The First Insights into the Phylogeny, Genomics, and Ecology of the Novel Bacterial Phylum *Armatimonadetes*

A thesis

submitted in fulfilment of the requirements for the degree

of

Doctor of Philosophy in Biology

at

The University of Waikato

by

Chien-Yu Kevin Lee

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2015

Executive Summary

Currently, a large proportion of novel microbial evolutionary lineages is poorly understood due to limited coverage of representative species. These "candidate" lineages represent significant gaps in our understanding of microbial function and ecology. This study focused on *Chthonomonas calidirosea*, the earliest isolated species within *Armatimonadetes*, the most recently-recognised bacterial phylum. The overall aim of this research was to start to understand the ecology and phylogeny of *Armatimonadetes*, and provide a foundation for future research into the phylum, with the benefit of narrowing the current knowledge gaps on microbial diversity. This was achieved by integrating multiple data types (phylogenetics, genomics, and community profiling metagenomics).

The initial stage of this research aimed to address and clarify conflicts in reported phylogeny of the phylum *Armatimonadetes*. This study generated a comprehensive reference phylogenetic tree of 16S rRNA genes for the phylum, so that the phylogenetic position of newly-identified phylotypes can be reliably associated across studies. Multiple robust statistical methods were used to arrive at a consensus on the partitioning of classes and neighbouring phyla. The process also helped to identify and exclude candidate phyla previously misattributed to *Armatimonadetes*, thus better defining the phylum for future studies.

The deeply-branching phylogenetic relationship of *Armatimonadetes* with other bacterial phyla was resolved by the sequencing of *C. calidirosea* T49^T genome and analysing concatenated amino acid sequences of conserved genes against homologs in other prokaryotic genomes. The phylogenomic analysis showed *Chloroflexi* to be the most closely related formal phylum to *Armatimonadetes*. This publication was the first analysis of a genome from the phylum *Armatimonadetes*, and provided evolutionarily- and genetically-distinct insights to the overall knowledge of microbial genetic diversity. Analysis of the genome showed a metabolism geared towards non-cellulosic carbohydrates as the carbon and energy source, which coincides with previous culture-based physiological experiments (Lee et al., 2011). Genetic mechanisms behind leucine auxotrophy and narrow pH growth range were also identified. These observations supported the theory that *C. calidirosea* T49^T occupies the niche of a scavenger of diverse species of carbohydrates within geothermal environments, in association with cellulolytic community members. In addition, the genome exhibited an unusual disorganisation of functionally-related genes typically found in conserved operons. The relatively high abundance of sigma factors (relative to genome size) in strain T49^T may play an important role in gene regulation and coordination of metabolic pathways to compensate for the scattering of operons. Overall, this research built upon the previous physiological

characterisation *C. calidirosea* T49^T (Lee et al., 2011), resulting in a more in-depth and integrated analysis of the bacterium through both phenotypic and genotypic information.

Finally, to investigate the genome dynamics of the species (particularly in genome organisation and adaptation to various environments as a scavenger), and to provide ecological and evolutionary context beyond the single genome analysed, the genomes of three additional *C. calidirosea* isolates cultured from diverse locations across the Taupō Volcanic Zone were extracted, sequenced, and compared to T49^T. The genomes exhibited higher within-species conservation than other thermophilic species such as *Thermus thermophilus* (Henne et al., 2004; Jiang et al., 2013; Oshima & Ariga, 1975) and *Sulfolobus islandicus* (Reno et al., 2009) isolated from similar geographical distance. No genomic rearrangements were identified between *C. calidirosea* isolates. The majority of variation was limited to single nucleotide polymorphisms, with a limited number of horizontally-transferred genes and differentially-present fast-evolving genes, such as restriction modification system. The phylogeny and carbohydrate utilisation profiles of the isolates correlated with the geographical relationship between the sample sites rather than with other factors, such as soil geochemistry or microbial communities of the sites. The correlation between geography and phylogeny, low abundance of *C. calidirosea* at all sample sites (ranging from 0.006 % to 0.3 %), and the high genomic conservation indicated rapid aeolian dispersal and localised extinction as the most probable causes of homogeneity between the populations. The findings contribute to a better understanding of the genome dynamics and ecology of *C. calidirosea*, as well as the dispersal possibilities of free-living bacteria between distinct and discrete habitats.

These studies addressed the overarching aim to investigate the ecology and phylogeny of *Armatimonadetes* through the research outlined above. This body of work contributed greatly to our understanding of *Armatimonadetes* phylogeny, both by clarifying its internal taxonomy and its position relative to neighbouring clades. Furthermore, it contributed to understanding of *Armatimonadetes* ecology by richly describing the ecological niche, genome, and lifestyle of *C. calidirosea*. Not only does this work greatly increase our understanding of the newest of the 30 prokaryotic phyla (Euzéby, 2011, Retrieved in December 2014), it also provides a rich foundation for future study.

Publications Resulting from this Thesis

Dunfield, P. F., Tamas, I., **Lee, K. C**., Morgan, X. C., McDonald, I. R., & Stott, M. B. (2012). Electing a candidate: a speculative history of the bacterial phylum OP10. *Environmental Microbiology*, *14*(12), 3069–80. doi:10.1111/j.1462-2920.2012.02742.x

Lee, K. C.-Y., Stott, M. B., & Dunfield, P. F. (2013). Phylum *Armatimonadetes*. In E. Rosenberg, E. F. DeLong, F. Thompson, S. Lory, & E. Stackebrandt (Eds.), *The Prokaryotes* (4th ed.). Springer.

Lee, K. C. Y., Herbold, C. W., Dunfield, P. F., Morgan, X. C., McDonald, I. R., & Stott, M. B. (2013). Phylogenetic delineation of the novel phylum *Armatimonadetes* (former Candidate Division OP10) and definition of two novel candidate divisions. *Applied and Environmental Microbiology*, *79*(7), 2484–7. doi:10.1128/AEM.03333-12

Lee, K. C., Morgan, X. C., Dunfield, P. F., Tamas, I., McDonald, I. R., & Stott, M. B. (2014). Genomic analysis of *Chthonomonas calidirosea*, the first sequenced isolate of the phylum *Armatimonadetes*. *The ISME Journal*, 8, 1522–1533. doi:10.1038/ismej.2013.251

Lee, **K. C**., Stott, M. B., Dunfield, P. F., Huttenhower, C., McDonald, I. R., & Morgan, X. C. Comparative genomics and metagenomics of geographically-diverse *Chthonomonas calidirosea* isolates. *Applied and Environmental Microbiology* (In Revision).

Table of Contents

List of Figures

Supplementary figures cited within the thesis are located in Chapter 8 - Appendices.

List of Tables

Supplementary tables cited within the thesis are located in Chapter 8 - Appendices.

Acknowledgements

Whenever I recall my days in Taupō, the memories always remind me of how fortunate I've been. This was thanks in no small part to the wonderful people I've encountered since the first day I arrived in the lakeside town of Taupō with a box of clothes and ramen noodles. Six addresses later, this chapter of my life is finally coming to an end, and I would like to acknowledge the people who made this thesis possible with my gratitude.

First and foremost, I would like to thank my supervisors, Matt, Xochitl, and Ian, for keeping me oriented and (mostly) sane throughout this great journey. It's hard to summarise how much you've helped me through the last 5.5 years of my time in Taupō. I have been extremely lucky and I couldn't have asked for better PhD supervisors.

Thanks Mum and Dad for sending me the holiday photos (they certainly kept me motivated), and all of the support and love you've shown me. Nothing made me miss you more than working well into the morning on marathon experiments.

Thank you Nellie for putting up with working in the same office as me. Despite our different fields, you've made great impact on my work as an excellent listener and by providing perspectives not affected by the proximity to the subject matters. Thank you Jean and Karen. It was a privilege demonstrating how to make media when you started - now you know more on how to work with thermophilic cultures than I do. Your work ethic inspire me, and your presence (as well as the delicious food you made) have made my time at GNS Science so much more enjoyable. Best of luck for your PhDs.

Thank you Gabriel, Jeff, Ryland, "Dr Bogan" Dave, Mariam, and Caleb among many others for the life part of work-life balance.

Special thanks to my co-author Craig Herbold, who taught me so much on phylogeny, which became an underpinning theme throughout this thesis. Thanks to Alexander Kmoch, Michael Rosenberg, Timothy Tickle, Magali Moreau, and Georgia Wakerley for the help in my last research chapter.

Of course, all of this would not be possible without the support of the Sarah Beanland Memorial Scholarship from GNS Science, or the great people and facilities at GNS Wairakei. Thank you.

Notes for the Readers

This thesis, as part of the PhD requirement for the University of Waikato, is structured in the following order: The introduction chapter (Chapter 1) aims to introduce fundamental subjects which will be found throughout this thesis. The literature review in Chapter 2 provides a comprehensive overview of pre-existing publications, including the taxonomy, physiology, genomics, and ecology, of the novel phylum *Armatimonadetes* and member species therein. Knowledge gaps within pre-existing research are highlighted. The subsequent research chapters (Chapter 4-6) aims to address the key questions related to the knowledge gaps (Figure 0.1). Drawing from the research chapters, the thesis concludes in Chapter 7 to provide a synthesis of findings and future outlooks of the research on *C. calidirosea* and *Armatimonadetes*. The work presented in this thesis was the result from several collaborative efforts. I have made substantial contributions to all work presented. All co-authored work contribution within this PhD study are outlined in Section 3.3 - Co-authorship forms. Full citations of author contributions are detailed in the prefaces of respective research chapters.

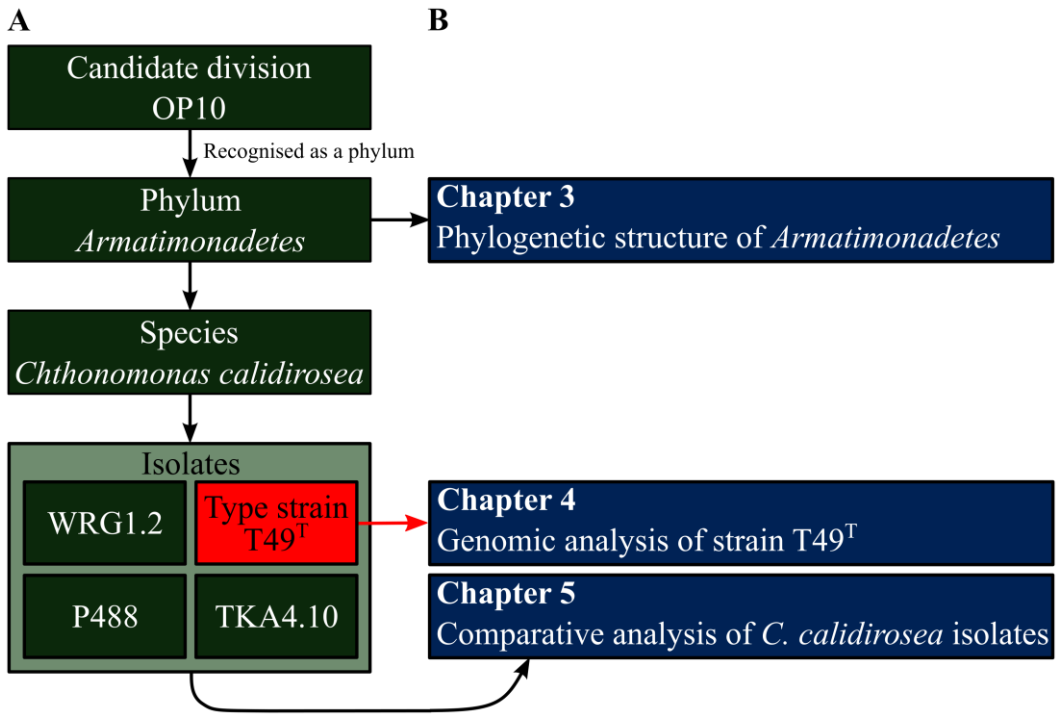


Figure 0.1 - Relationships between sources of data and chapters within this thesis. (A) Sources of data with arrows denoting their hierarchical or historical relationships. *Chthonomonas calidirosea* type strain T49^T is emphasised in red. (B) Research chapters dealing with particular aspect of phylum *Armatimonadetes*, utilising data derived from data sources indicated by arrows.

List of Abbreviations

ABC – ATP-binding cassette

BI – Bayesian inference

CAE - Carbohydrate active enzyme

CE - Cellulose esterase

COG - Clusters of orthologous group

CRISPR - Clustered regularly interspaced short palindromic repeats

DDBJ - DNA Data Bank of Japan

ECF - Extracytoplasmic function

EMBL - European Molecular Biology Laboratory

GH – Glycosyl hydrolase

HGT – Horizontal gene transfer

IMG/ER - Integrated Microbial Genomes / Expert Review system

MFS - Major facilitator superfamily

MIP - Major intrinsic protein

ML – Maximum-likelihood

MTT - 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide

NCBI – National Center for Biotechnology Information

NGS – Next-Generation Sequencing

NJ – Neighbour-joining

ORF – Open reading frame

OTU – Operational taxonomic unit

PCR – Polymerase chain reaction

PL - Pectin lyase

PP – Posterior probability

rRNA – Ribosomal RNA

SH - Shimodaira-Hasegawa test

SNP – Single nucleotide polymorphism

SOC – Soluble organic carbon

SSU – Small subunit

TKA – Te Kopia

TKT - Tikitere

TVZ - Taupō Volcanic Zone, New Zealand

WKT - Waikite

WRG - Wairakei

XRD – X-ray diffraction

XRF – X-ray fluorescence

YNP – Yellowstone National Park, USA

Chapter 1 Introduction

1.1 - The need to characterise the great prokaryotic unknown

Microorganisms play important roles in a wide range of global to organism-scale processes that are fundamental to the study of biology. These processes range from the Great Oxidation Event which changed the geochemical composition of the planet and gave rise to aerobic life forms (Sessions et al., 2009), to driving the oceanic microbial carbon cycle (Azam, 1998), to the coevolution between the human immune system and various microbial communities (Lee & Mazmanian, 2010). The metabolic diversity and genetic fluidity of microorganisms (Pace, 1997) enables them to thrive in extreme reaches of life from deep underground (Chivian et al., 2008) to the troposphere (DeLeon-Rodriguez et al., 2013). The prokaryotes (*Archaea* and *Bacteria*) represent a large proportion of global biomass, with estimations ranging from 51.7 Pg (Kallmeyer et al., 2012) to 546.2 Pg of carbon (Whitman et al., 1998). In comparison, global organic carbon content of plant life is estimated to be around 561.8 Pg (Whitman et al., 1998). The wide range of these estimates not only highlights the significance of prokaryotic biomass, but also reflects the lack of information on microbial communities in the environment, despite their abundance and ubiquitous nature.

In recent decades, research into microbial diversity has shown that microorganisms are a rich source of biochemical resources. The adaptation of microorganisms to diverse environments provides a biotechnological wealth, including the famous heat-resistant *Taq* polymerase which revolutionised molecular biology (Chien et al., 1976), novel antibiotics to address the drug-resistance problem (Lewis et al., 2010), and even the use of entire microbial communities to facilitate industrial processes (Antoni et al., 2007) or remediate pollution (Dojka et al., 1998). Aside from decomposition, the complex ecological and biochemical interactions within microbial communities may play other important roles within the ecosystems. Recent research has only begun to uncover the interactions between animals and plants and their closely-associated microbial communities, such as plant rhizospheres (Mendes et al., 2011) and the human microbiome (Gevers et al., 2012). Perturbations within these communities have been implicated with diseases (e.g., loss of microbial diversity and pathology caused by dominance of previously low-abundance species) or greater wellbeing (e.g., recruitment of protective species against pathogens) of the host organisms (Bäckhed et al., 2005; Berendsen et al., 2012). Free-living microorganisms such as those found in soil or oceanic microbiomes (Gilbert et al., 2011) also have important roles in the underlying functioning of their respective ecosystems. Despite the importance and prevalence of microorganisms and their communities, many of the underlying biochemical and ecological mechanisms are poorly-understood.

Identifying and characterising the microbial species and the mechanisms involved, such as harnessing human microbiomes as a source of antimicrobial agents (Fischbach & Walsh, 2009), may provide a wealth of resources for biotechnological developments.

In order to gain a better understanding of any ecosystem or harness the diverse metabolic capabilities from microorganisms, it is crucial to address the present lack of information of the vast microbial landscape. Microbial ecology can be formulated as examining the identities of various taxonomic groupings, their capabilities, as well as the actions and interactions taking place over spatial and temporal distributions. Overall, these aspects rely on the identification of evolutionary lineages in order to form a cohesive picture (Hugenholtz & Pace, 1996). The epistemological framework "*…there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - the ones we don't know we don't know*" by former US Secretary of Defence Donald Rumsfeld, originally used to describe limits of military intelligence, has been appropriated to describe the state of microbial biogeography, the science of investigating spatial and temporal patterns of microbial diversity (Fierer, 2008). Based on environmental surveys using molecular markers, a majority (Rappé & Giovannoni, 2003; Yarza et al., 2014) of bacterial phyla-level taxa were not represented by any cultured strain (see Section 1.3 The development of DNA sequencing technologies and community profiling**)**. These lineages ("candidate divisions") represent the "known unknowns" in that their existence is known, but other detailed information is lacking. This thesis aims to address the current paucity of knowledge concerning one such recently-described phylum, *Armatimonadetes*, previously known as Candidate Division OP10. This research is outlined in Section 1.7 Research aims. Beyond environmental surveys, the "unknown unknowns" are moving targets in the realm of speculations and projections. However, if the history of molecular and environmental microbiology is any indication, the acceleration of research progress in this field is certain to reveal many exciting discoveries ahead.

1.2 - A brief introduction to molecular and environmental microbiology

Since the dawn of the discipline in the 17th century, microbiology has faced unique challenges from the diminutive nature of the subject matter. Microorganisms, such as bacteria, were often difficult to identify due to indistinct or confounding features. We now know these are partly due to the fluid genetic landscape of these organisms with rapid convergent evolution and horizontal gene transfer (Woese, 1987). In addition to hurdles in systematic identification and establishment of relationships between known

strains, the discovery of novel diversity has been hampered by our limited ability to isolate and cultivate novel environmental microorganisms, which in turn has led to an incomplete perspective of the microbial world. The cultivation barrier, famously demonstrated through the "great plate count anomaly" (Staley & Konopka, 1985) in which only a small subset of environmental cells could be successfully cultivated, presents an invisible and persistent obstacle in traditional microbiological research efforts, that relied on the isolation and cultivation of microorganisms from their heterogeneous environment. Novel species can be difficult to isolate and cultivate due to unknown physiological requirements and/or competition from the dominant species. Furthermore, many microbial lineages may simply be unknown, precluding targeting for study. Culture-independent methods were developed to circumvent these issues by targeting microbial phylogenetic diversity directly from the environment, without selection bias from attempting to cultivate the species from a community (Hugenholtz, Goebel, et al., 1998). This new approach borrowed heavily from other disciplines, such as the developing field of molecular biology. This enabled unbiased discrimination of taxa by molecular means rather than by pure cultures and phenotypic traits. Molecular biology and its influence on evolutionary theory, particularly the neutral theory of evolution (Kimura, 1985), prompted microbiologists to utilise molecular markers such as the highly conserved 16S rRNA gene as biological "barcodes" (Hebert et al., 2003) to identify lineages and infer their evolutionary relationships (Lane et al., 1985).

Compared to the phenotypical approach in traditional microbiology, where morphology or biochemical characteristics are measured, molecular phylogeny based on the 16S rRNA gene sequence offered several advantages (Ludwig & Schleifer, 1994). First, this gene is highly conserved and present in all known bacteria and archaea. Therefore, it presents a consistent orthologous feature, in which divergence in the sequences can be used to identify the taxa as well as inferring the phylogeny of the organisms. Second, the gene contains conserved regions at which broad-specificity "universal" PCR primers can be used to amplify 16S rRNA genes from the environment, targeting both known and unknown organisms without the trial and error of cultivation attempts. Finally, nine hypervariable regions provide phylogenetic resolution for identifying related species (Chakravorty et al., 2007). In contrast, physiological characteristic-based methods such as the processes outlined in Bergey's Manual of Determinative Bacteriology (Holt, 1994) are limited to pure, cultivated species, and the biochemical characteristics tested may be confounded due to convergent evolution or lateral gene transfer. In addition, environmental 16S rRNA gene sequencing offers a unified framework where microbial diversity can be quickly compared with pre-existing entries in primary databases such as NCBI GenBank (Benson et al., 2011) and EMBL-Bank (Leinonen et al., 2011), or

curated databases such as SILVA (Pruesse et al., 2007) and Greengenes (DeSantis et al., 2006).

Due to these advantages of utilising the 16S rRNA gene for identification and phylogenetic analysis, it is the most commonly used molecular marker for environmental prokaryotes (Stackebrandt & Goebel, 1994). In order to produce compatible sequencing data, the popularity of the marker itself also influences future studies. Based on the SILVA database release 119 (released in July 2014), quality-controlled 16S rRNA sequences increased from 101,781 to 4,346,367 in the last decade (Pruesse et al., 2007). Evolutionary relationships may be inferred from 16S rRNA gene sequences, which can provide useful criteria for higher-order taxonomy (Garrity et al., 2004; Stackebrandt et al., 1994). The utilisation of molecular systematics resulted in the establishment of the new domain *Archaea* from parts of the biodiversity previously thought to be *Bacteria* (Woese et al., 1990; Woese & Fox, 1977), and brought forth the three-domain taxonomic system commonly in used today. Methods of phylogenetic inference have evolved greatly along with the amount of data generated and the advancement of bioinformatics computation since the 1980s. Simple distance-matrices-based methods (Felsenstein, 1989) were supplemented by more complex (and computationally intensive) approaches (Holder & Lewis, 2003) like maximum parsimony, maximum likelihood, and Bayesian inference.

The commercialisation of the polymerase chain reaction (PCR) and maturation of Sanger sequencing technology in the early 1990s broadened applicable targets for environmental surveys, and enabled the profiling of complex soil communities (Stackebrandt et al., 1993) as well as low abundance oligotrophic marine environments (Britschgi & Giovannoni, 1991; Giovannoni et al., 1990; Schmidt et al., 1991). The explosion of microbial diversity surveyed through molecular markers, such as lipid profile and 16S RNA genes, has led to the recognition that the majority of microbial diversity has evaded isolation and cultivation. As a means to address these detected diversity, which lacked representative type strains, Hugenholtz and colleagues (1998) integrated molecular phylogeny with taxonomy and defined "candidate divisions" as putative phylum-level lineages (< 85 % 16S rRNA gene sequence similarity to known sequences) that are unaffiliated with other phyla in multiple phylogenetic analyses and datasets. This definition has since then become a pivotal approach in dealing with uncultured diversity and has been widely-applied.

1.3 - The development of DNA sequencing technologies and community profiling

As environmental surveys targeting molecular markers (e.g., 16S rRNA gene) became prevalent, the number of candidate divisions grew more rapidly than those which went on to become recognised phyla (through the isolation and description of a type species). By 2003, half of approximately 52 bacterial phyla-level lineages were candidate divisions, which lacked in physiological and genomic information from representative isolates (Rappé et al., 2003). The curated database for 16S rRNA sequences SILVA Release 119, published in July 2014, contained 62 bacterial phyla-level lineages, of which 29 contained cultured and characterised isolates (Pruesse et al., 2007). It is worth noting that the definition of taxonomic group above class level is not covered by the current Rules of the Bacteriological Code (1990 Revision) (Lapage et al., 1992) and the number of candidate divisions and recognised phyla may vary between curators and criteria used. Nonetheless, the figures shown here illustrated that the majority of microbial diversity is poorly-understood, and these shadow clades were described as the "uncultured microbial majority" (Rappé et al., 2003) and "biological dark matter" (Marcy et al., 2007). Thanks to these community surveys, previously uninvestigated microbial diversity and their associated environments can now be systematically targeted for further investigations.

While the PCR-Sanger sequencing approach has led to the proliferation of environmental surveys, the method was limited by sequencing throughput, due to its reliance on clone libraries in order to isolate 16S rRNA genes. The process chain, despite automation, remained labour-intensive and costly to representatively sample an environment. To address these problems, the late 2000s experienced a rapid growth in competing high-throughput Next-Generation Sequencing (NGS) technologies; the differences in performance and technical details of these are well described in the 2010 review by Metzker (2010). Of these competing technologies, 454 pyrosequencing (Roche Applied Science, Penzberg, Germany) became the most widely adopted technology for community profiling (Quince et al., 2009), with the advantage of relatively long read length among NGS technologies. For example, in 2012, 454 GS FLX was capable of producing read length of around 700 base pairs (bp) long, while Sanger sequencing was capable of up to around 900 bp (Liu et al., 2012). However, in the 2010s, Illumina (Illumina Inc. San Diego, USA) surpassed 454, becoming the dominant sequencing platform with ~60 % market share by revenue (Mohamed & Syed, 2013). Illumina technology is favoured due to high throughput and low cost per base (Liu et al., 2012). These advantages enabled deep sequencing of complex communities and large number of samples through multiplexing. In addition, maturation of the platform as well as data processing methods enables novel approaches to move beyond simply recovering 16S

rRNA genes and made metagenomics (Handelsman, 2004) and metatranscriptomics (Moran, 2009) financially viable compared to previous Sanger sequencing.

The sequencing of functional genes directly from the environment enabled not only identification of the community members, but also inference of, physiological information from genomic data, while bypassing the cultivation barrier. To date, both metagenomic (Handelsman, 2004) and single cell genomic (Lasken, 2012) approaches have led to the retrieval and reconstruction of genomes (with varying degrees of completeness) from uncultivated organisms within candidate taxa, such as OP1 (Takami et al., 2012), TM6 (McLean et al., 2013), TM7 (Marcy et al., 2007), and "Nanohaloarchaea" a novel lineage of unusually small archaea (Narasingarao et al., 2012). Metatranscriptomics has uncovered underlying, time-dependent metabolic activities in communities such as the human gut microbiome (Gosalbes et al., 2011) and subtropical ocean gyres (Poretsky et al., 2009). Many of these putative species also contribute to a larger group of "Candidatus" taxa (http://www.bacterio.net/-candidatus.html) which have been characterised without being maintained in culture collections. The genomic analysis of uncultivated organisms provides insights into their metabolisms and allows ecological inferences of the roles of the organism in its environment. As current high throughput technologies continue to become more economical, another generation of sequencing technologies (Buermans & den Dunnen, 2014; van Dijk et al., 2014) is being developed to address varying market demands. The development and convergence of first, second, and third-generation sequencing technology has enabled environmental microbiology to uncover the "unknown majority" through culture-independent methods in addition to the traditional culture-based approaches.

Despite the power of DNA sequencing and bioinformatic analysis on genetic data, much of the *in silico* approach still relies on sequence (DNA, RNA, amino acid) homology to known grounding information. Obtaining this grounding information relied on relatively labour-intensive, low-throughput processes, such as physiological characterisation of a cultivated strain or biochemical analysis of an enzyme. In order to understand a microorganism and its ecological interactions, its phylogeny alone is insufficient without the understanding of its physiology. This information is often difficult to obtain because cultivation and characterising fastidious species remained a problem. Some species may be resistant to conventional laboratory setup due to factors such as closely coupled-metabolism of an obligate symbiont, faster-growing competitive species, or simply due to the numerous and unknown environmental parameters that the cultivation setup failed to

recreate, without the physiological characteristics of the organism that is being targeted (Stewart, 2012).

Overall, the culture-independent approach has provided perspectives on microbial genetic and phylogenetic diversity from the environment, which in turn helps culture-based methods to target and characterise novel species. Culture-based methods generate valuable experimental results such as physiological characteristics, and enables further investigations such as genomic and transcriptomic studies. Knowledge from culture-based approach provides grounding observations for the predictive models (e.g., gene calling and annotation) of the culture-independent approach. Therefore, the two approaches should be viewed as complementary in addressing a common goal of exploring unknown microbial diversity. Additionally, microbial species that are challenging to investigation may be of particular scientific interest[a], due to their likely distinctiveness from known organisms and previously developed techniques. In order to provide an encompassing perspective on the subject matter, this thesis aimed to utilise both approaches (culture and culture-independent) to target a novel microbial lineage from an environment of particular interest (geothermal sites in the Taupō Volcanic Zone).

1.4 - Extremophiles

Extremophiles have long been associated with the search of novel microbial diversity. Extreme environments (in temperature, pH, salinity etc.) present several desirable characteristics for the investigation of microbiology (see above) including uncommon selection pressures and the distinct and isolated habitat from the surrounding environments (Whitaker et al., 2003), factors which often lead to distinct and novel lineages that have been resistant to cultivation attempts (Rappé et al., 2003). Additionally, simple community structures and isolated habitats present useful environments to investigate microbial biogeography and biological interactions (Whitaker et al., 2003). Investigation into extremophiles may provide insights into the early evolutionary history of life on Earth (Schwartzman & Lineweaver, 2004).

Extremophiles can be classified into several major categories based on the "extremes" of physicochemical attributes within the host environments. These include temperature (thermophiles, psychrophiles), pH (acidophiles, alkaliphiles), and salinity (halophiles), as

[a] such as the rationale behind the "Genomic Encyclopedia of Bacteria and Archaea" project (Wu, Hugenholtz, et al., 2009) which aims to provide a better genomic representation of prokaryotic lineages currently negatively biased by cultivation barriers

well as organisms known for resistance to ionising radiation and desiccation, among many other smaller and less-understood groups (Rothschild & Mancinelli, 2001). An organism may possess more than one of these attributes, for example, *Anaerobranca gottschalkii* is a thermoalkaliphilic bacterium with temperature optimum of 50-55 °C and pH optimum of 9.5 (Prowe & Antranikian, 2001), which *Sulfolobus islandicus* (Brock et al., 1972), a thermoacidophilic archaeon with temperature optimum of 70-75 °C and pH optimum of 2-3.

A wide range of biochemical features evolved to enable these organisms to exist in environments hostile to other "normal" life forms. These adaptations remained targets for active ongoing research with prospect of biotechnological applications (Rothschild et al., 2001). For example, thermophiles and psychrophiles present adaptations for the two ends of the protein stability and reactivity spectrum in order to counteract protein denaturation at high temperature and slow enzyme kinetics at low temperature. Additionally, lipid membranes are also adapted to ensure suitable membrane fluidity. Many psychrophiles produce antifreezing proteins to prevent phase transition of the inter- and intracellular water molecules (Feller & Gerday, 2003). Cell membranes also play an important role in survival in extreme pH range and salinity by decreasing proton permeability and/or coupling energetics with the less permeable sodium in order to maintain the proton gradient within a viable range (Konings et al., 2002). Other adaptations may also lead to resistance to more than one environmental factor. For example, the radiation resistance of *Deinococcus radiodurans*, which far exceeds the level present on Earth, is also closely related to the ability of the bacterium to survive prolonged desiccation (Mattimore & Battista, 1996).

1.5 - Candidate Division OP10 and Phylum *Armatimonadetes*

In 1996 and 1998, surveys were conducted by amplifying 16S rRNA genes from sediment samples of Obsidian Pool in YNP (Barns et al., 1996; Hugenholtz, Pitulle, et al., 1998). The studies identified unusually rich microbial diversity from the geothermal pool, including 12 previously unknown phylum-level bacterial clades (named OP1 to OP12) and an archaeal clade ("Korarchaeota"). At the time, no representative isolates existed within these clades. Therefore, the highly-dissimilar lineages were termed "candidate divisions" for their putative taxonomic status. Some of these clades, based on newer phylogenetic information, were later merged into pre-existing taxa (e.g., OP12), while others (e.g., OP4, 6, and 7) remained small, with few similar sequences identified from other environments (Dunfield et al., 2012). However, some of the clades persisted and grew as phylum-level clusters, as more environmental microbial diversity was discovered and attributed to these groups. Candidate Division OP10 was the most frequently-detected OP groups (Dunfield et al., 2012). As a rough metric of abundance of the group,

the SILVA database SSU release 119 (Pruesse et al., 2007) contained 6100 OP10 sequences out of a total of 3,845,937 sequences within the bacterial domain. OP10 therefore represents 1 of every 1000 bacterial 16S rRNA gene sequences sampled and published within the SILVA database, coinciding with an earlier estimate by Dunfield et al. (2012) [b].

OP10 remained a candidate division with no cultivated representative species, but grew in attributed environmental 16S rRNA gene sequences until 2008, when a thermophilic species was isolated and cultivated from the Taupō Volcanic Zone (Stott et al., 2008). The first formal characterisation of an OP10 species, however, did not occur until three years later when the description of *Armatimonas rosea* strain YO-36^T was published (Tamaki et al., 2011); the candidate division, now a proper phylum *Armatimonadetes* took its name from the publication. In the same year, a strain (closely isolated to the strain described by Stott et al. in 2008) was formally described and named *Chthonomonas calidirosea* strain T49^T (Lee et al., 2011) [c]. The third species, *Fimbriimonas ginsengisoli* strain Gsoil 348^T was published one year later (Im et al., 2012). Each of these strains is the type strains of its respective class (*Armatimonadia*, *Chthonomonadetes*, and *Fimbriimonadia*). *Armatimonas rosea* YO-36^T is the type strain of the phylum. Presently, *C. calidirosea* is the only thermophilic and acidophilic species within *Armatimonadetes*, and is the particular focus of the research in this study (outlined in Chapter 3). In the next chapter, as part of a literature review of available publications on the phylum *Armatimonadetes*, the three *Armatimonadetes* species are described and compared in detail.

[b] I am a co-author of this mini-review.

[c] I am the first author of this study.

Chapter 2

Literature Review - "Phylum *Armatimonadetes*"

Kevin C. Y. Lee[1,2], Peter F. Dunfield[3], Matthew B. Stott[1]

[1]Extremophile Research Group, GNS Science, Taupō, New Zealand

[2]School of Science, University of Waikato, Hamilton, New Zealand

[3]Department of Biological Sciences, University of Calgary, Calgary, Canada

2.1 - Preface

This literature review is primarily based on the review article "Phylum *Armatimonadetes*" which I co-authored during the course of my PhD research. The article was published as part of the book series "The Prokaryotes (4th ed.)". The text has been revised for this thesis to reflect two important publications that appeared in print after the publication of this review. These include the recognition of *Fimbriimonas ginsengisoli* strain Gsoil 348^T (Im et al., 2012) as a valid species through the announcement in the *International Journal of Systematic and Evolutionary Microbiology* Validation List No. 147, as well as the publication of its genome (Hu et al., 2014). The knowledge gaps highlighted in this literature review shaped the research aims and hypotheses regarding *Armatimonadetes* outlined in the next chapter.

This review was recently published with the following citation:

Lee, K. C.-Y., Stott, M. B., & Dunfield, P. F. (2013). Phylum *Armatimonadetes*. In E. Rosenberg, E. F. DeLong, F. Thompson, S. Lory, & E. Stackebrandt (Eds.), *The Prokaryotes* (4th ed.). Springer.

2.2 - Overview

Armatimonadetes constitutes a moderately-abundant and phylogenetically-diverse bacterial phylum. Prior to the official description of the phylum by Tamaki et al. (2011), *Armatimonadetes* phylotypes were classified as Candidate Division OP10, first identified by Hugenholtz, Pitulle et al. (1998) in a molecular study conducted at Obsidian Pool, Yellowstone National Park. While c.a. 500 nearly full-length non-redundant public-domain16S rRNA gene sequences cluster into as many as 12 class-level groupings within this phylum, only four cultivated representatives have so far been described. The phylum *Armatimonadetes* is defined on a phylogenetic basis by comparative 16S rRNA gene sequence analysis of *Armatimonas rosea* YO-36^T (Tamaki et al., 2011; = NBRC 105658^T = DSM 23562^T; GenBank/EMBL/DDBJ accession number AB529679), *Chthonomonas calidirosea* T49^T (Lee et al., 2011; = DSM 23976^T = ICMP 18418^T; GenBank/EMBL/DDBJ accession number AM749780), *Chthonomonas*-like strain P488 (Stott et al., 2008; GenBank/EMBL/DDBJ accession number AM749768), and *Fimbriimonas ginsengisoli* strain Gsoil 348^T (Im et al., 2012; = KACC 14959^T = JCM 17079^T; GenBank/EMBL/DDBJ accession number GQ339893). There are few common features shared by the cultivated strains, including aerobic, oligotrophic metabolism and Gram-negative staining cells.

2.3 - Taxonomy, historical and current

Short description of the phylum and its classes

Armatimonadetes Tamaki, Tanaka, Matsuzawa, Muramatsu, Meng, Hanada, Mori, and Kamagata 2011, 1446

Armatimonadetes phyl. nov. Armatimonadetes (Ar.ma.ti.mo.na.de'tes. N.L. fem. pl. n. Armatimonadales type order of the phylum; N.L. fem. pl. n. Armatimonadetes phylum of the order Armatimonadales).

The phylum *Armatimonadetes* is principally defined on a phylogenetic basis by comparative 16S rRNA gene sequence analysis of type strain *Armatimonas rosea* (YO-36^T) (= NBRC 105658^T = DSM 23562^T; GenBank/EMBL/DDBJ accession no. AB5296790), type strain *Chthonomonas calidirosea* (T49^T) (= DSM 23976^T = ICMP 18418^T; GenBank/EMBL/DDBJ accession no. AM749780), type strain *Fimbriimonas ginsengisoli* (Gsoil 348^T) (= KACC 14959^T = JCM 17079^T; GenBank/EMBL/DDBJ accession no. GQ339893), and uncultured representatives from various terrestrial, aquatic habitats including temperate soils, human skin, anaerobic bioreactors and waste water treatment plans, geothermal soils and springs, and plant and animal symbionts.

Armatimonadetes is a recently-defined bacterial phylum described by Tamaki et al. (2011). Currently, the phylum contains only three type strains, the phylum type strain *Armatimonas rosea* YO-36^T (Tamaki et al., 2011) which also represents the class *Armatimonadia*, *Chthonomonas calidirosea* T49^T (Lee et al., 2011) which represents the class *Chthonomonadetes*, and *Fimbriimonas ginsengisoli* Gsoil 348^T (Im et al., 2012), which has recently been reported and is described as the type strain of the novel class *Fimbriimonadia.* The description of the novel class was validated through the announcement in Validation List No. 147 of the *International Journal of Systematic and Evolutionary Microbiology*. The member classes of *Armatimonadetes* can be distinguished from each other via specific 16S rRNA gene sequence signatures: class *Armatimonadia* is characterised by (*E. coli* position numbering) 316-U; 1201-G, class *Chthonomonadetes* by 974-U, 986-A, 1219-U, and class *Fimbriimonadia* is characterised by 340-A; 916-U; 964-A; 1082-G.

Phylotypes within *Armatimonadetes* were first identified as "candidate division OP10" by Hugenholtz, Pitulle et al. (1998). Candidate Division OP10 was one among 12 phylum-level groups identified at the locale by the culture-independent phylogenetic survey of 16S rRNA gene sequences recovered from Obsidian Pool, Yellowstone National Park, USA. The molecular phylogenetic landscape has shifted with the continual advancement in sequencing and computational solutions in the following decade, however, Candidate Division OP10 remained as a candidate phylum as some previous OP groups remained with few phylotypes, or were merged into pre-existing phyla based on latter analyses with the benefit of additional data. The first cultivation of Candidate Division OP10 strains was reported by Stott et al. (2008), but OP10 was not formally recognised as a phylum until the publication of the phylum type strain *A. rosea* YO-36^T by Tamaki et al. (2011). One of the strains isolated by Stott and colleagues (2008), strain T49^T, was soon after formally described by Lee et al. (2011) as *C. calidirosea*.

2.4 - Phylogenetic structure of the phylum

2.4.1 - The phylogenetic relationship of *Armatimonadetes* with other bacterial phyla

The phylogenetic relationship of *Armatimonadetes* with other bacterial phyla is poorly-defined due to the limited number of cultivated strains. Phylogenetic analysis of *Armatimonadetes* is therefore primarily limited to 16S rRNA gene sequences.

Early studies of Candidate Division OP10, based on a very limited number of 16S rRNA gene sequences, suggested that Candidate Division OP10 was adjacent to *Dictyoglomus* and *Thermodesulfobacteria* (Hugenholtz, Pitulle, et al., 1998) or *Actinobacteria* (Rappé et

al., 2003). More recently, with more "OP10-like" sequences in public databases, phylogenetic analyses indicate that *Armatimonadetes* is probably more closely related to *Chloroflexi* (Lee et al., 2011; Tamaki et al., 2011) or *Actinobacteria* and *Firmicutes* (Portillo & Gonzalez, 2008). This assessment also generally reflects the guide tree in ARB-SILVA database (SSU Ref NR 111) (Pruesse et al., 2007) which places *Armatimonadetes* near the *Deinococcus-Thermus*, *Cyanobacteria*, and *Chloroflexi* phyla (Figure 2.1). A different assessment is offered by The Living Tree Project (LTPs108, July 2012) (Yarza et al., 2010), which focused on 16S rRNA gene sequences of most of the cultivated bacterial type strains. The LTP indicated that the two *Armatimonadetes* strains, *A. rosea* YO-36^T and *C. calidirosea* T49^T were closest to the only type strain of the phylum *Elusimicrobia*, *Elusimicrobium minutum*, and the phyla *Fusobacteria* and *Spirochetes*, thus placing *Armatimonadetes* near *Proteobacteria*.

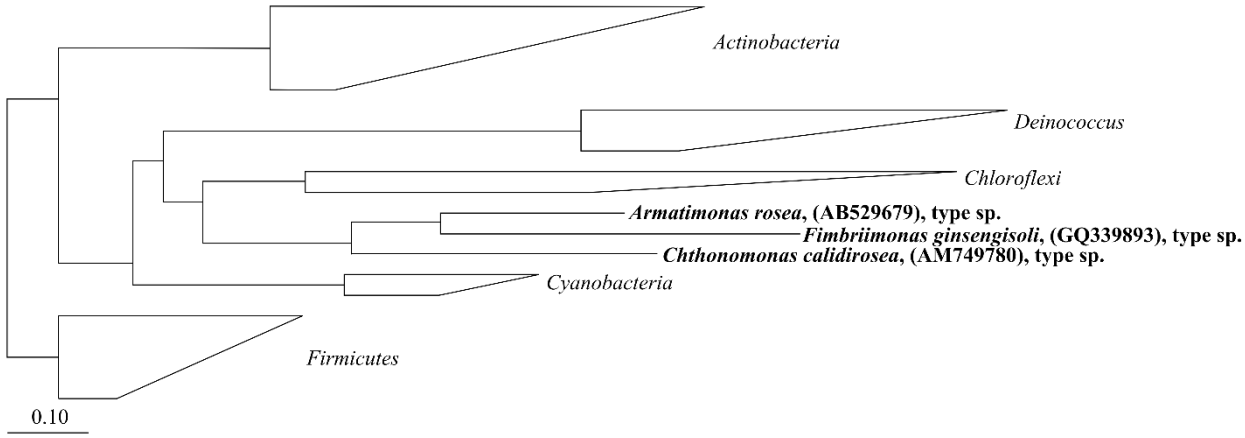


Figure 2.1 - Phylogenetic reconstruction of the phylum *Armatimonadetes* based on the maximum likelihood algorithm RAxML (Stamatakis, 2006). Sequence dataset and alignments according ARB-SILVA database (SSU Ref NR 111) (Pruesse et al., 2007) The selection of phyla neighbouring *Armatimonadetes* was based on Dunfield et al. (2012). Representative sequences from close relative genera were used to stabilise the tree topology. In addition, a 40 %maximum frequency filter was applied to remove hypervariable positions from the alignment. Scale bar indicates estimated sequence divergence.

The uncertainty surrounding the consensus phylogenetic position of *Armatimonadetes* within the domain *Bacteria* highlights the potential flaws of phylogenetic inference based solely on 16S rRNA gene sequences (Gupta & Griffiths, 2002; Henz et al., 2005). As an alternative approach, Dunfield et al. (2012) conducted concatenated multi-gene phylogeny analysis with *C. calidirosea* T49^T amino acid sequences from 29 universal core proteins derived from the draft genome and based on the method by Wu and Eisen,

(2008). The results from the study indicated a strong support for *Chloroflexi* as the nearest neighbouring phylum. This assessment may gain more support when *Armatimonadetes* is better represented with more genomes for further analysis.

2.4.2 - The phylogenetic relationships within *Armatimonadetes*

Armatimonadetes is a phylogenetically-diverse phylum which currently encompasses around 500 non-redundant (nr) near full length and high quality phylotypes (Pruesse et al., 2007). A pairwise comparison of the current *Armatimonadetes* isolates is presented in Table 2.1.

Table 2.1 - Pairwise 16S rRNA gene sequence identities of the three strains - *Armatimonas rosea* YO-36 ^{T}, *Chthonomonas calidirosea* T49^{T}, and *Fimbriimonas ginsengisoli* Gsoil 348^{T}

	***A. rosea* YO-36 ^{T}**	***C. calidirosea* T49^{T}**
***A. rosea* YO-36 ^{T}**	-	-
***C. calidirosea* T49^{T}**	79.4 %	-
***F. ginsengisoli* Gsoil 348^{T}**	80.2 %	77.1 %

The recent formal recognition of *Armatimonadetes* as a phylum has meant the phylogenetic structure within *Armatimonadetes* is in its infancy and currently lacks consistency. Recent publications have used differing selections and numbers of phylotypes, differing methodologies, and differing outgroups to build the *Armatimonadetes* phylogenetic structure (Dunfield et al., 2012; Im et al., 2012; Lee et al., 2011; Portillo et al., 2008; Tamaki et al., 2011). This has resulted in varying tree topologies and a lack of consistency in group/class nomenclature. Table 2.2 compares the differing nomenclature and numbers of primary groups identified in different studies, and includes key groups within *Armatimonadetes* such as the current isolates and the original Obsidian Pool OP10 clones. When reviewing the phylogenetic topologies of *Armatimonadetes* / Candidate Division OP10 in the literature, it is obvious that the numerical naming schemes between publications have not been entirely consistent (Table 2.2). However, the nomenclatures for Group 1, proposed as *Armatimonadia*, and Group 3 have been preserved across the publications since the phylum's acceptance (Tamaki et al., 2011). The authors also predicted the potential class status of the then yet to be characterised Group 3, which was subsequently confirmed by Lee et al. (2011) as the type class of *Chthonomonas calidirosea* T49^{T}. Im et al. (2012) proposed an additional class *Fimbriimonadia*, containing isolate *Fimbriimonas ginsengisoli*. Dunfield et al. (2012) has provided the most comprehensive phylogenetic study of the phylum to date which includes almost 500 non-redundant near-full length phylotypes, classified into 12 groups

including phylotypes from the original Obsidian Pool study (Hugenholtz, Pitulle, et al., 1998), as well as the deeply-branching clades such as Group 11 and 'Candidate Division WS1'.

Table 2.2 - Comparison of phylogenetic groupings defined in various studies.

	Publications (numbers of groups cited in publication)					Representative strains or phylotypes of significance within Group (Accession numbers)
	Tamaki et al., 2011 (4)	Dunfield et al., 2012 (12)	Portillo & Gonzalez, 2008	Lee et al., 2011 (7)	Im et al., 2012 (6)	
Group designations cited in publication	1	1	1	1	1	*Armatimonas rosea* YO-36^T (AB529679)
	2	5	6	n.d.	5	Various environmental phylotypes (e.g., AJ009490)
	3	3	7	3	3	*Chthonomonas calidirosea* T49^T (AM749780)
	4	9	5	n.d.	4	*Fimbriimonas ginsengisoli* Gsoil 348^T (GQ339893)
	n.d.	6	3	4	n.d.	Obsidian Pool clone, OPB50 (AF027092)
	n.d.	10	4/8	6	n.d.	Obsidian Pool clone, OPB90 (AF027090)
	n.d.	12	10	n.d.	n.d.	Formerly Candidate Division WS1 (AF050579 & AF050578)

n.d.: not defined.

2.5 - Molecular analyses

As of the writing of this review article (late 2013), no additional comparative molecular analysis (e.g., MLSA and MALDI-TOF) was available regarding the phylogeny within *Armatimonadetes*, and no representative genomes were publicly available or published.

2.6 - Phenotypic analyses

The main phenotypic characteristics of the cultivated members of *Armatimonadetes*, *A. rosea* YO-36^T, *C. calidirosea* T49^T and *F. ginsengisoli* Gsoil 348^T are listed in Table 2.3. Higher taxa such as order and family for these strains have yet to be described phylogenetically, and instead were defined via phenotypic description of the type genera. Considering the low 16S rRNA gene sequence similarity (Table 2.1), there are unsurprisingly few common phenotypic features of the three cultivated strains; currently these are an oligotrophic chemoheterotrophic metabolism, Gram-negative staining and obligate aerobicity. However, as discussed later in the Ecology section, the latter is unlikely to remain a conserved phenotype of the phylum, as large numbers of phylotypes are regularly detected in anaerobic environments. A comparison of general phenotypic features is presented in Table 2.3, substrate specificities in Table 2.4, a comparison of enzyme expression in Table 2.5, and a comparison of antibiotic sensitivities in Table 2.6. The type genera are listed below.

Table 2.3 - Distinct phenotypic features of the three known strains of *Armatimonadetes*.

Characteristics	*A. rosea* (YO-36^T)[1]	*C. calidirosea* (T49^T)[2-4]	*F. ginsengisoli* (Gsoil 348^T)[5]
Morphology	Rod to ovoid (1.4-1.8 x 2.4-3.2 µm)	Rod (0.5-0.7 x 2.5-3.0 µm)	Rod (0.5–0.7 x 2.5–5.0 µm)
Gram stain	Gram-negative	Gram-negative	Gram-negative
Motility	Non-motile	Motile	Non-motile
Flagella, fimbriae	-	-	fimbriae
GC Content	62.4 mol%	54.6 mol%	61.4 mol%
Atmosphere	Aerobic	Aerobic	Aerobic
Temperature range (optima)	20–40 °C (30–35 °C)	50–73 °C (68 °C)	15-30 °C (30 °C)
pH range (optima)	pH 5.5–8.5 (6.5)	pH 4.7–5.8 (5.3)	6.0–8.5 (7.0)
N sources	n.r.	NH_4^+, casamino acids	n.r.
Major fatty acids	16:0 (39.2 %), 16:1 (28.0 %), 14:0 (24.5 %), 15:0 (8.3 %)	16:0 (25.8 %), i17:0 (19.3 %), ai17:0 (13.5 %), cis-16:1D5 (8.8 %), cis-i17:1D5 (6.8 %), 5,6-methylene 16:0 (5.2 %)	i15:0 (30.9 %), i17:0 (19.5 %), 16:0 (17.1 %), cis-16:1D5 (11.3 %), i13:0 3-OH (5.8 %)
Primary quinones	MK-12	MK-8	MK-11, MK-10
Salt tolerance	0.5 % (w/v)	2.0 % (w/v)	> 1.0 % (w/v)
Catalase activity	-	+	+
Oxidase activity	-	-	-

[1] Tamaki et al. (2011); [2] Lee et al. (2011); [3] Stott et al. (2008); [4] Vyssotski et al. (2011) [d]; [5] Im et al. (2012); n.r., not reported.

[d] I am a co-author of this study.

Table 2.4 - Substrate specificity of cultivated *Armatimonadetes* strains.

Substrates		*Armatimonadetes* strains		
		***A. rosea* (YO-36 ^{T})[1]**	***C. calidirosea* (T49^{T})[2]**	***F. ginsengisoli* (Gsoil 348^{T})[3]**
Monosaccharides	D-arabinose	+	+	-
	D-xylose	-	+	-
	D- ribose	-	+	-
	D-fructose	-	+	n.d.
	D-glucose	-	+	-
	D-mannose	-	+	n.d.
	D-rhamnose	-	+	-
	D-galactose	-	+	n.d.
	D-N-acetylglucosamine	n.d.	+	-
	Sorbitol	-	+	n.d.
Di- & trisaccharides	Maltose	+	+	-
	Sucrose	+	+	-
	Gentiobiose	+	n.d.	n.d.
	Lactose	-	+	n.d.
	Trehalose	-	+	-
	Cellobiose	-	+	n.d.
	Raffinose	+	+	-
Polysaccharides	Gellan gum	+	+	n.d.
	Xanthan gum	+	+	n.d.
	Starch	-	+	n.d.
	Glycogen	-	+	-
	Dextrin	n.d.	+	-
	CMC	-	+	n.d.
	Galactomannan	n.d.	+	n.d.
	Chitin	n.d.	+	n.d.
	Pullulan	n.d.	+	n.d.
	Pectin	+	+	n.d.
Protein	Yeast extract	+	-	+
	Casamino acids	-	-	+
	Peptone	-	-	+

+, positive; -, negative; n.d., not determined

[1] Tamaki et al. (2011); [2] Lee et al. (2011); [3] Im et al. (2012).

Table 2.5 - Diagnostic differences in enzyme expression (bioMérieux API ZYM) by cultivated *Armatimonadetes* strains.

Enzyme assayed	*Armatimonadetes* strains		
	***A. rosea* (YO-36^T)[1]**	***C. calidirosea* (T49^T)[2]**	***F. ginsengisoli* (Gsoil 348^T)[3]**
Alkaline phosphatase	+	+	+
Esterase	+	+	w
Esterase-lipase	+	+	w
Lipase	-	-	-
Leucine arylamidase	+	-	+
Valine arylamidase	+	-	w
Cysteine arylamidase	-	-	-
Trypsin	+	-	w
α-chymotrypsin	+	-	+
Acid phosphatase	+	+	n.r.
Naphthol-AS-BI-phosphohydrolase	+	+	+
α-galactosidase	-	+	+
β -galactosidase	+	+	+
β -glucuronidase	-	-	w
α glucosidases	-	+	n.r.
β-glucosidase	+	+	+
N-acetyl- β -glucosaminidase	-	+	+
α -mannosidase	-	+	+
α-fucosidase	-	+	+

+, positive; -, negative; w, weakly positive; n.r., not reported

[1] Tamaki et al. (2011); [2] Lee et al. (2011); [3] Im et al. (2012).

Table 2.6 - Antibiotic sensitivities of *A. rosea* (YO-36^T) and *C. calidirosea* (T49^T).

Antibiotics tested (concentration tested)	*Armatimonadetes* strains [1]	
	***A. rosea* (YO-36 ^{T})[2]**	***C. calidirosea* (T49^T)[3]**
Ampicillin	+ (50 µg mL^{-1})	+ (10 µg mL^{-1})
Kanamycin	- (50 µg mL^{-1})	+ (10 µg mL^{-1})
Streptomycin	n.d.	+ (10 µg mL^{-1})
Polymyxin B	n.d.	+ (100 µg mL^{-1})
Trimethoprim	n.d.	+ (100 µg mL^{-1})
Metronidazole	n.d.	- (10 & 100 µg mL^{-1})
Lasalocid A	n.d.	- (10 & 100 µg mL^{-1})
Neomycin	n.d.	- (10 & 100 µg mL^{-1})
Rifampicin	+ (50 µg mL^{-1})	-[4] (10 & 100 µg mL^{-1})
Vancomycin	+ (50 µg mL^{-1})	n.d.
Chloramphenicol	+ (50 µg mL^{-1})	n.d.
Tetracycline	+ (50 µg mL^{-1})	n.d.

+, sensitive/growth inhibited; -, not sensitive/growth; n.d., not determined

[1] antibiotic sensitivities for *F. ginsengisoli* Gsoil 348^T have not been reported; [2] Tamaki et al. (2011); [3] Lee et al. (2011); [4] Unpublished data by Lee KC (2012).

2.6.1 - *Armatimonas*

Tamaki, Tanaka, Matsuzawa, Muramatsu, Meng, Hanada, Mori, and Kamagata 2011, 1446^{VP}

Armatimonas [Ar.ma.ti.mo'nas. L. adj. armatus armoured or armour-clad; L. fem. n. monas a unit; N.L. fem. n. *Armatimonas* an armour-clad unit, referring to the hard colonies]

Armatimonas is the type genus for the class *Armatimonadia.* The type strain *Armatimonas rosea* YO-36^T (=NBRC 105658^T =DSM 23562^T) was isolated from the rhizoplane of *Phragmites australis* (common reed) in a fresh water mesophilic lake in Yamanashi prefecture, Japan (Tamaki et al., 2011). The strain is negative for both catalase and oxidase activities. Morphologically, the cells are 1.4–1.8 µm wide and 2.4–3.2 µm long, and ovoid- to rod-shaped. After one week of incubation at 30 °C, strain YO-36^T forms smooth, hard, circular, pink colonies that are 1-2 mm in diameter on R2A agar plates. The strain is mesophilic with a growth temperature range of 20-40 °C (optimum at 30-35 °C),

and neutrophilic with a growth pH range of 5.5–8.5 (optimum pH 6.5). Growth occurs at 0-0.5 % NaCl (w/v). The cells are Gram-negative, non-motile, and no spore formation has been observed.

Strain YO-36^T is an aerobic chemoheterotroph able to utilise a limited range of substrates. The strain is able to utilise at D-arabinose (2 mM), raffinose (2 mM), maltose (2 mM), sucrose (10 mM), yeast extract (0.5 g L^{-1}), pectin (0.5 g L^{-1}) and gellan gum (0.5 g L^{-1}). Weak growth was also observed using gentiobiose (2 mM and 10 mM). No significant growth was observed at either 2 mM or 10 mM concentration of D-glucose, D-ribose, D-xylose, D-fructose, D-galactose, D-mannose, lactose, trehalose, L-rhamnose, turanose, melibiose, cellobiose, erythritol, mannitol, D-sorbitol, gelatin (0.5 g L^{-1}), glycogen (0.5 g L^{-1}), starch (0.5 g L^{-1}), agar (0.5 g L^{-1}), CM-cellulose (0.5 g L^{-1}), xylan (0.5 g L^{-1}), alginate (0.5 g l21), L-arginine, L-tyrosine, L-valine, L-ornithine, L-aspartate, L-histidine, L-phenyl- alanine, L-alanine, L-serine, glycine, L-leucine, L-proline, L-threonine, L-glutamine, L-cysteine, L-methionine, formate, acetate, propionate, butyrate, valerate, caproate, heptanoate, caprylate, pelargonate, caprate, laurate, pyruvate, L-malate, L-lactate, fumarate, succinate, methanol, ethanol, 1-propanol, glycerol, benzoate (1 mM and 10 mM), phenol (1 mM and 10 mM), a-ketovalerate, b-hydroxybutyrate, citrate, and casamino acids (0.5 g L^{-1}).

Using the bioMérieux API 20E, 20NE, and API ZYM systems, strain YO-36^T showed positive activities for alkaline phosphatase, esterase, esterase-lipase, leucine arylamidase, valine arylamidase, trypsin, chymotrypsin, acid phosphatase, naphthol-AS-BI-phosphohydrolase, β -galactosidase, β-glucosidase, and gelatinase. A comparison of the enzyme expression by all three *Armatimonadetes* isolates using the API ZYM assay kit is presented in Table 2.5, and a list of antibiotic sensitivities of YO-36^T is presented in Table 2.6.

The mol% G+C content is 62.4.

2.6.2 - *Chthonomonas*

Lee, Dunfield, Morgan, Crowe, Houghton, Vyssotski, Ryan, Lagutin, McDonald, and Stott 2011, 2487-2488^{VP}

Chthonomonas [Chtho.no.mo'nas. Gr. n. Chthōn Chthonos earth, soil, land; Gr. fem. n. monas a unit, monad; N.L. fem. n. Chthonomonas a unit (bacterium) from soil].

