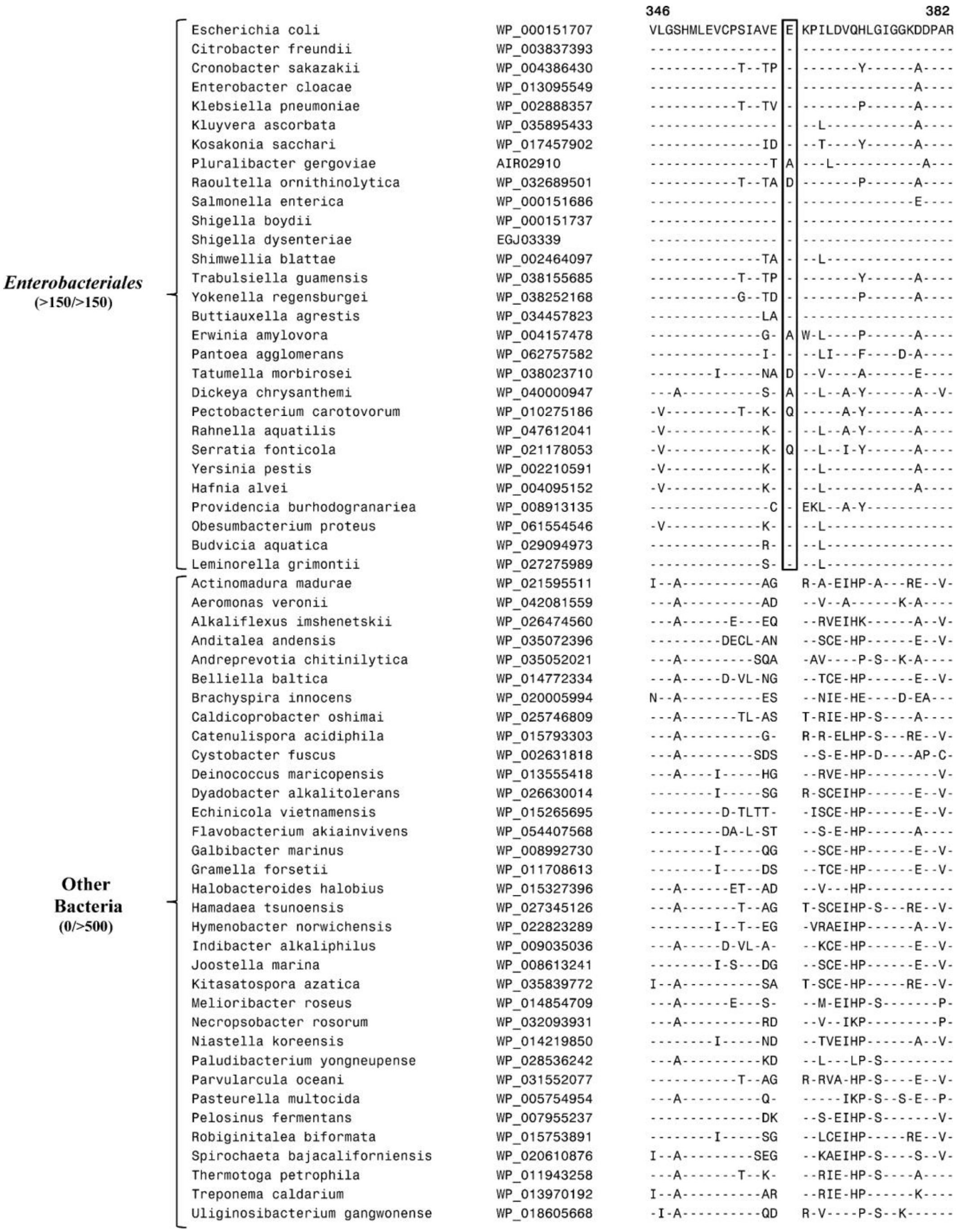

Enterobacter-Escherichia clade
Erwinia-Pantoea clade
Pectobacterium-Dickeya clade
Yersinia-Serratia clade
Hafnia-Edwardsiella clade
Proteus-Xenorhabdus clade
Budvicia clade
0%
100%

Figure 2



			346		382
Enterobacteriales **(>150/>150)**	Escherichia coli	WP_000151707	VLGSHMLEVCPSIAVE	E	KPILDVQHLGIGGKDDPAR
	Citrobacter freundii	WP_003837393	----------------	-	-------------------
	Cronobacter sakazakii	WP_004386430	-----------T--TP	-	-------Y------A----
	Enterobacter cloacae	WP_013095549	----------------	-	--------------A----
	Klebsiella pneumoniae	WP_002888357	-----------T--TV	-	-------P------A----
	Kluyvera ascorbata	WP_035895433	----------------	-	--L-----------A----
	Kosakonia sacchari	WP_017457902	--------------ID	-	--T----Y------A----
	Pluralibacter gergoviae	AIR02910	---------------T	A	---L-----------A---
	Raoultella ornithinolytica	WP_032689501	-----------T--TA	D	-------P------A----
	Salmonella enterica	WP_000151686	----------------	-	--------------E----
	Shigella boydii	WP_000151737	----------------	-	-------------------
	Shigella dysenteriae	EGJ03339	----------------	-	-------------------
	Shimwellia blattae	WP_002464097	--------------TA	-	--L----------------
	Trabulsiella guamensis	WP_038155685	-----------T--TP	-	-------Y------A----
	Yokenella regensburgei	WP_038252168	-----------G--TD	-	-------P------A----
	Buttiauxella agrestis	WP_034457823	--------------LA	-	-------------------
	Erwinia amylovora	WP_004157478	--------------G-	A	W-L----P------A----
	Pantoea agglomerans	WP_062757582	--------------I-	-	--LI---F----D-A----
	Tatumella morbirosei	WP_038023710	--------I-----NA	D	--V----A------E----
	Dickeya chrysanthemi	WP_040000947	---A----------S-	A	--L--A-Y------A--V-
	Pectobacterium carotovorum	WP_010275186	-V---------T--K-	Q	-----A-Y------A----
	Rahnella aquatilis	WP_047612041	-V------------K-	-	--L--A-Y------A----
	Serratia fonticola	WP_021178053	-V------------K-	Q	--L--I-Y------A----
	Yersinia pestis	WP_002210591	-V------------K-	-	--L-----------A----
	Hafnia alvei	WP_004095152	-V------------K-	-	--L-----------A----
	Providencia burhodogranariea	WP_008913135	---------------C	-	EKL--A-Y-----------
	Obesumbacterium proteus	WP_061554546	-V------------K-	-	--L----------------
	Budvicia aquatica	WP_029094973	--------------R-	-	--L----------------
	Leminorella grimontii	WP_027275989	--------------S-	-	--L----------------
Other Bacteria **(0/>500)**	Actinomadura madurae	WP_021595511	I--A----------AG		R-A-EIHP-A---RE--V-
	Aeromonas veronii	WP_042081559	---A----------AD		--V--A------K-A----
	Alkaliflexus imshenetskii	WP_026474560	---A------E---EQ		--RVEIHK------A--V-
	Anditalea andensis	WP_035072396	---------DECL-AN		--SCE-HP------E--V-
	Andreprevotia chitinilytica	WP_035052021	---A---------SQA		-AV----P-S--K-A----
	Belliella baltica	WP_014772334	---A-----D-VL-NG		--TCE-HP------E--V-
	Brachyspira innocens	WP_020005994	N--A----------ES		--NIE-HE----D-EA---
	Caldicoprobacter oshimai	WP_025746809	---A-------TL-AS		T-RIE-HP-S----A----
	Catenulispora acidiphila	WP_015793303	---A----------G-		R-R-ELHP-S---RE--V-
	Cystobacter fuscus	WP_002631818	---A---------SDS		--S-E-HP-D----AP-C-
	Deinococcus maricopensis	WP_013555418	---A----I-----HG		--RVE-HP---------V-
	Dyadobacter alkalitolerans	WP_026630014	--------I-----SG		R-SCEIHP------E--V-
	Echinicola vietnamensis	WP_015265695	---------D-TLTT-		-ISCE-HP------E--V-
	Flavobacterium akiainvivens	WP_054407568	---------DA-L-ST		--S-E-HP------A----
	Galbibacter marinus	WP_008992730	--------I-----QG		--SCE-HP------E--V-
	Gramella forsetii	WP_011708613	--------I-----DS		--TCE-HP------E--V-
	Halobacteroides halobius	WP_015327396	---A------ET--AD		--V---HP-----------
	Hamadaea tsunoensis	WP_027345126	---A-------T--AG		T-SCEIHP-S---RE--V-
	Hymenobacter norwichensis	WP_022823289	--------I--T--EG		-VRAEIHP------A--V-
	Indibacter alkaliphilus	WP_009035036	---A-----D-VL-A-		--KCE-HP------E--V-
	Joostella marina	WP_008613241	--------I-S---DG		--SCE-HP------E--V-
	Kitasatospora azatica	WP_035839772	I--A----------SA		T-SCE-HP-----RE--V-
	Melioribacter roseus	WP_014854709	---A------E---S-		--M-EIHP-S-------P-
	Necropsobacter rosorum	WP_032093931	---A----------RD		--V--IKP---------P-
	Niastella koreensis	WP_014219850	--------I-----ND		--TVEIHP------A--V-
	Paludibacterium yongneupense	WP_028536242	---A----------KD		--L---LP-S---------
	Parvularcula oceani	WP_031552077	-----------T--AG		R-RVA-HP-S----E--V-
	Pasteurella multocida	WP_005754954	---A----------Q-		-----IKP-S--S-E--P-
	Pelosinus fermentans	WP_007955237	--------------DK		--S-EIHP-S-------V-
	Robiginitalea biformata	WP_015753891	--------I-----SG		--LCEIHP-----RE--V-
	Spirochaeta bajacaliforniensis	WP_020610876	I--A---------SEG		--KAEIHP-S----S--V-
	Thermotoga petrophila	WP_011943258	---A-------T--K-		--RIE-HP-S----A----
	Treponema caldarium	WP_013970192	I--A----------AR		--RIE-HP------K----
	Uliginosibacterium gangwonense	WP_018605668	-I-A----------QD		R-V----P-S--K------

