Studies on Genomic Organization, Gene Order, and Recombination in Prokaryotes, Concentrating on Members of the Thermotogales, an Order of Hyperthermophilic Bacteria

by

Mary Ellen Rose Boudreau

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
April 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-27157-5
Our file Notre référence
ISBN: 978-0-494-27157-5

F.M.C.

They say it takes a village to raise a child. Lucky for us we can make up our own village.

This is for the Boudreau clan, both young and old, blood or embraced. Together we can raise each other up to the highest of mountains.

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Comparative genomic analyses of prokaryotes reveal extensive variation in gene content among organisms classified as the same, or closely related, "species". Strains of the same species, as defined by SSU rRNA, can contain 0 to 25% strain-specific genes. Prokaryotic genome plasticity allows for gene shuffling within a genome, along with lateral gene transfer, potentially disrupting conserved gene clusters or operons.

This thesis examined a range of strains from the order Thermotogales at specific areas of the genome to look for rearrangements and unsual features, using *Thermotoga maritima* MSB8 as a comparison point. A long walk PCR technique revealed gene content downstream of *prfA*, which codes for a conserved protein release factor. In *T. maritima* MSB8 and several related taxa, two different flagellar protein gene clusters were found, revealing recombination and different gene histories than the SSU rRNA gene, while maintaining the overall structure of the cluster. Analysis of several clusters of ribosomal protein genes, using genome comparison utilities and sequence analysis from the Thermotogales, reveals gene histories that differ from the SSU rRNA gene, and shuffling of clusters throughout bacteria and archaea. However, in spite of recombination, higher order cluster structure is maintained.

Spatial autocorrelation analysis, most often used for biogeographic studies, was adapted for a novel use in circular prokaryotic genomes, to assess the distribution of functional gene categories within genomes. Because only one Thermotogales genome sequence is available (*T. maritima* MSB8), 26 additional bacterial strains, from six species groups, were used to examine functional genomic architecture. Of particular interest was the distribution of ORFans (orphaned open reading frames), which are hypothetical genes with no known function. In *T. maritima* MSB8, and one other strain, *Chlamydia pneumoniae* AR39, hyperdispersal of true hypothetical proteins was observed, indicating that they were likely misannotated intervening sequence. Conserved hypothetical proteins in all strain groups except *T. maritima* MSB8 and *Prochlorococcus marinus* strains show clustered distributions, implying that they may code for functional clusters not yet discovered, but maintained within strain groups. The random distributions within *T. maritima* MSB8 and *P. marinus* strains likely result from gene insertions and recombinations within existing functional clusters.

List of Abbreviations and Symbols Used

A	adenosine
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pair(s)
°C	degree(s) Celsius
C	cytidine
Cat.	(gene) Category
CHP	conserved hypothetical protein
CMR	Comprehensive Microbial Resource (at TIGR)
D	not C (A, G or T nucleotide)
ddH2O	distilled, deionized H2O
DFC	distal flagellar cluster
dGTP	deoxyguanosine triphosphate
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphate
EDTA	ethylene-diamine-tetra-acetic acid
g	gram
G	guanosine
h	hour(s)
indel	insertion/deletion (of nucleotide or protein sequence)
kbp	kilobase pair(s)

LB	Luria Bertani media
LGT	lateral gene transfer
LSU	large subunit (of the ribosome)
LWPCR	long walk PCR
μL	microliter
μM	micromolar
mg	milligram
min	minute(s)
mL	milliliter
mM	millimolar
N	any (A, C, G or T nucleotide)
ng	nanogram
ORF	open reading frame
PCR	polymerase chain reaction
pH	$-\log_{10}[H^+]$
pmol	picomole
PFC	proximal flagellar cluster
R	purine (A or G nucleotide)
RNA	ribonucleic acid
RNase	ribonuclease
rRNA	ribosomal RNA
SDS	sodium dodecyl sulphate
sp.	species

SSH	suppressive subtractive hybridization
SSU	small subunit (of the ribosome)
str	strain
T	thymidine
TAE	Tris/acetic acid/EDTA
Taq	Thermus aquaticus (DNA polymerase)
TdT	terminal deoxynucleotidyl transferase
TE	Tris-EDTA
TIGR	The Institute for Genomic Research
Tris	tris (hydroxymethyl) aminomethane
THP	true hypothetical protein
U	unit (of enzyme activity)
V	volt
Y	pyrimidine (C or T nucleotide)

ACKNOWLEDGEMENTS

Here, in what is arguably the most important part of the thesis, I will follow the traditional verbose Boudreau format. In no hard and fast order, I thank those who have held me up and brought me forward these past years.

First, to my mentors, Ford and Camilla. By accepting me into the lab, you have shown me what it's like to be amongst greatness. Ford – I am truly honoured to have been part of such a vibrant, exciting research program. Camilla – you are an exceptional scientist, and an example to women in research.

To all Doolittle lab members, past and present – I thank you for making it such a joy to come to work each and every day. In particular, I have to mention John Archibald and Joel Dacks – you both took extra time to make me feel welcome, and you showed me that real people work here.

Katrin Sommerfeld, Marlena Dlutek, and Wanda Danlichuk – I thank you for keeping the lab running, and also keeping us on the ball. I would not have survived without your help in planning, DNA sequencing, and general lab support. You are all gems.

To Faylene Lunn – my fellow grad student, neighbor, and dear friend. I thank you for being there to share in everything non-academic that comes with being a grad student. Sometimes I think we've both been given much more to handle than average twenty-somethings, but you've always shown impeccable poise and grace.

To Amy Coleman – my childhood friend. I thank you for being there for me, through thick and thin, with a smile and a hug. If it weren't for you, I might have been homeless for the last part of this writing process. Over the miles, and through the years, you are indeed a true friend.

To Dr. Beth Retallack – I thank you for providing an example of the harmony that can be created from teaching, research, student advising, and family life. You encouraged me to follow my own path, and helped give me the tools to clear the way.

To the people who led me to Dalhousie, and supported my academic aspirations from a very early age - every student has teachers that they mark as central to their development as both students and people, and I am especially blessed in this area. The late Mary Ellen Cash had the dubious task of helping a precocious fourth grader through her first set of challenges (including long division, and the passing of dear Papa), with a kind hand and unwavering love and encouragement. Donna Burke, who treated all of her English students as prodigies, taught me never to suppress my creativity. I know she wanted me to be a writer; this thesis may not be what she had in mind, but I write with my heart because of her. Finally, the "triumvirate" of Riverview High School – Rick Coleman, Jim Burke, and Kevin Deveaux, who taught Honours Chemistry, Physics, and

Biology; thank you for giving me a great love for the sciences. I cannot chose between them. Rick, who never slows down, keeps me 'energetically' inspired, and instilled in me a love of teaching. Jim, whose very first teaching experience involved 15 or so headstrong overachievers, has grown into a distinguished department head. He showed me the importance of patience and respect between student and teacher, which I couldn't fully understand until I was on the other side of the desk, so to speak. Kevin has left Riverview to lead our rival Sydney Academy as principal, but he is directly responsible for my present interest in molecular biology – his first class on DNA transcription and translation stuck with me, and I still refer to his handouts.

And, finally, to my family. I give them my love and thanks in no particular order, because sibling/spousal/parental rivalry is the last thing I want to create.

Rob – I had hoped we would end up writing our theses at the same time, and I have to admit that it's been difficult to be left behind. I thank you for paving the way, and for showing me that it can, indeed, be done. Your time in graduate school was not easy; in fact, it seemed like people were purposefully trying to knock you down. You refused to fall, and ended up soaring. Randy is a lucky woman.

John – thank you for reminding me what it took for me to get here, and the place I was in not so long ago. We started on the same road, but you had the harder journey. Your perseverance has paid off, and I am so proud of you. You are a shining example of why we should fight for our dreams, but not forget to thank God for unanswered prayers.

Mike – thank you for being you. We got to know each other as colleagues, then as friends, and now we're husband and wife. You understand, probably better than anyone, what this expedition has meant, and I could not have done it without your love and support. You are the greatest gift that graduate school has given me.

Nana – All you've known for the past ten years is that we've been working hard, and that school keeps us away from home more than you'd like. Thank you for encouraging us to take the time to relax and visit, and for being a constant in our lives.

Mom – thank you for keeping me grounded, and making me realize there's more to life than academics. You truly are the keystone to our family, and the most difficult of days can be made better just by hearing your voice.

Dad – you often lament how hard work goes unrecognized these days. As students, we pour our blood, sweat, and tears into our education. We try to fulfill the requirements we're given, and to excel beyond the expected, but the rules inevitably change; you have to sit back and watch as the rug gets pulled out from under our feet. Sometimes it seems like we can never quite get the recognition or reward that we deserve. But we, as your children, know every day how proud you are of us.

That is my reward.

CHAPTER 1: INTRODUCTION

1.1: Perspective

The classification of prokaryotic organisms has long been problematic, particularly before the advent of molecular genetics and genomics. As the field moves from traditional microbiology to molecular techniques and DNA sequence assessment, many problems have arisen, particularly with the criteria used to determine the identity of organisms, the appropriateness of various genetic markers, as well as their evolutionary histories, and sources of novel genetic material. This thesis concentrates on small groups of closely related prokaryotic organisms, and attempts to show that, upon closer examination, most of these criteria are inappropriate and not applicable in any but a very general sense. However, groups that are considered to be closely related tend to be cohesive on a more general level, and therefore such groupings are still a useful starting point as new strains are discovered.

1.2: Traditional Microbiology and Molecular Classification Techniques

Traditional bacteriology or microbiology uses tests to determine what family or type of organisms are present; this can include light microscopy (to sort rod, coccus, and spirochaete morphologies), simple enzymatic activity tests, selective staining, and antibiotic resistance. However, these techniques often cannot distinguish between subtle strain differences, including potential virulence factors that may be present in one strain and absent in others.

The advent of DNA sequencing opened up the field to more detailed identification of organisms and their genomic content. Molecules shared amongst organisms could be examined and compared, to determine their relatedness.

In order to have a standard that could be used across all divisions of prokaryotes, a molecule needed to be chosen that was conserved and essential across all life, and could therefore be found in all organisms assayed; the small subunit ribosomal RNA molecule (SSU rRNA) can provide such a standard to assay microbial relationships and evolution (Olsen *et al.* 1986). Ribosomes are necessary for protein translation across all forms of life, and the very nature and complexity of the translation machinery requires conservation. Woese and others (Woese 1987) postulated that the SSU rRNA in prokaryotes, which is approximately 1500 nucleotides long, evolved in a regular, clock-like fashion, and is therefore suitable for phylogenetic analysis.

1.2.1: SSU rRNA as a Phylogenetic Marker

As molecular sequences (both DNA and amino acid) were found to be useful in constructing phylogenetic relationships, SSU rRNA came to the forefront as the most informative and useful molecule to determine the evolutionary history of large groups of microorganisms (Fox *et al.* 1980; Olsen *et al.* 1994; Pace 1997). Hugenholz and colleagues (Hugenholtz *et al.* 1998) stated that the advent of molecular phylogeny removed the need to culture all organisms in order to infer their characteristics; only a small percentage of the prokaryotic world is cultured or culturable, therefore resulting in a large amount of diversity being missed (the inability to culture organisms can be circumvented, in a way, by newer metagenomic techniques, which will be discussed later). Because of the essential function of SSU rRNA as an informational molecule (i.e. involved in the replication, maintenance, and transmission of genetic information) in all living systems, it can provide the basis for many legitimate assessments of bacterial diversity. For example, SSU rRNA sequence phylogeny initiated a major advance toward the establishment of the present three-domain system of life – the Bacteria, the Archaea, and the Eucarya (Woese and Fox 1977; Woese *et al.* 1990). The evolutionary history of the SSU rRNA agrees with broader morphological and biochemical characteristics that define the three major kingdoms of life, such as the presence of a nucleus in eukaryotes, or ether lipids in archaeal membranes. As a result of this apparent utility, large catalogues of SSU rRNA sequences have been amassed (Larsen *et al.* 1993), and used to construct universal phylogenetic trees, which attempt to include all (or a representation of) prokaryotes and indeed all life (Olsen *et al.* 1994). Ribosomal RNAs (and SSU rRNAs in particular) seem, on the surface, to change slowly and consistently. Their evolution can therefore theoretically give a good overall picture of universal history of life, but fine differences can be lost.

1.3: Problems with SSU rRNA

If SSU rRNA did actually evolve as a molecular clock, and was inherited solely and faithfully in a vertical fashion, it could be used without fault to establish a universal history of life, and illustrate the relationship between all extant organisms, provided the majority of other molecules were also vertically inherited. However, using SSU rRNA for organismal phylogeny is not always an easy task; methods of phylogenetic reconstruction can give conflicting answers, depending on the dataset and method used. For example, Van de Peer and colleagues (Van de Peer *et al.* 1994) showed that the order of divergence of the major clusters or bacterial groups is not constant in phylogenetic trees, but changes as both the dataset increases in size and the methods of phylogenetic reconstruction are varied. We need to understand, then, what forces may act upon the SSU rRNA molecule, as well as other molecules, to cause changes to the "normal" pattern of vertical inheritance that may confound phylogenetic methods.

1.3.1: Chimeras May Form When Obtaining SSU rRNA Sequences from Unknown, Uncultured Organisms or Environmental Samples

The polymerase chain reaction (PCR) is often used to obtain SSU rRNA sequences from total environmental DNA, or mixed populations of organisms, to discover what unknown or uncultured species may be living in that particular environment. Because of the conservation of the molecule, recombination or chimera formation may take place during the reaction, resulting in a hybrid sequence, containing domains, or segments, with differing evolutionary histories. This makes it extremely difficult to distinguish between a chimeric sequence created by PCR and a truly unique SSU rRNA sequence, depending on the degree of phylogenetic relatedness of the SSU rRNAs involved in the recombination (Ashelford *et al.* 2006; Lu *et al.* 2006; Yu *et al.* 2006). As well, some rRNA genes may be preferentially amplified when using universal primers, giving a skewed representation of the species that are actually living there.

1.3.2: SSU rRNA is not Always Homogeneous Within any Given Organism

Many prokaryotes, such as *E. coli* or the archaeal group Halobacteriales, have multiple rRNA operons within the same cell; it is usually assumed that these operons are homogeneous within a cell or a species, but this is not always the case (for an example, see Boucher *et al.* 2004), where they found intragenomic SSU rRNA sequence variation of over 5%. Dealing with such species presents similar problems to environmental rRNA PCR assays; there are two possible erroneous results: (1) PCR, cloning, and sequencing the SSU rRNA gene may only give a subset of the rRNA operons within the cell; and, (2) chimeras created in the PCR from different regions from an individual cell's different rRNA operons may result, giving an inaccurate picture of the diversity of the operons within that cell, as well as the identity and relationship of that organism to other organisms.

1.3.3: SSU rRNA Identification may not Give any Indication of Unique Characteristics of an Organism

Based on existing database sequences, obtaining an SSU rRNA sequence and determining its closest relatives may not give much, or any, information on the organism itself, its metabolism, lifestyle, or characteristics. Welch and colleagues did a comparison of three completely sequenced *E. coli* genomes (Welch *et al.* 2002), illustrating the extraordinary differences present between organisms considered to be strains of the same species. Two of the three strains are pathogenic, and of the complete non-redundant protein set, only 39.2% of the complete set was shared by all three organisms. Many of the differential genes in the pathogenic strains reside on islands thought to have been acquired by lateral gene transfer, and involved in pathogenicity. Other studies have found unique pathogenicity islands in other strains of *E. coli* (Bingen-Bidois *et al.* 2002; Dobrindt *et al.* 2002; Hejnova *et al.* 2005; Kao *et al.* 1997).

Such large amounts of inter-strain diversity are not limited to pathogens, even though such variability is thought to be a hallmark of the pathogenic lifestyle. As more studies are completed on non-pathogens or environmental strains, significant amounts of differentiation between strains are being seen. For example, Nesbø and colleagues examined two strains of *Thermotoga maritima*, *T. maritima* MSB8 and *T.* sp. RQ2, which are 99.7% identical in SSU rRNA sequence. They determined that the non-sequenced strain had as much as 20% of its genome occupied by completely different genes than those of the sequenced strain, not counting divergent homologs that were also present. Many of the differential genes in *T.* sp. RQ2 were homologous to known sugar metabolism genes, suggesting that this strain may have obtained these novel functions as an adaptation to its environment (Nesbø *et al.* 2002).

1.4: Lateral Gene Transfer

SSU rRNA sequencing, along with more traditional classification techniques, can all be called into question when genetic variation is introduced in ways other than vertical inheritance of mutations and/or duplication and divergence events. This exchange of DNA between separate species, termed lateral gene transfer (or LGT) can theoretically erase phylogenetic signal present in any marker, including SSU rRNA. LGT can be facilitated by three general, established mechanisms:

Transformation. Cells take up DNA from the environment and incorporate it into the genome by homologous or non-homologous recombination.

Conjugation. Cells transfer DNA unidirectionally by a cell-surface bridge, called a pilus. A segment of the donor's chromosome can then be incorporated into the recipient's genome via recombination.

Transduction. Bacteriophages serve as intermediaries in transferring genetic information between a donor and a recipient; the size of the DNA fragment that can be carried by phage is small compared to the other modes of exchange, but it can undergo recombination in the same fashion.

The extent to which LGT erases genetic signal, and how we recognize this, depends on the types of genetic information that tend to be transferred, the evolutionary relationship between donor and recipient, and how the recipient lineage deals with the new information.

Differential gene histories, which can be assessed in a number of ways, provide a way to identify laterally transferred genetic material, and the organisms or families affected. When the sequence of a gene and its homologs are available from a large number of organisms, extensive phylogenetic analyses will show the relatedness of the various homologs, which can reveal the history of that particular gene. Put simplistically, if a gene's history is different from that of the SSU rRNA gene, or any other gene, and one can rule out differential loss of paralogs, some form of transfer has likely taken place.

Genomic subtraction, or suppressive subtractive hybridization (SSH), is an example of a method that can be useful in determining genomic content that is present in one strain and absent from another (Akopyants *et al.* 1998; Lan and Reeves 1996; Nesbø *et al.* 2002; Straus and Ausubel 1990), although not as easily as the more expensive option of complete genome sequencing. Genes that are present in only one of a large group of related organisms may have a unique evolutionary history, and are likely candidates for transfer or recombination events.

Interestingly, SSU rRNA itself is not completely resistant to lateral gene transfer. The very nature of the molecule and its function could in fact make it more likely to be transferred amongst very close relatives, because the same or very similar molecules, such as ribosomal proteins or translation factors, would have evolved in concert with the existing machinery. Ribosomal RNA operons have been exchanged or recombined experimentally, as well as in nature (Amador *et al.* 2000; Asai *et al.* 1999a; Asai *et al.* 1999b; Boucher *et al.* 2004; Hashimoto *et al.* 2003; Suzuki *et al.* 2001), resulting in a fully functional organism, but it is the nature of the molecule that leads some to the conclusion that it should, instead, be resistant to LGT.

1.5: The Complexity Hypothesis

The likelihood of a gene being successfully transferred between two different strains or species of prokaryotes depends on a number of factors. SSU rRNA was originally chosen as the gold standard in phylogenetic markers for a number of reasons. As mentioned, it was thought to evolve at a fairly clocklike rate, with nucleotide substitutions being fairly rare. Also, it is considered to be an informational, as opposed to operational, gene. Informational genes, such as those for the ribosomal RNAs, code for molecules essential to the basic function of the cell, such as DNA transcription and translation. The presumed resistance to transfer and incorporation of informational genes from foreign sources has been termed the "complexity hypothesis" – systems that are essential to the cell and have evolved to have tight regulation are too complex and integrated to be functional in another organism, and are therefore resistant to a different (i.e. laterally transferred) version of any of the components (Jain *et al.* 1999). Although there are many systems within the cell that are complex, informational genes are usually called upon to support this hypothesis, which itself is being increasingly scrutinized as of late.

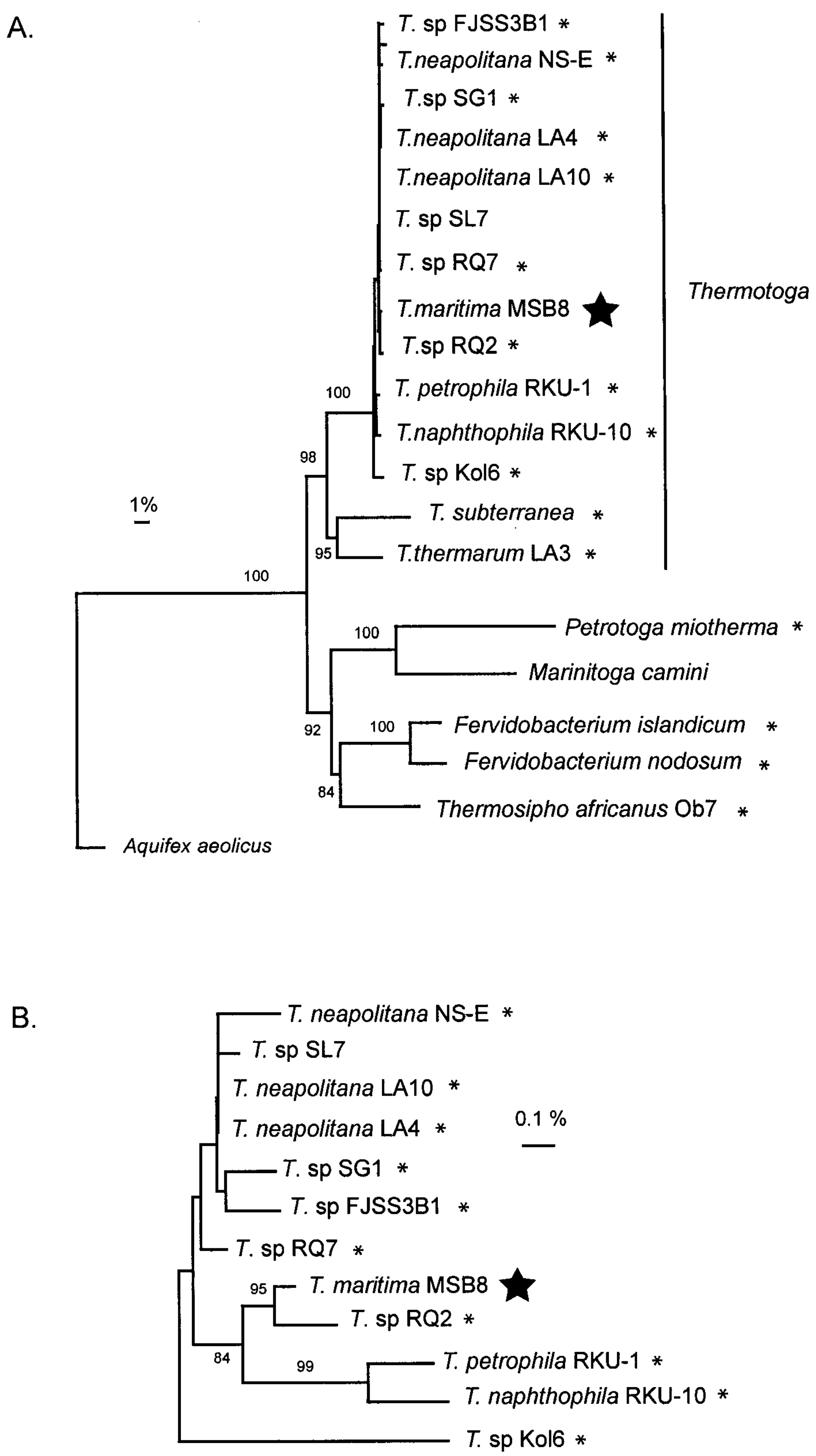
1.6: Thermotogales, a Hyperthermophilic Order of Bacteria, as a Study System

The Thermotogales (Figure 1.1) were chosen as a study system for various reasons, including their unique habitat, biology, and potential mosaic origin of the genome.