Chthonomonas is the type genus for the class *Chthonomonadetes.* The type strain *Chthonomonas calidirosea* T49^T (=ICMP 18418^T =DSM 23976^T) was isolated from geothermally-heated soil at Tikitere, New Zealand (Stott et al., 2008) and was characterised by Lee et al. (2011). The strain is positive for catalase and negative for oxidase activities. Morphologically, the cells are rod-shaped, 0.5-0.7 µm in width and 2.5-3.0 µm in length. The cells exhibit an irregular, corrugated outer membrane. Strain T49^T was grown on oligotrophic AOM1 plates (Stott et al., 2008) at 60 °C, and form circular, pink, and convex colonies in c. three days. The colonies darken in colour from pink to orange as the colonies age. Spectrographic analysis of the extracted pigment exhibited a similar spectral profile of *Thermomicrobium roseum*, a pink- pigmented thermophile of the phylum *Chloroflexi* (Wu, Raymond, et al., 2009).

The strain is thermophilic with a growth temperature range of 50-73 °C (optimum at 68 °C), and moderately acidophilic with a growth pH range of 4.7–5.8 (optimum pH 5.3). Growth occurs at 0-2.0 % NaCl (w/v). The cells are Gram-negative and non-spore forming. While strain T49^T is mostly non-motile, motile cells are occasionally observed (unpublished, Lee et al., 2011).

Strain T49^T is an aerobic chemoheterotroph able to utilise most mono- disaccharides and branched or amorphous polysaccharides. Using a defined minimal salts medium strain T49^T was able to utilise the following substrates (0.5 g L^{-1}): DL-arabinose, D-xylose, D-ribose, D-fructose, D-glucose, D-mannose, D-rhamnose, D-galactose, sucrose, lactose, maltose, trehalose, cellobiose, raffinose, D-N-acetylglucosamine, as well as polysaccharides including starch, glycogen, dextrin, CMC, xylan, galactomannan, pectin, gellan, xanthan, chitin and pullulan. Strain T49^T was unable to utilise D-galacturonic acid, sodium alginate, agarose, cotton, Whatman™ filter paper, Avicel™, crude lignin extract, non-commercial lignocellulosic pulp preparations, and alcohols including methanol, ethanol, 1-propanol, and 2-propanol. Strain T49^T utilises ammonium salts, yeast extract and casamino acids as assimilatory nitrogen sources, but not nitrate or dinitrogen gas. A summary of the enzyme expression of *C. calidirosea* T49^T using the API ZYM (bioMérieux) assay is presented in Table 2.5, and antibiotics sensitivity in Table 2.6.

The mol% G+C content is 54.6.

2.6.3 - *Fimbriimonas*

Im, Hu, Kim, Rhee, Meng, Lee, and Quan 2012, 315

Fimbriimonas [Fim.bri.i.mo'nas. L. pl. n. fimbriae, fibres, threads, fringe, and in biology, fimbriae; L. fem. n. monas, a unit, monad; N.L. fem. n. Fimbriimonas, fimbriae-shaped monad].

Fimbriimonas, the type genus for the class *Fimbriimonadia*. The type strain *Fimbriimonas ginsengisoli* Gsoil 348^T (= KACC 14959^T = JCM 17079^T) was isolated from a ginseng field soil sample in Pocheon province, South Korea (Im et al., 2012). The strain is positive for catalase and negative for oxidase activities. Morphologically, the cells are rod shaped and are 0.5–0.7 µm in diameter and 2.5–5.0 µm length. Strain Gsoil 348^T was grown on one-half strength R2A agar plates for two weeks at 30 °C, resulting in raised, ivory-pigmented, circular colonies with a greasy surface, and 1–2 mm in diameter. The colony morphology was also described as "highly mucoid". The strain is mesophilic with a growth temperature range of 15-30 °C (optimum at 30 °C), and neutrophilic with growth pH range at 6.0–8.5 (optimum of pH 7.0). Growth occurs without supplementary NaCl, and inhibition occurs at 1.0 % (w/v). The cells are Gram-negative and exhibit an abundance of fine peritrichous fibrils. No spore formation or cell motility was observed.

Substrate utilisation was determined via growth with defined basal salt medium (Im et al., 2012), substrate concentration 0.05 % (w/v), at 30 °C for 10 days. Strain Gsoil 348^T has a very limited substrate range which it can utilise. The strain showed positive utilisation for peptone, casamino acid, and yeast extract. The strain showed no significant growth for L-rhamnose, propionate, D-xylose, D-fucose, D-glucose, ethanol, D-arabinose, L-arabinose, D-ribose, L-xylose, N-acetyl-glucosamine, pyruvic acid, formic acid, 3-hydroxybutyrate, valerate, caprate, maleic acid, fumaric acid, L-sorbose, phenyl acetate, benzoic acid, 3-hydroxybenzoate, 4-hydroxybenzoate, salicin, citrate, lactate, succinic acid, suberate, D-cellobiose, D-lactose, D-maltose, D-melibiose, D-sucrose, D-trehalose, D-raffinose, gluconate, D-dulcitol, inositol, D-mannitol, xylitol, amygdalin, glycerol, methanol, glycogen, inulin, dextran, and L-amino acids. Enzyme expression assays using API ZYM (bioMérieux) for *F. ginsengisoli* is presented in Table 2.5.

The mol% G+C content is 61.4.

2.7 - Isolation, enrichment and maintenance procedures

2.7.1 - *Armatimonas rosea* YO-36^T

Armatimonas rosea YO-36^T was isolated from the root surface of the plant *Phragmites australis* (Common reed) using DTS plate medium, which contains 0.17 g L^{-1} Bacto

tryptone (Difco), 0.03 g L^{-1} Bacto soytone (Difco), 0.025 g L^{-1} glucose, 0.05 g L^{-1} NaCl and 0.025 g L^{-1} K_2HPO_4, at pH 7.0. The soil from the reed root was washed with sterile DTS medium to remove soil and the roots homogenised. The homogenates were then diluted 1:10 in DTS and 50 µL of suspension was inoculated solid medium and incubated in the dark at 25 °C for 30 days (Tamaki et al., 2011; Tanaka et al., 2012). The isolate can be maintained with on R2A (Difco) agar plates. Colonies develop at 30 °C after one week of incubation.

The strain is oligotrophic and is sensitive to nutrient-rich media such as Trypticase Soy Agar (TSA) and Lysogeny Broth (LB) medium, but exhibits good growth with diluted media, i.e. 1:10 strength LB, 1:100 strength TSA, and 1:10 strength R2A medium, confirming its oligotrophic phenotype.

No information is currently available on the preservation of strain YO-36^T. However lyophilisation is the viable long term preservation method used by the DSMZ and NBRC.

2.7.2 - *Chthonomonas calidirosea* T49^T

Chthonomonas calidirosea T49^T was isolated from steam-affected geothermal soil in Tikitere, New Zealand (Stott et al., 2008). The sample was collected at a depth of around 15 cm, had an *in situ* temperature of approximately 55 °C, and a pH of 4.3. Isolation of strain T49^T was achieved by streaking collected soil samples on oligotrophic solid FS3V medium (pH 4.5 and 10 % v/v CO_2 headspace). AOM1 solidified medium (10 % CO_2 headspace) also facilitated the enrichment and isolation of other *C. calidirosea* strains such as P488 (Stott et al., 2008). The plates were incubated in sealed jars at 60 °C over the duration of four weeks. Colonies formed during this period were sub-cultured onto new medium until pure cultures were achieved. Isolate identity and purity were confirmed via PCR targeting 16S rRNA gene sequences. The gellan gum served as both a gelling agent (replacing agar) as well as a carbon source for strain T49^T.

FS3V medium contains (L^{-1}): 0.4 g NH_4Cl, 0.1 g KH_2PO_4, 0.02 g $CaCl_2 \cdot 6H_2O$, 1.0 g $MgCl_2 \cdot 6H_2O$ and 0.04 g $MgSO_4 \cdot 7H_2O$, 15 g gellan gum (Phytagel ™, P8169 – Sigma), 3 ml of FeEDTA solution (see below), 3 ml of trace elements solution 1 (see below), 1 ml of trace elements solution 2 (see below), and 100 mg of yeast extract. To make the solidified medium, all minimal medium salts and metal solutions minus the gellan gum and yeast extract are combined into 500 ml of distilled H_2O. The acidity of the medium is adjusted to pH 5.0 using 1 M H_2SO_4. The use of HCl instead of H_2SO_4 is not advised. The gellan gum is added to a separate 500-ml dH_2O volume with no pH adjustment. Both

media are then sterilised at 121 ºC, 100 kPa, for 20 minutes. The media are then combined and mixed once cooled to c. 60 ºC. The yeast extract is then passed through a sterile 0.2-µm filter into the FS3V medium, and the plates are poured immediately.

AOM1 medium plates are made by combining (L^{-1}): 4 g $NH_4SO_4 \cdot 7H_2O$, 0.25 g $NaHCO_3$, 5 mg $CaCl_2 \cdot 6H_2O$, 0.05 g KH_2PO_4, 0.666 g $MgSO_4 \cdot 7H_2O$, 100 mg yeast extract, 100 mg vitamins (see below), 1 ml FeEDTA solution (see below), 1 ml trace elements solution 2 (see below), and 15 g gellan gum (Phytagel – Sigma). The pH of the medium is adjusted to pH 6.5 (H_2SO_4) and sterilised at 121 ºC, 100 kPa, for 20 minutes. 10 mg of yeast extract and 10 mg of B vitamins (see below) then filtered through a sterile 0.2-µm filter into the medium prior to pouring.

Recent observations have noted the vitamin mixture is *not* required for successful cultivation. However, the initial isolation media was supplemented with a commercial vitamin mixture containing: (per 100 mg): 1 mg folic acid, 8 mg vitamin B1, 4 mg vitamin B2, 1 mg niacin, 10 mg niacinamide, 15 mg pantothenate, 15 mg pyridoxine, 5 mg cobalamin, 5 mg biotin, 15 mg choline, 15 mg inositol, and 6 mg para-amino benzoic acid. The FeEDTA solution contains (L^{-1}): 1.54 g $FeSO_4 \cdot 7H_2O$, 2.06 g Na_2EDTA. The trace element solution 1 contains (L^{-1}): 0.44 g $ZnSO_4 \cdot 7H_2O$, 0.20 g $CuSO_4 \cdot 5H_2O$, 0.19 g $MnCl_4 \cdot H_2O$, 0.06 g $Na_2MoO_4 \cdot 2H_2O$, 0.1 g H_3BO_3, 0.08 g $CoCl_2 \cdot 6H_2O$. The trace elements solution 2 contains (L^{-1}): 1.5 g nitrilotriacetic acid, 0.2 g $Fe(NH_4)_2(SO_4)_2 \cdot 6H_2O$, 0.2g Na_2SeO_3, 0.1 g $CoCl_2 \cdot 6H_2O$, 0.1 g $MnSO_4 \cdot 2H_2O$, 0.1 g $Na_2MoO_4 \cdot 2H_2O$, 0.1 g $Na_2WO_4 \cdot 2H_2O$, 0.1 g $ZnSO_4 \cdot 7H_2O$, 0.04 g $AlCl_3 \cdot 6H_2O$, 0.025 g $NiCl_2 \cdot 6H_2O$, 0.01 g H_3BO_3, 0.01 g $CuSO_4 \cdot 5H_2O$. No pH adjustments were made to the media and they are stored at 4 ºC without sterilisation.

Strain T49^T can be regularly cultured with gellan-solidified AOM1. Weekly subculture is advised. Colonies form pits in the solidified medium approximately one week as the gellan gum is consumed. Successful subculturing requires a large amount of inoculum compared to average bacterial cultures. Homogenising and suspension of cell pellets collected from plate culture also improves subsequent subculturing yield. The strain can also be maintained in liquid medium where the gellan gum concentration is reduced to 1 g L^{-1} and is used only a carbon and energy source. Strain T49^T cell pellets can be cryopreserved by suspending the cell pellets in 5 % DMSO solution and kept at -80 °C (unpublished, Lee et al., 2011).

2.7.3 - *Fimbriimonas ginsengisoli* Gsoil 348^T

Fimbriimonas ginsengisoli Gsoil 348^T was isolated from a soil sample obtained from a ginseng field (Im et al., 2012). The sample was suspended and diluted in 50 mM phosphate buffer (pH 7.0) and then spread on 1:5 strength R2A agar plates, containing (L^{-1}): 0.25 g tryptone, 0.25 g peptone, 0.25 g yeast extract, 0.125 g malt extract, 0.125 g beef extract, 0.25 g casamino acid, 0.25 g soytone, 0.5 g dextrose, 0.3 g soluble starch, 0.2 g xylan, 0.3 g sodium pyruvate, 0.3 g K_2HPO_4, 0.05 g $MgSO_4 \cdot 7H_2O$, 0.05 g $CaCl_2$ and 15 g agar. The plates were incubated at room temperature over the period of two months. During this period, microcolonies formed were subcultured on R2A or 1:2 strength R2A plates and further incubated at room temperature. Strain Gsoil 348^T can be preserved in 20 % glycerol solution (w/v) at -70 °C.

2.8 - Ecology

Because of the lack of cultivated representatives, information on the ecology of *Armatimonadetes* strains can only be reasonably estimated via the analysis of the ecosystems in which *Armatimonadetes* / OP10 strains and phylotypes have been detected. To date, publications on the three cultivated strains are limited to taxonomic descriptions and generally have only noted the environment from which the *Armatimonadetes* representatives have been obtained. Two of the current *Armatimonadetes* isolates, *A. rosea* YO-36^T (Tamaki et al., 2011) and *F. ginsengisoli* Gsoil 348^T(Im et al., 2012), were isolated from similar mesophilic soil environments with close associations with plant root systems. Thermophilic representatives, *C. calidirosea* T49^T (Lee et al., 2011) and the closely-related strain P488 (Stott et al., 2008) were isolated from geothermally-affected soils rich in organic materials. Both YO-36^T and T49^T are reported to have a carbohydrate-based metabolism, with T49^T also exhibiting an ability to hydrolyse a reasonably broad range of C5 and C6 oligosaccharides and also amorphous polysaccharides. In contrast, Gsoil 348^T is reported to have a very restricted and obligately proteolytic-based metabolism. Interestingly, both YO-36^T and T49^T also have an obligate requirement for trace concentrations of yeast extract or casamino acids for growth, with Lee et al. (2011) suggesting that this may represent a limited ability to synthesise amino acids. Based on these limited phenotypic data, we can only reasonably note that all the currently available *Armatimonadetes* strains have a soil-based ecological niche and are aerobic oligotrophs sensitive to nutrient-rich culture media. Relatively dilute media are required for cultivation. The phenotypic differences between the three isolates noted above are not surprising considering the 16S rRNA gene sequence similarity of < 81 % (Table 2.1), and thus it is anticipated that their individual ecological function would also be substantially different.

A broad range of ecological function is also expected to extend across the phylum. The phylum contains ca. 500 distinct high-quality phylotypes, based on the ARB-SILVA database, SSU Ref 111 NR (Pruesse et al., 2007). By collating the source environment information associated with OP10 or *Armatimonadetes* phylotypes, Dunfield and colleagues (2012) reported that the most dominant source environment was temperate soils (37 %), followed by skin swabs clones (17 %) from the Human Skin Microbiome (Grice et al., 2009) and the Mouse Wound Microbiota projects (Grice et al., 2010), aerobic and anaerobic bioreactors and waste treatment facilities (12 %), and fresh- waters and freshwater sediments (7 %). Surprisingly, clones from geothermal environments only contribute around 6 % of the total *Armatimonadetes* / OP10 clones. These proportions of phylotype niches may not be directly representative of the distribution of *Armatimonadetes* in the environment, as extensive studies of specific environments (e.g., skin microbiome) can contribute large amounts of data into the database, thus skewing the environmental distribution through sampling bias. The issue has been partially addressed by the non-redundant sequence filter in ARB-SILVA (Pruesse et al., 2007), which removed redundant sequences based a 98 % similarity criterion. Nevertheless, studies with larger sampling depth will have their rare diversity overrepresented. These rare phylotypes may represent genuinely rare members of the communities, or they may represent dormant or contaminating species removed from their natural habitat. This may explain the presence of *Armatimonadetes* phylotypes associated with skin and wound microbiome projects as the number of phylotypes associated with *Armatimonadetes* / OP10 is very low within the total dataset of these projects (< 0.05 % of the total), suggesting a random or contaminating presence perhaps via soil or dust (Dunfield et al., 2012). Conversely, several studies have identified environments where *Armatimonadetes* appear to dominate the microbial communities. *Armatimonadetes* / OP10 16S rRNA gene sequences were reported as the dominant phylotypes (50 %, 50 % and 39 % of the populations) in an c.a. 50-57 °C green bacterial mat (Portillo et al., 2009), in a metal-rich freshwater reservoir (Stein et al., 2002) and a lab-scale wastewater batch bioreactor (Crocetti et al., 2002) respectively.

An examination of the broad range of optimal temperature and pH conditions of the cultured species (Figure 2.2) gives some indication of the diverse environmental conditions at which *Armatimonadetes* are detected. *Armatimonadetes* phylotypes have been detected across low- to high-temperature environments. The maximum temperature in which *Armatimonadetes* have been detected is c.a. 79 °C at Obsidian Pool (Hugenholtz, Pitulle, et al., 1998), but they are also detected in arctic and alpine environments such as high elevation Andean soils at -5 °C (Costello et al., 2009) and at -3 °C in alpine tundra soils (Nemergut et al., 2008). With only one exception,

Armatimonadetes appear to have a neutrophilic pH requirement with c.a. pH 4.5 as the lower limit in various soil ecosystems (Faoro et al., 2010; Lee et al., 2011), and pH 8.6 as the upper limit in a 72 °C geothermal sediment sample (Tomova et al., 2010) and pH 8.5 lab-scale wastewater batch bioreactor (Crocetti et al., 2002). The single pH exception was a single clone (accession number: FN870192) detected in the sediment of acidic mine tailings with a bulk soil pH of c.a. pH 2.5 (Lu et al., 2010), and may represent an acidophilic variant of *Armatimonadetes.* However, considering that this was a single phylotype in a sample containing known sulfate-reducing bacteria (SRB), it is more likely that this phylotype was contained within a soil micro-niche with an elevated pH generated via SRB activity, or a non-resident organism in that habitat.

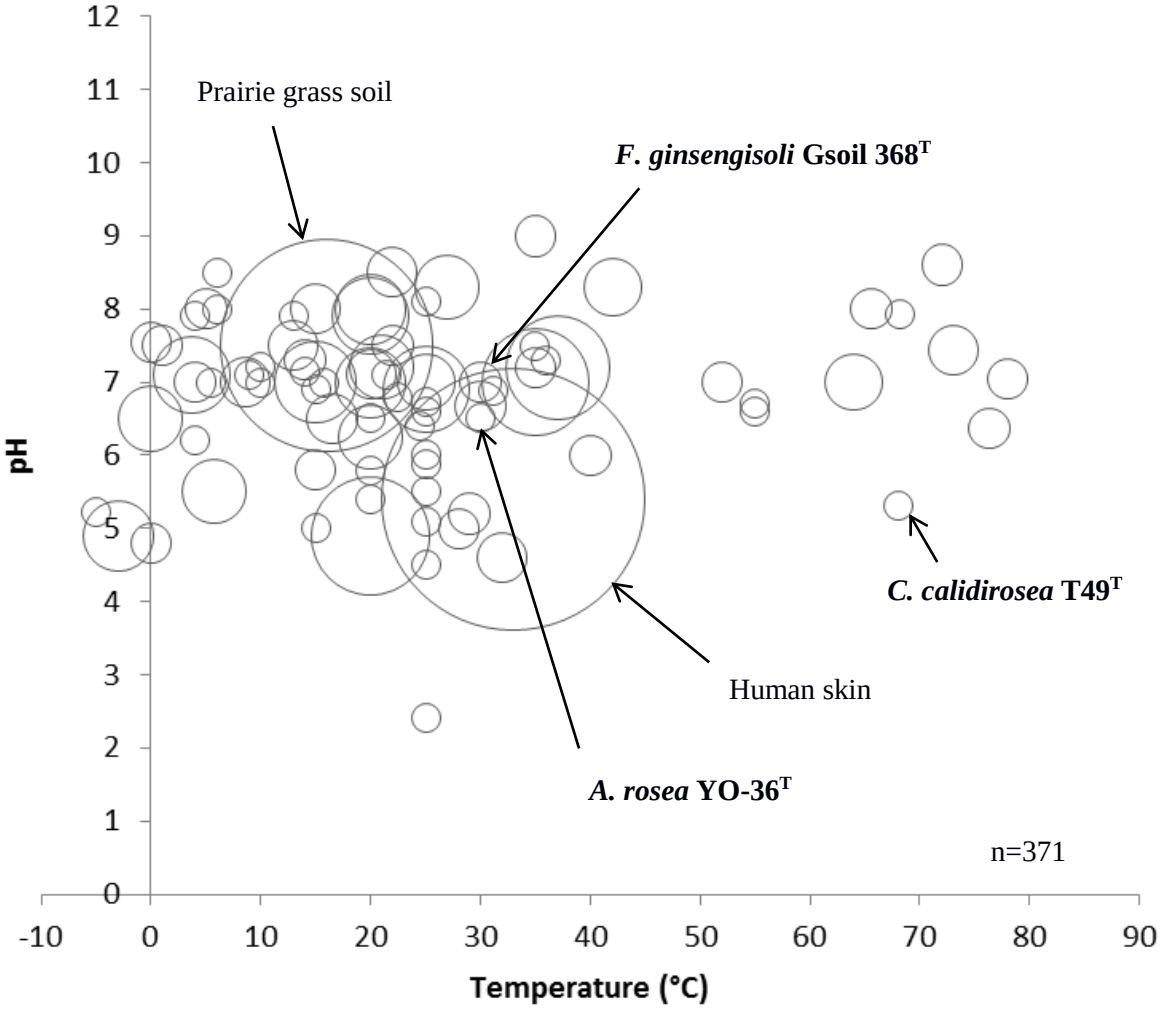


Figure 2.2 - Distribution of *Armatimonadetes* phylotypes as a function of environmental temperature and pH. Metadata for this figure were generated from publications linked to the phylotype accession numbers presented in Dunfield et al. (2012). We were able to determine niche pH and/or temperature from c.a. 75 % of the phylotypes. The large data sets such as the Human skin microbiome project (Grice et al., 2009) and a Prairie soil diversity study (Elshahed et al., 2008) are indicated, along with the three current *Armatimonadetes* isolates. The isolates data points are based on their reported optimal pH and temperature growth conditions. The relative size of the bubbles reflects the abundance of phylotypes detected at the individual niche temperature and pH.

Clustering of phylotypes originating from similar environments on a phylogenetic tree demonstrates the broad diversity of niche environments in which *Armatimonadetes* / OP10 phylotypes have been detected. Figure 2.3 broadly represents the niche diversity of *Armatimonadetes* / OP10 phylotypes based upon the data presented by Dunfield et al. (2012). Most of the groups within, including those reflecting the novel classes (Group 1 - *Armatimonadia*, Group 3 - *Chthonomonadetes*, and Group 9 - *Fimbriimonadia*) showed no single environmental variable the individual clades could be strongly associated with. On the other hand, groups such as Group 10 and Group 12 were apparently present predominantly in geothermal environments and anaerobic environments, respectively. Group 12 in particular were exclusively anaerobic sediments, saturated soils or anaerobic wastewater treatment facilities and bioreactors, which suggests that an anaerobic metabolism is a conserved trait within this group.

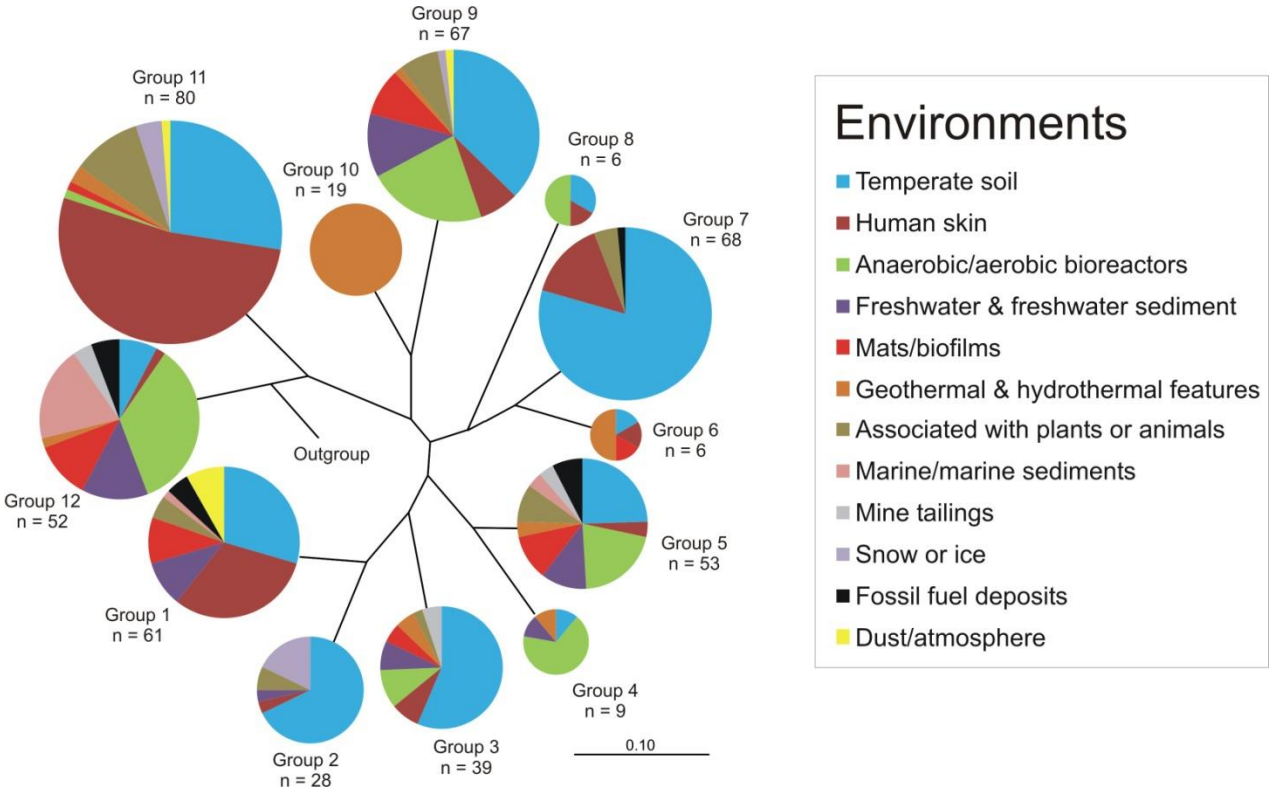


Figure 2.3 - Maximum likelihood (ML) 16S rRNA gene-based phylogenetic tree showing class-level groupings and associated niche environment distributions of *Armatimonadetes*. Environmental niches have been adapted from Dunfield et al. (2012), and pie charts are scaled relative to phylotype abundance of each grouping. Individual phylotypes have been colour-coded according to environmental niche. *Armatimonas rosea* YO-36^{T} places in Group 1 (class *Armatimonadia*), *Chthonomonas calidirosea* T49^{T} places in Group 3 (class *Chthonomonadetes*), and *Fimbriimonas ginsengisoli* Gsoil 348^{T} (class *Fimbriimonadia*) places in Group 9. The phylogenetic tree was generated using 490 near-full length sequences. The branch lengths are indicative of the distance to the deepest branching phylotypes to the individual groups. The scale bar represents 0.1 changes per nucleotide position. The complete phylogenetic tree, including the accession numbers of phylotypes used can be found in the supporting information section of Dunfield et al. (2012).

While a unifying trait(s) of *Armatimonadetes* has yet to be established, it is tempting to speculate that *Armatimonadetes* strains are broadly involved in the degradation of plant materials, polysaccharide-based substances or photosynthetic biomass. Many of the *Armatimonadetes* phylotypes are detected in soils, associated with plants or involved in the degradation of plant materials either in bioreactors (Wang et al., 2010) or ruminants (Brulc et al., 2009). In addition, *Armatimonadetes* are commonly detected, often in high abundance, in areas predominated by photosynthetic bacteria or eukarya, including stromatolites (Burns et al. 2004), hypersaline photosynthetic mats (Isenbarger et al., 2008; Ley et al., 2006), photosynthetic hypolithic communities (de la Torre et al., 2003; Wong et al., 2010), geothermal spring mats (Portillo et al., 2009; Stott et al., 2008), macrophyte-dominated eutrophic sediment (Shao et al., 2011), and freshwater cyanobacterial blooms (Pope & Patel, 2008). Some role in scavenging products of photosynthetic organisms therefore seems probable, but cultivation of a broader range of *Armatimonadetes* strains and more focused ecological studies are needed to verify this speculation.

2.9 - Pathogenicity, clinical relevance

To date, there are no studies available relating *Armatimonadetes* clones with pathogenicity in plants or animals. Aside from 16S rRNA gene sequence clones associated with the skin and wound microbiome projects (Grice et al., 2009, 2010), the presence of which was discussed in the previous section, few clones have any direct relationship with animals. Out of c.a. 500 phylotypes, only few sequences are associated with GI tract in ruminant (Brulc et al., 2009; Ley et al., 2008), catfish (Wu et al., 2010), or larval gut (accession number: EU344940). Some clones were also found in proximity of plant roots and leaves (Chelius & Triplett, 2001; Delmotte et al., 2009; Im et al., 2012; Tanaka et al., 2012), sponges (Gernert et al., 2005), and deep sea corals (accession number: DQ395456). An *Armatimonadetes* phylotype has also been detected in the synovial fluid of some arthritis patients (Siala et al., 2009), although these are probably stray contaminant or opportunistic colonisers rather than causative agents.

Chapter 3
Research Aims, Hypotheses, and Overview

3.1 - Research aims

The isolation of novel *Armatimonadetes* species represented a valuable first glimpse into a globally-distributed, deeply-branching lineage (Dunfield et al., 2012) within the tree of life. Overall, the aim of this thesis is to gain the first phylogenetic, genetic, and ecological insights to the poorly-understood phylum *Armatimonadetes*, through the representative species *Chthonomonas calidirosea*. From this starting point, and based on the knowledge gaps outlined in Introduction (Chapter 1) and Literature Review (Chapter 2), several subjects presented themselves as critical in understanding this new phylum, as well as providing valuable foundations for future investigations.

1. **Resolving the internal and external phylogenetic structures of *Armatimonadetes*** – As outlined in literature review (Section 2.4 -Phylogenetic structure of the phylum), previous publications involving Candidate Division OP10 or *Armatimonadetes* have presented incompatible and partial phylogenetic descriptions due to the selection of phylotypes and phylogenetic methods. While the taxonomy of the phylum and member classes have been established based on the three type strains, the phylogeny of the phylum, which included the type strains and the environmental phylotypes (based on 16S rRNA gene sequences) remained poorly-defined. Rigorous elucidation of the phylogeny of *Armatimonadetes* is important as it provides both the basis to connect previous findings and the framework to accurately associate future phylotypes (to avoid misidentification and taxa "boundary creep"). To address this issue, I conducted analyses specifically targeting the phylogeny of *Armatimonadetes* and associated taxonomic groupings at class and phylum-level. The goal of this investigation was to resolve the internal and external phylogenetic structure of the phylum, i.e. identifying *Armatimonadetes*-associated clades and resolve their relationships. Within the phylum, the groups included the three described classes (*Armatimonadia*, *Fimbriimonadia*, and *Chthonomonadetes*), as well as clades currently without representative isolates. External phylogeny to be assessed included identifying a suitable boundary between *Armatimonadetes* and neighbouring clades that should be excluded from the phylum. The overall analysis methodology aims to address issues from previous publications, by using a comprehensive phylotype dataset and utilising multiple phylogenetic methods to validate the identification of reliable monophyletic clades, and the elucidation of their relationships.

2. **Genomic analysis of a representative of *Armatimonadetes*** – As highlighted in the literature, no genetic information regarding *Armatimonadetes* was publicly available beyond environmental 16S rRNA gene sequences (see Section 2.5 - Molecular analyses). This represented a significant knowledge gap within our understanding of *Armatimonadetes.* Genomic data is the blueprint within the "Central Dogma" of molecular biology (Crick, 1970) from which phenotypes are derived. Genomic analysis therefore, provides a valuable perspective on the nature of an organism. *C. calidirosea* as one of two then-recognised *Armatimonadetes* species and type species of class *Chthonomonadetes*, presented itself as a suitable representative species of the genetic diversity of *Armatimonadetes*. Previous characterisation (Lee et al., 2011) identified a variety of physiological traits, including a narrow pH growth range and a diverse range of soluble carbohydrates as the primary carbon and energy source for *C. calidirosea* T49^{T}. These phenotypes, as well as the physicochemistry of the particular geothermal environments, such as the presence of degraded cellulosic materials, may influence the ecological niche of this bacterium. Therefore, the identification of the genetic basis of these phenotypic observations contributes to a better inference of the ecology of *C. calidirosea*. In addition, phylogeny based on 16S rRNA genes and characteristics of pigment spectrophotometry (Lee et al., 2011) suggests *Armatimonadetes* are closely related to *Chloroflexi*. With the whole genomic sequence available, phylogenomic analysis can be conducted to infer a robust phylogeny for *Armatimonadetes.*

3. **Comparative analysis of *C. calidirosea* genomes, environmental physicochemistry and the community structure of the host environments** – Thermophilic prokaryotes are generally under-investigated compared to mesophilic species due to limitations in methodologies, remoteness of natural habitat, and historical precedence. Similarly, the relationships between the thermophiles and their environment also remained poorly-understood. *C. calidirosea* isolates have been found in several distinct geothermal sites within the TVZ (e.g., Lee et al., 2011; Stott et al., 2008). The distinct and discontinuous nature of the geothermal habitats is thought to be an ideal system for investigating biogeographical patterns (Reno et al., 2009). It raises questions regarding the drivers (environmental factors versus dispersal) influencing the genetic, phylogenetic, and physiological characteristics of the *C. calidirosea* isolates. Previous population genomic studies (Reno et al., 2009; Whitaker et al., 2003) on the thermophilic archaeon *Sulfolobus islandicus* observed

biogeographical patterns among the genomes due to dispersal constraints, even within localised geographical scale comparable to that between *C. calidirosea* sample sites. This research may provide a better understanding of the ecology of *C. calidirosea*, and by extension, microbial ecology of the genome evolution of thermophilic bacteria.

3.2 - Hypotheses

This chapter has reviewed the state of published research into phylum *Armatimonadetes*, as well as the outstanding knowledge gaps which will be addressed by the work in this thesis, as outlined in Section 3.1 -Research aims, set out to address. As stated in this literature review, much of the data generated regarding this phylum have few external points of reference, since most *Armatimonadetes* phylotypes are only associated with cultured bacteria at the highest taxonomic level. Therefore, explorative research into this new territory is both inevitable and desirable. As the research in this thesis was based upon prior physiological characterisation of *C. calidirosea* T49^T (Lee et al., 2011) as well as pre-existing 16S rRNA gene sequences, it aimed to address the following hypotheses:

1. **Phylogenetic structure of *Armatimonadetes* (Chapter 4)** – Due to the lack of characterised species and internal taxonomy, Candidate Division OP10 has accumulated environmental 16S rRNA gene sequences attributed to the group. As the group became a recognised phylum (*Armatimonadetes*) the lack of a thorough phylogenetic investigation has resulted in ambiguity in the boundaries and relationships of taxonomic groups described. In order to resolve these issues, so that the identities of *Armatimonadetes*-related lineages may be reliably related between studies. I conducted a study of phylotypes attributed to *Armatimonadetes* and identified phylogenetic relationships based on a representative dataset as well as multiple methodologies. For the outcome of this study, I hypothesised that these phylotypes in fact consisted of phylum *Armatimonadetes* and additional lineages. In order to test this hypothesis, I utilised multiple phylogenetic methods to interrogate a comprehensive set of phylotypes attributed to Candidate Division OP10/*Armatimonadetes* to identify phylogenetic clades internal and external to the phylum.

2. **Genomic analysis of *C. calidirosea* T49^T (Chapter 5)** – The genomic sequencing of *C. calidirosea* T49^T represented the first publicly available genome within *Armatimonadetes*. Based on previous literature information, two hypotheses were formulated.

a. Previous phylogenetic analyses based on 16S rRNA genes concluded that *Armatimonadetes* may be most closely related to either *Actinobacteria* (Portillo et al., 2008), *Elusimicrobia* (Yarza et al., 2010), or *Chloroflexi* (Tamaki et al., 2011). Interestingly, *C. calidirosea* T49^T exhibited pigments with spectral characteristics similar to that of the *Chloroflexi* species *Thermomicrobium roseum* (Lee et al., 2011). In order to better resolve the deeply-branching relationships of *Armatimonadetes* with neighbouring phyla, I conducted phylogenetic analysis using concatenated amino acid sequences of conserved proteins from the genomes of a broad representative distribution of prokaryotic species. For the outcome of this study, I hypothesised that *Armatimonadetes*, as represented by *C. calidirosea* T49^T, is most closely related to *Chloroflexi*.

b. Previous physiological analysis has shown *C. calidirosea* T49^T is capable of utilising a wide range of soluble carbohydrates (Lee et al., 2011). In order to determine the genetic capability of the bacterium, I conducted sequencing and analysis of the genome, and assigned annotated genes into relevant biochemical pathways. For this study, I hypothesised that the genome of *C. calidirosea* T49^T reflects the phenotypic traits with a diverse distribution of carbohydrate metabolic pathways.

3. **Comparative analysis of *C. calidirosea* genomes, environmental physicochemistry and the community structure of the host environments (Chapter 6)** – Due to the disorganisation (lack of operonic structures) of the strain T49^T genome, I hypothesise that *C. calidirosea* as a species would exhibit a high degree of genome divergence from the close coupling of the metabolism with the immediate environment. Therefore, I sequenced the genomes of three additional *C. calidirosea* isolates, each isolate originated from a geographically distinct geothermal site within the TVZ, and comparatively analysed the genome as well as the biotic and abiotic factors between the environments.

3.3 - Co-authorship forms

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 858 5096
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Phylum Armatimonadetes

Nature of contribution by PhD candidate	Literature review, writing of manuscript
Extent of contribution by PhD candidate (%)	90

CO-AUTHORS

Name	Nature of Contribution
Matthew B. Stott	manuscript editing, discussion
Peter F. Dunfield	manuscript editing, discussion

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

Name	Signature	Date
Matthew B. Stott		8 Oct, 2014
Peter F. Dunfield		09 Oct 2014

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 858 5096
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work. Phylogenetic Delineation of the Novel Phylum *Armatimonadetes* (Former Candidate Division OP10) and Definition of Two Novel Candidate Divisions

Nature of contribution by PhD candidate	Experiment design, conducting analysis, and writing of manuscript
Extent of contribution by PhD candidate (%)	90

CO-AUTHORS

Name	Nature of Contribution
Craig W. Herbold	Experiment design, data analysis, discussion
Peter D. Dunfield	Experiment design, discussion
Xochitl C. Morgan	Experiment design, manuscript editing, discussions
Ian R. McDonald	Experiment design, manuscript editing, discussions
Matthew B. Stott	Experiment design, manuscript editing, discussions

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

Name	Signature	Date
Craig W. Herbold		14 Oct 2014
Peter D. Dunfield		09 Nov 2014
Xochitl C. Morgan		7 Oct 2014
Ian R. McDonald		14 Oct 2014
Matthew B. Stott		8 Oct 2014

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 858 5096
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work. Genomic analysis of *Chthonomonas calidirosea*, the first sequenced isolate of the phylum *Armatimonadetes*

Nature of contribution by PhD candidate	Experiment design, conducting experiments, analysis, and writing of manuscript
Extent of contribution by PhD candidate (%)	85

CO-AUTHORS

Name	Nature of Contribution
Xochitl C. Morgan	Experiment design, genome assembly, manuscript editing, discussion
Peter F. Dunfield	Alien_Hunter HGT, NCBI genome sample size bias analysis, manuscript editing, discussion
Ivica Tamas	Alien_Hunter HGT, NCBI genome sample size bias analysis, discussion
Ian R. McDonald	Experiment design, manuscript editing, discussion
Matthew B. Stott	Experiment design, analysis of glycoside hydrolases, manuscript editing, discussion

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

Name	Signature	Date
Xochitl C. Morgan	Xochitl Morgan	7 Oct 2014
Peter F. Dunfield		09 Nov 204
Ivica Tamas	Ivica Tamas	09 Nov 2014
Ian R. McDonald	Ian R McDonald	14 Oct 2014
Matthew B. Stott		8 Oct. 2014

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 858 5096
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Comparative genomics and metagenomics of geographically diverse *Chthonomonas calidirosea* isolates

Nature of contribution by PhD candidate	Experiment design, conducting experiments, analysis, and writing of manuscript
Extent of contribution by PhD candidate (%)	90

CO-AUTHORS

Name	Nature of Contribution
Matthew B. Stott	Experiment design, manuscript editing, discussions
Peter F. Dunfield	Experiment design, manuscript editing, discussions
Curtis Huttenhower	Experiment design, manuscript editing, discussions
Ian R. McDonald	Experiment design, manuscript editing, discussions
Xochitl C. Morgan	Experiment design, genome sequencing, manuscript editing, discussions

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- in cases where the PhD candidate was the lead author of the work that the candidate wrote the text.

Name	Signature	Date
Matthew B. Stott		8 Oct 2014
Peter F. Dunfield		09 Nov 2014
Curtis Huttenhower		7 Oct 2014
Ian R. McDonald		14 Oct 2014
Xochitl C. Morgan		7 Oct 2014

Chapter 4

Phylogenetic Delineation of the Novel Phylum *Armatimonadetes* (Former Candidate Division OP10) and Definition of Two Novel Candidate Divisions

Lee, K.C.Y.[1,2#], Herbold, C.[2], Dunfield, P.F.[3], Morgan, X.C.[1,4], McDonald, I.R.[2], and Stott, M.B.[1]

[1] GNS Science, Extremophile Research Group, Private Bag 2000, Taupō 3352, New Zealand.

[2] School of Biological Sciences, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.

[3] Department of Biological Sciences, University of Calgary, 2500 University Dr. NW, Calgary, T2N 1N4 Canada.

[4] Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston MA 02115, USA

[#] Corresponding author: E-mail: k.lee@gns.cri.nz

4.1 - Preface

This chapter documents the phylogenetic and taxonomic classification of the phylum *Armatimonadetes*. Portions of this chapter were subsequently published in the journal *Applied and Environmental Microbiology*. The published short-form version of this manuscript is included within the appendices (Section 8.1.1) along with supplementary information, while the original longer-form version of the research has been used for this chapter. Supplementary tables too large for printed form are referenced to digital files attached to this thesis.

This research was conducted due to the need for a robust phylogenetic framework to relate newly-described taxa (classes *Armatimonadia*, *Fimbriimonadia*, and *Chthonomonadetes*) within the equally-novel phylum *Armatimonadetes*. Prior to this publication, previous works represented the phylum with limited and inconsistent selections of phylotypes and with various phylogenetic inference methods. Due to this variance in methodologies, the descriptions of clades within the phylum, the boundaries between the phylum and neighbouring taxa (candidate divisions and phyla), and the relationships between the classes in pre-existing literature were conflicting and/or ambiguous (see Section 2.4 -Phylogenetic structure of the phylum).

To address these issues, we conducted a critical assessment of the phylogeny of *Armatimonadetes* utilising the most extensive collection of high-quality phylotype sequences attributed to the phylum then available. This allowed us to relate some of the previously-identified phylotype groupings to those shown in this publication through the overlaps of phylotypes used (i.e. a grouping in this publication may contain all the phylotypes used in a previously-identified grouping, thus suggesting congruence of the identity of the clade). A single phylogenetically diverse outgroup containing 46 phylotypes from 13 bacterial phyla, based on those used by Dalevi et al. (2001), was selected to avoid the influence of biased outgroup selection on the resulting phylogenetic inference. In order to extensively assess the phylogenetic relationships of phylotypes, we integrating tree nodal support values from multiple inference paradigms including neighbour-joining, maximum likelihood, and Bayesian phylogenetic inference. During this process, we identified previously-hidden chimeric sequences due to their inconsistent phylogenetic signals between methods.

Through these analyses, ten well-supported class-level clades within *Armatimonadetes* were identified, leaving one poorly-supported class-level clade (Group 5), which contained no obvious monophyletic structure. Based on the internal clades and sequence divergence values in established taxa as a guideline (Yarza et al., 2010), we delineated a monophyletic boundary of the phylum *Armatimonadetes*, and identified two neighbouring candidate divisions previously often misattributed to *Armatimonadetes*/OP10. The hypothesis of identifying *Armatimonadetes* as well as additional lineages within phylotypes commonly attributed to the phylum (see Section 3.2) was supported by the identification of deeply-rooted and internally diverse clades (WS1 and FBP) with high bifurcation supports. These clades, with distinct sources of phylotype detection (activated sludge and anaerobic sediment (Dunfield et al., 2012), were substantially dissimilar to the *Armatimonadetes* type strain, with 16S rRNA gene sequence dissimilarities above the range of previously-established prokaryotic phyla.

This research generated a robust reference phylogenetic tree of *Armatimonadetes* and neighbouring taxa, and allowed the determination of evolutionary relatedness of new phylotypes to *Armatimonadetes* as well as clades within the phylum. We applied a "first published phylotype" clade naming rule consistent with major SSU rRNA databases (SILVA, Greengenes) and provided a means to unify various previously-identified clades. With the increasing influx of community 16S rRNA gene sequence data, we provided the groundwork in classifying the microbial diversity in a wide range of environments in which *Armatimonadetes* were identified. This research integrated many recommendations from previous environmental 16S rRNA gene phylogeny studies (Dalevi et al., 2001;

Hugenholtz, Goebel, et al., 1998; McDonald, Price, et al., 2012), utilised additional up-to-date methods e.g., Bayesian inference (Ronquist & Huelsenbeck, 2003), and integrated tree branch bipartition test with Bali-Phy (Suchard & Redelings, 2006) in order to resolve problematic deep-branching taxonomic relationships.

Ultimately, phylogenetic trees are dynamic structures dependant on available data. We recognise that inferred relationships are subject to change as more phylotypes are sequenced. However, we aimed to provide sufficient high-quality phylotypes and high-confidence class-level clades in our study as a useful reference point for future refinements. Phylogenetic inference has become one of the most valuable tools in defining high-level taxa. The hierarchical prokaryotic taxonomy is an evolving system with shifting paradigms, resulting in conflicts or ambiguities. However, the system remained a valuable tool for microbiologists in cataloguing and identifying microbial diversity. We were able to associate, with degrees of confidence, taxonomic groupings with phylogeny. This paper represents part of a general trend in utilising molecular phylogeny to provide a more consistent taxonomy of microbial diversity and deriving information between lineages and their associated traits (Yarza et al., 2014).

As the primary author, I designed the experiments, conducted and optimised phylogenetic analyses, and wrote the majority of this chapter. Craig W. Herbold provided input on experiment design, discussion on phylogeny, and tested bipartitions within bootstrap trees to generate support values using Bali-Phy. Peter F. Dunfield provided input on the phylogeny of the phylum and editing of the manuscript. My supervisors Ian R. McDonald, Xochitl C. Morgan, and Matthew B. Stott reviewed and edited the content.

The published form of this research can be found as an appendix in Chapter 8, with the following citation:

Lee, K. C. Y., Herbold, C. W., Dunfield, P. F., Morgan, X. C., McDonald, I. R., & Stott, M. B. (2013). Phylogenetic delineation of the novel phylum *Armatimonadetes* (former Candidate Division OP10) and definition of two novel candidate divisions. *Applied and Environmental Microbiology*, *79*(7), 2484–7. doi:10.1128/AEM.03333-12

License to reproduce this material in this thesis has been granted by the publisher under ASM Journals Statement of Authors Rights, appended here after this preface.

4.2 - Abstract

Approximately 500 near full-length non-redundant SSU rRNA gene sequences are currently associated with the bacterial phylum *Armatimonadetes*, formerly Candidate Division OP10. However, inconsistent phylogenetic treatments and phylotype selections have yielded conflicting tree topologies, highlighting the need for an in-depth phylogenetic assessment and consensus of the *Armatimonadetes* phylum. We used a comprehensive dataset of phylotypes associated with *Armatimonadetes*, and applied multiple tree-calculating methodologies. We determined the relationships between *Armatimonadetes* and other division-level (phylum) clades across the domain *Bacteria*, and also clarified the affiliation of previously ambiguous groups such as Group 11 and Group 12, to the rest of the *Armatimonadetes* (Dunfield et al., 2012). Of these two groups, we confirmed the status of Group 12, as a candidate division, WS1, and reclassified Group 11 as a novel candidate division, which we termed Candidate Division FBP based on the first phylotype (clone FBP249, AY250868) within the group. Monophyly of *Armatimonadetes* was supported with the taxa boundary outlined in this study. The data also strongly supported monophyly for the three previously-characterised classes (*Armatimonadia*, *Chthonomonadetes* and *Fimbriimonadia*), and seven other putative class-level groups, redefining the boundary of *Armatimonadetes* in relation to several clades previously associated with Candidate Division OP10.

4.3 - Introduction

Although the Bacteriological Code does not define the division level ranking immediately above class (i.e. "phylum") in the domains *Bacteria* and *Archaea* (Lapage et al., 1992), organising microbiological strains into an interpretable taxonomic framework necessitates a division-level schema. To this end, two broadly-accepted definitions of a bacterial division-level clade are "a lineage consisting of two or more 16S rRNA gene sequences that are reproducibly monophyletic and unaffiliated with all other division-level relatedness groups that constitute the bacterial domain" (Hugenholtz, Goebel, et al., 1998), and "an unaffiliated lineage in multiple analyses of datasets with varying types and number of taxa and having < 85 % identity to reported sequences, indicating its potential to represent a new bacterial division" (Hugenholtz, Pitulle, et al., 1998). Using these loose definitions, there are estimated to be between 39 to > 70 bacterial divisions (Achtman & Wagner, 2008; Glöckner et al., 2010; Pace, 2009); 29 of these contain at least one described and cultivated strain, and are thus accepted as bacterial phyla (retrieved in September 2012; Euzéby, 1997). The majority of division-level clades have a paucity of cultivated strains (Keller & Zengler, 2004; Pace, 2009), or none, in which case they are known as candidate divisions (Hugenholtz, Pitulle, et al., 1998). Candidate divisions and those with few cultivated representatives can thus only be realistically defined by the systematics of the phylogenetic comparison of highly conserved genes, in most cases by comparison of the SSU rRNA gene sequence phylogeny (Ludwig & Klenk, 2005).

A case in point is the former Candidate Division OP10, now known as the phylum *Armatimonadetes* (Tamaki et al., 2011). Since the description of the (candidate) division in 1998 by Hugenholtz, Pitulle, and colleagues, *Armatimonadetes* has grown considerably as environmental sequences were attributed to the division. The original SSU rRNA gene sequences "Hot-spring clone OPB50" (AF027092) and "Hot-spring clone OPB80" (AF027089) were used as the ingroup set to define monophyly of subsequent sequences placed within the candidate taxon (Dalevi et al., 2001). As of April 2012, 364, 568, and 653 SSU rRNA gene sequences were classified as either *Armatimonadetes* or Candidate Division OP10 in the EMBL, SILVA Reference (SSU Ref), and RDP databases respectively. Using the SILVA database numbers, the sequences within the division have a maximum SSU rRNA gene sequence dissimilarity of ~29 % and span a broad range of source environmental niches including human skin, hot springs, temperate soils and aerobic/anaerobic bioreactors (Dunfield et al., 2012). However, the phylogeny of *Armatimonadetes* is still poorly defined, as recent publications (Dalevi et al., 2001; Dunfield et al., 2012; Im et al., 2012; Lee et al., 2011; Portillo et al., 2008; Stott et al., 2008; Tamaki et al., 2011) have not agreed upon a consistent and well-supported

consensus tree. These publications have used a variety of methodologies and outgroup/ ingroup sequence selections, which has resulted in the description of between four and 12 (in some cases poorly-supported) sub-groupings. Thus, the phylogenetic diversity and sub-divisional architecture of *Armatimonadetes* remains uncertain.

In order to clarify the phylogenetic relationships within the *Armatimonadetes* and establish well-supported taxa boundaries, we conducted phylogenetic inference using multiple methodologies to i) confirm the division boundaries, and ii) define the sub-phylum-level group structure(s), including the previously identified classes *Armatimonadia* (Tamaki et al., 2011), *Chthonomonadetes* (Lee et al., 2011) and the recently validated class of *Fimbriimonadia* (Im et al., 2012).

4.4 - Methodology

4.4.1 - Ingroup and outgroup sequence selection:

SSU rRNA genes putatively identified as belonging to *Armatimonadetes* or Candidate Division OP10 were selected from the SILVA SSU NR (non-redundant) Release 108 database (Pruesse et al., 2007) for initial sequence alignment. In the NR database, redundant sequences from uncultured clones were removed from the raw database by clustering with UCLUST (Edgar, 2010) using 99 % identity criterion to increase efficiency in phylogenetic analysis while maintaining a representative dataset. Construction of phylogenetic trees and additional manual refinement of the alignment were performed in the ARB software environment (Ludwig et al., 2004). Sequences with low quality (alignment quality value < 75) and/or Pintail scores (< 50) (Ashelford et al., 2002) were excluded in this study (automated analyses conducted in the SILVA database). In total, the ingroup dataset consisted of 492 sequences (Supplementary Table 8.1), excluding 15 chimeric sequences (Supplementary Table 8.2) found during the analysis outlined in this paper. The outgroup consisted of 46 sequences, which was a combination of the three outgroup sets described by Dalevi et al. (2001), named OP10A (15 sequences), OP10B (17 sequences), and OP10C (14 sequences). The combined outgroup was replicated with only minor adjustments where sequences were not included in the SILVA NR database. In this case, the most closely-related sequences in the NR database were used. A list of the accession numbers for the sequences used in the outgroup is available in Supplementary Table 8.3.

4.4.2 - Methods of phylogenetic analysis

This study utilised five different phylogenetic methods to define the nodal support for class level groupings. The rationale for the selection of algorithms is set out in the

discussion. Bootstrap analyses were performed using neighbour-joining (NJ) and maximum likelihood (ML) methods. NJ was performed using ARB (Ludwig et al., 2004) with Jukes-Cantor and Olsen substitution models and with 2,000 bootstrap replicates. ML was performed using PhyML v2.4.5 (Guindon & Gascuel, 2003) using a general time reversible (GTR) model with gamma-distributed rate heterogeneity and an estimated proportion of invariable sites (assuming four substitution rate categories), and RAxML v7.0.3 (Stamatakis, 2006) using the GTRMIX model with the rapid bootstrapping algorithm. Both ML methods were performed with 500 bootstrap replicates. PhyML and RAxML were interfaced using the ARB software environment. The GTR model was selected as the best-fit ML model based on the results of jModelTest (Posada, 2008) with the ingroup dataset. Bayesian Inference (BI) was performed using MrBayes v3.2.1(Ronquist et al., 2003) with GTR model with gamma-distributed variation with six gamma categories. BI was performed using Metropolis-Coupled Markov Chain Monte Carlo (MC^3) for 3,100,000 generations with four chains (three chains were heated at chain temperature = 0.2 with a subsampling frequency of 400. Posterior probabilities were calculated from the last 1000 trees from each analysis (last 400,000 iterations), which is an effective relative burn-in of 87 %. RAxML trees of the corresponding datasets were used as the starting tree for BI analysis. The starting trees were scrambled with random perturbations (nperts = 10). For each given dataset, two simultaneous independent analyses were performed with two differently perturbed starting trees to improve statistical confidence.

4.4.3 - Determination of monophyletic groups

For the sake of clarity, we used the sub-divisional group nomenclature (Group 1-12) as defined by Dunfield et al. (2012) as a starting reference point. As these groupings may not have been the deepest-branching monophyletic groups within *Armatimonadetes*, we tested the monophyletic support of various hypothetical sets of groups. This was accomplished by using the "trees-bootstrap" command in Bali-Phy (Suchard et al., 2006) which queries the many raw trees generated by BI and ML methods for the occurrence of the hypothetical sets, and generate support values even when the phylogenetic relationship was not shown in the consensus trees. These extracted support values (Supplementary Table 8.4 & Supplementary Table 8.5) were then used for creating a most probable consensus tree combining support values from all the methods (Figure 4.1 in Results section). Specifically, the extracted values were used to establish the relationship of intermediary nodes within the *Armatimonadetes* phylogenetic tree. In order to avoid confusion of generic group names and facilitate communication between studies, we employed a similar approach to Greengenes (McDonald, Price, et al., 2012). Monophyletic groupings, once established, were subsequently given unique identifiers

based on the name of the first validly-published phylotypes within said group (e.g., "OPB50" for Group 6). Correspondence of the two naming schemes can be found at Figure 4.1.

Monophyly of clades represented in Figure 4.1 were called using ≥ 70 % ML nonparametric bootstrap proportion as the threshold (Hillis & Bull, 1993). While Bayesian posterior probabilities (PP) are usually strongly correlated with ML bootstrap proportions (Alfaro et al., 2003), it is generally recognised that PP values tend to be higher than that of ML nonparametric bootstrap proportion (Alfaro & Holder, 2006; Taylor & Piel, 2004; and references therein). Here we used ≥ 95 %, as a commonly suggested PP threshold (Murphy et al., 2001; Wilcox et al., 2002). Where branch nodes were not supported by confidence values of greater or equal to the threshold, the branches were multifurcated (Peplies et al., 2008). Sequence dissimilarities between the phylum type strain *A. rosea* YO-36^T and individual key phylotypes, as well as mean sequence dissimilarities between the type strain and key groups were generated in ARB to assess the boundary of the phylum based on uncorrected sequence distance from the type strain (Figure 4.2). The average maximum sequence dissimilarity of phyla across the bacterial domain, with 95 % confidence intervals (Yarza et al., 2010) are provided as a comparative guide.

4.4.4 - Detection of chimeric sequences

Out of 492 OP10/*Armatimonadetes* sequences extracted from the SILVA database, 15 chimeric sequences were identified (Supplementary Table 8.2). These sequences were not initially flagged by the automated quality control in SILVA using Pintail. A majority of these sequences (13 out of 15) failed to place consistently within a single group when included in initial dataset calculations, exhibiting what we termed as "jumping" behaviour (see 4.6 -Discussion section). These unflagged sequences were manually re-analysed using Bellerophon (Huber et al., 2004) and then inspected in ARB for poorly-aligned segments. The putative rogue insertions were tested using BLAST (Altschul et al., 1990). Most of the "jumping" sequences had probable insertions, suggesting that they were chimeric and thus were excluded from further inclusion in ingroup dataset calculations.