Figure 3

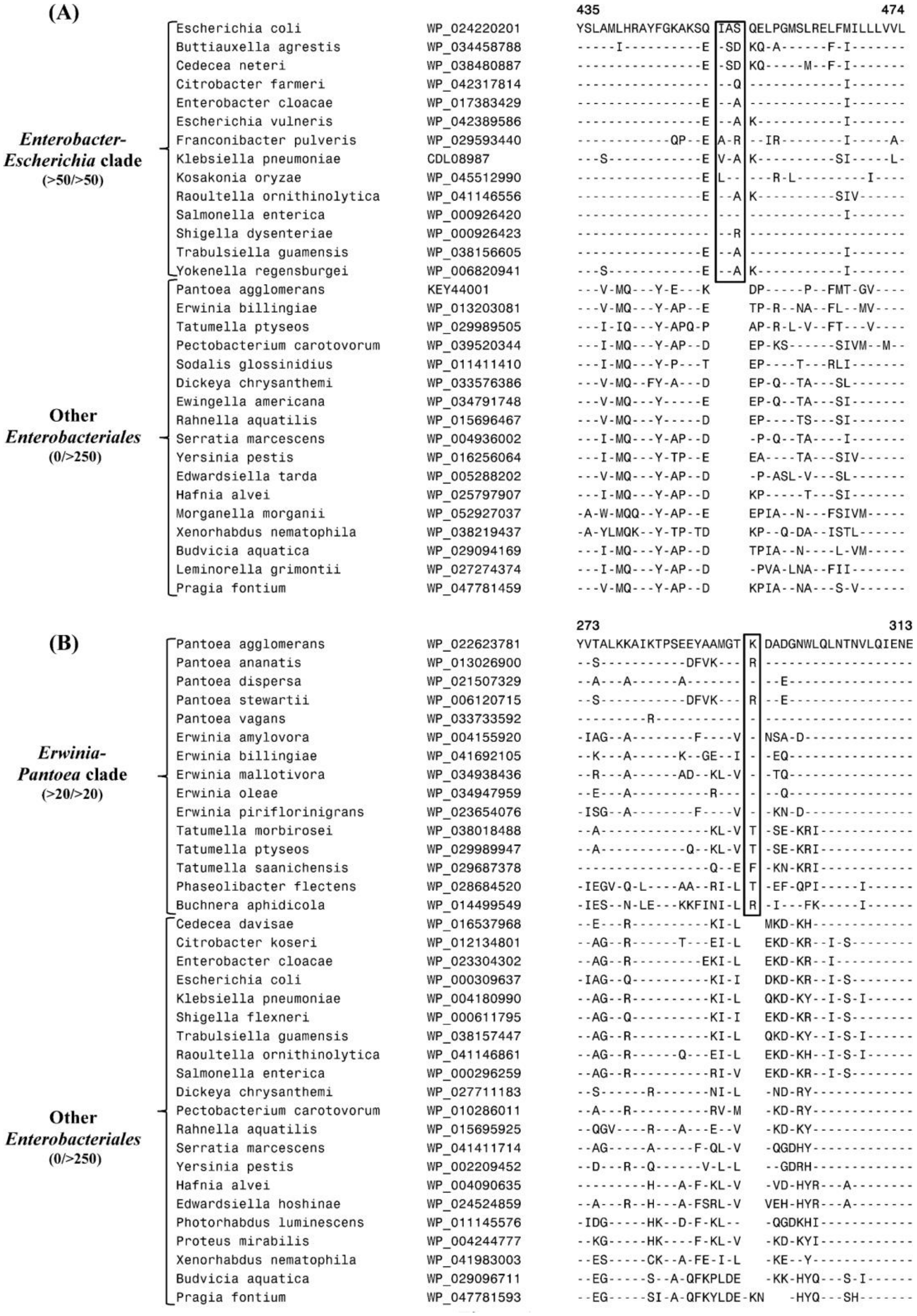

(A)
435
474
Escherichia coli WP_024220201 YSLAMLHRAYFGKAKSQ IAS QELPGMSLRELFMILLLVVL
Buttiauxella agrestis WP_034458788 -----I----------E -SD KQ-A------F-I-------
Cedecea neteri WP_038480887 ----------------E -SD KQ-----M--F-I-------
Citrobacter farmeri WP_042317814 ----------------- --Q ------------I-------
Enterobacter cloacae WP_017383429 ----------------E --A ------------I-------
Escherichia vulneris WP_042389586 ----------------E --A K-----------I-------
Franconibacter pulveris WP_029593440 -----------QP--E A-R --IR--------I-----A-
Klebsiella pneumoniae CDL08987 ---S-----------E V-A K----------SI-----L-
Kosakonia oryzae WP_045512990 ----------------E L-- ---R-L---------I----
Raoultella ornithinolytica WP_041146556 ----------------E --A K----------SIV------
Salmonella enterica WP_000926420 ----------------- --- ------------I-------
Shigella dysenteriae WP_000926423 ----------------- --R --------------------
Trabulsiella guamensis WP_038156605 ----------------E --A ------------I-------
Yokenella regensburgei WP_006820941 ---S-----------E --A K-----------I-------
Pantoea agglomerans KEY44001 ---V-MQ---Y-E---K DP-----P--FMT-GV----
Erwinia billingiae WP_013203081 ---V-MQ---Y-AP--E TP-R--NA--FL--MV----
Tatumella ptyseos WP_029989505 ---I-IQ---Y-APQ-P AP-R-L-V--FT---V----
Pectobacterium carotovorum WP_039520344 ---I-MQ---Y-AP--D EP-KS------SIVM--M--
Sodalis glossinidius WP_011411410 ---I-MQ---Y-P---T EP----T---RLI-------
Dickeya chrysanthemi WP_033576386 ---V-MQ--FY-A---D EP-Q--TA---SL-------
Ewingella americana WP_034791748 ---V-MQ---Y-----E EP-Q--TA---SI-------
Rahnella aquatilis WP_015696467 ---V-MQ---Y-----D EP----TS---SI-------
Serratia marcescens WP_004936002 ---I-MQ---Y-AP--D -P-Q--TA----I-------
Yersinia pestis WP_016256064 ---I-MQ---Y-TP--E EA----TA---SIV------
Edwardsiella tarda WP_005288202 ---V-MQ---Y-AP--D -P-ASL-V---SL-------
Hafnia alvei WP_025797907 ---I-MQ---Y-AP--D KP-----T---SI-------
Morganella morganii WP_052927037 -A-W-MQQ--Y-AP--E EPIA--N---FSIVM-----
Xenorhabdus nematophila WP_038219437 -A-YLMQK--Y-TP-TD KP--Q-DA--ISTL------
Budvicia aquatica WP_029094169 ---I-MQ---Y-AP--D TPIA--N----L-VM-----
Leminorella grimontii WP_027274374 ---I-MQ---Y-AP--D -PVA-LNA--FII-------
Pragia fontium WP_047781459 ---V-MQ---Y-AP--D KPIA--NA---S-V------
Enterobacter-Escherichia clade
(>50/>50)
Other Enterobacteriales
(0/>250)
(B)
273
313
Pantoea agglomerans WP_022623781 YVTALKKAIKTPSEEYAAMGT K DADGNWLQLNTNVLQIENE
Pantoea ananatis WP_013026900 --S-----------DFVK--- R -------------------
Pantoea dispersa WP_021507329 --A---A------A------- - --E----------------
Pantoea stewartii WP_006120715 --S-----------DFVK--- R --E----------------
Pantoea vagans WP_033733592 ---------R----------- - -------------------
Erwinia amylovora WP_004155920 -IAG--A--------F----V - NSA-D--------------
Erwinia billingiae WP_041692105 --K---A------K--GE--I - -EQ----------------
Erwinia mallotivora WP_034938436 --R---A------AD--KL-V - -TQ----------------
Erwinia oleae WP_034947959 --E---A----------R--- - --Q----------------
Erwinia piriflorinigrans WP_023654076 -ISG--A--------F----V - -KN-D--------------
Tatumella morbirosei WP_038018488 --A--------------KL-V T -SE-KRI------------
Tatumella ptyseos WP_029989947 --A-----------Q--KL-V T -SE-KRI------------
Tatumella saanichensis WP_029687378 -----------------Q--E F -KN-KRI------------
Phaseolibacter flectens WP_028684520 -IEGV-Q-L----AA--RI-L T -EF-QPI-----I------
Buchnera aphidicola WP_014499549 -IES--N-LE---KKFINI-L R -I---FK-----I------
Cedecea davisae WP_016537968 --E---R----------KI-L MKD-KH-------------
Citrobacter koseri WP_012134801 --AG--R------T---EI-L EKD-KR--I-S--------
Enterobacter cloacae WP_023304302 --AG--R---------EKI-L EKD-KR--I----------
Escherichia coli WP_000309637 -IAG--Q----------KI-I DKD-KR--I-S--------
Klebsiella pneumoniae WP_004180990 --AG--R----------KI-L QKD-KY--I-S-I------
Shigella flexneri WP_000611795 --AG--Q----------KI-I EKD-KR--I-S--------
Trabulsiella guamensis WP_038157447 --AG--R----------KI-L QKD-KY--I-S-I------
Raoultella ornithinolytica WP_041146861 --AG--R------Q---EI-L EKD-KH--I-S-I------
Salmonella enterica WP_000296259 --AG--R----------RI-V EKD-KR--I-S--------
Dickeya chrysanthemi WP_027711183 --S------R-------NI-L -ND-RY-------------
Pectobacterium carotovorum WP_010286011 --A---R----------RV-M -KD-RY-------------
Rahnella aquatilis WP_015695925 --QGV----R---A---E--V -KD-KY-------------
Serratia marcescens WP_041411714 --AG-----A-----F-QL-V -QGDHY-------------
Yersinia pestis WP_002209452 --D---R--Q------V-L-L --GDRH-------------
Hafnia alvei WP_004090635 ---------H---A-F-KL-V -VD-HYR---A--------
Edwardsiella hoshinae WP_024524859 --A---R--H---A-FSRL-V VEH-HYR---A--------
Photorhabdus luminescens WP_011145576 -IDG-----HK--D-F-KL-- -QGDKHI------------
Proteus mirabilis WP_004244777 --KG-----HK----F-KL-V -KD-KYI------------
Xenorhabdus nematophila WP_041983003 --ES-----CK--A-FE-I-L -KE--Y-------------
Budvicia aquatica WP_029096711 --EG-----S--A-QFKPLDE -KK-HYQ---S-I------
Pragia fontium WP_047781593 --EG-----SI-A-QFKYLDE-KN -HYQ---SH-------
Erwinia-Pantoea clade
(>20/>20)
Other Enterobacteriales
(0/>250)