This order comprises thermophilic and hyperthermophilic gram-negative rods, which typically have an outer sheath-like membrane, or "toga" (Reysenbach 2001). They are non spore-forming and heterotrophic, and occupy a wide variety of habitats, including geothermally heated vents and oil reservoirs (Fardeau *et al.* 1997; Jeanthon *et al.* 1995; Ravot *et al.* 1995; Takahata *et al.* 2001). These habitats were long thought to be the realm of only archaea, which are typically more extremeophilic than bacteria.

Thermotogales and other hyperthermophiles often branch near the root of the SSU rRNA tree, indicating an early evolutionary origin, owing to either a) a true deep evolutionary history or b) G+C content attraction in the tree, resulting from an abundance of G+C, required for thermostability of RNA secondary structure (Galtier and Lobry 1997; Thoma *et al.* 1998). As well, thermostable proteins often have an excess of cysteine or charged residues, due to the stabilizing nature of disulfide bridges and salt bridges, respectively (Thoma *et al.* 1998).

Figure 1.1 SSU rRNA phylogeny of the Thermotogales, a diverse group of thermophilic and hyperthermophilic bacteria. A) Logdet tree of SSU rRNA sequences of Thermotogales, based on a tree from Nesbø et al, 2001 (Nesbø *et al.* 2001). Thermotoga maritima MSB8, whose genome sequence is available in TIGR, is indicated by a star (★), while additional strains used in this thesis are marked with an asterisk (*). Clades indicated within the genus Thermotoga are 1) the M/N clade, which comprises all strains thought to be either *T. maritima* or *T. neapolitana*, 2) the S/T clade, which comprises *T. subterranea* and *T. thermarum*. B) Phylogeny of the M/N clade, showing the relationships among members. This tree clearly shows the division between a group of *T. maritima* strains and *T. neapolitana* strains, rooted using *T.* sp. Kol6.


A.
T. sp FJSS3B1 *
T.neapolitana NS-E *
T.sp SG1 *
T.neapolitana LA4 *
T.neapolitana LA10 *
T. sp SL7
T. sp RQ7 *
T.maritima MSB8 ★
T.sp RQ2 *
T. petrophila RKU-1 *
T.naphthophila RKU-10 *
T. sp Kol6 *
T. subterranea *
T.thermarum LA3 *
Thermotoga
100
98
95
100
1%
Petrotoga miotherma *
Marinitoga camini
Fervidobacterium islandicum *
Fervidobacterium nodosum *
Thermosipho africanus Ob7 *
100
100
92
84
Aquifex aeolicus
B.
T. neapolitana NS-E *
T. sp SL7
T. neapolitana LA10 *
T. neapolitana LA4 *
0.1 %
T. sp SG1 *
T. sp FJSS3B1 *
T. sp RQ7 *
95
T. maritima MSB8 ★
T. sp RQ2 *
84
99
T. petrophila RKU-1 *
T. naphthophila RKU-10 *
T. sp Kol6 *

The genome of the type strain, *Thermotoga maritima* MSB8, was sequenced (Nelson *et al.* 1999), and analysis found that 24% of its open reading frames had closest database matches to archaeal proteins, suggesting that a great deal of its present genome was acquired by lateral gene transfers from archaea that live in the same environment. This initial assessment has given rise to a rather extensive study of this strain and its relatives. However, the top database match (in the form of a BLAST hit) is not always conclusive in identifying the source of a transferred gene (Koski *et al.* 2001), and so more detailed studies using these genes as starting points can tell us more about the evolution of this group as a whole.

Nesbø and colleagues (Nesbø *et al.* 2001) completed a detailed study on two potential candidates for lateral gene transfer, *gltB* and *ino1*, which revealed that these genes were indeed acquired from Archaea during the divergence of the Thermotogales order. One of these genes, *gltB*, was also transferred from Archaea into three other unrelated bacterial species.

Thermophiles and hyperthermophiles in general present an interesting problem evolutionarily, because of the constraints that are placed on their biomolecules and cellular environment. Hyperthermophiles were originally thought to share only one protein amongst all strains (Forterre 2002), reverse gyrase, which is thought to be necessary to prevent excess unwinding of the DNA double helix at extreme temperatures; it has since been proven unnecessary to maintain a hyperthermophilic lifestyle (Atomi *et al.* 2004) and there does not seem to be any other core set of genes that can be said to define hyperthermophily.

1.7: The Utility of Completely Sequenced Microbial Genomes

On a very grand scale, it is conceivable that large-scale prokaryotic sequencing projects that encompass both a wide range of prokaryotic groups, as well as significant depth of sampling within taxa could begin to answer many questions about the evolution and relatedness of these organisms. Prokaryotic genome sequencing is coming down in price, effort, and time commitment, but it is still difficult to justify the expense of sequencing many strains from a particular group, especially if there are no economic or medical benefits. Many strains have been sequenced, for example, of the well known *E. coli* or *Salmonella* groups, as well as *Borrelia*, the causative agent of Lyme disease, and *Helicobacter pylori*, implicated in ulcers. Genome sequences of several related organisms can give a better picture of strain-specific genetic content, which may be important to the biology of the organisms, as in the case of pathogenicity islands present in virulent strains of *E. coli* but absent from typical avirulent laboratory strains. While useful for assessing pathogenic capabilities and evolution, sequencing the genomes of many closely related pathogens does not contribute to the depth of coverage of distantly related or non-pathogenic organisms.

When the complete genome sequence is only available for one member of a particular prokaryotic group, it is a good starting point, but it may be said that it is only useful for finding out about the biology of that one organism. It may be difficult to predict, outside of the basic informational genes, whether that strain is representative of the group, or which regions of its genome may be adaptive to its own particular environment, pathogenic lifestyle, or metabolic specialty.

If there is at least one completely sequenced genome representing a bacterial group, we can use various methods to estimate the strain-specific genes present in any other related organism, but these are estimates and can only apply to that one member. For example, Nesbø and colleagues (Nesbø *et al.* 2002) used suppressive subtractive hybridization, which is a PCR-based method that uses genomic DNA from two organisms, one of which has been sequenced, to enrich for strain-specific sequences in the unsequenced tester strain. As mentioned previously, using the genome sequence of *T. maritima* MSB8 (Nelson *et al.* 1999) as a reference, they determined that *T.* sp. RQ2, which differs only by 0.3% from the completely sequenced MSB8 strain in its SSU rRNA sequence, contained upwards of 20% strain-specific genes, particularly unique sugar metabolism genes. Until and unless we completely sequence the genome of many other strains, we won't know if these particular functions are (a) unique to *T.* sp. RQ2 or (b) simply missing from *T. maritima* MSB8, but present in other members of the Thermotogales. We need more than one or two sequenced genomes to give a fair picture of the diversity of any given group, or its evolutionary origins, even if we have SSU rRNA sequences from many members.

1.8: The Species Genome Concept

The complete genome sequence of any given organism is just that – the sequence for that organism, not of the species, or even necessarily the strain. In 2000, Lan and Reeves introduced the 'species genome concept', which refers to all the DNA important for a species as a whole, which is distinct from the genome of any given individual (Lan and Reeves 2000). As was found previously (Nesbø *et al.* 2002), there can be huge regions of a genome found either in all but one of the known strains of a species, or unique to one member only. For example, the species genome of *E. coli* may be said to contain a wide variety of environmental and pathogenicity islands, enabling different strains to colonize and infect different hosts or niches (Bingen-Bidois *et al.* 2002; Dobrindt *et al.* 2002; Hejnova *et al.* 2005; Kao *et al.* 1997). The *Thermotoga maritima* species genome could be said to include a wide array of sugar metabolism genes and capabilities, which depend on where the organism lives, as well as both bacterial- and archaeal-type ATPases and *mutS* homologs. Most members of the species would only require one or two of the genes available for each function, but may harbour multiple copies that have been acquired by LGT and have not been degraded yet, even though they may or may not be expressed. The 'species genome' allows inclusion of all sequence data at the present time available within a particular group, and can give an overall view of the biology of that group, but it cannot definitively predict what will be found in future sequenced genomes.

1.9: Metagenomics

The inability to culture the diversity of microorganisms that are sometimes seen under the microscope has led to the study of "metagenomics"; the metagenome is the collective genetic information from a given environment. Total environmental DNA is isolated and sequenced, in small plasmids or larger cosmids, fosmids, or BAC libraries, and the organisms present are extrapolated from the gene identity found (often SSU rRNA sequences are used as a starting point). Perhaps the most famous metagenome to date was created from the Sargasso Sea (Venter *et al.* 2004), where 1.045 billion basepairs of sequence were assembled into scaffolds, and the authors determined that there were approximately 1800 genomic species, and 1.2 million previously unknown genes present in this environment (which, ironically, was chosen because it is nutrient-poor, and therefore likely to harbour less diversity). The usefulness of sequencing billions of basepairs of environmental DNA must be there, but is hidden and must be carefully sought after. Venter and colleagues sought to assemble complete genomes from hundreds of genomes worth of DNA sequence fragments, but perhaps this is not a reasonable expectation of this type of dataset. It is possible that huge stretches of DNA, found in areas with unexpectedly high levels of biodiversity for a low-productivity area of the ocean, may belong to several distinct species or populations. For example, sequence islands may be shared by different populations or species, and behave as artifactual anchor points for assembly of chimeric genomes.

Such metagenomic projects are useful in that they may tell us about the types of metabolic processes and lifestyles that predominate in any one environment, but they may also miss a great deal of information. With a finite number of sequences being generated, rare but essential metabolisms may be missed when amassing the metagenome.

1.10: Operons and Gene Order

Another feature of genomes that can be used to assess the relatedness of strains or species is the presence/absence of gene clusters or operons, and the order of genes within a given cluster. While gene order is seen to be consistent at some level, particularly among closely related organisms (Casjens *et al.* 1995; Ojaimi *et al.* 1994), prokaryotic genomes (operons and gene order) are known to be relatively plastic, and unstable over long periods of time and evolutionary distance (Itoh *et al.* 1999; Watanabe *et al.* 1997). Even in highly regulated systems, where expression of component proteins is tightly controlled, operon shuffling can occur.

Some state that gene order, particularly within operons, is conserved in closely related species and/or in certain regions of the genome (Casjens *et al.* 1995; Ojaimi *et al.* 1994). Clustering may be important to maintain if the expression of component gene products must be tightly controlled, but even in the case of established operons, gene order does not seem to be an essential feature. Upon examination of gene pairs, interesting patterns can be seen (Dandekar *et al.* 1998); while overall gene order shows little to no conservation, proteins encoded by conserved gene pairs do tend to interact physically, implying that functionality is assisted by proximity within the genome.

For example, the assembly of the flagellar apparatus is tightly regulated, and the operons involved are regulated in a hierarchical fashion (Macnab 2003). Genes for components required at a particular stage of assembly may shuffle around within that particular cluster, but the expression of gene clusters corresponding to these stages is often sequential, and corresponds to the actual construction of the apparatus from the membrane outward. The regulon-type expression of the operons results in proximal proteins being expressed and laid down first, which then form an export-type apparatus that exports the gene products of the distal operon, which are then assembled nearer to the outside of the membranes.

Gene order has utility when examining organisms that are closely related – in the same family or order – but genome plasticity will likely erode important characteristics beyond that. The conservation of order, or of a common gene inserted within a cluster in a number of strains can reveal common history or a shared habitat, and could lead to further study of the gene cluster or inserted genes, as well as the species or groups implicated. The following examples serve to illustrate this phenomenon.

1.10.1: Flagellar Gene Clusters in Prokaryotes

Synthesis of the bacterial flagellum is a complex and metabolically expensive process (Macnab 2003), which involves dozens of proteins that must interact in both homo- and heterocomplexes. It is structurally different from both the archaeal and eukaryal flagella as well. In order to ensure proper assembly order and stoichiometry of the components, genes are often organized into both operons and larger regulons (i.e. operons that are expressed in a sequential fashion). Because of this tight regulation, in species where flagella are present, it could be thought to be a complex system resistant to recombination or lateral gene transfer. However, if mutations or recombinations do not negatively affect protein-protein interactions, replacements might be accommodated in order to ensure flagellar assembly and function if motility is essential for the strain. Operon rearrangements also might be permitted, provided they maintain the proper expression order and stoichiometry.

1.10.2: Ribosomal Protein Gene Clusters in Prokaryotes

Because functionally related genes sometimes tend to cluster together on a genome more often than unrelated genes (Tamames *et al.* 1997), ribosomal proteins would be a likely group to be found in clusters, or close proximity on a genome. Many gene clustering studies as well as ribosomal structural examinations have been done in prokaryotes (Ban *et al.* 2000; Yusupov *et al.* 2001). For instance, a fairly thorough examination of prokaryotic diversity in several clusters (S10, spc and alpha ribosomal protein gene clusters) was completed recently; while the content of the clusters (presence within the cluster, presence elsewhere in the genome, complete absence) is somewhat plastic, the order within the clusters themselves is fairly conserved. However, detailed phylogenetic analysis of several proteins does indicate that LGT was a factor in their evolution (Coenye and Vandamme 2005). Also, there seems to be precedent for ribosomal protein gene clusters being interrupted by non-ribosomal protein genes. Previous work has indicated that ribosomal proteins S14 and L27 have been transferred horizontally, thus further complicating the evolution of these clusters (Brochier *et al.* 2000; Garcia-Vallve *et al.* 2002; Makarova *et al.* 2001; Matte-Tailliez *et al.* 2002).

The proteins of interest in this project have had their homologs mapped on the large subunit in *E. coli* (Yusupov *et al.* 2001). The cluster proteins (L13, L21, and L27) are structural proteins, on the backside of the subunit, and do not directly contact the small subunit.

Ribosomal proteins have been used in the past in concert with SSU rRNA for universal phylogenies, with varying success. Matte-Tailliez *et al.* (Matte-Tailliez *et al.* 2002) used a concatenation of 53 ribosomal proteins to create a universal archaeal phylogeny, and found that only 8 of those proteins had phylogenetic signal that contradicted that of the SSU rRNA tree (implying that they were acquired or exchanged by LGT). The 45 concordant proteins supported the SSU relationship quite well, and the LGT seemed to be biased to organisms that live in the same environment. Brochier *et al.*

(Brochier *et al.* 2000) examined a single conserved protein in bacteria (RpS14) and its genomic context, and found that LGT did indeed take place. Interestingly, this protein is near the peptidyl transferase centre in the fully assembled ribosome, and according to the complexity hypothesis should be resistant to transfer, owing to its multiple (protein and RNA) interacting partners. This gene showed both transfer between organisms, as well as operon or gene order shuffling within single genomes.

Ribosomal proteins have also been transferred together with operational genes, as was shown by Garcia-Vallve *et al.* (Garcia-Vallve *et al.* 2002). Protein L27, the homolog of which is of interest to us in *T. maritima* MSB8, appears to have been transferred into *Arthrobacter* sp., an Actinomycete, from an unknown *Bacillus* species, along with a cluster of six genes responsible for creatinine and sarcosine degradation. While L27 would be a non-transferrable gene by the complexity hypothesis, amino acid degradation gene clusters are not, and are ideal for transfer in that they can confer a new metabolic function on the recipient lineage.

A study of a much more extensive area of the genome was done with only two strains, *Sinorhizobium meliloti* and *Bacillus subtilis* (Barloy-Hubler *et al.* 2001). The authors indicate that three different assessments (DNA, amino acid, and gene order/organization) give divergent results. A similar pattern of clustering is found in these two species; however, the authors claim that it is as a result of functional convergence and not any phylogenetic relationship. Work done by Klein and colleagues (Klein *et al.* 2004) compares the structure of the large ribosomal subunit from both *Haloarcula marismortui* (an archaeon) and *Deinococcus radiodurans* (a bacterium). They found that by looking at the structure and location of the proteins making up the large subunit, one can see many cases of molecular mimicry and functional convergence, where different proteins can be used successfully to stabilize identical RNA structures. If the proteins themselves are there solely to maintain structural integrity, then it is more the characteristics of the protein rather than a precise sequence that is necessarily conserved. This theory would then allow for more flexibility in terms of transfer, provided the key contact properties were maintained (e.g. glycine/ arginine/lysine rich regions that occur quite frequently in extensions of the protein structure, as opposed to the globular proteins that are rich in alanine, valine and aspartate (Klein *et al.* 2004)).

1.11: ORFans in Microbial Genomes

When a new genome is sequenced, a large percentage of the open reading frames (ORFs) are unique in the database (that is, have no sequence match in the available databases) although they may be present in more than one genome or strain. These ORFs have been termed ORFans (Fischer and Eisenberg 1999). ORFans are a source of untapped, and possibly difficult to assess, diversity in bacterial genomes. They can represent anywhere from 0-60% of genes in a genome, depending on its similarity to previously sequenced genomes. ORFans that are conserved in a select group of strains or species, or those found in several distantly related bacterial groups, can provide interesting puzzles as to their origins, as well as the evolutionary implications when considering distant relatives that share ORFs of unknown function that were previously thought to be ORFans within one genome.

ORFans fall into several different categories: (1) singleton ORFans, or true ORFans (true hypothetical proteins THP), which have no match anywhere; (2)

orthologous ORFans, often known as conserved hypothetical proteins (CHP), have matches in other genomes but not to any sequence with known function; (3) paralogous ORFans, with matches to other ORFs in the same genome, but not to ORFs in other genomes; and, (4) ORFan modules, which are unmatched regions within proteins of known function, and may represent domains with a novel function (Siew and Fischer 2003a). ORFan populations in the databases show unique dynamics; the overall number of ORFans is still increasing with the addition of prokaryotic genomes, but at a lower rate than the number of ORFs in total. As more diverse strains are added, singleton ORFans find matches and are either placed in orthologous ORFan families, or are assigned function (Siew and Fischer 2003a). However, each new genome does add new ORFans to the database, making it difficult to determine when, if ever, the percent of ORFans out of total ORFs will decrease or reach a stable level.

ORFans present both a unique puzzle and a potentially useful tool. While it is likely that a significant proportion of ORFans are missannotated sequence or junk DNA that has no function, some will represent unique functions and genomic signatures, particularly orthologous ORFans that are found in only a select few genomes (Siew and Fischer 2003a). As previously stated, gene order and/or operon conservation can sometimes be used to successfully predict the function of ORFans (Wolf *et al.* 2001).

A detailed assessment of ORFans was completed for several *E. coli* strains and related bacterial groups (Daubin and Ochman 2004), and the authors concluded that most of the ORFans present had a traceable history, and did in fact perform functions within the genome. The source of the ORFans, in this case, was thought to be bacteriophage. The ORFans in the *E. coli* lineage had several features that distinguished them from genes ancestral to the gamma-proteobacteria and other sporadically distributed genes; the majority are short, A-T rich and fast evolving, perhaps implying that they are derived from bacteriophage genes that establish themselves in the genome by adopting roles in cellular functions.

Shared ORFans within a group of strains (related or unrelated) implies a relationship among those organisms, whether it be phylogenetic or environmental, and can be used to tease out shared histories. A conserved intact orthologous ORFan would suggest that the ORF/gene is coding and functional, and provides a beneficial function to the organism; examination of codon usage and evolutionary rates of it and its flanking genes can reveal its history within the group of organisms.

1.12: Functional Biogeography in Microbial Genomes

Assessment of the maintenance of gene clusters can be problematic considering the plasticity of most prokaryotic genomes. Rearrangement of functional operons can be seen, and oftentimes synteny is destroyed even within groups of what are thought to be closely related strains. However, if the genomes themselves are examined not as a single organism, but a community of interacting units (genes), a more general, global assessment of gene order is possible, using methods most often reserved for biogeographical studies. Spatial autocorrelation analyses are used to determine the level at which members of a community interact, and if the interaction is biologically and statistically significant. These are typically used to look at the biogeography of higher organisms, including but not limited to badger, salmon, beetle, deer, soybean, eucalyptus, and pine (Epperson and Allard 1989; Jones *et al.* 2007; Kuehn *et al.* 2007; Kuroda *et al.*

2006; Pope *et al.* 2006; Primmer *et al.* 2006; Schmuki *et al.* 2006). Such analyses have been applied to prokaryotes, to determine community structure in salt marshes (Franklin *et al.* 2002), *Pseudomonas* communities in soil (Cho and Tiedje 2000), general arable soil communities (Nunan *et al.* 2001; Nunan *et al.* 2002) and agricultural fields (Franklin and Mills 2003). Biogeographic analyses examine interactions in two- and three-dimensional space, but spatial autocorrelation methods have also been used within single genomes to assess clustering of mutations, though typically in human and mouse genomes (Firneisz *et al.* 2003; Gaffney and Keightley 2005; von Grunberg *et al.* 2004). This can localize mutationally active regions of a genome, or mutational islands. In each of these cases, however, assumptions of distribution, population size, and coverage must be made, which can affect the outcome of the analysis.

While genes within prokaryotic genomes are typically grouped into categories during annotation, any information on the physical distribution of different categories is limited to comparison to conserved operon and regulon structures found in other organisms, along with other stretches of syntenic genes. Once gene order synteny is lost, however, clustering cannot be easily seen, nor can other types of gene distribution. However, if prokaryotic genomes are considered analogous to one-dimensional geographic features, such as coastlines, it could allow the assessment of any physical clustering or unusual distribution of functionally similar genes.