4.5 - Results

We assessed monophyly of the phylum and the (class-level) sub-divisional groupings of *Armatimonadetes* using a total of 477 (492 from SILVA database, with 15 chimeric sequences removed) SSU rRNA gene ingroup sequences, and five different phylogenetic inference methods. Bootstrapping or MC^3 were used to test the robustness and reproducibility of potential monophyly identified within the trees. A consensus of five separate phylogenetic inferences identified monophyletic groups as per threshold criteria (Supplementary Table 8.4 & Supplementary Table 8.5). Based on the current ingroup dataset, two deeply-branching nodes (node A and node B in Figure 4.1) are shown to have strong monophyly support. Node A, which includes all three *Armatimonadetes* classes as well as the original OP10 phylotypes, but excludes Group 11 and 12, is supported by four out of five methods. Node B which represents all the ingroup sequences was supported by all five of the phylogenetic methods used in this study. Method support was defined as ≥ 70 % for ML and NJ methods, and ≥ 95 % for BI.

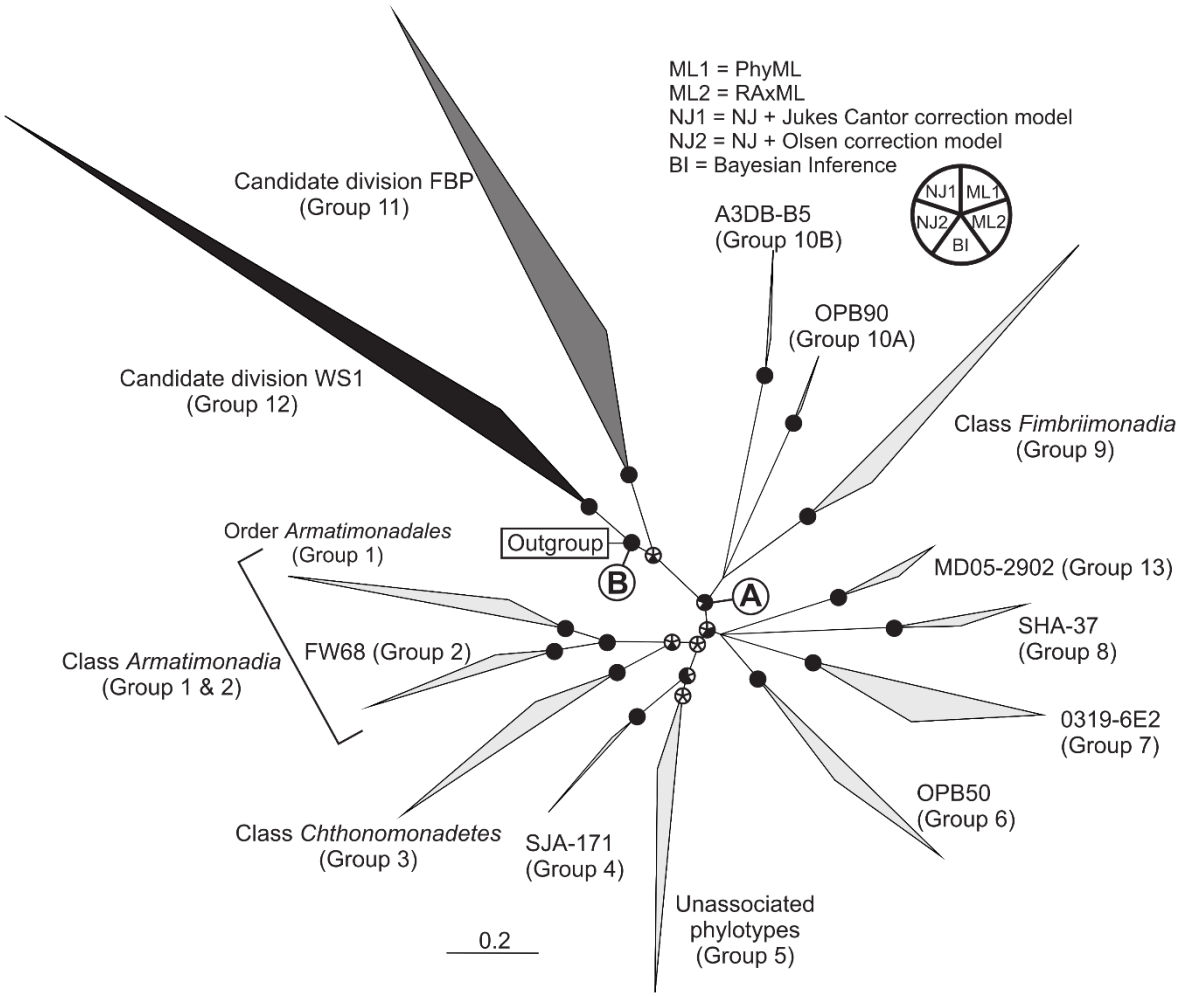


Figure 4.1 - Unrooted consensus tree showing the phylum *Armatimonadetes* and affiliated groups. The groups were based on those described in Dunfield et al. (2012), the list of phylotypes (accession numbers) within each group are listed in Supplementary Table 8.1. The analyses of support for the groups are outlined in Table 4.1 and further detailed in Supplementary Table 8.4 & Supplementary Table 8.5. The nodal support here displays supports by five phylogenetic methods (2× NJ, 2× ML, and BI) and are represented by pie charts on each bifurcation node. Support is defined as ≥ 70 % bootstrap proportion value (NJ and ML) and ≥ 95 % posterior probability value (BI). A node supported by the specific method has the corresponding portion of the pie chart shaded black. Multifurcations were manually introduced to nodes without support from any of the methods (Peplies et al., 2008). Taxa designations are shown based on the position of type strains and the supported monophyly of nodes. On the consensus tree, node marked "A" represents the phylum *Armatimonadetes* and the node marked "B" represents the super-phylum from the inclusion of *Armatimonadetes* with the Candidate Division FBP (Group 11) and WS1 (Group 12). The scale bar represents 0.2 nucleotide substitutions per site.

Uncorrected sequence distances shows that Group 11 exceeded the upper 95 % confidence interval range of the average maximum phyla boundary calculated by Yarza *et al*. (2010) from curated bacterial family type strains. Group 12 fell just within that confidence interval, but still exceeded the boundary itself (Figure 4.2). Other key phylotypes (including the class type strains and the original OP10 clones) and associated groupings were less than average maximum phyla boundary as a result of a lesser SSU rRNA gene sequence dissimilarity to the phylum type strain *A. rosea* YO-36^{T}.

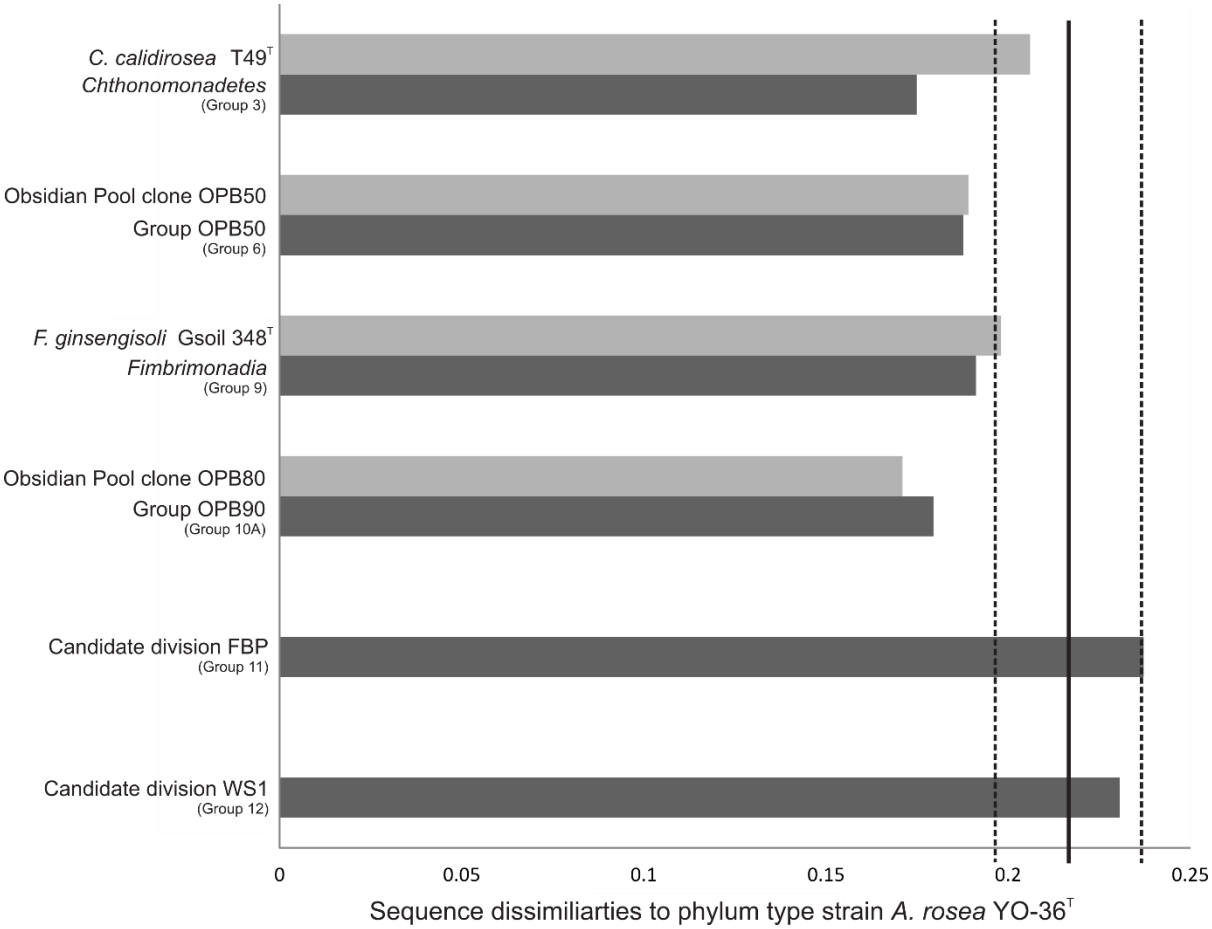


Figure 4.2 - Bar graph displaying sequence dissimilarities between *Armatimonadetes* type strain *A. rosea*^{T} and key phylotypes (*Armatimonadetes* isolates and the original OP10 clones) as the lighter bars. The graph also displays mean sequence dissimilarities between *A. rosea*^{T} and groups containing the key phylotypes (Group 6, Group 9, Group 10A, Group 11, and Group 12) as the darker bars. The groups were based on those described in Dunfield et al. 2012, the list of phylotypes (accession numbers) within each group are listed in Supplementary Table 8.1.The solid vertical line marking the value of 2.16 indicates the average maximum phylum sequence dissimilarity calculated by Yarza et al. (2010). The flanking dotted line indicates the 95 % confidence interval (± 0.02).

Bipartition support values of groups within *Armatimonadetes* (shown in Table 4.1 and further detailed in Supplementary Table 8.4 & Supplementary Table 8.5) confirms the

monophyly of the previously described classes, *Armatimonadia*, *Chthonomonadetes*, and *Fimbriimonadia* as defined by respective authors (Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011). These taxa correspond to Group 1-2, Group 3, and Group 9 shown in the consensus tree (Figure 4.1) However, the analysis suggests that the boundary of the class *Armatimonadia* should in fact include Group 2 due to the strong support for monophyly of the combined "super-group" (Groups 1 and 2). In addition, based on the putative groups identified by Dunfield et al. (2012), individual monophyly of Group 6 (containing original OP10 sequence OPB50, AF027092 from Hugenholtz, Pitulle, et al., 1998), Group 4 (SJA-171), Group 7 (0319-6E2), Group 8 (SHA-37), Group 11 (FBP), Group 12 (also described as elsewhere as Candidate Division WS1 (Dojka et al., 1998)), and a small previously unidentified Group 13 (MD05-2902) were supported by the analyses (Table 4.1). The original Group 10 (Dunfield et al., 2012) failed to resolve unambiguously and was instead split into two strongly monophyletic groups, Group 10A (OPB-90) and 10B (A3DB-B5). Group 10A contains the original OP10 sequence OPB80, AF027089. In addition, none of the methods supported the monophyly of Group 5 (Table 4.1).

Table 4.1 - Summary of support values of key groups associated with *Armatimonadetes* generated from neighbour joining (NJ), maximum likelihood (ML) and Bayesian inference (BI) methods. NJ method displayed the results from using two distance correction methods (Jukes-Cantor and Olsen). A unique identifier for each monophyletic group was given based on the name of the first validly published phylotypes within said group.

Groups tested	**Methodologies**			
	NJ (JC,Olsen)	PhyML	RAxML	BI
Phylum				
Armatimonadetes (excluding Group 11 and 12)	70, 68	85.2	98.2	100
(Class level) sub-divisional groupings				
Group 1-2 (*Armatimonadia*)	99, 99	97.2	99.6	100
Group 3 (*Chthonomonadetes*)	99, 99	99.8	99.8	100
Group 4 (SJA-171)	99, 99	100	99.8	100
Group 5 (Unassociated phylotypes)	U, U	32	31.6	58.7
Group 6 (OPB50)	92, 91	89	96	100
Group 7 (0319-6E2)	99, 99	100	100	100
Group 8 (SHA-37)	99, 99	100	100	100
Group 9 (*Fimbriimonadia*)	99, 99	100	100	100
Group10A (OPB90)	99, 99	100	100	100
Group10B (A3DB-B5)	99, 99	100	100	100
Group13 (MD05-2902)	99, 99	100	100	100
Super-phylum				
Armatimonadetes, Group 12 (WS1) & Group 11 (FBP)	78, 80	79.6	91.4	100
Candidate divisions				
Group 11 (FBP)	98, 98	99.4	99.8	100
Group 12 (WS1)	76, 79	90.4	99.8	100

U; unsupported

Phylogenetic relationships of nodes between higher order groupings (i.e. combining aforementioned groups) were poorly supported by all methodologies with the exception of BI (Supplementary Table 8.4 & Supplementary Table 8.5). Of all the hypothetical sets tested for example, Groups 6 (OPB50), 7 (0319-6E2), 8 (SHA-37) and 13 (MD05-2902) (Figure 4.1), only the node formed by Group 1 (Order *Armatimonadales*) and Group 2 (FW68), and the phylum *Armatimonadetes* in its entirety was decisively monophyletic (Figure 4.1; Table 4.1). A radial consensus phylogenetic tree summarising the final grouping nomenclature and relationships is presented in Figure 4.1.

4.6 - Discussion

The inconsistent application of SSU rRNA gene sequences and use of different methodologies to generate phylogenetic frameworks for the former Candidate Division OP10/*Armatimonadetes* (Im et al., 2012; Lee et al., 2011; Portillo et al., 2008; Stott et al., 2008) has led to difficulty in comparing published tree topologies and group nomenclatures, as the different selection of ingroups meant that only limited consistency in group nomenclature between different analyses in publications. Because of this, it is perhaps unsurprising that the previous phylogenetic frameworks have failed to reach a coherent consensus on the phylogeny of *Armatimonadetes*. The reported support and architecture of the phylum (and the groups within it) vary from apparently well-supported to poorly supported (Dunfield et al., 2012; Lee et al., 2011; Portillo et al., 2008; Tamaki et al., 2011) depending on the individual phylogenetic methods used and the selection of ingroup and outgroup phylotypes.

The goal of this research was twofold; i) to confirm the current paradigm of *Armatimonadetes* monophyly, and ii) to generate a consensus tree topology and a uniform group nomenclature. To achieve this, we first selected only representative high-quality sequences and then utilised five different phylogenetic methods to define the nodal support for class level groupings. These methods belong to three separate phylogenetic paradigms (NJ, ML, and BI), and within these approaches, different implementations were used. These implementations include two correction models for NJ and two executions of the ML approach in the form of PhyML and RAxML. More than one implementation or correction model were used as they may highlight potential issues in phylogenetic inference, such as the sensitivity of NJ to incorrect model assumptions (Huelsenbeck & Hillis, 1993). The maximum parsimony (MP) method was not selected for use in this study as MP is known to suffer from long branch attraction, which becomes prominent in large and highly divergent datasets such as this one (Alfaro et al., 2003; Felsenstein, 1978; Ludwig et al., 2005). As a widely-used methodology, ML was selected

based on reports of correct tree recovery in simulated datasets (Huelsenbeck et al., 1993; Huelsenbeck, 1995), robustness to model selection error (Yang et al., 1994), and desirable statistical properties (Rogers, 2001). PhyML uses an improved form of NJ (BIONJ) to produce a starting tree which then undergoes iteration via Nearest Neighbour Interchanges (NNIs) topological moves and then a preferred tree is determined via ML optimality criterion. RAxML on the other hand uses MP to produce a starting tree and utilises Lazy Subtree Rearrangement (LSR) topological moves. Previous research has demonstrated that the search algorithms (NNI vs. LSR) implemented by PhyML and RAxML respectively result in lower bootstrap values for PhyML using the same data sets (Stamatakis, 2006).

The two ML methods (PhyML and RAxML) showed little disagreement in nodal support values and sequence clustering, with the only exception of the intermediary node "G9-G10A-G10B-G11-G12-Outgroup", where the bootstrap support values of two methods disagreed with each other by 37.8 % (Supplementary Table 8.5). While there is no simple way to determine which method made a more accurate assessment of the nodal support, the differences may be the result of the way in which the methods implement the maximum-likelihood calculations. The NJ algorithm was used as a further assessment of the *Armatimonadetes* dataset as an alternative to the ML method. NJ computations are much faster than ML and thus remained applicable over large datasets, but the methodology has several drawbacks compared to newer and more computationally intensive methods (Holder et al., 2003 and references therein). Nonetheless, this well-established technique provides an alternative approach for comparison against other methods. A comparison of the NJ and PhyML results show that the trees shared similar topologies and shared identical sequence clustering. The Olsen and Jukes-Cantor correction methods made little difference to tree topologies and support values. Finally, BI was also applied to the *Armatimonadetes* dataset. While BI shares a strong connection to ML, the two methods differ by the different statistical formulations (Bayesian statistics vs. frequentist statistics). As a relatively new method, BI is still comparatively uncommon in bacterial phylogenetic studies. However, it is gaining popularity due to some advantages over ML (Alfaro et al., 2003, 2006; Holder et al., 2003), as well as the development of phylogenetic applications such as MrBayes (Ronquist et al., 2003) and BEAST (Drummond & Rambaut, 2007). An analysis of the PP values for BI showed shared support for groups as identified by the ML methods. However, BI showed higher support values for deeper nodes between groups. These intermediary nodes were rarely supported by any other methods except in the case of Group1-Group2, which was universally supported. These results highlight the supported monophyly of the sub-phylum-level groups presented in Figure 4.1, and usefulness of BI approach to detect

sensitive phylogenetic signals. However, we also note that there is a degree of on-going debate on whether BI may overestimate uncertain signals (Cummings et al., 2003; Erixon et al., 2003; Taylor et al., 2004), and thus the inter-group structure generated from the support of intermediary nodes should be treated with a degree of caution. In summary, the methods were in general agreement in their support for the groupings examined, and the use of the multiple methods also provided confidence in the phylogenetic conclusions.

The multiple phylogenetic methods approach was also crucial in identifying unflagged problematic sequences. These potentially chimeric or low quality sequences (15 sequences in total) appeared to "jump" between different clusters when comparing the BI and other methodologies. Conversely, the remainder of the phylotypes (525, including those in the outgroup) and the broader tree topologies remained essentially the same for all applications of these methods. This behaviour was initially thought to be an artefact of the BI method, possibly due to low chain convergence. However, the phenomenon persisted despite adjustment in parameters in BI including increased chain length and using a random starting tree (instead of using ML tree from RAxML as initially used). Curation by SILVA database with Pintail (Ashelford et al., 2002) and subsequent reanalysis using Bellerophon (Huber et al., 2004) failed to show the sequences as chimeric. Nevertheless, upon manual examination of the multiple alignments of these "jumping" sequences, putative inserted sequences (poorly aligned and highly dissimilar to known SSU rRNA gene sequences), were identified and interrogated using BLAST (listed in Supplementary Table 8.2).

The interplay of multiple methods approach helps highlight problematic sequences as well as assist in troubleshooting anomalous nodes (i.e. nodes with inconsistent support values) by adjusting parameters (e.g., chain temperature in BI and removal of chimeric sequences). Even so and despite extensive analyses, the resolution of resulting phylogenetic inference may still be limited by factors such as the computational resources available, and the limits of phylogenetic signals within the datasets. For example, we were unable to resolve the monophyly of Group 10 as a single group, and only when the group was subdivided into two independent groups, Group 10A (OPB-50) and Group 10B (A3DB-B5), was monophyly supported. Interestingly, both Group 10A and Group 10B consisted of clones entirely from geothermal environments (Dunfield et al., 2012) lending some evidence to a closer phylogenetic relationship not currently shown by SSU rRNA gene phylogeny. Alternatively, this could be the result of convergent evolution and the verdict awaits new data and further investigations. Group 5 was also unsupported as a monophyletic group, but unlike Group 10, we were unable to identify any clearly distinguishable monophyletic subgroups, and instead its incumbents should be currently

considered a collection of not-specifically associated *Armatimonadetes* phylotypes. However, it is probable that the currently ambiguous phylogenetic relationship of these phylotypes may become more evident as more phylotypes are deposited in databases and more cultured *Armatimonadetes* isolates are characterised and published.

The type strain of *Armatimonadetes*, *Armatimonas rosea* YO-36^T (AB529679) (Tamaki et al., 2011) placed as previously reported in Group 1 (Dunfield et al., 2012; Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011). However, a deeper and strongly-supported branching is shared between Group 1 and Group 2 (Dunfield et al., 2012; Im et al., 2012; Lee et al., 2011), suggesting the presence of a single amalgamated monophyletic group (Group 1-2). These data indicate that this combined grouping should represent the current boundary of the class *Armatimonadia*. This amalgamation is consistent with the variety of ecological niches reported previously, including temperate soils, freshwaters, human skin, and microbial biofilms/mats (Dunfield et al., 2012). Furthermore, considering the expanded class *Armatimonadia*, we also suggest that Group 1 as cited by several authors (Dunfield et al., 2012; Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011) now represents the order *Armatimonadales* (Figure 4.1). Group 3 includes the type strain, *Chthonomonas calidirosea* T49^T (AM749780), a thermophilic heterotroph isolated from New Zealand (Stott et al., 2008). This group had strong support for monophyly from all phylogenetic methods used. The phylogenetic relationship of Group 3 with other groups generated by the datasets had support values below the threshold and thus was not resolved and confirms the designation of Group 3 as the class *Chthonomonadetes* (Dunfield et al., 2012; Im et al., 2012; Lee et al., 2011; Portillo et al., 2008; Tamaki et al., 2011). A third recently described strain belonging to *Armatimonadetes*, *Fimbriimonas ginsengisoli* Gsoil 348^T (GQ339893) is the first cultivated representative of Group 9 (class *Fimbriimonadia*), an extensive clade containing a variety of host niches (Dunfield et al., 2012; Im et al., 2012).

This study has confirmed the previous assertion (Dunfield et al., 2012) that Candidate Division WS1 is the deepest-branching group (Group 12) relating to other *Armatimonadetes*-associated groups (node B, Figure 4.1). If we apply the often used criterion for a phylum/candidate division, which requires consistent monophyly of the deepest resolvable branch that is independent from all other division-level taxa (Hugenholtz, Goebel, et al., 1998), then node B would be a likely boundary of the phylum *Armatimonadetes.* However, when sequence dissimilarity is taken into consideration, the two deepest branching groups within the ingroup of this study, Group 11 and Group 12, are shown to be either at (Group 12), or exceeding (Group 11) the upper limit (95 %

confidence interval) of the average maximum bacterial phylum boundary (Yarza et al., 2010; Figure 4.2).

It is also worth noting that since the sequence dissimilarity is uncorrected using a distance correction model, the real genetic distance tends to be underestimated, especially when comparing highly divergent sequences due to saturation (Hamilton, 2009, p. 243). This means small changes of uncorrected genetic distance between highly divergent sequences translate to larger changes in the actual genetic distance. Group 11 (Candidate Division FBP) and Group 12 (Candidate Division WS1) therefore individually represent monophyletic groups which have substantially above average distance to the closest phylum type strain compared to groups in other bacterial phyla. Furthermore, the source environments of phylotypes within Group 11 and Group 12 exhibit distinct distributions (Dunfield et al., 2012). Group 11 consisted mostly of clones from soil or skin swab samples, and Group 12 contained mostly of clones from anaerobic environments such as oceanic sediment and anaerobic bioreactors. The contrast of source environments suggests separate ecology and underlying phylogeny. Based on the distance of these two groups from the phylum type strain, strong individual phylogenetic support, and their distinct environmental distribution, we conclude that the two groups would likely to be separate candidate divisions outside of *Armatimonadetes*. Group 12 was previously identified as Candidate Division WS1 (Dojka et al., 1998), and here we term Group 11 as Candidate Division FBP based on the earliest published phylotype, clone FBP249 (AY250868) (de la Torre et al., 2003). The two candidate divisions also share a well-supported relationship with *Armatimonadetes*, as demonstrated by Node B in Figure 4.1, this relationship would therefore likely represent the basal node for a putative highly-divergent super-phylum.

In summary, this research has confirmed the monophyly of *Armatimonadetes*, which currently includes 346 (excluding the 15 chimeric sequences, and those belonging to Candidate Divisions WS1 and FBP) near-full length non-redundant phylotypes and contains 10 sub-phylum monophyletic groupings. These groups are the classes *Armatimonadia*, *Chthonomonadetes* and *Fimbriimonadia*, and Groups 4, 6-8, 10A and 10B, and Groups 13 (Figure 4.1). We also identified two additional phylum-level monophyletic groups, Candidate Division FPB (Group 11) and the previously identified Candidate Division WS1 (Group 12) which group with *Armatimonadetes* to form a super-phylum. It is difficult to make a direct comparison of the phylogenetic assessment here with some of the previous studies involving *Armatimonadetes*/OP10 due to the difference in selection of ingroup sequences and the absence of some *Armatimonadetes* groups. However, if we only focus on the report of monophyly for "key groups" (i.e. groups

containing isolate sequences and original OP10 clones phylotypes) that are present in various studies, this paper is in agreement in regards to the monophyly of the three *Armatimonadetes* classes (Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011). We have also made some refinements on the groupings identified by Dunfield et al. (2012), with additional phylotypes and the application of multiple phylogenetic methods.

The ever-increasing numbers of phylotypes submitted to publicly-available databases represent a rich source of data available to help resolve phylogenetic landscapes. In order to accurately understand complex phylogenetic relationships, it is essential that assessments include a broad selection of quality curated sequences, the use of multiple methodologies and the selection of a diverse selection of outgroup representatives in order to place a confidence in the assessment (Hugenholtz, Goebel, et al., 1998; Peplies et al., 2008). These analyses should be on-going and should challenge current phylogenetic frameworks as the incorporation of further phylotypes into the Tree of Life will continue to shift and redefine the boundaries of different taxa. High-level microbial taxa, where ecological diversity is often broad and classifications are often based on phylogeny alone, should be considered dynamic as groupings are both discovered and amalgamated. This study provides a consistent phylogenetic framework for *Armatimonadetes* upon which future studies involving this phylum can be based.

Chapter 5

Genomic Analysis of *Chthonomonas calidirosea*, the First Sequenced Isolate of the Phylum *Armatimonadetes*

Kevin C-Y. Lee[a,b], Xochitl C. Morgan[a,c], Peter F. Dunfield[a,d], Ivica Tamas[d], Ian R. McDonald[b] and Matthew B. Stott[a,1].

[a] GNS Science, Extremophiles Research Group, Private Bag 2000, Taupō 3352, New Zealand.

[b] Department of Biological Sciences, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.

[c] Department of Biostatistics, Harvard School of Public Health, Boston, MA USA.

[d] Department of Biological Sciences, University of Calgary, 2500 University Dr. NW, Calgary, T2N 1N4 Canada.

5.1 - Preface

This chapter focuses on the characterisation of the *Chthonomonas calidirosea* T49^T genome, the first published genome from *Armatimonadetes* (Lee et al., 2014). The chapter consists of the manuscript submitted for publication, along with supplementary information (see appendix Section 8.2). Supplementary tables too large for printed form are referenced to digital files attached to this thesis. Genomic data deposited in publicly available databases are referred to using accession numbers.

Chthonomonas calidirosea T49^T, the type strain of the class *Chthonomonadetes*, represented one of three formally-described *Armatimonadetes* species. The phenotypic analysis of the strain has identified characteristics such as narrow pH growth range, isoleucine auxotrophy, and a carbohydrate-dominated metabolism. These characteristics indicated T49^T may be a scavenger within a thermophilic cellulolytic environment, which targets degradation by-products, such as soluble carbohydrates from cellulose hydrolysis, within the community. These findings represent some of the earliest insights into the genetic, metabolic, and ecological functions within the novel phylum. In addition, we aim to answer the hypothesis on the phylogeny of the phylum based on this strain (see Section 3.2), as well as investigating the genetic basis of the physiological characteristics (such as carbohydrate-based metabolism). To achieve this, we sequenced the *C. calidirosea* T49^T genome. After *de novo* assembly into contigs and scaffolds, we performed genome finishing, including primer walking to close gaps between contigs and generate a high-quality genome sequence for the downstream analysis.

Among the putative protein coding genes in *C. calidirosea* T49^T genome, 21.45 % had no functional-prediction. As the only publicly available genome within *Armatimonadetes* at

the time, genes from *C. calidirosea* T49^T exhibited overall low similarity to references in literatures and databases. The median protein sequence identity of the T49^T genes (*with* functional prediction) to entries in NCBI RefSeq protein database[e] was 46.18 % and no higher than 91.89 % (from the highly-conserved ribosomal protein L36P, to a homolog in *Candidatus* Solibacter usitatus, from the phylum *Acidobacteria*). For a graphical comparison of amino acid sequence similarity when comparing between phyla (T49^T to other phyla), species, and strains, please refer to Figure 7.1.

The hypothesis on the phylogenetic relationship of *Armatimonadetes* and *Chloroflexi* was addressed via a comprehensive analysis of a large number (~400) of conserved genes, among many genomes (3,737) within the bacterial domain. The findings supported *Chloroflexi* to be the closest-related formal phylum to *Armatimonadetes*. However, the analysis also showed an even closer relationship with an uncultivated representative of TM7, based on partial genomic data obtained through metagenomics sequencing (Podar et al., 2007).

The phylogenetic distinctiveness of *C. calidirosea* meant the genomic content was also overall distinct from known species. The novelty of the phylum meant that no closely-, or even moderately-related species with well-studied genes could be used to conclusively infer gene functions. Furthermore, while bacterial genomes are typically arranged in functionally-linked operons to facilitate simultaneous transcription and translation of metabolically-related genes, the proximity of genes ("gene neighbourhood") in *C. calidirosea* was not reliable for inference of metabolically-associated functions. Rather, its genome appeared to be highly disorganised, such that operon-like sequences did not appear to be functionally linked, while functions typically contained within operons (e.g., tryptophan biosynthesis) instead had their component genes scattered throughout the genome. To address this issue, we carefully examined putative genes with available databases (e.g., Pfam and NCBI GenBank databases) interfaced through the Integrated Microbial Genomes / Expert Review (IMG/ER) system in order to establish likely metabolic pathways and capabilities of the bacterium. In addition, due to the limitations of gene prediction for a novel phylum, we conducted cultured-based physiological experiments (carbohydrate and amino acid utilisation, and carotenoid analysis) to validate the results of genomic analysis.

[e] Retrieved in October 2014, excluding *Armatimonadetes* hits as those new entries did not exist when *C. calidirosea* T49^T was published

Through these methods, we identified putative genes and other genomic features, contributing to a fuller understanding of the ecological role of *C. calidirosea* T49^T. Our results showed that the strain had a carbohydrate-based lifestyle with an abundance of diverse transporters and associated pathways. In addition, we identified possible mechanisms leading to a narrow pH growth range and isoleucine auxotrophy. The high abundance of σ-factors may serve as the primary means for gene regulation, and may provide an explanation for the apparently-disorganised genome (see above section regarding gene neighbourhood) and the lack of commonly-found operons. Overall, the genome analysis supported the hypothesis (see Section 3.2 -that the genomic content reflects the carbohydrate-utilising phenotypic traits with a diverse range of carbohydrate metabolism genes. Furthermore the carbohydrate-based phenotypic traits were also reflected through an abundance of transporters and hydrolases to degrade and transport carbohydrates. In addition, the lack of putative endoglucanases provided a mechanistic explanation for the inability of *C. calidirosea* T49^T to utilise structured/crystalline cellulose.

This research represents the first genome and the first integration between physiological and genomic data within the phylum *Armatimonadetes*. The analysis provided a genetic context to our ecological knowledge. Additionally, the genomic data contributes to our overall understanding of microbial genetic diversity by providing grounding information for future predictive models in processes such as gene annotation. This research opens up opportunity for comparative analyses with other members of *Armatimonadetes* in order to gain a representative perspective of this novel lineage. The second genome of an *Armatimonadetes* species (and the first within class *Fimbriimonadia*) was recently published (Hu et al., 2014). A brief comparison with *C. calidirosea* T49^T is outlined in Section 6.8 - Addendum of the next chapter.

As the primary author in this publication, I designed the experiments, cultivated *C. calidirosea* T49^{T}, and contributed genome assembly through primer-walking and resolution of gaps in the genome, performed analysis of the genome, wrote the majority of the associated manuscript, and compiled contributions from the co-authors. Xochitl C. Morgan developed a method for extraction of high-quality genomic DNA from *C. calidirosea*, assembled the genome, and contributed to the analysis of the genome, particularly in stress response and gene organisation. Peter F. Dunfield and Ivica Tamas performed the analysis of horizontal gene transfer through Alien_Hunter (Vernikos & Parkhill, 2006) and NCBI genome sample size bias analysis. Matthew B. Stott contributed to the analysis of the genome, particularly in the carbohydrate active enzymes (CAEs). In addition, my supervisors (Xochitl C. Morgan, Ian R. McDonald, and Matthew B. Stott) as well as Peter F. Dunfield also contributed in manuscript revision and discussion.

This chapter has been published as an original research paper with the following citation:

Lee, K. C., Morgan, X. C., Dunfield, P. F., Tamas, I., McDonald, I. R., & Stott, M. B. (2014). Genomic analysis of *Chthonomonas calidirosea*, the first sequenced isolate of the phylum *Armatimonadetes*. *The ISME Journal*, 8, 1522–1533. doi:10.1038/ismej.2013.251

5.2 - Abstract

Most of the lineages of *Bacteria* have remained unknown beyond environmental surveys using molecular markers. Until the recent characterisation of several strains, the phylum *Armatimonadetes* (formerly known as "Candidate Division OP10") was a dominant and globally-distributed lineage within this "uncultured majority". Here we report the first *Armatimonadetes* genome from the thermophile *Chthonomonas calidirosea* T49^T and its role as a saccharide scavenger in a geothermal steam-affected soil environment. Phylogenomic analysis indicates T49^T to be related closely to the phylum *Chloroflexi*. The predicted genes encoding for carbohydrate transporters (27 carbohydrate ATP-binding cassette transporter-related genes) and carbohydrate-metabolising enzymes (including at least 55 putative enzymes with glycosyl hydrolase domains) within the 3.43 Mb genome help explain its ability to utilise a wide range of carbohydrates as well as its inability to break down extracellular cellulose. The presence of only a single class of branched amino acid transporter appears to be the causative step for the requirement of isoleucine for growth. The genome lacks many commonly conserved operons (e.g., *lac* and *trp*). Potential causes for this, such as dispersion of functionally related genes via horizontal gene transfer from distant taxa, or recent genome recombination were rejected. Evidence suggests T49^T relies on relatively abundant σ-factors, instead of operonic organisation, as the primary means of transcriptional regulation. Examination of the genome with physiological data and environmental dynamics (including interspecific interactions) reveals ecological factors behind the apparent elusiveness of T49^T to cultivation, and by extension, the remaining "uncultured majority" that have so far evaded conventional microbiological techniques.

5.3 - Introduction

Despite a growing scientific and industrial drive to cultivate and characterise the "uncultured majority", the bulk of known microbial diversity has evaded attempts at cultivation (Fox, 2005; Keller et al., 2004). In fact, microbial strains from just five divisions of *Bacteria* (*Actinobacteria*, *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, and *Cyanobacteria*) represent 95 % of cultivated species (Keller and Zengler, 2004) and 88 % of sequenced genomes (Liolios et al., 2010). Yet cultured strains represent only a minor fraction of the global microbial and genetic diversity (Whitman et al., 1998; Wu, Hugenholtz, et al., 2009). This is a significant knowledge gap and fundamentally limits our ability to understand ecological and biogeochemical processes, and to discover novel proteins with biotechnological value (Wu, Hugenholtz, et al., 2009). The importance of investigating representatives from undersampled phyla or first cultivars from phyla with no cultivated representatives (candidate divisions) can therefore not be overstated. Here we provide the first genomic analysis of *Chthonomonas calidirosea* T49^T, the first cultivated strain of the newly-described phylum *Armatimonadetes* (formerly Candidate Division OP10) and use these data to speculate on its ecological role.

Candidate Division OP10 was a dominant bacterial candidate phylum first detected in a 16S rRNA gene survey of Obsidian Pool, Yellowstone National Park (Hugenholtz, Pitulle, et al., 1998). Until recently, this division had no cultivated strains, yet contained just over 600 full length phylotypes in the SILVA database version 114 (Dunfield et al., 2012; Pruesse et al., 2007). Phylotypes grouping within Candidate Division OP10 were detected in a wide range of environments including soil, geothermal springs, freshwater sediments, and bioreactors (Bond et al., 1995; Dunfield et al., 2012; Hugenholtz, Pitulle, et al., 1998; Lehours et al., 2007; Stott et al., 2008) suggesting a broad range of metabolic capabilities and ecological roles. The first reported cultivated strains of Candidate Division OP10, strains T49^T and P488, were isolated from geothermally-heated soils in New Zealand (Stott et al., 2008). To date, three species from the *Armatimonadetes* have been formally described: *Chthonomonas calidirosea* T49^T (Lee et al., 2011), *Fimbriimonas ginsengisoli* GSoil348^T (Im et al., 2012), and the phylum type species, *Armatimonas rosea* YO-36^T (Tamaki et al., 2011). These strains are only distantly related, with SSU rRNA gene sequence similarities of < 80 % (Dunfield et al., 2012).

A phenotypic comparison of the three cultivated strains identifies a number of common traits; all strains are chemoheterotrophic with a carbohydrate-based metabolism, strictly aerobic, Gram-negative, and pink-rose pigmented. However, in contrast to the mesophilic and neutrophilic *A. rosea* and *F. ginsengisoli*, T49^T is a thermophile (T_{opt} 68 °C, range: 50-73 °C) and a moderate acidophile (growth pH 5.3, range: pH 4.7-5.8). The three

strains differ in mol% G+C content, cell morphology, fatty acid content, quinone, and salt tolerance (Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011). In addition, T49^T appears to have a much broader carbohydrate utilisation range, including a partially cellulolytic phenotype; it can hydrolyse amorphous polysaccharides, but not linear polysaccharides (Lee et al., 2011).

Here we present the genomic analysis of T49^T and use these data along with experimental work to provide insight into the metabolism and ecology of this strain. These data allow us to infer its ecological role and speculate on possible reasons for the rarity in detection and cultivation of *Armatimonadetes* species.

5.4 - Materials and methods

5.4.1 - Genomic DNA extraction

The type strain T49^T (=DSM 23976^T = ICMP 18418^T) was grown for 7 days on solid medium (AOM1, pH 6.2) at 60 °C (Stott et al., 2008). Cell biomass was collected, washed seven times in sterile water to remove gellan gum and exopolysaccharides, and freeze-dried. Total dry cell weight was ~10 mg. Cells were resuspended in 500 µL TE buffer with 30 µL 10 % SDS and 20 µL lysozyme at 37 °C and 700 r.p.m. for 1 hour, before proteinase K (100 µL) was added and incubation was continued overnight. Genomic DNA was obtained by phenol:chloroform extraction. RNA was removed with RNAse A. The product was re-extracted with phenol:choloform, precipitated with ethanol and acetate, and resuspended in 10 mM TE buffer.

5.4.2 - DNA sequencing

A paired-end (8 kb) Titanium chemistry library of the T49^T genome was constructed and then sequenced through 454 GS-FLX system (Roche, Branford, CT, USA), which generated 171,649 reads with an average size of 400 bp. The reads assembled into three scaffolds (N50 = 3456 kb) and 60 contigs (N50 = 162 kb) using Newbler and resulted in 20× sequencing coverage. A single draft scaffold was assembled from the contigs with the assistance of the paired-end data, and the remaining gaps were closed by primer-walking and Sanger sequencing. The genomic data has been submitted to GenBank/EMBL/DDBJ databases under BioProject PRJEB1573 and accession number HF951689.

5.4.3 - Genome annotation and analysis

Gene prediction, annotation, and additional analysis were performed using the Integrated Microbial Genome-Expert Review (IMG-ER) (Markowitz et al., 2010) pipeline, anomalies in gene prediction were manually curated with the assistance of GenePRIMP (Pati et al., 2010). Carbohydrate-active enzymes (CAE) were identified by hidden-Markov model (HMM) profiles through IMG-ER and externally cross-referenced with MetaCyc (Caspi et al., 2012) and the CAZy database (Cantarel et al., 2009). Putative CRISPRs were identified via the CRISPRFinder webserver (Grissa et al., 2007) and σ-factors were identified and categorised based on HMM profile hits of conserved σ-factor regions.

Putative genes from horizontal gene transfer (HGT) were identified using Alien_hunter (Vernikos et al., 2006) using a threshold of 12.015 and a window size of 5,000 nt. BLASTX search of the output identified 357 open reading frames (ORFs). Due to the large multi-gene windows used by Alien_hunter, not all candidate genes identified were HGT products. HGT was also assessed via IslandViewer (Langille & Brinkman, 2009). The package implements "IslandPath-DIMOB", a dinucleotide sequence position bias and mobility genes analysis (Hsiao et al., 2005), and "SIGI-Hidden Markov Model", a codon usage hidden Markov model analysis (Waack et al., 2006). Finally, we utilised the deep phylogenetic roots of $T49^T$ and its lack of any known close neighbours to screen for candidate HGT genes by BLAST, selecting only genes with E value lower than $1e^{-100}$, as high similarity to genes in other phyla is unexpected for vertically-transferred genes. In this way, both the compositional biases of the DNA (Alien_hunter) and sequence similarities (BLAST) were applied in order to identify the most likely putative HGT genes. A custom Perl script was used to BLASTP search for top hits of all the predicted protein sequences against a "balanced" genome dataset (Supplementary Table 8.6) to test the effect of sample size bias for *Firmicutes* in NCBI database.

Phylogenetic inference was conducted using PhyloPhlAn (Segata et al., 2013), which uses USEARCH (Edgar, 2010) and MUSCLE (Edgar, 2004) to identify and align the conserved proteins in a new genome against its built-in database, and FastTree (Price et al., 2010) to generate an approximate maximum-likelihood tree with local support values using Shimodaira-Hasegawa test (Shimodaira & Hasegawa, 1999). Pairwise genome comparisons were conducted by assigning predicted ORFs from $T49^T$ to orthologous groups (OGs) with the highest BLAST identity within the eggNOG 2.0 (Muller et al., 2010) database with a threshold of $1e^{-10}$. The resulting 2080 ORFs with OG assignment were compared with a subset of the database. All bacteria in the "core" eggNOG database belonging to the phyla *Actinobacteria*, *Cyanobacteria*, *Aquificae*, *Thermotogae*,

Deinococcus-Thermus, *Fusobacteria*, *Chloroflexi*, and *Firmicutes* (as *Bacilli* and *Clostridia*) were selected for comparison based on the phylogenetic position of T49^T. A matrix of OG occurrences was created for each genome and pairwise genome comparisons made using three similarity indices: presence-absence (Jacaard and Sørensen indices) or abundance (Bray-Curtis index) (Legendre & Legendre, 1998).

5.4.4 - Amino acid assimilation

The requirement of amino acid supplementation for growth was tested by growing T49^T on a modified mineral salt medium FS1V (Stott et al., 2008) containing (g L^{-1}): NH_4Cl; 4.0,KH_2PO_4 ; 0.5, $MgSO_4.7H_2O$; 0.2, $CaCl_2$; 0.1, mannose; 3.0, and previously reported trace metal solutions. The medium was adjusted to pH 5.7 prior to sterilisation (121 °C, 15 psi, 15 minutes). Single amino acids (0.1 g L^{-1}) were added aseptically via 0.22-µm filter after steam sterilisation. The liquid cultures were incubated in sealed bottles with 5:1 air to medium headspace at 60 °C and 180 r.p.m.

5.4.5 - Carbon catabolite repression

Carbon catabolite repression was tested using the aforementioned modified FS1V without individual amino acid addition. Casamino acids (0.2 g L^{-1}) and carbohydrate source(s) (3.0 g L^{-1}) were added aseptically following steam sterilisation. Growth medium was supplemented with either a single or two-carbohydrate mixture. Single carbohydrate media (glucose, mannose, xylose, lactose, or galactose) were used as controls for the baseline utilisation of the respective carbohydrates. Two-carbohydrate media utilised galactose, mannose, xylose or lactose with glucose in 1:1 mixtures (w/w; sum total 3.0 g L^{-1}). Biomass generation was determined via optical density (600 nm) and carbohydrate utilisation via high performance liquid chromatography.

5.5 - Results and discussion

5.5.1 - General genome characteristics

The T49^T genome is a single circular chromosome of 3.43 Mb (Figure 5.1) with an average mol% G+C content of 54.4 and the total proportion of coding bases at 90.7 %. The G+C content of the coding regions (90.1 % of bases) was higher (55.4 %) than that of the non-coding regions (49.1 %). The genome contains 2877 predicted protein-coding genes; a functional prediction could be assigned for 2248 genes (~78 %), while 629 genes had no functional prediction of any kind. The putative origin of replication (OriC) was identified via DNA-Box-Complex using OriFinder (Gao & Zhang, 2008) and was in the vicinity of the GC skew minimum and adjacent to *recF* (CCALI_ 01940). The genome contained one complete rRNA operon (16S, 5S, 23S) and one additional copy of the 16S

rRNA gene; both 16S rRNA genes (1416 bp) had identical nucleotide sequences. Forty-six tRNAs representing 43 anticodons were encoded in the genome (Supplementary Table 8.7).

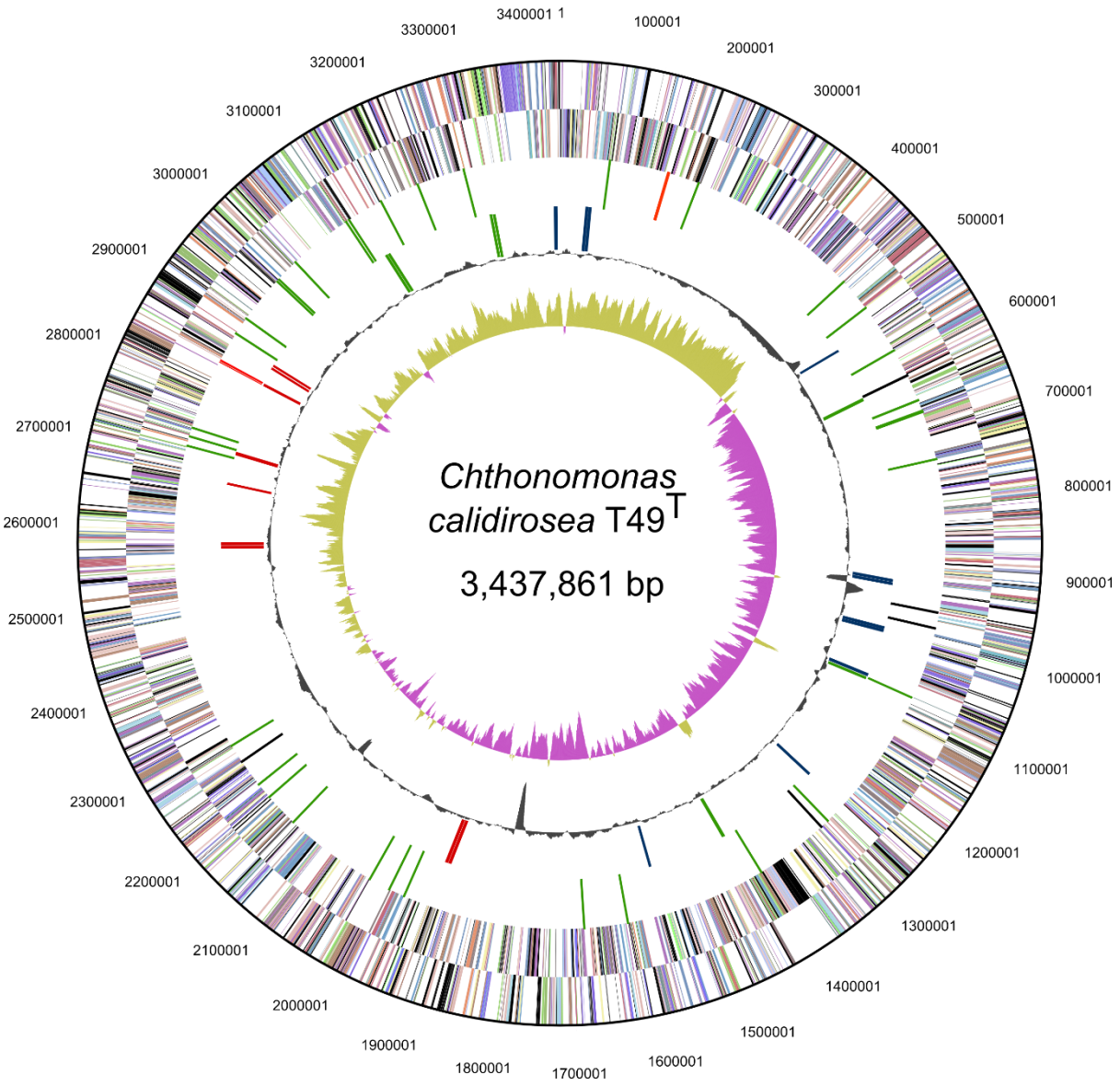


Figure 5.1 - Circular representation of the *C. calidirosea* T49^T genome. From outside to the centre: (1) Genes on forward strand (colour by COG categories); (2) Genes on reverse strand (colour by COG categories); (3) RNA genes (tRNAs green, rRNAs red, other RNAs black); (4) Genes involved in: histidine biosynthesis (red), tryptophan biosynthesis (light green), purine biosynthesis (blue); (5) GC content; (6) GC skew

A putative CRISPR element which consisted of a single spacer region (99 bp) and a pair of flanking direct repeats (46 bp each) was identified in the region 2,334,212 - 2,334,402. In addition, a cluster of CRISPR associated proteins were found at a distant locus from the putative CRISPR. Interestingly, all the genes within the cluster (*cmr1*; CCALI_02398, *cmr2*; CCALI_02399 and CCALI_02400, and *cmr3*; CCALI_02401) appear to be pseudogenes due to frameshifts and premature stop codons.

5.5.2 - Phylogenetic analysis

The placement of T49^T within the phylum *Armatimonadetes* was previously investigated through SSU rRNA-based analysis (Lee et al., 2011). However, the SSU rRNA gene alone does not reliably resolve the relationships among deeply-branching phyla (Delsuc et al., 2005). Here we utilised the recently-reported phylogenetic pipeline "PhyloPhlAn" (Segata et al., 2013) to ensure a robust phylogenetic placement of T49^T within the entire bacterial domain. The PhyloPhlAn pipeline presents an expansion in the scope of analysis, with the inclusion of both a large number (~400) of highly-conserved proteins and (currently 3,737) microbial genomes. The analysis (Figure 5.2; Supplementary Figure 8.1) confirmed previous findings (Dunfield et al., 2012) of the deeply-branching nature of T49^T/*Armatimonadetes* lineages and the strong association, and likely common ancestry with the phylum *Chloroflexi*. Additionally, the analysis confirmed our previously-inferred cladal relationships with other phyla, including *Actinobacteria*, *Deinococcus-Thermus*, and *Cyanobacteria*. Interestingly, the closest phylogenetic neighbour of T49^T was the partial genome sequence of an uncultivated representative of TM7 (Podar et al., 2007), a broadly-distributed candidate division that has been detected in activated sludge, human oral cavities, soils, and plant rhizospheres (Hugenholtz et al., 2001). Similarly to the three formally described *Armatimonadetes*, the recently sequenced TM7 metagenome had a saccharolytic-based metabolism (Albertsen et al., 2013). Finally, COGs-based functional similarity comparison with genomes from the most closely related phyla (see 5.4 - Materials and methods) showed that the highest gene content similarities were with thermophiles of diverse phylogenetic backgrounds (Supplementary Table 8.8), indicating that T49^T lacks high similarity to any other single described phylum, and thus represents a novel and distinct taxon.

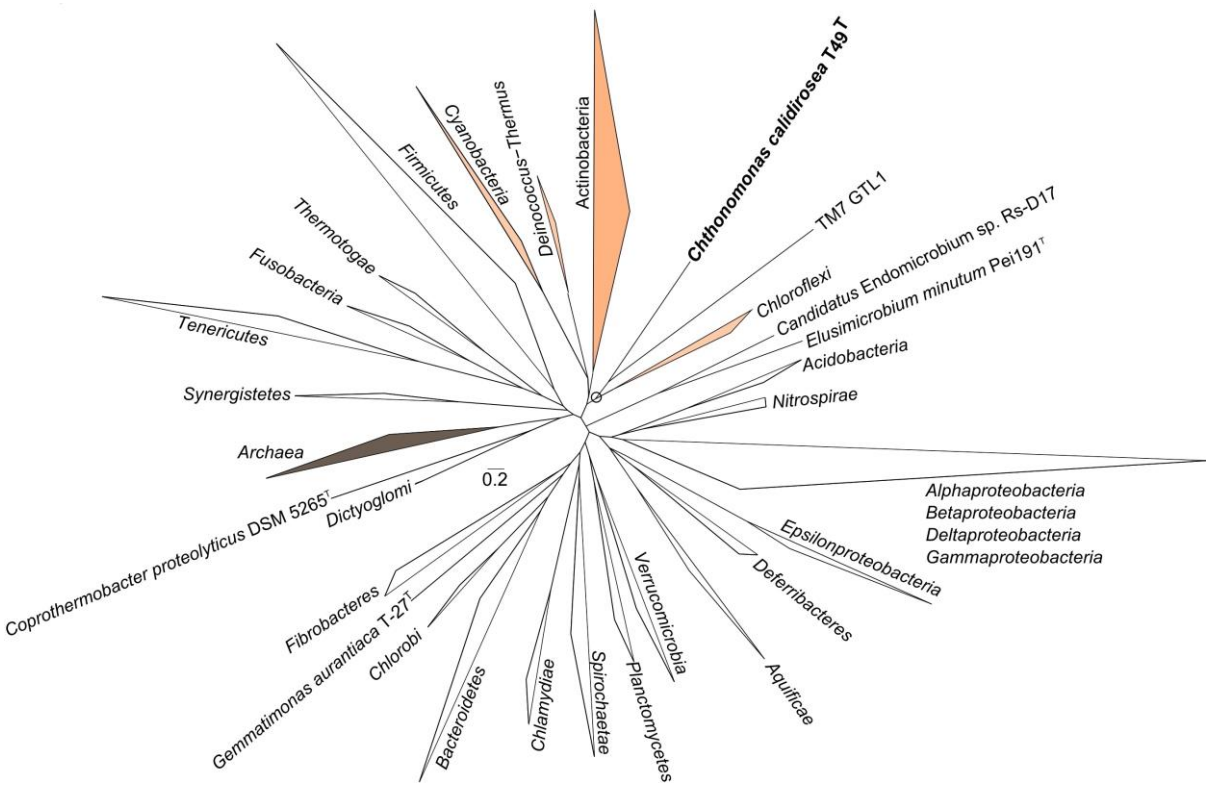


Figure 5.2 - Unrooted tree representing the phylogenetic position of *C. calidirosea* T49^T with major lineages (phyla) within the bacterial domain. The tree was constructed using PhyloPhlAn (Segata et al., 2013) with concatenated amino acid sequences of ~400 conserved proteins among 3,737 genomes. The circled node indicates the bifurcation between *C. calidirosea* T49^T and TM7 GTL1 genomes with *Chloroflexi*, the closest other formal phylum (with SH-like local support value 85 %). Scale indicates normalised fraction of total branch length. The full phylogenetic tree is shown in Supplementary Figure 8.1.

5.5.3 - Genome organisation

Many genes commonly found in conserved gene clusters such as the histidine, tryptophan, and purine biosynthesis operons (Supplementary Figure 8.2) were scattered throughout the T49^T genome (Figure 5.1). This scattering was not universal for all operon-like clusters as genes encoding for flagella biosynthesis (CCALI_01237-44, 02077-85), coenzyme PQQ biosynthesis (CCALI_00346-48), and some ribosomal proteins (CCALI_02857-95) were present in conserved gene clusters. We examined the possibility that wide-spread HGT could explain the lack of commonly-conserved operons in the genome. It has previously been reported (Davids & Zhang, 2008) that compared to strain and species-specific genes, HGT genes are the least likely to be found in operons. However, very little HGT was detected in the T49^T genome. Only 74 putative HGT genes (Supplementary Table 8.9) in small islands consisting of up to seven genes were identified via DNA compositional bias using Alien_hunter, and IslandViewer detected no genomic islands. According to their COG categories (Tatusov et al., 2003), most of the identified genes encode proteins involved in DNA replication, recombination and repair, and carbohydrate transport and metabolism, although there were also a number of hypothetical proteins. Based on protein sequence similarities, the putative sources of these genes appear to be a variety of bacterial groups. With the exception of *Actinobacteria*, the neighbouring phyla of *Armatimonadetes* (*Chloroflexi*, *Deinococcus-Thermus*, and *Cyanobacteria*) appeared to be well represented. However, the lack of other *Armatimonadetes* genomes hinders the confirmation of candidate HGT genes, as genuine HGT events can be difficult to distinguish from highly-conserved genes. Indeed, upon further examination, most of the putative HGT genes identified via sequence similarity and/or DNA compositional bias analysis, including those most similar to homologs from distant phyla such as *Aquificae* (CCALI_00187) and *Proteobacteria* (CCALI_00606), appeared to be a result of conserved vertical inheritance instead of recent HGT events. These genes showed equally-high similarity with homologs from a wide range of bacterial phyla. There were two notable exceptions:, a type I restriction-modification system methyltransferase subunit (CCALI_00805) and an adenine-specific DNA methylase (CCALI_00449), which were both most similar to homologs to the firmicute *Candidatus* "Desulforudis audaxviator" (Chivian et al., 2008). These two putative HGT genes shared exceptionally high similarity scores (64 % and 71 % identity) with their best hits (accession numbers: YP_001716579 and YP_001716684), and significantly lower similarity scores (32 % and 45 % identity) in subsequent hits, suggesting a close phylogenetic relationship between the homologs rather than inter-phyla gene conservation.