Figure4

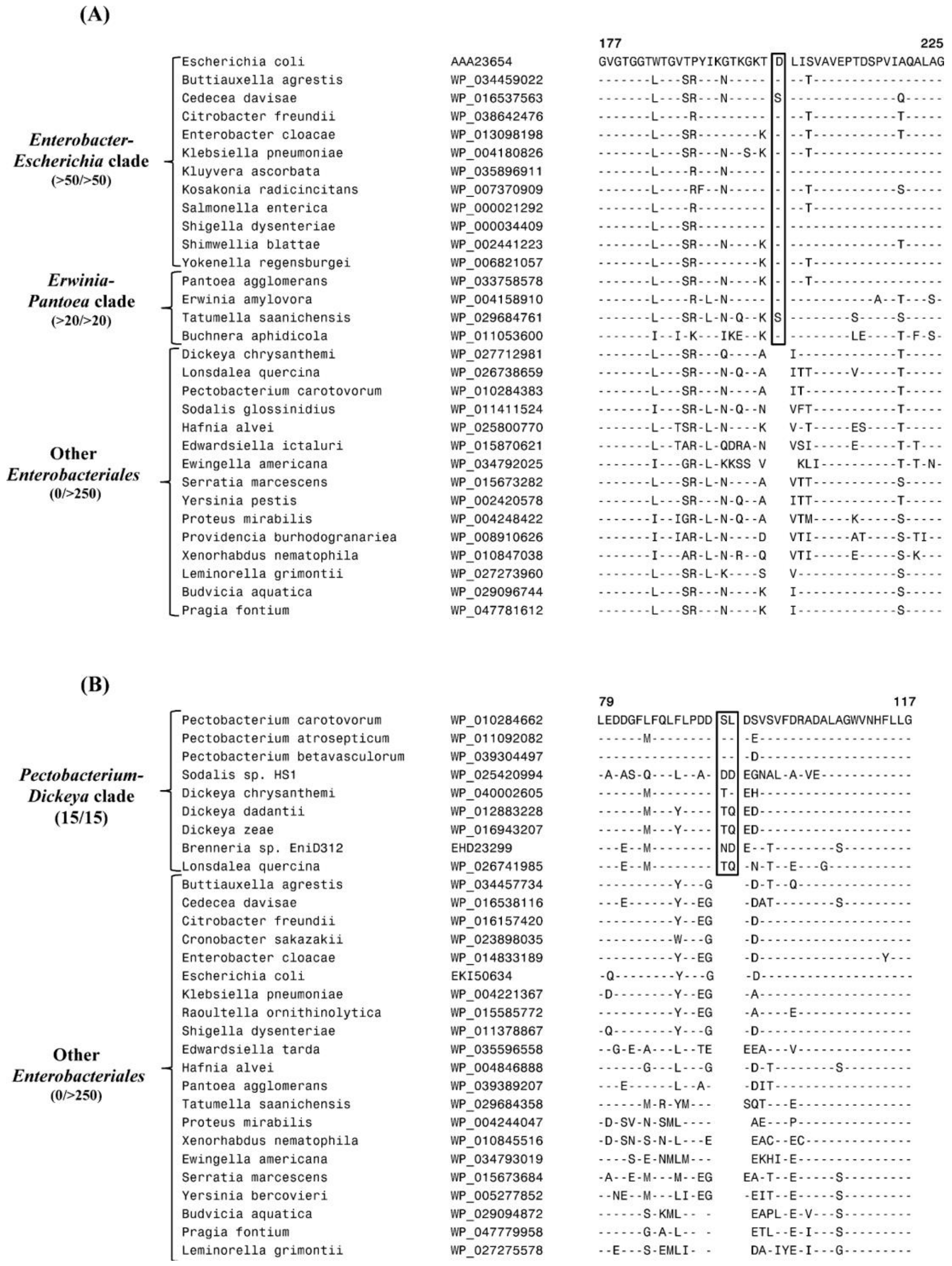

(A)
177
225
Enterobacter-Escherichia clade (>50/>50)
Erwinia-Pantoea clade (>20/>20)
Other Enterobacteriales (0/>250)
Escherichia coli AAA23654 GVGTGGTWTGVTPYIKGTKGKT D LISVAVEPTDSPVIAQALAG
Buttiauxella agrestis WP_034459022 -------L---SR---N----- - --T-----------------
Cedecea davisae WP_016537563 -------L---SR---N----- S --------------Q-----
Citrobacter freundii WP_038642476 -------L----R--------- - --T-----------T-----
Enterobacter cloacae WP_013098198 -------L---SR--------K - --T-----------T-----
Klebsiella pneumoniae WP_004180826 -------L---SR---N--S-K - --T-----------------
Kluyvera ascorbata WP_035896911 -------L----R---N----- - --------------------
Kosakonia radicincitans WP_007370909 -------L----RF--N----- - --T-----------S-----
Salmonella enterica WP_000021292 -------L----R--------- - --T-----------------
Shigella dysenteriae WP_000034409 -------L---SR--------- - --------------------
Shimwellia blattae WP_002441223 -------L---SR---N----K - --------------T-----
Yokenella regensburgei WP_006821057 -------L---SR--------K - --T-----------------
Pantoea agglomerans WP_033758578 -------L---SR---N----K - --T-----------------
Erwinia amylovora WP_004158910 -------L----R-L-N----- - -----------A--T---S-
Tatumella saanichensis WP_029684761 -------L---SR-L-N-Q--K S --------S-----S-----
Buchnera aphidicola WP_011053600 -------I--I-K---IKE--K - --------LE----T-F-S-
Dickeya chrysanthemi WP_027712981 -------L---SR---Q----A I-------------T-----
Lonsdalea quercina WP_026738659 -------L---SR---N-Q--A ITT-----V-----T-----
Pectobacterium carotovorum WP_010284383 -------L---SR---N----A IT------------T-----
Sodalis glossinidius WP_011411524 -------I---SR-L-N-Q--N VFT-----------T-----
Hafnia alvei WP_025800770 -------L--TSR-L-N----K V-T-----ES----T-----
Edwardsiella ictaluri WP_015870621 -------L--TAR-L-QDRA-N VSI-----E-----T-T---
Ewingella americana WP_034792025 -------I---GR-L-KKSS V KLI----------T-T-N-
Serratia marcescens WP_015673282 -------L---SR-L-N----A VTT-----------S-----
Yersinia pestis WP_002420578 -------L---SR---N-Q--A ITT-----------T-----
Proteus mirabilis WP_004248422 -------I--IGR-L-N-Q--A VTM-----K-----S-----
Providencia burhodogranariea WP_008910626 -------I--IAR-L-N----D VTI-----AT----S-TI--
Xenorhabdus nematophila WP_010847038 -------I---AR-L-N-R--Q VTI-----E-----S-K---
Leminorella grimontii WP_027273960 -------L---SR-L-K----S V-------------S-----
Budvicia aquatica WP_029096744 -------L---SR---N----K I-------------S-----
Pragia fontium WP_047781612 -------L---SR---N----K I-------------S-----
(B)
79
117
Pectobacterium-Dickeya clade (15/15)
Other Enterobacteriales (0/>250)
Pectobacterium carotovorum WP_010284662 LEDDGFLFQLFLPDD SL DSVSVFDRADALAGWVNHFLLG
Pectobacterium atrosepticum WP_011092082 ------M-------- -- -E--------------------
Pectobacterium betavasculorum WP_039304497 --------------- -- -D--------------------
Sodalis sp. HS1 WP_025420994 -A-AS-Q---L--A- DD EGNAL-A-VE------------
Dickeya chrysanthemi WP_040002605 ------M-------- T- EH--------------------
Dickeya dadantii WP_012883228 ------M---Y---- TQ ED--------------------
Dickeya zeae WP_016943207 ------M---Y---- TQ ED--------------------
Brenneria sp. EniD312 EHD23299 ---E--M-------- ND E--T--------S---------
Lonsdalea quercina WP_026741985 ---E--M-------- TQ -N-T--E---G-----------
Buttiauxella agrestis WP_034457734 ----------Y---G -D-T--Q---------------
Cedecea davisae WP_016538116 ---E------Y--EG -DAT--------S---------
Citrobacter freundii WP_016157420 ----------Y--EG -D--------------------
Cronobacter sakazakii WP_023898035 ----------W---G -D--------------------
Enterobacter cloacae WP_014833189 ----------Y--EG -D----------------Y---
Escherichia coli EKI50634 -Q--------Y---G -D--------------------
Klebsiella pneumoniae WP_004221367 -D--------Y--EG -A--------------------
Raoultella ornithinolytica WP_015585772 ----------Y--EG -A---E----------------
Shigella dysenteriae WP_011378867 -Q--------Y---G -D--------------------
Edwardsiella tarda WP_035596558 --G-E-A---L--TE EEA---V---------------
Hafnia alvei WP_004846888 ------G---L---G -D-T--------S---------
Pantoea agglomerans WP_039389207 ---E------L--A- -DIT------------------
Tatumella saanichensis WP_029684358 ------M-R-YM--- SQT---E---------------
Proteus mirabilis WP_004244047 -D-SV-N-SML---- AE---P---------------
Xenorhabdus nematophila WP_010845516 -D-SN-S-N-L---E EAC--EC-------------
Ewingella americana WP_034793019 ----S-E-NMLM--- EKHI-E---------------
Serratia marcescens WP_015673684 -A--E-M---M--EG EA-T--E-----S---------
Yersinia bercovieri WP_005277852 --NE--M---LI-EG -EIT--E-----S---------
Budvicia aquatica WP_029094872 ------S-KML-- - EAPL-E-V---S---------
Pragia fontium WP_047779958 ------G-A-L-- - ETL--E-I---S---------
Leminorella grimontii WP_027275578 --E---S-EMLI- - DA-IYE-I---G---------

Figure 5

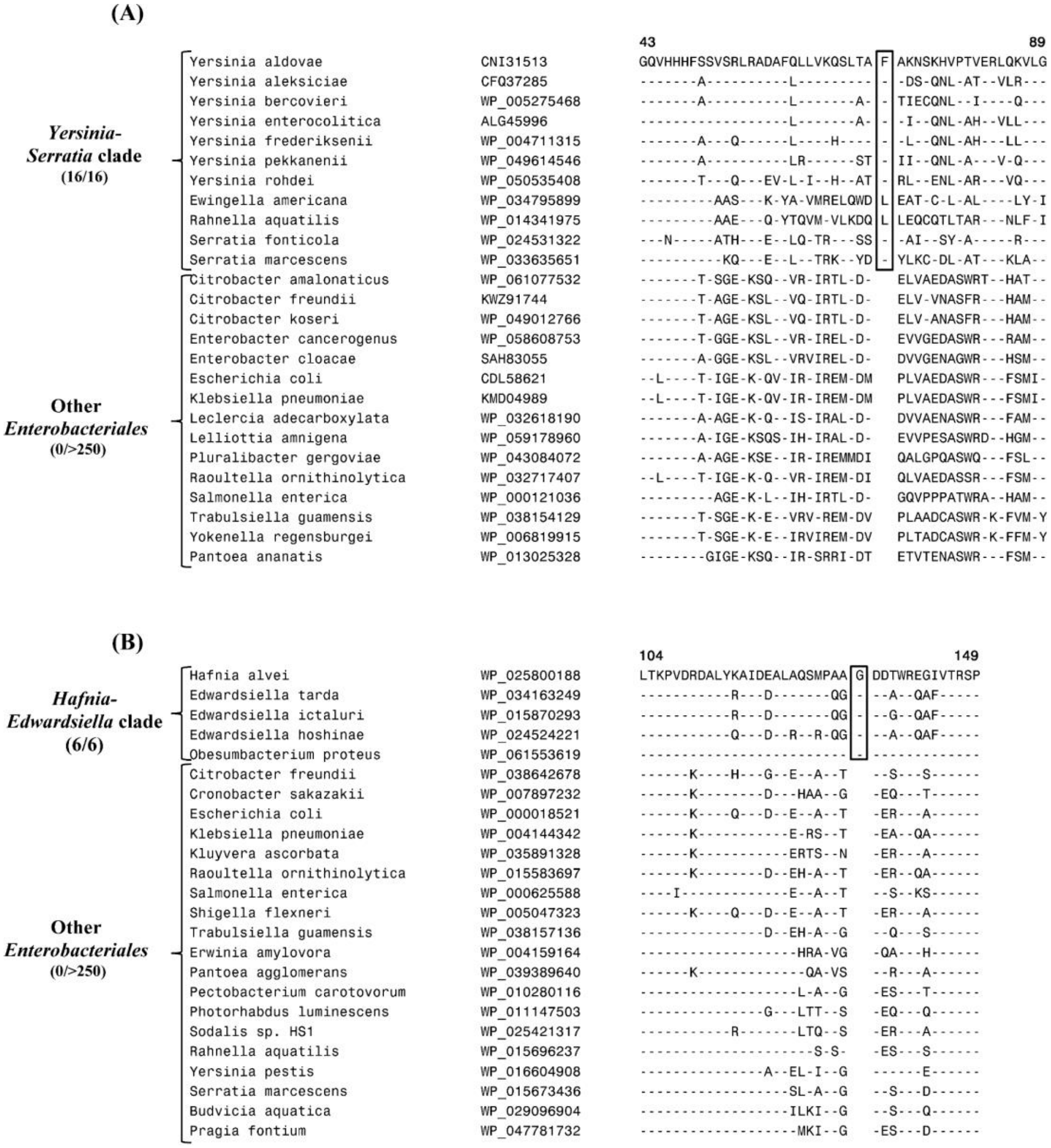
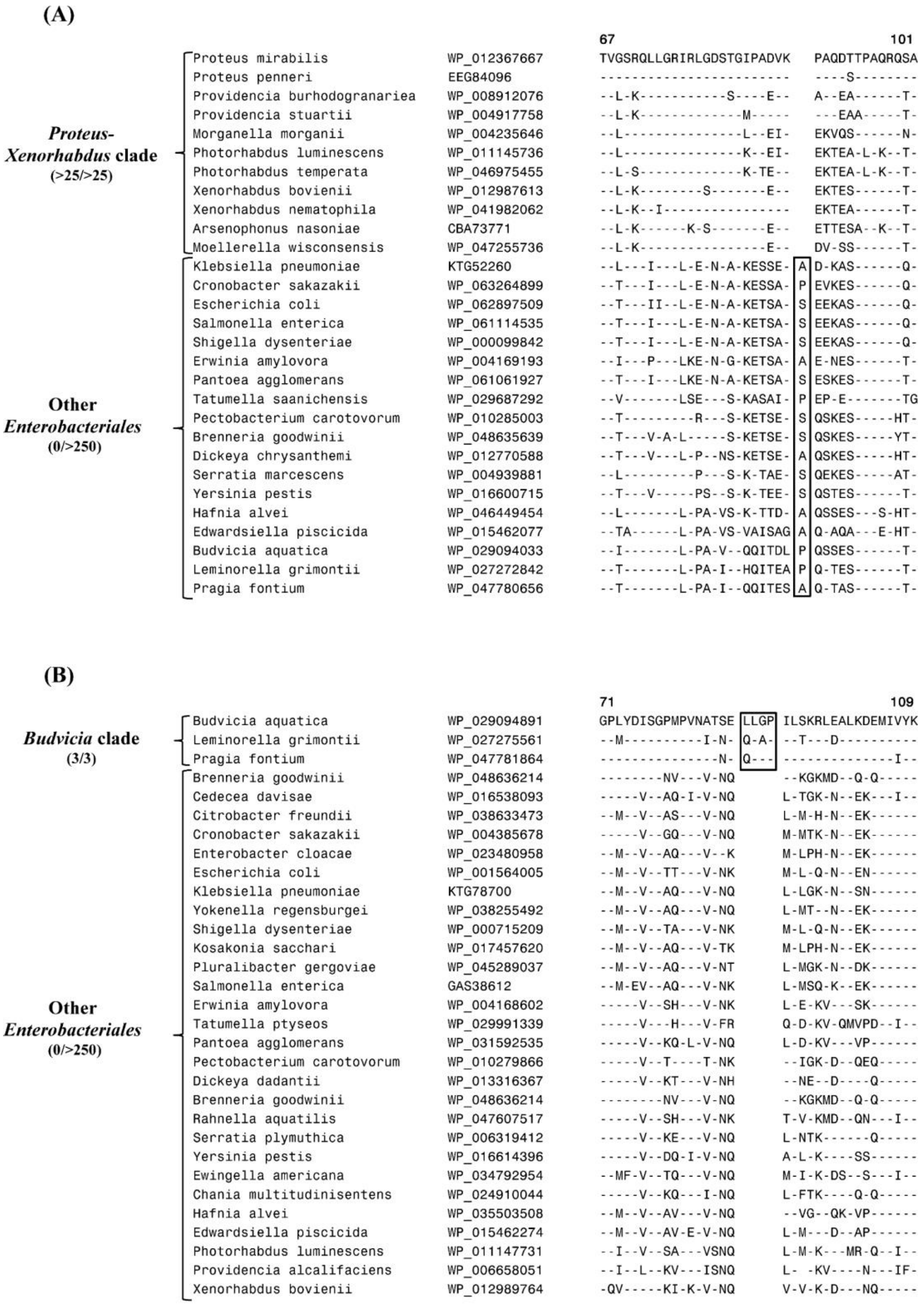

(A)
43
89
Yersinia-Serratia clade
(16/16)
Other Enterobacteriales
(0/>250)
Yersinia aldovae CNI31513 GQVHHHFSSVSRLRADAFQLLVKQSLTA F AKNSKHVPTVERLQKVLG
Yersinia aleksiciae CFQ37285 -------A----------L--------- - -DS-QNL-AT--VLR---
Yersinia bercovieri WP_005275468 -------A----------L-------A- - TIECQNL--I----Q---
Yersinia enterocolitica ALG45996 ------------------L-------A- - -I--QNL-AH--VLL---
Yersinia frederiksenii WP_004711315 -------A---Q------L----H---- - -L--QNL-AH---LL---
Yersinia pekkanenii WP_049614546 -------A----------LR------ST - II--QNL-A---V-Q---
Yersinia rohdei WP_050535408 -------T---Q---EV-L-I--H--AT - RL--ENL-AR---VQ---
Ewingella americana WP_034795899 ---------AAS---K-YA-VMRELQWD L EAT-C-L-AL----LY-I
Rahnella aquatilis WP_014341975 ---------AAE---Q-YTQVM-VLKDQ L LEQCQTLTAR---NLF-I
Serratia fonticola WP_024531322 ---N-----ATH---E--LQ-TR---SS - -AI--SY-A-----R---
Serratia marcescens WP_033635651 ----------KQ---E--L--TRK--YD - YLKC-DL-AT---KLA--
Citrobacter amalonaticus WP_061077532 -------T-SGE-KSQ--VR-IRTL-D- ELVAEDASWRT--HAT--
Citrobacter freundii KWZ91744 -------T-AGE-KSL--VQ-IRTL-D- ELV-VNASFR---HAM--
Citrobacter koseri WP_049012766 -------T-AGE-KSL--VQ-IRTL-D- ELV-ANASFR---HAM--
Enterobacter cancerogenus WP_058608753 -------T-GGE-KSL--VR-IREL-D- EVVGEDASWR---RAM--
Enterobacter cloacae SAH83055 -------A-GGE-KSL--VRVIREL-D- DVVGENAGWR---HSM--
Escherichia coli CDL58621 --L----T-IGE-K-QV-IR-IREM-DM PLVAEDASWR---FSMI-
Klebsiella pneumoniae KMD04989 --L----T-IGE-K-QV-IR-IREM-DM PLVAEDASWR---FSMI-
Leclercia adecarboxylata WP_032618190 -------A-AGE-K-Q--IS-IRAL-D- DVVAENASWR---FAM--
Lelliottia amnigena WP_059178960 -------A-IGE-KSQS-IH-IRAL-D- EVVPESASWRD--HGM--
Pluralibacter gergoviae WP_043084072 -------A-AGE-KSE--IR-IREMMDI QALGPQASWQ---FSL--
Raoultella ornithinolytica WP_032717407 --L----T-IGE-K-Q--VR-IREM-DI QLVAEDASSR---FSM--
Salmonella enterica WP_000121036 ---------AGE-K-L--IH-IRTL-D- GQVPPPATWRA--HAM--
Trabulsiella guamensis WP_038154129 -------T-SGE-K-E--VRV-REM-DV PLAADCASWR-K-FVM-Y
Yokenella regensburgei WP_006819915 -------T-SGE-K-E--IRVIREM-DV PLTADCASWR-K-FFM-Y
Pantoea ananatis WP_013025328 --------GIGE-KSQ--IR-SRRI-DT ETVTENASWR---FSM--
(B)
104
149
Hafnia-Edwardsiella clade
(6/6)
Other Enterobacteriales
(0/>250)
Hafnia alvei WP_025800188 LTKPVDRDALYKAIDEALAQSMPAA G DDTWREGIVTRSP
Edwardsiella tarda WP_034163249 -----------R---D-------QG - --A--QAF-----
Edwardsiella ictaluri WP_015870293 -----------R---D-------QG - --G--QAF-----
Edwardsiella hoshinae WP_024524221 -----------Q---D--R--R-QG - --A--QAF-----
Obesumbacterium proteus WP_061553619 ------------------------- - -------------
Citrobacter freundii WP_038642678 ------K----H---G--E--A--T --S---S------
Cronobacter sakazakii WP_007897232 ------K--------D---HAA--G -EQ---T------
Escherichia coli WP_000018521 ------K----Q---D--E--A--T -ER---A------
Klebsiella pneumoniae WP_004144342 ------K-----------E-RS--T -EA--QA------
Kluyvera ascorbata WP_035891328 ------K-----------ERTS--N -ER---A------
Raoultella ornithinolytica WP_015583697 ------K--------D--EH-A--T -ER--QA------
Salmonella enterica WP_000625588 ----I-------------E--A--T --S--KS------
Shigella flexneri WP_005047323 ------K----Q---D--E--A--T -ER---A------
Trabulsiella guamensis WP_038157136 ---------------D--EH-A--G --Q---S------
Erwinia amylovora WP_004159164 -------------------HRA-VG -QA---H------
Pantoea agglomerans WP_039389640 ------K-------------QA-VS --R---A------
Pectobacterium carotovorum WP_010280116 -------------------L-A--G -ES---T------
Photorhabdus luminescens WP_011147503 ---------------G---LTT--S -EQ---Q------
Sodalis sp. HS1 WP_025421317 -----------R-------LTQ--S -ER---A------
Rahnella aquatilis WP_015696237 ---------------------S-S- -ES---S------
Yersinia pestis WP_016604908 ---------------A--EL-I--G ------E------
Serratia marcescens WP_015673436 ------------------SL-A--G --S---D------
Budvicia aquatica WP_029096904 ------------------ILKI--G --S---Q------
Pragia fontium WP_047781732 -------------------MKI--G -ES---D------