1.13: Project Rationale

Two areas of the genome of *T. maritima* MSB8 were targeted, for both their commonalities and their differing properties. Both regions of the genome constitute clusters, as opposed to single genes. The first project examines a cluster of flagellar genes within the Thermotogales. Flagellation for motility can be thought of as non-essential, depending on the environment in which a particular species lives, but each flagellar component gene could be considered essential when the system is taken as a whole (i.e. a partial flagellum is not useful or functional, and the system is a good example of "irreducible complexity"). At the 3′ end of the cluster, a single hypothetical ORF is present, and conserved in the strains that contain an intact cluster. Thermotogales are anaerobic, and therefore do not need motility to access oxygen, for example, and motility itself is not a characteristic of all members of this order (Reysenbach 2001). The second project examines a cluster of ribosomal protein genes. These gene products, and the resulting ribosome structure is indeed essential to life, but the individual components can be functional even without all members present. Structural ribosomal proteins serve as a scaffold for the ribosome, and homologous proteins are not always universally present across all domains of life (Ban *et al.* 2000; Barloy-Hubler *et al.* 2001; Klein *et al.* 2004; Matte-Tailliez *et al.* 2002; Willumeit *et al.* 2001; Yusupov *et al.* 2001).

The first project, long walk PCR downstream of locus TM1363, or *prfA*, as identified in *T. maritima* MSB8, successfully combines two ideas. Firstly, *prfA* provides an essential function, as a protein release factor in the process of protein translation, and would thus be important and resistant to change in and around it, genomically. Secondly, this gene sits immediately upstream of a cluster of flagellar genes, which frequently form clusters in prokaryotes and are highly regulated.

The second project, which involves characterization of a cluster of ribosomal protein genes, combines examination of a potentially conserved cluster of informational genes with the existence and location of potential ORFans. The cluster of three ribosomal protein genes (L21, L27 and L13) present in MSB8 is interrupted by two separate hypothetical ORFs; the first is a conserved hypothetical ORF, which is present in several sequenced genomes but has no known function, while the second is a true hypothetical ORF (or singleton ORFan), with no detectable similarity to anything else in the databases. Through protein structure prediction and database searching, as well as the discovery of this ORF in other Thermotogales, we may be able to establish its function and distribution in these organisms.

The third project involves the maintenance of functional clusters within prokaryotic genomes. The genome of *T. maritima* MSB8 was analyzed, and is being used as the comparison point for all of our studies, but it is the only completely sequenced member of the Thermotogales at present. Analyses of several different groups of closely related strains (*Bacillus anthracis*, *Campylobacter jejuni*, *Chlamydia pneumoniae*, *Escherichia coli*, *Legionella pneumophila*, and *Prochlorococcus marinus*) were completed to determine the conservation of potential ORFans in their genomes.

1.13.1: Flagellar Gene Clusters in the Thermotogales

PrfA, which codes for a protein release factor used in protein translation termination, is located approximately 3/4 of the way through the genome of *T. maritima* MSB8, from the origin of replication. This gene was chosen as the anchor point for walking PCR studies (modified from (Katz *et al.* 2000)). In *T. maritima* MSB8, *prfA* is flanked by flagellar genes, and of particular interest is the downstream cluster, which consists of three flagellar genes (which code for proteins in the proximal region of the flagellum – Figure 1.2) and a conserved hypothetical ORF. At first glance, the cluster is present intact in all the strains studied from the *maritima/neapolitana* clade.

Phylogenetic analysis of the individual genes, or ORFs, indicates a potential recombination with the more distantly related strain, *T.* sp. Kol6, but a sliding window analysis (data not shown) revealed a more complicated pattern of recombination. Recombination amongst the four closest SSU rRNA relatives, *T. maritima* MSB8, *T.* sp. RQ2, *T. naphthophila* RKU10 and *T. petrophila* RKU1, supports a second branching pattern, placing MSB8 with RKU10 and RQ2 with RKU1. Intra-strain recombination, which would normally be missed with traditional presence/absence analyses, is evident here, indicating that common ancestors of these strains, which live in very different environments, may have been in contact. As well, not all members of our strain collection are known to be flagellated and motile; if motility is not an essential function, it was likely lost quite recently, as in the close relatives, the entire cluster is present, intact, and the changes are conservative. Interestingly, the sister clade to the *maritima/neapolitana* clade also contains a flagellar gene cluster immediately downstream of *prfA*, but it is a completely different set of genes, which in *T. maritima* MSB8 are located 165 kbp downstream. Of the two strains, *T. thermarum* is known to be motile, but the motility of *T. subterranea* has not been determined. However, the ORFs seem to be intact, albeit very distantly related to the MSB8 genes.

The reasoning for the tendency of *prfA* to neighbor flagellar genes or gene clusters is not known. Expression of flagellar proteins is known to be highly regulated, both in order of expression and stoichiometry of the components, so perhaps the presence of bits of the transcription/translation machinery in close proximity on the genome helps

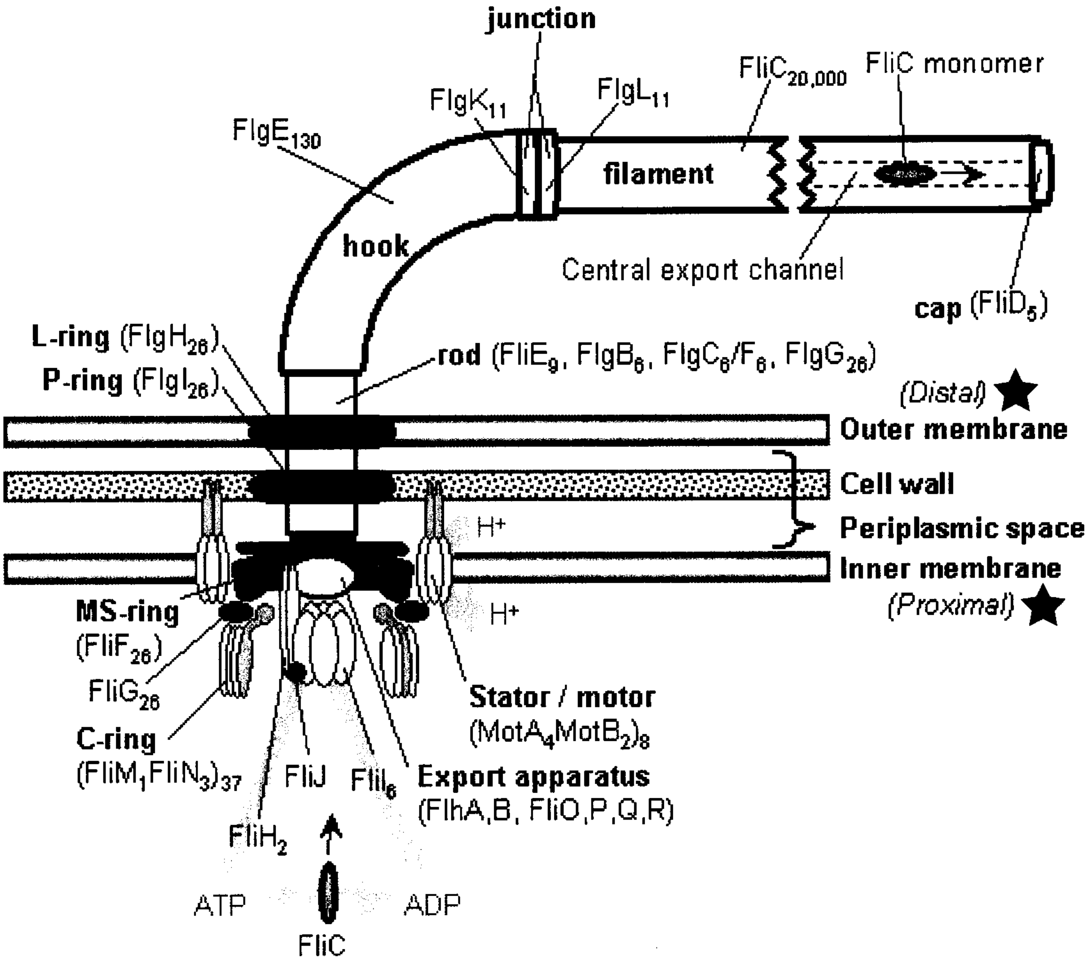


Figure 1.2 Illustration of the flagellar apparatus in bacteria. This illustration was taken from http://www.talkdesign.org/faqs/flagellum.html). The proximal region is indicated by a red star, and the distal region is indicated by a blue star. Proximal and distal are used to refer to the location of the gene product relative to the cytoplasm of the cell.

to ensure proper and timely assembly. To complete this analysis, I will be using conventional PCR to assay our strain collection for the presence of this second operon, as well as potential recombination events.

1.13.2: Ribosomal Protein Gene Clusters in the Thermotogales and in Other Prokaryotes

A second gene cluster was targeted for analysis, based on its occurrence in the MSB8 genome. The ORF cluster at loci TM1454-TM1458 consists of three ribosomal protein genes, a conserved hypothetical ORF and a true hypothetical ORF. We are hoping to take advantage of the presence of the potential ORFans (TM1455 and TM1457) to investigate the history of this cluster in our strain collection. The presence and conservation of these ORFs, along with protein structure prediction and domain analysis, can elucidate their function, and possibly determine their origin, whether they were vertically inherited, or recombined into the ancestral genomes. The first two ribosomal proteins, TM1454 and TM1456, respectively, code for two structural proteins (L13 and L27), and in most prokaryotic genomes, these two proteins are found flanking one another, or very close in the genome; in the case of *T. maritima* MSB8, they are interrupted by a conserved hypothetical protein, TM1455. The third ribosomal protein, TM1458 (L21), is also a structural protein, but is most often found further away in any given prokaryotic genome, and is not as highly conserved. This thesis investigates the evolutionary history and conservation of L13, L21 and L27 in the Thermotogales, as well as other ribosomal protein genes in other prokaryotes.

1.13.3: Spatial Autocorrelation of Functional Gene Categories, and ORFans, Within *Thermotoga maritima* MSB8, and Other Groups of Closely Related Strains of Bacteria

Identifying clusters of functionally similar genes can be difficult because of the plasticity of prokaryotic genomes. Using a simple spatial autocorrelation analysis, the distribution of different functional categories of genes can be evaluated for any prokaryotic genome. Here, several groups of closely related organisms were evaluated, concentrating on two types of ORFans that are annotated as functional categories within the TIGR Comprehensive Microbial Resource: conserved hypothetical proteins (also known as orthologous ORFans) and true hypothetical proteins (also known as singleton ORFans). The type of distribution of these two categories varies between different strain groups, and in several cases, extreme hyperdispersion of the ORFs can be seen, indicating possible misannotation in the database. Any significant clustering of ORFans, whether conserved or true hypothetical proteins, may also indicate islands of transferred genes, or islands of novel function.

1.13.4: Evaluating Higher Order Genomic Structure in the Face of Lateral Gene Transfer and Genome Rearrangements

Lateral gene transfer, along with the plastic nature of prokaryotic genomes, can serve to reduce or eliminate any higher order architecture within the genomes. However, because operon structures are conserved to some degree, even with internal shuffling, there would seem to be a force maintaining a level of genomic structure or framework. By examining specific gene clusters or operons that are conserved among a diverse group of prokaryotes, even when involved in various lateral gene transfer or recombination events, one can determine if such a framework exists.

In addition to examining specific clusters in detail, generating physical maps of functional gene categories found in prokaryotic genomes can also assess higher order architecture. Such analyses can also give insight into the distribution and possible function of open reading frames having no known function.

CHAPTER 2: MATERIALS AND METHODS

2.1: DNA Acquisition

Genomic DNA for 20 Thermotogales strains, as presented in Figure 1.1, was obtained from various sources (Table 2.1). For those strains available as cell mass, genomic DNA extractions were performed, using a modification of (Charbonnier and Forterre 1994). The protocol was scaled down for small quantities of cell mass, and was done as follows.

2.1.1 DNA Extraction

A small pellet of cell mass, approximately 100 µL in volume, was placed into a 1.5 mL microcentrifuge tube, and resuspended in 800 µL of TNE at pH 7.5. To this mixture, 100 µL of N-lauroylsarcosine was added, and the tube was inverted several times to mix. Then, 100 µL of 10% SDS was added and the tube was inverted to mix. To this 50 µL of a 20 mg/mL proteinase K solution was added, and incubated at 50°C, taped to the mechanism of a rotating hybridization oven, for 3 -12 h. Five µL of RNase was added, and the solution incubated for 1 h at 37°C.

The solution was then transferred, in two equal parts, to two fresh microcentrifuge tubes, and each tube was treated as per the following: TE-phenol, 650 µL, was added, and the solution was agitated at 37°C for 10 min, spun in a microcentrifuge at room temperature for 5 min, and the aqueous phase was transferred to a fresh microcentrifuge tube. The TE-phenol treatment was repeated an additional two times.

Table 2.1 Thermotogales strains studied. DNA and bacterial cell masses used in this thesis were kind gifts from Dr. Camilla Nesbø, Dr. N. Glansdorff , Dr. H. Morgan, Dr. K.O. Stetter, and Dr. Yoh Takahata. In cases were cell mass was used, the DNA was extracted as per the method of Charbonnier and Forterre (Charbonnier and Forterre 1994) (Please see Section 2.1.1).

Strain	Source
Thermotoga	
T. maritima MSB8T	DNA from Dr. Camilla Nesbø, Dalhousie University
T. maritima SL7	DNA from Dr. Camilla Nesbø
T. maritima FjSS3B1	Bacterial cell mass from Dr. K.O. Stetter, University of Regensburg, Germany
Thermotoga petrophila RKU1	DNA from Dr. Yoh Takahata, Taisei Research Institute, Japan
Thermotoga naphthophila RKU10	DNA from Dr. Yoh Takahata
Thermotoga sp. RQ2	Bacterial cell mass from Dr. H. Morgan, University of Waikato, New Zealand
Thermotoga sp. RQ7	Bacterial cell mass from Dr. H. Morgan
Thermotoga sp. SG1	Bacterial cell mass from Dr. H. Morgan
Thermotoga sp. kol 6K	Bacterial cell mass from Dr. H. Morgan
T. neapolitana LA4	Bacterial cell mass from Dr. H. Morgan
T. neapolitana LA10	Bacterial cell mass from Dr. H. Morgan
T. neapolitana NS-ET	Bacterial cell mass from Dr. H. Morgan
T. thermarum LA3	Bacterial cell mass from Dr. Camilla Nesbø
T. subterranea SL1	Bacterial cell mass from Dr. Camilla Nesbø
Thermosipho	
T. africanus Ob7	Bacterial cell mass from Dr. H. Morgan
Fervidobacterium	
F. islandicum H12	Bacterial cell mass from Dr. H. Morgan
F. nodosum	DNA from Dr. Camilla Nesbø
Petrotoga	
P. miotherma	DNA from Dr. Camilla Nesbø
P. mobilis	DNA from Dr. N. Glansdorff, University Libre de Bruxelles, Belgium

A solution of chloroform:isoamyl alcohol (24:1 ratio) was added to the aqueous phase at a volume of 650 μL, and agitated for 10 min at 37°C, spun in a microcentrifuge at room temperature for 5 min, and the aqueous phase transferred to a fresh microcentrifuge tube. The chloroform:isoamyl alcohol treatment was repeated one additional time.

At this point, two volumes of 100% ethanol were added to the aqueous phase, the tube was inverted several times, and incubated at -20°C for 1 h. The tube was then spun for 5 min in a microcentrifuge and the ethanol removed. The pellet was then washed in 500 μL of 70% ethanol, spun for 5 min at room temperature, and the ethanol removed. This 70% ethanol treatment was repeated one additional time, and the pellet left to air-dry for 1 h. The resulting pellet was resuspended in a Tris-Cl solution.

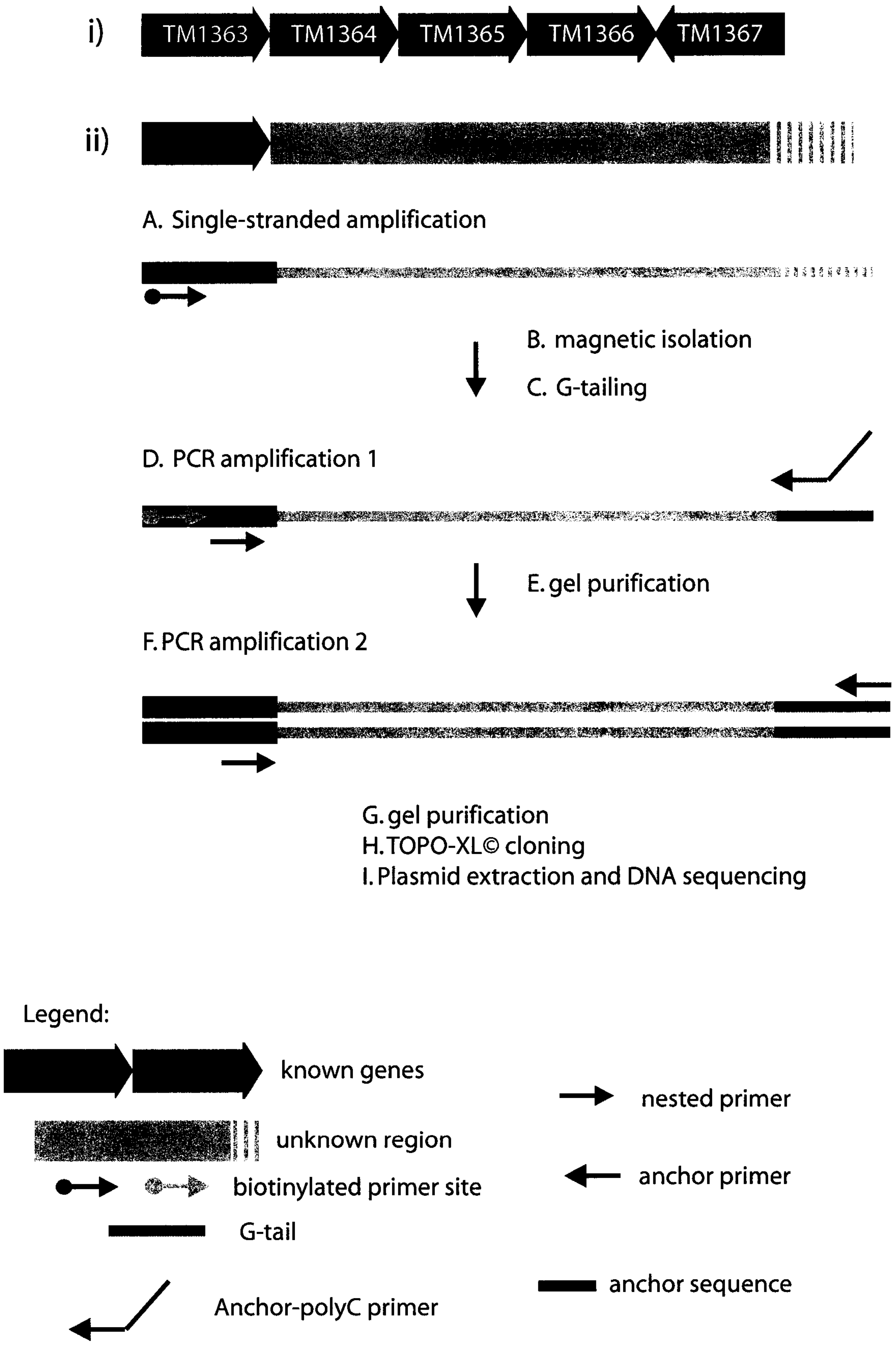
2.2: Long Walk PCR of Proximal Flagellar Cluster (PFC)

A modification of the long walk PCR protocol of Katz *et al.* (Katz *et al.* 2000) has been outlined in Figure 2.1, and used as follows.

2.2A: Linear Amplification

A single degenerate primer, of sequence 5′-AAGTTCTTTCTCRTAYTTYTC-3′ was designed to the 3′ end of the *prfA* gene (TM1363) in *Thermotoga maritima* MSB8. This primer was biotinylated at the 5′ end to facilitate magnetic isolation of amplification products, using streptavidin-coated beads. An initial linear amplification was performed, using 40 to 80 ng of genomic DNA in 100 μL total reaction volume, containing the

Figure 2.1 Long Walk PCR analysis of the PFC region in the Thermotogales. A schematic representation of the method, based on Katz et al, 2000 (Katz *et al.* 2000). i) ORFs found within *T. maritima* MSB8 The gene for *prfA* is indicated in black, and serves as a conserved priming site. ii) Potential genomic context within tester strains. A portion of *prfA*, encompassing the 3′ end of the gene, is shown in black, while the unknown downstream flanking region is shown in grey. A) A linear amplification, or primer extension, is performed using a biotinylated primer, designed near the 3′ end of *prfA*. B) The single stranded DNA product is then isolated using streptavidin-coated paramagnetic beads, which bind to the biotinylated primer. C) The single-stranded product is G-tailed. D) PCR Amplification 1 is performed, using a nested primer, designed six nucleotides downstream of the priming site of the biotinylated primer, and a poly-C primer with an anchor sequence. E) The double stranded DNA product of PCR amplification 1 is gel-purified for use as template. F) PCR amplification 2 is performed, using the products cleaned in E) as template, with the nested primer and an anchor primer identical to the anchor sequence of the poly-C anchor primer. Products are then G) gel purified, H) cloned and I) sequenced.


i)
TM1363
TM1364
TM1365
TM1366
TM1367
ii)
A. Single-stranded amplification
B. magnetic isolation
C. G-tailing
D. PCR amplification 1
E. gel purification
F. PCR amplification 2
G. gel purification
H. TOPO-XL© cloning
I. Plasmid extraction and DNA sequencing
Legend:
known genes
unknown region
biotinylated primer site
G-tail
Anchor-polyC primer
nested primer
anchor primer
anchor sequence

following components: 1 U Platinum Taq Hi-Fidelity DNA polymerase (Invitrogen Cat. No. #11304-029), 10 µL Hi-Fi buffer, 400 µM dNTP mix, and 100 mM primer. The reaction was performed in a thermocycler with a temperature profile of 3 min of initial denaturation at 94°C, followed by 35 cycles of 30 sec denaturation @ 94°C, 30 sec annealing @ 43°C, and 5 min elongation @ 72°C, followed by a final elongation of 10 min @ 68°C. The resulting single stranded products were then immediately bound to streptavidin-coated paramagnetic beads (Promega Cat. No. Z5481) for magnetic isolation as follows.