Some bacterial phyla can be influenced by large-scale HGT from another phylum. This was demonstrated for the *Thermotogae,* which have apparently received many genes from *Firmicutes* (Zhaxybayeva et al., 2009). However, such a trend was not evident for T49^T. When its inferred proteome was searched by BLASTP against the NCBI non-redundant protein database, most hits were to *Firmicutes* rather than to the phylogenetically closer *Chloroflexi* (Supplementary Table 8.6), however this probably reflects a large bias in the database rather than extensive HGT with *Firmicutes*. IMG-ER presently (August 2013) lists 1606 sequenced *Firmicutes* genomes and only 22 *Chloroflexi* genomes. BLASTP searches against a more balanced custom dataset showed that most genes matched more closely to orthologs in *Chloroflexi* than *Firmicutes*, as expected based on the phylogenetic analyses (Supplementary Table 8.6). In addition, the eggNOG gene content analysis indicated that the mean gene content similarity of T49^T to members of any other phylum was well below the mean similarity among members of that phylum (Supplementary Table 8.10). This suggests that T49^T has a unique assemblage of genes and is not greatly influenced by HGT from another single phylum. In contrast, the *Thermotogae* show a much higher gene content similarity to *Firmicutes* than to other groups (Supplementary Table 8.10), in line with the theory that *Thermotogae* have been greatly affected by HGT from *Firmicutes* (Zhaxybayeva et al., 2009). These analyses support the view that the *Armatimonadetes* is a unique phylum with low affiliation to any other phylum. Based on all of these analyses, putative HGT events seem to be rare and are unlikely to explain the lack of gene clusters and conserved bacterial operons in the genome of T49^T. The dispersal of genes is extensive, distributed through the genome, and does not correlate with the putative HGT sites.

5.5.4 - Sigma Factors

The genome of T49^T contains a total of 30 sigma-70-like proteins including 22 extracytoplasmic function (ECF) σ-factors (Supplementary Table 8.11), and has a high σ-factor to genome size (σ/Mb) ratio (Table 5.1). Diverse σ-factors play an important role in global transcriptional regulation by coordinating metabolic response genes, such as polysaccharide-degrading glycosyl hydrolases (GHs) and exopolysaccharide biosynthesis in *Bacteroides thetaiotaomicron* (Xu et al., 2004). Organisms lacking in commonly conserved gene clusters such as the marine bacterium *Pirellula* sp. strain 1 (Glöckner et al., 2003) and T49^T, may rely more heavily on σ-factors to coordinate critical metabolic processes. Indeed, this hypothesis is supported by the identification of putative promoters and sigma-70 family transcription factor binding sites of the dispersed histidine, tryptophan, and purine biosynthesis genes (Supplementary Figure 8.3).

Table 5.1 - A comparison of the number of σ-factors versus genome size for selected bacteria

Species	Genome size (MB)	No. of σ-factors	Ratio of no. of σ-factors to genome size
Escherichia coli (Blattner et al., 1997)	4.7	18	3.8
Pseudomonas aeruginosa (Potvin et al., 2008)	6.3	24	3.8
Solibacter usitatus (Ward et al., 2009)	10	58	5.8
Pirellula sp. (Glöckner et al., 2003)	7.1	51	7.1
Streptomyces coelicolor (Bentley et al., 2002)	8.7	67	7.7
Bacteroides thetaiotaomicron (Xu et al., 2003)	6.3	54	8.6
Chthonomonas calidirosea T49^T	3.4	30	8.8
Nitrosomonas europaea (Chain et al., 2003)	2.8	29	10.3

Bacterial strains isolated from soils (*C. calidirosea* T49^T, *S. usitatus*, and *S. coelicolor*) are included as a comparison with model organisms *E. coli* and *P. aeruginosa*.

5.5.5 - Primary metabolism

T49^T central metabolism and carbon fixation proceeds via routine glycolysis and the tricarboxylic acid cycle. Alpha- and β-D-glucose enter the glycolysis pathway via a glucose-6-phosphate isomerase and glucokinase. Acetyl CoA is synthesised from pyruvate via a complete pyruvate dehydrogenase complex, although the two copies of complex E3 (dihydrolipoamide dehydrogenase- CCALI_00365 and 00724) that convert dihydrolipoamide-E to lipoamide-E are located distant from the other complex-encoding genes (CCALI_01202-04). Succinyl-CoA is generated via iso-citrate dehydrogenase and 2-oxoglutarate dehydrogenase complex.

T49^T has previously been reported as exhibiting obligate chemoheterotrophic metabolism (Lee et al., 2011). In addition to these previous cultivation experiments, we predicted and experimentally confirmed its ability to utilise several less-common carbohydrates and associated derivatives, including sorbitol and galactan. Superoxide dismutase (CCALI_01521), catalase (CCALI_02206), and a complete oxidative phosphorylation pathway with an F-type ATPase were also identified, supporting the observed aerobic phenotype with no evidence of complete dissimilatory anaerobic respiration or fermentation pathways. No genes relating to photosynthesis were found in the genome, and genes related to major carbon fixation pathways were limited and fragmented. However, a complete pathway for anaplerotic CO_2 fixation appears to be in place, proceeding via the reaction of phosphoenolpyruvate (PEP) with CO_2 by phosphoenolpyruvate carboxylase (PEPC) to form oxaloacetate in a manner similar to the Wood-Werkman reaction. We have previously observed that growth of *C. calidirosea* strain P488 was improved when CO_2 was supplemented into the aerobic headspace during growth (Stott et al., 2008). While PEPC is a critical enzyme for the first step of carbon fixation in C4 and crassulacean acid metabolism plants (O'Leary & Diaz, 1982), the role of PEPC here appears to be limited to an anaplerotic role in maintaining the balance of intermediary molecules within the TCA cycle (Dunn, 2011); this is logical as the fixation of carbon dioxide is energetically-unfavourable compared to the organic carbon substrates from which *C. calidirosea* T49^T is capable of deriving energy.

Nitrogen uptake by T49^T was experimentally determined to be via ammonia assimilation and /or amino acid uptake (Lee et al., 2011). It appears that ammonia can be assimilated via glutamate or glycine and modified by a complete tetrahydrofolate one-carbon pathway. T49^T possesses two ammonia permeases (CCALI_00383 & 01636). Interestingly, genes encoding both assimilatory nitrite reductase components (*nirA* and *nirD*) were identified, but no other genes encoding for nitrate dissimilation or assimilation, nor aerobic or anaerobic ammonia oxidation were found. We experimentally tested the assimilation of nitrite and nitrate, but no positive growth was detected, suggesting that the putative nitrite reductase genes may instead function as a means for sulfite/sulfate uptake.

Previous growth experiments also showed that T49^T obligately requires supplementary amino acid addition in the form of either yeast extract or casamino acids. No conclusive pathway deficiencies were identified. However, the requirement for supplementary amino acids may be a result of regulation difficulties or loss-of-function mutations rather than defunct pathways. The genome of T49^T encodes only two amino acid transporters (Figure 5.3); a branched amino acid ABC transporter complex (CCALI_00418-20), and a second oligopeptide transporter complex (CCALI_01041-45) proximal to an isoleucine synthetase. We tested various amino acid combinations in basal nutrient medium, including the branched amino acids valine, leucine, and isoleucine, and determined that isoleucine alone can meet the amino acid requirement of T49^T for growth. As a side-note, we believe that these observations highlight the limitations of non-cultivation-based genome sequencing and associated genome prediction that has become prevalent since the advent of high-throughput sequencing technologies. An *in silico*-only based analysis of the genome would have had difficulty identifying this metabolic deficiency and any resulting cultivation attempts based on a non-cultivation-based genome assembly would have been unsuccessful based on the lack of supplementary amino acids in the enrichment medium.

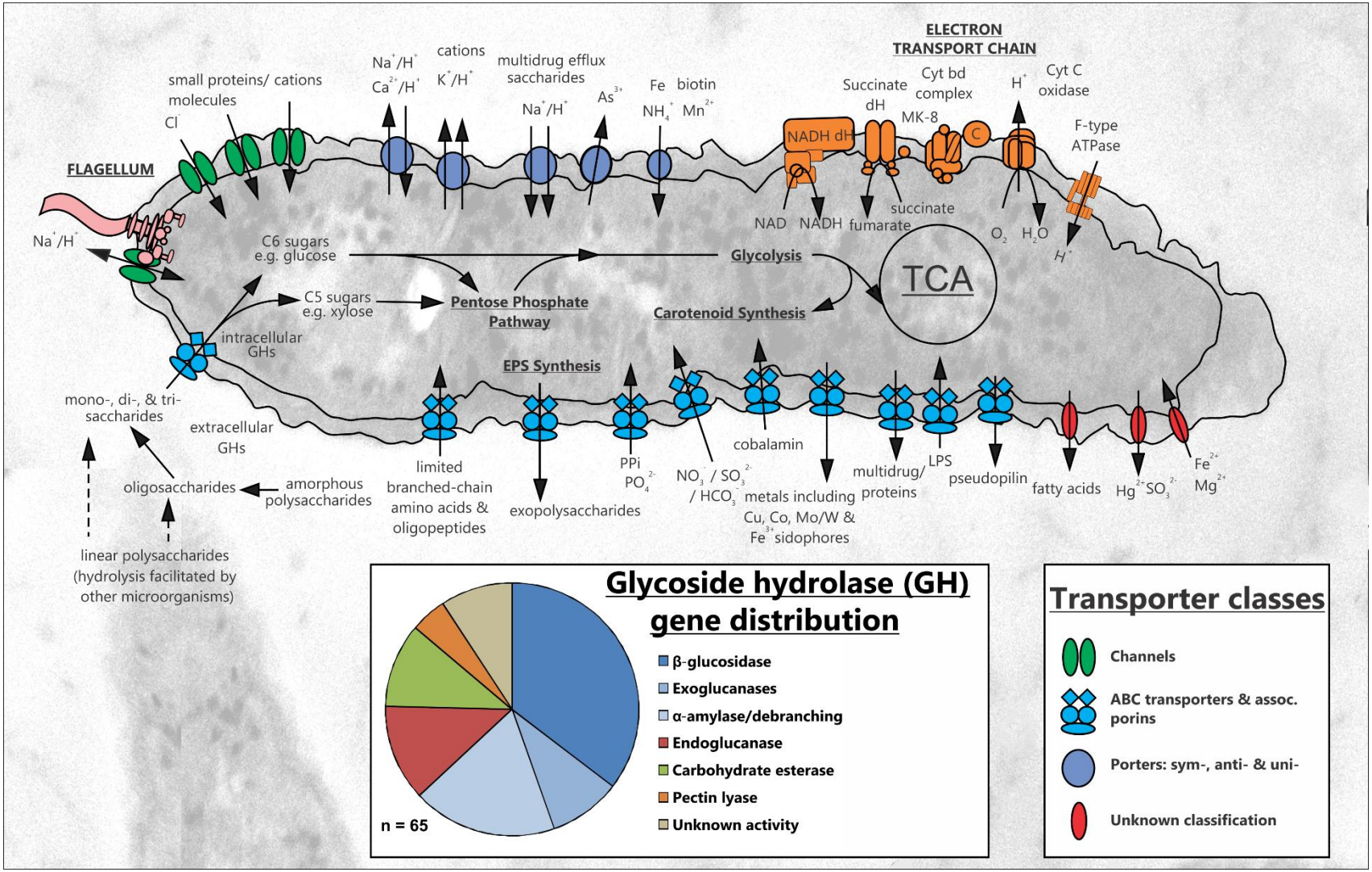


Figure 5.3 - The metabolic pathways of *C. calidirosea* T49^T illustrating predicted channels, ABC transporters and associated proteins, symporters, antiporters and uniporters, and membrane transport proteins of unknown classification. A wide range of putative intracellular and extracellular GHs have been identified. The large number of carbohydrate ABC transporters (n = > 27) highlight the scavenging capability of *C. calidirosea* T49^T for soluble carbohydrates, assisted by GH hydrolysis. The component sugars lead to the hexose and pentose pathways, glycolysis, TCA cycle, and the electron transport chain. The high copy number of prepilin genes (n = 57) may reflect the exportation of exopolysaccharides and the formation of biofilm by *C. calidirosea* T49^T.

5.5.6 - Secondary metabolism features

An interesting feature of T49^T is its growth within a narrow pH range, from 4.7-5.8 (Lee et al., 2011). This may result from a limited capacity to buffer cytoplasmic pH against external variation. A K^+ transporting ATPase complex (CCALI_01552-54) may serve to reduce the electrical component ($\Delta\Psi$) of the proton motive force and facilitate the acidophilic phenotype. However, as expected from its narrow pH tolerance range, nearly all of the well-known inducible pH homeostasis mechanisms (Jain & Sinha, 2009; Krulwich et al., 2011) were missing from the genome. There was no evidence of glutamate, arginine or lysine decarboxylases for dealing with acid stress, nor of a urease system, agmatine deiminase, or malolactic fermentation. The genome lacks any clear hydrogenases for proton removal with the possible exception of CCALI_00215, a protein of unknown function that contains several putative hydrogenase-like-homolog domains. T49^T cannot grow on acetate or lactate and lacks acetyl-CoA synthetase, so small organic acids might easily uncouple the membrane potential. There is no tryptophanase to compensate for alkali stress. However, T49^T does have several other amino acid deaminases and two NhaP type Na^+/H^+ (or K^+/H^+) antiporters (CCALI_00262 & 02512) that may respond to alkali stress. In general, the genome reflects a bacterium adapted to a pH-stable environment, with limited ability to respond to pH changes.

Another feature of T49^T is the pink/orange pigmentation of cells associated with pellicle formation in aqueous medium (Lee et al., 2011). No pigmentation or pellicle formation has been noted in lag and exponential phase growth (including growth on solid medium). However, in stationary or decline phase, cells become pigmented and aggregate rapidly. This change can be brought about by low oxygen saturation, or more commonly, by medium acidification due to carbohydrate hydrolysis, with a concomitant increase in carotenoid production. We have identified multiple genes related to carotenoid biosynthesis, including two phytoene desaturases (CCALI_00263 & 00231), a phytoene/squalene synthetase (CCALI_01286), a ζ-carotene desaturase (CCALI_01316), and a putative chlorobactene glucosyltransferase (CCALI_02059). Like many other metabolic features of the genome, these carotenoid biosynthesis genes were not arranged in operons as seen in other bacterial species. The spectrophotometric profile of the T49^T acetone carotenoid extract is similar to that of the primary *Thermomicrobium roseum* carotenoid, oscillaxanthin (Jackson et al., 1973; Wu, Raymond, et al., 2009). However, no homologs were identified for the 1,1′ hydroxylase-acyl transferase fusion gene of *T. roseum*, nor for *cruF* and *cruD*, the non-fusion equivalents in *Salinibacter ruber*, which play a key role in the modification of lycopene to oscillaxanthin. In addition, a wide array of genes present in the genome appears to be related to the production of exopolysaccharides, consistent with the observed pellicle formation. An operon-like gene

cluster starting with exopolysaccharide synthesis protein (CCALI_01270) was identified. The cluster contains two polysaccharide export-related proteins (CCALI_01265 and 01269), a glycosyltransferase (CCALI_01268), an EpsI family protein (CCALI_01266), and a transmembrane exosortase (CCALI_01267). Together in a putative operon, these components bear some functional resemblance to known exopolysaccharide operons such as the '*eps*' genes in *Bacillus subtilis* (Nagorska et al., 2010). Additional exopolysaccharide-related genes were also identified outside this cluster, including exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase (CCALI_02642) and two putative alginate export channels (CCALI_00999 and 02397).

5.5.7 - Regulation of carbohydrate metabolism

Carbohydrate utilisation experiments were conducted to test carbon catabolite repression in T49^T. Growth was tested in various combinations of two-carbohydrate media (glucose + mannose, glucose + galactose, glucose + lactose, and glucose + xylose). No diauxic shifts were observed and carbohydrates were simultaneously utilised in all of the carbohydrate combinations tested (Supplementary Figure 8.4), indicating a lack of carbon catabolite repression and non-diauxic growth. Non-diauxic growth has been observed previously in *Sulfolobus acidocaldarius* and *Thermoanaerobacter thermohydrosulfuricus* and was attributed to the lack of a carbohydrate phosphotransferase system (PTS) (Cook et al., 1993, 1994; Joshua et al., 2011). The T49^T genome also lacks any transporter genes associated with PTS. Many other microbial species (including most *Archaea*) lack the PTS system (Koning *et al*., 2002; Silva *et al*., 2005), but the prevalence of diauxic vs. non-diauxic growth as a physiological trait within the bacterial and archaeal domains is currently not well described. Archaeal and bacterial species without PTS transporters employ alternative carbohydrate transporter systems, including ATP-binding cassette (ABC) transporters, and secondary transporters such as the major facilitator superfamily (MFS) and major intrinsic protein (MIP) (Koning *et al*., 2002; Silva *et al*., 2005; Joshua *et al*., 2011). These alternative transport systems are thought to be less involved in the regulation of metabolism and transcription than PTS (Saier, 2001). Eleven MFS and 99 ABC transporter-related genes have been identified in the T49^T genome (Supplementary Table 8.12). Of the 99 ABC transporter genes identified, at least one quarter (27 genes) are putatively involved in carbohydrate transport. The presence of these alternative carbohydrate transporters may explain the concurrent importation of carbohydrates in the absence of PTS transporters, and thus the lack of diauxic growth in the two-carbohydrate conditions tested. MFS transporters were less abundant than ABC transporters, with only nine genes related to carbohydrate transport. No MIP transporters were identified.

5.5.8 - Carbohydrate-active enzymes

The genome contains at least 65 genes encoding CAEs including GHs, cellulose esterases (CE), and a pectin lyase (PL) (Supplementary Table 8.13), but excluding glycosyl transferases. The GHs belong to 26 described GH families; a further 16 putative CAEs were not associated with currently-described families from the CAZy database (Cantarel *et al*., 2009).

Previous metabolic characterisation of T49^T demonstrated that it can grow on a broad array of simple hemicellulosic (C5) and cellulosic (C6) carbohydrates, including monosaccharides, oligosaccharides, and amorphous polysaccharides (e.g., starch, xylan, glucomannan, pectin, and carboxymethyl cellulose) (Lee *et al*., 2011). In contrast, cells were unable to hydrolyse linear polysaccharides such as cotton, Avicel^{TM}, or lignocellulosic pulp preparations. The hydrolysis of linear or crystalline cellulose typically requires the synergistic action of endoglucanases, exoglucanases, and β-glucosidases in order to obtain complete degradation (Lynd *et al*., 2002; Wang *et al*., 2011). In particular, endoglucanases from GH families 9 and/or 48 appear to be obligately required for linear cellulose hydrolysis (Tolonen *et al*., 2009; Olson *et al*., 2010), although GH families 5, 6, 12, 44, 45 and 74 have also been reported to be active against polysaccharides (Kaoutari *et al*., 2013). In agreement with the observed carbohydrate hydrolysis activity of T49^T, few putative endoglucanase GHs were detected. Of these, seven endoglucanases/mannanases (GH5, 6 and 44) and three β-1,4-xylanases (GH10) were detected and confirm the utilisation capability of glucomannan and components of xylan for growth. In addition, three PLs and two families of CE were detected and these most likely act upon plant cell wall polysaccharides. However, consistent with observations that T49^T frequently lacks operon-based genome organisation, we were unable to identify polysaccharide utilisation loci previously reported (Martens *et al*., 2011) as essential for the degradation of plant pectins. The presence of endoglucanase-acting mannanases, xylanases, and pectin lyases may reflect an adaption to the elevated mannan and hemicellulose contents of New Zealand native plants and exotic pines which are abundant in the geothermal locations from which other *Chthonomonas* isolates were isolated (Stott *et al*., 2008).

Nearly two-thirds of the annotated T49^T GHs were related to β-glucosidases, exoglucanases, and/or de-branching/starch-hydrolysing activities (Supplementary Table 8.13). These enzymes include β-galactosidases (GH families 2 and 42; 10 copies), α-arabinofuranosidases (GH family 51; four copies), and α and β-amylases (GH families 13, 14 and 57; six copies). The high number of genes encoding these enzymes and the lack of crystalline cellulose-degrading endoglucanases suggests that *C. calidirosea* T49^T may rely on the cellulolytic activity of other microorganisms to supply carbohydrate oligomers. A similar scenario may also apply during the degradation of the pectin and xylan components of complex plant wall polysaccharides, as we were unable to identify the full complement of genes reportedly required for complete hydrolysis. Thus, it is reasonable to infer that T49^T does not act as a primary biomass degrader in its host ecosystem, but rather forms consortia with other soil-based cellulolytic bacteria to allow for complete hydrolysis.

5.5.9 - Inferred ecology

In combination with the results of previous community and physiological studies (Dunfield et al., 2012; Lee et al., 2011; Stott et al., 2008), the genomic analysis presented here has allowed us to more accurately define the ecological role of T49^T and provided some insight into the ecology of the phylum *Armatimonadetes*. Physiological data, characteristics of carbohydrate-active enzymes, and a large array of carbohydrate transporters all indicate that T49^T has a chemoheterotrophic carbohydrate-based primary metabolism that targets soluble/amorphous carbohydrates.

The inability of T49^T to directly utilise the complex and predominantly plant-based polysaccharides that are the primary carbohydrate source in its host environment suggests that it occupies its ecosystem niche as a scavenger, relying on the diffusion of various hydrolysis products generated by cellulose digesters, and possibly even acting as an oxygen barrier for cellulolytic anaerobes such as *Clostridium thermocellum*. In addition, T49^T may contribute to a mutualistic relationship by removing C5 sugars (components of hemicellulose which many cellulolytic species cannot utilise) from the environment, thus facilitating further cellulose degradation. The regulation of carbohydrate utilisation in T49^T does not appear to function via the standard well-described operon-based models of *Escherichia coli* and *B. subtilis*, but may reflect an ecological role as a scavenger in an environment in which the spectrum of available carbohydrate species is highly heterogeneous, in low concentration, and in flux. Such environments would be detrimental to metabolisms adapted to maximise the utilisation of a single carbon source over others. Despite the relative lack of operons, the high abundance of σ-factors suggests that T49^T can coordinate functionally-related-but-dispersed genes to rapidly respond to

ecosystem fluxes. In particular, the numerous ECF σ-factors may play a significant role in environment sensing (Butcher et al., 2008) and facilitation of biofilm formation (Bordi & de Bentzmann, 2011). Similar to many other thermophilic aerobic bacteria, T49^T produces carotenoids, likely to be regulated through stress response σ-factors in response to oxidative stress caused by the hostile environment. Surprisingly, despite the propensity for HGT within soil and biofilm environments, T49^T showed few obvious signs of HGT, and the genome topology based on GC skew gave no evidence of recent genomic rearrangement, despite a genome frequently lacking in typically conserved operon structures.

The physiology of T49^T suggests a tight coupling with its environment due to its narrow pH range and specific nutrient (carbohydrate and branched amino acids) requirements. In addition, geothermal environments within the suitable pH range (~4-5) for T49^T are uncommon, as typical geochemical interactions cause a bimodal distribution of pH in geothermal features, with peaks at pH 2 and 7 (Brock, 1971). The combination of the fastidiousness of T49^T and the rarity of suitable environments may explain why the species evaded isolation for almost a decade following the identification of the first phylotype within *Armatimonadetes*/Candidate Division OP10 (Hugenholtz, Pitulle, et al., 1998).

5.5.10 - Conclusion

Our analysis of the T49^T genome has revealed adaptations of a thermophilic bacterium in a complex geothermal soil-biofilm environment, as well as the first glimpse into the genomic landscape of the novel phylum *Armatimonadetes*. Currently, T49^T is the only complete *Armatimonadetes* genome that is publicly available, and it has undergone extensive automated and manual improvement to remove gaps and resolve problematic regions, in order to conform to the definition of "improved high quality draft" (Chain et al., 2009). The prospect of additional *Armatimonadetes* genomic data (in particular the two known characterised species, *A. rosea* YO-36^T (Tamaki et al., 2011) and *F. ginsengisoli* GSoil348^{Tf} (Im et al., 2012), provides an exciting opportunity to further understand common genomic features and evolution of this newly-described phylum.

[f] The genome of *F. ginsengisoli* GSoil348^T was recently published (Hu et al., 2014). A brief comparison with *C. calidirosea* T49^T is outlined in Section 6.8 - Addendum of the next chapter.

Chapter 6

Comparative Genomics and Metagenomics of Geographically-Diverse *Chthonomonas calidirosea* Isolates

Kevin C. Lee[a,b], Matthew B. Stott[a#], Peter F. Dunfield[a,c], Curtis Huttenhower[d,e], Ian R. McDonald[b], and Xochitl C. Morgan[a,d,e]

[a] GNS Science, Extremophiles Research Group, Wairakei Research Centre, Taupō, New Zealand

[b] School of Science, University of Waikato, Hamilton, New Zealand

[c] Department of Biological Sciences, University of Calgary, Calgary, AB, Canada

[d] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[e] Broad Institute of Harvard and MIT, Cambridge, MA, USA

6.1 - Preface

The overall goal of the research outlined in this chapter was to assess variations within the genomes of *C. calidirosea* and to correlate these to changes within its habitat. Therefore, this chapter documents the comparative analysis of genomes and physiological profiles of *C. calidirosea* isolates, as well as the microbial communities and geochemistry of their host environments. The chapter consists of the manuscript submitted to the journal *Applied and Environmental Microbiology*, along with supplementary information (Section 8.3). Supplementary tables too large for printed form are referenced to digital files attached to this thesis. Genomic data deposited in publicly-available databases are referred to using accession numbers.

The genome of *Chthonomonas calidirosea* T49^T was characterised in the previous chapter. The analysis of this genome supported the ecological role of the bacterium as a versatile carbohydrate scavenger. The analysis noted that the genome appeared disorganised, where many genes commonly in conserved operons were found scattered within T49^T genome. Rapid genomic diversification due to close metabolic coupling with the immediate environment was postulated to be the underlying cause.

In order to test this new hypothesis, the divergence of *C. calidirosea* genomes among different environments was analysed. We conducted an integrated analysis incorporating multiple data types (genomics, geochemistry, physiology, and community). Genomes from three additional *C. calidirosea* isolates (P488, WRG1.2, TKA4.10) from geographically-distinct sites within the TVZ, were sequenced, assembled, and analysed. Carbohydrate utilisation profiles of the four isolates (including T49^T) were compared. The structures of the associated microbial communities as well as the soil geochemistry

of the sample sites were also compared, in order to identify likely factors influencing the genomic composition of the species.

In contrast to the anticipated results, the genomes were highly similar even from geographically isolated geothermal sites. Neither the phylogeny of the isolates inferred from conserved genes, nor the pattern of isolate-variant genes corresponded to the patterns of soil chemistry or microbial community profiles of the sample sites. Rather, the evolutionary history of the isolates corresponded to the geographical relationship between the sites along the TVZ. We therefore rejected the hypothesis of environmental selection as the primary driver of the $T49^T$ genome, due to the very high genome conservancy between the isolates, and the single low abundance operational taxonomic unit (OTU) representing genus *Chthonomonas* shared among the sample sites. We propose that the homogeneity of genomes of the low-abundance *C. calidirosea* may instead be maintained between sites by a rapid aeolian dispersal, frequent localised extinction process. The findings contributes to a better understanding of the genomic dynamics of C. *calidirosea* as well as the dispersal of free-living bacteria between distinct suitable habitats.

As the primary author, my contributions to this project included the cultivation of all *C. calidirosea* isolates, the extraction of *C. calidirosea* genomic DNA for sequencing, assembly of the genomes, the preparation of soil samples for community amplicon sequencing, profiling of *C. calidirosea* carbohydrate metabolism, and the processing and analysis of all the data generated and writing of the majority of the manuscript. Matthew B. Stott, Peter F. Dunfield, Curtis Huttenhower, Ian R. McDonald and Xochitl C. Morgan contributed with experiment design, discussions, and editing of the manuscript. Xochitl C. Morgan also contributed by sequencing the genomes of isolate P488 and WRG1.2.

This chapter is been submitted and is currently under revision as an original research paper with the following citation:

Lee, **K. C**., Stott, M. B., Dunfield, P. F., Huttenhower, C., McDonald, I. R., & Morgan, X. C. Comparative genomics and metagenomics of geographically-diverse *Chthonomonas calidirosea* isolates. *Applied and Environmental Microbiology* (In Revision).

6.2 - Abstract

Chthonomonas calidirosea T49^{T} is a thermophilic bacterium with an apparently disorganised genomic structure and an ecological role of a low-abundance scavenger of carbohydrates in geothermal soil environments. We hypothesised that these traits are highly responsive to its immediate environment, resulting in the divergence of genomic content and organization. In order to test this hypothesis, we sequenced the genomes of four *C. calidirosea* isolates from the Taupō Volcanic Zone in New Zealand. We defined the structure of the associated microbial communities and the physicochemical attributes of the host ecosystems through 16S rRNA gene sequencing and X-ray and UV-Vis spectrometries. Comparative analyses showed that the genomes exhibited very low divergence (maximum 1.17 %) despite their ecological and geographical isolation. Variations among the isolates consisted of single nucleotide polymorphisms, isolate-specific restriction modification systems, and mobile elements. While the community abundance of *C. calidirosea* appeared to relate to variations in environmental characteristics, these factors did not result in substantial differentiation of genomic content. Conversely, similarity of carbohydrate utilization profiles between isolates appeared to follow their inferred phylogenies, which in turn resembled the geographical relationships between the sample sites. This suggests that the genotypes present in a site were primarily determined by stochastic dispersal rather than adaptive evolution. These comparative analyses point to aeolian dispersal and localised extinction as possible mechanisms for the conservation of *C. calidirosea* genomes, which allow for rapid population homogenization with little restriction by geographical barriers. These findings may illuminate the population dynamics for many low-abundance free-living microorganisms.

6.3 - Introduction

To date, the poorly-characterised phylum *Armatimonadetes* is described primarily by environmental 16S rRNA marker gene data (Dunfield et al., 2012; Hugenholtz, Pitulle, et al., 1998; Lee et al., 2013; Portillo et al., 2008) and has only three characterised type species (Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011), of which two genomes are sequenced (Hu et al., 2014; Lee et al., 2014). These genome sequences have revealed that *Armatimonadetes* are most closely related to the candidate lineages FBP, WS1, and to the phylum *Chloroflexi* (Lee et al., 2014, 2013). *Armatimonadetes* have three described classes, each represented by a single type strain: *Armatimonadia* (*Armatimonas rosea* YO-36^T) (Tamaki et al., 2011), *Fimbriimonadia* (*Fimbriimonas ginsengisoli* Gsoil 348^T) (Im et al., 2012), and *Chthonomonadetes* (*Chthonomonas calidirosea* T49^T) (Lee et al., 2011). The phylum contains at least additional eight class-level phylogenetic lineages without cultivated representatives (Dunfield et al., 2012).

Chthonomonas calidirosea strain T49^T was isolated from geothermal soil within the Taupō Volcanic Zone (TVZ), New Zealand (Stott et al., 2008), an area rich in geothermal systems and surface hydrothermal features. Strain T49^T is an aerobic, moderately-acidophilic thermophile. It produces extracellular polymeric substances as well as extracellular saccharolytic enzymes, which allow it to utilise most carbohydrates, with the exception of crystalline polymers such as cellulose (Lee et al., 2011). Genomic analysis of strain T49^T identified a wide range of glycosyl hydrolases and carbohydrate ATP-binding cassette transporters, as well as many extracytoplasmic function sigma factors (Lee et al., 2014). Previous studies concluded, based on genomic and physiological data, that the environmental role of *C. calidirosea* T49^T was likely that of a scavenger, utilizing heterogeneous carbohydrates from degraded biomass within the environment (Lee et al., 2014, 2011).

Interestingly, in strain T49^T, many gene functions for which component genes were typically organised into operons (e.g., histidine, tryptophan, and purine biosynthesis) instead had their component genes spread throughout the genome in an apparently disorganised manner. This genome disorganization complicated metabolic pathway predictions, although the abundant sigma factors observed in the genome provided a potential explanatory mechanism for gene regulation (Lee et al., 2014). We hypothesised that traits such as extensive carbohydrate utilization and genomic disorganization (potentially corresponding to frequent genomic rearrangement) would be highly responsive to the immediate environment, and that the selection pressures of local environments would therefore be reflected within the genomes of isolates of *C. calidirosea* from distinct geothermal sites. Therefore, in order to assess the degree of flux

within the *C. calidirosea* genome in response to environmental selection pressure, we compared the genomes of three additional *C. calidirosea* isolates cultured from geographically-isolated sites (averaging 46.2 km from the site of strain T49^T isolation) with T49^T. In addition, we collected environmental geochemistry data and sequenced the 16S rRNA genes from the microbial communities for an integrated understanding of the biotic and abiotic environment of each site.

Here we present a comparative study on the genomic contents and organisation of *C. calidirosea* isolates, their physiological profiles, and the microbial communities and geochemistry of their sample sites of origin. We found a highly-conserved core of 2797 genes and a total pangenome of 2945 genes across the four strains. Most functionally-annotated strain-specific genes were associated with restriction modification systems and other nucleotide-interacting enzymes. We analysed the phylogeny of the four strains based on 327 conserved gene sequences in their genomes. Physiologically, strains shared a highly similar carbohydrate metabolism profile, which also exhibited a pattern of pairwise distance similar to the inferred phylogenies and geographical distances between the sample sites. We found that *Archaea* dominated the microbial community composition at each of the sample sites, and that *C. calidirosea* was a minor bacterial species. Geochemically, the sample sites exhibited a range of geothermal influence and clay content. Overall, the pairwise similarity patterns of the above elements led us to conclude that the evolutionary history of the strains was more related to geographical distance than to the community or physicochemical profile of the sample sites. Our analysis of a range of comparative data suggests the hypothesis that most likely evolutionary forces on the genome dynamics of *C. calidirosea* are frequent stochastic aeolian dispersal and localised extinction, leading to the high conservation of genomic content and organisation in *C. calidirosea* isolates across the TVZ.

6.4 - Methods

6.4.1 - Cultivation of *Chthonomonas calidirosea* isolates

Figure 6.1 details the relationships of methods and data generated in this study. Based on the detection of *Armatimonadetes*/Candidate Division OP10 in clone libraries, four *C. calidirosea* pure culture isolates (T49^T, P488, WRG1.2, and TKA4.10) were isolated from four geographically-separated geothermal systems (Tikitere - TKT, Nov 2006; Waikite - WKT, July 2006; Wairakei - WRG, April 2009; Te Kopia - TKA, September 2010) within the TVZ. The isolates were cultivated employing the method and conditions as outlined by Stott and colleagues for a diverse range of oligotrophic carbohydrate-utilizing thermophiles (Stott et al., 2008).

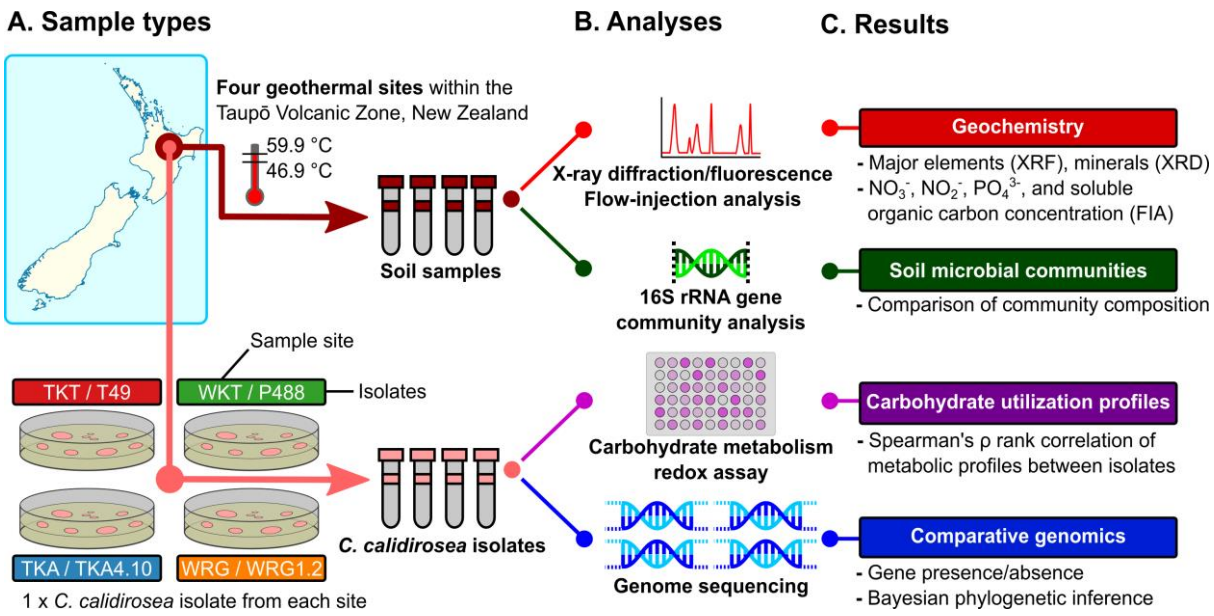


Figure 6.1 - Overview of data collection and analysis. (A) Soil samples were collected from four sites across the Taupō Volcanic Zone (TVZ): TKT, WKT, WRG, and TKA. *C. calidirosea* was isolated from each site. (B) The geochemistry of the soil samples was analysed by X-Ray diffraction/fluorescence (XRD/XRF), and flow injection analysis (FIA) to determine the mineral content. The DNA was extracted and subjected to 16S rRNA gene analysis to determine the composition of the microbial community. *C. calidirosea* isolates were metabolically profiled by BIOLOG PM1 phenotype microarray, and their genomes were sequenced. (C) The data types and / or analytical results of these methods are indicated in column C. The sample sites and the *C. calidirosea* isolates are colour-coded. The colour scheme is also used in other figures and tables in this paper.

6.4.2 - Genome sequencing, assembly, and quality assessment

6.4.2.1 Sequencing of C. calidirosea isolate genomes

Isolates P488, WRG1.2, and TKA4.10 were cultivated, and their genomic DNA was extracted using the phenol-chloroform method previously described for strain T49^T (Lee et al., 2014). The four genomes were sequenced via different technologies (Table 6.1). The reference genome T49^T was sequenced prior to this study using the 454 GS-FLX system with primer walking to resolve ambiguous regions (Lee et al., 2014). The genomes of isolates P488 and WRG1.2 were sequenced using the Illumina MiSeq platform. Libraries were constructed using the Nextera XT DNA Sample Preparation kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's protocol, and 2 x 150-bp paired-end sequencing was performed using a 300-cycle sequencing kit (v 1.0). The genome of isolate TKA4.10 was sequenced using the Ion Torrent platform (Thermo Fisher Scientific Inc., Waltham, MA, USA). The Ion Torrent library was constructed

using the Ion Xpress Plus Fragment Library Kit with 100 ng of input DNA. The beads were prepared using the Ion One Touch 200bp v2 DL kit, and the sequencing was performed with the Ion PGM 300 bp Sequencing Kit.

6.4.2.2 Assembly and annotation of isolate genomes

The genome of the type strain T49^T (Lee et al., 2014) was used as the template for mapping the assembly of isolates P488, WRG1.2, and TKA4.10 using MIRA assembler version 4.0rc3 (Chevreux et al., 1999). Each of the genomes was represented by a single scaffold, with regions unmapped by reads (indicating possible gene deletion compared with the reference genome or gaps in assembly) marked as repeating Ns (IUPAC notation). The resulting assemblies were aligned using progressiveMauve (Darling et al., 2010). Assembly fold coverage calculations (Table 6.1) excluded contigs ≥ 5000 bp to avoid inflation of the figures. The genome sequence and assembly data of the *C. calidirosea* isolates were deposited at EMBL-Bank with the following accession numbers: isolate P488 (PRJEB4907), isolate WRG1.2 (PRJEB4936), and isolate TKA4.10 (PRJEB4937). Gene prediction and annotation of the genomes were performed using the Integrated Microbial Genome-Expert Review (IMG-ER) pipeline (Markowitz et al., 2010).

6.4.2.3 Quality assessment of isolate genomes

Assembly paired-end length violation assessment was conducted using Hagfish (https://github.com/mfiers/hagfish/). Reads from paired-end sequencing were mapped back onto the assembled genome sequence. Read pairs mapped further away from expected insertion size range (assessed via Bioanalyzer) would indicate poor assembly due to problems such as repetitive regions or the mismatch between reference and the genome being assembled (i.e. genome rearrangement). This process excluded the TKA4.10 genome because the Ion Torrent sequencing was single-ended. Additionally, in order to assess assembly bias due to using MIRA and T49^T as mapping reference, the SPAdes 3.1 (Bankevich et al., 2012) assembler was used to assemble contigs *de novo* for P488, WRG1.2, and TKA4.10 genomes. The resulting contigs were mapped against the T49^T genome with CONTIGuator2 (Galardini et al., 2011).

6.4.3 - Genome comparison and phylogenetic analysis

6.4.3.1 Identification of gene homologs across C. calidirosea isolates

In order to identify isolate-specific genes among *C. calidirosea* genomes, a reciprocal BLASTP best-hit search was conducted using the "Phylogenetic Profiler" function in the IMG-ER system (Markowitz et al., 2012) with default search parameters. To validate

uniqueness of functionally annotated genes, BLASTN alignments against subject genomes were manually inspected.

6.4.3.2 Identification of putative genes in unmapped reads

Potential genes omitted in the mapping assembly with strain T49^T as the reference genome were identified via *de novo*, per-genome, MIRA assembly of reads left over from initial mapping assembly. Resulting contigs ≥ 500 bp in length and had more than 1/3 of average coverage of the *de novo* assemblies were searched against the NCBI nr database using default BLASTN parameters in order to locate putative genes that were unassembled and not present in T49^T. The selection criterion was chosen because shorter or lower coverage nucleotide sequences are more likely to represent repetitive motifs instead of actual genes. The identification of potential unassembled genome content, with no homologs found in T49^T, acted as a means to assess bias introduced from using the reference genome for assembly.

6.4.3.3 Determination of isolate phylogeny

The PhyloPhlAn gene collection of 400 ubiquitous and phylogenetically-informative genes (Segata et al., 2013) was used for phylogenetic analysis of isolates. The *C. calidirosea* genomes contained 327 of these genes (Supplementary Table 8.14). Due to the high similarity of the isolates, nucleotide sequence-based Bayesian phylogeny inference was conducted (instead of filtered amino acid sequence used by PhyloPhlAn) to increase phylogenetic resolution at an isolate level and to enable the assessment of nucleotide sequence divergence (substitution per site). Each gene was first separately aligned (aligning the four homologs of the gene from the four isolates) using MUSCLE (Edgar, 2004) without the removal of invariant sites. The multiple alignments were then concatenated and used for phylogenetic inference with MrBayes (Huelsenbeck & Ronquist, 2001). The genes were divided into partitions and were unlinked so that each partition had its own set of parameters for estimation during the Markov chain Monte Carlo (MCMC) process. Each partition used the Generalised Time-Reversible substitution model with heterogeneity rate determined by the invariable site plus gamma distribution (GTR + I + Γ). The model was selected using jModeltest (Posada, 2008) with Akaike information criterion (AIC). Two MCMC runs were conducted. Each MCMC process ran for 140 million generations, sampling every 1,000 generations, with 35 million (25 %) generations as burn-in. MCMC performance diagnostics were conducted via Tracer (http://beast.bio.ed.ac.uk/software/tracer/).

6.4.3.4 Genomic rearrangement comparison with other thermophiles

Genomic sequences of *Sulfolobus islandicus* (strain Y.N.15.51 and Y.G.57.14) and *Thermus thermophilus* (strains HB-8, HB-27 and JL-18) strains were downloaded from NCBI GenBank. In order to compare the degree of genomic rearrangement observed in other thermophilic species isolated from defined locations, the genomes of each species (including *C. calidirosea*) were aligned using progressiveMauve (Darling et al., 2010)

6.4.4 - Characterisation of *C. calidirosea* isolate metabolism with BIOLOG phenotype microarrays

All four *C. calidirosea* isolates were maintained using 4.5NZS solid medium (Stott et al., 2008). Colonies collected from the medium after one week of incubation at 60 °C were used to inoculate 250 mL of 4.5NZS liquid medium (pH 5.5), with 3 g/L maltose and 0.2 g/L casamino acids in a 500 mL container i.e. 1:1 (by volume) air headspace/medium ratio in total vessel volume. The liquid cultures were incubated at 60 °C for ~40 h. 50 mL of cells were centrifuged (5190 g for 15 min) and washed with sterile water three times before resuspension in a maltose-free 4.5NZS liquid medium to an $OD_{600\ nm}$ of 0.6 - 0.8. An 100 µL aliquot of the resuspended culture and 5 µL (5 g/L) MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) was added to each of the 96 wells on a BIOLOG PM1 plate (Biolog, Inc., Hayward, CA, USA). Each well was thoroughly mixed using a pipette to ensure uniform distribution of the cells and medium components. The plates were sealed and incubated in darkness at 60 °C for ~23 h. After incubation, 900 µL of dimethyl sulfoxide was added to each well to solubilise the formazan formed from the reduction of MTT (Mosmann, 1983). The mixture was measured with a visible spectrophotometer (Dynamica HALO VIS-10) at 540 nm. Each isolate was tested on two PM1 plates from separate cultivation batches as experimental replicates. Mean responses were calculated from the corresponding wells of each isolate, and tie-corrected Spearman rank-order correlations between isolates were calculated with the R-function cor.test (R Core Team, 2013).

6.4.5 - Physicochemical analyses of soil samples

Soil temperature was measured on-site using a Fluke 50S thermocouple. Sample pH was measured in the laboratory at 25 °C, by mixing 1 g of the sample in 10 mL of deionised water.

Analysis of moisture content and major oxides was performed by CRL Energy Spectrachem Analytical (Wellington, New Zealand) via X-ray fluorescence (XRF). The moisture content of the samples was measured through loss on drying (i.e. change in

sample weight after drying in an oven at 110 °C until a constant weight is achieved). Major oxides were detected by XRF with borate fusion of the samples. Loss on ignition of the samples was also quantified in the process.

Soluble organic carbon analysis was conducted by Landcare Research (Palmerston North, New Zealand). Samples were transported overnight with icepacks to prevent degradation of the organic carbon content. Each sample was mixed with 25 mL of 0.05 M K_2SO_4 and filtered to extract total soluble carbon. Total carbon in the solution was measured by a Multi N/C 3100 analyser (Analytik Jena AG, Jena, Germany), which measures the carbon dioxide generated through the combustion of the solution with a cerium dioxide catalyst. The inorganic carbon content of the sample was measured separately by quantifying the carbon dioxide generated by the acidification of an aliquot of the original sample (American Public Health Association, 2005). Soluble organic carbon is expressed as mg per gram of unmodified soil.

X-ray diffraction (XRD) analysis to measure mineral content was conducted by GNS Science (Wairakei, New Zealand). Samples were prepared using zinc oxide (10 % w/w) as an internal reference. XRD spectra were collected using a Philip X'Pert Pro instrument with a CoKα radiation source. The mineralogy of the samples was determined using the interpretive software X'Pert High Score (Spectris plc, England). Quantitative mineral percentages were determined using the quantitative software Siroquant (Sietronics Pty. Ltd. Canberra, Australia). Nitrate, nitrite, ammonia, and dissolved reactive phosphate in the soil samples were determined by flow injection analysis (FIA). Analytes were extracted by mixing 1 g of soil sample with 50 mL of 1 M KCl. The mixture was centrifuged and the resulting supernatant was passed through a 0.2-µm filter. The concentrations of the aforementioned ionic species were determined by GNS Science (Wairakei, New Zealand) using QuikChem Method 31-107-04-1-A, Method 31-107-06-1-B and Method 31-115-01-1-H with an automated Lachat QuikChem Series 2 Flow-injection system (Hach Company, Loveland, CO, USA).

6.4.6 - Community 16S rRNA gene-targeted sequencing and processing

Total soil DNA was extracted with a NucleoSpin Soil DNA kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany). Each soil sample was mixed with 500 µL of 50 g/L sterile skim milk solution (Becton, Dickinson and Company, NJ, USA) to reduce DNA binding to clay soil before extraction as per the manufacturer's instructions. Fusion primers were constructed to amplify a 350-bp amplicon on the V4 16S rRNA gene region, using the "universal" prokaryotic primers f515 and r806 (Caporaso et al., 2012). The forward fusion primer consisted of the "A" adaptor, an Ion Xpress barcode sequence, a GAT

barcode adaptor and the f515 primer. The reverse fusion primer contained the P1 sequence and r806 primer (Myllykangas et al., 2012).PCR reactions were performed in 15-μL reactions, each containing 0.304 μL (10 μM concentration) of forward and reverse primers, 0.61 μL of BSA (10 mg/mL), 1.83 μL of dNTPs (8 mM), 1.83 μL of 10X PCR buffer, 1.83 μL of $MgCl_2$ (50 mM), and 1 μL of soil community DNA as a template. Prior to the addition of primers and template, the PCR master mix was treated with 0.53 μL of ethidium monoazide bromide (1/200 dilution from 1 mg/mL stock) per reaction, and with 1 min of photoactivation using a lamp to remove trace DNA contaminants (Rueckert & Morgan, 2007). The PCR reagents, including ethidium monoazide bromide, were supplied by Invitrogen (Invitrogen, Carlsbad, CA, USA). PCR was conducted using the following thermocycling parameters after the initial 3 min of 94 °C denaturation: 30 cycles (94 °C, 45 sec; 50 °C, 1 min; 72 °C, 1.5 min), followed by a final 10 min incubation at 72 °C.

The amplicons from triplicate PCR products were pooled and purified using SPRIselect (Beckman Coulter Inc.) to remove small nucleic acid fragments (left-hand selection). Amplicon concentration and quality were verified and adjusted to 26 pM with a Qubit fluorimeter (high-sensitivity) (Life Technologies) and a 9100 BioAnalyzer (Agilent Technologies) prior to pooling amplicons from all samples together for emulsion PCR. DNA sequencing was conducted at the Waikato DNA Sequencing Facility (University of Waikato, New Zealand) using Ion Torrent PGM (Life Technologies) with an Ion 318v2 chip, using 400 bp chemistry.

UPARSE (Edgar, 2013) was used to process 216,944 raw sequencing reads, with the following workflow to remove anomalous sequences. First, reads with lengths outside 276-344 bp range were removed. Next, the forward primer and barcode segments of the reads were trimmed off, and the resulting reads were subjected to Q-score filtering to remove reads with maximum expected error > 3. Singleton reads were then removed, and the sequences were globally trimmed to 250 bp. A total of 85,063 of 85,063 high quality sequences were obtained after quality control from the four sample sites, with a mean of 20,733 sequences and a range of 17,941 to 25,656 sequences per sample. The sequences were clustered *de novo* into operational taxonomic units (OTUs) with minimum identity of 97 %, and the representative sequences were taxonomically assigned using UCLUST consensus taxonomy assigner implemented in QIIME 1.8.0 (Caporaso, Kuczynski, et al., 2010). Greengenes release 13_8 (DeSantis et al., 2006) was used as the reference taxonomic database, including the classification of candidate taxa (e.g., YNPFFA and MBG-A). Beta diversity was measured through Bray-Curtis dissimilarity and weighted and unweighted Unifrac indices. A weighted-UniFrac hierarchical clustering tree was

constructed using the unweighted pair group method with arithmetic mean (UPGMA). Jackknife analysis with the UPGMA tree was conducted with 200 repetitions and 12,000 sequences per sample. Beta-diversity calculations and downstream analyses were conducted using QIIME (Caporaso, Kuczynski, et al., 2010). The OTU table was rarefied to 17,941 sequences per sample to remove sample heterogeneity. For phylogenetic-based metrics, PyNAST alignment (Caporaso, Bittinger, et al., 2010) and FastTree (Price et al., 2010) were used to generate the required phylogenetic trees.

6.5 - Results

In order to understand the influence of the environment on the evolution of the *C. calidirosea* genome, we sequenced the genomes of four isolates from geographically isolated geothermal systems, as well as analysing the soil chemistry and microbial community structure of each sample site (Figure 6.1)**.** For each isolate genome, we performed nucleotide sequence divergence assessment, phylogenetic inference, and identified gene presence/absence patterns among the genomes. Finally, we investigated the physiological profile of the isolates. We aimed to connect genome dynamics with external factors and to assess the influence of genome divergence on phenotypic similarities between isolates in order to better understand *C. calidirosea* and its ecological roles.

6.5.1 - Genome content and organisation shows high levels of conservation between isolates

6.5.1.1 Assembly

The genomes of *C. calidirosea* isolates P488, WRG1.2, and TKA4.10 were sequenced using the Illumina MiSeq and Ion Torrent platforms (Table 6.1) and assembled using the previously sequenced T49^T genome (Lee et al., 2014) as a scaffold. The 16S rRNA genes (each genome has two copies) were identical to one another, with the exception of P488, which has 1 base difference in both copies. The TKA4.10 genome shared the same assembly length and GC content as T49^T, which is likely due to assembly bias from the reference genome and the lack of information from single-ended Ion Torrent reads to extend past the reference contig boundaries, while P488 and WRG1.2 had slightly longer genome assemblies and lower GC % (Table 6.1).

Table 6.1 - Genome statistics of the four *Chthonomonas calidirosea* isolates.

Isolates	**T49^T ***	**P488**	**WRG1.2**	**TKA4.10**
Source	Hell's Gate, Tikitere, North Island, NZ	Waikite Valley, North Island, NZ	Wairakei, North Island, NZ	Te Kopia, North Island, NZ
Sequencing technology	454 Titanium + Sanger	Illumina MiSeq	Illumina MiSeq	Ion Torrent
Reads assembled	171,649	2,533,602	3,272,279	954,245
Fold Coverage	20 x	88 x	112 x	39 x
Genome Size (bp)	3,437,861	3,438,278	3,438,088	3,437,861
GC %	54.41	53.93	53.57	54.41
Assembled protein coding genes	2877	2876	2855	2885
5S, 16S, 23S rRNA genes	1,2,1	1,2,1	1,2,1	1,2,1
tRNA genes	46	46	46	46

6.5.1.2 Shared synteny between C. calidirosea genomes

The quality and synteny of paired-end sequenced P488 and WRG1.2 genome assemblies was assessed by measuring the distance of paired-end reads mapped onto the respective assembled genome sequences. The vast majority of the genomes consisted of mapped read-pairs within the insertion size range of the Illumina libraries at high sequencing coverage (Supplementary Figure 8.5), indicating actual shared synteny over the majority of genomic regions between P488 or WRG1.2 and the reference T49^T genome, rather than bias introduced from the assembly backbone. Nevertheless, the mapped genome coverage plot also showed a small number of regions with poor read coverage and regions with mapped reads longer than expected pair-end length, indicating possible hypervariable regions, or genomic rearrangement breakpoints. Indeed, mapping assembly would not be able to assemble contigs substantially different to the reference genome into the new scaffold. To address this issue, we also assessed contigs that were not assembled into the scaffold (see next section).

6.5.1.3 Validation of assembly method

Additionally, in order to assess potential biases from using MIRA assembler and the reference genome, the consensus sequence of the mapped assemblies were compared with *de novo* assembled contigs (assembled using SPAdes) (Supplementary Figure 8.6). The contigs showed no indication of genome rearrangement breakpoints and few putative insertions/deletions (indels). Most of the putative indels were located at the edge of the

contigs, suggesting assembly artefacts due to gaps in the reference genome or repetitive sequences. The three large putative insertions found only in P488 contained genes found in T49^T (e.g., hydroxypyruvate isomerase and prepilin protein), which may suggest assembly error or potential gene duplication sites. Overall, the validation from SPAdes assembly exhibited high consistency with the results of MIRA, indicating high overall genome synteny and low isolate-variant gene contents.

6.5.1.4 Functionally-annotated isolate-variant genes are primarily DNA-interacting enzymes

We assessed the similarity of genomic content between the genomic assemblies by using reciprocal best-hits BLASTP analysis of predicted protein sequences (see Methods). The majority of homologs (2797) were present in all four *C. calidirosea* isolates. Through this automated process, we identified 148 putative isolate-variant genes (i.e. not present in all four isolates), including genes with no known functional association, such as hypothetical proteins and domains of unknown function (Supplementary Table 8.15). We manually confirmed gene presence and absence of all functionally annotated genes (excluding hypothetical proteins) by using BLASTN to identify either the presence or absence of each gene within each genome, and the presence of its orthologous flanking regions (Supplementary Table 8.16). With the exception of a β-1,4 xylanase absent in the P488 genome, the majority of these 17 isolate-variant genes (Figure 6.2) were putative nucleic acid-interacting enzymes such as DNA-methyltransferases and restriction enzymes. Isolates TKA4.10 and T49^T shared the highest overlap in annotated, isolate-variant genes, followed by isolate P488. This trend of gene overlap holds true for putative genes in Supplementary Table 8.15 as well.

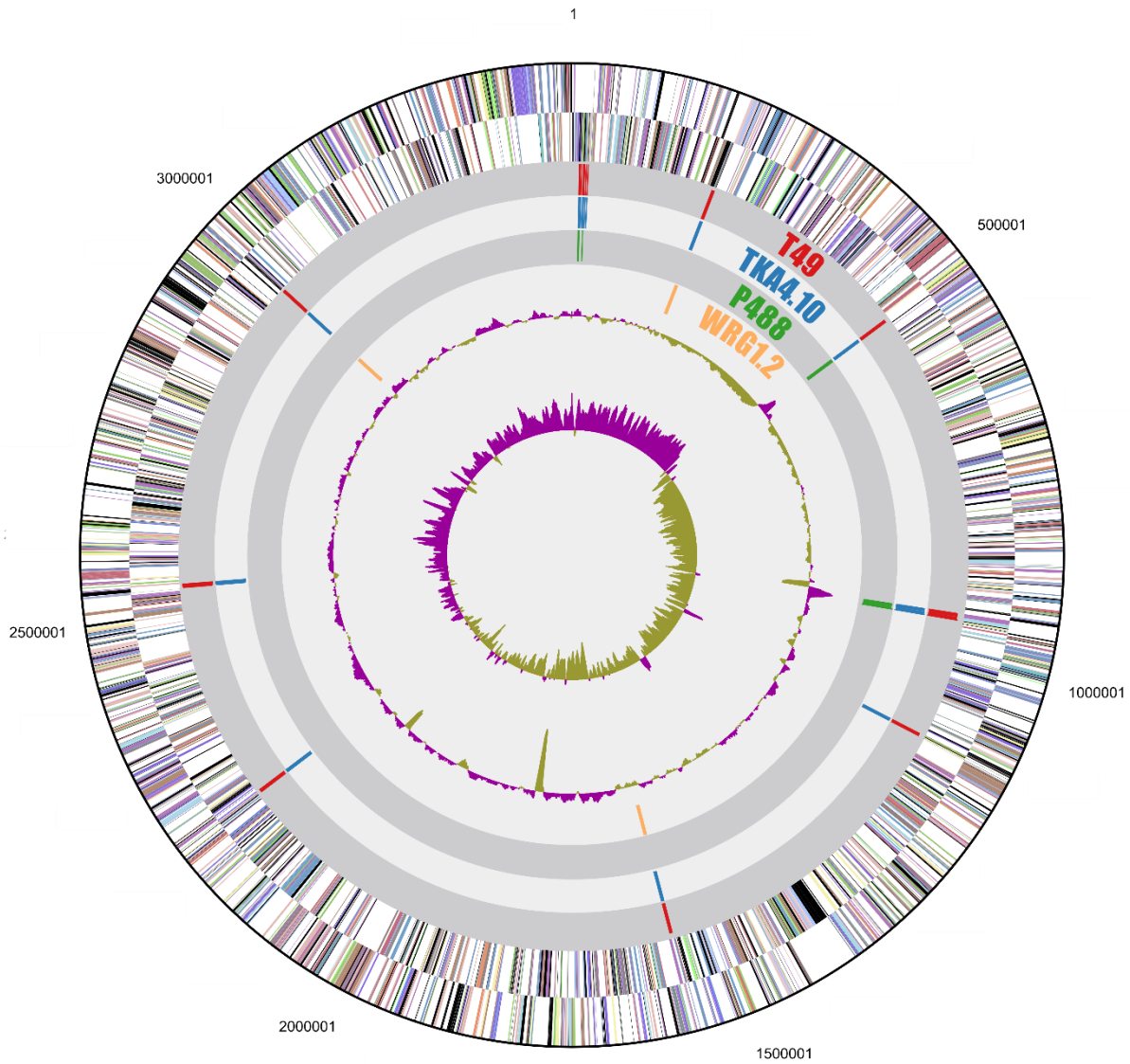


Figure 6.2 - Isolate-variant genes in the *C. calidirosea* isolates. *C. calidirosea* isolate-variant genes are plotted on the graphical circular map of T49^T reference genome. Only the genes with functional annotations are shown. From inside to outside, the rings indicate: (1) GC skew (2) GC content; isolate-variant genes that are included in the genomes of isolates (3) WRG1.2 (orange), (4) P488 (green) (5) TKA4.10 (blue), (6) T49^T (red); genes on T49^T genome coloured by COG category, (7) reverse strand, and (8) forward strand.