Figure 6



			67		101
Proteus-Xenorhabdus clade (>25/>25)	Proteus mirabilis	WP_012367667	TVGSRQLLGRIRLGDSTGIPADVK		PAQDTTPAQRQSA
	Proteus penneri	EEG84096	------------------------		----S--------
	Providencia burhodogranariea	WP_008912076	--L-K-----------S----E--		A--EA------T-
	Providencia stuartii	WP_004917758	--L-K-------------M-----		---EAA-----T-
	Morganella morganii	WP_004235646	--L---------------L--EI-		EKVQS------N-
	Photorhabdus luminescens	WP_011145736	--L---------------K--EI-		EKTEA-L-K--T-
	Photorhabdus temperata	WP_046975455	--L-S-------------K-TE--		EKTEA-L-K--T-
	Xenorhabdus bovienii	WP_012987613	--L-K--------S-------E--		EKTES------T-
	Xenorhabdus nematophila	WP_041982062	--L-K--I----------------		EKTEA------T-
	Arsenophonus nasoniae	CBA73771	--L-K------K-S-------E--		ETTESA--K--T-
	Moellerella wisconsensis	WP_047255736	--L-K----------------E--		DV-SS------T-
Other *Enterobacteriales* (0/>250)	Klebsiella pneumoniae	KTG52260	--L---I---L-E-N-A-KESSE-	A	D-KAS------Q-
	Cronobacter sakazakii	WP_063264899	--T---I---L-E-N-A-KESSA-	P	EVKES------Q-
	Escherichia coli	WP_062897509	--T---II--L-E-N-A-KETSA-	S	EEKAS------Q-
	Salmonella enterica	WP_061114535	--T---I---L-E-N-A-KETSA-	S	EEKAS------Q-
	Shigella dysenteriae	WP_000099842	--T---I---L-E-N-A-KETSA-	S	EEKAS------Q-
	Erwinia amylovora	WP_004169193	--I---P---LKE-N-G-KETSA-	A	E-NES------T-
	Pantoea agglomerans	WP_061061927	--T---I---LKE-N-A-KETSA-	S	ESKES------T-
	Tatumella saanichensis	WP_029687292	--V-------LSE---S-KASAI-	P	EP-E-------TG
	Pectobacterium carotovorum	WP_010285003	--T---------R---S-KETSE-	S	QSKES-----HT-
	Brenneria goodwinii	WP_048635639	--T---V-A-L-----S-KETSE-	S	QSKES-----YT-
	Dickeya chrysanthemi	WP_012770588	--T---V---L-P--NS-KETSE-	A	QSKES-----HT-
	Serratia marcescens	WP_004939881	--L---------P---S-K-TAE-	S	QEKES-----AT-
	Yersinia pestis	WP_016600715	--T---V-----PS--S-K-TEE-	S	QSTES------T-
	Hafnia alvei	WP_046449454	--L-------L-PA-VS-K-TTD-	A	QSSES---S-HT-
	Edwardsiella piscicida	WP_015462077	--TA------L-PA-VS-VAISAG	A	Q-AQA---E-HT-
	Budvicia aquatica	WP_029094033	--I-------L-PA-V--QQITDL	P	QSSES------T-
	Leminorella grimontii	WP_027272842	--T-------L-PA-I--HQITEA	P	Q-TES------T-
	Pragia fontium	WP_047780656	--T-------L-PA-I--QQITES	A	Q-TAS------T-

			71		109
Budvicia clade (3/3)	Budvicia aquatica	WP_029094891	GPLYDISGPMPVNATSE	LLGP	ILSKRLEALKDEMIVYK
	Leminorella grimontii	WP_027275561	--M----------I-N-	Q-A-	--T---D----------
	Pragia fontium	WP_047781864	---------------N-	Q---	--------------I--
Other *Enterobacteriales* (0/>250)	Brenneria goodwinii	WP_048636214	--------NV---V-NQ		--KGKMD--Q-Q-----
	Cedecea davisae	WP_016538093	-----V--AQ-I-V-NQ		L-TGK-N--EK---I--
	Citrobacter freundii	WP_038633473	--M--V--AS---V-NQ		L-M-H-N--EK------
	Cronobacter sakazakii	WP_004385678	-----V--GQ---V-NQ		M-MTK-N--EK------
	Enterobacter cloacae	WP_023480958	--M--V--AQ---V--K		M-LPH-N--EK------
	Escherichia coli	WP_001564005	--M--V--TT---V-NK		M-L-Q-N--EN------
	Klebsiella pneumoniae	KTG78700	--M--V--AQ---V-NQ		L-LGK-N--SN------
	Yokenella regensburgei	WP_038255492	--M--V--AQ---V-NQ		L-MT--N--EK------
	Shigella dysenteriae	WP_000715209	--M--V--TA---V-NK		M-L-Q-N--EK------
	Kosakonia sacchari	WP_017457620	--M--V--AQ---V-TK		M-LPH-N--EK------
	Pluralibacter gergoviae	WP_045289037	--M--V--AQ---V-NT		L-MGK-N--DK------
	Salmonella enterica	GAS38612	--M-EV--AQ---V-NK		L-MSQ-K--EK------
	Erwinia amylovora	WP_004168602	-----V--SH---V-NK		L-E-KV---SK------
	Tatumella ptyseos	WP_029991339	-----V---H---V-FR		Q-D-KV-QMVPD--I--
	Pantoea agglomerans	WP_031592535	-----V--KQ-L-V-NQ		L-D-KV---VP------
	Pectobacterium carotovorum	WP_010279866	-----V--T----T-NK		--IGK-D--QEQ-----
	Dickeya dadantii	WP_013316367	-----V--KT---V-NH		--NE--D----Q-----
	Brenneria goodwinii	WP_048636214	--------NV---V-NQ		--KGKMD--Q-Q-----
	Rahnella aquatilis	WP_047607517	-----V--SH---V-NK		T-V-KMD--QN---I--
	Serratia plymuthica	WP_006319412	-----V--KE---V-NQ		L-NTK------Q-----
	Yersinia pestis	WP_016614396	-----V--DQ-I-V-NQ		A-L-K----SS------
	Ewingella americana	WP_034792954	--MF-V--TQ---V-NQ		M-I-K-DS--S---I--
	Chania multitudinisentens	WP_024910044	-----V--KQ---I-NQ		L-FTK----Q-Q-----
	Hafnia alvei	WP_035503508	--M--V--AV---V-NQ		--VG--QK-VP------
	Edwardsiella piscicida	WP_015462274	--M--V--AV-E-V-NQ		L-M---D--AP------
	Photorhabdus luminescens	WP_011147731	--I--V--SA---VSNQ		L-M-K---MR-Q--I--
	Providencia alcalifaciens	WP_006658051	--I--L--KV---ISNQ		L- -KV----N---IF-
	Xenorhabdus bovienii	WP_012989764	-QV-----KI-K-V-NQ		V-V-K-D---NQ-----

Figure 7

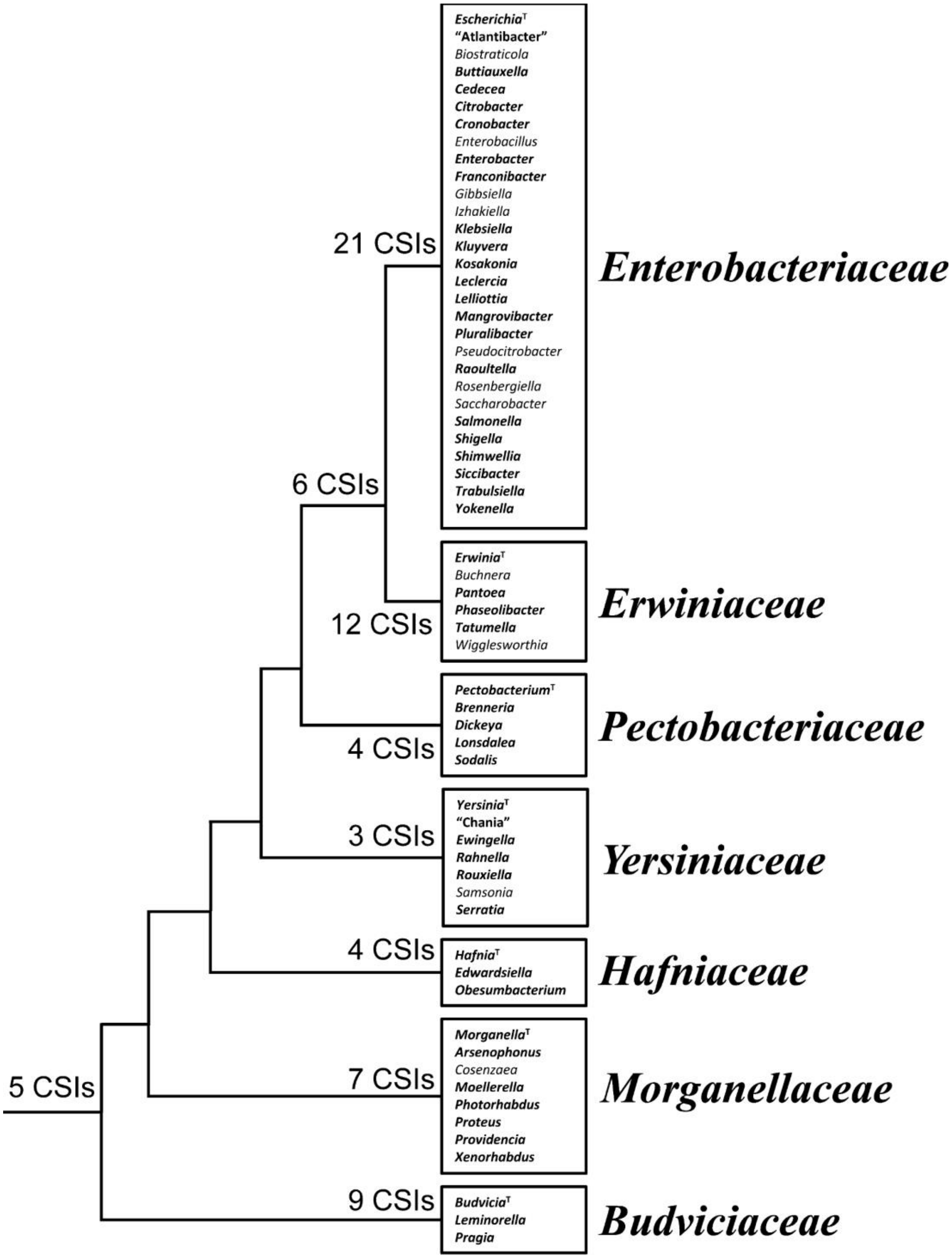

Escherichia^T
"Atlantibacter"
Biostraticola
Buttiauxella
Cedecea
Citrobacter
Cronobacter
Enterobacillus
Enterobacter
Franconibacter
Gibbsiella
Izhakiella
Klebsiella
Kluyvera
Kosakonia
Leclercia
Lelliottia
Mangrovibacter
Pluralibacter
Pseudocitrobacter
Raoultella
Rosenbergiella
Saccharobacter
Salmonella
Shigella
Shimwellia
Siccibacter
Trabulsiella
Yokenella
21 CSIs
Enterobacteriaceae
6 CSIs
Erwinia^T
Buchnera
Pantoea
Phaseolibacter
Tatumella
Wigglesworthia
12 CSIs
Erwiniaceae
Pectobacterium^T
Brenneria
Dickeya
Lonsdalea
Sodalis
4 CSIs
Pectobacteriaceae
Yersinia^T
"Chania"
Ewingella
Rahnella
Rouxiella
Samsonia
Serratia
3 CSIs
Yersiniaceae
Hafnia^T
Edwardsiella
Obesumbacterium
4 CSIs
Hafniaceae
Morganella^T
Arsenophonus
Cosenzaea
Moellerella
Photorhabdus
Proteus
Providencia
Xenorhabdus
7 CSIs
Morganellaceae
5 CSIs
Budvicia^T
Leminorella
Pragia
9 CSIs
Budviciaceae

Figure 8

CHAPTER 8

Discussion and Conclusions

The Impact of Genome Based Phylogeny and Taxonomy

Elucidating the evolutionary history of an organism provides insights into the current, past, and potential future states of the ecological, phenotypic, physiological, molecular, and biochemical characteristics of that organism. Thus, biological classifications and taxonomy, the primary means by which the evolutionary relationships between organisms are systematized and conveyed, are centrally important to Biology as a whole. However, the bases by which prokaryotic taxonomic classifications are determined are often subjective and contain several drawbacks (Schleifer, 2009; Jones, 2012; Vandamme & Peeters, 2014; Sutcliffe, 2015; Thompson et al., 2015). Most notably, the phenotypic and biochemical assays used in traditional, polyphasic taxonomic descriptions produce results which exhibit high variability and poor reproducibility, and the characteristics which these assays are used to assess are often highly plastic and can vary between strains of a single species (Vandamme & Peeters, 2014; Sutcliffe, 2015; Thompson et al., 2015). Thus, modern prokaryotic taxonomy is heavily reliant on the genetic component of taxonomic descriptions, which are often solely limited to analysis of the 16S rRNA gene (Schleifer, 2009; Sutcliffe, 2015).