2.2B: Magnetic Isolation

One tube of bead suspension was used per 4 PCR reactions, and cleaned using the manufacturer's instructions (Cat. No. Z5481), resuspended in 100 µL of 0.5X SSC, and aliquoted into four 1.5 mL tubes (25 µL per sample). Each 100 µL completed linear reaction was then added to a tube with beads, incubated at 37°C for 10 minutes with shaking, and washed 3X with 0.1X SSC and 1X with 100 µL of 1X Tdt buffer (Promega Cat. No. M1871) to prepare for G-tailing.

2.2C: G-tailing of Single Stranded Products

Beads bound to the linear product were then mixed with 2 µL of 5X Tdt buffer (Promega), 5 µL of 20 µL dGTP, 4 µL of ddH_2O, placed in a 70°C water bath for 15 sec, then mixed with 1 U of TdT enzyme. The reaction mixture was incubated for 1-3 h at 37°C with shaking, then stopped by adding 2 µL of 0.5M EDTA and incubating at 65°C. The beads were captured on the magnet, washed twice with Tris-EDTA buffer ("TE" buffer, QIAquick Gel extraction kit, Cat. No. 28704) and resuspended in 20 μL of TE, to serve as template for the first true PCR amplification.

2.2D: PCR Amplification 1

Five μL of G-tailed, single stranded product was used as template for the first true PCR. The primers used were a nested degenerate primer of sequence 5′-CAGTTCATTTTCNCCYTCYTC-3′, designed 6bp downstream of the biotinylated primer (for specificity) and an anchor-polyCytosine primer of sequence 5′-CCACGCGTCGACTAGTAATTCCCCCCCCCCCCDN-3′. The poly-C portion anneals to the G-tail of the single stranded template, while the dinucleotide, DN, serves to target the primer to the joint between the original single-stranded product and the G-tail (Figure 2.1D). Each 100 μL reaction mix consisted of 5 μL of G-tailed single stranded template bound to streptavidin-coated beads, 1 U Platinum Taq Hi-Fidelity DNA polymerase, 10 μL Hi-Fi buffer, 400 μM dNTP mix, 100 mM of each primer, and ddH_20 to a final volume of 100 μL. The temperature profile was as follows: initial denaturation of 3 min @ 94°C, followed by two stages of amplification. The first stage had 15 cycles of 30 sec denaturation @ 94°C, 30 sec annealing @ 43°C, 5 min elongation @ 72°C; the second had 10 cycles of 30 sec denaturation @ 94°C, 30 sec of annealing at @ 55°C, 5 min of elongation @ 72°C and a final elongation of 10 min @ 68°C .

2.2E: Gel Purification 1

Amplification products from PCR Amplification 1 were electrophoresed on a 1% agarose gel containing crystal violet (Invitrogen TOPO®-XL cloning kit, Cat. No.

K4750-20), in TAE buffer at 60 V for approximately 30 min, so that molecular markers were resolved in the region of 2-10 kbp. Because the template for this reaction consists of single-stranded products of various sizes, the gel shows a very faint, sometimes almost imperceptible, smear. A wide gel slice corresponding to the region of 2-10 kbp was cut from the gel with a sterile razor blade, and the DNA extracted via the Qiagen Min-Elute protocol (Cat. No. 28004). Briefly, the gel slice was weighed in a microcentrifuge tube, and 3 volumes of buffer QG were added to the tube. The slice was then melted at 50°C for 10 min, with occasional inversion. Isopropanol was then added (one gel volume) and the solution mixed by inversion. This solution was then applied to the Min-elute column, set in a 2 mL collection tube, and spun in a microcentrifuge for 1 min. The flow through was discarded, the column washed with 750 μL of PE buffer, and spun for 1 min. The wash buffer was discarded, and the column spun for 1 min to dry. The DNA was then eluted by applying 10 μL of EB buffer to the column membrane, placing the column into a clean 1.5 mL microcentrifuge tube, letting it sit for 1 min, and then spinning in a microcentrifuge to collect the eluant.

2.2F: PCR Amplification 2

A second full PCR amplification was performed, using gel-purified PCR products from the first amplification as template. Also, the primer pair used differed in the anchor primer sequence – the poly-C portion was left out, leaving a primer of sequence 5′-CCACGCGTCGACTAGTAATT-3′. Reaction volume was reduced to 50 μL, and contained 5 μL of template, 0.2 U Platinum Taq Hi-Fidelity DNA polymerase, 5 μL Hi-Fi buffer, 400 μM dNTP mix, 100 mM of each primer, and ddH_20 to a final volume of 50 μL. The temperature profile used was as follows: an initial denaturation of 3 min @ 94°C, followed by 35 cycles of 30 sec denaturation @ 94°C, 30 sec of annealing at @ 55°C, 5 min of elongation @ 72°C and a final elongation of 10 min @ 68°C . Products were run out on a gel as per Section 2.2E (Gel purification 1), and the purified products used for TOPO®-XL cloning.

2.2G: Gel Purification 2

Products were purified as per section 2.2E, and used for ligation and cloning.

2.2H: TOPO®-XL Cloning and Transformation

Gel-purified PCR products were cloned into the TOPO® XL vector, following the manufacturer's instructions. Briefly, 4 μL of gel-purified PCR product was incubated with 1 μL of the pCR®-XL-TOPO® vector for 5 min at room temperature, then 1 μL of the provided 6X TOPO® Cloning Stop solution was added, and the solution mixed and placed on ice. One Shot ® TOP10 chemically competent *E. coli* cells were then transformed by the addition of 2 μL of the ligated vector solution, followed by incubation on ice for 30 min, heat shocking at 42°C for 30 sec, and incubation in ice for 2 min. S.O.C. medium, provided in the kit, was added at a volume of 250 μL, and the cells recovered with gentle shaking at 37°C for 1 h. Two volumes of cell suspension, 50 μL and 100 μL, were plated onto separate LB plates containing 50 μg/mL of kanamycin. Plates were incubated overnight at 37°C, and colonies picked for subsequent plasmid extraction and sequencing.

2.2I: Plasmid Extraction and DNA Sequencing

Individual colonies were used to inoculate 3 mL cultures of sterile LB media, containing 50 µg/mL of kanamycin. Cultures were incubated in a rotating drum overnight at 37°C, and plasmid DNA was extracted using the QIAprep Spin Miniprep Kit (Qiagen Cat. No. 27106). Briefly, 2 mL of culture was spun in a microcentrifuge and the supernatant discarded. The cell pellet was resuspended in 250 µL of buffer P1, 250 µL of buffer P2 was added, and the tube inverted 4-6 times. This solution was allowed to sit for no more than 5 min, at which point 350 µL of buffer N3 was added, and the tube inverted 4-6 times to mix. The resulting solution was spun in a microcentrifuge for 10 min, and the supernatant applied to the QIAprep spin column. The column was then spun for 1 min, and the flowthrough discarded. Wash buffer PE (750 µL) was applied to the column, spun through, and discarded. The column was dried with a 1 min spin, and the plasmid DNA eluted with 50 µL of buffer EB.

Plasmid preps were then submitted for in-house sequencing to Marlena Dlutek (Dalhousie University). DNA template was mixed with 6.4 pmol of primer (T7: 5′-TAATACGACTCACTATAGGG-3′; M13 Reverse: 5′-CAGGAAACAGCTATGAC-3′), and the volume adjusted to 15 µL with ddH$_2$O. Sequencing was done via a PCR-based cycle sequencing protocol using the BigDye® Terminator V3.1 Cycle Sequencing kit (Applied Biosystems), and samples run on an ABI Prism^{TM} 377 Automated DNA sequencer.

2.3: Ribosomal Protein Gene Clusters in the Thermotogales and Other Prokaryotes

2.3.1: Global Ribosomal Protein Gene Cluster Analysis

Coenye *et al.* (Coenye and Vandamme 2005) completed an analysis on three separate ribosomal protein gene clusters found in prokaryotes. The authors examined a subset of prokaryotic genomes; in particular, when there was more than one genome available for a group of closely related strains or species, one was chosen as representative.

Here, all genomes present in the TIGR database were examined. This strategy limited sampling bias as much as possible, although it should be emplasized that certain groups of organisms are overrepresented within the database of complete prokaryotic genomes.

2.3.1A: Determining Genome Region Information

Each member of the *s10*, *spc*, and *alpha* ribosomal protein gene clusters was used for a genome region search using the TIGR CMR database. Genomes used included those in the database at the time of analysis (246 genomes: 225 bacterial and 21 archaeal).

Genomic context was determined for each member of the three clusters, and could take one of three possible states: 1) present in the genome, in the same relative position, within the cluster; 2) present in the genome but absent from the cluster; or 3) absent from the genome. Data were summarized separately for bacterial and archaeal genomes and are presented in Results Table 3.2 and Figure 3.7.

2.3.2: L21 Ribosomal Protein Gene Cluster in the Thermotogales

An additional cluster of ribosomal proteins, present in *T. maritima* MSB8, was chosen for analysis within the genomes of several strains present in the lab (Table 2.1). This cluster consists of three ribosomal protein genes: TM1454, or L13; TM1456, or L27; and TM1458, or L21. The cluster is interrupted by a conserved hypothetical protein/orthologous ORFan (TM1455) and a true hypothetical protein/true ORFan (TM1457). For ease of reference, it will be called the L21 cluster, referring to the first gene (see Figure 2.2).

This cluster was chosen for amplification and analysis from the Thermotogales strains present in the lab because of its manageable size (1500 bp). A degenerate PCR approach was chosen to ensure amplification of the correct region, and a schematic is presented in Figure 2.2.

Nested degenerate PCR was used to balance the difficulties in obtaining amplification of the correct targets, while allowing for sequence differences that may occur between different genomes. Utilizing the degeneracy of the genetic code allows for synonymous changes in the DNA sequence that are likely to happen when comparing homologous genes in closely related organisms, and genes with conserved protein sequences in distantly related organisms. Adding a second, nested PCR accounts for the likelihood that the degenerate primers of the first reaction are not specific enough; i.e. they may hybridize to an area of the genome separate from the gene(s) of interest. The

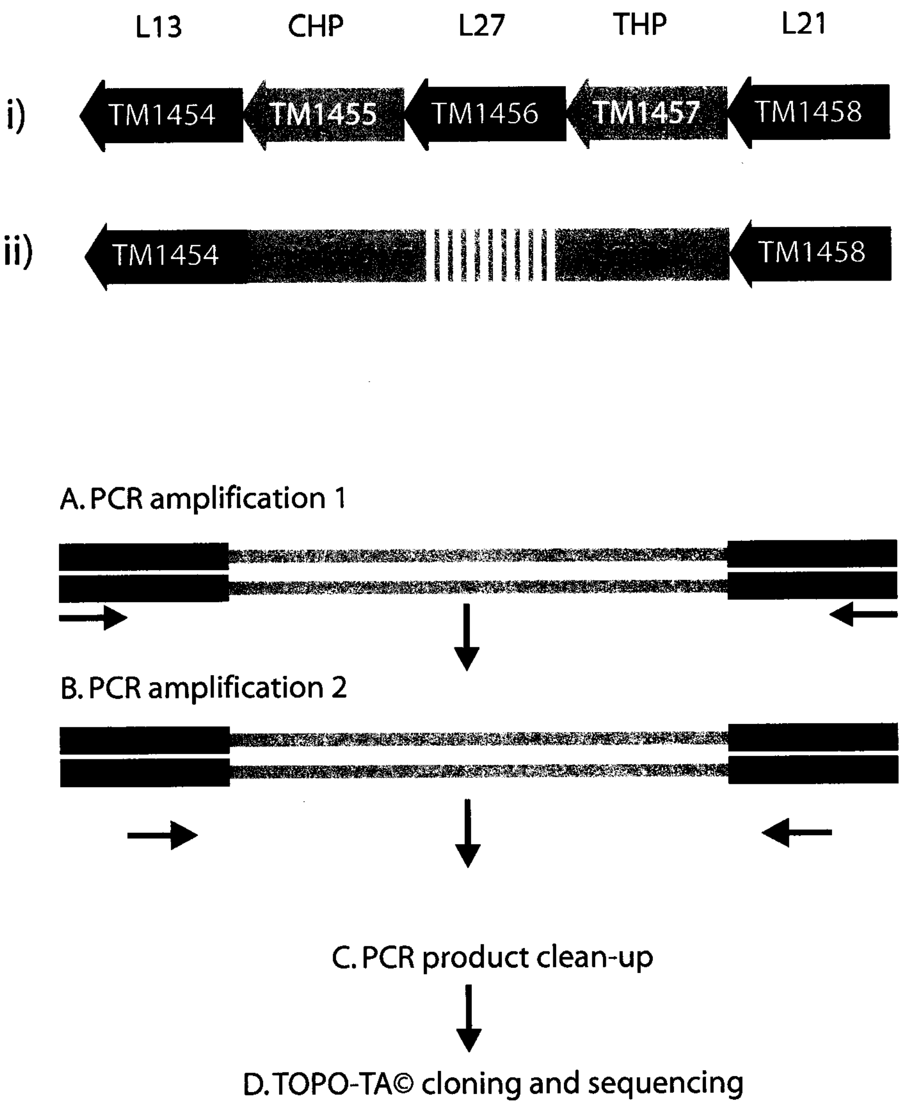


Figure 2.2 Nested degenerate PCR analysis of the L21 ribosomal protein gene cluster in Thermotogales. Illustrated here is a schematic representation of the nested degenerate PCR strategy employed to amplify the L21 gene cluster in Thermotogales. i) L21 gene cluster as found in T. *maritima* MSB8. Ribosomal protein genes are shown in black, and hypothetical proteins are shown in grey; genes are labeled both with TIGR annotation (TMXXXX) and the name of the protein (L21, L27 or L13). ii) Potential genomic context of the region between L21 and L13. Homologs to the genes in *T. maritima* MSB8 are shown in black, while unknown sequence is shown in grey. A) PCR amplification 1 is performed using degenerate primers targeted to either the 5′ end of L21 or the 3′ end of L13. B) PCR Amplification 2 is performed, using the products from PCR amplification 1 as template, along with nested degenerate primers, targeted inside those from amplification 1. C) Products from PCR amplification 2 are cleaned. D) Products are then cloned and sequenced.

primers used in this set of experiments were designed to regions of TM1454 and TM1458 that are conserved in homologous ORFs from other bacteria

2.3.2A: PCR Amplification 1

Primers designed to the 5′ end of TM1458 and the 3′ end of TM1454 were used for the first PCR amplification: TM1458F01 (5′-GTACGCCATTGTNGARACNGC-3′) and TM1454R01 (5′-TCACAGTTCAATNGGYTCNGG-3′). Each 15-μL reaction mix consisted of 2 μL of genomic DNA, 0.5 U Platinum Taq Hi-Fidelity DNA polymerase, 10 μL Hi-Fi buffer, 0.45 μL $MgSO_4$, 400 μM dNTP mix, 100 mM of each primer, and ddH_20 to a final volume of 15 μL. The temperature profile used was as follows: an initial denaturation of 3 min @ 94°C, followed by 35 cycles of 30 sec denaturation @ 94°C, 30 sec of annealing at @ 53°C, 2 min of elongation @ 72°C and a final elongation of 10 min @ 68°C.

2.3.2B: PCR Amplification 2

Nested degenerate primers were designed adjacent to TM1458F01 and TM1454R01 (see Figure 2.2 for placement), and used for a nested PCR amplification: TM1458F02 (5′-GCAGTACAGAGTNGARGARGG-3′) and TM1454R03 (5′-CTTCTGATCGAGYTTYTTNCC-3′). Each 45 μL reaction contained 5 μL of product from PCR Amplification 1, 0.5 U of Platinum Taq Hi-Fidelity DNA polymerase, 4.5 μL of Hi-Fi buffer, 400 mM dNTP mix, 100 mM of each primer, and ddH_20 to a final volume of 45 μL. The temperature profile used was as follows: an initial denaturation of

3 min @ 94°C, followed by 35 cycles of 30 sec denaturation @ 94°C, 30 sec of annealing at @ 53°C, 2 min of elongation @ 72°C and a final elongation of 10 min @ 68°C.

2.3.2C: PCR Product Clean-Up

PCR products from PCR amplification 2 were cleaned using Millipore Montage® PCR Filter Units (Cat. No. UFC7 PCR 50). Briefly, the PCR reaction mix was added to 450 µL of ddH_20 and applied to the filter column, seated in a microcentrifuge tube, and spun in a microcentrifuge for 15 min. The filter was then inverted and placed into a clean tube, 20 µL of TE buffer was added to the top, and the unit spun for 2 min to recover the PCR products for use in TOPO®-TA cloning.

2.3.2D: TOPO®-TA Cloning and Transformation

Cleaned PCR products were cloned into the TOPO®-TA vector, following manufacturer's instructions. Briefly, 4 µL of PCR product was incubated with 1 µL of TOPO® vector and 1 µL of salt solution for 5 min at room temperature. The solution was then placed on ice, and transformation was performed as per TOPO®-XL cloning, described above in Section 2.2G.

2.3.2E: Determining Genomic Context of the L21 Cluster in Other Bacteria

The genomic context for these five ORFs was determined as per Section 2.3.1A, to assess conservation of cluster structure among bacteria and archaea. When this analysis was completed, 266 bacterial genomes were available, and the results are presented in Figure 3.9.

2.4: Spatial Autocorrelation of Functional Categories

A spatial autocorrelation analysis was used to examine the physical distribution of functional categories of genes within bacterial genomes.

2.4A: Downloading Genome Information

All genomic information was downloaded from the Comprehensive Microbial Resource (CMR), at TIGR [http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi]. A list of genomes analyzed can be found in Table 2.2.

The CMR was used to ensure continuity in annotation; each completed genome that is included in this database is annotated using TIGR criteria, both for genomes sequenced at TIGR and those sequenced elsewhere. Data are easily downloaded and arranged in separate functional categories, facilitating analysis. All genome annotations used in this analysis are TIGR annotations.

The following gene attributes were downloaded from [http://cmr.tigr.org/tigr-scripts/CMR/shared/MakeFrontPages.cgi?page=geneattribute]. Options given on the webpage are indicated in italics, while the choices that were made are indicated in bold.

STEP 1: *Choose your gene selection method:* Retrieve attributes for the specified DNA feature within a specific organism and/or a specific role category

Choose the organism(s) of interest: **(Organisms of interest, up to 5 at a time, were chosen)**

Choose the DNA feature of interest: **Primary and TIGR Annotations (Default)**

Choose the role category/categories of interest: **(this was left blank to enable the script to download all categories)**

Table 2.2 Bacterial strains used for Genespat v.4 analyses of spatial autocorrelation of functional groups

Species group	Strains used (Reference, where available)
Bacillus anthracis	*B. anthracis A0039*
	B. anthracis Ames (Read *et al.* 2003)
	B. anthracis Ames Ancestor (Read *et al.* 2003)
	B. anthracis Sterne
	B. anthracis str. France (Fouet *et al.* 2002)
	B. anthracis str. Kruger B
	B. anthracis Vollum
	B. anthracis Western North America USA6153
Campylobacter jejuni	*C. jejuni NCTC 11168* (Parkhill *et al.* 2000)
	C. jejuni RM1221 (Fouts *et al.* 2005)
Chlamydia pneumoniae	*C. pneumoniae AR39* (Read *et al.* 2000)
	C. pneumoniae CWL029 (Kalman *et al.* 1999)
	C. pneumoniae J138 (Shirai *et al.* 2000)
	C. pneumoniae TW-183
Escherichia coli	*E. coli CFT073* (Welch *et al.* 2002)
	E. coli K12-MG1655 (Blattner *et al.* 1997)
	E. coli O157:H7 EDL933 (Perna *et al.* 2001)
	E. coli O157:H7 VT2-Sakai (Hayashi *et al.* 2001)
Legionella pneumophila	*L. pneumophila Lens* (Cazalet *et al.* 2004)
	L. pneumophila Paris (Cazalet *et al.* 2004)
	L. pneumophila Philadelphia 1 (Chien *et al.* 2004)
Prochlorococcus marinus	*P. marinus CCMP1375* (Dufresne *et al.* 2003)
	P. marinus CCMP1378 MED4 (Rocap *et al.* 2003)
	P. marinus MIT 9312
	P. marinus MIT9313 (Rocap *et al.* 2003)
	P. marinus NATL2A
Thermotoga maritima	*T. maritima MSB8* (Nelson *et al.* 1999)

STEP 2: *Choose your gene attributes:*

Choose General Gene Attributes: **Organism Name, DNA Molecule**

Choose TIGR Annotation Gene Attributes: **TIGR Locus Name**

Choose TIGR Annotation Gene Attributes: **TIGR Locus Name, Common Name, Gene Symbol, Cellular Role: Mainrole, Cellular Role: Subrole**

Choose Primary Annotation Gene Attributes: **Primary Locus Name**

Choose Other Gene Attributes: **GenBank ID**

The table generated from this script was then downloaded and opened in Microsoft Excel for ORF coding.

2.4B: ORF Coding

A sample data download is presented in Table 2.3. Data in columns C, D, E, and F are not used further in the spatial autocorrelation analysis, but rather are kept for future analysis on genome clusters of interest. Column J (DNA molecule) is used only to determine which ORFS are on the main chromosome, as plasmids are too small to be dealt with here.

The input file for calculating the joint count statistic (see Section 2.6C below) consists of two columns of numbers – the first representing the gene position, and the second representing the categories being assessed, in this case, main role functional categories. Because of annotation discrepancies, coding requires several steps, and these are illustrated here with the data from Table 2.3.

Table 2.3 Sample genome information and functional category download. Fictional data were created to illustrate the coding process, as explained in Section 2.4B. Column B and Column H form the input file for calculating joint count statistics. Main Role Abbreviations are as follows: AABS – Amino Acid Biosynthesis; BSCPC – Biosynthesis of cofactors, prosthetic groups, and carriers; CE – Cell Envelope; CP – Cellular processes; CIM – Central Intermediary Metabolism; DNAM – DNA Metabolism; PS – Protein Synthesis. Subroles are not given for space reasons.