We assessed the genomic content from reads that were not assembled into scaffolds using T49^T genome, and potentially absent in the reference genome, by conducting a *de novo* assembly of these unassembled reads for each genome and BLASTN search with NCBI nr database. This resulted in 14 additional putative functionally-annotated genes absent from T49^T (Supplementary Table 8.16). Using this approach, we identified a gene similar to Mrub_2453 from *Meiothermus ruber* DSM 1279 (Tindall et al., 2010), which was shared by P488 and WRG but not T49^T and TKA, in contrast to the typically-observed

gene presence-absence pattern of reduced genome content from the former two isolates. The identification of these genes showed that gene gain/loss was not entirely directional (i.e. does not follow the pattern of decreasing gene content of T49 - TKA4.10 - P488 - WRG), and may indicate an indel event separating the T49^T - TKA4.10 and P488 - WRG1.2 clade seen in the inferred phylogenetic tree (Figure 6.3).

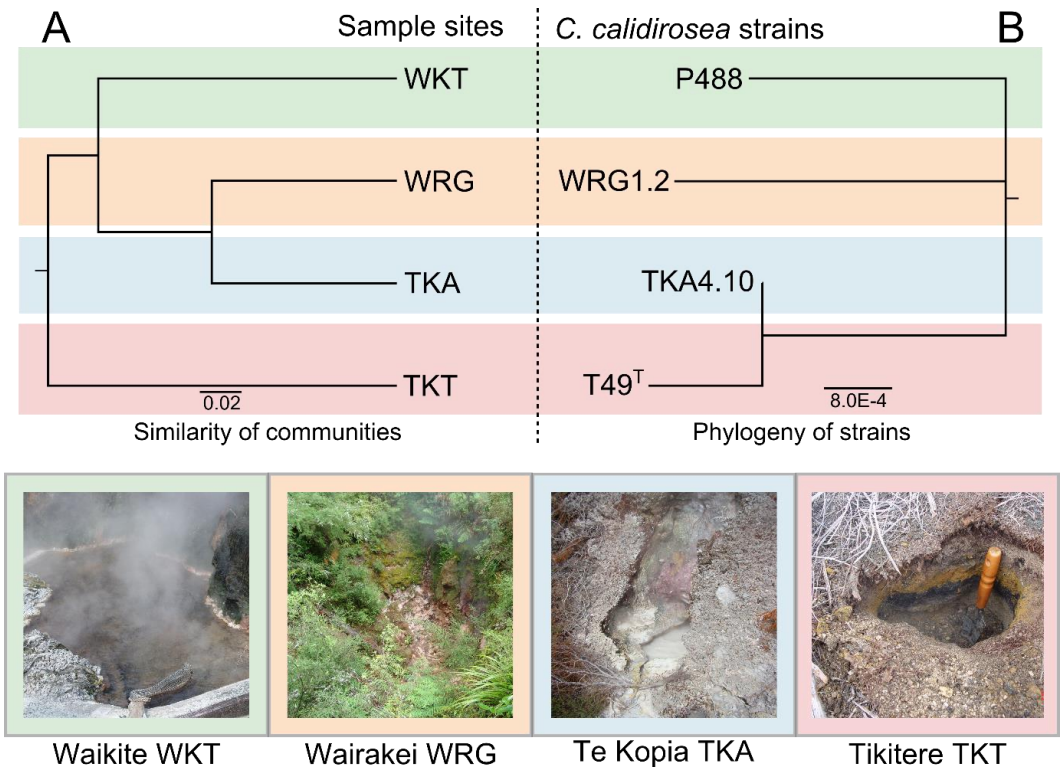


Figure 6.3 - Similarity between the communities of sample sites and phylogeny of *C. calidirosea* isolates. (A) The similarity between the communities of all four sites is shown by a jackknifed-UPGMA tree, created using weighted Unifrac distance. The community tree has high bootstrap support values (75-100 %) for all internal nodes. (B) The phylogenetic relationships between the four *C. calidirosea* isolates are shown by an unrooted Bayesian nucleotide phylogenetic tree of 327 concatenated conserved genes shared by the four isolates. The branching of the phylogenetic tree is supported by 100 % posterior probability, with the branch length indicating expected nucleotide changes per site. Photos of the sample sites are shown below the trees.

6.5.1.5 Two putative horizontally-transferred regions are absent in isolate WRG1.2

While the four *C. calidirosea* genomes showed a high degree of synteny in most genomic regions, previous studies identified two putative horizontally-transferred regions: A [CCALI_00447-449] and B [CCALI_00804 – CCALI00807] (Lee et al., 2014) (Supplementary Figure 8.7). Both Region A and Region B were conserved between isolates T49^T and TKA4.10, but were absent in isolate WRG1.2. In contrast, isolate P488 contained only a small deletion in Region B, and the sequence was slightly less

conserved. This hyper-variability may provide important clues on the gene-flow history among the isolates. It suggests that T49^T and TKA4.10 show low divergence from one other either due to recent dispersal or frequent population homogenization. Isolate P488 appeared to be more divergent, as deletions, truncations, and frameshift mutations in genes CCALI_00449 and CCALI_00805 indicate degradation.

6.5.1.6 Comparison of genomic rate of divergence with other thermophiles

The genomes of the four *C. calidirosea* isolates exhibited low divergence over a large geographic distance. We compared its genomic divergence to that of two other well-studied thermophile genomes isolated over geographic scales comparable to those of the *C. calidirosea* isolates in this study: *Thermus thermophilus* and *Sulfolobus islandicus.* The two isolates of *S. islandicus* were isolated over a distance of approximately 4.7 km (from Geyser Creek to Norris Geyser Basin) across Yellowstone National Park (Reno et al., 2009), while the three isolates of *T. thermophilus* were isolated over a distance of approximately 17 km, from Mine hot spring (Henne et al., 2004; Oshima et al., 1975) to Atagawa hot spring (based on Jiang et al (Jiang et al., 2013) and ATCC record) across Izu peninsula, Japan. We compared the rearrangement in overall genome structure **(**Supplementary Figure 8.8) and nucleotide divergence in conserved genes in all isolates (Table 6.2, Supplementary Table 8.17). *T. thermophilus* exhibited higher nucleotide divergence and rearrangement than between all *C. calidirosea* isolates (Supplementary Figure 8.9). In contrast, while the two *S. islandicus* strains showed lower nucleotide divergence in conserved genes than between some *C. calidirosea* isolates, the two archaeal genomes exhibited more genome rearrangements (Supplementary Figure 8.10) with more isolate-variant genes (329 or 163 depending on which genome was used for reference for IMG-ER phylogenetic profiler analysis) than between any two *C. calidirosea* genomes (Supplementary Table 8.16). These comparisons demonstrate that *C. calidirosea* has a high degree of conservancy in overall genome synteny as well as low nucleotide divergence.

Table 6.2 - Total number of base differences and sequence identity of 327 conserved genes. Average base difference per 100 bases is italicised. The multiple sequence alignment was 489,735 bases in length.

Isolates (ungapped seq. size)	**T49^T**	**P488**	**WRG1.2**
T49^T (488,944)			
P488 (488,769)	5,039 - av. *1.02 %*		
WRG1.2 (488,350)	5,736 - av. *1.17 %*	3,769 - av. *0.77 %*	
TKA4.10 (488,323)	1,627 - av. *0.33 %*	3,815 - av. *0.78 %*	4,518 - av. *0.92 %*

6.5.1.7 Phylogenetic analysis shows close relationship between T49^T and TKA4.10

Nucleotide sequence alignment of concatenated conserved genes (Supplementary Table 8.14) shows overall high pairwise sequence identities between the isolates, with an average of 99.2 % (Table 6.2). This coincides with the conservation in the 16S rRNA gene sequences. Based on this SNP data, T49^T shares the highest similarity with TKA, followed by P488, with WRG1.2 being the most distantly related. Using the concatenated conserved gene nucleotide sequences, Bayesian phylogenetic inference (Figure 6.3) supported the unrooted tree topology with > 99 % posterior probability value. Within this tree, TKA4.10 node showed a very short branch length from the internal node shared with T49^T. WRG1.2 has the longest terminal branch length.

6.5.1.8 C. calidirosea isolates have similar carbohydrate metabolism, but also isolate-specific profiles

Strain T49^T was physiologically and genomically characterised as a scavenger capable of utilizing a wide range of carbohydrates (Lee et al., 2014, 2011). In order to determine the physiological adaptation or divergence of other *C. calidirosea* isolates to their respective environments, we conducted carbohydrate utilization redox assays with BIOLOG PM1 carbon source microtitre plates. The plates contained 95 substrates ranging from mono- and di-saccharides, amino acids, sugar alcohols, nucleosides, organic acids, to other potential carbon sources. The assay showed that the isolates shared similar redox activities with the tested carbohydrates. All isolates induced particularly high redox response to hemicellulosic sugars such as xylose and mannose, and negligible responses to amino acids as the sole carbon and energy source. The overall pattern of high/low redox response to substrates of the three isolates closely resembled that of the reference strain T49^T. Even though the carbon utilization profiles of isolates were broadly similar, the relative intensity of substrate response appeared isolate-specific. In order to assess isolate substrate response similarity (note that the intensities of the individual isolate responses were reproducible), we calculated the Spearman rank order correlation

coefficients of isolates carbon utilisation profiles (Supplementary Figure 8.11). We observed the lowest correlation between isolates $T49^T$ and WRG1.2 ($\rho = 0.768$), and highest correlation between P488 and WRG1.2 ($\rho = 0.943$). These pairwise correlation differences do not appear to be due to noticeable variation in genes related to carbon metabolism (see above section). They may be therefore due to polymorphisms in functional genes, or differences in regulatory elements, which lead to isolate-specific differences in rates of substrate utilization.

6.5.2 - The sample sites exhibited differences in hydrothermal activities and clay content

The four sample sites were distributed along the TVZ (Supplementary Figure 8.12), with a maximum linear distance of 67.4 km and minimum of 12.0 km between sites. All samples sites were geothermally-affected soils, with acidic pH ranging from 3.5 to 4.8 and temperature from 46.9 to 59.9 °C. The isolates were isolated independently.

In order to determine the relationships between the geothermal environments and the abundance, genomic characteristics and physiology of *C. calidirosea* isolates, we first characterised the mineral compositions of the sample sites using semi-quantitative XRD and XRF analyses (Supplementary Figure 8.13; Supplementary Table 8.18). The soil samples were composed primarily of quartz and silica minerals and of clay minerals, the latter of which are sensitive to hydrothermal alteration. As shown in Supplementary Figure 8.13, the minerals present at site TKA and TKT were almost entirely of primary minerals (e.g., magnetite, biotite, pyroxene, and plagioclase), which indicates that these samples were only weakly affected by hydrothermal activity. In contrast, the clay content of samples WKT and WRG consisted mostly of secondary clay minerals (e.g., alunite-kaolinite and montmorillonite), which were formed from reaction of rock with acidic steam-heated water (Browne, 1978; Reyes, 1990; Steiner, 1977). The remnant magnetite and hydrothermal montmorillonite in WKT are indicators of weak alteration in mildly acidic conditions, while the alunite-kaolinite-amorphous silica phases in WRG point to more strongly acidic conditions. However, observed soil pH at the time of sample collection was similar to the other samples (Table 6.3).

Table 6.3 - General physicochemistry of the soil samples.

Sample sites (isolates)	**TKT ($T49^T$)**	**WKT (P488)**	**WRG (WRG1.2)**	**TKA (TKA4.10)**
Temperature °C	52.5	50.7	46.9	59.9
pH	4.3	4.5	4.8	3.5
Moisture %	28.0	59.9	37.1	28.8
Soluble organic carbon (SOC) mg/g	1.47	2.33	1.06	1.59
Ammonia mg/g	145.0	105.0	28.5	23.5
Nitrate mg/g	2	6	2	1
Nitrite mg/g	0.15	0.35	0.40	1.50
Dissolved reactive phosphorous mg/g	BDL	BDL	BDL	BDL

The relative abundance of clay minerals was variable between sites (Supplementary Figure 8.12). WKT had the highest clay content, followed by WRG, TKT, and finally TKA, which consisted mostly of quartz (86 % w/w). Among the four sites, the higher clay content correlated with a higher relative abundance of *C. calidirosea*. In addition to having the highest clay content, WKT was also much richer in soluble organic carbon (SOC) (2.33 mg/g soil) than the other sites (range: 1.06 – 1.59 mg/g soil) (Table 6.3). The SOC concentration was within the distribution seen in mesophilic soil environments such as the nutrient-rich crop field soil (6.1 mg/g) (Kanchikerimath & Singh, 2001) and nutrient-poor pine plantation forest soil (0.08-0.17 mg/g) (Li et al., 2014), indicating availability of organic carbon at the sample sites.

6.5.3 - *C. calidirosea*-associated communities are dominated by *Crenarchaeota*/ *Thaumarchaeota*

To determine the relationships between sample site chemistry, sample site community composition, and the genomic content of *C. calidirosea* isolates, we used 16S rRNA gene sequencing on the Ion Torrent platform to measure the relative abundance of *C. calidirosea* and other taxa at each sample site. After quality control and OTU clustering (see 6.4 -Methods), we observed a total of 190 OTUs among the four sites (Supplementary Table 8.19). Rarefaction curves of the sample sites showed plateauing of observed OTUs (WRG: 140, WKT: 78, TKT: 107, TKA: 65) (Supplementary Figure 8.14), indicating the majority of diversity has been sampled. Inasmuch as OTU counts are influenced by both extraction method and primer bias (Milani et al., 2013), here they are best considered as a relative indicator of site alpha diversity rather than as absolute values

of all OTUs present. The relative taxa abundance of the four communities is outlined in Supplementary Table 8.20, including abundance of OTUs unassigned to taxa within the Greengenes database.

6.5.3.1 C. calidirosea abundance

A single OTU representing *Chthonomonas* was detected at low abundance in all four communities, contributing from 0.006 % (TKA) to 0.3 % (WKT) of the total reads (Supplementary Table 8.20). The 250 bp OTU sequence was virtually identical to the corresponding 16S rRNA gene region in the four *C. calidirosea* isolate genomes, with only one base difference. Although the short read length and error rate of ion semiconductor sequencing does not allow fine-resolution comparison, this does indicate that a single species of the *Chthonomonadetes* was present in these sites, with limited phylogenetic diversity. Interestingly, another *Armatimonadetes* lineage, Group FW68 (Lee et al., 2013) of the class *Armatimonadia* (Tamaki et al., 2011) was detected at TKA only (0.2 %) suggesting that this site does present different niches for these bacteria than other sites. When we compared the communities where the *C. calidirosea* OTU was most abundant (WKT) to those where it was least abundant (TKA), we observed that *C. calidirosea* was more abundant in more nutrient-rich soil, as reflected by higher soluble organic carbon content, nitrate concentration, and overall clay mineral content. In contrast, *C. calidirosea* was less abundant in sites consisting mainly of fresh hydrothermal deposits rich in quartz and amorphous silica, as represented by the SiO_2 concentration (Supplementary Table 8.18), and lacking in soluble organic carbon. Soil with higher clay and organic matter content also has a buffering capacity (Osman, 2012); this may facilitate the survival of *C. calidirosea*, which is a pH-sensitive species (Lee et al., 2014, 2011).

6.5.3.2 Comparing the diversity between communities

To assess the relationship between the microbial communities in contrast to the dissimilarities of *C. calidirosea* isolates, phylogenetic (weighted and unweighted Unifrac) and non-phylogenetic (Bray-Curtis) beta diversity metrics were used to assess relationships among communities (Supplementary Table 8.21) and to construct a community hierarchical clustering (Figure 6.3). When the abundance of OTUs was considered (weighted Unifrac and Bray-Curtis), WRG and TKA were most similar, followed by WKT, and TKT was the most distant (Figure 6.3). This was likely due to the change of ratio in the two archaeal clades (MBG-A and *Thermoprotei*) in TKT compared to the other communities. When unweighted Unifrac distance was used to compare communities (taxa abundance not considered) the TKT and WRG communities were most similar as these communities had a relatively large number of low-abundance taxa

in common when compared WKT and TKA, which may play more subtle roles in community function.

6.5.3.3 Bacterial diversity

The bacterial taxa represented in the communities consisted of phyla associated with the super-phylum "Terrabacteria," (Rinke et al., 2013) (i.e. *Armatimonadetes*, *Chloroflexi*, *Firmicutes*, *Actinobacteria*, and *Thermus*), as well as soil-associated lineages such as *Planctomycetes*, *Bacteroidetes*, and *Proteobacteria* (Figure 6.4). Overall, no single bacterial taxon was highly abundant in all samples. However, based on their affiliations to characterised type strains, the majority of bacterial taxa fulfilled similar ecological roles. They were acidophilic and putatively capable of chemolithotrophic and aerobic growth via mechanisms such as iron and sulfur oxidation in family *Sulfobacillaceae* (Norris et al., 1996), iron oxidation in *Acidimicrobiaceae* (Cleaver et al., 2007), or methanotrophy with *Candidatus* Methylacidiphilum (Dunfield et al., 2007). The most abundant bacterial taxa were *Deltaproteobacteria* (12.1 %) in TKT, Candidate Division WPS-2 (12.1 %) in WKT, *Alphaproteobacteria* (specifically *Rhodospirillales* (5.7 %)) in WRG, and *Actinobacteria* (specifically *Acidimicrobiales* (8.8 %)) in TKA. Candidate Division WPS-2 has been associated with iron oxidation (Grasby et al., 2013), and *Rhodospirillales*, specifically those within family *Acetobacteraceae*, are commonly acetogenic (Kersters et al., 2006). The more abundant bacterial taxa identified here reflect those previously identified through Sanger sequencing of 16S rRNA gene clone libraries (Stott et al., 2008), such as the presence of family *Alicyclobacillaceae* of *Firmicutes* and *Acetobacteraceae* of *Alphaproteobacteria*.

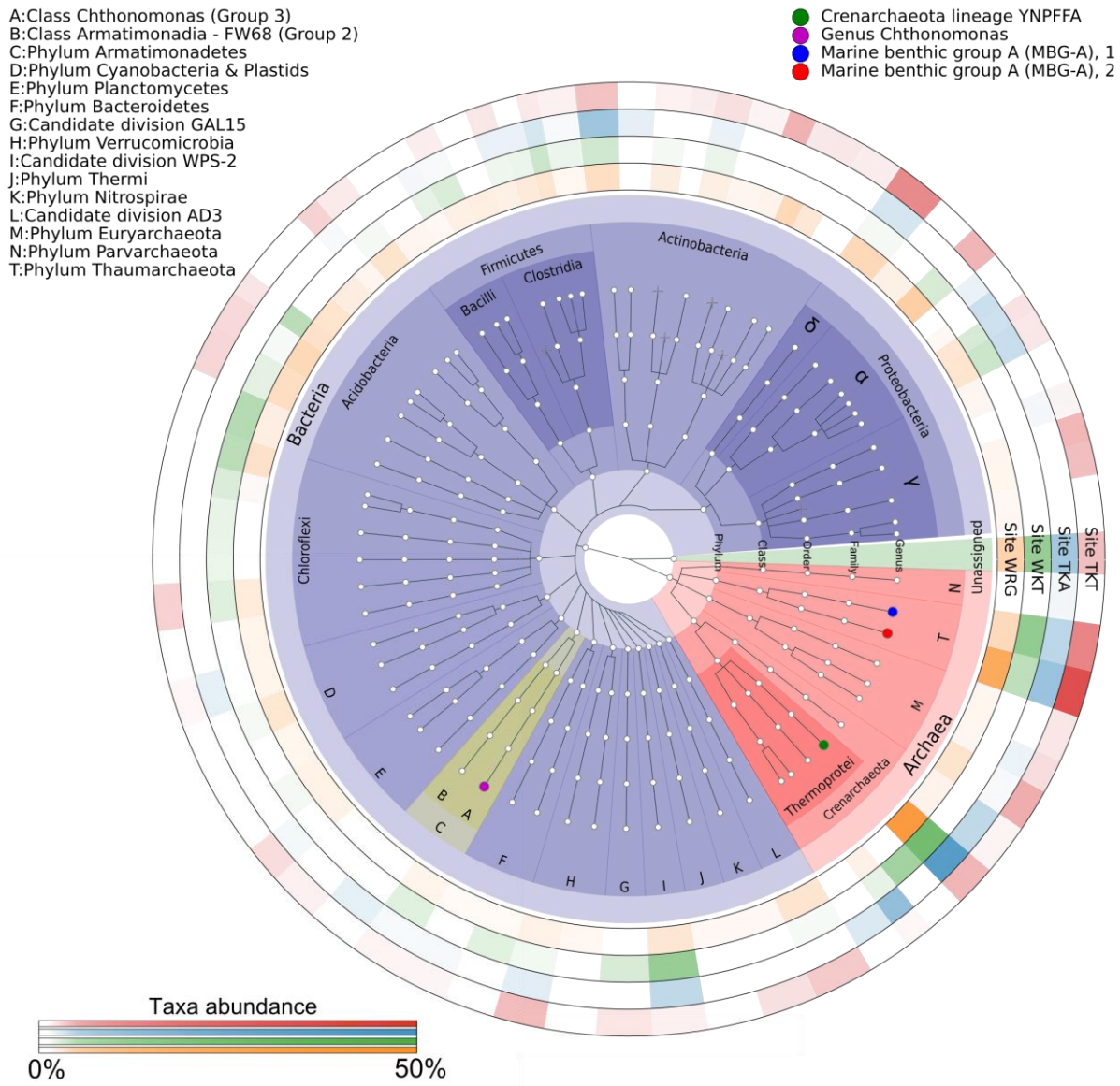


Figure 6.4 - Taxonomic relationship of organisms present at the sample sites and their relative abundance. Each ring on the outside of the tree corresponds to a sample site, and the colour intensity at each position on the ring indicates the relative abundance of the corresponding taxa at that sample site. The colour intensity is scaled nonlinearly to emphasise low abundance taxa by transforming the relative abundance with the exponent (⅓). Cross symbols at terminal nodes indicate representative reads with only higher level taxa assignment, and may contain multiple genera. The bacterial phylum *Armatimonadetes* is highlighted in yellow, and the member genus *Chthonomonas* of class *Chthonomonadetes* (labelled "A") is coloured in magenta. The relative abundance of *Chthonomonas*-related OTUs at each of the four sample sites was as following: WRG: 4.08×10^{-4}, WKT: 3.02×10^{-3}, TKA: 5.7×10^{-7}, and TKT: 1.17×10^{-4}.

6.5.3.4 Dominant taxa

At all four sites, the majority of the community was archaeal, ranging from a low of 56 % of the detected OTUs at WKT to a high of 75 % at TKA. However, only 21 archaeal

OTUs were observed, and only five of these were present at all sites. The *Euryarchaeota* consisted of 10 OTUs, all within the thermophilic acidophilic class *Thermoplasmata* (Reysenbach, 2001, p. 169), but represented less than 5 % of any of the communities. In contrast, *Crenarchaeota* and *Thaumarchaeota* dominated all four communities, representing more than half the relative abundance of each community (Figure 6.4). These few shared archaeal OTUs were the most dominant microbial diversity to coincide with the presence of *C. calidirosea*, which may indicate an underlying ecological relationship crucial for the survival of the bacterium.

6.5.3.5 Archaeal and bacterial diversity

The *Crenarchaeota* diversity consisted of a class-level lineage, *Thermoprotei* (9 OTUs) (Vetriani et al., 1999). The *Thermoprotei* are broadly distributed and are represented by a number of thermophilic strains such as *Sulfolobus* spp., which are primarily found in acidic terrestrial geothermal sites. Another class-level lineage, MBG-A (12 OTUs) was first identified from deep-sea marine benthic microbial communities and was assigned to *Crenarchaeota* (Vetriani et al., 1999). However, recent analysis (Offre et al., 2013) has attributed the group to the newly-recognised phylum *Thaumarchaeota* (Stieglmeier et al., 2014). With the exception of the TKT community, *Thermoprotei* was the most dominant group, comprising 39-58 % of total sequences, while MBG-A comprised 13-28 %. In the TKT community, the community ratio was roughly reversed, as 57 % of OTUs were from MBG-A, and 3 % were from the *Thermoprotei.* In this study, *Thermoprotei* was predominantly represented by the lineage YNPFFA, which is also associated with geothermal features at Yellowstone National Park, USA (Kan et al., 2011). Unfortunately, the lack of cultivated representatives within these *Crenarchaeota* and *Thaumarchaeota* lineages impedes any attempt to infer their ecological roles.

There were distinct differences in the four communities at a finer taxonomic resolution. Of 190 total OTUs, only 25 were present in all four sites (Supplementary Table 8.19). These taxa included the *Sulfobacillaceae, Rhodospirillales,* and *Acidimicrobiaceae*. Some predominant OTUs were detected in only one or two sites. For example, one archaeal OTU from clade SK322 (OTU_0003) was present at 10.3 % of the total reads in WKT, but was not detected in TKT, and one *Acinetobacter* OTU (OTU_158) was present in WKT at 8.5 % of total reads, but was undetected in the other soils (Supplementary Table 8.19). These observations stress that there are four distinct microbial communities in the sample set, and that ecological or stochastic differences across the four sites select for these differences.

6.6 - Discussion

6.6.1 - The role of *C. calidirosea* as a heterotrophic scavenger in microbial communities

Our 16S rRNA gene analysis of the four *C. calidirosea*-associated communities revealed that all communities were dominated by *Crenarchaeota* and *Thaumarchaeota*, with *Bacteria* as minor components of the communities. This contrasted with previous culture-based studies of WKT (Stott et al., 2008), which used gellan gum-based media and resulted in the isolation of several bacterial taxa (including *C. calidirosea*), but no archaeal species. There is a paucity of physiological data available regarding the dominant archaeal lineages detected in this study, due to the limited numbers of characterised isolates. However, the metabolic capabilities of described strains within related lineages (e.g., *Sulfolobales*, *Nitrososphaerales* and *Nitrosopumilales*) suggest possible autotrophic-based metabolisms (Berg et al., 2010; Huber & Prangishvili, 2006; Könneke et al., 2014; Villanueva et al., 2014).

We therefore speculate that the dominant crenarchaeotal and thaumarchaeal phylotypes detected in this study may occupy the niche of autotrophic primary producers, supporting the diverse, yet low-abundance, chemolithotrophic and heterotrophic bacterial species, which include *C. calidirosea.* The low relative abundance of putatively heterotrophic bacteria also likely reflects the low soluble soil carbon content available as a substrate. The scavenger phenotype of *C. calidirosea* isolates may be well-suited to persistence in an ecosystem with minimal or inconsistent saccharide sources (Lee et al., 2014, 2011). A recent study of bacterial communities in a Thailand hot spring also indicated a potential relationship between *Armatimonadetes* and autotrophs in geothermal environments. This survey detected abundant *Armatimonadetes* and *Chloroflexi* OTUs associated with *Cyanobacteria* (Cuecas et al., 2014). While the phylogenetic relatedness of the *Armatimonadetes* OTUs identified in the study (accession: HQ416751 to HQ416755 and HQ416757) to *C. calidirosea* is currently unknown, pairwise discontinuous megaBLAST showed that the OTUs shared approximately similar low sequence similarities (75-84 %) to the type strains of the three *Armatimonadetes* classes (Supplementary Table 8.22).

6.6.2 - Low genomic diversity in the face of geographical isolation

The genomes of the four *C. calidirosea* isolates originating from disparate sites across the TVZ shared striking similarities. Phylogeny inferred from SNPs and horizontal gene transfer indicated a close relationship between all isolates, particularly between strains T49^{T} and isolate TKA4.10. Gene presence/absence analysis identified few isolate-variant genes (Supplementary Table 8.15; Supplementary Table 8.16), such as restriction-

modification systems. These genes are known to be fast-evolving (Stern & Sorek, 2011) and mobile (Kobayashi, 2001), which may explain their presence in a pangenome which is otherwise highly conserved. Physiologically, all four *C. calidirosea* isolates shared similar carbohydrate metabolism and exhibited high rank-order correlation of redox responses from the phenotype microarrays. Pairwise correlation matrix of the carbon source utilisation profiles shared a similar pattern to the inferred phylogeny i.e. the close relatedness of $T49^T$ to TKA4.10, while $T49^T$ to WRG1.2 having the lowest correlation and greatest phylogenetic tree distance. Future comparative transcriptomic and proteomic analyses of the isolates may shed light on subtle underlying regulatory processes resulting in the variations of phenotypic response despite the highly conserved genomes.

6.6.3 - Potential mechanisms underpinning genomic conservation across geographic distance

The reason for low genomic divergence between the *C. calidirosea* isolates despite their geographical and ecological isolation is not immediately clear. One possible explanation for the high genomic conservation is a relatively recent sympatric speciation of *C. calidirosea*, leading to its occupation of a new niche in geothermal environments from temperate soils. A recent review of the environmental distribution of *Armatimonadetes* phylotypes indicated that temperate soil is the most dominant environment for the phylum (Dunfield et al., 2012). Of 39 phylotypes identified representing class *Chthonomonadetes* (Group 2), only two phylogenetically-distant clones were associated with thermophilic environments. In contrast, Group 10 (A&B) consists entirely of phylotypes from geographically disparate geothermal environments. Thus, the occasional occurrence of thermophily in the *Chthonomonadetes* may indicate sampling bias or recently acquired thermophilic adaptation. In this study, we detected only a single *Chthonomonadetes* OTU in our community analyses, which supports the low-diversity hypothesis of thermophilic *Chthonomonadetes.*

Although this hypothesis may explain the low diversity from an evolution and sampling perspective, the mechanism alone is inadequate to explain the very high similarity in content of the genomes from factors such as genetic drift, unless the speciation and dispersal of *C. calidirosea* isolates was extremely recent. The close relationship between strains $T49^T$ and isolate TKA4.10, as demonstrated by the low occurrence of SNPs in concatenated highly-conserved gene sequences in particular, suggests recent dispersal and low genetic drift between the two isolates. A recent study found that thermophilic microbes can be globally dispersed via aeolian transport, but are selected by their environments (Herbold et al., 2014). In dynamic geothermal environments, taxa with very low abundance such as *C. calidirosea* may become locally extinct, but subsequent

recolonization from another site would result in high genetic similarity between two populations. Over a sufficiently rapid timespan of extinction and colonization, the genomes would reflect stochastic aeolian dispersal, rather than being shaped by selection pressure of the specific environment. We believe the phylogeny and the low diversity of *C. calidirosea* genomes reflect these scenarios. While surface water flow presents an alternative mechanism of dispersal between sites, we believe this is unlikely due to the arrangement of catchments flowing towards the Waikato River and away from the sites (Supplementary Figure 8.12). Additionally, the inability to sporulate, as well as the physiological requirement (pH and temperature growth range) restricts the persistence of *C. calidirosea* in the environment and favours rapid modes of transport able to disperse the bacterium between suitable habitats.

6.7 - Conclusion

In order to investigate the evolutionary history of *C. calidirosea*, we have performed an integrative analysis incorporating comparative genomics and isolate physiology, as well as community and environmental geochemistry data. We propose a potential mechanism of genome conservation for low-abundance microbial taxa. We have shown that *C. calidirosea*, a thermophilic non-spore-forming bacterium, is capable of dispersal across geographical barriers, and exhibits detectable, albeit minor genome differences in relation to geographic distance. Our approach shows the value of augmenting amplicon profiling of poorly-characterised communities with additional data such as the physiological characterization of isolates and geochemical analysis in order to provide context and validation of ecological inferences.

6.8 - Addendum

The genome of *F. ginsengisoli* Gsoil 348^T (Hu et al., 2014), type strain of class *Fimbriimonadia* was recently published during the editing of this thesis. The genome showed low overall synteny and limited shared orthologous clusters with the *C. calidirosea* T49^T genome. Like *C. calidirosea*, Gsoil 348^T appears to lack many conserved operons (e.g., purine, tryptophan, histidine biosynthesis), and displayed similar degree of disorganisation in clustering of functionally-related genes. The similarity of the *C. calidirosea* isolate genomes, including the gene organisation, suggests that the genome of the species was not subjected to as much rearrangement as previously hypothesised in Chapter 5, possibly due to the mechanisms described in this chapter. However, the low degree of gene organisation shared between *F. ginsengisoli* Gsoil 348^T and *C. calidirosea* indicates that gene cluster conservation is only present at a lower taxonomic level.

It is also worth noting that the primer pair used in this study f515 and r806 (Caporaso et al., 2012), may overestimate some archaeal abundance (Kittelmann et al., 2013) as well as bias against some taxa such as Crenarchaeota/Thaumarchaeota and clade SAR11. This has prompted the Earth Microbiome Project, which have utilised these primers heavily to design newer and more degenerate primers (http://www.earthmicrobiome.org/emp-standard-protocols/16s/). The microbial diversity of the four TVZ sample sites may be better estimated with more primers capturing sequence variance as well targeting different segment of the 16S rRNA gene.

Chapter 7 Synthesis and Conclusion

7.1 - Significance

In the last five years, the phylum *Armatimonadetes* has transitioned from a candidate division to a recognised taxon with three species, each representing a separate class (Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011). This PhD research contributes to the body of knowledge regarding this novel phylum in multiple aspects. First, *Armatimonadetes* now has a well-defined phylogenetic and taxonomic structure. Approaches utilising phylogeny to assist with the delineation of higher-order taxonomy (including the candidate taxa), as demonstrated in Chapter 4, have become increasingly prevalent, as they allow (A) an objective comparison of lineages, and (B) the ability to infer the coherence of genetic, phenotypic, and ecological traits associated with clade-based taxa (Yarza et al., 2014). While *Armatimonadetes* currently only consists of few very distantly-related type strains and environmental phylotypes, based on the sources of DNA isolation of the environmental phylotypes, the clades already exhibited distinct ecological themes (Dunfield et al., 2012). In the future, as *Armatimonadetes* becomes better understood through the characterisation of novel isolates that are representative of the clades within the phylum, other shared traits within each clade may become better understood through this phylogenetic framework. The clarification of *Armatimonadetes* taxonomy, including the identification of both internal (7 class-level taxa without cultured representative) and external (two neighbouring candidate divisions) candidate taxa, provides a robust framework for future research into one of the 30 generally-recognised bacterial phyla (Euzéby, 2011, Retrieved in December 2014). This framework will enable more rapid and accurate classification of newly-cultured *Armatimonadetes* strains within this highly-diverse phylum. The comprehensive selection of representative phylotypes, as well as multiple commonly used phylogenetic methods, meant that the results are robust, repeatable, and can be expanded upon as more phylotypes are identified, thus facilitating future research of *Armatimonadetes* through standardisation and reduction of duplicated efforts.

Second, this work builds on previous physiological characterisation (Lee et al., 2011) and contributes to an integrative view on the biology of a representative *Armatimonadetes* species (*C. calidirosea*) through genomic characterisation (Chapter 5) and investigation of the genome-population dynamics of the species (Chapter 6). The research reveals a thermophilic carbohydrate scavenger with restrictive physiology suitable for limited and dynamic habitats. The interactions between physiology, the ecological niche, and the population-genome dynamics provide an unusual insight into a member of the low abundance “rare biosphere”.

Finally, this study contributes to fundamental knowledge of the physiological, genetic, and ecological diversity of the microbial world. Extremophilic organisms in general are poorly understood despite their unique physiology and distinct evolutionary histories (Rothschild et al., 2001). Although phylum *Armatimonadetes* (Candidate Division OP10) was originally detected in a Yellowstone hot spring, *C. calidirosea* is currently the only cultivated thermophile within *Armatimonadetes*. Furthermore, based on the environmental distribution of the uncultured phylotypes (Dunfield et al., 2012), *C. calidirosea* appears to be unusual within its clade (class *Chthonomonadetes*) with no closely-related phylotypes detected from thermophilic environments. This may be due to genuine rarity, sampling bias, or even recently acquired thermophily specific to the lineage. Investigation into the genetic and phylogenetic distinctiveness of this organism (Figure 7.1**,** see also preface of Chapter 5) raises interesting questions (see 7.3 - Future directions below), but also provides the body of knowledge towards future questions we do not yet know to ask, or the "unknown unknowns" (see section 1.1).

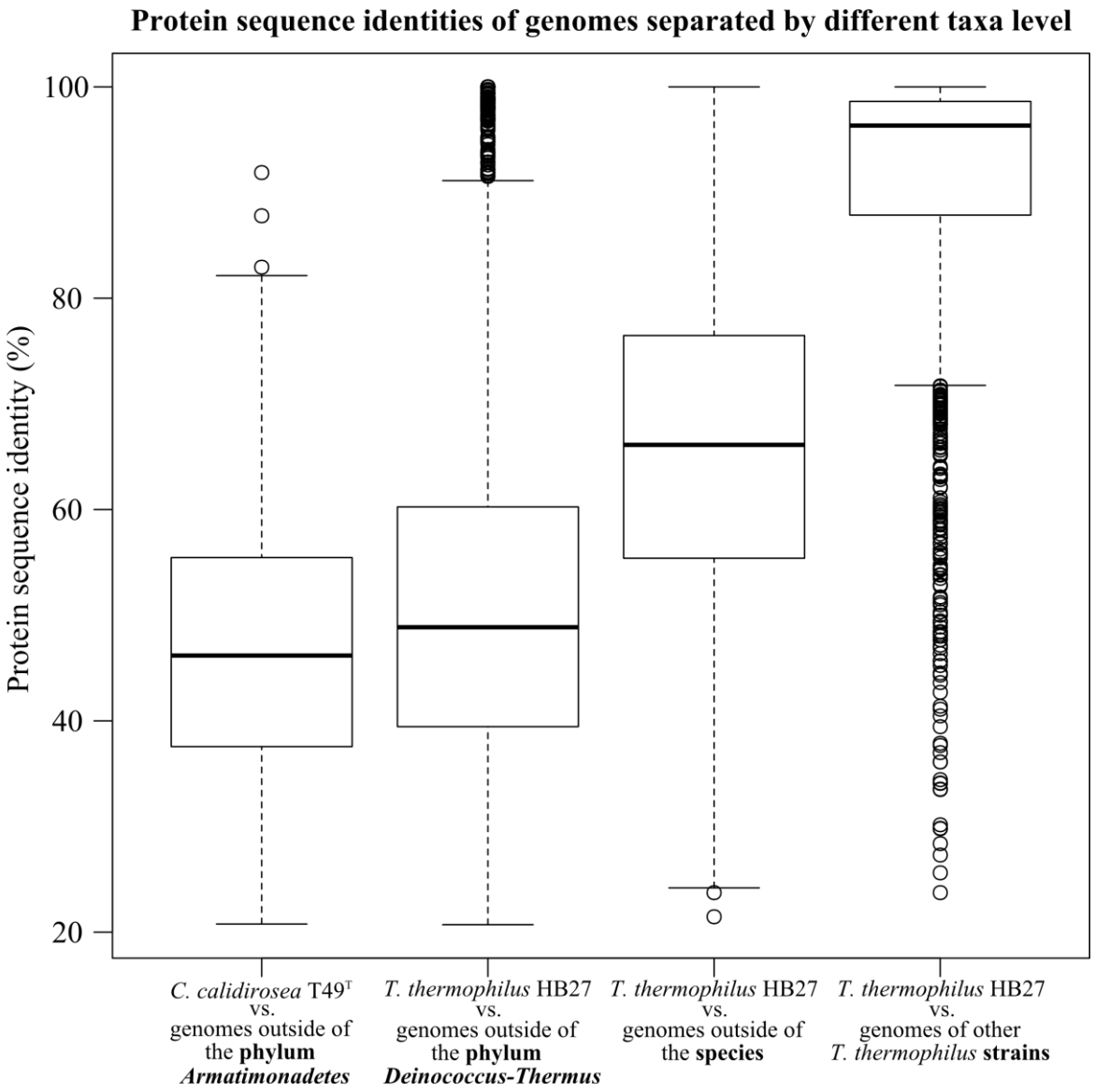


Figure 7.1 - *C. calidirosea* T49^T showed low overall amino acid sequence identities to known proteins. Top amino acid sequence percentage identities of predicted proteins from the genomes of *C. calidirosea* T49^T and *Thermus thermophilus* HB27 against NBCI RefSeq protein database (October 2014). BLASTP hits of *C. calidirosea* T49^T were filtered to exclude results from *Armatimonadetes* to reflect the reality when the genome was published in January 2014. From left to right, strain T49^T was the only genome available within *Armatimonadetes* and exhibited low overall sequence identity to known proteins in the database. For comparison of sequence dissimilarity at different taxonomic scale, *T. thermophilus* HB27 was searched against protein sequences outside of the phylum *Deinococcus-Thermus*, genus *Thermus,* and outside the species *T. thermophilus*, which exhibited highest sequence similarity due to BLASTP hits with other *T. thermophilus* strains.

7.2 - Overview of results and future questions

Prior to this PhD research, the knowledge of *Armatimonadetes* was restricted to environmental 16S rRNA gene-based phylotypes and three characterised *Armatimonadetes* species: *A. rosea*, *C. calidirosea*, and *F. ginsengisoli*. All three strains are oligotrophic soil-based bacteria, but *C. calidirosea* T49^T is distinct from other species as the only thermophilic isolate, and it exhibits a much wider carbohydrate utilisation profile (see Chapter 2 Literature review). Both the phylogenetic structure of monophyletic classes within the phylum itself, and the relationship between the phylum and neighbouring phyla were unclear due to inconsistent selection of phylotypes used for phylogenetic inferences (Im et al., 2012; Lee et al., 2011; Tamaki et al., 2011). Knowledge of the ecology of all three *Armatimonadetes* representative species was limited to their physiological characterisation, as no genomic information of *Armatimonadetes* was available. Therefore, knowledge of genomic variation with any of the species was also unavailable. Thus, our entire understanding of the phylum *Armatimonadetes* was limited to individual snapshots of three highly-diverse representative type strains of distantly-related taxa.

7.2.1 - Phylogeny and taxonomy of *Armatimonadetes*

The research in this thesis contributed to the current state of knowledge of *Armatimonadetes* and microbial diversity in the following areas, which correspond to the chapters in this thesis. In Chapter 4 (Lee et al., 2013), the internal and external phylogeny of the phylum was better defined, and a clear taxonomic and phylogenetic framework for future studies was provided. The study identified 10 class-level taxa and two neighbouring candidate divisions. The analysis also showed that classes *Armatimonadia* and *Chthonomonadetes* are more closely related to one another than to class *Fimbriimonadia*. As sequencing of novel environmental phylotypes continues to outpace the cultivation and description of novel species, this study represents a general trend in microbiology in reconciling taxonomy (traditionally based on cultivated type strains) with phylogeny involving environmental phylotypes.

In a recent publication, Yarza et al. (2014) used a taxonomic threshold-based strategy for phylogenetic inference similar to that demonstrated in this study (Chapter 4), and proposed a generalised classification system for assessing high taxonomic ranks based on 16S rRNA gene sequences. Compared to the methodology in Chapter 4, the approach applied by the group differs mainly in the treatment of taxonomic thresholds as evidence for potential taxonomic ranking. In Chapter 4, maximum sequence distance of a monophyletic group to the member type strain was used as the taxonomic boundary; whereas in Yarza et al., the taxonomic boundaries are applied as thresholds for hierarchical clustering of phylotype sequences to generate clusters of certain taxonomic

size (e.g., a "genus size[g]" cluster of sequences), before monophyletic support of the cluster is considered. Since hierarchical clustering is based on all of the relative distances of the sequences being analysed, taxonomic boundaries are not influenced by the arbitrary phylogenetic position of the type strain (i.e. the type strain may not be within the middle of sequence diversity of the taxa).

Analogously to the way in which operational taxonomic units (OTUs) are generated based on sequence clustering, and operational phylogenetic unit (OPUs) are generated from monophyletic clades in phylogenetic analysis, the group proposed "candidate taxonomic unit" (CTU) as means to achieve the goal of encompassing cultured and uncultured microorganisms under one unified classification. However, due to different implementations of Yarza's guideline, certain degrees of subjectivity may persist. Identification of CTU is heavily influenced by input phylotype and phylogenetic analysis method. Indeed, perhaps due to the utilisation of only 16S rRNA gene sequences from type strains, the LTP Project (Yarza et al., 2010), upon which the Yarza et al. (2014) analysis was based, associated the three type strains of *Armatimonadetes* most closely with the *Elusimicrobia* and *Fusobacteria*, and in proximity with *Proteobacteria*. Furthermore, Yarza et al. (2014) used *Deinococcus-Thermus* as an outgroup despite the fact that previous studies (Dunfield et al., 2012; Lee et al., 2014) have demonstrated that *Chloroflexi* is the closest-related phylum.

Nonetheless, this system may bring clarity by creating a better-defined line of evidence (CTUs) which joins two sources of phylogenetic diversity (cultured and uncultured) and supplement other information (e.g., genetic, phenotypic, and environmental) in defining biological taxa, as the authors intended. Ultimately, the goal of biological classification is to create a construct that is both meaningful (for reflecting relationships between genetic, phenotypic, environmental diversity) and useful for scientific research and communication (Rosselló-Móra, 2012). This trend towards combining two sources of information on phylogenetic diversity will likely assist in identifying patterns (e.g., congruence in phenotypes and environmental distribution) and facilitate future research. The systematic formulation and standardisation (including naming of CTU) should help increase analysis throughput and decrease subjectivity and ambiguity in high-order microbial taxonomy facing an abundance of environmental phylotypic data. Overall, Yarza's analysis on *Armatimonadetes* agreed with the findings in Chapter 4 (Lee et al.,

[g] Sequence identity thresholds of various taxonomic levels were calculated from 16S rRNA genes of preexisting taxonomic groups, curated in The Living Tree Project (Yarza et al., 2008)

2013), with general agreements on clades identified as well as their phylogenetic relationships. Both studies have found *Armatimonadetes* to be a very divergent phylum but differ in proposed taxonomic treatments based on normative sequence identity thresholds. Study in Chapter 4 took a relatively conservative approach and proposed the partitioning of two candidate phyla in order to maintain *Armatimonadetes* within the upper-boundary of sequence dissimilarity of prokaryotic phyla. In contrast, Yarza et al. took a more generalised approach (see above) and recommended the separation of *Armatimonadetes* into five phyla-level groups

7.2.2 - Genomic analysis of *C. calidirosea* T49^T

Chapter 5 documents the analysis of *C. calidirosea* T49^T genome, which was the first description of an *Armatimonadetes* genome. Phylogenetic inference from concatenated amino acid sequences of conserved genes supported the hypothesis that *Chloroflexi* is the most closely-related formal phylum. However, the analysis also showed that TM7, a candidate division represented by partially reconstructed genomic sequence, may be even closer to *Armatimonadetes*.

Analysis of gene annotations found diverse carbohydrate metabolism pathways, transporters, and extracellular glycosyl hydrolases, corresponding to a carbohydrate-based metabolism, supporting the hypothesis regarding a diverse carbohydrate-utilisation capability observed in growth experiments (Lee et al., 2011). In addition, putative genetic features related to other physiological observations, such as carotenoid production and the leucine auxotrophy were identified. These findings corroborate previous physiological characterisation and contribute to a better understanding of the ecological niche of *C. calidirosea.* Finally, genomic features with no direct relation to clearly-observable phenotypes were recognised, including the apparent genome disorganisation (i.e. the lack of commonly-conserved operons) and the high abundance of sigma factors. These findings progressed our understanding of *C. calidirosea* beyond presence or absence of biochemical pathways and raised interesting questions regarding genomic regulation of the bacterium. Since there are few reports (Glöckner et al., 2003) to this unusual scattering of operonic genes, we set forth to compare the genomes of three additional *C. calidirosea* isolates (Chapter 6) in order to understand the underlying dynamics and factors influencing the *C. calidirosea* genomes.

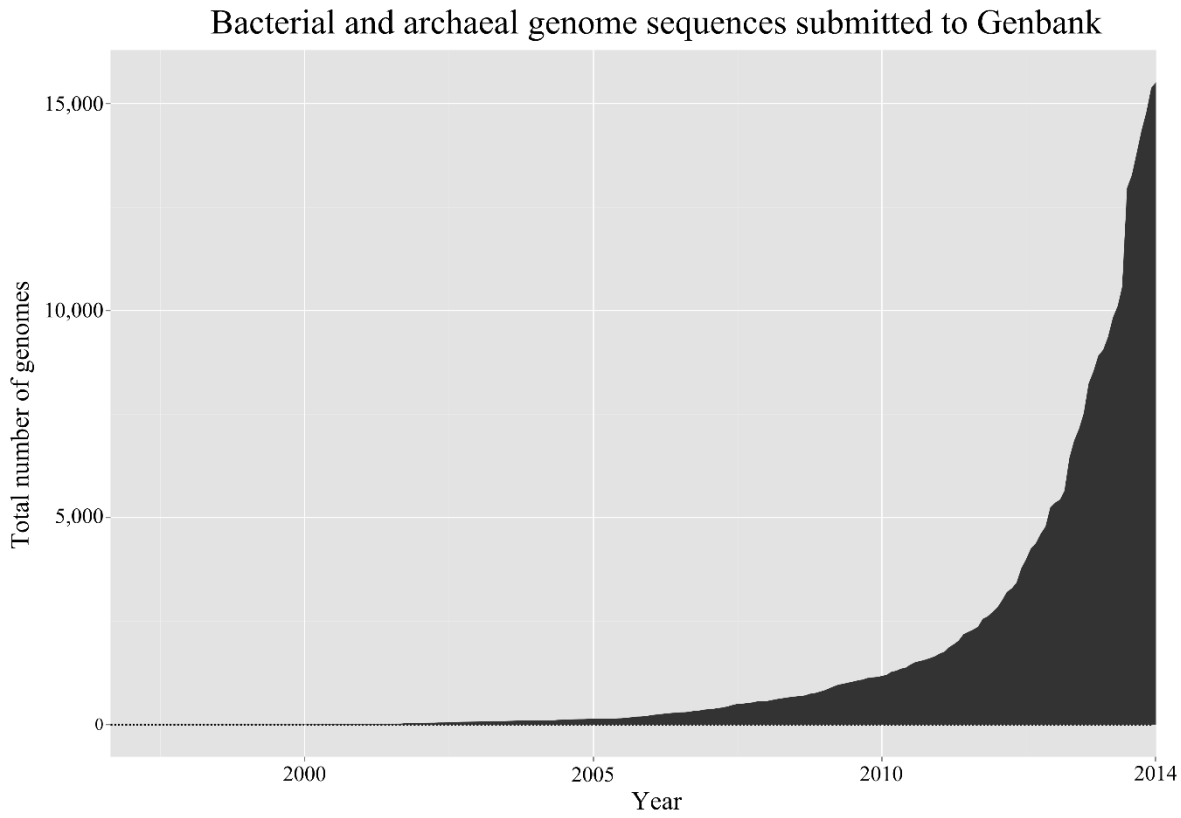


Figure 7.2 - The number of bacterial and archaeal genome sequences in GenBank database has increased exponentially. Data source: NCBI Genome Report retrieved 17th October, 2014.

As a result of the research described in this thesis, the novel bacterial phylum *Armatimonadetes* and in particular the species *Chthonomonas calidirosea* are now much better understood than they were at the time of discovery and cultivation of the first isolate from phylum (Stott et al., 2008). In the last decade, other efforts toward understanding microbial "dark matter" have also resulted in some success. New phyla such as *Caldisericum* (Candidate Division OP5) (Mori et al., 2009) and *Elusimicrobia* (Geissinger et al., 2009) have been established with the characterisation analysis of their respective type species[h]. However, genomic descriptions of these valuable type strains may be absent or may trail behind the publication of the phenotypic description, as in the cases of *Caldisericum exile* and *Armatimonas rosea*. Nonetheless, thanks to the development of sequencing technologies (see Section 1.3), genomic sequencing of novel

[h] *Synergistetes* (Jumas-Bilak et al., 2009) represents a special case, as the novel phylum was delineated through the resolution of previously misclassified taxa via phylogenetic analysis (Hugenholtz et al., 2009). An issue covered in Chapter 4 and paragraphs at the beginning of this section.

bacteria or archaea has become increasing affordable. Publicly-available genomic sequences have been increasing exponentially every year since 1998 (Figure 7.2**)**.Strain isolation has traditionally been the primary barrier to characterising "dark matter". Once a novel microorganism has successfully been isolated and cultured, integrating its genomic data with its physiological profile may soon become a matter of course, if not a requirement in the description of novel microbial diversity, instead of physiological characterisation alone.

7.2.3 - Within-species genomic variations of *C. calidirosea*

In Chapter 6, the comparative analysis of *C. calidirosea* isolate genomes as well as their associated sample sites led to a better understanding of the relationships between *C. calidirosea* and various environmental parameters. The study showed that the *C. calidirosea* isolates shared highly-conserved genomes. From the genomic sequences, the inferred phylogeny exhibited a pattern similar to the geographical relationships of the sample sites. In addition, while ~17 isolate-variant genes were detected, the carbohydrate metabolic profile of the isolates exhibited a rank correlation pattern (for the tested set of carbohydrate substrates) that was similar to the inferred phylogeny among the isolates. The relative abundances of the isolates were low in their respective communities (ranging from 0.006 % to 0.3 %). The correlation between genome phylogeny and physiology with geographical pattern suggests that the genomic homogeneity of the low-abundance *C. calidirosea* was caused by local extinction and rapid aeolian dispersal. Community profile analysis also showed that no phylotypes closely related to *C. calidirosea* were found among the four sites. In addition, a previous analysis of the ecological distribution of *Armatimonadetes* phylotypes (Dunfield et al., 2012) indicated that no environmental phylotypes closely related to *C. calidirosea* were associated with thermophilic environments. This evidence points to the ecological isolation of *C. calidirosea* to closely-related species due to recently-acquired thermophily, perhaps analogous to the evolution of thermophily seen in *Synechococcus* strains (Miller & Castenholz, 2000). This may act as another possible mechanism contributing to the conservancy of the genomes, as horizontal gene transfer tends to occur between closely related taxa (Reno et al., 2009).

Thermophilic microorganisms are often considered ideal models for testing microbial ecological theories (Hreggvidsson et al., 2012; Reno et al., 2009), particularly those regarding microbial biogeography (e.g., the spatial and temporal distribution of microbial diversity), due to their distinct and isolated habitats and relatively simple communities. The remoteness of suitable habitats and difficulty in isolating novel strains however, present challenges for culture-dependent studies. As closely-related strains exhibit few

phenotypic differences. In order to investigate genetic divergence in relation to the environment, earlier studies relied on molecular profiling methods such as ribotyping, DNA:DNA hybridisation, or pulse field gel electrophoresis (e.g., Moreira et al., 1997; Renders et al., 1996; Williams et al., 1996). The development of DNA sequencing technologies has enabled more direct comparison of 16S rRNA gene sequences, conserved functional genes (Kiewitz & Tümmler, 2000; Whitaker et al., 2003), as well as the entire genome (Calteau et al., 2014; Makarova et al., 2006; Parkhill et al., 2003; Reno et al., 2009). Comparative genomic analysis allows the identification of evolutionary processes such as divergence at the scale of individual nucleotides, and helps to associate phenotypic variations to underlying genetic variations. Furthermore, genomic DNA sequences are more easily compared between studies than are electrophoresis banding patterns. Unfortunately, few studies of this kind have been published regarding thermophilic prokaryotes. Therefore, the comparative analysis of *C. calidirosea* genomes and the associated environment may provide valuable and contrasting perspectives on thermophilic microbial ecology to the more established taxa such as *Thermus* spp. (Brüggemann & Chen, 2006; Moreira et al., 1997), *Sulfolobus* spp. (Nayak, 2013; Reno et al., 2009), and thermophilic *Cyanobacteria* (Miller et al., 2000), in order to identify commonalities due to shared conditions (thermophily) as well as lineage-specific ecological interactions.

7.2.4 - Questions arising from this study

Overall, this thesis provided foundational insights into the phylogeny, genomics, and ecology of *Chthonomonas calidirosea* and *Armatimonadetes.* The findings from the research also raise new questions and hypotheses to be answered or tested. An overview of these questions are outlined below.

1. **What are the conserved and variable genetic features among *Armatimonadetes* species?** Other species within *Armatimonadetes* (*A. rosea* and *F. ginsengisoli*, as well as future novel species) provide interesting targets for comparative analysis. An integrative comparison of physiology and the underlying genetics may provide insights into the characteristics specific to the *Armatimonadetes* lineage. So far, analysis of ecological distribution based on where environmental phylotypes were detected as well as the three characterised species suggest that *Armatimonadetes* are mainly distributed in soil and sediment (Dunfield et al., 2012). Investigation into the genetics of *Armatimonadetes* therefore may yield clues on the physiologies which enable the phylum to survive in these environments.

2. **How endemic is *C. calidirosea*?** As noted earlier, no closely-related phylotypes of *C. calidirosea* appear to originate from thermophilic environments. Compared to other thermophilic prokaryotes such as *Thermus thermophilus* and *Sulfolobus islandicus*, the genomes of *C. calidirosea* appear to be highly conserved within the geographical scale investigated (Chapter 6). *C. calidirosea* has not been identified outside of the TVZ, which may act as an ecologically isolated "island". In order to identify closely related phylotypes, the 16S rRNA gene sequence of *C. calidirosea* $T49^T$ was searched against NCBI GenBank (Benson et al., 2011) nr database (Accessed December 2014) as well as the Earth Microbiome Project (EMP) dataset (Gilbert et al., 2011; Release January 2014). Nucleotide BLAST searches of these two datasets (NGS surveys and Sanger reads) found no closely-related phylotypes outside of Taupō Volcanic Zone. Furthermore, none of the closest hits were associated with thermophilic environments. The closest hits related to $T49^T$ were approximately within the same order (*Chthonomonadales*), based on the 89-91 % sequence identity, EMP annotation data, and analysis linking 16S rRNA gene sequence identity and taxonomic assignment (Konstantinidis & Tiedje, 2005). The environmental distribution of the closet phylotypes hits supported the findings by Dunfield et al. (2012), consisting of temperate soil environments. If the species is present outside of this area, the hypothesis on genome conservancy can be tested at a much larger geographical scale.