The use of Genome sequence data in prokaryotic taxonomy, as seen in the studies described in Chapters 2, 3, 4, 5, and 7 of this thesis, has several promising advantages over genetic and traditional polyphasic taxonomy, and provides a sufficient basis to build a robust and reliable taxonomic framework for most prokaryotes (Chun & Rainey, 2014; Rossello-Mora & Amann, 2015; Sutcliffe, 2015; Whitman, 2015b). Firstly, the taxonomic thresholds established for measures of genomic distance, including those discussed in Chapter 1 of this thesis, provide a comprehensive representation of the average rate of divergence between two organisms. Taxonomic thresholds based on genetic distance, such as those based on the 16S rRNA gene, reflect the rate of divergence of a single gene, which may be under different evolutionary pressures than the remainder of the genome. Secondly, genome based taxonomic inferences can be informed by reliable and robust phylogenetic trees based on the entire shared core genome of a group, rather than phylogenetic trees based on a single gene. Phylogenomic trees utilized in genome based taxonomy can be produced using fast, simple, and automated tools such as the GLIMPS pipeline, discussed in Chapter 6 of this thesis. Furthermore, genome sequence data can be used to predict metabolic, physiological, and biochemical capabilities of an organism; largely eliminating the need for traditional biochemical and chemotaxonomic assays (Sutcliffe et al., 2013; Thompson et al., 2015). Lastly, the application of comparative genomic analysis techniques to genome sequence data enables the identification of rare genomic changes useful in characterizing related groups of organisms (Rokas & Holland, 2000; Rokas et al., 2003; Delsuc et al., 2005). These rare genomic changes include discrete genetic events which can be readily identified from genomic sequence data such as gene rearrangements, gene fusions and fissions, gene duplication, and, most importantly for prokaryotic taxonomy, the occurrence of insertions and deletions in amino acid sequences (CSIs) such as those described in Chapters 2, 3, 4, 5, and 7 of this thesis.

The Utilization of Molecular Signatures in Phylogeny and Taxonomy

The phylum Spirochaetes and the class *Betaproteobacteria* are large groups of diverse bacteria, classified primarily on the basis of 16S rRNA gene analysis. Until recently, the phylum Spirochaetes was comprised of a single class, *Spirochaetia*, containing a single order, *Spirochaetales*, which was made up of four families (Paster, 2011). In my work we have identified 38 CSIs which are specific for either all members of the phylum Spirochaetes or its different main clades. The relationships between the members of the phylum Spirocheates suggested by the identified CSIs are strongly supported by neighbour-joining and maximum-likelihood phylogenetic trees, based upon the concatenated sequences of 22 conserved proteins. On the basis of these findings, we have proposed that the four families within the phylum Spirocheates should be elevated to the order level taxonomic ranks (viz. *Spirochaetales*, *Brevinematales*, *Brachyspiriales*, and *Leptospiriales*) and that the genera *Borrelia* and *Cristispira* be transferred to a new family *Borreliaceae* within the order *Spirochaetales* (Gupta et al., 2013b). Additionally, we have identified 53 CSIs and 25 CSPs which distinguish the two groups of clinically distinct organisms within the genus *Borrelia*, the Lyme disease related *Borrelia* and the relapsing fever related *Borrelia*. The distinctiveness of the two groups of *Borrelia* is supported by average nucleotide identity analysis and phylogenetic analysis based upon the 16S rRNA gene and the concatenated sequences for 25 conserved proteins. On the basis of these results, we have proposed a division of the genus *Borrelia* into two genera, limiting the genus *Borrelia* to only the members of the relapsing fever *Borrelia* group, and transferring the members of the Lyme disease *Borrelia* group to the genus *Borreliella* (Adeolu & Gupta, 2014).

Within the class *Betaproteobacteria*, we have examined the phylogeny of the order *Neisseriales*, a group containing the causative agent of the increasingly drug resistant sexually transmitted infection gonorrhea and a number of other highly prevalent pathogenic and environmental bacteria classified as a single family (Stephens et al., 2007; Cohn et al., 2010; World Health Organization, 2011). In my work, we have identified 54 CSIs in widely distributed proteins that are specific for either all of the *Neisseriales*, or which differentiate its subgroups. Importantly, the identified CSIs were able to distinguish a group of obligate host-associated *Neisseriales*, containing the important pathogens in the order, from all other members of the order *Neisseriales*. This distinction is also supported by 16S rRNA and concatenated protein based phylogenetic trees. Additionally, the association of many of the identified CSIs with the obligate host-associated organisms in the order suggests that the CSIs may play a functional role in the evolution of obligate host-association within this order. On the basis of these findings, we have proposed a taxonomic revision limiting the family *Neisseriaceae* to the obligate host-associated members of the order *Neisseriales* and transfering the other genera within the order *Neisseriales* to the novel family *Chromobacteriaceae* (Adeolu & Gupta, 2013). We have also examined the phylogeny of the genus *Burkholderia*, a group of over 70 species of soil bacteria which are ubiquitous in the environment and have varying pathogenic potential (White, 2003; Workowski et al., 2008; Lipuma, 2010). My work on the genus *Burkholderia* has led to the identification of 42 highly specific CSIs that delineate a number of well-defined groups of *Burkholderia*. Importantly, six of these CSIs are specific for a group of *Burkholderia* containing all clinically relevant members of the genus. Within clinically relevant groups we have also identified multiple CSIs that serve to clearly demarcate the *B. cepacia* complex, the *B. pseudomallei* group, and the phytopathogenic *Burkholderia*. A division between the clinically relevant members of the genus *Burkholderia* and the plant-beneficial and environmental *Burkholderia* is also observed in phylogenetic trees based upon concatenated sequences for 21 conserved proteins and the 16S rRNA gene. Based upon the identified CSIs, the pathogenicity profile of *Burkholderia* species, and phylogenetic analyses, we proposed that the genus *Burkholderia* should be limited to the clinically relevant group within the genus and that the plant-beneficial and environmental *Burkholderia* should be transferred to the novel genus *Paraburkholderia* (Sawana et al., 2014). In addition to the groups described in this thesis, I have also been involved in published evolutionary and systematic studies of the phylum Chlamydiae (Gupta et al., 2015b), the class *Coriobacteriia* (Gupta et al., 2013a), the class *Negativicutes* (Campbell et al., 2015), the class

Halobacteria (Gupta et al., 2016), the order *Xanthomonadales* (Naushad et al., 2015b), the order *Bifidobacteriales* (Zhang et al., 2016), and the family *Pasteurellaceae* (Naushad et al., 2015a).

In each of these cases, molecular signatures provide a novel and powerful means for the unambiguous delineation of distinct monophyletic evolutionary linages, and provide support for elevated taxonomic status. Additionally, phylogenetic inferences derived from CSIs and CSPs are independent of gene or genome based phylogenetic trees, and are generally robust against long-branch attraction, compositional biases, differences in evolutionary rates, lateral gene transfers, and other artifacts in the construction of phylogenetic trees (Delsuc et al., 2005; Gupta, 2014). Evolutionarily informative CSIs also have an extremely reliable specificity for a given group of organisms. Notably, many CSIs were first identified when genome sequences were available for less than 100 species (Gupta, 1998; Gupta, 2001; Gupta & Griffiths, 2002). However, despite the availability of over 50 000 sequenced genomes today, these markers have retained their specificity for the indicated groups and are found in other, newly sequenced members of the indicated groups, providing evidence of their predictive ability (Bhandari et al., 2012; Gupta, 2014; Gupta, 2016). The CSIs and CSPs described here are predicted to have similar specificity and reliability for members of their group as the availability of sequence information continues to grow. The long-term specificity and reliability of similar CSIs and CSPs has facilitated their use in taxonomic revisions and descriptions of prokaryotic groups ranging from species to phylum level taxa (Bhandari et al., 2013; Gupta & Lali, 2013; Naushad & Gupta, 2013; Bhandari & Gupta, 2014; Howard-Azzeh et al., 2014; Naushad et al., 2014; Gupta et al., 2015a). Thus, the CSIs and CSPs described here also represent novel tools for the taxonomic placement of new members of these groups as they are discovered.

Phylogenomics and the path forward

The core strength of genome based systematic studies lies in the scale of data brought to bear in resolving phylogenetic relationships. In Chapter 7 of this thesis, I described an example of the use of genome scale data to resolve the phylogenetic relationships among the members of the order *Enterobacteriales*. The order *Enterobacteriales* is a large and diverse group of non-spore-forming rods within the class *Gammaproteobacteria*. The taxonomy of the order *Enterobacteriales* is based, primarily, on the 16S rRNA gene (Hauben et al., 1998; Spröer et al., 1999; Francino et al., 2006; Naum et al., 2008). However, the 16S rRNA gene has low discriminatory power and interrelationships of the members of the order *Enterobacteriales* are poorly resolved in 16S rRNA gene based phylogenetic trees (Hauben et al., 1998; Naum et al., 2008; Octavia & Lan, 2014). Consequently, the >250 species within the order *Enterobacteriales* are all placed into a single family.

We have identified 69 CSIs, in widely distributed proteins, which are unique characteristics of seven different groups within the order

Enterobacteriales. Independent of the identification of CSIs, we have also employed the GLIMPS pipeline, detailed in Chapter 6, to construct a highly robust phylogenetic tree based on 1548 shared core proteins from the whole genome sequences of 179 representative genome sequenced members of the order *Enterobacteriales*, as well as phylogenetic trees based on 53 ribosomal proteins and 4 MLSA proteins. Unlike phylogenetic trees based on the 16S rRNA gene, each of these phylogenetic trees supports the presence of the seven main groups suggested by the identified CSIs. Additionally, the proportion of shared protein families in the analyzed genomes (POCP), one of the measures of genomic distance discussed in Chapter 1, also supports the presence of seven main groups within the order *Enterobacteriales*. On the basis of these analyses, we are proposing a division of the order *Enterobacteriales* into seven families.

The limited ability of the 16S rRNA gene to resolve phylogenetic relationships within the order *Enterobacteriales* has been a long-standing issue in bacterial classification (Hauben et al., 1998; Naum et al., 2008; Octavia & Lan, 2014). We were able to employ independent means of analyzing genomic sequence data (viz. a supermatrix based phylogenomic tree, concatenated universal protein based phylogenetic trees, a measure of genomic distance based on shared protein families, and rare genomic changes) to show that the order *Enterobacteriales* possesses a robust and discernable phylogenetic structure. This study represents a powerful example of the strengths of genome based taxonomy. As genome based systematic research becomes increasingly prevalent, we expect evolutionary and systematic studies to utilize similar multipronged approaches to genomic sequence analysis and for systematic studies, such as the one described in Chapter 7, to become the overarching basis for prokaryotic classification and taxonomy.

Future Directions

Genome sequence based evolution and systematics research is paving the way for future biological classifications and taxonomic frameworks (Chun & Rainey, 2014; Rossello-Mora & Amann, 2015; Sutcliffe, 2015; Whitman, 2015b). In this thesis, I have described the application of molecular signatures and phylogenomic techniques to the identification, differentiation, and classification of several distinct prokaryotic groups. The systematic studies presented here serve as exemplars for the utility of genomic sequence analysis in prokaryotic taxonomy. Further, I have described the GLIMPS phylogenomic analysis pipeline, an integrated software pipeline that produces supermatrix based phylogenomic trees and calculates genomic distance using multiple methodologies. We have made the GLIMPS pipeline freely available and easy to use for the wider research community. Recently, we have also made several tools for the identification of CSIs available on the Gleans.net website (Gupta, 2014). We hope that the availability of these tools will enable more researchers to attempt genome sequence based evolutionary research, and to identify novel and informative molecular signatures.