A. TIGR ORF #	B. Actual position	C. Common name	D. Gene symbol	E. Primary locus name	F. Genbank Acc. No.	G. Main role	H. Role category number	I. Sub role	J. DNA molecule
XX0001	1	~	*xxxX*	~	~	AABS	1	aromatics	main
XX0002	2	~	*xxxX*	~	~	BSCPC	2	molybdopterin	main
XX0003	3	~	*xxxX*	~	~	CE	3	surface carbs	main
XX0004	4	~	*xxxX*	~	~	CP	4	toxin resistance	main
XX0005	5	~	*xxxX*	~	~	CP	4	toxin resistance	main
XX0005	5	~	*xxxX*	~	~	CP	4	pathogenicity	main
XX0007	6	~	*xxxX*	~	~	CIM	5	polyamines	main
XX0008	7	~	*xxxX*	~	~	CIM	5	polyamines	main
XX0009	8	~	*xxxX*	~	~	DNAM	6	replication	main
XX0010	9	~	*xxxX*	~	~	DNAM	6	replication	main
XX0011	10	~	*xxxX*	~	~	DNAM	6	replication	main
XX0014	11	~	*xxxX*	~	~	DNAM	6	degradation	main
XX0015	12	~	*xxxX*	~	~	CIM	5	P compounds	main
XX0016	13	~	*xxxX*	~	~	DNAM	6	degradation	main
XX0017	14	~	*xxxX*	~	~	DNAM	6	degradation	main
XX0018	15	~	*xxxX*	~	~	DNAM	6	degradation	main
XX0019	16	~	*xxxX*	~	~	CIM	5	polyamines	main
XX0020	17	~	*xxxX*	~	~	CIM	5	polyamines	main
XX0021	18	~	*xxxX*	~	~	CP	4	cell division	main
XX0021	18	~	*xxxX*	~	~	PS	15	ribosomal	main

1. ORFs were sorted by “TIGR ORF #” (Column A) – this represents their sequential occurrence on the circular chromosome.
2. Actual positions were determined, beginning at the first gene annotated after the origin of replication (Column B) – this step is necessary since numbers are often duplicated or left out in the final annotation process. For example:
    a. ORF0005 and ORF0021 each have two separate annotations. In the sample dataset, ORF0005 would only be counted once, as both annotations belong to the same main role category. ORF0021 is counted twice, as it was annotated with two distinct main roles, and could therefore theoretically be functioning in two separate pathways.
    b. There is no ORF0006, so ORF0007 would be counted as if it immediately followed ORF0005.
3. The resulting data in Columns B (Actual position) and H (Role category number) were then exported as a text file to be used in the calculation of the joint count statistics.

2.4C: Calculating Joint Count Statistics

To assess the patterning of functional categories of genes in a bacterial genome, join count statistics were calculated. The statistic is calculated for each role category as follows:

$$J_{rr}(d) = \frac{1}{2}\left[\sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\delta_{ij}(d)x_{ri}x_{rj}\right] \qquad \text{(Equation 2.1)}$$

Where:

i and *j* are the members of the gene pair being compared

x_{ij} is the attribute of the gene pair (same category = 1, different category = 0)

d_{ij} is the indication of connectivity of genes i and j

r refers to the role category evaluated

(*d*) indicates the distance class either in terms of *d* neighbours or *d* Euclidean distance classes at which the sampling units are to be considered connected (1) or not (0).

The calculation determines, based on the total number of genes and the number of genes in role category r, whether the members of the category are a) randomly distributed, b) hyperdispersed, or c) patchily clustered (see Figure 2.3 for a schematic).

For a single genome dataset, the distribution of each functional category, numbered from 1 to 23, was calculated separately. Distance classes were evaluated up to 100 genes apart, in a genome of size *G*. While all possible distance classes could be calculated from 1 to ½ *G* (i.e. points directly opposite on the chromosome), most clustering signal would likely degrade by a distance of 100 genes, which was the largest distance class calculated. In most cases, signal degraded by a distance of 50 genes, so data presented only includes those distance classes unless otherwise indicated.

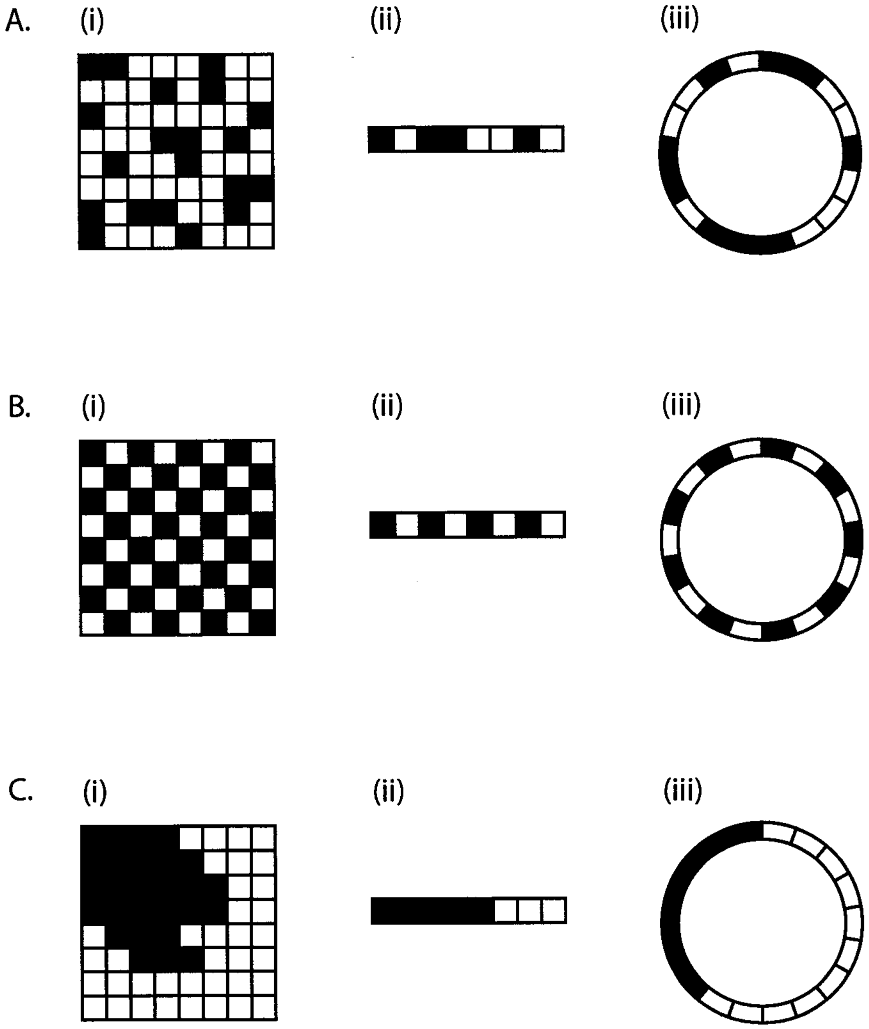


Figure 2.3 Illustration of possible spatial distributions resulting from spatial autocorrelation analysis. Shown here are examples of three spatial distributions of black and white squares on (i) a two-dimensional sampling grid (such as a forest); (ii) a one-dimensional sampling line (such as a coastline); (iii) a circular sample (such as a bacterial genome, a circular piece of DNA) that could result from spatial autocorrelation analysis by joint count.

A. Random distribution: here there is no discernable patterning to the black squares

B. Hyperdispersal of black squares: Black squares are never found adjacent to one another, and therefore are considered to be negatively associated.

C. Clustered or patchy distribution: Black squares are found adjacent to one another more often than would be expected randomly, and are therefore considered to be positively associated.

2.4D: Genespat v.4

In order to efficiently perform the joint count calculations, a script was written in Pascal by Dr. Robert Latta, and compiled for execution on an Apple Terminal platform (as "Genespat" v.4) by Dr. David Spencer.

Genespat gives a table as output, with the following information given for each category:

i. Role category

ii. Frequency - the frequency of each role category, out of all functionally annotated ORFs

iii. Number – the total number of functionally annotated ORFs in the category

iv. Count: the number of joins at distance (*d*) found

v. Exp: the expected number of joins at distance (*d*) based on a normal distribution

vi. Snd (z): Z-score for comparison to the normal distribution.

2.4E: Visualization of Distributions

For illustrative purposes, sample datasets were generated, with random, hyperdispersed, and clustered distributions as shown in Figure 2.3, and put through Genespat (see Table 3.4 for sample datasets, and Figure 3.11 for plots of each dataset). Visualizing patterns of distribution was accomplished by plotting the Z-scores for each category against the distance (d). A positive Z-score at each distance indicates that there are more occurrences of genes of the same category at that distance than would be expected, where a negative Z-score indicates a dissociation, i.e. genes at that distance are likely to be of different categories. In all cases, values >|1.96| are considered significant.

The Z-scores for each distribution gives a distinct curve: (i) randomly distributed genes result in Z-scores that cycle near zero (See Results, Fig 3.11A); (ii) hyperdispersed genes result in Z-scores that are initially negative, but spike significantly to indicate spacing (Fig 3.11B); (iii) clustered genes result in Z-scores that start off positive at close distances, but decrease to zero (Fig 3.11C).

Plots were generated for functional categories in T. *maritima* MSB8, and grouped into random, hyperdispersed, and clustered categories (See Results Figure 3.12)

2.4F: Functional Gene Category Frequency Distributions and Genome Comparisons in Six Groups of Closely Related Bacterial Strains

Groups of closely related prokaryotes were chosen for comparison of the evolution of functional architecture, because only one completely sequenced Thermotogales genome was available. In cases where more than one strain was available in the TIGR database, all members of a strain group were downloaded as per Section 2.4A and 2.4B; each one was coded as per Section 2.4C, and analyzed with Genespat v.4 as per section 2.4D.

The frequencies of each functional category within groups of genomes were compared to determine if groups of genomes devote similar amounts of their functional genome architecture to particular functional categories. Frequency values for each group of closely related strains or species were plotted in bar-graph format to visualize the variation between the genomes (see Appendix 2). As well, distribution plots were generated for the conserved hypothetical proteins and true hypothetical proteins from each of the strains analyzed (see Appendix 3 and 4).

CHAPTER 3: RESULTS

3.1: Long Walk PCR Analysis of the Proximal Flagellar Cluster

3.1.1: Modifications of the Method of Katz *et al.* (2000)

This project utilized aspects of the walking PCR method of Katz *et al.* (Katz *et al.* 2000), but with several key modifications. The authors used an avadin/agarose capture system, which was changed here to streptavidin/biotin with molecule capture using magnetic beads (Section 2.2). This system is often used for the capture of mRNA transcripts, and the affinity of streptavidin for biotin is strong, enabling efficient capture of biotin-tagged targets.

Secondly, the original method was used to obtain terminal sequences of genes, and exact-match primers were used in both the initial linear amplification and the subsequent nested PCR. The protocol was modified here to use degenerate primers, to enable amplification in more than one strain where there may be DNA sequence differences underlying a conserved protein sequence.

Thirdly, a long-range, high-fidelity Taq polymerase (Invitrogen Cat. No. 11304-029) was employed. The original method was used to obtain sequence data for small stretches of DNA. Use of a high-fidelity polymerase enables accurate amplification of longer regions, resulting in more sequence data extending into unknown regions.

Table 3.1 Thermotogales strains used in successful amplification of the PFC, using Long Walk PCR.

Genus	Strain amplified
Thermotoga	*T. maritima* SL7
	T. maritima FjSS3B1
	Thermotoga petrophila RKU1
	Thermotoga naphthophila RKU10
	Thermotoga sp. RQ2
	Thermotoga sp. RQ7
	Thermotoga sp. SG1
	Thermotoga sp. kol 6K
	T. neapolitana LA4
	T. neapolitana LA10
	T. neapolitana NS-E
	T. thermarum LA3
	T. subterranea SL1
Thermosipho	*T. africanus* Ob7
Fervidobacterium	*F. islandicum* H12
	F. nodosum
Petrotoga	*P. miotherma*

3.1.2: Amplification Results Using the Modified Long Walk PCR Protocol

Long Walk PCR (LWPCR) was successful in 17 strains (Table 3.1). For each PCR amplification performed (Figure 2.1D and 2.1F), crystal violet staining was used to size-select products via staining of the molecular marker. In all cases, no product was visible due to the low yield and low sensitivity of crystal violet staining. Figure 3.1 shows a schematic of the electrophoresis results from PCR Amplification 1 and 2 (Figure 2.1D and F) and demonstrates the length distribution of obtained PCR products. When the protocol works properly, a smear of products ranging from ~100 bp up to 20 kbp should be obtained, reflecting the tendency of individual polymerase molecules to drop off at random extension lengths. Because the amplification is targeting an unknown region of the genome, a region of the smear from 4-12 kbp was isolated to obtain products that were as long as possible.

3.1.3: Open Reading Frame Identification

Sequence data from each strain that was successfully amplified were used to perform a database search using TIGR BLAST [http://tigrblast.tigr.org/cmr-blast/], and the resulting gene identities were mapped onto the SSU rRNA trees generated by Nesbø *et al.* (Nesbø *et al.* 2002) (see Figure 3.2). The entire *maritima/neapolitana* clade contained the full PFC cluster, as is present in *T. maritima* MSB8, and this was sufficiently conserved to allow for further DNA sequence comparison (please see Section 3.1.4).

The sister clade (as defined by SSU rRNA) that consists of *T. thermarum* LA3 and *T. subterranea* contained a second set of flagellar protein genes, which code for

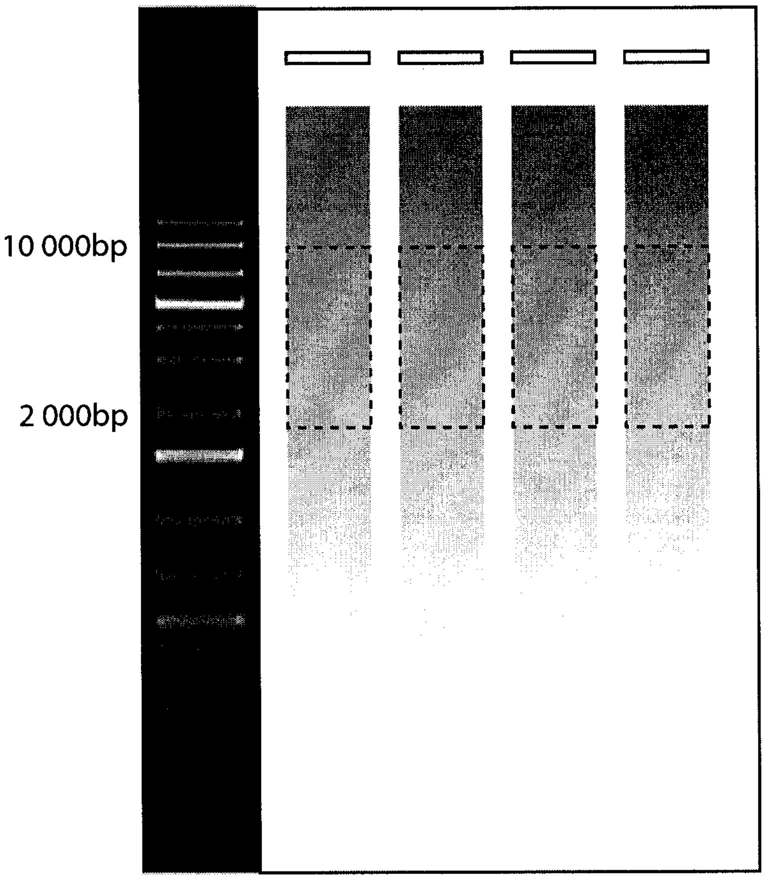
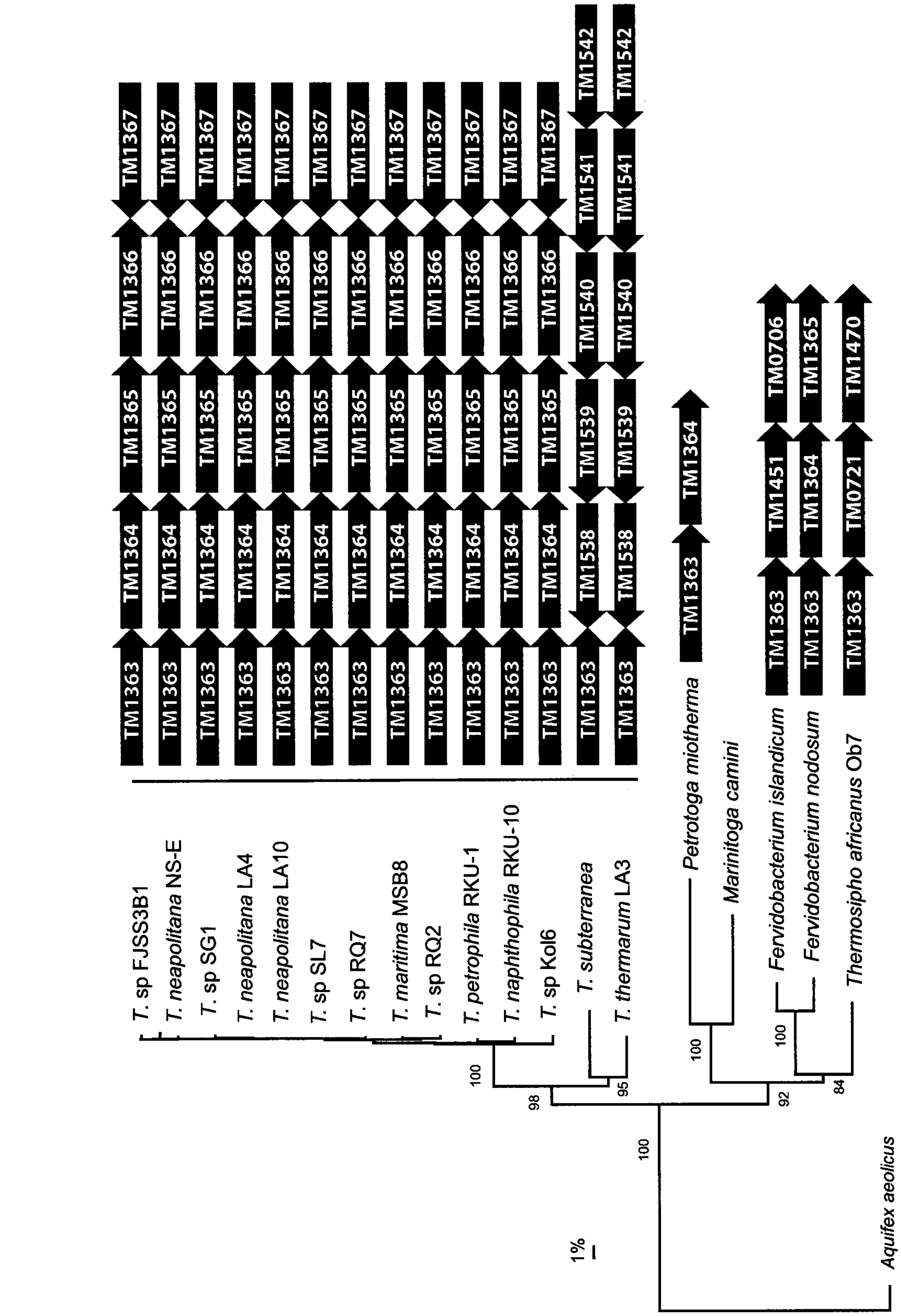


Figure 3.1 Illustration of the electrophoretic pattern of Long Walk PCR products. Smears of amplified DNA are shown as diffuse gray rectangles; gel slices were cut at the position of the dotted lines, and PCR products extracted as per Section 2.2E.

Figure 3.2 ORF identification in Long Walk PCR products in the Thermotogales. Genes from the proximal flagellar cluster (PFC) are indicated in red, while genes from the distal flagellar cluster (DFC) are indicated in blue. Other open reading frames with homology to genes from T. *maritima* MSB8 are indicated in green, and *prfA*, the anchor gene, is indicated in black.


T. sp FJSS3B1
T. neapolitana NS-E
T. sp SG1
T. neapolitana LA4
T. neapolitana LA10
T. sp SL7
T. sp RQ7
T. maritima MSB8
T. sp RQ2
T. petrophila RKU-1
T. naphthophila RKU-10
T. sp Kol6
T. subterranea
T. thermarum LA3
Petrotoga miotherma
Marinitoga camini
Fervidobacterium islandicum
Fervidobacterium nodosum
Thermosipho africanus Ob7
Aquifex aeolicus
100
98
95
100
100
100
92
84
1%
TM1363
TM1364
TM1365
TM1366
TM1367
TM1538
TM1539
TM1540
TM1541
TM1542
TM1451
TM0706
TM0721
TM1470

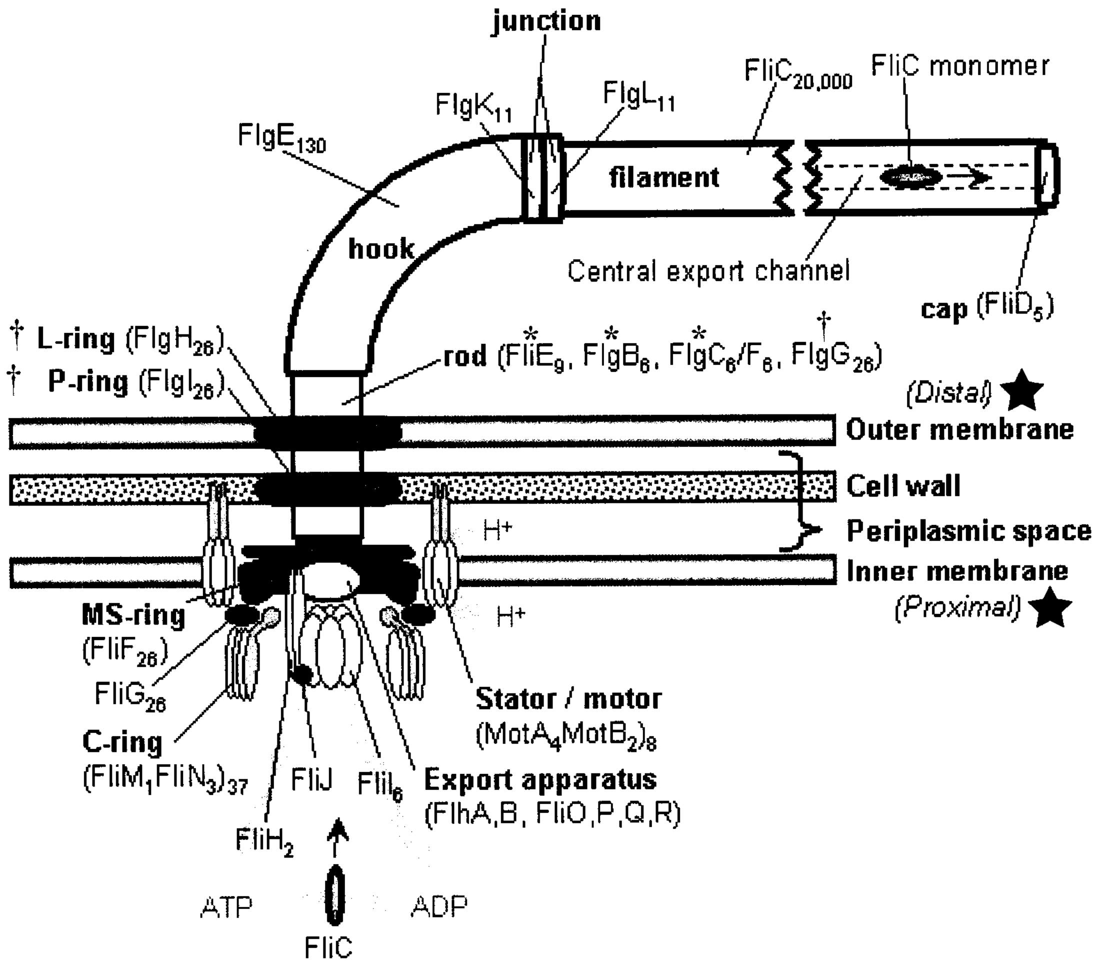


Figure 3.3 Location of the PFC and DFC gene products in the bacterial flagellar apparatus. Illustration was taken from http://www.talkdesign.org/faqs/flagellum.html). PFC genes with homologs in the Thermotogales are indicated by a red asterisk (*), and genes from the DFC are indicated by a blue dagger (†).

proteins present in the distal regions of the flagellar structure, or DFC (see Figure 3.3). The DNA sequence of the DFC in these two strains is conserved within the clade, but different than that of the distal cluster in *T. maritima* MSB8. However, the 3′ end of *prfA* is intact and conserved, as would be expected from the successful amplification.