3. **How are genes regulated in *C. calidirosea*?** Specifically, carbohydrate metabolism in response to multiple substrates as well as carotenoid and biofilm production as stress responses. Sigma factors are hypothesised as the more dominant mean of gene expression regulation due to the lack of typically conserved operons in *C. calidirosea*. Identifying regulatory patterns in response to these environmental stimuli may provide a better ecological context to the physiology and genetics of *C. calidirosea*.

7.3 - Future directions

Outside the direct scope of this research (phylum *Armatimonadetes*), there has been a matching growing trend in characterising other previously unknown microbial lineages (Section 7.1), as the roles microorganisms play as well as the scope of the biological "dark matter" became better understood. Investigations into novel microorganisms and their genomic, physiological, and ecological diversity all lead to the broadening of perspectives in microbiology from the better studied model species.

As outlined in this chapter, this study represents many first insights (phylogenetic, genomic, and ecological) into phylum *Armatimonadetes*. The research also raises questions I believe warrant further attention (Section 7.2.4). As such, we propose possible research directions to address these new questions, with some technical considerations:

1. **Community survey of *Armatimonadetes* phylotypes** – Currently, little information on *Armatimonadetes* is available outside of the three representative species, aside from clades (based on 16S rRNA gene sequences) which have exhibited some congruence with trends in ecological distributions (Dunfield et al., 2012). With in-depth and targeted surveys connecting the community profile and physicochemistry of habitats, the relationships between genetic, phenotypic, ecological traits, and inferred phylogeny of *Armatimonadetes* may be further illuminated (Yarza et al., 2014). The identification of *Armatimonadetes* in different environments may also provide potential targets for isolation of novel species.

   Since the recognition of *Armatimonadetes*, the phylum is detected regularly in environmental surveys. In 2014 alone, Google Scholar has indexed approximately 100 publications identifying the phylum in environments such as epibiotic bacteria in freshwater lakes (He et al., 2014), pinewood forest soil (Shi et al., 2015), wastewater sludge (Weissbrodt et al., 2014), and soil from a creosote-contaminated site (Mukherjee et al., 2014). Despite the prevalence of community surveying through high-throughput sequencing, the resulting data currently lacked publicly-available curated databases, analogous to SILVA (Pruesse et al., 2007), Greengenes (DeSantis et al., 2006), or RDP Project (Larsen et al., 1993) for full-length or near full-length 16S rRNA gene sequences that were extracted from primary INSDC (Karsch-Mizrachi et al., 2012) nucleotide databases (i.e. DDBJ (Tateno et al., 2002), GenBank (Benson et al., 2011), and ENA (Leinonen et al., 2011)[i]. Direct comparison of high-throughput sequencing reads from across studies and platforms therefore, required extensive data gathering and curation from the end-users. Consequently, current community studies through high-throughput sequencing are mostly self-contained, where short sequences are generated for the study itself, and pre-existing full length 16S

[i] It is worth noting that both SILVA and RDP Project offer analysis pipelines for user-supplied high-throughput sequencing datasets against their curated database of Sanger 16S rRNA gene sequences.

rRNA gene reads from curated databases are utilised for taxonomic reference. As a result, meta-analysis through the extraction of short reads and abundance data (i.e. OTU tables) from various sources remains a challenging process.

Standard operating procedures (SOPs) from Earth Microbiome Project (Gilbert et al., 2011), QIIME (Caporaso, Kuczynski, et al., 2010), and Mothur (Schloss et al., 2009), as well as the Biological Observational Matrix (BIOM) format standard for OTU tables (McDonald, Clemente, et al., 2012) represent some of many current efforts in addressing the issues and facilitate data analysis. These efforts may allow large-scale data mining, which relates phylogeny with ecology from metadata, a possibility in the foreseeable future. Meanwhile, to produce comparable data (community profile and environmental parameters), a standalone study profiling sample sites where *Armatimonadetes* were previously detected (e.g., see above environments) may yield valuable insight on the relationship between community compositions, phylogeny of *Armatimonadetes* phylotypes, and the metadata of the host environments. Inferred ecology from this process can also assist the cultivation of novel *Armatimonadetes* species.

2. **Culture-independent genomics through metagenomics** – In order to gain information on *Armatimonadetes* without cultivation (which has shown to be labour-intensive and the success in isolation capricious), the emerging metagenomic approach (Narasingarao et al., 2012; Pell et al., 2012; Sharon & Banfield, 2013) offers a compelling alternative. On one level, metagenomics may provide information on the functional genes involved in these communities in addition to the use of 16S rRNA maker genes to characterise communities associated with *Armatimonadetes*. Genomes of novel species may be reconstructed from communities with simple structure, or with the assistance of cell sorting and whole-genome amplification (Marcy et al., 2007). Recent developments in binning methods meant even the genomes of less abundant community members may be recovered (Albertsen et al., 2013; Wu et al., 2014). The recovery of functional genes in the community yields clues to the overall ecological interactions involved while genome reconstruction provides genotypic information on the specific organism which can be used for comparative genomic analysis between *Armatimonadetes* species (see above Question 1). In addition, both methods help guide strategies to isolate and cultivate novel *Armatimonadetes*.

3. **Isolation, cultivation, and characterisation of novel *Armatimonadetes* spp.** – In the various aspects of investigating the microbial unknown, the cultivation of novel prokaryotes remains a particularly challenging process (see Section 7.1). Nonetheless, knowledge of (A) the location of the targeted *Armatimonadetes* species from community surveys, and (B) inferred ecology of the targeted species from metagenomic analysis, may provide vital clues towards the isolation and cultivation strategies. The cultivation of the novel species is a critical step leading to many potential downstream methodologies such as physiological characterisation, as well as genomic, proteomic, and transcriptomic analyses, where fundamental aspects of the phylum can be examined at a species level.

4. **Transcriptomics** – *C. calidirosea* T49^T cultures have been noted to produce pigments and exopolysaccharides in later growth phase, possibly as a stress response to culture acidification or lowering of oxygen saturation (Lee et al., 2014; Lee et al., 2011). In Chapter 5, *C. calidirosea* T49^T was noted for the lack of diauxic growth, which indicated the absence of carbon catabolite repression with the substrates tested. It was hypothesised that genes involved in carbohydrate metabolism, including extracellular glycosyl hydrolases and membrane bound transporters, are mostly constituently expressed in order to utilise a wide range of possible carbohydrates from the environment without the expression delays. An alternative hypothesis suggests that many of the genes, which were not organised in functionally related clusters (or operons), may instead be regulated by the abundant sigma factors. Gene regulation of *C. calidirosea* T49^T in response to environmental stimuli, such as the metabolism of multiple carbohydrates, or stress responses may be investigated through the comparison of the transcriptomes of *C. calidirosea* T49^T cultures subjected to these contrasting conditions (carbohydrate availability and stress stimuli). Synchronisation of growth phases between replicates would be a major challenge due to difficulties in controlling starting inoculum (liquid-to-liquid sub-culturing results in poor growth, while biomass form plate cultures are difficult to standardise). Furthermore, growth response in liquid cultures, even with pH control and aeration through a bioreactor, has shown to be inconsistent. Batches of cultures with the same starting optical density may result in different doubling times and varying cell density in stationary phase, some of which entered stationary phase early and resulted in failed cultures. These problems may be due to batch-specific concentrations of metabolic by-products inhibiting growth, or quorum sensing in *C. calidirosea* isolates, which required high initial cell concentration in order to transition from lag phase to exponential phase.

Synchronisation problem of growth phases may be addressed via a bioreactor which balances influx nutrient and efflux of culture, also known as a chemostat or turbidostat (Novick & Szilard, 1950), in order to achieve a stable exponential phase. Theoretically, concentration and cell state of the initial inoculum, as well as metabolite build-up would not be a problem once culture enters exponential phase in such condition, and thus reproducible and manipulable cell conditions may be achieved for transcriptomic analysis. In addition, if the culture (and cell state) is able to be manipulated between phases or nutrient conditions, then a chemostat would simplify experimental design by avoiding conducting several independent batches in order to produce replicates. If these issues can be addressed, contrasting expression levels of regulatory genes (e.g., sigma factors and riboswitches) as well as functional genes (e.g., carotenoid biosynthesis genes, chaperones, carbohydrate metabolic pathways) may provide a better connection between environmental stimuli and the genetic potential of *C. calidirosea* T49^{T} through gene regulation, aspects of which (e.g., operon vs. sigma factors) formed yet-to-be-tested hypotheses originating from genomic analysis. Investigations into gene regulation in response to stress and different energy sources will help us better understand the ecology of *C. calidirosea* T49^{T}, and may provide insights into the survival of bacteria within extreme environments.

Chapter 8 Appendices

8.1 - Supplementary materials for Chapter 4

Supplementary Table 8.1 - A list of GenBank/EMBL/DDBJ accession numbers of phylotypes and associated groupings used in this study. The naming of the groups (Group 1 to 12) was based on Dunfield et al. (2012).

Group 1: 62 sequences						
AB529679 #	AB128880	AB630923	AF316757	AF507702	AJ867895	AM697438
AY192276	AY212563	DQ395456	DQ532193	EF126246	EF126255	EF516002
EF516134	EF516140	EF516423	EF516651	EF683077	EU134986	EU134997
EU676408	FJ478742	FJ891045	FJ891043	FM872555	FM873242	FN421827
FN811253	FR667324	FR667332	GQ264304	GQ340247	GQ397039	GU214136
GU214138	GU214172	HQ008579	HQ910267	HQ910302	JF417814	JF449948
JF449949	GQ051146	GQ093431	HM263172	HM289313	HM335998	JF096489
JF115297	JF116006	JF122162	JF124785	JF159835	JF168907	JF175139
JF176789	JF198606	JF198700	JF198859	JF228055	JF235970	
Group 2: 28 sequences						
AF524022	AY661983	DQ984576	DQ984582	EF516412	EU861876	EU861894
EU861902	EU861928	EU861934	FJ465978	FJ466011	FJ466040	FJ466062
FJ466076	FJ466088	FJ466094	FJ466092	FR749826	GQ402731	HM459623
HQ595210	HQ595215	HQ622728	HQ622735	HQ622751	HQ674891	JF227074
Group 3: 39 sequences						
AM749780 #	AB630937	AB630939	AF523930	AJ009456	AY555775	EF032776
EF515204	EF515236	EF515962	EU134983	EU134985	EU134984	EU134993
EU134988	EU134994	EU134990	EU134987	EU134999	EU134995	EU135023
EU135024	EU135025	EU135026	EU445226	EU907898	FM866305	FN870192
FR749792	GQ487984	GU214142	GU454980	HQ397564	HQ397563	JF429056
JF833608	HM266973	JF168484	JF229088			
Group 4: 10 sequences						
AB630922	AJ009501	CU917629	CU918094	CU919397	CU919520	CU920897
CU927392	EF205585	HM187268				
Group 5: 53 sequences						
AB183862	AB240295	AB240383	AB300092	AB300089	AB300096	AB300112
AB300087	AB630921	AB630924	AB630925	AJ009490	AJ009504	AM490693
AY214187	CU918470	CU919218	CU919158	CU919205	CU923234	CU925187
CU924856	CU927522	DQ330148	DQ330650	DQ330654	DQ404749	DQ404748
DQ404796	DQ404826	DQ787723	EF205560	EU246197	EU266862	EU266864
EU266913	EU335186	EU471620	EU473674	EU843463	EU887975	FJ484643
FJ793161	FJ873291	GQ487914	GU363034	HM069114	HQ183987	HQ183986
HQ697775	HQ904206	GQ105507	HM334038			
Group 6: 6 sequences						
AF027092 #	DQ450815	EU289437	HM845887	HM845963	JF181234	

Group 7: 68 sequences						
GQ339893	AB286372	AF368184	AF368186	AF368188	AF418946	AJ271048
CU921283	CU923669	DQ330645	DQ450814	DQ501336	DQ975217	EF032775
EF205445	EF515998	EF516079	EF516213	EF516982	EF580941	EF648093
EU134991	EU134989	EU134992	EU332819	EU907880	FJ478889	FJ479523
FJ536900	FJ562143	FJ710638	FJ936919	GQ263839	GQ263888	GQ264155
GQ264378	AB240372	GQ359995	GQ396875	GQ396936	GQ402559	GU389936
GU444068	GU455127	GU983354	HM186240	HM187116	HM187171	HM187213
HM444982	HM445386	HM445430	HM445522	HQ114053	HQ397190	HQ640633
HQ827956	HQ827973	HQ828003	HQ864199	JF429001	JF833904	HM312995
HM319220	HM332632	JF095934	JF175681			
Group 8: 6 sequences						
AJ306784	AY188316	EU528175	FJ535562	GU389309	JF174971	
Group 9: 67 sequences						
GQ339893 #	AB286372	AF368184	AF368186	AF368188	AF418946	AJ271048
CU921283	CU923669	DQ330645	DQ450814	DQ501336	DQ975217	EF032775
EF205445	EF515998	EF516079	EF516213	EF516982	EF580941	EF648093
EU134991	EU134989	EU134992	EU332819	EU907880	FJ478889	FJ479523
FJ536900	FJ562143	FJ710638	FJ936919	GQ263839	GQ263888	GQ264155
GQ264378	AB240372	GQ359995	GQ396875	GQ396936	GQ402559	GU389936
GU444068	GU455127	GU983354	HM186240	HM187116	HM187171	HM187213
HM444982	HM445386	HM445430	HM445522	HQ114053	HQ397190	HQ640633
HQ827956	HQ827973	HQ828003	HQ864199	JF429001	JF833904	HM312995
HM319220	HM332632	JF095934	JF175681			
Group 10A: 2 sequences						
AF027090 #	EF205558					
Group 10B: 17 sequences						
AF445740	DQ324867	DQ490004	DQ645244	DQ645248	EU635941	EU635952
FM164957	FM164959	FN545885	FN545892	GU437340	GU437353	HM448246
HM448264	HM640993	HM640997				
Group 11: 80 sequences						
AB374366	AB374381	AJ863240	AM982676	AY218624	AY218769	AY250868
AY289488	AY796037	AY898020	DQ129347	DQ248296	DQ990931	EF019331
EF522341	EU135310	EU135315	EU344940	FJ592715	FJ592716	FJ790607
FJ790619	FJ821646	FN421489	FN811251	FR749715	FR749786	FR749810
FR749775	GQ379560	HM187141	HM187237	HM187251	HM238160	HM445432
HM559209	HQ327283	JF776932	GQ002367	GQ008023	GQ072311	GQ108732
GQ116227	GQ116329	HM262861	HM263359	HM265371	HM266914	HM268864
HM269093	HM269099	HM269200	HM269858	HM270410	HM277030	HM277599
HM278725	HM298801	HM330993	HM845933	JF108214	JF129009	JF144518
JF154853	JF164962	JF167709	JF167832	JF167878	JF168255	JF170899
JF174680	JF175029	JF176688	JF176758	JF176901	JF181172	JF199240
JF219978	JF228178	JF232332				

Group 12: 51 sequences						
AB015272	AB177187	AB234287	AB177336	AB177145	AB300064	AJ009463
AY197424	CU917740	CU918550	CU919290	CU919342	CU919775	CU922588
CU924237	CU921809	CU923024	CU927602	CU926664	CU927585	DQ676417
EF515388	EF515601	EF515625	EU181506	EU245416	EU245570	EU245588
EU245988	EU266867	EU385899	EU385881	EU591645	FJ517033	FJ645697
GQ472455	GU302427	HM041925	HM185833	HM243824	HM992561	HQ183990
HQ183991	HQ588564	HQ588576	DQ330736	FJ485378	FJ517017	HM187449
AJ543756	GQ354974					
Group 13: 3 sequences						
EU385871	FJ484560	FJ901637				

Key phylotypes of *Armatimonadetes*: AB529679 *A. rosea*, Tamaki et al. (2011); AM749780 *C. calidirosea,* Lee et al. (2011); GQ339893 *F.ginsengisoli*, Im et al. (2012), and the original OP10 clones AF027090 and AF027092 (Hugenholtz et al., 1998).

Supplementary Table 8.2 - Phylotypes that changed grouping between BI method and other methods (referred as "jumping" sequences) are also presented. These "problematic" sequences were removed phylogenetic analyses.

"Jumping" /Chimeric sequences: 15 sequences						
AB630924	EF515962	EU134983	EU134985	EU134984	EU134986	EU134982
EU134993	EU134991	EU134989	EU861851	HM069114	JF198649	DQ330148
CU919520						

Supplementary Table 8.3 - GenBank/EMBL/DDBJ accession numbers in the OP10A, OP10B and OP10C outgroups used in this research. The outgroups were adapted from Dalevi et al. (2001).

Outgroup phylotypes derived from OP10A dataset : 15 sequences						
AF050559	AF027005	CP001743	CU928160	L09663	L10659	M21774
M11223	M34132	M21413	M59231	M59176	U75602	X90515
X95744						
Outgroup phylotypes derived from OP10B dataset: 17 sequences						
AF027004	AF050564	AJ308500	CP001807	D16296	D26171	FN178468
L11703	M34115	M59051	M83548	U75647	X60514	X71836
X77215	X82559	X86776				
Outgroup phylotypes derived from OP10C dataset: 14 sequences						
AF050559	AF027005	AF050564	AJ308500	CP001743	D16296	FN178468
L09663	M34115	M21413	M83548	U75647	X60514	X90515

Supplementary Table 8.4 - Support values of phylogenetic groupings (Groups, hereby abbreviated as G[number]). G1 to G12 are based on the definition of Dunfield et al. (2012). The outgroup datasets (OP10A,B,C) were adapted from Dalevi et al. (2001). Group10A and 10B are subgroups derived from phylotypes from Group10. Green coloured cells indicate support values meeting the threshold condition (≥ 70 % for ML and NJ, ≥ 95 % for BI), red coloured cells indicate support values not meeting those conditions, "U" indicates the grouping was absent in the analysis (unresolved).

Analysis	NJ		PhyML	RAxML	BI
Model	JC	Olsen	GTR	GTRMIX	GTR
Bootstrap / Chain Length	2000	2000	500	500	3100000
G1	99	99	98	99.2	1
G2	>99	>99	100	100	100
G3	99	99	99.8	99.8	100
G4	>99	>99	100	99.8	100
G5	U	U	32	31.6	58.7
G6	92	91	89	96	100
G7	99	99	100	100	100
G8	>99	>99	100	100	100
G9	>99	>99	100	100	100
G10	U	U	41.4	69.2	100
G10A	>99	>99	100	100	100
G10B	>99	>99	100	100	100
G11	98	98	99.4	99.8	100
G12	76	79	90.4	99.8	100
G13	>99	>99	100	100	100

Supplementary Table 8.5 - Support values of hypothetical sets of groups, including the entire ingroup dataset. For NJ, the support values were collated directly from the consensus trees generated, when the consensus tree topology did not show the hypothetical set under a single clade, it was recorded as (U)nresolved. For ML and BI methods, support values were calculated using BAli-Phy (Suchard et al., 2006) which examined the raw trees directly for the occurrence of the particular hypothetical set as a monophyletic group. Green coloured cells indicate support values meeting the threshold condition (≥ 70 % for ML and NJ, ≥ 95 % for BI), red coloured cells indicate support values not meeting those conditions.

Analysis	**NJ**		**PhyML**	**RAxML**	**BI**
Model	**JC**	**Olsen**	**GTR**	**GTRMIX**	**GTR**
Bootstrap / Chain Length	**2000**	**2000**	**500**	**500**	**3100000**
G1-G2	99	99	97.2	99.6	100
G2-G3	U	U	0.6	0.2	0
G1-G2-G3	43	41	44.8	50.8	100
G4-G5	81	80	7.2	57.2	100
G1-G2-G3-G4-G5	30	28	29	22.2	72.5
G7-G8	U	U	23.6	19.6	61.1
G7-G13	25	23	13.2	16	51
G6-G7-G13	21	20	5.8	14.8	2.2
G7-G8-G13	U	U	59.6	29.4	84.7
G6-G7-G8-G13	22	22	18.8	19	95.4
G9-G10AB	U	U	35	57.2	100
G10AB-G11-G12-Outgroup	26	U	10.8	25.4	0
G9-G10A-G10B-G11-G12-Outgroup	48	49	38.4	76.2	100
G11-G12	34	37	37.6	22.8	0
G11-Outgroup	U	U	1.8	1	0
G12-Outgroup	U	U	38.4	69.8	100
G11-G12-Outgroup	70	68	85.2	98.2	100
Ingroup	78	80	79.6	91.4	100

AEM
Journals.ASM.org

Phylogenetic Delineation of the Novel Phylum *Armatimonadetes* (Former Candidate Division OP10) and Definition of Two Novel Candidate Divisions

K. C. Y. Lee,[a,b] C. W. Herbold,[b] P. F. Dunfield,[c] X. C. Morgan,[a,d] I. R. McDonald,[b] M. B. Stott[a]

GNS Science, Extremophile Research Group, Taupo, New Zealand[a]; School of Biological Sciences, University of Waikato, Hamilton, New Zealand[b]; Department of Biological Sciences, University of Calgary, Calgary, Canada[c]; Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA[d]

Small-subunit (SSU) rRNA gene sequences associated with the phylum *Armatimonadetes* were analyzed using multiple phylogenetic methods, clarifying both the phylum boundary and the affiliation of previously ambiguous groupings. Here we define the *Armatimonadetes* as 10 class-level groups and reclassify two previously associated groups as candidate divisions WS1 and FBP.

Candidate division OP10, first identified in 1998 (1), was renamed as the phylum *Armatimonadetes* with the characterization of the type strain *Armatimonas rosea* YO-36^T (2). Hundreds of phylotypes from a broad range of environmental niches have been added to the group since its inception (1, 3); as of April 2012, 568 sequences in the SILVA database alone were classified as *Armatimonadetes* or OP10, and 364, 568, and 653 small-subunit (SSU) rRNA gene sequences were classified as either *Armatimonadetes* or candidate division OP10 in the EMBL, SILVA, and RDP databases, respectively. Using the SILVA database numbers, the sequences have a maximum SSU rRNA gene sequence dissimilarity of ~29%, a significantly diverse phylum compared to the average phylum diversity of 19.3% (4). However, the phylogeny of *Armatimonadetes* is still poorly defined, as recent publications (2, 5–9) have not been able to agree upon a consistent and well-supported consensus tree. These publications have used a variety of methods and outgroup/ingroup sequence selections, which has resulted in the description of between 4 and 12 (in some cases poorly supported) subgroupings. Currently, it is difficult to compare published tree topologies and group nomenclatures due to the inconsistent application of SSU rRNA gene sequences and the application of different methods to generate phylogenetic frameworks for the former candidate division OP10/*Armatimonadetes*. Thus, the phylogenetic diversity and subdivisional architecture of *Armatimonadetes* have remained uncertain.

In order to clarify the phylogenetic relationships within the *Armatimonadetes* and establish well-supported taxon boundaries, we used multiple phylogenetic methods to (i) confirm the division boundaries and (ii) define the subphylum level group structure(s), including the previously identified classes *Armatimonadia* (2) and *Chthonomonadetes* (8) and the recently validated class *Fimbriimonadia* (7).

For phylogenetic inference, near-full-length SSU rRNA genes putatively identified as *Armatimonadetes* or OP10 were selected from the SILVA SSU NR (nonredundant) Release 108 database (10) for initial sequence alignment. This ingroup data set consisted of 492 sequences. The outgroup consisted of 46 sequences, which were a combination of the three outgroup sets described by Dalevi et al. (5), named OP10A, OP10B, and OP10C. The combined outgroup was replicated with only minor adjustments: if sequences were not included in the SILVA NR database, the most closely related sequences in the SILVA SSU NR database were used. All sequence accession numbers here are listed in Data Set S1 in the supplemental material. The subdivisional group numbering (groups 1 to 12) recently defined by Dunfield et al. (6) was used as a default, nonpresumptuous naming scheme for hypothesis testing. Monophyletic groupings, once established, were subsequently given unique identifiers based on the name of the first validly published phylotype within said group, as described in the Greengenes database (11).

Two applications of neighbor-joining (NJ) and of maximum-likelihood (ML) methods and one application of Bayesian inference (BI) were utilized to define the nodal support for class level groupings, the details of which are described in File S1 in the supplemental material. Of 492 OP10/*Armatimonadetes* sequences extracted from the SILVA database, 15 flagged as chimeric using Pintail (12) were identified (see Data Set S1 in the supplemental material) and excluded from further calculations.

A radial consensus phylogenetic tree summarizing the final grouping nomenclature and relationships is presented in Fig. 1 along with the support of the nodes by the five methods used. Support values of the groupings as well as tested hypothetical combinations of these groupings are presented in Data Set S2 in the supplemental material. A total of 12 monophyletic groupings were defined from the *Armatimonadetes*/OP10 ingroup data set, which included two proposed candidate divisions and 10 class-level divisions (Fig. 1). Bipartition support values of groups within *Armatimonadetes* confirm the monophyly of the previously described classes *Armatimonadia* (group 1 to 2), *Chthonomonadetes* (group 3), and *Fimbriimonadia* (group 9), as defined by the respective authors (2, 7, 8). The boundary of the class *Armatimonadia* (group 1) should also include group 2 due to the strong

Received 29 October 2012 **Accepted** 25 January 2013
Published ahead of print 1 February 2013
Address correspondence to K. C. Y. Lee, k.lee@gns.cri.nz.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.03333-12.

doi:10.1128/AEM.03333-12

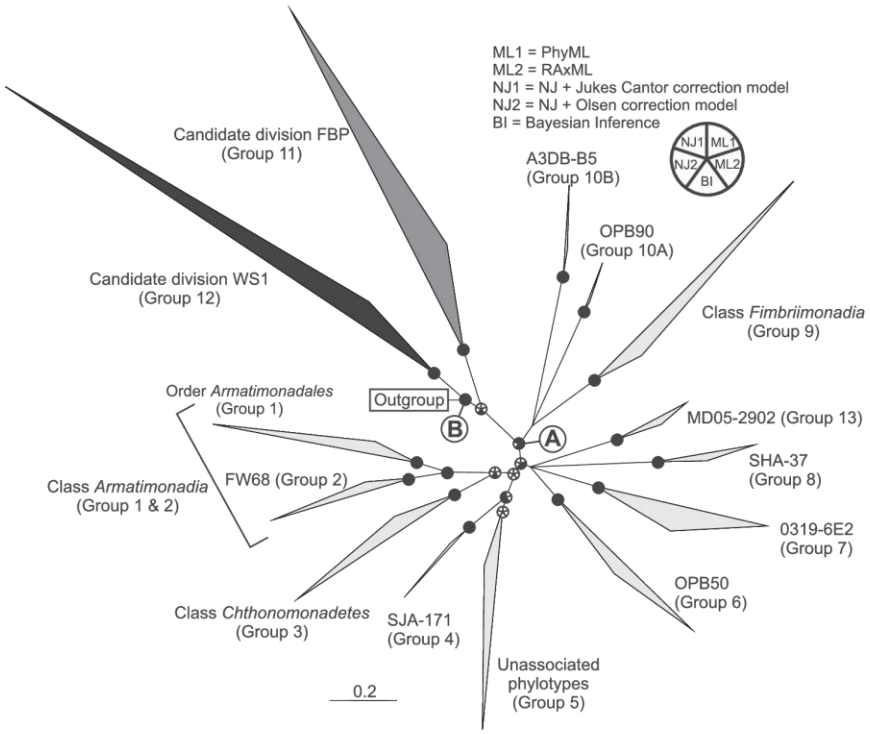


FIG 1 Unrooted consensus tree showing the phylum *Armatimonadetes* and affiliated groups. The groups were based those described by Dunfield et al. (6); the phylotypes (accession numbers) within each group are listed in Data Set S1 in the supplemental material. The analyses of support for the groups are detailed in Data Set S2. The nodal support here displays supports by five phylogenetic methods (two NJ, two ML, and one BI analysis), which are represented by pie charts on the individual bifurcation nodes. Support is defined as a ≥70% bootstrap proportion value (NJ and ML) and a ≥95% posterior probability value (BI). A node supported by the specific method has the corresponding portion of the pie chart shaded black. Multifurcations were manually introduced into nodes without support from any of the methods (15). Taxon designations are shown based on the position of type strains and the supported monophyly of nodes. On the consensus tree, node A represents the phylum *Armatimonadetes* and node B represents the superphylum resulting from the inclusion of *Armatimonadetes* with the candidate divisions FBP (group 11) and WS1 (group 12). The scale bar represents 0.2 nucleotide substitution per site.

support for monophyly of the combined "supergroup" (groups 1 and 2); this amalgamation is consistent with the variety of ecological niches reported previously, including temperate soils, fresh water, human skin, and microbial biofilms (6). Furthermore, we suggest that that group 1 of the expanded class *Armatimonadia* (2, 6–8) now represent the order *Armatimonadales* (Fig. 1).

Figure 1 shows two deeply branching nodes, A and B, with strong monophyly support. Node B included all the ingroup sequences and was supported by all five of the phylogenetic methods. Node A, which was supported by four of five methods, included the original OP10 phylotypes and all three described *Armatimonadetes* classes but consistently excluded groups 11 and 12. The phylogenetic distance of these two groups from the phylum type strain, strong and reproducibly monophyletic support (3), and distinct environmental distributions (6) indicate that groups 11 and 12 should be considered candidate divisions distinct from *Armatimonadetes*. Sequence dissimilarity analysis (Fig. 2) corroborates this assertion, as maximum and mean similarities equal (group 12) or exceed (group 11) the upper limit (95% confidence interval) of the previously published average maximum bacterial phylum boundary (21.6% ± 2.0%) (4). Although group 12 was recently included within the *Armatimonadetes* in the SILVA database, it was previously identified as candidate division WS1 (13). Here we term group 11 as candidate division FBP based on the earliest published, near-full-length (>1.4-kb) phylotype, clone FBP249 (AY250868) (14). The two candidate divisions also share a well-supported relationship with *Armatimonadetes* (node B), a relationship which would therefore likely represent the basal node for a putative highly divergent superphylum.

In this study, applying multiple methods to discern phylogenetic relationships within *Armatimonadetes* has highlighted problematic sequences as well as nodes with inconsistent support values. Despite extensive analyses, the phylogenetic resolution of *Armatimonadetes* should be considered ongoing and is still limited by factors such as the limits of phylogenetic signals within the current SSU rRNA gene data sets. Evidence suggests the independent monophyly of the groups OPB90 (group 10A) and A3DB-B5 (group 10B), although the two groups were previously associated and the clones in both groups were found only in geothermal environments (6); whether this is a result of convergent evolution or limitations in the data set available awaits future investigations. Group 5 was also unsupported as a monophyletic group, but unlike with group 10, we were unable to identify any clearly distinguishable

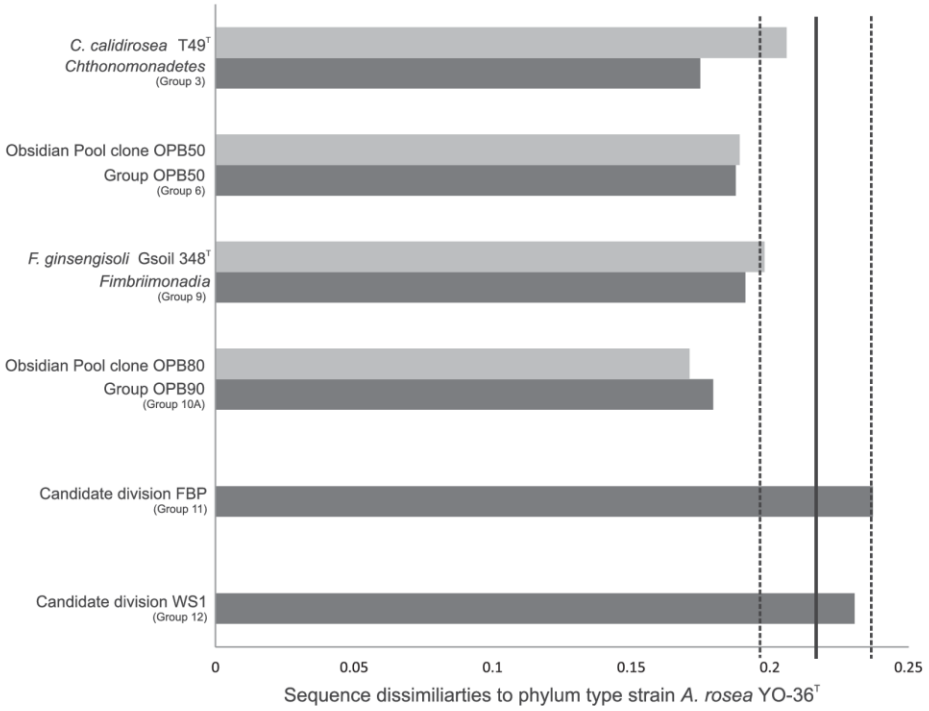


FIG 2 Bar graph displaying sequence dissimilarities between *Armatimonadetes* type strain *A. rosea* YO-36^T and key phylotypes (*Armatimonadetes* isolates and the original OP10 clones). The graph also displays mean sequence dissimilarities between *A. rosea* YO-36^T and groups containing the key phylotypes *Chthonomonadetes* (group 3), OPB50 (group 6), *Fimbriimonadia* (group 9), and OPB90 (group 10A), candidate division FBP (group 11), and candidate division WS1 (group 12). The phylotypes (and accession numbers) within each group are listed in Data Set S1 in the supplemental material. The solid vertical line marking the value of 0.216 indicates the average maximum phylum sequence dissimilarity calculated by Yarza et al. (4). The flanking dotted line indicates the 95% confidence interval (±0.02).

monophyletic subgroups. Therefore, its incumbents should be considered a collection of unassociated phylotypes.

In summary, the data presented in this study clarify the phylogenetic architecture of the phylum *Armatimonadetes*. The research confirms the monophyly of the three described *Armatimonadetes* classes along with seven additional class-level groupings and defines the basal node of the phylum. In addition, two clades (WS1 and FBP) were demonstrated to be consistently monophyletic and hence were confirmed as candidate divisions.

ACKNOWLEDGMENTS

This work was supported by the Sarah Beanland Memorial Scholarship and Geothermal Resources of New Zealand (GRN) funding at GNS Science.

REFERENCES

1. **Hugenholtz P, Pitulle C, Hershberger KL, Pace NR.** 1998. Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180**:366–376.
2. **Tamaki H, Tanaka Y, Matsuzawa H, Muramatsu M, Meng X-Y, Hanada S, Mori K, Kamagata Y.** 2011. *Armatimonas rosea* gen. nov., sp. nov., of a novel bacterial phylum, *Armatimonadetes* phyl. nov., formally called the candidate phylum OP10. Int. J. Syst. Evol. Microbiol. **61**:1442–1447.
3. **Hugenholtz P, Goebel BM, Pace NR.** 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. **180**:4765–4774.
4. **Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R.** 2010. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst. Appl. Microbiol. **33**:291–299.
5. **Dalevi D, Hugenholtz P, Blackall LL.** 2001. A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. Int. J. Syst. Evol. Microbiol. **51**:385–391.
6. **Dunfield PF, Tamas I, Lee KC, Morgan XC, McDonald IR, Stott MB.** 2012. Electing a candidate: a speculative history of the bacterial phylum OP10. Environ. Microbiol. **14**:3069–3080.
7. **Im W-T, Hu Z-Y, Kim K-H, Rhee S-K, Meng H, Lee S-T, Quan Z-X.** 2012. Description of *Fimbriimonas ginsengisoli* gen. nov., sp. nov. within the *Fimbriimonadia* class nov., of the phylum *Armatimonadetes*. Antonie van Leeuwenhoek **102**:307–317.
8. **Lee KC, Dunfield PF, Morgan XC, Crowe MA, Houghton KM, Vyssotski M, Ryan JL, Lagutin K, McDonald IR, Stott MB.** 2011. *Chthonomonas calidirosea* gen. nov., sp. nov., an aerobic, pigmented, thermophilic microorganism of a novel bacterial class, *Chthonomonadetes* classis nov., of the newly described phylum *Armatimonadetes* originally designated candidate division OP10. Int. J. Syst. Evol. Microbiol. **61**:2482–2490.
9. **Portillo MC, Gonzalez JM.** 2008. Members of the candidate division OP10 are spread in a variety of environments. World J. Microbiol. Biotechnol. **25**:347–353.
10. **Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. **35**:7188–7196.
11. **McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P.** 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. **6**:610–618.
12. **Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ.** 2005. At least 1 in 20 16S rRNA sequence records currently held in public repos-

itories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. **71**:7724–7736.
13. **Dojka MA, Hugenholtz P, Haack SK, Pace NR.** 1998. Microbial diversity in a hydrocarbon- and chlorinated solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl. Environ. Microbiol. **64**:3869–3877.
14. **de la Torre JR, Goebel BM, Friedmann EI, Pace NR.** 2003. Microbial diversity of cryptoendolithic communities from the McMurdo Dry Valleys, Antarctica. Appl. Environ. Microbiol. **69**:3858–3867.
15. **Peplies J, Kottmann R, Ludwig W, Glockner FO.** 2008. A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. Syst. Appl. Microbiol. **31**:251–257.

8.1.2 - Supplementary materials for the short-form version

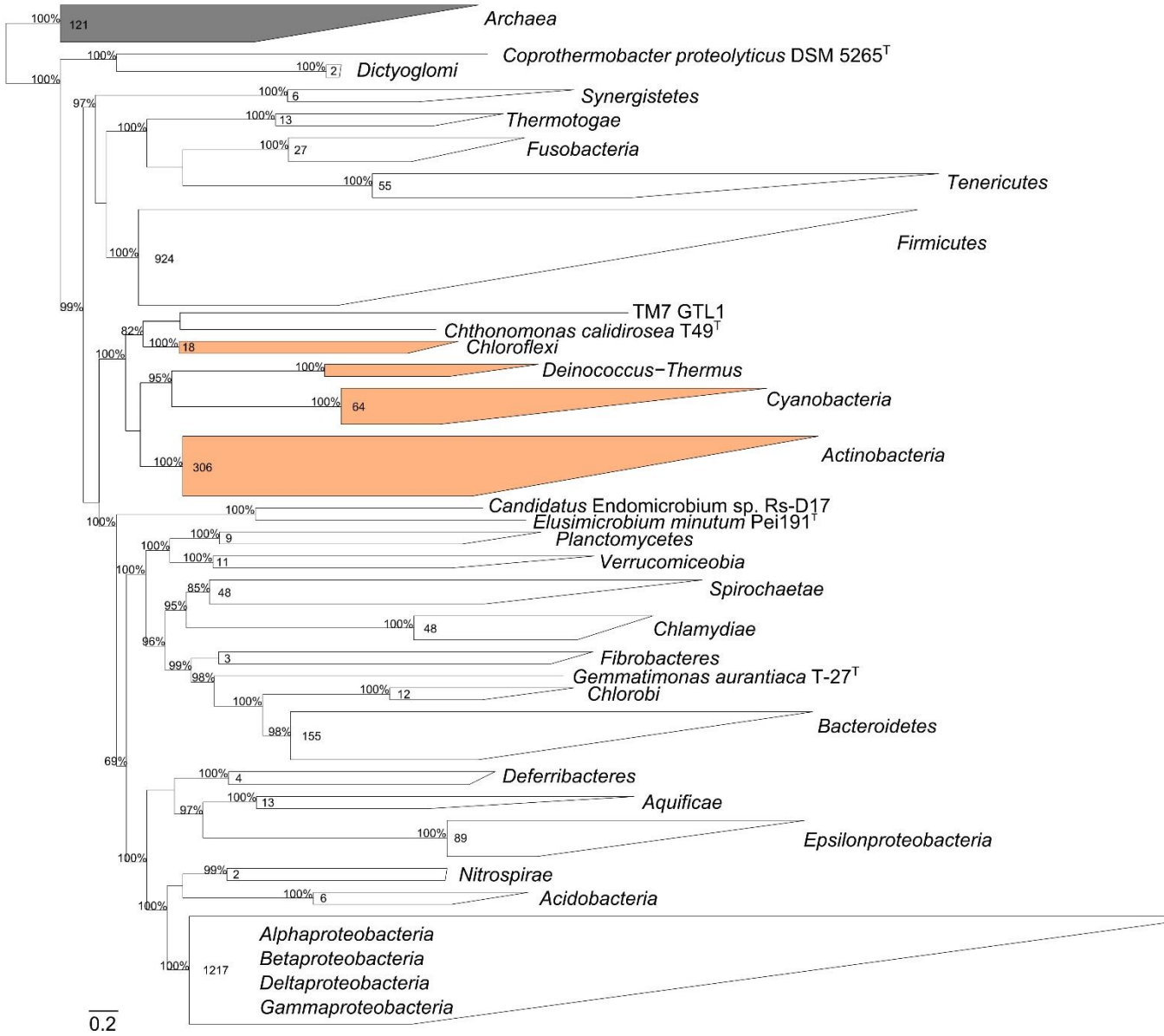
Supplementary dataset S1, S2, as well as additional description of the phylogenetic methods and the support value thresholds used in the study referenced in the short-form manuscript that was published in *Applied and Environmental Microbiology* (Lee et al., 2013) can be found at http://dx.doi.org/10.1128/AEM.03333-12 and as digital supplementary files attached to this thesis with the following names:

Chapter 4 – Short Form – Supplementary Methods.pdf
Chapter 4 – Short Form – SupplementaryDataset S1.xls
Chapter 4 – Short Form – SupplementaryDataset S2.xls

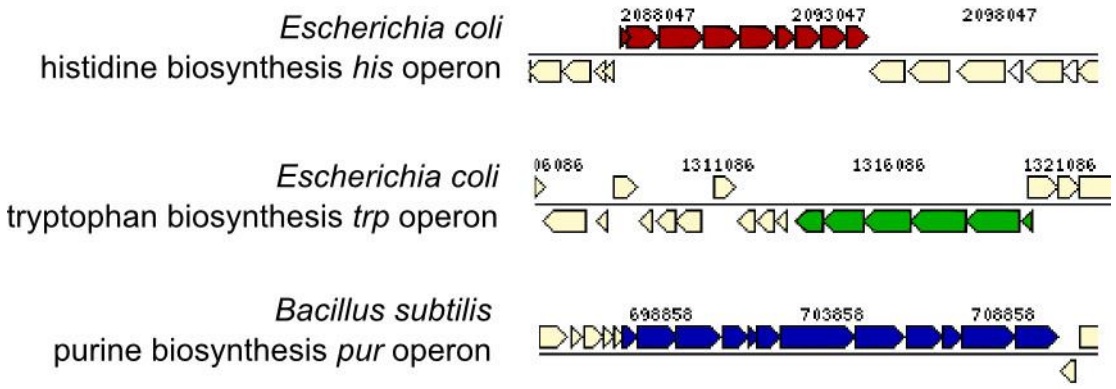
8.2 - Supplementary materials for Chapter 5

Supplementary materials associated with this study, published in *The ISME Journal* (Lee et al., 2014) can also be found in digital form at http://dx.doi.org/10.1038/ismej.2013.251.

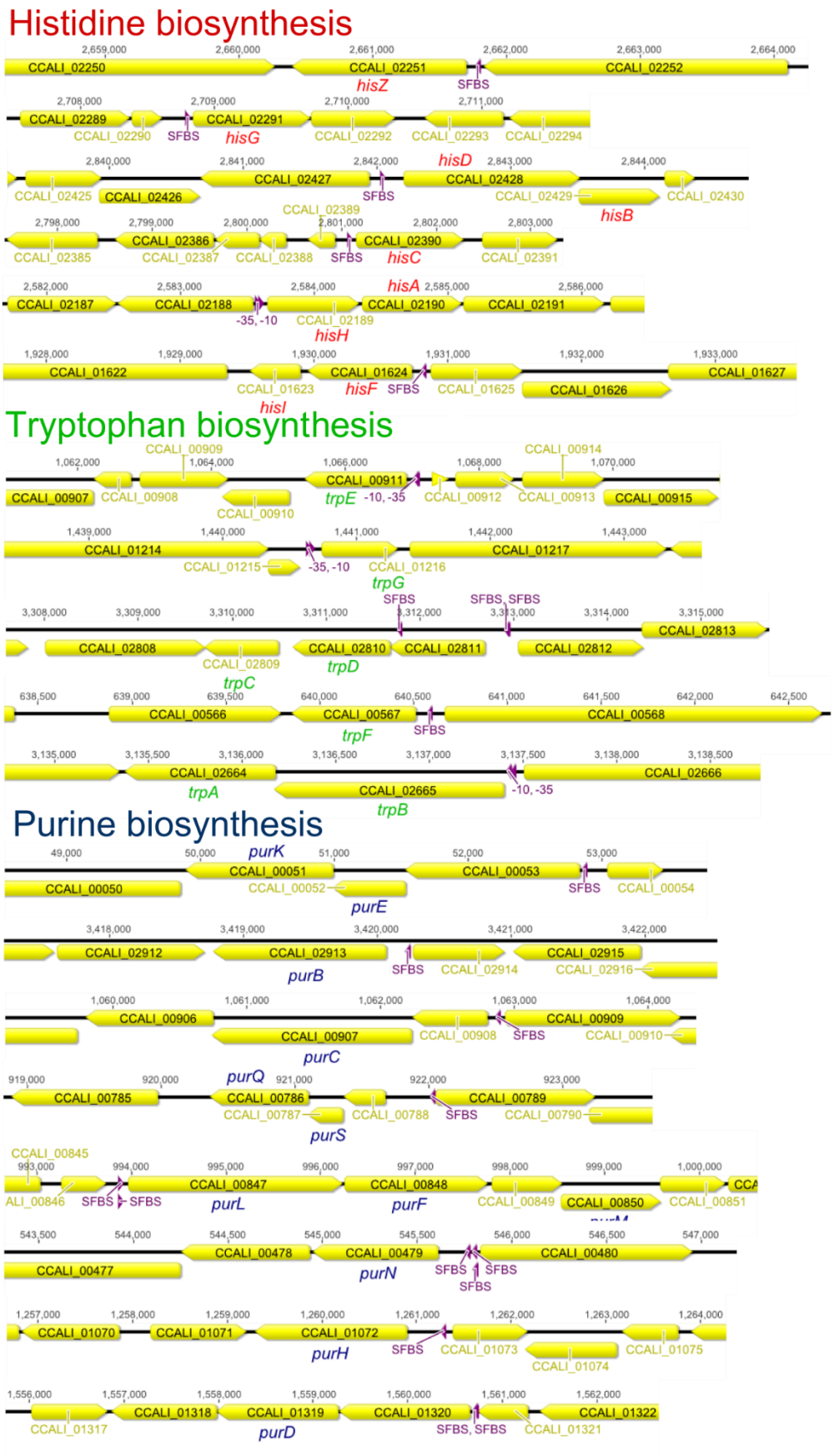
8.2.1 - Supplementary figures



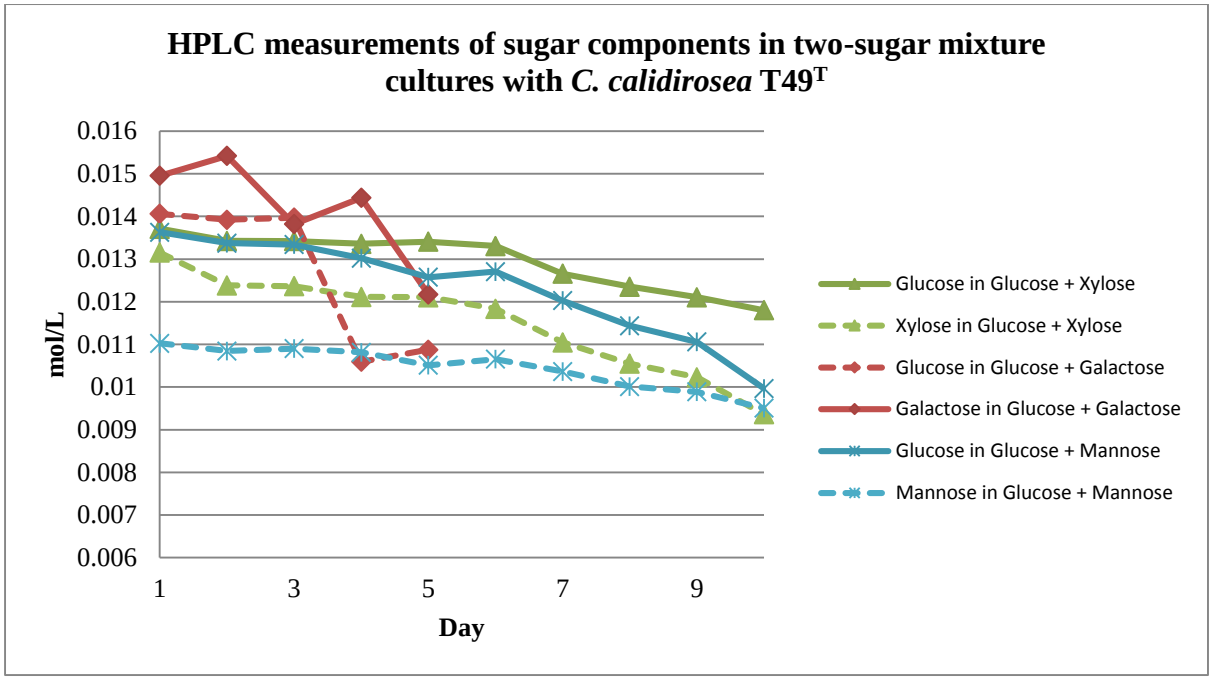
Supplementary Figure 8.1 - Phylogenetic tree showing the relationship of the *Chthonomonas calidirosea* T49^T with major lineages (phyla) within the bacterial domain. The tree was constructed using PhyloPhlAn (Segata et al., 2013) with concatenated amino acid sequences of ~400 conserved proteins and 3,737 bacterial genomes. Number within each clade indicates the number of genomes within the taxa used for analysis. Percentages along branch bifurcations indicate SH-like local support values, with only support values ≥ 70 % are shown here. Scale indicates normalised fraction of total branch length.



Supplementary Figure 8.2 - Genes involved in histidine, tryptophan, and purine biosynthesis in *E. coli* and *B. subtilis* organised in operons



Supplementary Figure 8.3 - *C. calidirosea* T49^{T} genes involved in histidine (red), tryptophan (green), and purine (blue) biosynthesis corresponding to operonic genes found in *Escherichia coli* K-12 and *Bacillus subtilis* 168. Sigma factor binding sites (SFBSs) upstream of the genes of interest were predicted using Putative sigma factor binding sites were predicted using the BPROM web server (http://www.softberry.com). Where no SFBSs were predicted, putative -35 and -10 promoter box motifs were shown.



Supplementary Figure 8.4 - HPLC measurements of sugar components in two-sugar mixture cultures with *Chthonomonas calidirosea* T49^{T}, demonstrating simultaneous uptake up both sugar species in mixtures by the bacterium.

8.2.2 - Supplementary tables

Supplementary Table 8.6 - The top BLAST search hits of the inferred *C. calidirosea* proteome to different bacterial phyla. Only the top three phyla are shown, and only hits detected above an E value cutoff of e^{-10} are considered. The entire *C. calidirosea* proteome was searched against the non-redundant NCBI database and against a more balanced database consisting of 10 *Firmicutes*, 5 *Chloroflexi* and 5 *Cyanobacteria* proteomes. The tendency for *Firmicutes* to arise as the top BLAST hits against the nr database contradicts the phylogenetic analysis, and points to an obvious bias toward *Firmicutes* in the database. This bias is removed by using a more balanced BLAST reference database.

Phylum	Percent top BLAST search results against the nr database	Percent top BLAST search results against the balanced database[1]
Chloroflexi	10.66%	49.56%
Firmicutes	27.84%	35.04%
Cyanobacteria	5.49%	15.38%

[1] The balanced BLAST database consisted of the *Firmicutes*: *Bacillus cereus* ATCC 10987 (AE017194), *Clostridium perfringens* ATCC 13124 (CP000246.1), *Enterococcus faecalis* V583 (AE016833.1), *Geobacillus thermodenitrificans* NG80-2 (CP000558.1) Staphylococcus aureus RF122 (J938182.1), *Lactobacillus acidophilus* NCFM (CP000033), *Leuconostoc citreum* KM20 (DQ489736), *Streptococcus agalactiae* 2603V/R (AE009948), *Eubacterium eligens* ATCC 27750 (CP001104); *Thermoanaerobacter ethanolicus* CCSD1 (ACXY0000000); the *Chloroflexi: Chloroflexus aggregans* DSM 9485 (CP001337), *Herpetosiphon aurantiacus* ATCC 23779 (CP000875), *Dehalococcoides ethenogenes* 195 (CP000027), *Thermobaculum terrenum* ATCC BAA-798 (CP001825, CP001826), *Ktedonobacter racemifer* SOSP1-21 (ADVG00000000); and the *Cyanobacteria*: CS-328 (ABYK00000000), *Cyanothece* sp. CCY 0110 (AAXW00000000), *Nostoc* sp. PCC 7120 (BA000019), *Prochlorococcus marinus* MIT 9312 (CP000111), *Synechococcus elongatus* PCC 6301 (AP0082314).

Supplementary Table 8.7 - tRNA genes and associated anticodons encoded by the genome of *C. calidirosea* T49^T.

Amino Acid	Number encoded	Locus tag	anticodon
Alanine (Ala)	3	CCALI_01930 CCALI_00443 CCALI_02433	CGC GGC TGC
Arginine (Arg)	5	CCALI_02538 CCALI_00398 CCALI_02744 CCALI_01802 CCALI_00578	CCG CCT GCG TCG TCT
Asparagine (Asn)	1	CCALI_01197	GTT
Aspartic acid (Asp)	1	CCALI_01657	GTC
Cysteine (Cys)	1	CCALI_02650	GCA
Glycine (Gly)	2	CCALI_01641 CCALI_02539	CTG TTG
Glutamine (Glu)	2	CCALI_01372 CCALI_02799	CTC TTC
Glycine (Gly)	3	CCALI_01107 CCALI_00577 CCALI_02651	GCC TCC CCC
Histadine (His)	1	CCALI_02698	GTG
Isoleucine (Ile)	1	CCALI_02432	GAT
Leucine (Leu)	5	CCALI_01425 CCALI_00925 CCALI_02800 CCALI_02315 CCALI_01106	CAA CAG GAG TAA TAG
Lysine (Lys)	2	CCALI_01843 CCALI_02643	CTT TTT
Methionine (Met)	3	CCALI_00185 CCALI_00543 CCALI_01864	CAT CAT CAT
Phenylalanine (Phe)	1	CCALI_02329	GAA
Proline (Pro)	3	CCALI_01687 CCALI_02302 CCALI_02470	CGG GGG TGG
Serine (Ser)	4	CCALI_00399 CCALI_02574 CCALI_00656 CCALI_00506	CGA GCT GGA TGA
Threonine (Thr)	3	CCALI_02804 CCALI_00600 CCALI_01863	CGT GGT TGT
Tryptophane (Trp)	1	CCALI_00597	CCA
Tyrosine (Tyr)	1	CCALI_00601	GTA
Valine (Val)	3	CCALI_00070 CCALI_02488 CCALI_02543	CAC GAC TAC

Supplementary Table 8.8 - Genome content similarities of *C. calidirosea* T49^T with selected species from the core eggNOG database, based on 3 similarity coefficients: Bray-Curtis, Jaccard, and Sørensen (Legendre et al., 1998). Species are sorted based on highest Bray-Curtis similarities to *C. calidirosea* T49^T.

NCBI Tax. ID	Species	Phylum-Class	Bray-Curtis	Jaccard	Sørensen
351607	Acidothermus cellulolyticus 11B	Actinobacteria	0.51	0.40	0.57
262724	Thermus thermophilus HB27	Thermus/Deinococcus	0.48	0.40	0.57
224324	Aquifex aeolicus VF5	Aquificae	0.48	0.40	0.58
197221	Thermosynechococcus elongatus BP-1	Cyanobacteria	0.47	0.36	0.53
266117	Rubrobacter xylanophilus DSM 9941	Actinobacteria	0.47	0.38	0.56
243274	Thermotoga maritima MSB8	Thermotogae	0.46	0.38	0.55
292459	Symbiobacterium thermophilum IAM 14863	Firmicutes-Clostridia	0.45	0.37	0.54
273068	Thermoanaerobacter tengcongensis MB4	Firmicutes-Clostridia	0.45	0.37	0.54
351627	Caldicellulosiruptor saccharolyticus DSM 8903	Firmicutes-Clostridia	0.45	0.38	0.55
243230	Deinococcus radiodurans R1	Thermus/Deinococcus	0.45	0.37	0.54
246194	Carboxydothermus hydrogenoformans Z-2901	Firmicutes-Clostridia	0.44	0.36	0.53
381764	Fervidobacterium nodosum Rt17-B1	Thermotogae	0.44	0.35	0.52
267747	Propionibacterium acnes KPA171202	Actinobacteria	0.44	0.35	0.52
264732	Moorella thermoacetica ATCC 39073	Firmicutes-Clostridia	0.44	0.36	0.53
235909	Geobacillus kaustophilus HTA426	Firmicutes/Bacilli	0.44	0.33	0.50
269800	Thermobifida fusca YX	Actinobacteria	0.44	0.33	0.50
370438	Pelotomaculum thermopropionicum SI	Firmicutes-Clostridia	0.43	0.35	0.51
74547	Prochlorococcus marinus str. MIT 9313	Cyanobacteria	0.43	0.32	0.49
221109	Oceanobacillus iheyensis HTE831	Firmicutes-Bacilli	0.43	0.34	0.50
391009	Thermosipho melanesiensis BI429	Thermotogae	0.43	0.35	0.52
281090	Leifsonia xyli subsp. xyli str. CTCB07	Actinobacteria	0.43	0.33	0.50
251221	Gloeobacter violaceus PCC 7421	Cyanobacteria	0.43	0.34	0.51
349161	Desulfotomaculum reducens MI-1	Firmicutes-Clostridia	0.42	0.35	0.52
383372	Roseiflexus castenholzii DSM 13941	Cyanobacteria	0.42	0.30	0.46
84588	Synechococcus sp. WH 8102	Cyanobacteria	0.42	0.31	0.48
443906	Clavibacter michiganensis subsp. michiganensis NCPPB 382	Actinobacteria	0.42	0.31	0.47
272562	Clostridium acetobutylicum ATCC 824	Firmicutes-Clostridia	0.42	0.35	0.51
335541	Syntrophomonas wolfei subsp. wolfei str. Goettingen	Firmicutes-Clostridia	0.42	0.34	0.51
266940	Kineococcus radiotolerans SRS30216	Actinobacteria	0.41	0.31	0.47
206672	Bifidobacterium longum NCC2705	Actinobacteria	0.41	0.31	0.48
224308	Bacillus subtilis subsp. subtilis str. 168	Firmicutes-Bacilli	0.40	0.31	0.47
158878	Staphylococcus aureus subsp. aureus Mu50	Firmicutes-Bacilli	0.40	0.31	0.47
203120	Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	Firmicutes-Bacilli	0.39	0.31	0.47
196164	Corynebacterium efficiens YS-314	Actinobacteria	0.39	0.30	0.47
272626	Listeria innocua Clip11262	Firmicutes-Bacilli	0.39	0.31	0.48
272623	Lactococcus lactis subsp. lactis Il1403	Firmicutes-Bacilli	0.39	0.32	0.48
190304	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	Fusobacteria	0.39	0.31	0.48
369723	Salinispora tropica CNB-440	Actinobacteria	0.38	0.31	0.47
203124	Trichodesmium erythraeum IMS101	Cyanobacteria	0.38	0.31	0.48
196162	Nocardioides sp. JS614	Actinobacteria	0.38	0.31	0.48
203123	Oenococcus oeni PSU-1	Firmicutes-Bacilli	0.38	0.29	0.45
293826	Alkaliphilus metalliredigens QYMF	Firmicutes-Clostridia	0.38	0.31	0.48
290340	Arthrobacter aurescens TC1	Actinobacteria	0.38	0.28	0.44

243164	Dehalococcoides ethenogenes 195	Chloroflexi	0.37	0.30	0.46
170187	Streptococcus pneumoniae TIGR4	Firmicutes-Bacilli	0.37	0.30	0.47
106370	*Frankia* sp. CcI3	Actinobacteria	0.37	0.27	0.43
1148	*Synechocystis* sp. PCC 6803	Cyanobacteria	0.37	0.26	0.42
220668	Lactobacillus plantarum WCFS1	Firmicutes-Bacilli	0.37	0.30	0.46
278197	Pediococcus pentosaceus ATCC 25745	Firmicutes-Bacilli	0.36	0.28	0.44
226185	Enterococcus faecalis V583	Firmicutes-Bacilli	0.36	0.29	0.45
138119	Desulfitobacterium hafniense Y51	Firmicutes-Clostridia	0.35	0.32	0.49
83332	Mycobacterium tuberculosis H37Rv	Actinobacteria	0.35	0.27	0.43
203267	Tropheryma whipplei str. Twist	Actinobacteria	0.34	0.28	0.44
240292	Anabaena variabilis ATCC 29413	Cyanobacteria	0.34	0.26	0.41
103690	*Nostoc* sp. PCC 7120	Cyanobacteria	0.33	0.25	0.40
247156	Nocardia farcinica IFM 10152	Actinobacteria	0.32	0.26	0.41
405948	Saccharopolyspora erythraea NRRL 2338	Actinobacteria	0.32	0.28	0.43
100226	Streptomyces coelicolor A3(2)	Actinobacteria	0.30	0.23	0.37
101510	Rhodococcus sp. RHA1	Actinobacteria	0.26	0.25	0.40

Supplementary Table 8.9 - Putative horizontally transferred genes identified from the genome of *Chthonomonas calidirosea* T49 using "Alien_hunter" algorithms (Vernikos & Parkhill, 2006). The genes are also characterised by their KEGG (Kanehisa et al., 2012) and COG (Tatusov et al., 2003) categories.