The use of the CSIs and CSPs described in this thesis is not limited to evolutionary and systematic studies. CSIs and CSPs possess a number of attractive attributes which make them ideal candidates for diagnostic probes. Firstly, due to the high level of sequence conservation within CSIs and CSPs, degenerate oligonucleotide PCR primers can be readily designed to specifically and reliably amplify CSI- or CSP-containing regions of DNA. CSIs and CSPs have also been shown to possess extremely reliable specificity for their given group of organisms as more genomes are sequenced (Gupta, 2016). Thus, their detection provides unambiguous evidence for the presence of a member of the group for which they are specific (Griffiths et al., 2005; Gupta & Griffiths, 2006). Consequently, highly robust diagnostic assays, based on CSIs found in certain proteins, have been developed for distinguishing strains of *Bacillus anthracis* from those of *Bacillus cereus* species/strains, and for the identification of enterohemorrhagic *E. coli* O157:H7 from other *E. coli* strains; bacterial strains which are not reliably distinguished from each other by most other means (Ahmod et al., 2011; Wong et al., 2014). Similarly, the sequences of CSIs and CSPs can be used to identify organisms based solely on genomic or metagenomic sequence data (Segata et al., 2012; Gupta & Sharma, 2015; Truong et al., 2015). Thus, the CSIs and CSPs described in this thesis have applications as novel diagnostic genomic markers.

The CSIs and CSPs described in this thesis also represent novel targets for functional studies. Prior work by our laboratory has shown that CSIs are essential for the proper function of the protein in the groups of bacteria in which they are found, and that their removal or any substantial changes in their sequences leads to a cessation of the cellular function of that protein (Singh & Gupta, 2009). Additionally, structural analyses of CSIs indicates that they are located within surface loops of the proteins, away from the active site (Hsing & Cherkasov, 2008; Singh & Gupta, 2009; Gupta & Khadka, 2015). Surface loops are highly accessible regions of the protein that are indicated to play important roles in protein-protein and protein-ligand interactions, which may be modified or modulated by the presence of the identified CSIs (Akiva et al., 2008; Hashimoto & Panchenko, 2010). While the functional role of many of the CSPs described here is currently unknown, their presence in all members of a group of organisms suggests that they likely have an adaptive function, protecting them from the effects of purifying selection. Further analyses of the CSIs and CSPs described here has the potential to lead to the discovery of novel functions in these organisms, mediated by CSIs and CSPs, which may provide important insights into the physiology, evolution, and adaptations of these groups.

Concluding Remarks

The increasing availability of genomic sequence data is providing researchers with an unparalleled wealth of information from which we can elucidate the evolutionary relationships of living organisms. The use of this data is revolutionizing the fields of biological classification and taxonomy. Importantly, this wealth of genome sequence data is enabling the detection of conserved molecular signatures, such as CSIs and CSPs, which are shared by evolutionarily related groups of organisms. Using these molecular signatures, it is possible to infer phylogenetic relationships, independent of gene and genome based phylogenetic trees. Thus, molecular signatures are powerful new tools for evolutionary studies. Additionally, these molecular signatures represent novel diagnostic markers for their specified group and further analyses of these molecular signatures should lead to the discovery of novel functions and biological characteristics, mediated by CSIs and CSPs, which will provide important insights into the physiology, evolution, and adaptations of these groups.

BIBLIOGRAPHY

Adams, D. A., Gallagher, K. M., Jajosky, R. A., Kriseman, J., Sharp, P., Anderson, W. J., Aranas, A. E., Mayes, M., Wodajo, M. S., Onweh, D. H., et al. (2013). Summary of Notifiable Diseases - United States, 2011. *MMWR Morb Mortal Wkly Rep, 60*(53), 1-117.

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

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

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

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*(36), 13292.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol, 55*(4), 539-552.

Beiko, R. G., Harlow, T. J., & Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A, 102*(40), 14332-14337.

Bellgard, M. I., Wanchanthuek, P., La, T., Ryan, K., Moolhuijzen, P., Albertyn, Z., Shaban, B., Motro, Y., Dunn, D. S., Schibeci, D., et al. (2009). Genome sequence of the pathogenic intestinal spirochete brachyspira hyodysenteriae reveals adaptations to its lifestyle in the porcine large intestine. *PLoS One, 4*(3), e4641.

Bergey, D., Breed, R., Hammer, B., Huntoon, F., Murray, E. G. D., & Harrison, F. C. (1934). *Bergey's Manual of Determinative Bacteriology* (4th ed.). Baltimore, USA: Williams and Wilkins.

Bergey, D., Harrison, F., Breed, R., Hammer, B., & Huntoon, F. (1923). *Bergey's Manual of Determinative Bacteriology* Baltimore, USA: Williams and Wilkins.

Bhandari, V., Ahmod, N. Z., Shah, H. N., & Gupta, R. S. (2013). Molecular signatures for *Bacillus* species: demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. *Int J Syst Evol Microbiol, 63*(7), 2712-2726.

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

Bhandari, V., Naushad, H. S., & Gupta, R. S. (2012). Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. *Frontiers in cellular and infection microbiology, 2*.

Bininda-Emonds, O. R. P. (2004). *Phylogenetic supertrees: combining information to reveal the tree of life* (Vol. 4): Springer Science & Business Media.

Bonham-Carter, O., Steele, J., & Bastola, D. (2014). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics, 15*(6), 890-905.

Bottacini, F., Medini, D., Pavesi, A., Turroni, F., Foroni, E., Riley, D., Giubellini, V., Tettelin, H., van Sinderen, D., & Ventura, M. (2010). Comparative genomics of the genus *Bifidobacterium*. *Microbiol-Sgm, 156*, 3243-3254.

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M., & Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J Bacteriol, 186*(12), 3980-3990.

Brady, C., Cleenwerck, I., Venter, S., Coutinho, T., & De Vos, P. (2013). Taxonomic evaluation of the genus *Enterobacter* based on multilocus sequence analysis (MLSA): proposal to reclassify *E. nimipressuralis* and *E. amnigenus* into *Lelliottia* gen. nov. as *Lelliottia nimipressuralis* comb. nov. and *Lelliottia amnigena* comb. nov., respectively, *E. gergoviae* and *E. pyrinus* into *Pluralibacter* gen. nov. as *Pluralibacter gergoviae* comb. nov. and *Pluralibacter pyrinus* comb. nov., respectively, *E. cowanii, E. radicincitans, E. oryzae* and *E. arachidis* into *Kosakonia* gen. nov. as *Kosakonia cowanii* comb. nov., *Kosakonia radicincitans* comb. nov., *Kosakonia oryzae* comb. nov. and *Kosakonia arachidis* comb. nov., respectively, and *E. turicensis*, *E. helveticus* and *E. pulveris* into *Cronobacter* as *Cronobacter zurichensis* nom. nov., *Cronobacter helveticus* comb. nov. and *Cronobacter pulveris* comb. nov., respectively, and emended description of the genera *Enterobacter* and *Cronobacter*. *Syst Appl Microbiol, 36*(5), 309-319.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet, 28*(3), 281-285.

Bryson, B. (2003). *A short history of nearly everything*: DC Books.

Campbell, C., Adeolu, M., & Gupta, R. S. (2015). A Genome Based Taxonomic Framework for the class Negativicutes: Division of the class Negativicutes into the orders Selenomonadales, Acidaminococcales ord. nov., and Veillonellales ord. nov. *Int J Syst Evol Microbiol, 65*(9), 3203-3215.

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972-1973.

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

Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M., & Ragan, M. A. (2014). Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep, 4*.

Chan, R. H., Chan, T. H., Yeung, H. M., & Wang, R. W. (2012). Composition vector method based on maximum entropy principle for sequence comparison. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 9*(1), 79-87.

Chang, J.-M., Di Tommaso, P., & Notredame, C. (2014). TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol*, msu117.

Chun, J., & Rainey, F. A. (2014). Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int J Syst Evol Microbiol, 64*(Pt 2), 316-324.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science, 311*(5765), 1283-1287.

Coenye, T., & Vandamme, P. (2003). Diversity and significance of Burkholderia species occupying diverse ecological niches. *Environ Microbiol, 5*(9), 719-729.

Cohen, S. S. (1948). The synthesis of bacterial viruses i. The synthesis of nucleic acid and protein in *Escherichia coli* B infected with T2r+ bacteriophage. *J Biol Chem, 174*(1), 281-293.

Cohn, A. C., MacNeil, J. R., Harrison, L. H., Hatcher, C., Theodore, J., Schmidt, M., Pondo, T., Arnold, K. E., Baumbach, J., Bennett, N., et al. (2010). Changes in *Neisseria meningitidis* disease epidemiology in the United States, 1998–2007: implications for prevention of meningococcal disease. *Clin Infect Dis, 50*(2), 184-191.

Cohn, F. (1872). Grundzüge einer neuen natürlichen Anordnung der kryptogamischen Pflanzen. *Jahresb Schles Ges Vaterl Kultur, 49*, 83-89.

Cohn, F. (1875). *Untersuchungen über Bacterien: I*: JU Kern.

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

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563.

Cummins, C. A., & McInerney, J. O. (2011). A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst Biol, 60*(6), 833-844.

Cutler, S. J. (2010). Relapsing fever: a forgotten disease revealed. *J Appl Microbiol, 108*(4), 1115-1122.

Dagan, T., & Martin, W. (2006). The tree of one percent. *Genome Biol, 7*(10), 118.

Deloger, M., El Karoui, M., & Petit, M. A. (2009). A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol, 191*(1), 91-99.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics, 6*(5), 361-375.

den Bakker, H. C., Cummings, C. A., Ferreira, V., Vatta, P., Orsi, R. H., Degoricija, L., Barker, M., Petrauskene, O., Furtado, M. R., & Wiedmann, M. (2010). Comparative genomics of the bacterial genus *Listeria*: Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics, 11*, 688.

Dias, U., Dias, Z., & Setubal, J. C. (2011). *Two new whole-genome distance measures*. Paper presented at the Proceedings of the 6th Brazilian Symposium on Bioinformatics (BSB'2011).

Dunn, C. W., Howison, M., & Zapata, F. (2013). Agalma: an automated phylogenomics workflow. *BMC Bioinformatics, 14*, 330.

Dutilh, B. E., Huynen, M. A., Bruno, W. J., & Snel, B. (2004). The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J Mol Evol, 58*(5), 527-539.

Dutilh, B. E., Snel, B., Ettema, T. J. G., & Huynen, M. A. (2008). Signature genes as a phylogenomic tool. *Mol Biol Evol, 25*(8), 1659.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755-763.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol, 7*(10), e1002195.

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

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

Efron, B. (1992). *Bootstrap methods: another look at the jackknife*: Springer.

Fang, G., Rocha, E. P. C., & Danchin, A. (2008). Persistence drives gene clustering in bacterial genomes. *BMC Genomics, 9*(1), 4.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol, 17*(6), 368-376.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Biol*, *19*(2), 99-113.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Biol*, *20*(4), 406-416.

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

Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S., & Woese, C. R. (1977a). Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences, 74*(10), 4537-4541.

Fox, G. E., Pechman, K. R., & Woese, C. R. (1977b). Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int J Syst Evol Microbiol*, *27*(1), 44-57.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., et al. (1980). The phylogeny of prokaryotes. *Science*, *209*(4455), 457-463.

Fox, G. E., Wisotzkey, J. D., & Jurtshuk, P. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol*, *42*(1), 166-170.

Francino, M. P., Santos, S. R., & Ochman, H. (2006). Phylogenetic relationships of bacteria with special reference to endosymbionts and enteric species The Prokaryotes (pp. 41-59): Springer.

Fred, E., & Wilson, P. (1934). On photosynthesis and free nitrogen assimilation by leguminous plants. *Proceedings of the National Academy of Sciences, 20*(7), 403-409.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics, 28*(23), 3150-3152.

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J Exp Zool Part B, 304B*(1), 64-74.

Gao, B., & Gupta, R. S. (2012a). Microbial systematics in the post-genomics era. *Anton Leeuw Int J G, 101*(1), 45-54.

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

Gao, B., Paramanathan, R., & Gupta, R. S. (2006). Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Anton Leeuw Int J G, 90*(1), 69-91.

Gao, L., Qi, J., Sun, J., & Hao, B. (2007). Prokaryote phylogeny meets taxonomy: an exhaustive comparison of composition vector trees with systematic bacteriology. *Science in China Series C: Life Sciences, 50*(5), 587-599.

Garrity, G. M., Bell, J. A., & Lilburn, T. (2005). The revised road map to the manual. In G. M. Garrity, D. J. Brenner, N. R. Krieg & J. T. Staley (Eds.), Bergey’s Manual® of Systematic Bacteriology (Vol. 2, pp. 159-187). New York: Springer.

Garrity, G. M., Winters, M., Kuo, A. W., & Searles, D. B. (2001). *Taxonomic Outline of the Procaryotes Bergey's Manual of Systematic Bacteriology*: New York: Springer-Verlag.

Gerbase, A. C., Rowley, J. T., Heymann, D. H. L., Berkley, S. F. B., & Piot, P. (1998). Global prevalence and incidence estimates of selected curable STDs. *Sex Trans Inf*, *74*(1), S12.

Gest, H., & Kamen, M. D. (1948). Studies on the phosphorus metabolism of green algae and purple bacteria in relation to photosynthesis. *J Biol Chem, 176*(1), 299-318.

Gest, H., Kamen, M. D., & Bregoff, H. M. (1950). Studies on the Metabolism of Photosynthetic Bacteria V. Photoproduction of Hydrogen and Nitrogen Fixation by *Rhodospirillum Rubrum*. *J Biol Chem, 182*(1), 153-170.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., & Thompson, F. L. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology, 3*(9), 733-739.

Glaeser, S. P., & Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst Appl Microbiol, 38*(4), 237-245.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol, 57*(1), 81-91.

Grant, J. R., & Katz, L. A. (2014). Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. *PLoS currents, 6*.

Greisen, K., Loeffelholz, M., Purohit, A., & Leong, D. (1994). PCR primers and probes for the 16S rRNA gene of most species of pathogenic bacteria, including bacteria found in cerebrospinal fluid. *J Clin Microbiol, 32*(2), 335-351.