3.1.4: Sequence Data for the PFC Genes in Thermotogales

All sequence data for the PFC genes amplified in selected Thermotogales strains were aligned, and used for preliminary phylogenetic analysis; trees are presented in Figure 3.4. The sequences were sufficiently close that not many groupings could be resolved, except for two interesting events involving *T. maritima* MSB8, *T.* sp. RQ2, and the two Japanese strains, *T. petrophila* RKU1 and *T. naphthophila* RKU10. The relationship set out by the SSU rRNA tree (Figure 1.1) groups *T. maritima* MSB8 with *T.* sp. RQ2, distinct from *T. petrophila* RKU1 and *T. naphthophila* RKU10, but the homologs of TM1364 and TM1367 are more closely related in *T. maritima* MSB8 and *T. petrophila* RKU 1, and subsequently in *T.* sp. RQ2 and *T. naphthophila* RKU10. All of the heterogeneities between these four strains are presented in Figure 3.5.

It is evident by visual inspection that a large portion of the heterogeneities in this alignment support the grouping of *T.* sp. RQ2 with *T. petrophila* RKU1, and *T. maritima* MSB8 with *T. naphthophila* RKU10, contrary to the SSU rRNA relationship, and these four strains were chosen for further analysis.

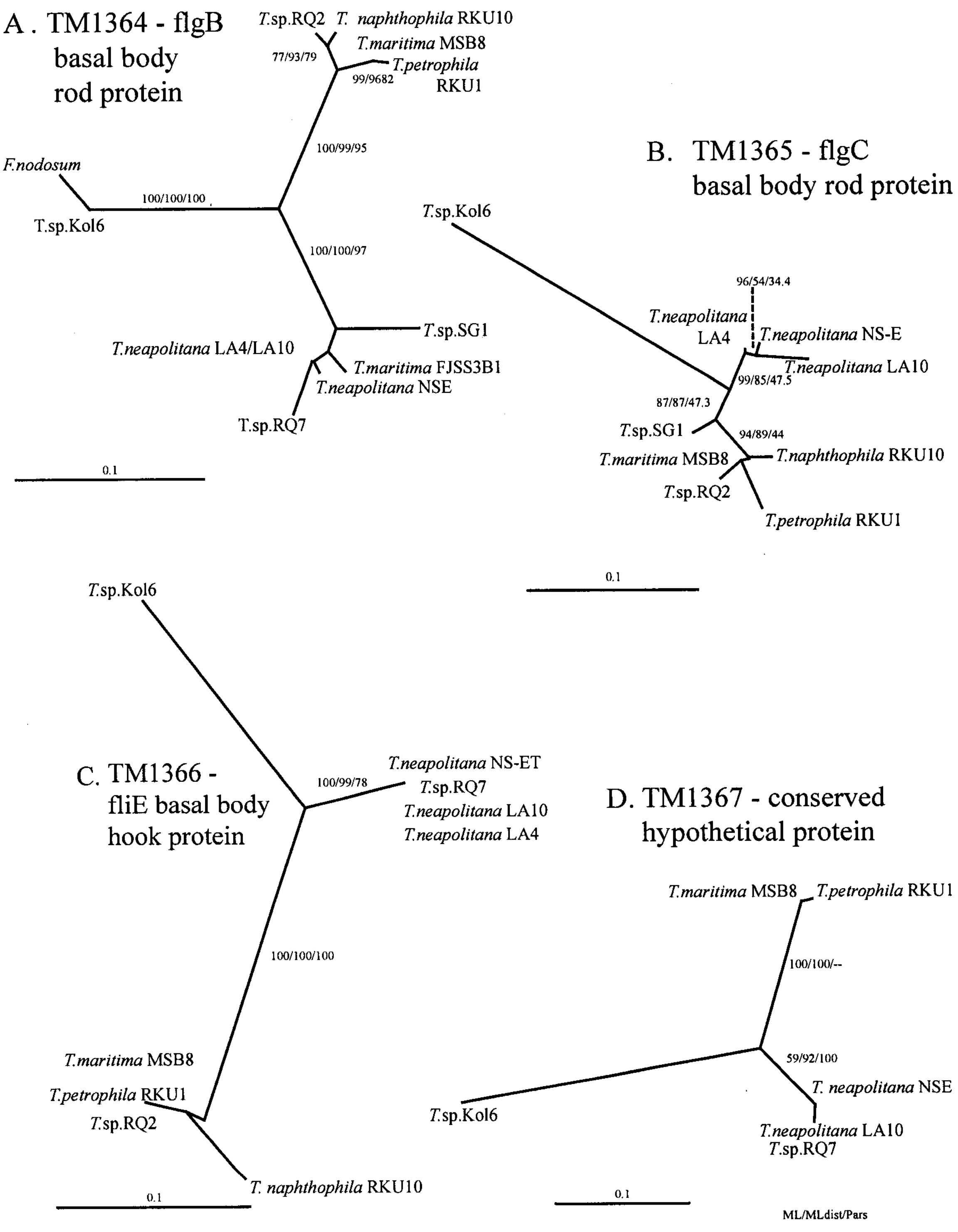
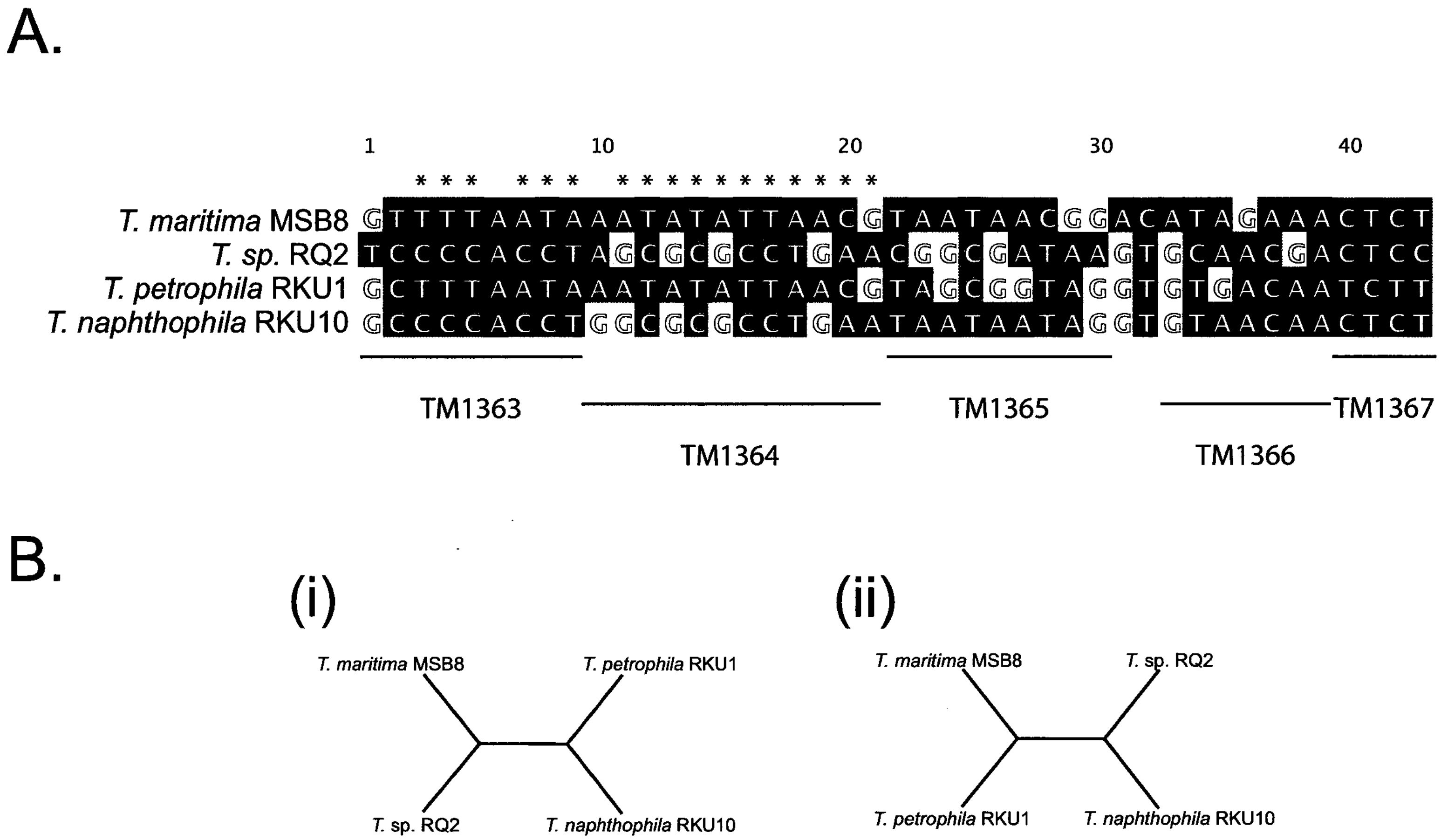


Figure 3.4 Phylogenetic analysis of PFC genes in the Thermotogales.

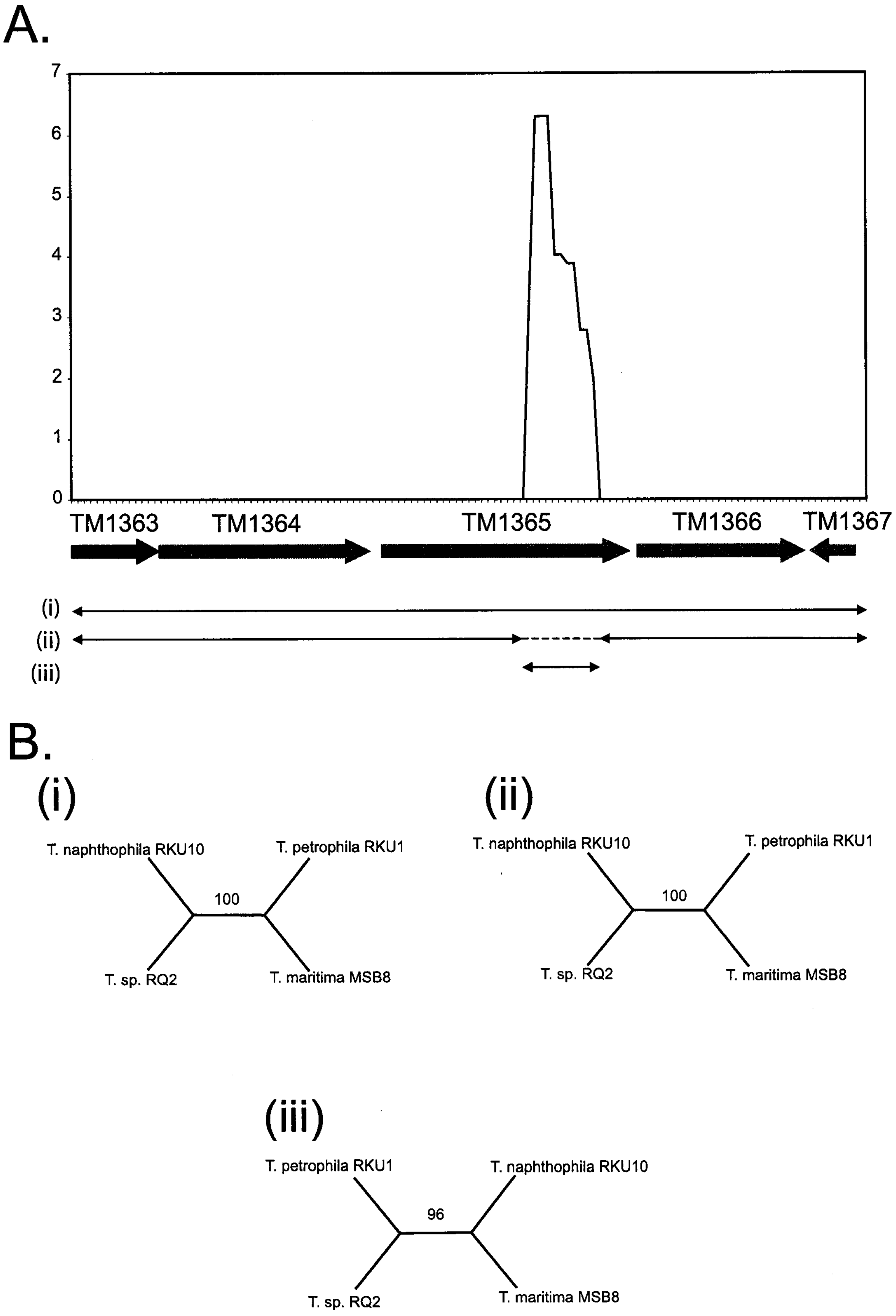
Figure 3.5 DNA heterogeneities in PFC genes from four Thermotogales strains. DNA heterogeneities were culled from the alignment of the PFC cluster in *T. maritima* MSB8, *T.* sp. RQ2, *T. petrophila* RKU1 and *T. naphthophila* RKU10. A) Alignment of heterogeneities of the four strains examined. Positions that support the grouping of *T. maritima* MSB8 with *T. petrophila* RKU1, and *T.* sp. RQ2 with *T. naphthophila* RKU10 are indicated by an asterisk (*). B) Unrooted four-taxon trees showing (i) the relationship of the four strains as determined by SSU rRNA sequence and (ii) the relationship as determined by the positions marked with an asterisk (*). The majority of the remaining heterogeneities are phylogenetically uninformative.


A.
1
10
20
30
40
T. maritima MSB8 GTTTTAATAAATATATTAACGTAATAACGGACATAGAAACTCT
T. sp. RQ2 TCCCCACCTAGCGCGCCTGAACGGCGATAAGTGCAACGACTCC
T. petrophila RKU1 GCTTTAATAAATATATTAACGTAGCGGTAGGTGTGACAATCTT
T. naphthophila RKU10 GCCCCACCTGGCGCGCCTGAATAATAATAGGTGTAACAACTCT
TM1363
TM1364
TM1365
TM1366
TM1367
B.
(i)
T. maritima MSB8
T. petrophila RKU1
T. sp. RQ2
T. naphthophila RKU10
(ii)
T. maritima MSB8
T. sp. RQ2
T. petrophila RKU1
T. naphthophila RKU10

3.1.5: Likewind Recombination Analysis

Recombination analysis was performed to determine if the region indicated by the heterogeneities (Figure 3.5) was significant and represented an actual recombination event in the history of these four organisms. The program (Archibald and Roger 2002) uses a sliding window approach to calculate the difference in likelihood between trees made from each window of the alignment (100 nt) and one made from the entire alignment. If a window of the entire alignment gives a maximum likelihood tree that has a different topology than that made from the main alignment, it can be considered to have been involved in a recombination event. Results of the analysis indicated that a short region of TM1365 (*flgC*) disagreed with the remainder of the alignment, indicating a likely recombination event. This region includes all but one of the heterogeneities present in the alignment of TM1365 (see Figure 3.5). The regions identified by Likewind analysis were then used to generate mini-phylogenies in PAUP* (Swofford 2002) of these four strains, to determine their histories. Figure 3.6A shows the delta-lnL plotted against the length of the alignment, and region of recombination can be seen as a spike. Figure 3.6B demonstrates the three mini-phylogenies generated, for i) the entire alignment, ii) the baseline (region outside the recombination) and iii) the region of recombination. Two important points can be gleaned from this figure – firstly, the entire region of the flagellar cluster has a separate history than that of the SSU rRNA in these organisms, and secondly, that the small region of recombination, encompassing part of ORFs *flgB* and *flgC* has yet a third history.

Figure 3.6 Likewind recombination analysis of the PFC genes from four members of the Thermotogales. A) Plot of delta lnL between lnL of trees created from a 100 bp window of the alignment as compared with lnL of the tree created by the entire alignment. The region encompassed by the spike has the greatest delta lnL, and therefore a different history than the rest of the alignment. B) Unrooted four-taxon trees built from different regions of the alignment: (i) based on the entire concatenated DNA sequence; (ii) based on the baseline regions; (iii) based on the region encompassed by the spike. The relationship based on both the entire alignment and the baseline is shown in B) (i) and (ii), with *T. maritima* MSB8 and *T. petrophila* RKU1 forming a clade, and *T.* sp. RQ2 and *T. naphthophila* RKU10 forming a separate clade. The relationship based on the dischordant region, shown in B)(iii), groups *T. maritima* MSB8 with *T. naphthophila* RKU10, and *T.* sp. RQ2 with *T. petrophila* RKU1. It should be noted that this region of recombination shows yet a third relationship, different from that determined by SSU rRNA sequence, and the DNA sequence of the remainder of this alignment.


A.
7
6
5
4
3
2
1
0
TM1363
TM1364
TM1365
TM1366
TM1367
(i)
(ii)
(iii)
B.
(i)
T. naphthophila RKU10
T. petrophila RKU1
100
T. sp. RQ2
T. maritima MSB8
(ii)
T. naphthophila RKU10
T. petrophila RKU1
100
T. sp. RQ2
T. maritima MSB8
(iii)
T. petrophila RKU1
T. naphthophila RKU10
96
T. sp. RQ2
T. maritima MSB8

3.1.6: Presence and Conservation of the Distal Flagellar Cluster (DFC)

Amplification of the DFC was attempted using a separate method, nested degenerate PCR, as described in Section 2.4. This method was chosen because ORF presence and order of the DFC was conserved between *T. maritima* MSB8 and the *thermarum/subterranea* clade, in the region amplified by long walk PCR. This enabled primers to be designed for the flanking genes of the cluster, TM1538 and TM1543, similar in location to those used to amplify the L21 cluster (see Figure 3.2). Unfortunately, amplification was unsuccessful in most of the strains available, and possible explanations will be presented in the Discussion.

3.2: Ribosomal Protein Gene Clusters

3.2.1: *s10*, *spc* and *alpha* Ribosomal Protein Gene Clusters in Prokaryotes

Data for the 266 prokaryotic genomes analyzed are presented in Table 3.2. The data for Bacteria and Archaea were considered separately, and are summarized in Figure 3.7. For Bacteria, 8 of the 11 members of the *s10* cluster are conserved, both in presence and position, in >85% of the genomes examined, while 7 of 15 in the *spc* cluster and 4 of 5 in the *alpha* cluster are conserved to this degree. In Archaea, the *s10* cluster members are found elsewhere, spread throughout the genome, but 8 of 15 genes in the *spc* cluster, and 4 of 5 genes in the *alpha* cluster are conserved in presence and position in >85% of the genomes examined.

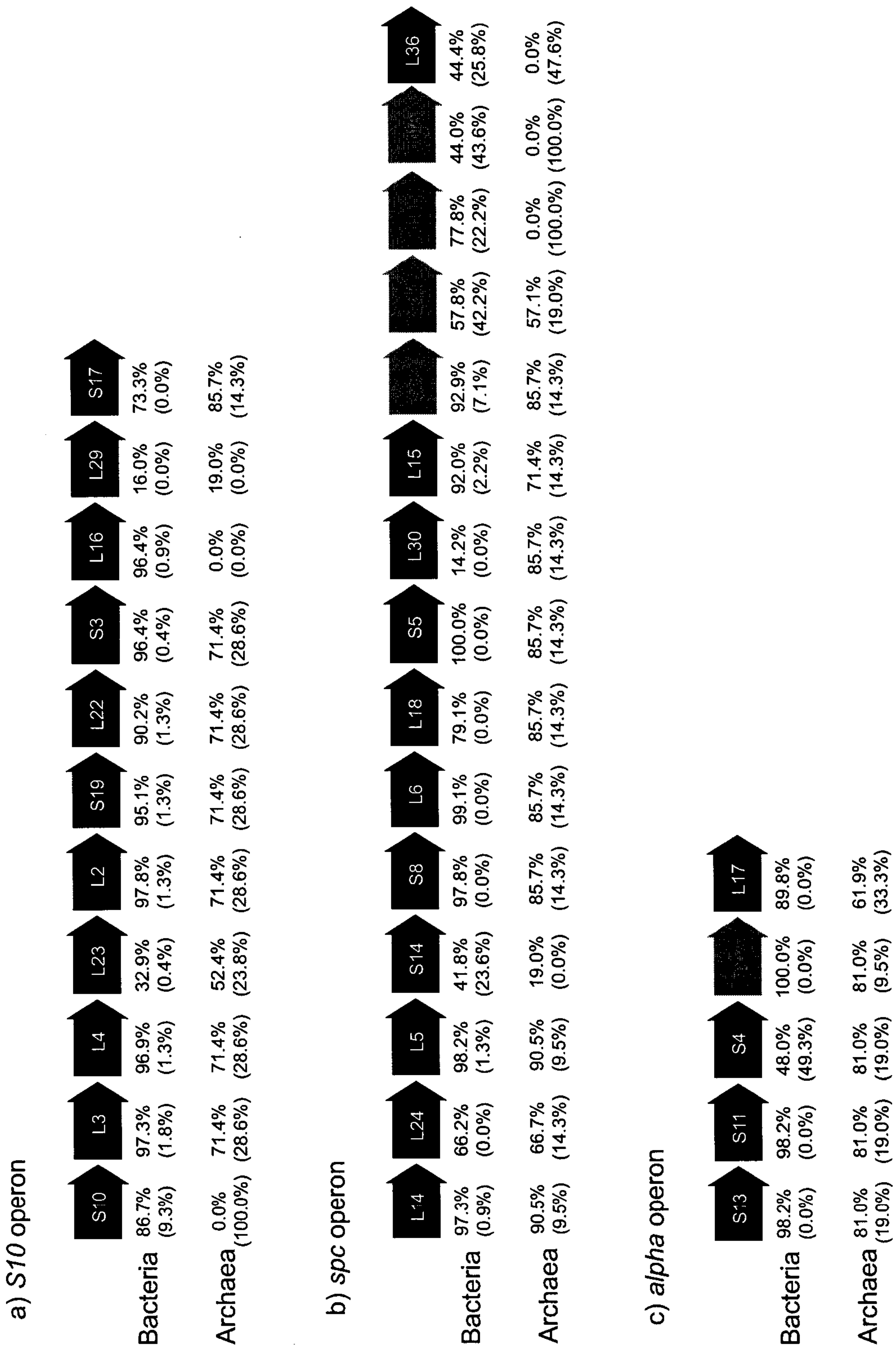
Table 3.2 Genomic context analysis of *s10*, *spc*, and *alpha*, three ribosomal protein gene clusters. # + bacteria – Number of bacterial genomes where the gene is both present and conserved in relative position to the other members of the cluster; # + archaea – Number of archaeal genomes where the gene is both present and conserved in relative position to the other members of the cluster; # X bacteria – number of bacterial genomes where the gene is present, but not found within the cluster; # X archaea – Number of archaeal genomes where the gene is present, but not found in the cluster. In archaea, within the alpha operon, the first four genes are present in the cluster for 16 strains, but S11 and S4 are swapped in position; also, all but one of the L17 proteins in archaea are annotated as L18, indicated by **.