The table is too large for printed form. Please refer to digital supplementary files attached to this thesis with the name:
Chapter 5 – Supplementary Table 5.4.xls

Supplementary Table 8.10 - Bray-Curtis similarity coefficients from comparison of the genome contents of *Chthonomonas calidirosea* T49 ^{T} and selected bacteria from the core eggNOG database (Muller et al., 2010). For all cells except those comparing *C. calidirosea* T49^{T}, data are averages of multiple species in a particular phylum or class. The numbers in parentheses indicate the number of species in each group. The diagonal (bold numbers) shows average within-group similarities.

	Actinobacteria	**Cyanobacteria**	**Bacilli**	**Clostridia**	**Thermotogae**	**Thermus**
Actinobacteria (19)	0.453					
Cyanobacteria (9)	0.332	0.481				
Firmicutes- Bacilli (12)	0.376	0.320	0.536			
Firmicutes-Clostridia (11)	0.355	0.342	0.409	0.529		
Thermotogae (3)	0.337	0.324	0.393	0.454	0.762	
Thermus (2)	0.412	0.395	0.421	0.428	0.442	0.595
C. calidirosea T49^{T}	0.389	0.398	0.390	0.424	0.445	0.465

Supplementary Table 8.11 - Sigma factors (σ) identified from the genome of *Chthonomonas calidirosea* T49^{T}.

	Locus_tag	Amino acid sequence length	Predicted homologs
σ-70 family			
Group 1 (Primary σ)	CCALI_00640	403aa	rpoD
Group 2	CCALI_00566	304aa	rpoD
	CCALI_01294	363aa	rpoD
	CCALI_01259	348aa	rpoD
Group 3	CCALI_02523	437aa	rpoE?
	CCALI_02370	228aa	sigH[1]
	CCALI_00991	256aa	sigD/fliA/whiG
	CCALI_00758	253aa	sigD/fliA/whiG
Group 4 Extracytoplasmic function (ECF)	CCALI_02009	193aa	sigE/rpoE[2]
	CCALI_02186	186aa	sigE/rpoE[2]
	CCALI_02236	209aa	sigE/rpoE[2]
	CCALI_02281	212aa	sigE/rpoE[2]
	CCALI_02293	195aa	?
	CCALI_02297	195aa	sigE/rpoE[2]
	CCALI_02356	189aa	?
	CCALI_01025	243aa	sigE/rpoE[2]
	CCALI_01093	235aa	sigE/rpoE[2]
	CCALI_01145	248aa	?
	CCALI_01207	204aa	sigE/rpoE[2]
	CCALI_01237	179aa	?
	CCALI_01442	253aa	No homolog found. Contains *gerE/luxR* domain
	CCALI_01683	188aa	sigE/rpoE[2]
	CCALI_01706	201aa	sigE/rpoE[2]
	CCALI_01897	195aa	sigE/rpoE[2]
	CCALI_00120	186aa	sigE/rpoE[2]
	CCALI_00471	240aa	sigE/rpoE[2]
	CCALI_02721	245aa	sigE/rpoE[2]
	CCALI_02375	242aa	?
	CCALI_02556	135aa	?
	CCALI_00103	219aa	?
σ -54 family			
	CCALI_01002	464aa	rpoN/sigL

[1] Phylogenetic position of *sigH* between Group 3 and Group 4 is currently unclear (Schmid et al., 2012), [2] Group 4 sigmas are highly diverse, *sigE* prediction was based on BLASTP search with the KEGG ortholog database (Kanehisa et al., 2012).

Supplementary Table 8.12 - ATP-binding cassette (ABC) transporters and major facilitator superfamily (MFS) transporters identified in *Chthonomonas calidirosea* T49^{T} genome. Genes related to carbohydrate transporter, based on COG classification (Tatusov et al., 2003), were highlighted in bold.

ATP-binding cassette (ABC) transporters

Locus_tag	Description
CCALI_00079	carbohydrate ABC transporter membrane protein 2, CUT1 family (TC 3.A.1.1.-)
CCALI_00081	carbohydrate ABC transporter membrane protein 1, CUT1 family (TC 3.A.1.1.-)
CCALI_00082	carbohydrate ABC transporter substrate-binding protein, CUT1 family (TC 3.A.1.1.-)
CCALI_00092	nucleoside ABC transporter membrane protein
CCALI_00093	nucleoside ABC transporter membrane protein
CCALI_00094	nucleoside ABC transporter ATP-binding protein (EC:3.6.3.17)
CCALI_00153	ABC-type Fe3+-hydroxamate transport system, periplasmic component
CCALI_00169	oligopeptide/dipeptide ABC transporter, ATP-binding protein, C-terminal domain
CCALI_00203	monosaccharide ABC transporter membrane protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00204	monosaccharide ABC transporter substrate-binding protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00257	ABC-type metal ion transport system, ATPase component
CCALI_00304	nucleoside ABC transporter membrane protein
CCALI_00305	nucleoside ABC transporter membrane protein
CCALI_00306	nucleoside ABC transporter ATP-binding protein (EC:3.6.3.17)
CCALI_00405	ABC-type sugar transport systems, permease components
CCALI_00408	ABC-type sugar transport system, permease component
CCALI_00418	amino acid/amide ABC transporter substrate-binding protein, HAAT family (TC 3.A.1.4.-)
CCALI_00419	amino acid/amide ABC transporter membrane protein 1, HAAT family (TC 3.A.1.4.-)
CCALI_00420	ABC-type branched-chain amino acid transport systems, ATPase component
CCALI_00421	amino acid/amide ABC transporter ATP-binding protein 2, HAAT family (TC 3.A.1.4.-)
CCALI_00503	ABC-2 family transporter protein.
CCALI_00504	ABC-type multidrug transport system, ATPase component
CCALI_00531	ABC-type cobalamin/Fe3+-siderophores transport systems, ATPase components (EC:3.6.3.34)
CCALI_00551	ABC-type polysaccharide/polyol phosphate transport system, ATPase component
CCALI_00552	ABC-type polysaccharide/polyol phosphate export systems, permease component
CCALI_00561	ABC-type transport system involved in cytochrome c biogenesis, permease component
CCALI_00562	ABC-type transport system involved in cytochrome c biogenesis, permease component
CCALI_00569	ABC-type antimicrobial peptide transport system, permease component
CCALI_00607	Iron-regulated ABC transporter permease protein SufD
CCALI_00608	Iron-regulated ABC transporter ATPase subunit SufC
CCALI_00609	Iron-regulated ABC transporter membrane component SufB
CCALI_00617	monosaccharide ABC transporter membrane protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00618	monosaccharide ABC transporter membrane protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00629	monosaccharide ABC transporter membrane protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00744	ABC-type antimicrobial peptide transport system, ATPase component
CCALI_00745	ABC-type transport system, involved in lipoprotein release, permease component
CCALI_00750	ABC-type antimicrobial peptide transport system, permease component
CCALI_00765	ABC-type Fe3+-hydroxamate transport system, periplasmic component
CCALI_00766	ABC-type multidrug transport system, ATPase component

CCALI_00773	monosaccharide ABC transporter membrane protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00774	monosaccharide ABC transporter substrate-binding protein, CUT2 family (TC 3.A.1.2.-)
CCALI_00820	ABC-type multidrug transport system, ATPase and permease components
CCALI_00827	ABC-type transport system involved in multi-copper enzyme maturation, permease component
CCALI_00828	ABC-type multidrug transport system, ATPase component
CCALI_00829	ABC-2 family transporter protein.
CCALI_00921	Excinuclease ABC subunit B
CCALI_00942	carbohydrate ABC transporter ATP-binding protein, CUT1 family (TC 3.A.1.1.-)
CCALI_00943	ABC-type antimicrobial peptide transport system, ATPase component
CCALI_01041	ABC-type uncharacterized transport system, duplicated ATPase component
CCALI_01042	oligopeptide/dipeptide ABC transporter, ATP-binding protein, C-terminal domain
CCALI_01043	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
CCALI_01044	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
CCALI_01045	ABC-type oligopeptide transport system, periplasmic component
CCALI_01060	ABC-type multidrug transport system, ATPase and permease components (EC:3.6.3.-)
CCALI_01061	ABC-type multidrug transport system, ATPase and permease components (EC:3.6.3.-)
CCALI_01101	ABC-type sugar transport system, ATPase component (EC:3.6.3.17)
CCALI_01265	ABC-type polysaccharide/polyol phosphate export systems, permease component
CCALI_01272	oligopeptide/dipeptide ABC transporter, ATP-binding protein, C-terminal domain
CCALI_01279	ABC-type dipeptide transport system, periplasmic component
CCALI_01280	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
CCALI_01281	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
CCALI_01322	ABC-type transport system involved in resistance to organic solvents, periplasmic component
CCALI_01323	ABC-type transport system involved in resistance to organic solvents, ATPase component
CCALI_01324	ABC-type transport system involved in resistance to organic solvents, permease component
CCALI_01377	ABC-type multidrug transport system, ATPase and permease components
CCALI_01418	monosaccharide ABC transporter substrate-binding protein, CUT2 family (TC 3.A.1.2.-)
CCALI_01461	ABC-type antimicrobial peptide transport system, permease component
CCALI_01463	ABC-type antimicrobial peptide transport system, ATPase component
CCALI_01499	Excinuclease ABC subunit C
CCALI_01508	ABC-type antimicrobial peptide transport system, permease component
CCALI_01547	cobalt ABC transporter, permease protein CbiQ
CCALI_01548	ABC-type cobalt transport system, ATPase component
CCALI_01647	ABC-type metal ion transport system, ATPase component
CCALI_01744	monosaccharide ABC transporter substrate-binding protein, CUT2 family (TC 3.A.1.2.-)
CCALI_01745	ribose ABC transporter membrane protein
CCALI_01782	ABC-type multidrug transport system, ATPase and permease components
CCALI_01788	ABC-type transport system involved in multi-copper enzyme maturation, permease component
CCALI_01959	ABC-type (unclassified) transport system, ATPase component (EC:3.6.3.-)
CCALI_01971	ABC-type Fe3+-siderophore transport system, permease component
CCALI_01972	ABC-type Fe3+-hydroxamate transport system, periplasmic component
CCALI_02053	ABC-2 family transporter protein.
CCALI_02120	excinuclease ABC, A subunit

CCALI_02227	ABC-type multidrug transport system, ATPase and permease components (EC:3.6.3.-)
CCALI_02240	ABC-type uncharacterized transport system, permease component
CCALI_02417	ABC-type polysaccharide/polyol phosphate transport system, ATPase component
CCALI_02439	ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic components
CCALI_02440	ABC-type nitrate/sulfonate/bicarbonate transport system, ATPase component
CCALI_02442	ABC-type nitrate/sulfonate/bicarbonate transport system, permease component
CCALI_02575	Predicted soluble lytic transglycosylase fused to an ABC-type amino acid-binding protein
CCALI_02591	ABC-type uncharacterized transport system, periplasmic component
CCALI_02695	ABC-type multidrug transport system, ATPase component
CCALI_02700	ABC-type antimicrobial peptide transport system, ATPase component
CCALI_02701	ABC-type antimicrobial peptide transport system, permease component
CCALI_02709	ATPase components of ABC transporters with duplicated ATPase domains
CCALI_02795	phosphate ABC transporter substrate-binding protein, PhoT family (TC 3.A.1.7.1)
CCALI_02796	phosphate ABC transporter, permease protein PstA/phosphate ABC transporter, permease protein PstC
CCALI_02798	phosphate ABC transporter ATP-binding protein, PhoT family (TC 3.A.1.7.1) (EC:3.6.3.27)
CCALI_02822	ABC-type transport system, involved in lipoprotein release, permease component
CCALI_02902	heme ABC exporter, ATP-binding protein CcmA (EC:3.6.3.41)

Major facilitator superfamily (MFS) transporters

Locus_tag	Description
CCALI_00053	Arabinose efflux permease
CCALI_00147	Arabinose efflux permease
CCALI_00324	Sugar phosphate permease
CCALI_00325	drug resistance transporter, EmrB/QacA subfamily
CCALI_00540	Arabinose efflux permease
CCALI_01379	FOG: HEAT repeat
CCALI_01627	Major Facilitator Superfamily.
CCALI_02072	Major Facilitator Superfamily.
CCALI_02323	Arabinose efflux permease
CCALI_02436	Arabinose efflux permease
CCALI_01089	MFS/sugar transport protein

Supplementary Table 8.13 - Gene loci and associated characteristics of carbohydrate active enzymes (CAEs) encoded in the genome of *Chthonomonas calidirosea* T49^{T}. Genes encoding CAE belonging to CAZy families (Cantarel et al., 2009) known for glycosyl hydrolase activity against the primary plant-based polysaccharides have been highlighted and include xylanases (yellow), cellulases, xyloglucanases and β-mannanases (orange) and pectinases (green). The predicted functions and general activities of cellulose esterases (CE), glycoside hydrolases (GH), carbohydrate-binding motifs (CBM) and pectin lyases (PL) have been defined along with predicted signal peptides, transmembrane helices and localisation. Putative glycoside hydrolases and pectin lyases appear at the bottom of the table with no predicted domain or activity.

Locus tag	Domains[1]	Known or predicted function[1]	Transmembrane helices[2]	Signal peptide[3]	Localisation[4]	E.C. number[1]	Activity-type classification[1]
CCALI_00319	CE4	xylanase/chitin deacetylase	yes	no	C	3.1.1.72 / 3.5.1.-	cellulose esterase
CCALI_02646	CE4	xylanase/chitin deacetylase	yes	no	C	3.1.1.72 / 3.5.1.-	cellulose esterase
CCALI_00310	CE4	xylanase/chitin deacetylase	yes	no	C	3.1.1.72 / 3.5.1.-	cellulose esterase
CCALI_01866	CE4	xylanase/chitin deacetylase	no	no	C	3.1.1.72 / 3.5.1.-	cellulose esterase
CCALI_00308	CE4	xylanase/chitin deacetylase	no	no	CM	3.1.1.72 / 3.5.1.-	cellulose esterase
CCALI_00469	CE4	uncharacterized protein conserved in bacteria	no	no	C	3.1.1.72 / 3.5.1.-	cellulose esterase
CCALI_00541	CE7	acetyl xylan esterase (AxeI)	yes	yes	U	3.1.1.72	cellulose esterase
CCALI_01582	GH2	β-galactosidase	yes	no	CM	3.2.1.23	β-glucosidase
CCALI_01726	GH2	β-galactosidase	no	yes	CM	3.2.1.23	β-glucosidase
CCALI_02561	GH2	β-galactosidase	no	yes	U	3.2.1.23	β-glucosidase
CCALI_00677	GH2	β-galactosidase/β-glucuronidase	no	no	C	3.2.1.23	β-glucosidase
CCALI_00372	GH2	β--mannosidase	no	no	C	3.2.1.25	exoglucanase (db)
CCALI_00761	GH3-CBM6	β-glucosidase-related glycosidase	no	no	PP	3.2.1.21	β-glucosidase
CCALI_01617	GH5	cellulase	no	no	U	n.d.	n.d.
CCALI_00143	GH5	cellulase	no	no	U	n.d.	n.d.
CCALI_02737	GH5	glycosyl hydrolase	yes	yes	U	n.d.	n.d.
CCALI_02246	GH5	endoglucanase/endomannanase	no	no	U	3.2.1.4	endoglucanase
CCALI_02041	GH5	endoglucanase	yes	no	U	3.2.1.4	endoglucanase
CCALI_02491	GH6	glycosyl hydrolase	no	no	U	n.d.	n.d.
CCALI_01336	GH10	β-1,4-xylanase	no	yes	U	3.2.1.8	endoglucanase
CCALI_02203	GH10	β-1,4-xylanase	no	yes	U	3.2.1.8	endoglucanase
CCALI_00619	GH10	β-1,4-xylanase	no	no	C	3.2.1.8	endoglucanase

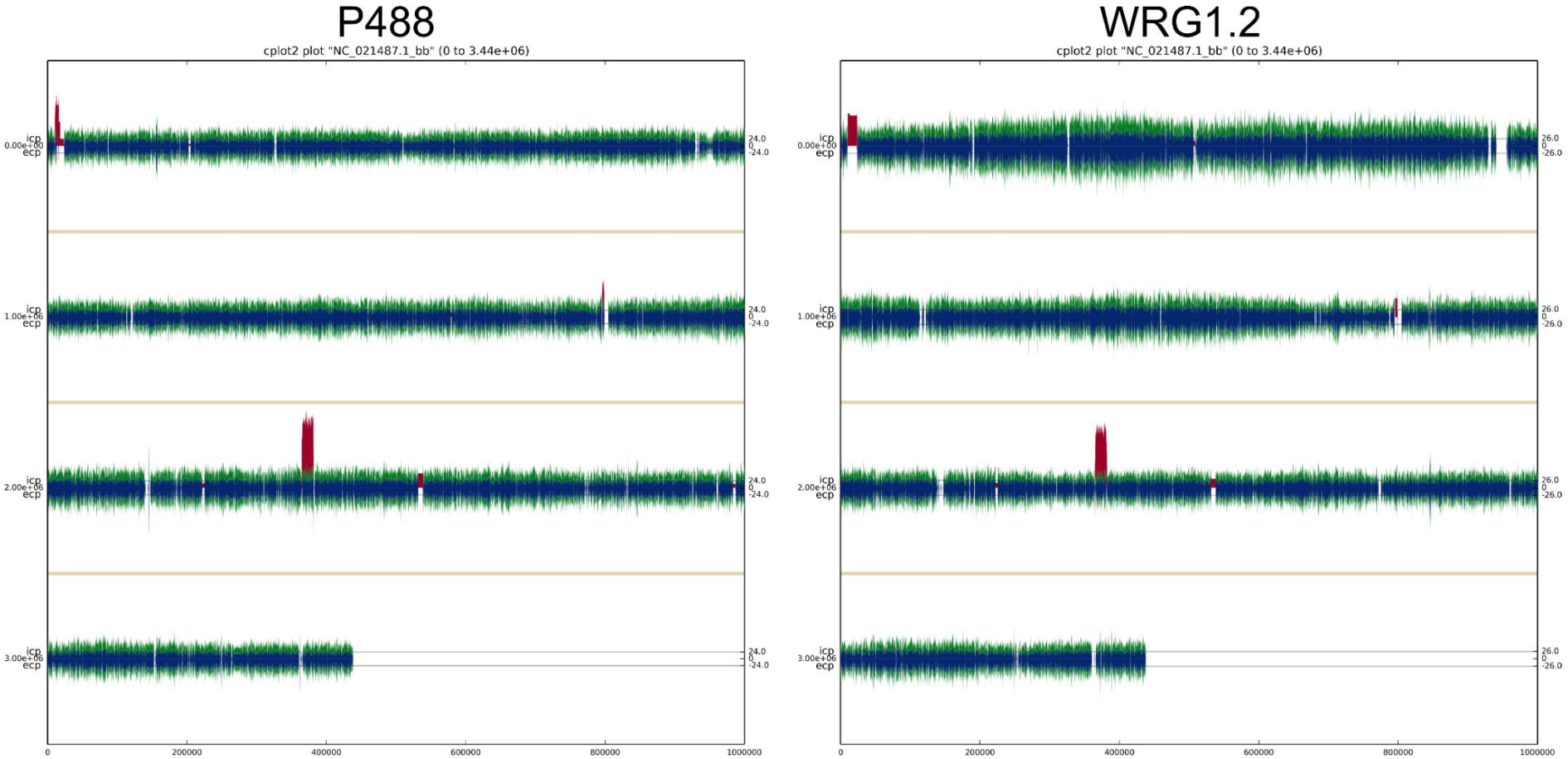
CCALI_00256	GH13	glycogen debranching enzyme/α-1,6-glucosidase	no	no	U	3.2.1.25	exoglucanase (db)
CCALI_01763	GH13-CBM48	α-amylase	no	yes	C	3.2.1.1	exoglucanase (db)
CCALI_02225	GH13-DUF1939	α-amylase	yes	no	CM	3.2.1.1	exoglucanase (db)
CCALI_01466	GH13-GH27	α-galactosidase	no	no	U	3.2.1.22	β-glucosidase
CCALI_01571	GH13-GH27-unknown CBM	α-galactosidase (melibiase)	no	no	U	3.2.1.22	β-glucosidase
CCALI_01939	GH14	β-amylase	no	no	U	3.2.1.2	exoglucanase (db)
CCALI_02392	GH15	glucoamylase	no	no	U	3.2.1.3	exoglucanase (db)
CCALI_01467	GH16	β-glucanase	yes	no	EC	n.d.	endoglucanase
CCALI_00557	GH20	β-N-acetylhexosaminidase	no	no	CM	3.2.1.-	n.d.
CCALI_01398	GH25	muramidase	no	no	U	3.2.1.17	exoglucanase
CCALI_02545	GH28/PL3	endopolygalacturonase	no	no	EC	3.2.1.15	endoglucanase
CCALI_02393	GH28/PL3	endopolygalacturonase	no	no	OM	3.2.1.15	endoglucanase
CCALI_01938	GH029	α-L-fucosidase	no	no	U	3.2.1.51	β-glucosidase
CCALI_00611	GH29	α-L-fucosidase	no	no	C	3.2.1.51	β-glucosidase
CCALI_00327	GH29-CBM6	α-L-fucosidase	no	yes	C	3.2.1.51	β-glucosidase
CCALI_02596	GH30	O-glycosyl hydrolase	no	no	CM	n.d.	β-glucosidase
CCALI_01285	GH35-GH2-CBM2	β-galactosidase	yes	yes	U	3.2.1.23	β-glucosidase
CCALI_01759	GH38	α -mannosidase	no	no	C	3.2.1.24	β-glucosidase
CCALI_02454	GH38	α-mannosidase	no	no	C	3.2.1.24	β-glucosidase
CCALI_00612	GH38	α-mannosidase	no	no	C	3.2.1.24	β-glucosidase
CCALI_02706	GH38	α-mannosidase	yes	no	CM	3.2.1.24	β-glucosidase
CCALI_01195	GH42	β-galactosidase	no	no	C	3.2.1.23	β-glucosidase
CCALI_02040	GH42	β-galactosidase	no	yes	CM	3.2.1.23	β-glucosidase
CCALI_02314	GH42	β-galactosidase	no	no	C	3.2.1.23	β-glucosidase
CCALI_00681	GH44-CBM2	glycoside hydrolase	yes	no	U	3.2.1.4/3.2.1.151	endoglucanase
CCALI_01373	GH51	α-L-arabinofuranosidase	no	no	C	3.2.1.55	exoglucanase
CCALI_00207	GH51	α-L-arabinofuranosidase	yes	yes	U	3.2.1.55	exoglucanase
CCALI_00907	GH51	α-L-arabinofuranosidase	no	no	C	3.2.1.55	exoglucanase
CCALI_02151	GH51	α-L-arabinofuranosidase	no	no	C	3.2.1.55	exoglucanase
CCALI_02697	GH053	endo-1,4-β-galactanase	yes	no	U	3.2.1.89	endoglucanase
CCALI_02107	GH057	α-amylase	no	no	C	3.2.1.1	exoglucanase (db)
CCALI_02121	GH057	α-amylase/α-mannosidase	no	no	C	3.2.1.1	exoglucanase (db)
CCALI_00903	GH063	mannosyl oligosaccharide glucosidase.	no	no	C	3.2.1.106	exoglucanase

CCALI_02048	GH078	α-L-rhamnosidase	no	no	C	3.2.1.40	exoglucanase (db)
CCALI_01328	GH078	α-L-rhamnosidase	no	yes	U	3.2.1.40	exoglucanase (db)
CCALI_01720	GH078	α-L-rhamnosidase	no	no	C	3.2.1.40	exoglucanase (db)
CCALI_01583	GH095	α-L-fucosidase	no	no	C	3.2.1.51	β-glucosidase
CCALI_00294	GH116	predicted bile acid β-glucosidase	no	no	C	3.2.1.45	β-glucosidase
CCALI_00696	GH116	predicted bile acid β-glucosidase	no	no	CM	3.2.1.45	β-glucosidase
CCALI_00937	GH116	predicted bile acid β-glucosidase	no	no	U	3.2.1.45	β-glucosidase
CCALI_02567	PL4	rhamnogalacturonate lyase	yes	no	EC	4.2.2.-	pectin lyase
CCALI_00556	n.d.	putative cellulase and CBM	no	yes	U	n.d.	n.d.
CCALI_02519	n.d.	putative cellobiose phophorylase (glycogen-debranching)	no	no	U	n.d.	exoglucanase (db)
CCALI_00922	n.d.	putative glycosyl hydrolase	no	yes	U	n.d.	n.d.
CCALI_00803	n.d.	putative glycosyl hydrolase	yes	yes	U	n.d.	n.d.
CCALI_02711	n.d.	putative glycosyl hydrolase	no	no	U	n.d.	n.d.
CCALI_01611	n.d.	putative trehalose & maltose hydrolases/phosphorylase	no	no	U	n.d.	n.d.
CCALI_02844	n.d.	predicted glycosylase	no	no	C	n.d.	n.d.
CCALI_01613	n.d.	putative glycosyl hydrolase	yes	no	CM	n.d.	n.d.
CCALI_01838	n.d.	putative glycosyl hydrolase	no	yes	U	n.d.	n.d.
CCALI_01844	n.d.	putative glycosyl hydrolase	no	yes	U	n.d.	n.d.
CCALI_02044	n.d.	putative glycosyl hydrolase	no	no	U	n.d.	n.d.
CCALI_02352	n.d.	putative glycosyl hydrolase	no	no	CM	n.d.	n.d.
CCALI_00734	n.d.	putative glycosyl hydrolase	no	yes	U	n.d.	n.d.
CCALI_01059	n.d.	unknown glycosyl hydrolase (DUF1680)	no	no	C	n.d.	n.d.
CCALI_00133	n.d.	putative pectin lyase	yes	yes	U	n.d.	n.d.
CCALI_01596	n.d.	putative pectin lyase	yes	no	U	n.d.	n.d.
CCALI_02549	n.d.	putative pectin lyase	yes	no	U	n.d.	n.d.
CCALI_00184	n.d.	putative pectin lyase	no	yes	U	n.d.	n.d.

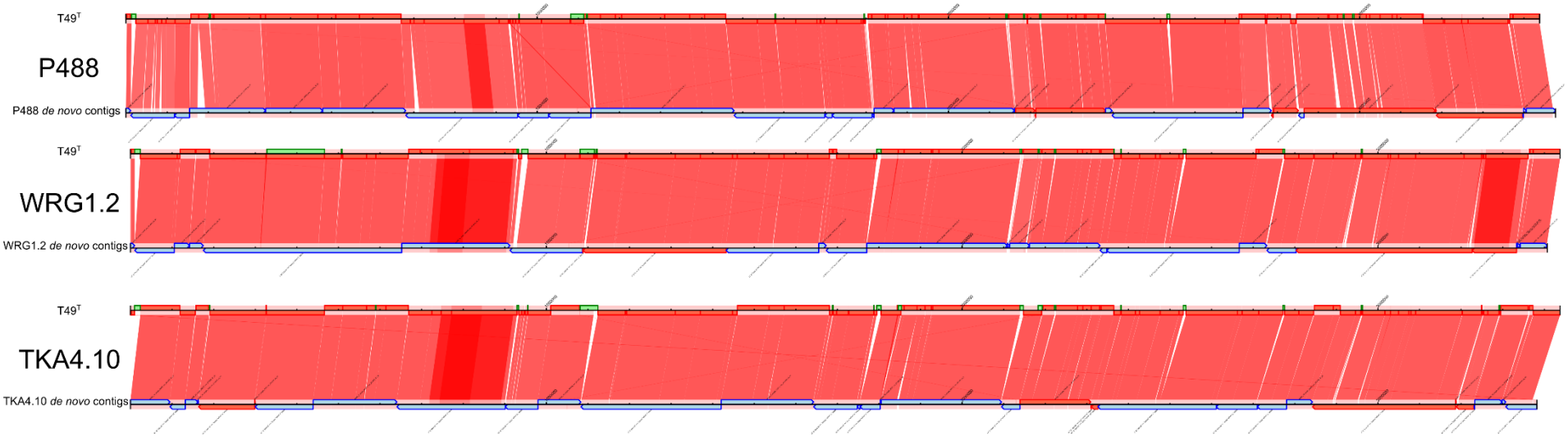
[1] the domains, predicted functions, activity-type classification and E.C. numbers were defined using a combination of the IMG/ER platform (Markowitz et al., 2010), the CAZy (Cantarel et al., 2009), MetaCyc (Caspi et al., 2012) databases, and the CAT analysis toolkit (Park et al., 2010). [2] determined using TMHMM v2.0 tool (Krogh et al., 2001) [3] determined using SignalP v4.0 tool (Petersen et al., 2011), [4] determined using PSORTb v3.0 tool (Yu et al., 2010). U: unknown, C: cytoplasm, CM: cytoplasmic membrane, PP: periplasm, EC: extracellular, n.d.: not determined, DUF: domain of unknown function, GH: glycoside hydrolase family, CE: cellulose esterase family, PL: pectin lyase family, CBM: cellulose-binding motif.

8.3 - Supplementary materials for Chapter 6

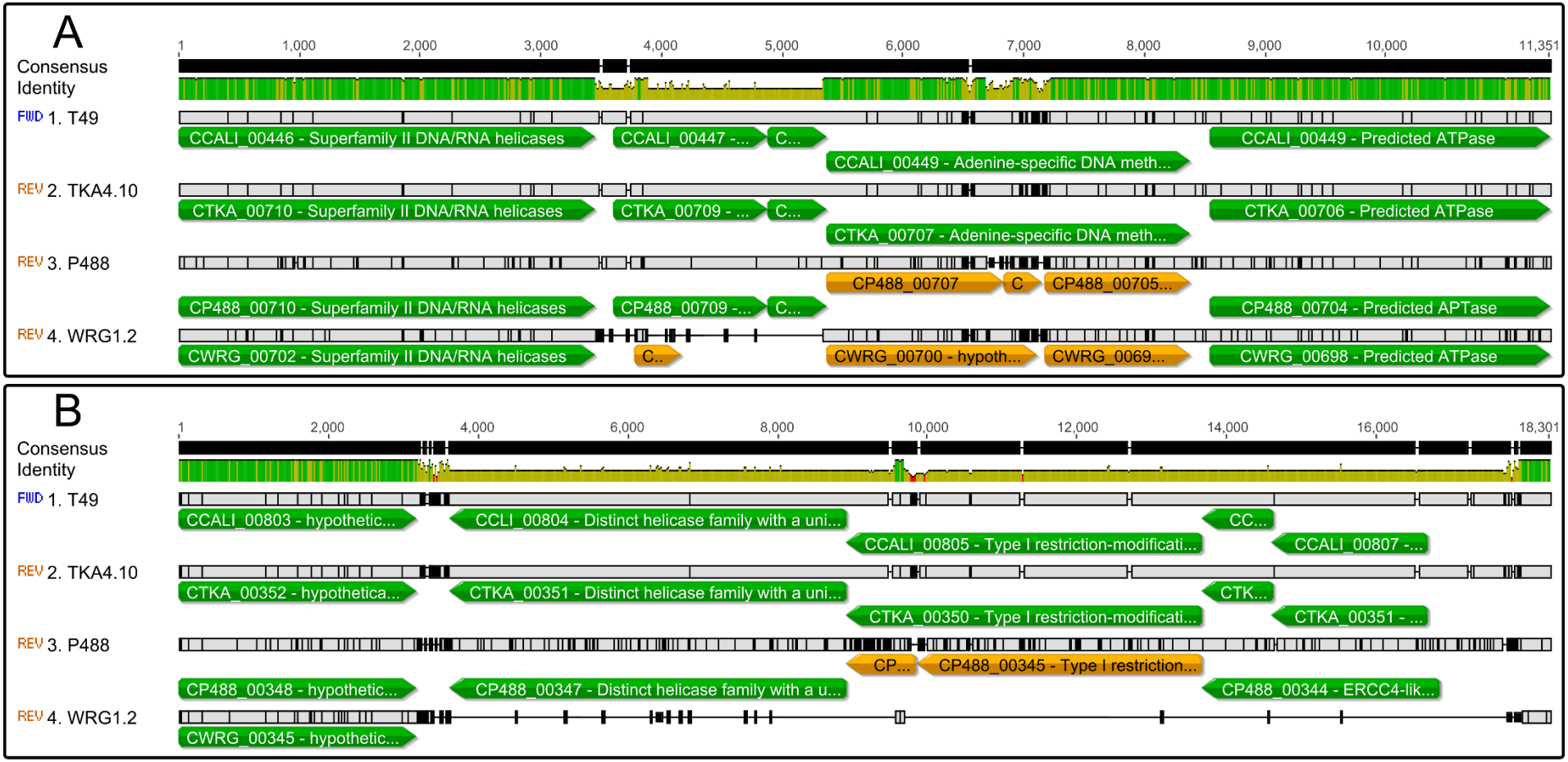
8.3.1 - Supplementary figures



Supplementary Figure 8.5 - Quality of paired-end genome assemblies. Genome coverage plot generated with Hagfish (https://github.com/mfiers/hagfish). Inclusive coverage and exclusive coverage histogram plots (above and below the centreline respectively) indicate the number of P488 or WRG1.2 paired-end reads mapped onto the reference T49^{T} genome. The inclusive coverage histogram includes the sequence between two paired-end reads, while the exclusive coverage histogram only covers the area of the actual pair-end reads. Green bars indicate read-pairs distance within expected range, blue bars indicate read-pairs shorter than expected paired-end distance, and red bars indicate read-pairs longer than expected distance. Regions with no pair-end data are shown without any peaks. Paired end violation of insertion size may be misassembled due to repetitive regions, regions of sequence not found in the reference genome, or a rearrangement breakpoint. Overall, the plots showed read pairs within expected insertion size range (green peaks). TKA assembly was not analysed because the genomes was sequenced with the Ion Torrent platform and therefore not pair-ended.



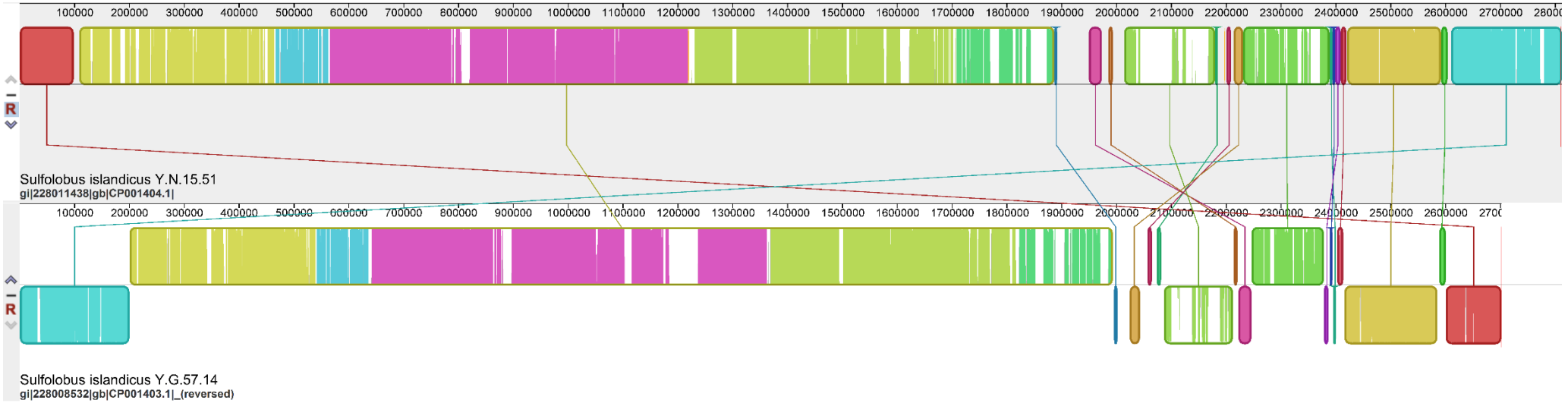
Supplementary Figure 8.6 - Synteny of *C. calidirosea* genome assemblies. P488, WRG1.2, TKA4.10 de novo assembled contigs were mapped against T49^T genome consensus sequence. The contigs the three isolates showed high degrees of synteny in all three comparisons, with no indication of genome rearrangement breakpoints. Most of the putative insertions/deletions were located at the edge of the contigs, suggesting assembly artefacts due to gaps in the reference genome or repetitive sequences. The three large putative insertions found only in P488 contained genes found in T49^T (e.g., hydroxypyruvate isomerase and prepilin protein), which may indicate assembly error or potential gene duplication sites.



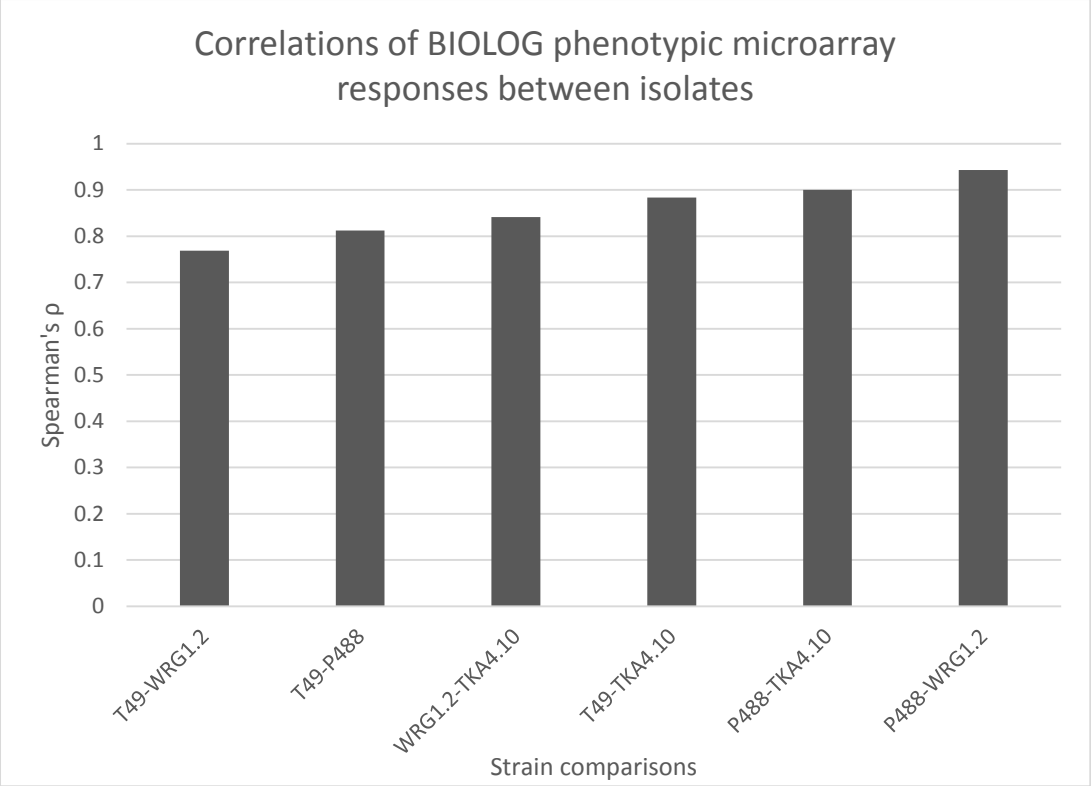
Supplementary Figure 8.7 - Alignments of putative HGT regions. Multiple MUSCLE (Edgar, 2004) alignments of putative HGT regions of *C. calidirosea* isolates (T49^T, TKA4.10, P488, WRG1.2) genomes, in relation to T49^T locus tags: (**A**) CCALI_00447-00449 (**B**) CCALI_00804-00807. Yellow segments indicate likely non-functional pseudogenes in P488 and WRG1.2 genome affected by deletion (putative DNA methylase, repair protein, and helicase). Black vertical boxes indicate regions of disagreement, and horizontal line indicate gaps in alignments. Note the high sequence agreement between T49^T and TKA4.10 and disagreement between T49^T and TKA4.10 with P488. Apart from the gaps, P488 shared similar nucleotide variation with WRG1.2.

Supplementary Figure 8.8 - progressiveMauve (Darling et al., 2010) genome alignment of *Chthonomonas calidirosea* isolates showing a high degree of synteny among the genomes. The green line shows the location of a putative HGT region examined in Supplementary Figure 8.7 (CCALI_00803 – CCALI_00807). Similarity plot within each Locally Collinear Block (LCB) indicates the sequence similarity among the genomes. The genomic sequences were assembled with MIRA (Chevreux et al., 1999) using T49^{T} genome as reference.

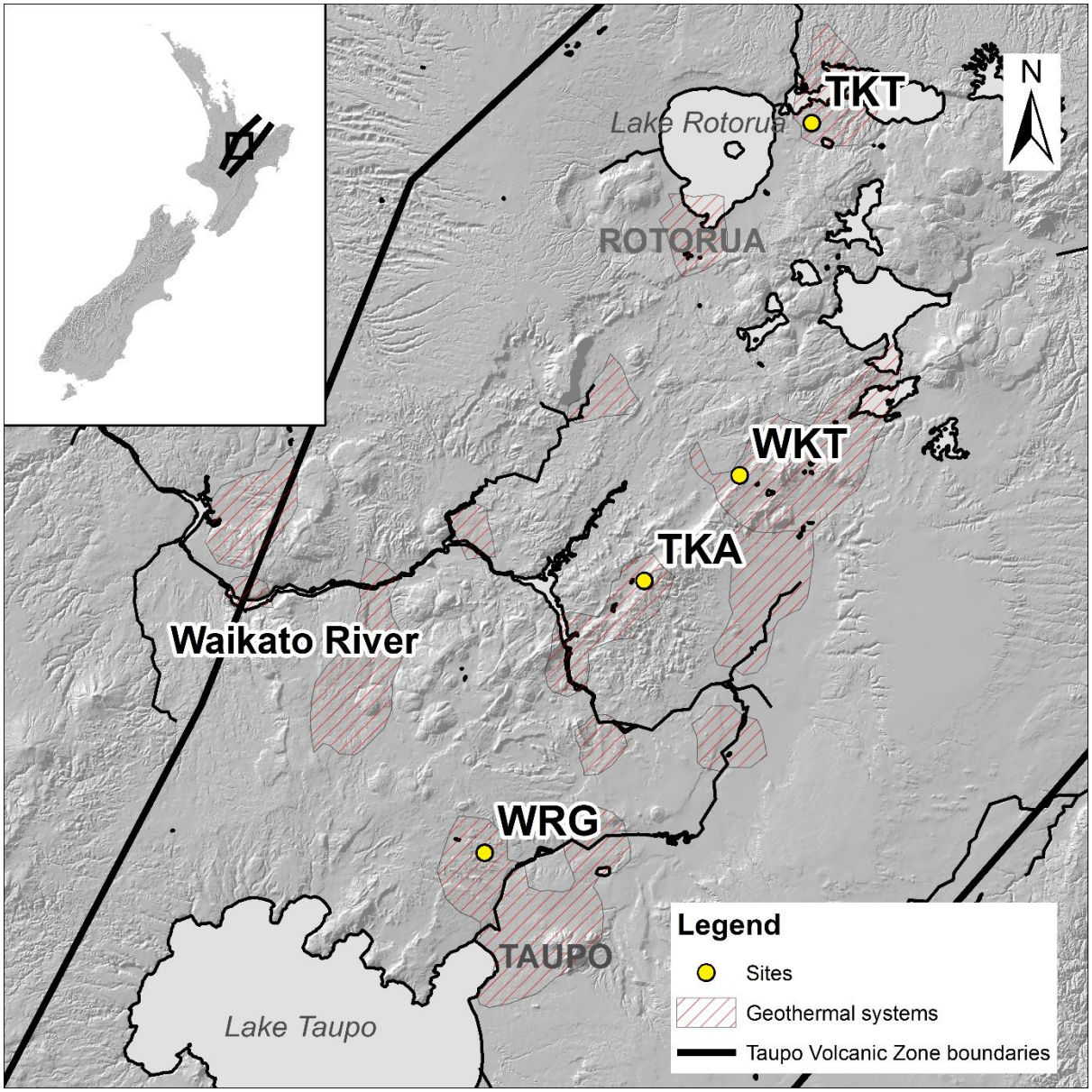
Supplementary Figure 8.9 - progressiveMauve (Darling et al., 2010) genome alignment of three *Thermus thermophilus* isolates. Strains HB-8 and HB-27 were isolated from the Izu Peninsula, Japan (Henne et al., 2004; Oshima et al., 1975), while strain JL-18 was isolated from Nevada, USA (Costa et al., 2009). The alignment shows the sites of rearrangements among the genomes. Note the position of the red Locally Collinear Block (LCB) in relation to other LCBs. The position of the LCBs above or below the median line for each genome indicates the relative sequence direction compared to the reference genome (HB-8). Similarity plot within each LCB indicates the sequence similarity among the genomes.



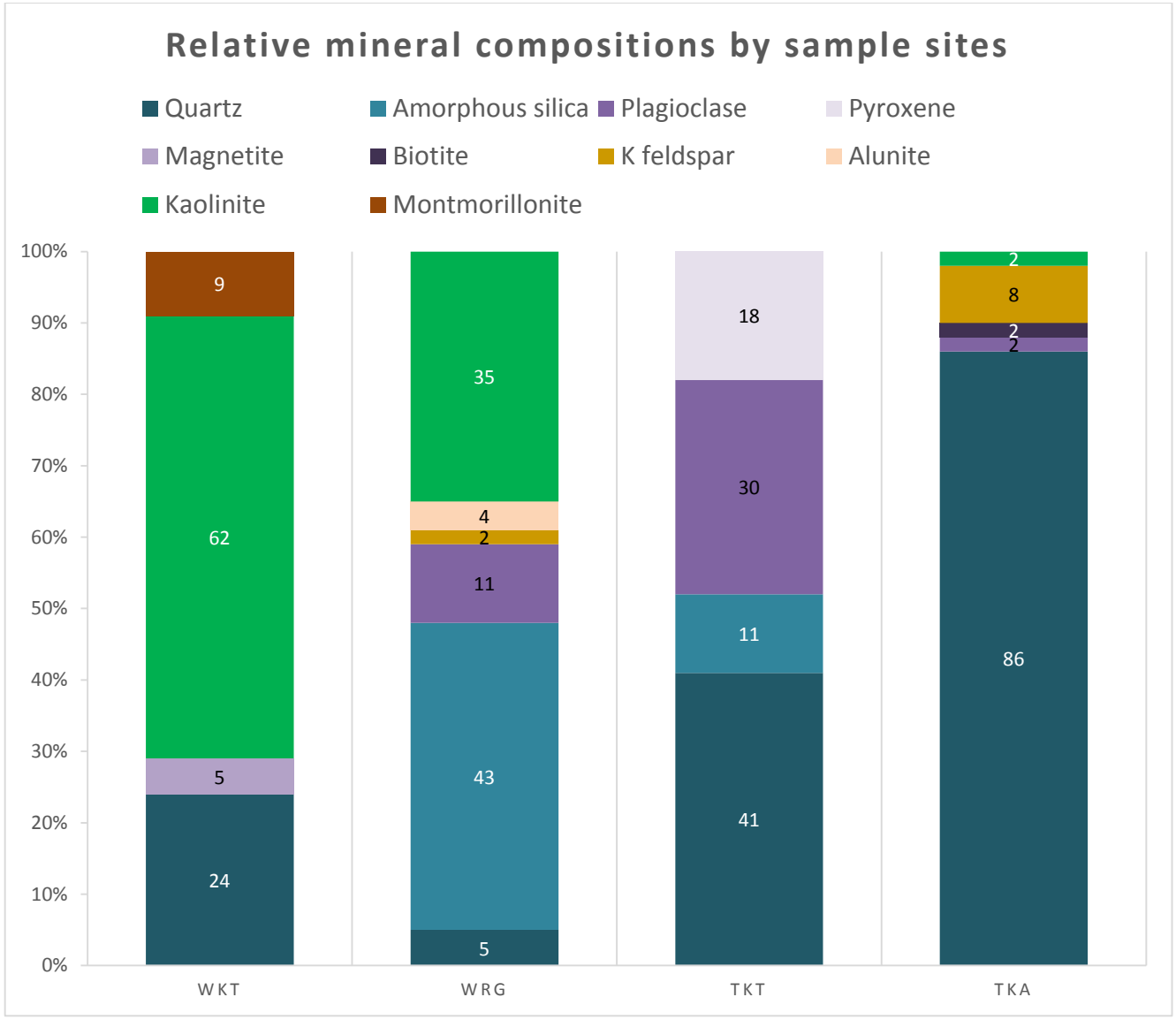
Supplementary Figure 8.10 - progressiveMauve (Darling et al., 2010) genome alignment of two *Sulfolobus islandicus* strains isolated from Yellowstone National Park, USA (Reno et al., 2009) showing regions of genome rearrangement and areas of low sequence similarity within the Locally Collinear Blocks (LCBs).



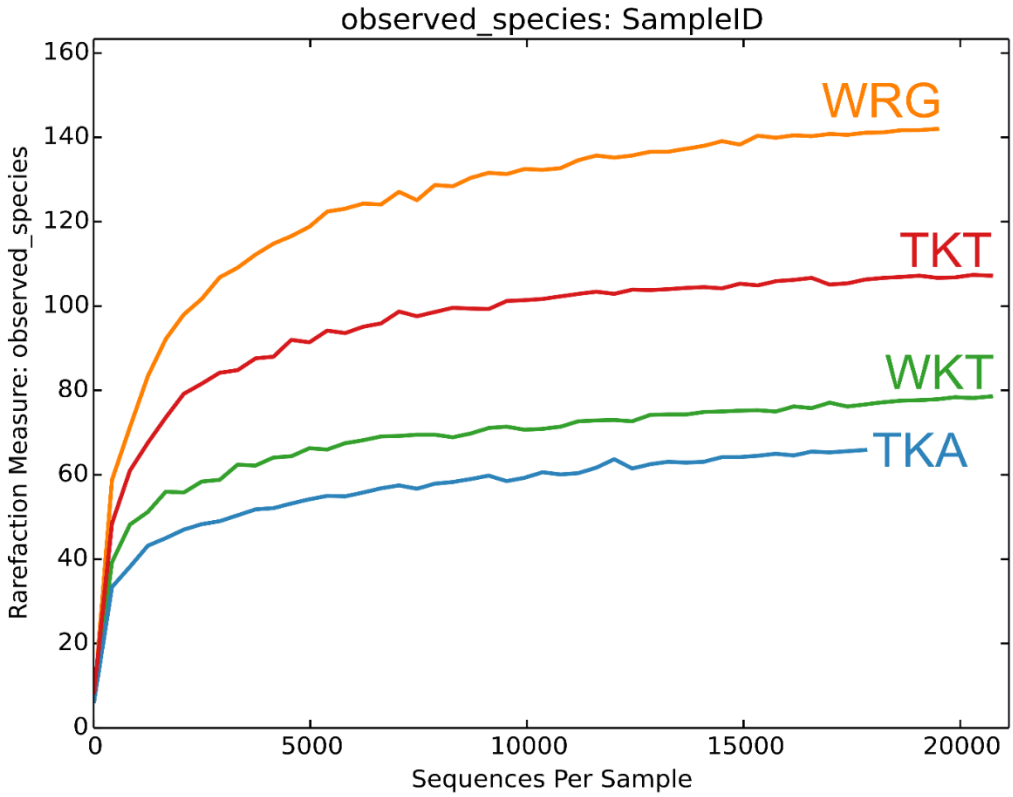
Supplementary Figure 8.11 - Correlation of BIOLOG phenotype microarray response between isolates. Tie-corrected Spearman's rho of redox dye responses to carbon sources (95 substrates including various carbohydrates, amino acids, and their derivatives) on the BIOLOG PM1 phenotype microarray plates (http://www.biolog.com/pdf/pm_lit/PM1-PM10.pdf), using means of duplicates of each isolate. Overall, the isolates demonstrated high physiological similarity in the utilization of carbohydrates. However, differences in Spearman's rank order correlation between isolates were still observed. Analysis of the range of rank-order correlations showed P488 and WRG1.2 with the highest correlation (0.94), and $T49^T$ and WRG1.2 with the lowest redox response profile correlation (0.77) in carbon source utilisation.



Supplementary Figure 8.12 - Map of the Taupō Volcanic Zone (TVZ) with GIS layers of geothermal fields, digital topography, and rivers/streams. The sample sites are indicated on the map. Wairakei – WRG, Te Kopia – TKA, Waikite – WKT, Tikitere – TKT.



Supplementary Figure 8.13 - Mineral composition of sample sites. The mineral compositions and associated compositions (percentage of total) of each sample site were determined by quantitative X-ray diffraction analysis. The mineral types have been grouped in like colours to reflect the relative effects of weathering and hydrothermal alterations. Yellow hues indicate primary minerals sensitive to hydrothermal alteration. Green hues indicate secondary clay minerals resulting from hydrothermal alteration. Blue hues indicate quartz and amorphous silica, non-clay minerals from which the effect of hydrothermal alteration could not be directly determined. The samples sites are arranged by relative abundance of clay minerals (inverse of quartz and amorphous silica abundance).



Supplementary Figure 8.14 - Observed species in the four sample sites. Rarefaction curves of observed species (as defined by OTU clustering with 97 % criterion) in the four sample site communities showing plateauing of observed species with increased numbers of sequences sampled, thus indicating a majority of targeted biodiversity were represented.

8.3.2 - Supplementary tables

Supplementary Table 8.14 - List of conserved genes used for phylogenetic inference. Information here is based on strain T49^T reference genome.

The table is too large for printed form. Please refer to digital supplementary files attached to this thesis with the name:
Chapter 6 – Supplementary Tables.xlsx → Under tab "Table S1"

Supplementary Table 8.15 - Best-Hits from reciprocal BLASTP analysis between genomes with E-value threshold 1e^{-5}, minimum identity 30 %. These tables contain all the outputs from BLASTP searches and contains hits due to variance in gene calling or hypothetical proteins with unknown putative functions. Due to the order of query versus subject, overlapped hits are present (referred using the locus tag of the query genome).

The table is too large for printed form. Please refer to digital supplementary files attached to this thesis with the name:
Chapter 6 – Supplementary Tables.xlsx → Under tab "Table S2"

Supplementary Table 8.16 – 8.16A - A list of isolate-variant genes. Presence of the gene (using T49^T locus tag as identifier) is indicated by coloured cells. Genes absent in the respective genome is marked by blank/white cells. COG functional groups and length of the T49^T homolog are also shown. Genes without annotation (i.e. hypothetical proteins) have been excluded in this table. Bolded locus tags indicate genes in proximity with putative HGT genes identified in Lee et al. (2014). **8.16B -** To assess potential bias from using T49^T reference genome as the assembly backbone. Reads that were not assembled into the isolate genomes were assembled de novo to form contigs. The contigs were analysed via BLASTN to identify genes with no close homolog to T49^T (as defined by not having a homolog from T49^T genome as the top BLASTN hit). Hits with no functional annotation and contaminants (e.g., cloning vectors) were excluded from this table.

The table is too large for printed form. Please refer to digital supplementary files attached to this thesis with the name:
Chapter 6 – Supplementary Tables.xlsx → Under tab "Table S3"

Supplementary Table 8.17 - Nucleotide divergence of concatenated conserved genes of *Thermus thermophilus* strains, JL-18, HB-27, BB-8 and ATCC33923. Top half of the distance matrix indicates base changes per 100 nucleotides. The bottom half shows base differences in the 6756 base multiple sequence alignments.

Strains	JL-18	HB-27	HB-8	ATCC 33923
JL-18		1.6134	1.5246	2.2351
HB-27	109		0.4144	1.0657
HB-8	103	28		0.9917
ATCC 33923	151	72	67	

Nucleotide divergence of concatenated conserved genes of *Sulfolobus islandicus* strains, Y.N.15.51 and Y.G.57.14. Top half of the distance matrix indicates base changes per 100 nucleotides. The bottom half shows base differences in the 4773 base sequence alignments.

Strains	Y.N.15.51	Y.G.57.14
Y.N.15.51		0.1257
Y.G.57.14	6	

NCBI BioProject Accessions:

T. thermophilus JL-18:	PRJNA66077
T. thermophilus HB-27:	PRJNA58033
T. thermophilus HB-8:	PRJNA13202
T. thermophilus ATCC 33923:	PRJNA196548
S. islandicus Y.N.15.51:	PRJNA18651
S. islandicus Y.G.57.14:	PRJNA19487

All genomes listed were downloaded from the Integrated Microbial Genome system (Markowitz et al., 2010).