Griffiths, E., Petrich, A. K., & Gupta, R. S. (2005). Conserved indels in essential proteins that are distinctive characteristics of *Chlamydiales* and provide novel means for their identification. *Microbiology, 151*(8), 2647-2657.

Grimont, P. A., Popoff, M. Y., Grimont, F., Coynault, C., & Lemelin, M. (1980). Reproducibility and correlation study of three deoxyribonucleic acid hybridization procedures. *Curr Microbiol, 4*(6), 325-330.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol, 59*(3), 307-321.

Gupta, A., & Sharma, V. K. (2015). Using the taxon-specific genes for the taxonomic classification of bacterial genomes. *BMC Genomics, 16*(1), 396.

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

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

Gupta, R. S. (2010). Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. *Photosynthesis Res, 104*(2), 357-372.

Gupta, R. S. (2014). Identification of conserved indels that are useful for classification and evolutionary studies Methods in Microbiology (Vol. 41). Oxford, UK: Academic Press.

Gupta, R. S. (2016). Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. *FEMS Microbiol Rev*.

Gupta, R. S., Chen, W. J., Adeolu, M., & Chai, Y. (2013a). Molecular signatures for the class *Coriobacteriia* and its different clades; proposal for division of the class *Coriobacteriia* into the emended order *Coriobacteriales,* containing the emended family *Coriobacteriaceae* and *Atopobiaceae* fam. nov., and *Eggerthellales* ord. nov., containing the family *Eggerthellaceae* fam. nov. *Int J Syst Evol Microbiol, 63*(Pt 9), 3379-3397.

Gupta, R. S., & Gao, B. (2010). Recent advances in understanding microbial systematics. In J. Xu (Ed.), Microbial population genetics (pp. 1-14). Norfolk, United Kingdom: Caister Academic Press.

Gupta, R. S., & Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theor Popul Biol, 61*(4), 423-434.

Gupta, R. S., & Griffiths, E. (2006). *Chlamydiae*-specific proteins and indels: novel tools for studies. *Trends Microbiol, 14*(12), 527-535.

Gupta, R. S., & Khadka, B. (2015). Evidence for the presence of key chlorophyll-biosynthesis-related proteins in the genus Rubrobacter (Phylum Actinobacteria) and its implications for the evolution and origin of photosynthesis. *Photosynth Res*.

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

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

Gupta, R. S., Naushad, S., & Baker, S. (2015a). Phylogenomic analyses and molecular signatures for the class Halobacteria and its two major clades: a proposal for division of the class Halobacteria into an emended order Halobacteriales and two new orders, Haloferacales ord. nov. and Natrialbales ord. nov., containing the novel families Haloferacaceae fam. nov. and Natrialbaceae fam. nov. *Int J Syst Evol Microbiol, 65*(Pt 3), 1050-1069.

Gupta, R. S., Naushad, S., Chokshi, C., Griffiths, E., & Adeolu, M. (2015b). A phylogenomic and molecular markers based analysis of the phylum Chlamydiae: proposal to divide the class *Chlamydiia* into two orders, *Chlamydiales* and *Parachlamydiales* ord. nov., and emended description of the class *Chlamydiia*. *Antonie Van Leeuwenhoek, 108*(3), 765-781.

Gupta, R. S., Naushad, S., Fabros, R., & Adeolu, M. (2016). A phylogenomic reappraisal of family-level divisions within the class *Halobacteria*: proposal to divide the order *Halobacteriales* into the families *Halobacteriaceae*, *Haloarculaceae* fam. nov., and *Halococcaceae* fam. nov., and the order *Haloferacales* into the families, *Haloferacaceae* and *Halorubraceae* fam nov. *Antonie Van Leeuwenhoek*, 1-23.

Hashimoto, K., & Panchenko, A. R. (2010). Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc Natl Acad Sci U S A, 107*(47), 20352-20357.

Hauben, L., Moore, E. R., Vauterin, L., Steenackers, M., Mergaert, J., Verdonck, L., & Swings, J. (1998). Phylogenetic position of phytopathogens within the *Enterobacteriaceae*. *Syst Appl Microbiol, 21*(3), 384-397.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., & Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics, 21*(10), 2329-2335.

Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology, 36*(1), 39-56.

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

Hsing, M., & Cherkasov, A. (2008). Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinformatics, 9*, 293.

Huddleson, I. (1947). Proceedings of 4th International Congress on Microbiology.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., & Ise, K. (2016). A new view of the tree of life. *Nature Microbiology, 1*, 16048.

Huss, V. A., Festl, H., & Schleifer, K. H. (1983). Studies on the spectrophotometric determination of DNA hybridization from renaturation rates. *Syst Appl Microbiol, 4*(2), 184-192.

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol, 45*(9), 2761-2764.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., & Cody, A. J. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology, 158*(4), 1005-1015.

Jolley, K. A., Chan, M.-S., & Maiden, M. C. (2004). mlstdbNet–distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics, 5*(1), 86.

Jolley, K. A., & Maiden, M. C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics, 11*(1), 595.

Jones, A. L. (2012). The Future of Taxonomy. In G. M. Gadd & S. Sariaslani (Eds.), Adv Appl Microbiol (1 ed., Vol. 80, pp. 23-35). San Diego: Academic Press Inc.

Jun, S.-R., Sims, G. E., Wu, G. A., & Kim, S.-H. (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proceedings of the National Academy of Sciences, 107*(1), 133-138.

Kämpfer, P. (2012). Systematics of prokaryotes: the state of the art. *Anton Leeuw Int J G, 101*(1), 3-11.

Kämpfer, P., & Glaeser, S. P. (2013). Prokaryote characterization and identification The Prokaryotes (pp. 123-147): Springer.

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

Katoh, K., & Toh, H. (2007). PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics, 23*(3), 372-374.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664.

Kim, M., Oh, H. S., Park, S. C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol, 64*(Pt 2), 346-351.

Klappenbach, J. A., Saxman, P. R., Cole, J. R., & Schmidt, T. M. (2001). rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res, 29*(1), 181-184.

Konstantinidis, K. T., & Tiedje, J. M. (2005). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol, 187*(18), 6258-6264.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., & Koonin, E. V. (2011). Computational methods for Gene Orthology inference. *Briefings in bioinformatics*, *12*(5), 379-391.

Kumar, S., Krabberod, A. K., Neumann, R. S., Michalickova, K., Zhao, S., Zhang, X., & Shalchian-Tabrizi, K. (2015). BIR Pipeline for Preparation of Phylogenomic Data. *Evol Bioinform Online, 11*, 79-83.

Kuo, C. H., & Ochman, H. (2009). The fate of new bacterial genes. *FEMS Microbiol Rev, 33*(1), 38-43.

Lang, J. M., Darling, A. E., & Eisen, J. A. (2013). Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS One, 8*(4), e62510.

Lapage, S., Clark, W., Lessel, E., Seeliger, H., & Sneath, P. (1973). Proposed revision of the International Code of Nomenclature of Bacteria. *Int J Syst Evol Microbiol, 23*(1), 83-108.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics, 12*(1), 1.

Lehman, K., & Neumann, R. (1896). Atlas und grundriss der bakeriologie und lehrbuch der speziellen bakteriologischen diagnositk. *Teil II, München: Lehmann*.

Lerat, E., Daubin, V., Ochman, H., & Moran, N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol, 3*(5), e130.

Lessel, E. F. (1971). Judicial Commission of the International Committee on Nomenclature of Bacteria. *Int J Syst Evol Microbiol, 21*(1), 100-103.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*(13), 1658-1659.

Lindgren, E., & Jaenson, T. G. (2006). Lyme borreliosis in Europe: influences of climate and climate change, epidemiology, ecology and adaptation measures. *WHO Regional Office for Europe Copenhagen*.

Linnaeus, C. v. (1758). Systema Naturae, edition X, vol. 1 (Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Tomus I. Editio decima, reformata). *Salvii, Holmiae*, 1.

Lipuma, J. J. (2010). The Changing Microbial Epidemiology in Cystic Fibrosis. *Clin Microbiol Rev, 23*(2), 299.

Liu, K., Linder, C. R., & Warnow, T. (2010a). Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Currents Tree of Life*.

Liu, K., Linder, C. R., & Warnow, T. (2011). RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One, 6*(11), e27731.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International, 2012*.

Liu, Y., Schmidt, B., & Maskell, D. L. (2010b). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics, 26*(16), 1958-1964.

Liu, Z., Meng, J., & Sun, X. (2008). A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping. *Biochem Biophys Res Commun, 368*(2), 223-230.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., & Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology, 10*(9), 599-606.

Löytynoja, A., & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science, 320*(5883), 1632-1635.

Lu, G., Zhang, S., & Fang, X. (2008). An improved string composition method for sequence comparison. *BMC Bioinformatics, 9*(Suppl 6), S15.

Ludwig, W., & Klenk, H. P. (2001). Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systematics. In G. M. Garrity (Ed.), (2 ed., Vol. 1, pp. 49-65). New York: Springer.

Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M., & Schleifer, K. H. (1998). Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis, 19*(4), 554-568.

Maiden, M. C. (2006). Multilocus sequence typing of bacteria. *Annu Rev Microbiol, 60*, 561-588.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., & Caugant, D. A. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences, 95*(6), 3140-3145.

Maiden, M. C., van Rensburg, M. J. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., & McCarthy, N. D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology, 11*(10), 728-736.

Marchesi, J. R., Sato, T., Weightman, A. J., Martin, T. A., Fry, J. C., Hiom, S. J., & Wade, W. G. (1998). Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Appl Environ Microbiol, 64*(2), 795-799.

McCarthy, B., & Bolton, E. (1963). An approach to the measurement of genetic relatedness among organisms. *Proceedings of the National Academy of Sciences, 50*(1), 156-164.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics, 14*(1), 1.

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Res, 15*(12), 1767-1776.

Mignard, S., & Flandrois, J. P. (2006). 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *J Microbiol Methods*, *67*(3), 574-581.

Narra, H. P., Cordes, M. H. J., & Ochman, H. (2008). Structural features and the persistence of acquired proteins. *Proteomics*, *8*(22), 4772-4781.

Naum, M., Brown, E. W., & Mason-Gamer, R. J. (2008). Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the *enterobacteriaceae*? *J Mol Evol, 66*(6), 630-642.

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

Naushad, H. S., Lee, B., & Gupta, R. S. (2014). Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*. *Int J Syst Evol Microbiol, 64*(2), 366-383.

Naushad, S., Adeolu, M., Goel, N., Khadka, B., Al-Dahwi, A., & Gupta, R. S. (2015a). Phylogenomic and Molecular Demarcation of the Core Members of the Polyphyletic *Pasteurellaceae* genera *Actinobacillus*, *Haemophilus*, and *Pasteurella*. *International journal of genomics, 2015*, 198560.

Naushad, S., Adeolu, M., Wong, S., Sohail, M., Schellhorn, H. E., & Gupta, R. S. (2015b). A phylogenomic and molecular marker based taxonomic framework for the order *Xanthomonadales*: proposal to transfer the families *Algiphilaceae* and *Solimonadaceae* to the order *Nevskiales* ord. nov. and to create a new family within the order *Xanthomonadales*, the family *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. *Antonie Van Leeuwenhoek, 107*(2), 467-485.

NCBI. (2016). NCBI Genome Database. http://www.ncbi.nlm.nih.gov/genome/

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol, 302*(1), 205-217.

Octavia, S., & Lan, R. (2014). The Family *Enterobacteriaceae*. *The Prokaryotes: gammaproteobacteria*, 225-286.

Ogden, T. H., & Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol, 55*(2), 314-328.

Olsen, G., Pace, N., Nuell, M., Kaine, B., Gupta, R., & Woese, C. (1985). Sequence of the 16S rRNA gene from the thermoacidophilic archaebacterium *Sulfolobus solfataricus* and its evolutionary implications. *J Mol Evol, 22*(4), 301-307.

Olsen, G. J., Woese, C. R., & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol, 176*(1), 1.

Oren, A., & Garrity, G. M. (2014). Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie Van Leeuwenhoek, 106*(1), 43-56.

Orla-Jensen, S. (1909). Die Hauptlinien des natürlichen Bakteriensystems. *Zentralbl Bakteriol Parasitenkd Infektionskr Hyg Abt, 2*(22), 305-346.

Otu, H. H., & Sayood, K. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics, 19*(16), 2122-2130.

Palleroni, N. J. (2005). Genus I. *Burkholderia* Yabuuchi et al. 1993, 398^{VP} (Effective publication: Yabuuchi et al. 1992, 1268) emend. Gillis et al. 1995, 286*. In D. J. Brenner, N. R. Krieg, G. M. Garrity & J. T. Staley (Eds.), Bergey's Manual of Systematic Bacteriology (2 ed., Vol. 2, pp. 575-600). New York: Springer.

Parte, A. C. (2014). LPSN--list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res, 42*(Database issue), D613-616.

Paster, B. J. (2011). Phylum XV. Spirochaetes Garrity and Holt 2001. In D. J. Brenner, N. R. Krieg, G. M. Garrity & J. T. Staley (Eds.), Bergey's Manual of Systematic Bacteriology (2 ed., Vol. 3, pp. 471-471). New York: Springer.

Patel, J. B. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Mol Diagn, 6*(4), 313-322.

Pearse, W. D., & Purvis, A. (2013). phyloGenerator: an automated phylogeny generation tool for ecologists. *Methods Ecol Evol, 4*(7), 692-698.

Peeters, C., Zlosnik, J. E., Spilker, T., Hird, T. J., LiPuma, J. J., & Vandamme, P. (2013). *Burkholderia pseudomultivorans* sp. nov., a novel *Burkholderia cepacia* complex species from human respiratory samples and the rhizosphere. *Syst Appl Microbiol, 36*(7), 483-489.