	# +	# X	Total	% +	% X	% -
Protein	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria	Bacteria
S10	195	21	225	86.7	9.3	4.0
L3	219	4	225	97.3	1.8	0.9
L4	218	3	225	96.9	1.3	1.8
L23	74	1	225	32.9	0.4	66.7
L2	220	3	225	97.8	1.3	0.9
S19	214	3	225	95.1	1.3	3.6
L22	203	3	225	90.2	1.3	8.4
S3	217	1	225	96.4	0.4	3.1
L16	217	2	225	96.4	0.9	2.7
L29	36	0	225	16.0	0.0	84.0
S17	165	0	225	73.3	0.0	26.7
L14	219	2	225	97.3	0.9	1.8
L24	149	0	225	66.2	0.0	33.8
L5	221	3	225	98.2	1.3	0.4
S14	94	53	225	41.8	23.6	34.7
S8	220	0	225	97.8	0.0	2.2
L6	223	0	225	99.1	0.0	0.9
L18	178	0	225	79.1	0.0	20.9
S5	225	0	225	100.0	0.0	0.0
L30	32	0	225	14.2	0.0	85.8
L15	207	5	225	92.0	2.2	5.8
secY	209	16	225	92.9	7.1	0.0
adk	130	95	225	57.8	42.2	0.0
map	175	50	225	77.8	22.2	0.0
infA	99	98	225	44.0	43.6	12.4
L36	100	58	225	44.4	25.8	29.8
S13	221	0	225	98.2	0.0	1.8
S11	221	0	225	98.2	0.0	1.8
S4	108	111	225	48.0	49.3	2.7
rpoA	225	0	225	100.0	0.0	0.0
L17	202	0	225	89.8	0.0	10.2

	# +	# X	Total	% +	% X	% -
Protein	Archaea	Archaea	Archaea	Archaea	Archaea	Archaea
S10	0	21	21	0.0	100.0	0.0
L3	15	6	21	71.4	28.6	0.0
L4	15	6	21	71.4	28.6	0.0
L23	11	5	21	52.4	23.8	23.8
L2	15	6	21	71.4	28.6	0.0
S19	15	6	21	71.4	28.6	0.0
L22	15	6	21	71.4	28.6	0.0
S3	15	6	21	71.4	28.6	0.0
L16	0	0	21	0.0	0.0	100.0
L29	4	0	21	19.0	0.0	81.0
S17	18	3	21	85.7	14.3	0.0
L14	19	2	21	90.5	9.5	0.0
L24	14	3	21	66.7	14.3	19.0
L5	19	2	21	90.5	9.5	0.0
S14	4	0	21	19.0	0.0	81.0
S8	18	3	21	85.7	14.3	0.0
L6	18	3	21	85.7	14.3	0.0
L18	18	3	21	85.7	14.3	0.0
S5	18	3	21	85.7	14.3	0.0
L30	18	3	21	85.7	14.3	0.0
L15	15	3	21	71.4	14.3	14.3
secY	18	3	21	85.7	14.3	0.0
adk	12	4	21	57.1	19.0	23.8
map	0	21	21	0.0	100.0	0.0
infA	0	21	21	0.0	100.0	0.0
L36	0	10	21	0.0	47.6	52.4
S13	17	4	21	81.0	19.0	0.0
S11	17	4	21	81.0	19.0	0.0
S4	17	4	21	81.0	19.0	0.0
rpoA	17	2	21	81.0	9.5	9.5
L17**	13	7	21	61.9	33.3	4.8

Figure 3.7 Illustration of the conservation of *s10*, *spc* and *alpha* ribosomal protein gene clusters in prokaryotes. Analysis of the *s10*, *spc* and *alpha* ribosomal protein gene clusters in prokaryotes was based on Coenye and Vandamme, 2005, with the inclusion of 147 additional genomes. Illustrated here is a representation of the frequency and organization of three ribosomal protein gene clusters found in prokaryotes. An all vs all BLAST analysis was performed using the TIGR CMR "genome comparison" utility to determine how often a particular gene is either present in bacterial or archaeal genomes in this cluster (top number) or present elsewhere in the genome (bottom number in parentheses). This analysis includes 246 completely sequenced genomes (225 bacteria and 21 archaea) that are found in the TIGR CMR Database.

Notes:
In the *alpha* operon, S11 and S4 are reversed to S4-S11 in all genomes containing the complete operon.
Two archaea (*Nanoarchaeon equitans* and *Pyrobaculum furiosus*) have little or no conserved organization.


a) S10 operon
Bacteria
Archaea
S10 86.7% (9.3%) 0.0% (100.0%)
L3 97.3% (1.8%) 71.4% (28.6%)
L4 96.9% (1.3%) 71.4% (28.6%)
L23 32.9% (0.4%) 52.4% (23.8%)
L2 97.8% (1.3%) 71.4% (28.6%)
S19 95.1% (1.3%) 71.4% (28.6%)
L22 90.2% (1.3%) 71.4% (28.6%)
S3 96.4% (0.4%) 71.4% (28.6%)
L16 96.4% (0.9%) 0.0% (0.0%)
L29 16.0% (0.0%) 19.0% (0.0%)
S17 73.3% (0.0%) 85.7% (14.3%)
b) spc operon
Bacteria
Archaea
L14 97.3% (0.9%) 90.5% (9.5%)
L24 66.2% (0.0%) 66.7% (14.3%)
L5 98.2% (1.3%) 90.5% (9.5%)
S14 41.8% (23.6%) 19.0% (0.0%)
S8 97.8% (0.0%) 85.7% (14.3%)
L6 99.1% (0.0%) 85.7% (14.3%)
L18 79.1% (0.0%) 85.7% (14.3%)
S5 100.0% (0.0%) 85.7% (14.3%)
L30 14.2% (0.0%) 85.7% (14.3%)
L15 92.0% (2.2%) 71.4% (14.3%)
92.9% (7.1%) 85.7% (14.3%)
57.8% (42.2%) 57.1% (19.0%)
77.8% (22.2%) 0.0% (100.0%)
44.0% (43.6%) 0.0% (100.0%)
L36 44.4% (25.8%) 0.0% (47.6%)
c) alpha operon
Bacteria
Archaea
S13 98.2% (0.0%) 81.0% (19.0%)
S11 98.2% (0.0%) 81.0% (19.0%)
S4 48.0% (49.3%) 81.0% (19.0%)
100.0% (0.0%) 81.0% (9.5%)
L17 89.8% (0.0%) 61.9% (33.3%)

Of the 31 genes examined, only three were universally present in all bacteria and archaea present: S5, *secY* and *map*, all from the *spc* cluster; only one of these (S5) codes for a structural ribosome protein. Three genes are absent from >60% of bacterial genomes (L23 and L29 from the *s10* cluster, and L30 from the *spc* cluster), and three are absent from >60% of archaeal genomes (L16 and L29 from the *s10* cluster, and S14 from the *spc* cluster).

3.2.2: L21 Ribosomal Protein Gene Cluster in Thermotogales

The full gene cluster was successfully amplified and sequenced for 8 of the 10 organisms assayed, and ORFs were determined by BLAST and mapped on the SSU rRNA tree (Figure 3.8).

In each case where the cluster was present and amplifiable under PCR conditions used, it was found to be intact – all five members were present, and their sequences were not interrupted by stop codons. Alignment of the sequence (including the 4 base pairs of intergenic spacer found between TM1455 and TM1456) revealed no indels.

Alignments of the heterogeneities within each of the ORFs from the L21 cluster in the eight strains assayed, as well as homologs from *T. maritima* MSB8, are shown in Figure 3.9 (A through E). Similar to the PFC cluster amplified in the Thermotogales, the DNA sequences are seen to have a high percent identity. Preliminary phylogenies were created in PAUP* (data not shown), and resulted in a split-star phylogeny, with five strains on either side – *T.* sp. RQ2, *T.* sp. RQ7, *T. thermarum* LA3, *T. africanus* ob7 and *T. maritima* MSB8 forming one clade, and *T.* sp. SG1, *T. neapolitana* NS-E, *T. neapolitana* LA4 and *T. neapolitana* LA10 forming the second. Differences are easily

Figure 3.8 L21 ribosomal protein gene cluster amplified from Thermotogales. Strains from which the L21 cluster was successfully amplified are indicated by a yellow star, on a modified SSU rRNA tree. The *maritima/neapolitana* clade topology is taken from Figure 1.1B. Because only one other strain was successfully amplified, the region of the tree containing the other Thermotogales strains is indicated using a dashed-line backbone, from Figure 1.1A. *T. maritima* MSB8 is indicated using a red star.

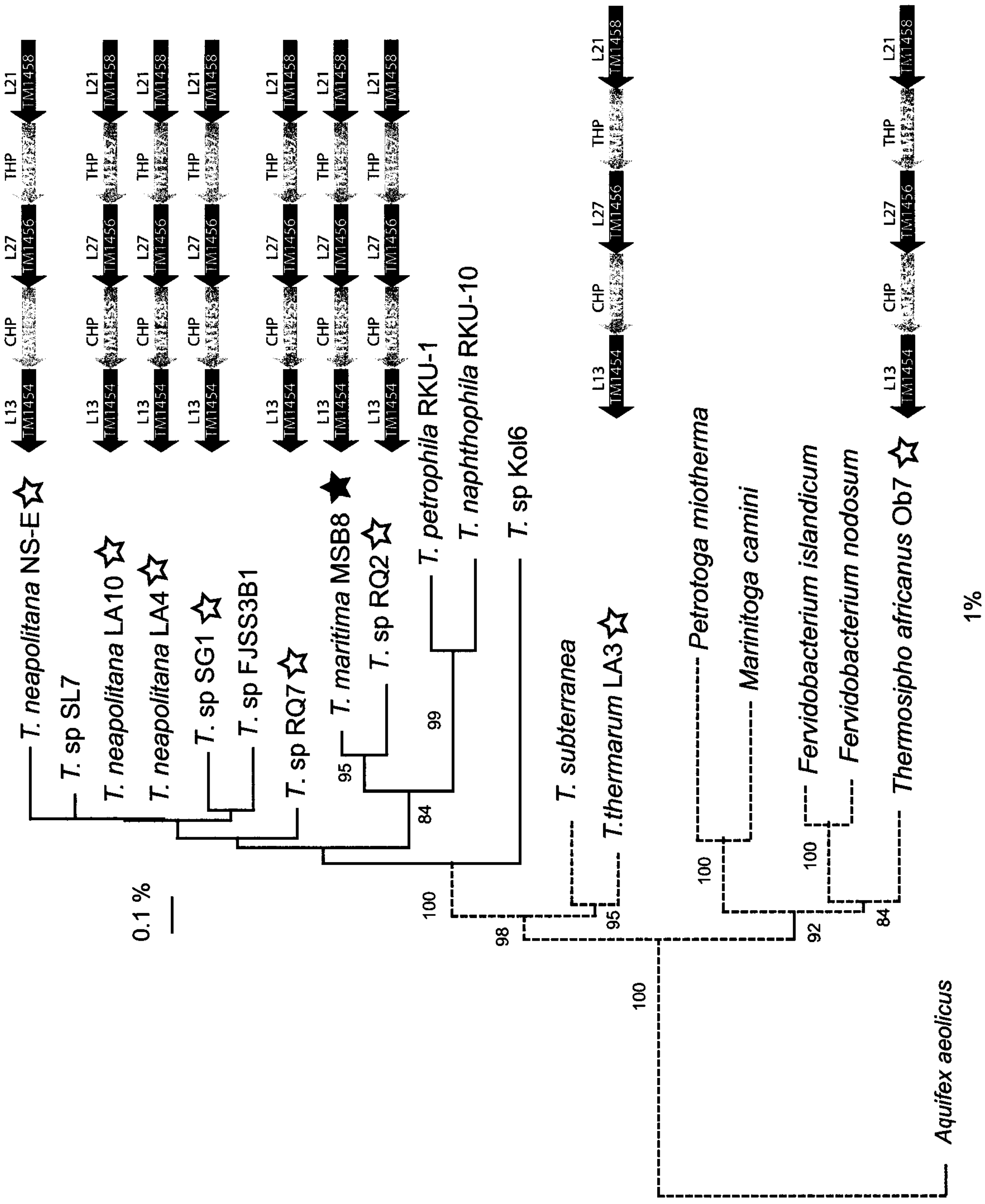
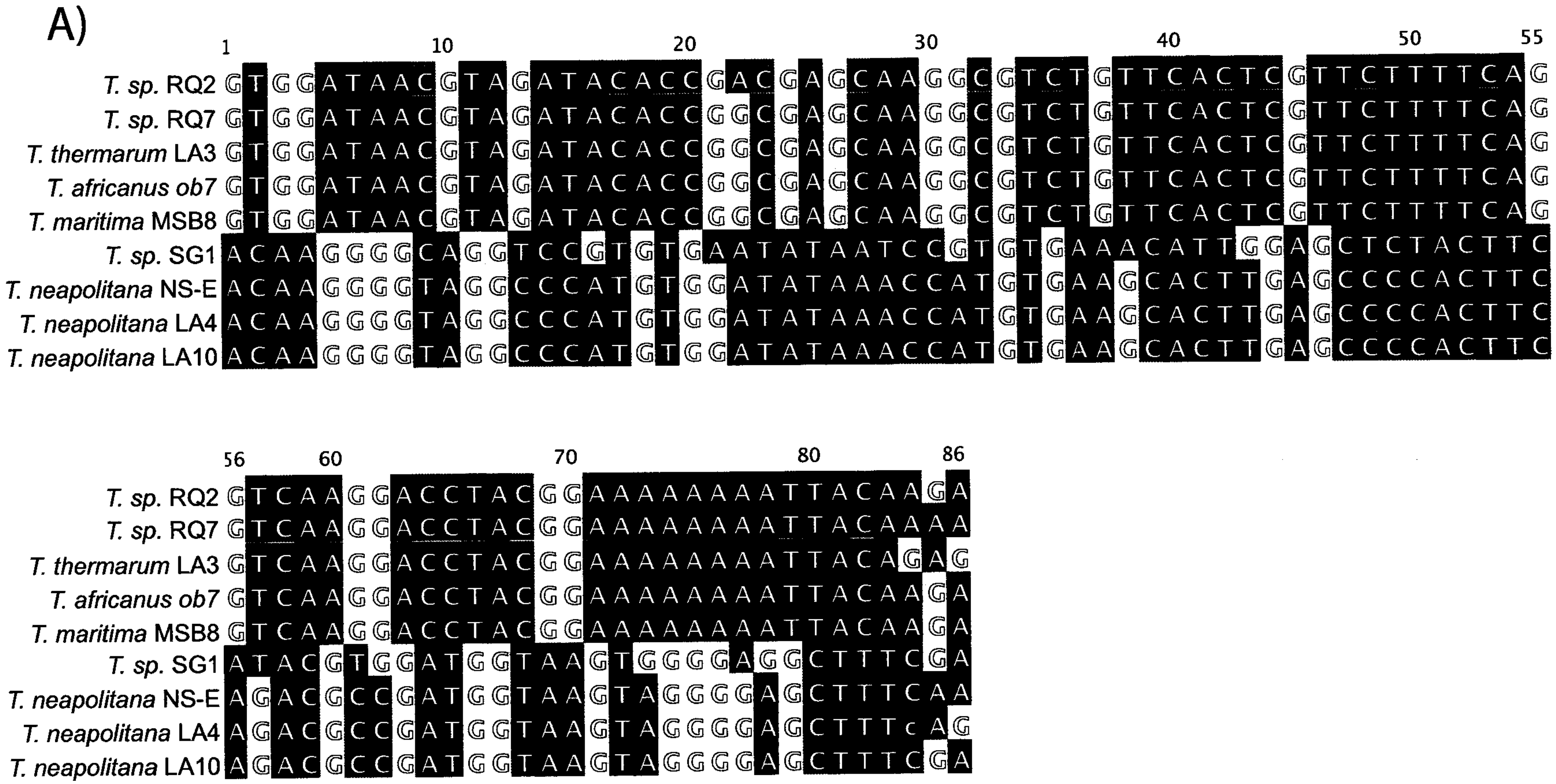
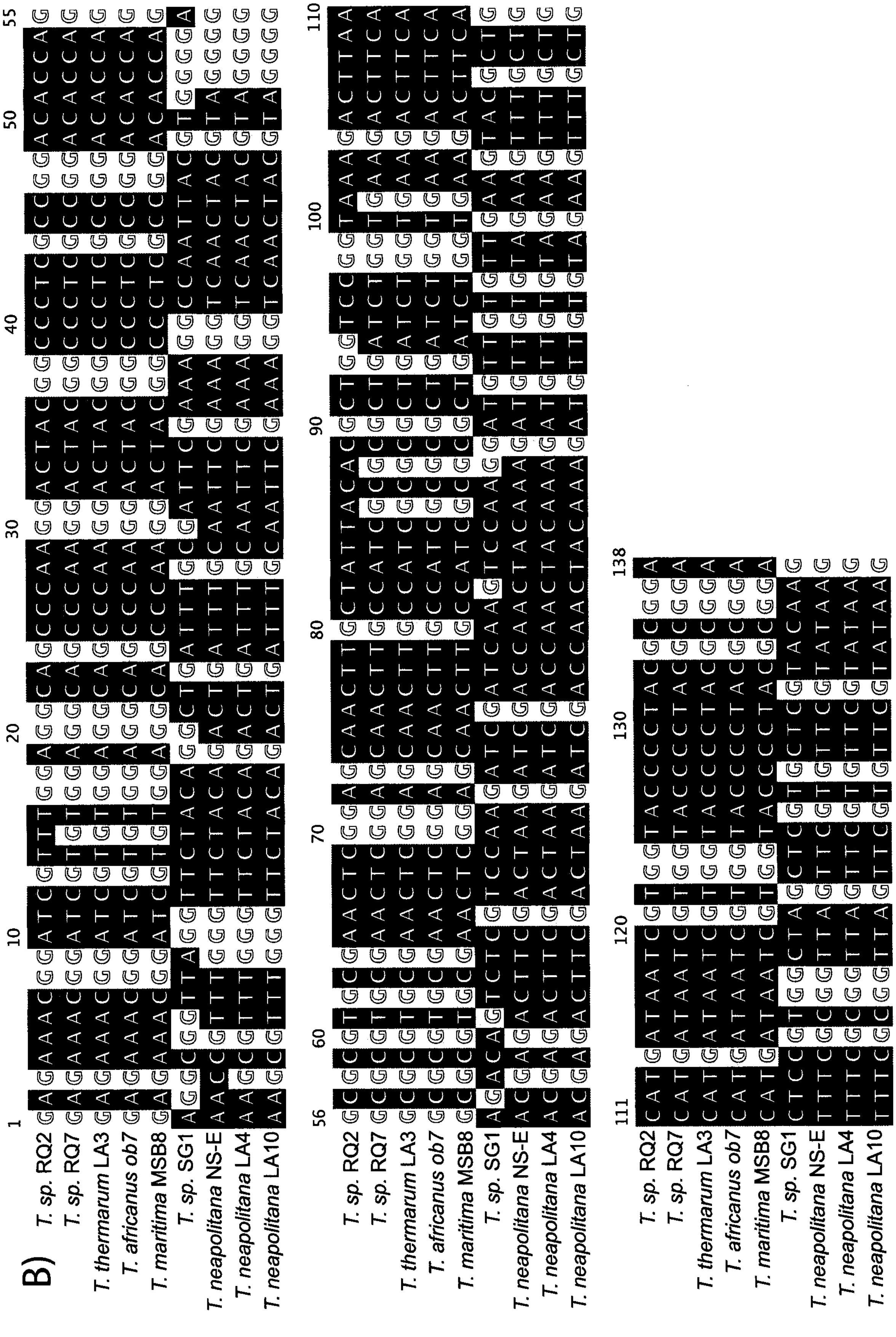
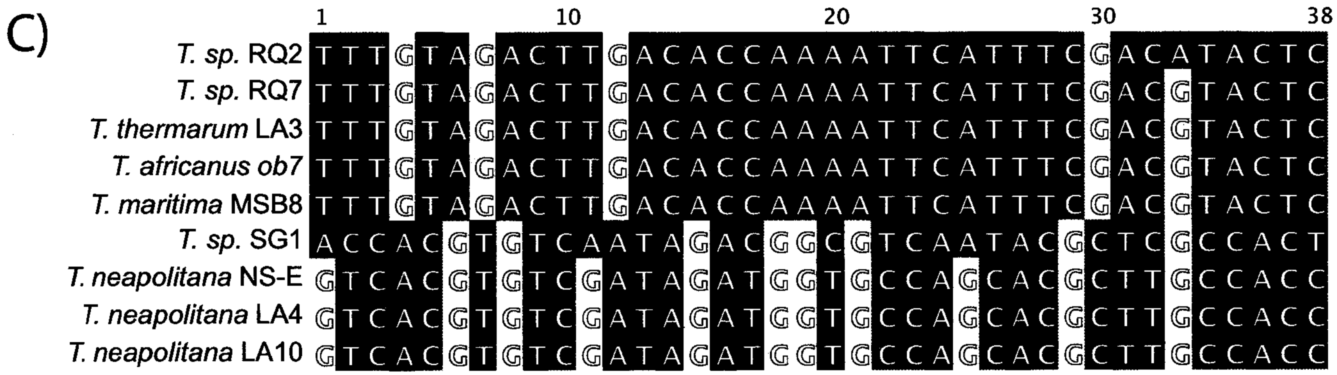

T. neapolitana NS-E
T. sp SL7
T. neapolitana LA10
T. neapolitana LA4
T. sp SG1
T. sp FJSS3B1
T. sp RQ7
T. maritima MSB8
T. sp RQ2
T. petrophila RKU-1
T. naphthophila RKU-10
T. sp Kol6
T. subterranea
T.thermarum LA3
Petrotoga miotherma
Marinitoga camini
Fervidobacterium islandicum
Fervidobacterium nodosum
Thermosipho africanus Ob7
Aquifex aeolicus
0.1 %
1%
95
99
84
100
98
95
100
100
92
84
100
L13
TM1454
CHP
L27
TM1456
THP
L21
TM1458

Figure 3.9 DNA heterogeneities found in members of the L21 ribosomal protein gene cluster in Thermotogales. A) TM1454 homologs; B) TM1455 homologs; C) TM1456 homologs; D) TM1457 homologs; E) TM1458 homologs. Visual inspection of all five gene alignments shows the strains falling into two groups: *T.* sp. RQ2 and RQ7, *T.* thermarum LA3, *T. africanus* ob7, and *T. maritima* MSB8 form one distinct clade, while *T.* sp. SG1 and *T. neapolitana* strains NS-E, LA4, and LA10 form a second clade. Two interesting recombinations are revealed – *T.* sp. RQ7, which is a strain of *neapolitana* as defined by SSU rRNA, shows nearly 100% DNA identity to *T. maritima* MSB8 and related strains. Secondly, *T. africanus* ob7 and *T.* thermarum LA3, which both show less than 90% DNA identity to *T. maritima* MSB8 in the SSU rRNA gene, show nearly 100% DNA identity in the five gene alignments presented here.