Conserved genes used:

T. thermophilus - DNA primase (dnaG), ribosome recycling factor (rrf), translation initiation factor IF-3 (infC), NusA antitermination factor (nusA), phosphoglycerate kinase (pgk), CTP synthase (py**r**G)
S. islandicus – small and large subunits of DNA primase, phosphoglycerate kinase (pgk), CTP synthase (pyrG)

Supplementary Table 8.18 - Relative weight percentage of major metal oxides detected in the samples from the four sites using X-ray fluorescence (XRF). LOI = Loss on ignition, indicates weight percentage of volatile substances (e.g., carbon dioxide from carbonate and water from hydrate) escaped during strong heating. LOD = Loss on drying, indicates moisture content of the samples as measured by weight change during mild heating

SAMPLE weight %	TKT	WKT	TKA	WRG
Fe_2O_3	6.23	5.29	3.44	3.33
MnO	0.04	1.61	0.04	0.04
TiO_2	0.53	0.60	0.35	0.46
CaO	1.36	0.32	0.04	1.07
K_2O	2.62	0.06	2.75	1.15
SO_3	< 0.01	< 0.01	< 0.01	not measured
P_2O_5	0.10	0.14	0.02	0.08
SiO_2	69.08	52.97	81.09	63.92
Al_2O_3	12.41	26.31	8.84	19.52
MgO	0.58	1.39	0.30	0.26
Na_2O	2.03	0.04	0.04	1.12
LOI	4.61	9.98	2.54	8.94
SUM	99.62	98.72	99.45	99.89
LOD	28.03	59.94	37.05	28.77

Supplementary Table 8.19 - Operation taxonomic units (OTU) table along with representative sequences of the OTUs identified in this study. OTU clustering was conducted using UPARSE with 97% criterion. OTU taxonomic assignment was based on the Greengenes database. Number of reads assigned to the specific OTU is also shown.

The table is too large for printed form. Please refer to digital supplementary files attached to this thesis with the name:
Chapter 6 – Supplementary Tables.xlsx → Under tab "Table S6"

Supplementary Table 8.20 – Relative taxa abundance of the four communities containing *Chthonomonas calidirosea*. Note each entry does not correspond to a single OTU but to the closest assignable taxonomic clade, therefore a class-level entry, for example, may contain several OTUs, while OTUs with closer resemblance to reference sequences in the data may be assigned to a family-level entry.

The table is too large for printed form. Please refer to digital supplementary files attached to this thesis with the name:
Chapter 6 – Supplementary Tables.xlsx → Under tab "Table S7"

Supplementary Table 8.21 - Beta diversity of the four communities measured using Bray-Curtis, weighted and unweighted Unifrac matrices.

Bray-Curtis dissimilarity matrix of the four geothermal communities

Sites / Communities	TKT	WKT	WRG
TKT			
WKT	0.836		
WRG	0.626	0.601	
TKA	0.820	0.574	0.430

Weighted Unifrac distance matrix of the four geothermal communities

Sites / Communities	TKT	WKT	WRG
TKT			
WKT	0.373		
WRG	0.250	0.260	
TKA	0.339	0.311	0.177

Unweighted Unifrac distance matrix of the four geothermal communities

Sites / Communities	TKT	WKT	WRG
TKT			
WKT	0.489		
WRG	0.386	0.484	
TKA	0.434	0.526	0.516

Supplementary Table 8.22 - Pairwise discontiguous megaBLAST comparison of 16S rRNA gene sequences between *Armatimonadetes* OTUs identified in a Thailand hot spring (Cuecas et al., 2014) and the type strains of the three described *Armatimonadetes* classes: *Armatimonas rosea* YO-36^T (Tamaki et al., 2011), *Fimbriimonas ginsengisoli* Gsoil 348^T (Im et al., 2012)[3], and *Chthonomonas calidirosea* T49^T (Lee et al., 2011)

Sequence identity†	NR_113009.1	NR_121726.1	NR_103954.1	HQ416751	HQ416752	HQ416753	HQ416754	HQ416755	HQ416757
NR_113009									
NR_121726	81%								
NR_103954	80%	78%							
HQ416751	76%	77%	78%						
HQ416752	80%	78%	76%	92%					
HQ416753	77%	78%	75%	93%	96%				
HQ416754	78%	78%	78%	97%	92%	92%			
HQ416755	78%	78%	78%	97%	93%	94%	97%		
HQ416757	84%	79%	81%	78%	77%	78%	94%	80%	

Accession numbers:

Armatimonas rosea YO-36^T	NR_113009.1
Fimbriimonas ginsengisoli Gsoil 348^T	NR_121726.1
Chthonomonas calidirosea T49^T	NR_103954.1
Thailand hot spring (Cuecas et al., 2014) OTUs	HQ416751 - HQ416755, HQ416757

Chapter 9 References

Achtman, M., & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews. Microbiology*, *6*(6), 431–40. doi:10.1038/nrmicro1872

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533–8. doi:10.1038/nbt.2579

Alfaro, M. E., & Holder, M. T. (2006). The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 19–42. doi:10.1146/annurev.ecolsys.37.091305.110021

Alfaro, M. E., Zoller, S., & Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, *20*(2), 255–66. doi:10.1093/molbev/msg028

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. doi:10.1016/S0022-2836(05)80360-2

American Public Health Association. (2005). *Standard Methods for the Examination of Water & Wastewater, Centennial Edition [Hardcover]* (21st ed.). Washington DC: American Public Health Association; 21 Har/Cdr edition.

Antoni, D., Zverlov, V. V, & Schwarz, W. H. (2007). Biofuels from microbes. *Applied Microbiology and Biotechnology*, *77*(1), 23–35. doi:10.1007/s00253-007-1163-x

Ashelford, K. E., Weightman, A. J., & Fry, J. C. (2002). PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Research*, *30*(15), 3481–9.

Azam, F. (1998). Microbial control of oceanic carbon flux: the plot thickens. *Science*, *280*(5364), 694–696. doi:10.1126/science.280.5364.694

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science (New York, N.Y.)*, *307*(5717), 1915–20. doi:10.1126/science.1104816

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *19*(5), 455–77. doi:10.1089/cmb.2012.0021

Barns, S. M., Delwiche, C. F., Palmer, J. D., & Pace, N. R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(17), 9188–93.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2011). GenBank. *Nucleic Acids Research*, *39*(Database issue), D32–7. doi:10.1093/nar/gkq1079

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., … Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, *417*(6885), 141–7. doi:10.1038/417141a

Berendsen, R. L., Pieterse, C. M. J., & Bakker, P. A. H. M. (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science*, *17*(8), 478–86. doi:10.1016/j.tplants.2012.04.001

Berg, I. a, Kockelkorn, D., Ramos-Vera, W. H., Say, R. F., Zarzycki, J., Hügler, M., … Fuchs, G. (2010). Autotrophic carbon fixation in archaea. *Nature Reviews. Microbiology*, *8*(6), 447–60. doi:10.1038/nrmicro2365

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., … Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science (New York, N.Y.)*, *277*(5331), 1453–62.

Bond, P. L., Hugenholtz, P., Keller, J., & Blackall, L. L. (1995). Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. *Applied and Environmental Microbiology*, *61*(5), 1910–1916.

Bordi, C., & de Bentzmann, S. (2011). Hacking into bacterial biofilms: a new therapeutic challenge. *Annals of Intensive Care*, *1*(1), 19. doi:10.1186/2110-5820-1-19

Britschgi, T. B., & Giovannoni, S. J. (1991). Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. *Applied and Environmental Microbiology*, *57*(6), 1707–13.

Brock, T. D. (1971). Bimodal distribution of pH values of thermal springs of the world. *Geological Society of America Bulletin*, *82*(5), 1393. doi:10.1130/0016-7606(1971)82[1393:BDOPVO]2.0.CO;2

Brock, T. D., Brock, K. M., Belly, R. T., & Weiss, R. L. (1972). Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. *Archiv Für Mikrobiologie*, *84*(1), 54–68.

Browne, P. R. L. (1978). Hydrothermal alteration in active geothermal fields. *Annual Review of Earth and Planetary Sciences*, *6*, 229–250. doi:10.1146/annurev.ea.06.050178.001305

Brüggemann, H., & Chen, C. (2006). Comparative genomics of Thermus thermophilus: plasticity of the megaplasmid and its contribution to a thermophilic lifestyle. *Journal of Biotechnology*, *124*(4), 654–61. doi:10.1016/j.jbiotec.2006.03.043

Brulc, J. M., Antonopoulos, D. A., Miller, M. E. B., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., … White, B. A. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 1948–53. doi:10.1073/pnas.0806191105

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta*, *1842*(10), 1932–1941. doi:10.1016/j.bbadis.2014.06.015

Burns, B. P., Goh, F., Allen, M., & Neilan, B. A. (2004). Microbial diversity of extant stromatolites in the hypersaline marine environment of Shark Bay, Australia. *Environmental Microbiology*, *6*(10), 1096–101. doi:10.1111/j.1462-2920.2004.00651.x

Butcher, B. G., Mascher, T., & Helmann, J. D. (2008). Environmental sensing and the role of extracytoplasmic function sigma factors. In W. El-Sharoud (Ed.), *Bacterial Physiology* (pp. 233–261). Springer.

Calteau, A., Fewer, D. P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., … Gugger, M. (2014). Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics*, *15*, 977. doi:10.1186/1471-2164-15-977

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for

glycogenomics. *Nucleic Acids Research*, *37*(Database issue), D233–8. doi:10.1093/nar/gkn663

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics (Oxford, England)*, *26*(2), 266–7. doi:10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–6. doi:10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., … Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, *6*(8), 1621–4. doi:10.1038/ismej.2012.8

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., … Karp, P. D. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, *40*(Database issue), D742–53. doi:10.1093/nar/gkr1014

Chain, P., Lamerdin, J., Larimer, F., Regala, W., Lao, V., Land, M., … Arp, D. (2003). Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *Journal of Bacteriology*, *185*(9), 2759–73. doi:10.1128/JB.185.9.2759-2773.2003

Chain, P. S. G., Grafham, D. V, Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., … Detter, J. C. (2009). Genomics. Genome project standards in a new era of sequencing. *Science (New York, N.Y.)*, *326*(5950), 236–7. doi:10.1126/science.1180614

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, *69*(2), 330–9. doi:10.1016/j.mimet.2007.02.005

Chelius, M. K., & Triplett, E. W. (2001). The diversity of archaea and bacteria in association with the roots of *Zea mays* L. *Microbial Ecology*, *41*(3), 252–263. doi:10.1007/s002480000087

Chevreux, B., Wetter, T., & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. In *German Conference on Bioinformatics* (pp. 45–56).

Chien, A., Edgar, D., & Trela, J. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *Journal of Bacteriology*, *127*(3), 1550.

Chivian, D., Brodie, E. L., Alm, E. J., Culley, D. E., Dehal, P. S., DeSantis, T. Z., … Onstott, T. C. (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science (New York, N.Y.)*, *322*(5899), 275–8. doi:10.1126/science.1155495

Cleaver, A. A., Burton, N. P., & Norris, P. R. (2007). A novel *Acidimicrobium* species in continuous cultures of moderately thermophilic, mineral-sulfide-oxidizing acidophiles. *Applied and Environmental Microbiology*, *73*(13), 4294–9. doi:10.1128/AEM.02658-06

Cook, G. M., Janssen, P. H., & Morgan, H. W. (1993). Uncoupler-resistant glucose uptake by the thermophilic glycolytic anaerobe *Thermoanaerobacter thermosulfuricus* (*Clostridium thermohydrosulfuricum*). *Applied and Environmental Microbiology*, *59*(9), 2984–90.

Cook, G. M., Janssen, P. H., Russell, J. B., & Morgan, W. (1994). Dual mechanisms of xylose uptake in the thermophilic bacterium *Thermoanaerobacter thermohydrosulfuricus*. *FEMS Microbiology Letters*, *116*, 257–262.

Costa, K. C., Navarro, J. B., Shock, E. L., Zhang, C. L., Soukup, D., & Hedlund, B. P. (2009). Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. *Extremophiles : Life under Extreme Conditions*, *13*(3), 447–59. doi:10.1007/s00792-009-0230-x

Costello, E. K., Halloy, S. R. P., Reed, S. C., Sowell, P., & Schmidt, S. K. (2009). Fumarole-supported islands of biodiversity within a hyperarid, high-elevation landscape on Socompa Volcano, Puna de Atacama, Andes. *Applied and Environmental Microbiology*, *75*(3), 735–47. doi:10.1128/AEM.01469-08

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–3.

Crocetti, G. R., Banfield, J. F., Keller, J., Bond, P. L., & Blackall, L. L. (2002). Glycogen-accumulating organisms in laboratory-scale and full-scale wastewater treatment processes. *Microbiology*, *148*(11), 3353–3364.

Cuecas, A., Portillo, M. C., Kanoksilapatham, W., & Gonzalez, J. M. (2014). Bacterial distribution along a 50 °C temperature gradient reveals a parceled out hot spring environment. *Microbial Ecology*. doi:10.1007/s00248-014-0437-y

Cummings, M. P., Handley, S. A., Myers, D. S., Reed, D. L., Rokas, A., & Winka, K. (2003). Comparing bootstrap and posterior probability values in the four-taxon case. *Systematic Biology*, *52*(4), 477–87.

Dalevi, D., Hugenholtz, P., & Blackall, L. L. (2001). A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. *International Journal of Systematic and Evolutionary Microbiology*, *51*(Pt 2), 385–91.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS One*, *5*(6), e11147. doi:10.1371/journal.pone.0011147

Davids, W., & Zhang, Z. (2008). The impact of horizontal gene transfer in shaping operons and protein interaction networks--direct evidence of preferential attachment. *BMC Evolutionary Biology*, *8*, 23. doi:10.1186/1471-2148-8-23

De la Torre, J. R., Goebel, B. M., Friedmann, E. I., & Pace, N. R. (2003). Microbial diversity of cryptoendolithic communities from the McMurdo Dry Valleys, Antarctica. *Applied and Environmental Microbiology*, *69*(7), 3858. doi:10.1128/AEM.69.7.3858

DeLeon-Rodriguez, N., Lathem, T. L., Rodriguez-R, L. M., Barazesh, J. M., Anderson, B. E., Beyersdorf, A. J., … Konstantinidis, K. T. (2013). Microbiome of the upper troposphere: species composition and prevalence, effects of tropical storms, and atmospheric implications. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(7), 2575–80. doi:10.1073/pnas.1212089110

Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., … Vorholt, J. A. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(38), 16428–33. doi:10.1073/pnas.0905240106

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, *6*(5), 361–375. doi:nrg1603 [pii] 10.1038/nrg1603

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–72. doi:10.1128/AEM.03006-05

Dojka, M. A., Hugenholtz, P., Haack, S. K., & Pace, N. R. (1998). Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Applied and Environmental Microbiology*, *64*(10), 3869–77.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. doi:10.1186/1471-2148-7-214

Dunfield, P. F., Tamas, I., Lee, K. C., Morgan, X. C., McDonald, I. R., & Stott, M. B. (2012). Electing a candidate: a speculative history of the bacterial phylum OP10. *Environmental Microbiology*, *14*(12), 3069–80. doi:10.1111/j.1462-2920.2012.02742.x

Dunfield, P. F., Yuryev, A., Senin, P., Smirnova, A. V, Stott, M. B., Hou, S., … Alam, M. (2007). Methane oxidation by an extremely acidophilic bacterium of the phylum *Verrucomicrobia*. *Nature*, *450*(7171), 879–82. doi:10.1038/nature06411

Dunn, M. F. (2011). Anaplerotic function of phosphoenolpyruvate carboxylase in *Bradyrhizobium japonicum* USDA110. *Current Microbiology*, *62*(6), 1782–1788. doi:10.1007/s00284-011-9928-y

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–7. doi:10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, *26*(19), 2460–1. doi:10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–8. doi:10.1038/nmeth.2604

Elshahed, M. S., Youssef, N. H., Spain, A. M., Sheik, C., Najar, F. Z., Sukharnikov, L. O., … Krumholz, L. R. (2008). Novelty and uniqueness patterns of rare members of the soil biosphere. *Applied and Environmental Microbiology*, *74*(17), 5422–8. doi:10.1128/AEM.00410-08

Erixon, P., Svennblad, B., Britton, T., & Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, *52*(5), 665–673. doi:10.1080/10635150390235485

Euzéby, J. P. (1997). List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *International Journal of Systematic Bacteriology*, *47*(2), 590–2.

Euzéby, J. P. (2011). List of Prokaryotic names with Standing in Nomenclature. Retrieved from www.bacterio.net

Faoro, H., Alves, A. C., Souza, E. M., Rigo, L. U., Cruz, L. M., Al-Janabi, S. M., … Pedrosa, F. O. (2010). Influence of soil characteristics on the diversity of bacteria in the Southern Brazilian Atlantic Forest. *Applied and Environmental Microbiology*, *76*(14), 4744–9. doi:10.1128/AEM.03025-09

Feller, G., & Gerday, C. (2003). Psychrophilic enzymes: hot topics in cold adaptation. *Nature Reviews. Microbiology*, *1*(3), 200–8. doi:10.1038/nrmicro773

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, *27*(4), 401–410.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164 – 166.

Fierer, N. (2008). Microbial biogeography: patterns in microbial diversity across space and time. In K. Zengler (Ed.), *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back* (1st ed., pp. 95–115). Washington DC: ASM Press.

Fischbach, M. A., & Walsh, C. T. (2009). Antibiotics for emerging pathogens. *Science (New York, N.Y.)*, *325*(5944), 1089–93. doi:10.1126/science.1176667

Fox, J. L. (2005). Current Topics: Ribosomal gene milestone met, already left in dust. *ASM News*, *71*(1), 6–7.

Galardini, M., Biondi, E. G., Bazzicalupo, M., & Mengoni, A. (2011). CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code for Biology and Medicine*, *6*(1), 11. doi:10.1186/1751-0473-6-11

Gao, F., & Zhang, C. T. (2008). Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. *BMC Bioinformatics*, *9*, 79. doi:10.1186/1471-2105-9-79

Garrity, G. M., Bell, J. A., & Lilburn, T. G. (2004). Taxonomic outline of the prokaryotes Release 5.0. In *Bergey's manual of systematic bacteriology* (2nd ed.). New York, NY: Springer-Verlag.

Geissinger, O., Herlemann, D. P. R., Mörschel, E., Maier, U. G., & Brune, A. (2009). The ultramicrobacterium "*Elusimicrobium minutum*" gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. *Applied and Environmental Microbiology*, *75*(9), 2831–40. doi:10.1128/AEM.02697-08

Gernert, C., Glöckner, F. O., Krohne, G., & Hentschel, U. (2005). Microbial diversity of the freshwater sponge *Spongilla lacustris*. *Microbial Ecology*, *50*(2), 206–12. doi:10.1007/s00248-004-0172-x

Gevers, D., Knight, R., Petrosino, J. F., Huang, K., McGuire, A. L., Birren, B. W., … Huttenhower, C. (2012). The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biology*, *10*(8), e1001377. doi:10.1371/journal.pbio.1001377

Gilbert, J. A., Bailey, M., Field, D., Fierer, N., Fuhrman, J. A., Hu, B., … Stevens, R. (2011). The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2011. *Standards in Genomic Sciences*, *5*(2), 243–247. doi:10.4056/sigs.2134923

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, *345*(6270), 60–3. doi:10.1038/345060a0

Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., … Reinhardt, R. (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8298–303. doi:10.1073/pnas.1431443100

Glöckner, J., Kube, M., Shrestha, P. M., Weber, M., Glöckner, F. O., Reinhardt, R., & Liesack, W. (2010). Phylogenetic diversity and metagenomics of candidate division OP3. *Environmental Microbiology*, *12*(5), 1218–29. doi:10.1111/j.1462-2920.2010.02164.x

Gosalbes, M. J., Durbán, A., Pignatelli, M., Abellan, J. J., Jiménez-Hernández, N., Pérez-Cobas, A. E., … Moya, A. (2011). Metatranscriptomic approach to analyze the functional human gut microbiota. *PloS One*, *6*(3), e17447. doi:10.1371/journal.pone.0017447

Grasby, S. E., Richards, B. C., Sharp, C. E., Brady, A. L., Jones, G. M., Dunfield, P. F., & Williamson, M.-C. (2013). The Paint Pots, Kootenay National Park, Canada — a natural acid spring analogue for Mars 1, 2. *Canadian Journal of Earth Sciences*, *50*(1), 94–108. doi:10.1139/e2012-060

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., … Segre, J. A. (2009). Topographical and temporal diversity of the human skin microbiome. *Science (New York, N.Y.)*, *324*(5931), 1190–2. doi:10.1126/science.1171700

Grice, E. A., Snitkin, E. S., Yockey, L. J., Bermudez, D. M., Liechty, K. W., & Segre, J. A. (2010). Longitudinal shift in diabetic wound microbiota correlates with prolonged skin defense response. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(33), 14799–804. doi:10.1073/pnas.1004204107

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(Web Server issue), W52–7. doi:10.1093/nar/gkm360

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696–704. doi:10.1080/10635150390235520

Gupta, R. S., & Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theoretical Population Biology*, *61*(4), 423–434. doi:10.1006/tpbi.2002.1589

Hamilton, M. (2009). *Population Genetics*. Hoboken, NJ: Wiley - Blackwell.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669. doi:10.1128/MBR.68.4.669–685.2004

He, D., Ren, L., & Wu, Q. L. (2014). Contrasting diversity of epibiotic bacteria and surrounding bacterioplankton of a common submerged macrophyte, Potamogeton crispus, in freshwater lakes. *FEMS Microbiology Ecology*, 1–12. doi:10.1111/1574-6941.12414

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological Sciences / The Royal Society*, *270*(1512), 313–21. doi:10.1098/rspb.2002.2218

Henne, A., Brüggemann, H., Raasch, C., Wiezer, A., Hartsch, T., Liesegang, H., … Fritz, H.-J. (2004). The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nature Biotechnology*, *22*(5), 547–53. doi:10.1038/nbt956

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., & Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics (Oxford, England)*, *21*(10), 2329–35. doi:10.1093/bioinformatics/bth324

Herbold, C. W., Lee, C. K., McDonald, I. R., & Cary, S. C. (2014). Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. *Nature Communications*, *5*, 3875. doi:10.1038/ncomms4875

Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, *42*(2), 182–192. doi:10.1093/sysbio/42.2.182

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews. Genetics*, *4*(4), 275–84. doi:10.1038/nrg1044

Holt, J. G. (1994). *Bergey's Manual of Determinative Bacteriology*. Lippincott Williams & Wilkins.

Hreggvidsson, G. O., Petursdottir, S. K., Björnsdottir, S. H., & Fridjonsson, O. H. (2012). Microbial speciation in the geothermal ecosystem. In H. Stan-Lotter & S. Fendrihan (Eds.), *Adaption of Microbial Life to Environmental Extremes* (1st ed., pp. 37–67). Vienna: Springer Vienna. doi:10.1007/978-3-211-99691-1

Hsiao, W. W. L., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B., & Brinkman, F. S. L. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genetics*, *1*(5), e62. doi:10.1371/journal.pgen.0010062

Hu, Z.-Y., Wang, Y.-Z., Im, W.-T., Wang, S.-Y., Zhao, G.-P., Zheng, H.-J., & Quan, Z.-X. (2014). The first complete genome sequence of the Class *Fimbriimonadia* in the Phylum *Armatimonadetes*. *PloS One*, *9*(6), e100794. doi:10.1371/journal.pone.0100794

Huber, H., & Prangishvili, D. (2006). Sulfolobales. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), *The Prokaryotes: Vol. 3 : Archaea. Bacteria: Firmicutes, Actinomycetes* (3rd ed., pp. 33–38). Springer Science & Business Media.

Huber, T., Faulkner, G., & Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics (Oxford, England)*, *20*(14), 2317–9. doi:10.1093/bioinformatics/bth226

Huelsenbeck, J., & Hillis, D. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology*, *44*(1), 17–48. doi:10.1093/sysbio/44.1.17

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, *17*(8), 754–5.

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, *180*(18), 4765–74.

Hugenholtz, P., Hooper, S. D., & Kyrpides, N. C. (2009). Focus: *Synergistetes*. *Environmental Microbiology*, *11*(6), 1327–9. doi:10.1111/j.1462-2920.2009.01949.x

Hugenholtz, P., & Pace, N. R. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology*, *14*(6), 190–197.

Hugenholtz, P., Pitulle, C., Hershberger, K. L., & Pace, N. R. (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *Journal of Bacteriology*, *180*(2), 366–76.

Hugenholtz, P., Tyson, G. W., Webb, R. I., Wagner, A. M., & Blackall, L. L. (2001). Investigation of candidate division TM7, a recently recognized major lineage of the domain *Bacteria* with no known pure-culture representatives. *Applied and Environmental Microbiology*, *67*(1), 411–419. doi:10.1128/AEM.67.1.411-419.2001

Im, W.-T., Hu, Z.-Y., Kim, K.-H., Rhee, S.-K., Meng, H., Lee, S.-T., & Quan, Z.-X. (2012). Description of *Fimbriimonas ginsengisoli* gen. nov., sp. nov. within the *Fimbriimonadia* class nov., of the phylum *Armatimonadetes*. *Antonie van Leeuwenhoek*, *102*(2), 307–17. doi:10.1007/s10482-012-9739-6

Isenbarger, T. A., Finney, M., Ríos-Velázquez, C., Handelsman, J., & Ruvkun, G. (2008). Miniprimer PCR, a new lens for viewing the microbial world. *Applied and Environmental Microbiology*, *74*(3), 840–9. doi:10.1128/AEM.01933-07

Jackson, T. J., Ramaley, R. F., & Meinschein, W. G. (1973). *Thermomicrobium*, a new genus of extremely thermophilic bacteria. *International Journal of Systematic Bacteriology*, *23*(1), 28–36. doi:10.1099/00207713-23-1-28

Jain, P., & Sinha, S. (2009). Neutrophiles: Acid challenge and comparison with acidophiles. *The Internet Journal of Microbiology*, *7*(1).

Jiang, L., Lin, M., Li, X., Cui, H., Xu, X., Li, S., & Huang, H. (2013). Genome Sequence of *Thermus thermophilus* ATCC 33923, a thermostable trehalose-producing strain. *Genome Announcements*, *1*(4). doi:10.1128/genomeA.00493-13

Joshua, C. J., Dahl, R., Benke, P. I., & Keasling, J. D. (2011). Absence of diauxie during simultaneous utilization of glucose and xylose by *Sulfolobus acidocaldarius*. *Journal of Bacteriology*, *193*(6), 1293–301. doi:10.1128/JB.01219-10

Jumas-Bilak, E., Roudière, L., & Marchandin, H. (2009). Description of "*Synergistetes*" phyl. nov. and emended description of the phylum "*Deferribacteres*" and of the family *Syntrophomonadaceae*, phylum "*Firmicutes*". *International Journal of Systematic and Evolutionary Microbiology*, *59*(Pt 5), 1028–35. doi:10.1099/ijs.0.006718-0

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(40), 16213–6. doi:10.1073/pnas.1203849109

Kan, J., Clingenpeel, S., Macur, R. E., Inskeep, W. P., Lovalvo, D., Varley, J., … Nealson, K. (2011). Archaea in Yellowstone Lake. *The ISME Journal*, *5*(11), 1784–95. doi:10.1038/ismej.2011.56

Kanchikerimath, M., & Singh, D. (2001). Soil organic matter and biological properties after 26 years of maize–wheat–cowpea cropping as affected by manure and fertilization in a Cambisol in semiarid region of India. *Agriculture, Ecosystems & Environment*, *86*(2), 155–162. doi:10.1016/S0167-8809(00)00280-2

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*(Database issue), D109–14. doi:10.1093/nar/gkr988

Kaoutari, A. E., Armougom, F., Gordon, J. I., Raoult, D., & Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews. Microbiology*, *11*(7), 497–504. doi:10.1038/nrmicro3050

Karsch-Mizrachi, I., Nakamura, Y., & Cochrane, G. (2012). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, *40*(Database issue), D33–7. doi:10.1093/nar/gkr1006

Keller, M., & Zengler, K. (2004). Tapping into microbial diversity. *Nature Reviews. Microbiology*, *2*(2), 141–50. doi:10.1038/nrmicro819

Kersters, K., Lisdiyanti, P., Komagata, K., & Swings, J. (2006). The family *Acetobacteraceae*: the genera *Acetobacter*, *Acidomonas*, *Asaia*, *Gluconacetobacter*, *Gluconobacter*, and *Kozakia*. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), *The Prokaryotes: Vol. 5: Proteobacteria: Alpha and Beta Subclasses* (pp. 163–200). New York, NY: Springer New York. doi:10.1007/0-387-30745-1

Kiewitz, C., & Tümmler, B. (2000). Sequence diversity of *Pseudomonas aeruginosa*: impact on population structure and genome evolution. *Journal of Bacteriology*, *182*(11), 3125–35.

Kimura, M. (1985). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kittelmann, S., Seedorf, H., Walters, W. A., Clemente, J. C., Knight, R., Gordon, J. I., & Janssen, P. H. (2013). Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PloS One*, *8*(2), e47879. doi:10.1371/journal.pone.0047879

Kobayashi, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research*, *29*(18), 3742–56.

Koning, S. M., Albers, S.-V., Konings, W. N., & Driessen, A. J. M. (2002). Sugar transport in (hyper)thermophilic archaea. *Research in Microbiology*, *153*(2), 61–7.

Konings, W. N., Albers, S.-V., Koning, S., & Driessen, A. J. M. (2002). The cell membrane plays a crucial role in survival of bacteria and archaea in extreme environments. *Antonie van Leeuwenhoek*, *81*(1-4), 61–72.

Könneke, M., Schubert, D. M., Brown, P. C., Hügler, M., Standfest, S., Schwander, T., … Berg, I. A. (2014). Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(22), 8239–44. doi:10.1073/pnas.1402028111

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2567–72. doi:10.1073/pnas.0409727102

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to

complete genomes. *Journal of Molecular Biology*, *305*(3), 567–80. doi:10.1006/jmbi.2000.4315

Krulwich, T. A., Sachs, G., & Padan, E. (2011). Molecular aspects of bacterial pH sensing and homeostasis. *Nature Review Microbiol*, *9*(5), 330–343. doi:nrmicro2549 [pii] 10.1038/nrmicro2549

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. a, Sogin, M. L., & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(20), 6955–9.

Langille, M. G. I., & Brinkman, F. S. L. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics (Oxford, England)*, *25*(5), 664–5. doi:10.1093/bioinformatics/btp030

Lapage, S., Sneath, P., Lessel, E., Skerman, V., Seeliger, H., & Clark, W. (1992). International Code of Nomenclature of Bacteria. ASM Press.

Larsen, N., Olsen, G. J., Maidak, B. L., McCaughey, M. J., Overbeek, R., Macke, T. J., … Woese, C. R. (1993). The ribosomal database project. *Nucleic Acids Research*, *21*(13), 3021–3.

Lasken, R. S. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews. Microbiology*, *10*(9), 631–40. doi:10.1038/nrmicro2857

Lee, K. C., Morgan, X. C., Dunfield, P. F., Tamas, I., McDonald, I. R., & Stott, M. B. (2014). Genomic analysis of *Chthonomonas calidirosea*, the first sequenced isolate of the phylum *Armatimonadetes*. *The ISME Journal*, *8*, 1522–1533. doi:10.1038/ismej.2013.251

Lee, K. C. Y., Herbold, C. W., Dunfield, P. F., Morgan, X. C., McDonald, I. R., & Stott, M. B. (2013). Phylogenetic delineation of the novel phylum *Armatimonadetes* (former candidate division OP10) and definition of two novel candidate divisions. *Applied and Environmental Microbiology*, *79*(7), 2484–7. doi:10.1128/AEM.03333-12

Lee, K. C.-Y., Dunfield, P. F., Morgan, X. C., Crowe, M. A., Houghton, K. M., Vyssotski, M., … Stott, M. B. (2011). *Chthonomonas calidirosea* gen. nov., sp. nov., an aerobic, pigmented, thermophilic micro-organism of a novel bacterial class, *Chthonomonadetes* classis nov., of the newly described phylum *Armatimonadetes* originally designated candidate division OP10. *International Journal of Systematic and Evolutionary Microbiology*, *61*(Pt 10), 2482–90. doi:10.1099/ijs.0.027235-0

Lee, Y. K., & Mazmanian, S. K. (2010). Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science (New York, N.Y.)*, *330*(6012), 1768–73. doi:10.1126/science.1195568

Legendre, P., & Legendre, L. F. J. (1998). *Numerical Ecology* (2nd ed., Vol. 1998). Amsterdam, New York: Elsevier.

Lehours, A. C., Evans, P., Bardot, C., Joblin, K., & Gerard, F. (2007). Phylogenetic diversity of archaea and bacteria in the anoxic zone of a meromictic lake (Lake Pavin, France). *Applied and Environmental Microbiology*, *73*(6), 2016–2019. doi:AEM.01490-06 [pii] 10.1128/AEM.01490-06

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, *39*(Database issue), D28–31. doi:10.1093/nar/gkq967

Lewis, K., Epstein, S., D'Onofrio, A., & Ling, L. L. (2010). Uncultured microorganisms as a source of secondary metabolites. *The Journal of Antibiotics*, *63*(8), 468–76. doi:10.1038/ja.2010.87

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., … Gordon, J. I. (2008). Evolution of mammals and their gut microbes. *Science (New York, N.Y.)*, *320*(5883), 1647–51. doi:10.1126/science.1155725

Ley, R. E., Harris, J. K., Wilcox, J., Spear, J. R., Miller, S. R., Bebout, B. M., … Pace, N. R. (2006). Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Applied and Environmental Microbiology*, *72*(5), 3685. doi:10.1128/AEM.72.5.3685

Li, W., Pan, K. W., Wu, N., Wang, J. C., Wang, Y. J., & Zhang, L. (2014). Effect of litter type on soil microbial parameters and dissolved organic carbon in a laboratory microcosm experiment. *Plant, Soil and Environment*, *60*(4), 170–176.

Liolios, K., Chen, I. M., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., & Kyrpides, N. C. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, *38*(Database issue), D346–54. doi:gkp848 [pii] 10.1093/nar/gkp848

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, *2012*, 251364. doi:10.1155/2012/251364

Lu, S., Gischkat, S., Reiche, M., Akob, D. M., Hallberg, K. B., & Küsel, K. (2010). Ecophysiology of Fe-cycling bacteria in acidic sediments. *Applied and Environmental Microbiology*, *76*(24), 8174–83. doi:10.1128/AEM.01931-10

Ludwig, W., & Klenk, H.-P. (2005). Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In *Bergey's Manual® of Systematic Bacteriology* (pp. 49–66). New York: Springer-Verlag. doi:10.1007/0-387-28021-9

Ludwig, W., & Schleifer, K. H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiology Reviews*, *15*(2-3), 155–173. doi:10.1111/j.1574-6976.1994.tb00132.x

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, … Schleifer, K.-H. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research*, *32*(4), 1363–71. doi:10.1093/nar/gkh293

Lynd, L. R., Weimer, P. J., van Zyl, W. H., & Pretorius, I. S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. *Microbiology and Molecular Biology Reviews : MMBR*, *66*(3), 506–77, table of contents.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., … Mills, D. (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15611–6. doi:10.1073/pnas.0607117103

Marcy, Y., Ouverney, C., Bik, E. M., Losekann, T., Ivanova, N., Martin, H. G., … Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 11889–11894. doi:0704662104 [pii] 10.1073/pnas.0704662104

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., … Kyrpides, N. C. (2010). The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Research*, *38*(Database issue), D382–90. doi:10.1093/nar/gkp887

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., … Kyrpides, N. C. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Research*, *40*(Database issue), D115–22. doi:10.1093/nar/gkr1044

Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. A., Wu, M., McNulty, N. P., … Gordon, J. I. (2011). Recognition and degradation of plant cell wall polysaccharides

by two human gut symbionts. *PLoS Biology*, *9*(12), e1001221. doi:10.1371/journal.pbio.1001221

Mattimore, V., & Battista, J. R. (1996). Radioresistance of *Deinococcus radiodurans*: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *Journal of Bacteriology*, *178*(3), 633–7.

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., … Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, *1*(1), 7. doi:10.1186/2047-217X-1-7

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., … Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–8. doi:10.1038/ismej.2011.139

McLean, J. S., Lombardo, M.-J., Badger, J. H., Edlund, A., Novotny, M., Yee-Greenbaum, J., … Lasken, R. S. (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(26), E2390–9. doi:10.1073/pnas.1219809110

Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J. H. M., … Raaijmakers, J. M. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science (New York, N.Y.)*, *332*(6033), 1097–100. doi:10.1126/science.1203980

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. doi:10.1038/nrg2626

Milani, C., Hevia, A., Foroni, E., Duranti, S., Turroni, F., Lugli, G. A., … Ventura, M. (2013). Assessing the fecal microbiota: an optimized ion torrent 16S rRNA gene-based analysis protocol. *PloS One*, *8*(7), e68739. doi:10.1371/journal.pone.0068739

Miller, S. R., & Castenholz, R. W. (2000). Evolution of thermotolerance in hot spring cyanobacteria of the genus *Synechococcus*. *Applied and Environmental Microbiology*, *66*(10), 4222–4229. doi:10.1128/AEM.66.10.4222-4229.2000

Mohamed, S., & Syed, B. A. (2013). Commercial prospects for genomic sequencing technologies. *Nature Reviews. Drug Discovery*, *12*(5), 341–2. doi:10.1038/nrd4006

Moran, M. A. (2009). Metatranscriptomics: eavesdropping on complex microbial communities. *Microbe*, *4*(7), 329–335.

Moreira, L. M., Da Costa, M. S., & Sá-Correia, I. (1997). Comparative genomic analysis of isolates belonging to the six species of the genus *Thermus* using pulsed-field gel electrophoresis and ribotyping. *Archives of Microbiology*, *168*(2), 92–101.

Mori, K., Yamaguchi, K., Sakiyama, Y., Urabe, T., & Suzuki, K. (2009). *Caldisericum exile* gen. nov., sp. nov., an anaerobic, thermophilic, filamentous bacterium of a novel bacterial phylum, *Caldiserica* phyl. nov., originally called the candidate phylum OP5, and description of *Caldisericaceae* fam. nov., *Caldisericales* ord. no. *International Journal of Systematic and Evolutionary Microbiology*, *59*(Pt 11), 2894–8. doi:10.1099/ijs.0.010033-0

Mosmann, T. (1983). Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays. *Journal of Immunological Methods*, *65*(1-2), 55–63. doi:10.1016/0022-1759(83)90303-4

Mukherjee, S., Juottonen, H., Siivonen, P., Lloret Quesada, C., Tuomi, P., Pulkkinen, P., & Yrjälä, K. (2014). Spatial patterns of microbial diversity and activity in an aged creosote-contaminated site. *The ISME Journal*, *8*(10), 2131–42. doi:10.1038/ismej.2014.151

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., … Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research*, *38*(Database issue), D190–5. doi:10.1093/nar/gkp951

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., … Springer, M. S. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science (New York, N.Y.)*, *294*(5550), 2348–51. doi:10.1126/science.1067179

Myllykangas, S., Buenrostro, J., & Ji, H. P. (2012). Bioinformatics for high throughput sequencing. In N. Rodríguez-Ezpeleta, M. Hackenberg, & A. M. Aransay (Eds.), (pp. 11–25). New York, NY: Springer New York. doi:10.1007/978-1-4614-0782-9

Nagorska, K., Ostrowski, a, Hinc, K., Holland, I. B., & Obuchowski, M. (2010). Importance of eps genes from *Bacillus subtilis* in biofilm formation and swarming. *Journal of Applied Genetics*, *51*(3), 369–81. doi:10.1007/BF03208867

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., … Allen, E. E. (2012). *De novo* metagenomic assembly reveals abundant novel major lineage of *Archaea* in hypersaline microbial communities. *The ISME Journal*, *6*(1), 81–93. doi:10.1038/ismej.2011.78

Nayak, K. C. (2013). Comparative genome sequence analysis of *Sulfolobus acidocaldarius* and 9 other isolates of its genus for factors influencing codon and amino acid usage. *Gene*, *513*(1), 163–73. doi:10.1016/j.gene.2012.10.024

Nemergut, D. R., Townsend, A. R., Sattin, S. R., Freeman, K. R., Fierer, N., Neff, J. C., … Schmidt, S. K. (2008). The effects of chronic nitrogen fertilization on alpine tundra soil microbial communities: implications for carbon and nitrogen cycling. *Environmental Microbiology*, *10*(11), 3093–105. doi:10.1111/j.1462-2920.2008.01735.x

Norris, P. R., Clark, D. A., Owen, J. P., & Waterhouse, S. (1996). Characteristics of *Sulfobacillus acidophilus* sp. nov. and other moderately thermophilic mineral-sulphide-oxidizing bacteria. *Microbiology*, *142*(4), 775–783. doi:10.1099/00221287-142-4-775

Novick, A., & Szilard, L. (1950). Description of the chemostat. *Science (New York, N.Y.)*, *112*(2920), 715–716. doi:10.1126/science.112.2920.715

O'Leary, M. H., & Diaz, E. (1982). Phosphoenol-3-bromopyruvate. A mechanism-based inhibitor of phosphoenolpyruvate carboxylase from maize. *The Journal of Biological Chemistry*, *257*(24), 14603–14605.

Offre, P., Spang, A., & Schleper, C. (2013). Archaea in biogeochemical cycles. *Annual Review of Microbiology*, *67*, 437–57. doi:10.1146/annurev-micro-092412-155614

Olson, D. G., Tripathi, S. A., Giannone, R. J., Lo, J., Caiazza, N. C., Hogsett, D. A., … Lynd, L. R. (2010). Deletion of the Cel48S cellulase from *Clostridium thermocellum*. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(41), 17727–17732. doi:10.1073/pnas.1003584107

Oshima, M., & Ariga, T. (1975). Omega-cyclohexyl fatty acids in acidophilic thermophilic bacteria. Studies on their presence, structure, and biosynthesis using precursors labeled with stable isotopes and radioisotopes. *Journal of Biological Chemistry*, *250*(17), 6963–6968.

Osman, K. T. (2012). Soils - Principles, Properties and Management. In *Soils - Principles, Properties and Management* (p. 109). Springer.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*(5313), 734.

Pace, N. R. (2009). Mapping the tree of life: progress and prospects. *Microbiology and Molecular Biology Reviews : MMBR*, *73*(4), 565–76. doi:10.1128/MMBR.00033-09

Park, B. H., Karpinets, T. V, Syed, M. H., Leuze, M. R., & Uberbacher, E. C. (2010). CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology*, *20*(12), 1574–1584. doi:10.1093/glycob/cwq106

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., … Maskell, D. J. (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genetics*, *35*(1), 32–40. doi:10.1038/ng1227

Pati, A., Ivanova, N. N., Mikhailova, N., Ovchinnikova, G., Hooper, S. D., Lykidis, A., & Kyrpides, N. C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature Methods*, *7*(6), 455–7. doi:10.1038/nmeth.1457

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., & Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(33), 13272–7. doi:10.1073/pnas.1121464109

Peplies, J., Kottmann, R., Ludwig, W., & Glöckner, F. O. (2008). A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. *Systematic and Applied Microbiology*, *31*(4), 251–7. doi:10.1016/j.syapm.2008.08.003

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–6. doi:10.1038/nmeth.1701

Podar, M., Abulencia, C. B., Walcher, M., Hutchison, D., Zengler, K., Garcia, J. A., … Keller, M. (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology*, *73*(10), 3205–14. doi:10.1128/AEM.02985-06

Pope, P. B., & Patel, B. K. C. (2008). Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiology Ecology*, *64*(1), 9–27. doi:10.1111/j.1574-6941.2008.00448.x

Poretsky, R. S., Hewson, I., Sun, S., Allen, A. E., Zehr, J. P., & Moran, M. A. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environmental Microbiology*, *11*(6), 1358–75. doi:10.1111/j.1462-2920.2008.01863.x

Portillo, M. C., & Gonzalez, J. M. (2008). Members of the Candidate Division OP10 are spread in a variety of environments. *World Journal of Microbiology and Biotechnology*, *25*(2), 347–353. doi:10.1007/s11274-008-9895-z

Portillo, M., Sririn, V., Kanoksilapatham, W., & Gonzalez, J. (2009). Pigment profiles and bacterial communities from Thailand thermal mats. *Antonie van Leeuwenhoek*, *96*(4), 559–67. doi:10.1007/s10482-009-9371-2

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, *25*(7), 1253–6. doi:10.1093/molbev/msn083

Potvin, E., Sanschagrin, F., & Levesque, R. C. (2008). Sigma factors in *Pseudomonas aeruginosa*. *FEMS Microbiology Reviews*, *32*(1), 38–55. doi:10.1111/j.1574-6976.2007.00092.x

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One*, *5*(3), e9490. doi:10.1371/journal.pone.0009490

Prowe, S. G., & Antranikian, G. (2001). Anaerobranca gottschalkii sp. nov., a novel thermoalkaliphilic bacterium that grows anaerobically at high pH and temperature. *International Journal of Systematic and Evolutionary Microbiology*, *51*(Pt 2), 457–65.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–96. doi:10.1093/nar/gkm864

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., … Sloan, W. T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, *6*(9), 639–41. doi:10.1038/nmeth.1361

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria.

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, *57*, 369–94. doi:10.1146/annurev.micro.57.030502.090759

Renders, N., Römling, Y., Verbrugh, H., & van Belkum, A. (1996). Comparative typing of Pseudomonas aeruginosa by random amplification of polymorphic DNA or pulsed-field gel electrophoresis of DNA macrorestriction fragments. *Journal of Clinical Microbiology*, *34*(12), 3190–5.

Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V, & Whitaker, R. J. (2009). Biogeography of the Sulfolobus islandicus pan-genome. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8605–10. doi:10.1073/pnas.0808945106

Reyes, A. G. (1990). Petrology of Philippine geothermal systems and the application of alteration mineralogy to their assessment. *Journal of Volcanology and Geothermal Research*, *43*(1-4), 279–309. doi:10.1016/0377-0273(90)90057-M

Reysenbach, A. L. (2001). Class IV. Thermoplasmata class. nov. In D. R. Boone & R. W. Castenholz (Eds.), *Bergey's Manual of Systematic Bacteriology vol. 1 (The Archaea and the deeply branching and phototrophic Bacteria)* (2nd ed.). New York: Springer-Verlag.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–7. doi:10.1038/nature12352

Rogers, J. S. (2001). Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Systematic Biology*, *50*(5), 713–22.

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)*, *19*(12), 1572–4.

Rosselló-Móra, R. (2012). Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. *Environmental Microbiology*, *14*(2), 318–34. doi:10.1111/j.1462-2920.2011.02599.x

Rothschild, L. J., & Mancinelli, R. L. (2001). Life in extreme environments. *Nature*, *409*(6823), 1092–101. doi:10.1038/35059215

Rueckert, A., & Morgan, H. W. (2007). Removal of contaminating DNA from polymerase chain reaction using ethidium monoazide. *Journal of Microbiological Methods*, *68*(3), 596–600. doi:10.1016/j.mimet.2006.11.006

Saier, M. H. (2001). The bacterial phosphotransferase system: structure, function, regulation and evolution. *Journal of Molecular Microbiology and Biotechnology*, *3*(3), 325–7.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–41. doi:10.1128/AEM.01541-09

Schmid, S., Bevilacqua, C., & Crutz-Le Coq, A.-M. (2012). Alternative sigma factor σH activates competence gene expression in Lactobacillus sakei. *BMC Microbiology*, *12*(1), 32. doi:10.1186/1471-2180-12-32

Schmidt, T. M., DeLong, E. F., & Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, *173*(14), 4371–8.

Schwartzman, D. W., & Lineweaver, C. H. (2004). The hyperthermophilic origin of life revisited. *Biochemical Society Transactions*, *32*(Pt 2), 168–71. doi:10.1042/

Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, *4*, 2304. doi:10.1038/ncomms3304

Sessions, A. L., Doughty, D. M., Welander, P. V, Summons, R. E., & Newman, D. K. (2009). The continuing puzzle of the great oxidation event. *Current Biology : CB*, *19*(14), R567–74. doi:10.1016/j.cub.2009.05.054

Shao, K., Gao, G., Qin, B., Tang, X., Wang, Y., Chi, K., & Dai, J. (2011). Comparing sediment bacterial communities in the macrophyte-dominated and algae-dominated areas of eutrophic Lake Taihu, China. *Canadian Journal of Microbiology*, *57*(4), 263–72. doi:10.1139/w11-003

Sharon, I., & Banfield, J. F. (2013). Microbiology. Genomes from metagenomics. *Science (New York, N.Y.)*, *342*(6162), 1057–8. doi:10.1126/science.1247023

Shi, C., Wang, C., Xu, X., Huang, B., Wu, L., & Yang, D. (2015). Comparison of bacterial communities in soil between nematode-infected and nematode-uninfected *Pinus massoniana* pinewood forest. *Applied Soil Ecology*, *85*, 11–20. doi:10.1016/j.apsoil.2014.08.008

Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, *16*(8), 1114–1116. doi:10.1093/oxfordjournals.molbev.a026201

Siala, M., Gdoura, R., Fourati, H., Rihl, M., Jaulhac, B., Younes, M., … Hammami, A. (2009). Broad-range PCR, cloning and sequencing of the full 16S rRNA gene for detection of bacterial DNA in synovial fluid samples of Tunisian patients with reactive and undifferentiated arthritis. *Arthritis Research & Therapy*, *11*(4), R102. doi:10.1186/ar2748

Silva, Z., Sampaio, M., Henne, A., Böhm, A., Gutzat, R., Boos, W., … Santos, H. (2005). The high-affinity maltose/trehalose ABC transporter in the extremely thermophilic bacterium *Thermus thermophilus* HB27 also recognizes sucrose and palatinose. *Journal of Bacteriology*, *187*(4), 1210–8. doi:10.1128/JB.187.4.1210-1218.2005

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. *International Journal of Systematic Bacteriology*, *44*(4), 846–849. doi:10.1099/00207713-44-4-846

Stackebrandt, E., Liesack, W., & Goebel, B. M. (1993). Bacterial diversity in a soil sample from a subtropical Australian environment as determined by 16S rDNA analysis. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *7*(1), 232–6.

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, *39*, 321–46. doi:10.1146/annurev.mi.39.100185.001541

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, *22*(21), 2688–90. doi:10.1093/bioinformatics/btl446

Stein, L. Y., Jones, G., Alexander, B., Elmund, K., Wright-jones, C., & Nealson, K. H. (2002). Intriguing microbial diversity associated with metal-rich particles from a freshwater reservoir. *FEMS Microbiology Ecology*, *42*, 431–440.

Steiner, A. (1977). The Wairakei geothermal area, North Island, New Zealand: its subsurface geology and hydrothermal rock alteration. In *New Zealand Geological Survey Bulletin 90* (p. 136). New Zealand Department of Scientific and Industrial Research.

Stern, A., & Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *33*(1), 43–51. doi:10.1002/bies.201000071

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of Bacteriology*, *194*(16), 4151–60. doi:10.1128/JB.00345-12

Stieglmeier, M., Klingl, A., Alves, R. J. E., Rittmann, S. K.-M. R., Melcher, M., Leisch, N., & Schleper, C. (2014). *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum *Thaumarchaeota*. *International Journal of Systematic and Evolutionary Microbiology*, *64*(Pt 8), 2738–52. doi:10.1099/ijs.0.063172-0

Stott, M. B., Crowe, M. a, Mountain, B. W., Smirnova, A. V, Hou, S., Alam, M., & Dunfield, P. F. (2008). Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand. *Environmental Microbiology*, *10*(8), 2030–41. doi:10.1111/j.1462-2920.2008.01621.x

Suchard, M. A., & Redelings, B. D. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics (Oxford, England)*, *22*(16), 2047–8. doi:10.1093/bioinformatics/btl175

Takami, H., Noguchi, H., Takaki, Y., Uchiyama, I., Toyoda, A., Nishi, S., … Takai, K. (2012). A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. *PloS One*, *7*(1), e30559. doi:10.1371/journal.pone.0030559

Tamaki, H., Tanaka, Y., Matsuzawa, H., Muramatsu, M., Meng, X.-Y., Hanada, S., … Kamagata, Y. (2011). *Armatimonas rosea* gen. nov., sp. nov., of a novel bacterial phylum, *Armatimonadetes* phyl. nov., formally called the candidate phylum OP10. *International Journal of Systematic and Evolutionary Microbiology*, *61*(Pt 6), 1442–7. doi:10.1099/ijs.0.025643-0

Tanaka, Y., Tamaki, H., Matsuzawa, H., Nigaya, M., Mori, K., & Kamagata, Y. (2012). Microbial community analysis in the roots of aquatic plants and isolation of novel microbes including an organism of the candidate phylum OP10. *Microbes and Environments / JSME*, *27*(2), 149–57.

Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., & Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Research*, *30*(1), 27–30.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V, … Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, *4*, 41. doi:10.1186/1471-2105-4-41

Taylor, D. J., & Piel, W. H. (2004). An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Molecular Biology and Evolution*, *21*(8), 1534–7. doi:10.1093/molbev/msh156

Tindall, B. J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Glavina Del Rio, T., … Lapidus, A. (2010). Complete genome sequence of Meiothermus ruber type strain (21). *Standards in Genomic Sciences*, *3*(1), 26–36. doi:10.4056/sigs.1032748

Tolonen, A. C., Chilaka, A. C., & Church, G. M. (2009). Targeted gene inactivation in *Clostridium phytofermentans* shows that cellulose degradation requires the family 9

hydrolase Cphy3367. *Molecular Microbiology*, *74*(6), 1300–1313. doi:DOI 10.1111/j.1365-2958.2009.06890.x

Tomova, I., Stoilova-Disheva, M., Lyutskanova, D., Pascual, J., Petrov, P., & Kambourova, M. (2010). Phylogenetic analysis of the bacterial community in a geothermal spring, Rupi Basin, Bulgaria. *World Journal of Microbiology and Biotechnology*, *26*(11), 2019–2028. doi:10.1007/s11274-010-0386-7

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426. doi:10.1016/j.tig.2014.07.001

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics (Oxford, England)*, *22*(18), 2196–203. doi:10.1093/bioinformatics/btl369

Vetriani, C., Jannasch, H. W., MacGregor, B. J., Stahl, D. A., & Reysenbach, A. L. (1999). Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Applied and Environmental Microbiology*, *65*(10), 4375–84.

Villanueva, L., Damsté, J. S. S., & Schouten, S. (2014). A re-evaluation of the archaeal membrane lipid biosynthetic pathway. *Nature Reviews. Microbiology*, *12*(6), 438–48. doi:10.1038/nrmicro3260

Vyssotski, M., Lee, K. C.-Y., Lagutin, K., Ryan, J., Morgan, X. C., & Stott, M. B. (2011). Fatty Acids of *Chthonomonas calidirosea*, of a novel class *Chthonomonadetes* from a recently described phylum *Armatimonadetes*. *Lipids*, *46*(12), 1155–61. doi:10.1007/s11745-011-3597-2

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., … Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, *7*(1), 142. doi:10.1186/1471-2105-7-142

Wang, H., Vuorela, M., Keränen, A.-L., Lehtinen, T. M., Lensu, A., Lehtomäki, A., & Rintala, J. (2010). Development of microbial populations in the anaerobic hydrolysis of grass silage for methane production. *FEMS Microbiology Ecology*, *72*(3), 496–506. doi:10.1111/j.1574-6941.2010.00850.x

Wang, P., Qi, M., Barboza, P., Leigh, M. B., Ungerfeld, E., Selinger, L. B., … Forster, R. J. (2011). Isolation of high-quality total RNA from rumen anaerobic bacteria and fungi, and subsequent detection of glycoside hydrolases. *Canadian Journal of Microbiology*, *57*(7), 590–598. doi:10.1139/W11-048

Ward, N. L., Challacombe, J. F., Janssen, P. H., Henrissat, B., Coutinho, P. M., Wu, M., … Kuske, C. R. (2009). Three genomes from the phylum *Acidobacteria* provide insight into the lifestyles of these microorganisms in soils. *Applied and Environmental Microbiology*, *75*(7), 2046–56. doi:10.1128/AEM.02294-08

Weissbrodt, D. G., Shani, N., & Holliger, C. (2014). Linking bacterial population dynamics and nutrient removal in the granular sludge biofilm ecosystem engineered for wastewater treatment. *FEMS Microbiology Ecology*, *88*(3), 579–95. doi:10.1111/1574-6941.12326

Whitaker, R. J., Grogan, D. W., & Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science (New York, N.Y.)*, *301*(5635), 976–8. doi:10.1126/science.1086909

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). *Prokaryotes*: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 6578–6583.

Wilcox, T. P., Zwickl, D. J., Heath, T. a, & Hillis, D. M. (2002). Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap

measures of phylogenetic support. *Molecular Phylogenetics and Evolution*, *25*(2), 361–71.

Williams, R. A., Smith, K. E., Welch, S. G., & Micallef, J. (1996). *Thermus oshimai* sp. nov., isolated from hot springs in Portugal, Iceland, and the Azores, and comment on the concept of a limited geographical distribution of *Thermus* species. *International Journal of Systematic Bacteriology*, *46*(2), 403–8.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, *51*(2), 221–71.

Woese, C. R., & Fox, G. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088.

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 of the United States of America*, *87*(12), 4576–9.

Wong, F. K. Y., Lacap, D. C., Lau, M. C. Y., Aitchison, J. C., Cowan, D. A., & Pointing, S. B. (2010). Hypolithic microbial community of quartz pavement in the high-altitude tundra of central Tibet. *Microbial Ecology*, *60*(4), 730–9. doi:10.1007/s00248-010-9653-2

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., … Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. *Nature*, *462*(7276), 1056–1060.

Wu, D., Raymond, J., Wu, M., Chatterji, S., Ren, Q., Graham, J. E., … Eisen, J. A. (2009). Complete genome sequence of the aerobic CO-oxidizing thermophile *Thermomicrobium roseum*. *PloS One*, *4*(1), e4207. doi:10.1371/journal.pone.0004207

Wu, M., & Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, *9*(10), R151. doi:gb-2008-9-10-r151 [pii] 10.1186/gb-2008-9-10-r151

Wu, S., Gao, T., Zheng, Y., Wang, W., Cheng, Y., & Wang, G. (2010). Microbial diversity of intestinal contents and mucus in yellow catfish (*Pelteobagrus fulvidraco*). *Aquaculture*, *303*(1-4), 1–7. doi:10.1016/j.aquaculture.2009.12.025

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., & Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, *2*, 26. doi:10.1186/2049-2618-2-26

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., … Gordon, J. I. (2003). A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science (New York, N.Y.)*, *299*(5615), 2074–6. doi:10.1126/science.1080029

Xu, J., Chiang, H., Bjursell, M., & Gordon, J. (2004). Message from a human gut symbiont: sensitivity is a prerequisite for sharing. *Trends in Microbiology*, *12*(1), 21–28. doi:10.1016/j.tim.2003.11.007

Yang, Z., Goldman, N., & Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum- likelihood phylogenetic estimation. *Mol. Biol. Evol.*, *11*(2), 316–324.

Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F. O., & Rosselló-Móra, R. (2010). Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Systematic and Applied Microbiology*, *33*(6), 291–9. doi:10.1016/j.syapm.2010.08.001

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., … Rosselló-Móra, R. (2008). The All-Species Living Tree project: a 16S rRNA-based

phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, *31*(4), 241–50. doi:10.1016/j.syapm.2008.07.001

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., … Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645. doi:10.1038/nrmicro3330

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., … Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)*, *26*(13), 1608–15. doi:10.1093/bioinformatics/btq249

Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., DeBoy, R. T., … Noll, K. M. (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the *Thermotogales*. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5865–5870. doi:0901260106 [pii] 10.1073/pnas.0901260106