Postic, D., Garnier, M., & Baranton, G. (2007). Multilocus sequence analysis of atypical *Borrelia burgdorferi* sensu lato isolates–description of *Borrelia californiensis* sp. nov., and genomospecies 1 and 2. *Int J Med Microbiol, 297*(4), 263-271.

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

Pringsheim, E. (1923). Zur kritik der Bakteriensystematik. *Lotos, 71*, 357-377.

Puigbo, P., Wolf, Y. I., & Koonin, E. V. (2009). Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol, 8*(6), 59.

Qin, Q. L., Xie, B. B., Zhang, X. Y., Chen, X. L., Zhou, B. C., Zhou, J., Oren, A., & Zhang, Y. Z. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol, 196*(12), 2210-2215.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res, 41*(D1), D590-D596.

Quenouille, M. H. (1949). *Approximate tests of correlation in time-series 3.* Paper presented at the Mathematical Proceedings of the Cambridge Philosophical Society.

Ramasamy, D., Mishra, A. K., Lagier, J. C., Padhmanabhan, R., Rossi, M., Sentausa, E., Raoult, D., & Fournier, P. E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol, 64*(Pt 2), 384-391.

Rannala, B., & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol, 43*(3), 304-311.

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., et al. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol, 190*(20), 6881-6893.

Reller, L. B., Weinstein, M. P., & Petti, C. A. (2007). Detection and identification of microorganisms by gene amplification and sequencing. *Clin Infect Dis, 44*(8), 1108.

Remm, M., Storm, C. E., & Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol, 314*(5), 1041-1052.

Ren, F. R., Tanaka, H., & Yang, Z. H. (2009). A likelihood look at the supermatrix-supertree controversy. *Gene, 441*(1-2), 119-125.

Richter, M., & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences, 106*(45), 19126-19131.

Rivera, M. C., & Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science, 257*(5066), 74-76.

Robbertse, B., Yoder, R. J., Boyd, A., Reeves, J., & Spatafora, J. W. (2011). Hal: an automated pipeline for phylogenetic analyses of genomic data. *PLoS currents, 3*, RRN1213.

Rodriguez-R, L. M., Grajales, A., Arrieta-Ortiz, M. L., Salazar, C., Restrepo, S., & Bernal, A. (2012). Genomes-based phylogeny of the genus Xanthomonas. *BMC Microbiol, 12*.

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

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

Rong, X., Guo, Y., & Huang, Y. (2009). Proposal to reclassify the *Streptomyces albidoflavus* clade on the basis of multilocus sequence analysis and DNA–DNA hybridization, and taxonomic elucidation of *Streptomyces griseus* subsp. *solvifaciens*. *Syst Appl Microbiol, 32*(5), 314-322.

Rosselló-Mora, R. (2005). Updating prokaryotic taxonomy. *J Bacteriol, 187*(18), 6255-6257.

Rosselló-Mora, R. (2006). DNA-DNA reassociation methods applied to microbial taxonomy and their critical evaluation. In E. Stackebrandt (Ed.), Molecular Identification, Systematics, and Population Structure of Prokaryotes (pp. 23-50): Springer.

Rossello-Mora, R., & Amann, R. (2015). Past and future species definitions for Bacteria and Archaea. *Syst Appl Microbiol, 38*(4), 209-216.

Sachse, K., Bavoil, P. M., Kaltenboeck, B., Stephens, R. S., Kuo, C.-C., Rosselló-Móra, R., & Horn, M. (2015). Emendation of the family *Chlamydiaceae*: proposal of a single genus, *Chlamydia*, to include all currently recognized species. *Syst Appl Microbiol, 38*(2), 99-103.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol, 4*(4), 406-425.

Sanger, F., Brownlee, G., & Barrell, B. (1965). A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol, 13*(2), 373-IN374.

Santos, S. R., & Ochman, H. (2004). Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol, 6*(7), 754-759.

Sawana, A., Adeolu, M., & Gupta, R. S. (2014). Molecular signatures and phylogenomic analysis of the genus Burkholderia: proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus Paraburkholderia gen. nov. harboring environmental species. *Frontiers in genetics, 5*, 429.

Schildkraut, C. L., Marmur, J., & Doty, P. (1961). The formation of hybrid DNA molecules and their use in studies of DNA homologies. *J Mol Biol, 3*(5), 595-IN516.

Schleifer, K. H. (2009). Classification of Bacteria and Archaea: past, present and future. *Syst Appl Microbiol, 32*(8), 533-542.

Schopf, J. W. (1978). The evolution of the earliest cells. *Sci Am, 239*(3), 111-138.

Segata, N., Bornigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications, 4*, 2304.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods, 9*(8), 811-814.

Seshadri, R., Myers, G. S. A., Tettelin, H., Eisen, J. A., Heidelberg, J. F., Dodson, R. J., Davidsen, T. M., DeBoy, R. T., Fouts, D. E., Haft, D. H., et al. (2004). Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc Natl Acad Sci U S A, 101*(15), 5646-5651.

Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol, 16*, 1114-1116.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol, 7*(1), 539.

Singh, B., & Gupta, R. S. (2009). Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. *Mol Genet Genomics, 281*(4), 361-373.

Skerman, V. B. D., McGowan, V., & Sneath, P. H. A. (1980). Approved Lists of Bacterial Names. *Int J Syst Bacteriol, 30*(1), 225-420.

Snel, B., Huynen, M. A., & Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annu Rev Microbiol, 59*, 191-209.

Spröer, C., Mendrock, U., Swiderski, J., Lang, E., & Stackebrandt, E. (1999). The phylogenetic position of *Serratia*, *Buttiauxella* and some other genera of the family *Enterobacteriaceae*. *Int J Syst Evol Microbiol, 49*(4), 1433-1438.

Stackebrandt, E. (2006). Defining Taxonomic Ranks. *The Prokaryotes: Symbiotic associations, biotechnology, applied microbiology, 1*, 29-57.

Stackebrandt, E. (2007). Forces shaping bacterial systematics. *Microbe - American Society for Microbiology, 2*(6), 283.

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

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kämpfer, P., Maiden, M. C., Nesme, X., Rosselló-Mora, R., Swings, J., & Trüper, H. G. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol, 52*(3), 1043-1047.

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bacteriol, 44*(4), 846-849.

Stackebrandt, E., Päuker, O., Steiner, U., Schumann, P., Sträubler, B., Heibei, S., & Lang, E. (2007). Taxonomic characterization of members of the genus *Corallococcus*: Molecular divergence versus phenotypic coherency. *Syst Appl Microbiol, 30*(2), 109-118.

Stackebrandt, E., & Schumann, P. (2006). Introduction to the taxonomy of actinobacteria The Prokaryotes (3 ed., pp. 297-321).

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313.

Stanier, R., & Van Niel, C. (1941). The main outlines of bacterial classification. *J Bacteriol, 42*(4), 437.

Stanier, R. Y., & Niel, C. v. (1962). The concept of a bacterium. *Arch Microbiol, 42*(1), 17-35.

Stanier, R. Y. A., Ingraham, E. A., Stanier, J. L. R. Y., Adelberg, E. A., & Ingraham, J. L. (1976). *The microbial world*.

Starkey, R. L., & Waksman, S. A. (1943). Fungi tolerant to extreme acidity and high concentrations of copper sulfate. *J Bacteriol, 45*(5), 509.

Stephens, D. S., Greenwood, B., & Brandtzaeg, P. (2007). Epidemic meningitis, meningococcaemia, and *Neisseria meningitidis*. *Lancet, 369*(9580), 2196-2210.

Sutcliffe, I., Trujillo, M., Whitman, W. B., & Goodfellow, M. (2013). A call to action for the International Committee on Systematics of Prokaryotes. *Trends Microbiol, 21*(2), 51-52.

Sutcliffe, I. C. (2015). Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again. *Frontiers in genetics, 6*, 218.

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol, 56*(4), 564-577.

Tang, Y. W., Ellis, N. M., Hopkins, M. K., Smith, D. H., Dodge, D. E., & Persing, D. H. (1998). Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. *J Clin Microbiol, 36*(12), 3674-3679.

Thompson, C. C., Amaral, G. R., Campeão, M., Edwards, R. A., Polz, M. F., Dutilh, B. E., Ussery, D. W., Sawabe, T., Swings, J., & Thompson, F. L. (2015). Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Arch Microbiol, 197*(3), 359-370.

Thompson, C. C., Chimetto, L., Edwards, R. A., Swings, J., Stackebrandt, E., & Thompson, F. L. (2013). Microbial genomic taxonomy. *BMC Genomics, 14*(1), 913.

Tindall, B. J., Kämpfer, P., Euzéby, J. P., & Oren, A. (2006). Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. *Int J Syst Evol Microbiol, 56*(11), 2715-2720.

Tindall, B. J., Rossello-Mora, R., Busse, H. J., Ludwig, W., & Kampfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol, 60*(Pt 1), 249-266.

Trachana, K., Larsson, T. A., Powell, S., Chen, W. H., Doerks, T., Muller, J., & Bork, P. (2011). Orthology prediction methods: a quality assessment using curated protein families. *Bioessays, 33*(10), 769-780.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods, 12*(10), 902-903.

Ukkonen, E. (1985). Finding approximate patterns in strings. *Journal of algorithms, 6*(1), 132-137.

Ulitsky, I., Burstein, D., Tuller, T., & Chor, B. (2006). The average common substring approach to phylogenomic reconstruction. *J Comput Biol, 13*(2), 336-350.

Valot, B., Guyeux, C., Rolland, J. Y., Mazouzi, K., Bertrand, X., & Hocquet, D. (2015). What It Takes to Be a *Pseudomonas aeruginosa*? The Core Genome of the Opportunistic Pathogen Updated. *PLoS One, 10*(5).

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet, 30*(9), 418-426.

Vandamme, P., & Peeters, C. (2014). Time to revisit polyphasic taxonomy. *Antonie Van Leeuwenhoek, 106*(1), 57-65.

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res*, gkv657.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., & Holt, R. A. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351.

Verma, M., Lal, D., Kaur, J., Saxena, A., Kaur, J., Anand, S., & Lal, R. (2013). Phylogenetic analyses of phylum Actinobacteria based on whole genome sequences. *Res Microbiol, 164*(7), 718-728.

Virtanen, A. I. (1947). The biology and chemistry of nitrogen fixation by legume bacteria. *Biological Reviews, 22*(3), 239-269.

Waksman, S. A. (1934). The distribution and conditions of existence of bacteria in the sea. *Ecol Monogr, 4*(4), 523-529.

Wang, L.-S., Leebens-Mack, J., Wall, P. K., Beckmann, K., DePamphilis, C. W., & Warnow, T. (2011). The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 8*(4), 1108-1119.

Wang, Y., & Qian, P.-Y. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One, 4*(10), e7401.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. *Nature, 171*(4356), 737-738.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res, 42*(Database issue), D581-591.

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M., Moore, L., Moore, W., Murray, R., & Stackebrandt, E. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Evol Microbiol, 37*(4), 463-464.

Wetterstrand, K. (2016). Data from the NHGRI Genome Sequencing Program (GSP).

White, N. J. (2003). Melioidosis. *Lancet, 361*(9370), 1715.

Whitman, W. B. (2015a). *Bergey's Manual of Systematics of Archaea and Bacteria*. Hoboken, New Jersey: Wiley.

Whitman, W. B. (2015b). Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Syst Appl Microbiol, 38*(4), 217-222.

Wilson, K. H. (1995). Molecular biology as a tool for taxonomy. *Clin Infect Dis, 20*(Supplement 2), S117.

Winogradsky, S. (1952). *On the classification of bacteria.* Paper presented at the Ann Inst Pasteur (Paris).

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev, 51*(2), 221.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences, 87*(12), 4576-4579.

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

Workowski, K. A., Berman, S. M., & Douglas, J. M. (2008). Emerging antimicrobial resistance in *Neisseria gonorrhoeae*: Urgent need to strengthen prevention strategies. *Ann Intern Med, 148*(8), 606-613.

World Health Organization. (2011). Prevalence and incidence of selected sexually transmitted infections, *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, syphilis and *Trichomonas vaginalis*: methods and results used by WHO to generate 2005 estimates. Geneva: the Organization; 2011. *World Health Organization, Geneva, SwitzerlandISBN, 978*(92), 4.

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

Wu, M., Chatterji, S., & Eisen, J. A. (2012). Accounting for alignment uncertainty in phylogenomics. *PLoS One, 7*(1), e30288.

Wu, M., & Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol, 9*(10), R151.

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics, 28*(7), 1033-1034.

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics, 13*(5), 303-314.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K. H., Ludwig, W., Glöckner, F. O., & Rosselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol, 31*(4), 241-250.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K. H., Whitman, W. B., Euzeby, J., Amann, R., & Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology, 12*(9), 635-645.

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

Zhang, G., Gao, B., Adeolu, M., Khadka, B., & Gupta, R. S. (2016). Phylogenomic Analyses and Comparative Studies on Genomes of the Bifidobacteriales: Identification of Molecular Signatures Specific for the Order Bifidobacteriales and its different Subclades. *Frontiers in microbiology, 7*.

Zhi, X.-Y., Zhao, W., Li, W.-J., & Zhao, G.-P. (2012). Prokaryotic systematics in the genomics era. *Anton Leeuw Int J G, 101*(1), 21-34.

Zmasek, C. M., & Eddy, S. R. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics, 17*(9), 821-828.

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol, 8*(2), 357-366.

Zuo, G., Xu, Z., & Hao, B. (2015). Phylogeny and Taxonomy of Archaea: A Comparison of the Whole-Genome-Based CVTree Approach with 16S rRNA Sequence Analysis. *Life, 5*(1), 949-968.