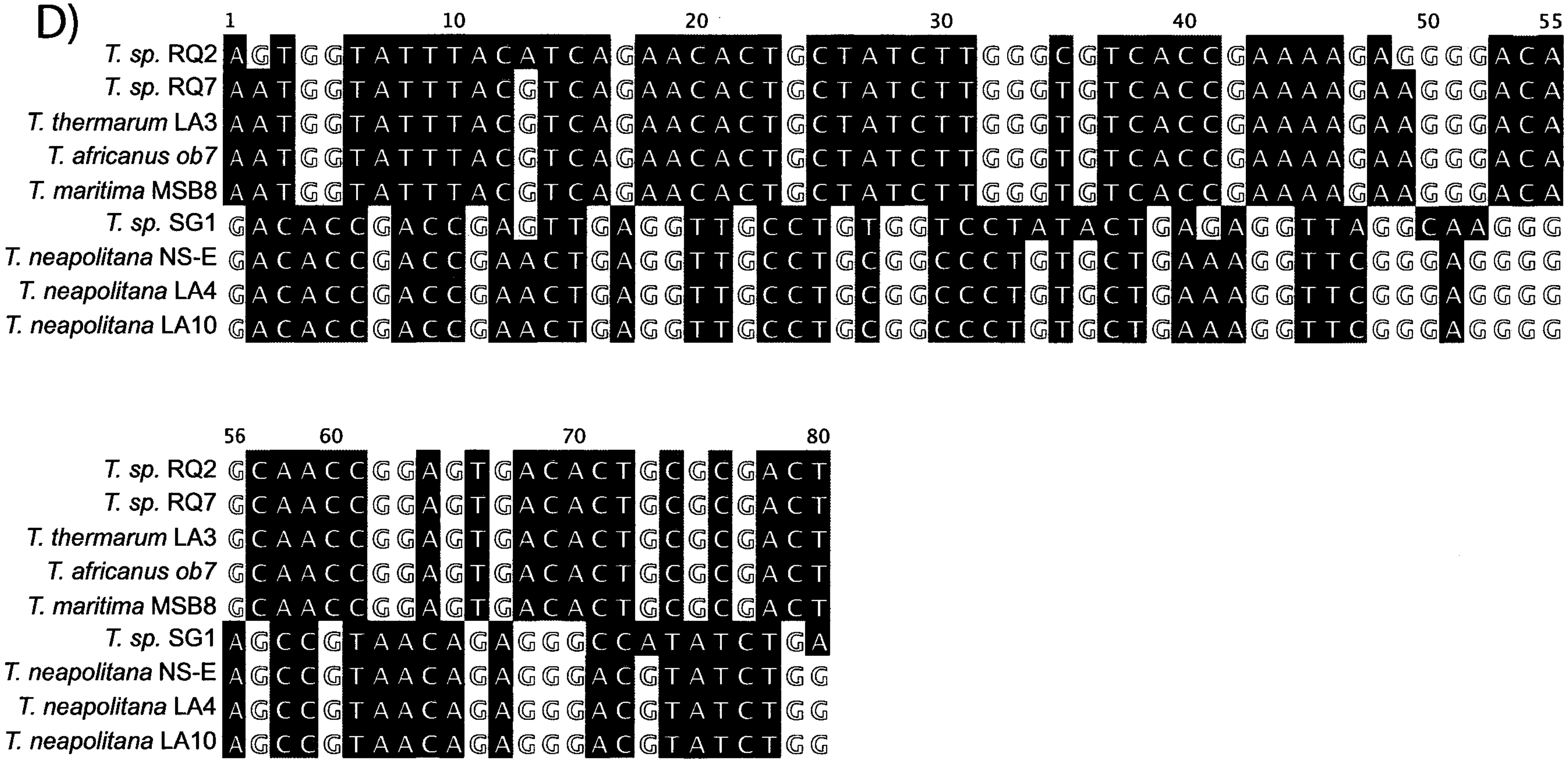

A)
1 10 20 30 40 50 55
T. sp. RQ2 GTGGATAACGTAGATACACCGACGAGCAAGGCGTCTGTTCACTCGTTCTTTTCAG
T. sp. RQ7 GTGGATAACGTAGATACACCGGCGAGCAAGGCGTCTGTTCACTCGTTCTTTTCAG
T. thermarum LA3 GTGGATAACGTAGATACACCGGCGAGCAAGGCGTCTGTTCACTCGTTCTTTTCAG
T. africanus ob7 GTGGATAACGTAGATACACCGGCGAGCAAGGCGTCTGTTCACTCGTTCTTTTCAG
T. maritima MSB8 GTGGATAACGTAGATACACCGGCGAGCAAGGCGTCTGTTCACTCGTTCTTTTCAG
T. sp. SG1 ACAAGGGGCAGGTCCGTGTGAATATAATCCGTGTGAAACATTGGAGCTCTACTTC
T. neapolitana NS-E ACAAGGGGTAGGCCCATGTGGATATAAACCATGTGAAGCACTTGAGCCCCACTTC
T. neapolitana LA4 ACAAGGGGTAGGCCCATGTGGATATAAACCATGTGAAGCACTTGAGCCCCACTTC
T. neapolitana LA10 ACAAGGGGTAGGCCCATGTGGATATAAACCATGTGAAGCACTTGAGCCCCACTTC
56 60 70 80 86
T. sp. RQ2 GTCAAGGACCTACGGAAAAAAAATTACAAGA
T. sp. RQ7 GTCAAGGACCTACGGAAAAAAAATTACAAAA
T. thermarum LA3 GTCAAGGACCTACGGAAAAAAAATTACAGAG
T. africanus ob7 GTCAAGGACCTACGGAAAAAAAATTACAAGA
T. maritima MSB8 GTCAAGGACCTACGGAAAAAAAATTACAAGA
T. sp. SG1 ATACGTGGATGGTAAGTGGGGAGGCTTTCGA
T. neapolitana NS-E AGACGCCGATGGTAAGTAGGGGAGCTTTCAA
T. neapolitana LA4 AGACGCCGATGGTAAGTAGGGGAGCTTTcAG
T. neapolitana LA10 AGACGCCGATGGTAAGTAGGGGAGCTTTCGA


B)
T. sp. RQ2
T. sp. RQ7
T. thermarum LA3
T. africanus ob7
T. maritima MSB8
T. sp. SG1
T. neapolitana NS-E
T. neapolitana LA4
T. neapolitana LA10
1
10
20
30
40
50
55
56
60
70
80
90
100
110
111
120
130
138



C)

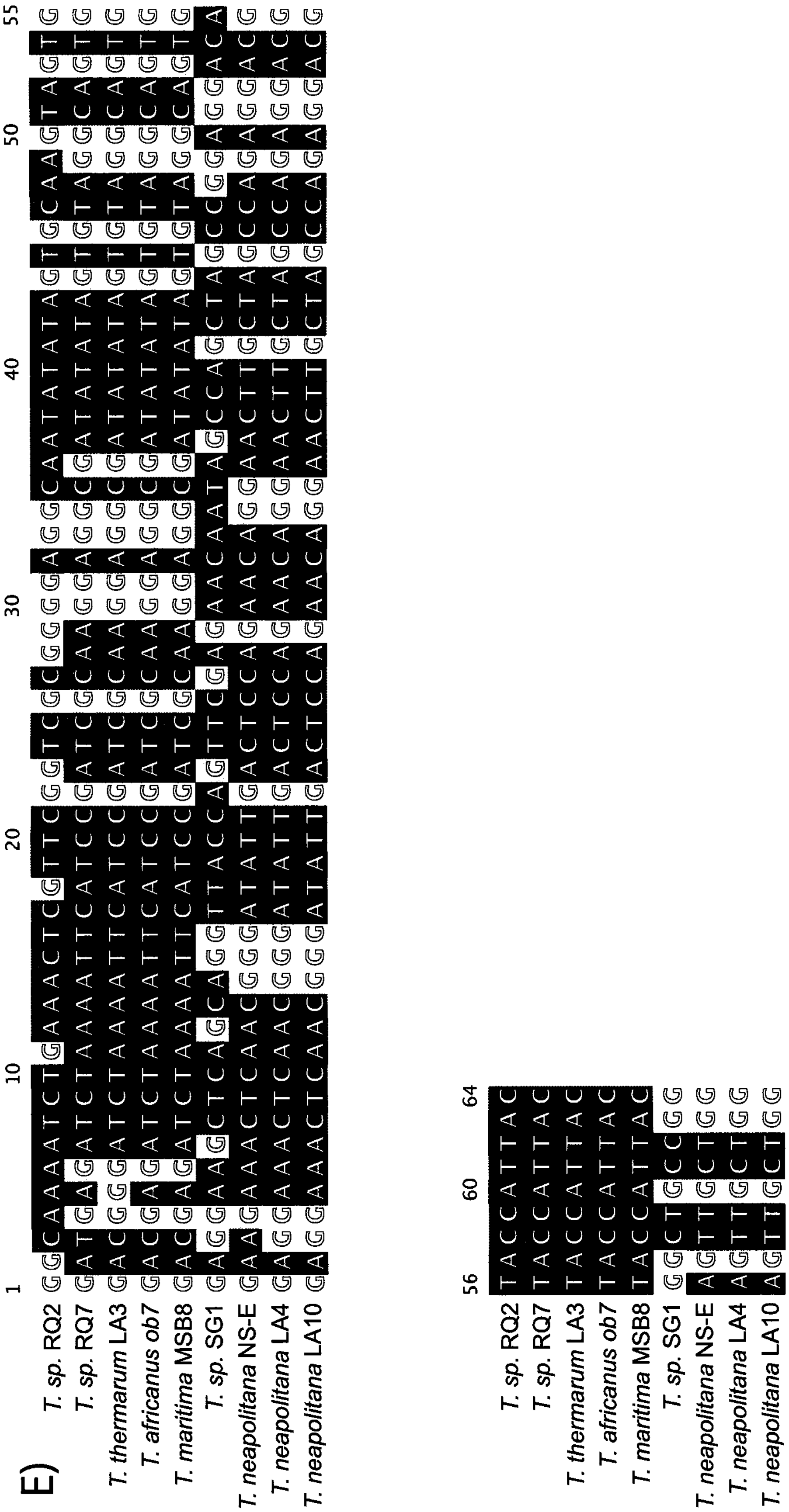
```
                          1        10        20        30      38
T. sp. RQ2               TTTGTAGACTTGACACCAAAATTCATTTCGACATACTC
T. sp. RQ7               TTTGTAGACTTGACACCAAAATTCATTTCGACGTACTC
T. thermarum LA3         TTTGTAGACTTGACACCAAAATTCATTTCGACGTACTC
T. africanus ob7         TTTGTAGACTTGACACCAAAATTCATTTCGACGTACTC
T. maritima MSB8         TTTGTAGACTTGACACCAAAATTCATTTCGACGTACTC
T. sp. SG1               ACCACGTGTCAATAGACGGCGTCAATACGCTCGCCACT
T. neapolitana NS-E      GTCACGTGTCGATAGATGGTGCCAGCACGCTTGCCACC
T. neapolitana LA4       GTCACGTGTCGATAGATGGTGCCAGCACGCTTGCCACC
T. neapolitana LA10      GTCACGTGTCGATAGATGGTGCCAGCACGCTTGCCACC
```



D)

```
                         1        10        20        30        40        50   55
T. sp. RQ2               AGTGGTATTTACATCAGAACACTGCTATCTTGGGCGTCACCGAAAAGAGGGGACA
T. sp. RQ7               AATGGTATTTACGTCAGAACACTGCTATCTTGGGTGTCACCGAAAAGAAGGGACA
T. thermarum LA3         AATGGTATTTACGTCAGAACACTGCTATCTTGGGTGTCACCGAAAAGAAGGGACA
T. africanus ob7         AATGGTATTTACGTCAGAACACTGCTATCTTGGGTGTCACCGAAAAGAAGGGACA
T. maritima MSB8         AATGGTATTTACGTCAGAACACTGCTATCTTGGGTGTCACCGAAAAGAAGGGACA
T. sp. SG1               GACACCGACCGAGTTGAGGTTGCCTGTGGTCCTATACTGAGAGGTTAGGCAAGGG
T. neapolitana NS-E      GACACCGACCGAACTGAGGTTGCCTGCGGCCCTGTGCTGAAAGGTTCGGGAGGGG
T. neapolitana LA4       GACACCGACCGAACTGAGGTTGCCTGCGGCCCTGTGCTGAAAGGTTCGGGAGGGG
T. neapolitana LA10      GACACCGACCGAACTGAGGTTGCCTGCGGCCCTGTGCTGAAAGGTTCGGGAGGGG

                         56  60        70        80
T. sp. RQ2               GCAACCGGAGTGACACTGCGCGACT
T. sp. RQ7               GCAACCGGAGTGACACTGCGCGACT
T. thermarum LA3         GCAACCGGAGTGACACTGCGCGACT
T. africanus ob7         GCAACCGGAGTGACACTGCGCGACT
T. maritima MSB8         GCAACCGGAGTGACACTGCGCGACT
T. sp. SG1               AGCCGTAACAGAGGGCCATATCTGA
T. neapolitana NS-E      AGCCGTAACAGAGGGACGTATCTGG
T. neapolitana LA4       AGCCGTAACAGAGGGACGTATCTGG
T. neapolitana LA10      AGCCGTAACAGAGGGACGTATCTGG
```


E)
T. sp. RQ2
T. sp. RQ7
T. thermarum LA3
T. africanus ob7
T. maritima MSB8
T. sp. SG1
T. neapolitana NS-E
T. neapolitana LA4
T. neapolitana LA10
1
10
20
30
40
50
55
56
60
64

seen by eye, and several strains show ORF history that differs from that of their SSU rRNA genes: 1) *T.* sp. RQ7, which according to SSU rRNA belongs to the *neapolitana* strain group, shows extraordinarily high percent identity (99%) in all five ORFS, to the homologs from *T. maritima* MSB8 and *T.* sp. RQ2, two *maritima* strains; 2) *T. thermarum* LA3 and *T. africanus* ob7, both of which sit far outside the *maritima/neapolitana* clade in the SSU rRNA tree, show nearly 100% identity to members of the *maritima* clade. This high level of conservation far exceeds that found in both the SSU rRNA sequence, and flagellar protein gene sequences, presented above. Also, the history of all genes within the L21 ribosomal protein gene cluster is different again from the SSU rRNA and flagellar protein genes.

3.2.3: Conservation of the L21 Ribosomal Protein Gene Cluster in Other Bacteria

Homologs of the L21 ORFs, including the ORFans, were found in bacterial genomes from the TIGR database, in a similar manner to the *s10*, *spc*, and *alpha* clusters. The clustering in all bacteria, compared to Thermotogales, is presented in Figure 3.10. L21 homologs were found in 263 of 266 genomes (98.9%), L27 homologs were found in all 266 (100%), and L13 homologs were found in 265 (99.6%) (but missing from a genome that was incompletely sequenced at the time of analysis). TM1455, the conserved hypothetical protein, had only 20 homologs in the database (i.e. was present in 7.5% of sequenced genomes), and TM1457 retained its status as a singleton ORFan, having no homologs. Conservation in archaea was too low to include in the analysis.

The conservation of position of the members of this cluster is less than would be expected. In 181 of the genomes (68.1%), L21 and L27 are found adjacent to one

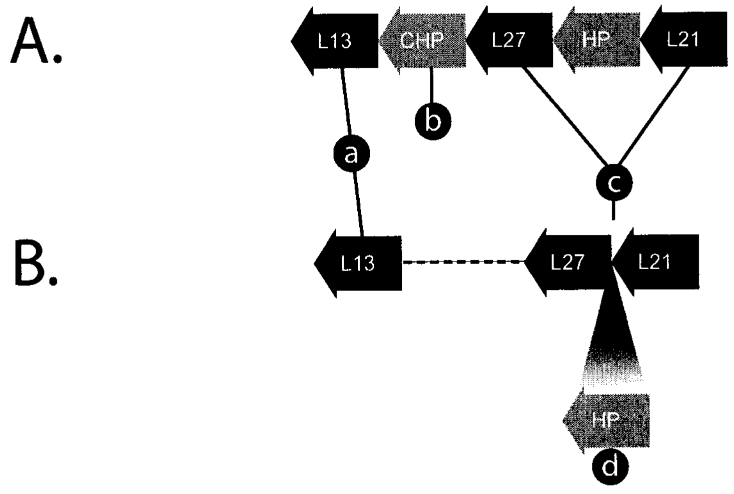


Figure 3.10 Conservation of the L21 ribosomal protein gene cluster in bacteria. A) L21 cluster in the Thermotogales. B) L21 cluster, as conserved in other bacteria. (a) L13 homologs are found in 99.6% of other sequenced bacterial genomes, but are never found associated with other ribosomal protein genes. (b) The conserved hypothetical protein coded for by TM1455 (indicated here by "CHP") had homologs in only 7.5% of bacterial genomes analyzed, and was also never found associated with ribosomal protein genes. (c) L21 and L27 were found to be adjacent in 68.1% of bacterial genomes analyzed, but in the remainder of the genomes, they were found within two genes of each other. (d) While homologs of TM1457 (indicated here by "HP") were never found, in 30.5% of the genomes analyzed, there was a place holder ORF in this location.

another, while in an additional 81 genomes (30.5%), they are found with anywhere between one and three intervening ORFS of various identities. The remaining four additional genomes either contain only an L27 homolog (three genomes) or have L21 and L27 completely unassociated (one genome). In none of the 20 genomes with homologs of TM1455 did this ORF associate with any ribosomal protein genes.

3.3: Spatial Autocorrelation of Functional Categories

Because of the plasticity of prokaryotic genomes, one can think of the individual genes as members of a dynamic population; for example, they interact, migrate, and undergo birth and death processes. The level of community interaction among multicellular organisms is often assessed using various spatial autocorrelation methods, one of which was successfully applied here to the 'gene' members of the 'genome' community.

3.3.1: Use of Joint Count Statistics to Map Physical Distribution of Gene Categories

Spatial autocorrelation and physical distribution of gene categories within prokaryotic genomes was easily be visualized using joint count statistics. Three sample datasets were generated, for random, hyperdispersed, and clustered distributions (See Table 3.3). The distribution of each category is visible by simply looking at the data, but the exaggeration of each type of possible result gives clear visualization of randomization, hyperdispersal and clustering. The output from Genespat v.4 was used to plot the distributions, which are presented in Figure 3.11. While the plot for a randomly distributed gene category does occasionally spike >|1.96|, cycling around a value of zero is clear. Hyperdispersed genes, in this case separated by four genes of another category, show a clear cycling from a significant negative value (indicating that genes are never found separated by these distances) to a significant positive value (indicating that genes within this category are always found separated by that specific distance). The plot for clustered gene categories indicates large positive z-scores at smaller distance classes, as genes within clusters would be found adjacent or nearly adjacent, with a steady decline toward zero.

Table 3.3 Sample datasets used for Genespat v.4 analysis. Three datasets were generated, with random, hyperdispersed, and clustered distributions. Five "gene" categories were used for each dataset. Only the first 50 positions are presented here; the full datasets used for calculating join count statistics, and for generating the sample plots in Figure 3.11. Each had 200 "genes"; the random "genome" was random throughout, whereas the hyperdispersed and clustered "genomes" simply repeated the first 50 genes an additional three times, for a total of 200 "genes".

gene position	random genome	hyper-dispersed genome	clustered genome
1	1	1	1
2	3	2	1
3	1	3	1
4	2	4	1
5	4	5	1
6	2	1	1
7	4	2	1
8	5	3	1
9	4	4	1
10	2	5	1
11	3	1	2
12	4	2	2
13	3	3	2
14	4	4	2
15	1	5	2
16	2	1	2
17	3	2	2
18	2	3	2
19	1	4	2
20	3	5	2
21	5	1	3
22	2	2	3
23	1	3	3
24	1	4	3
25	2	5	3
26	3	1	3
27	1	2	3
28	4	3	3
29	2	4	3

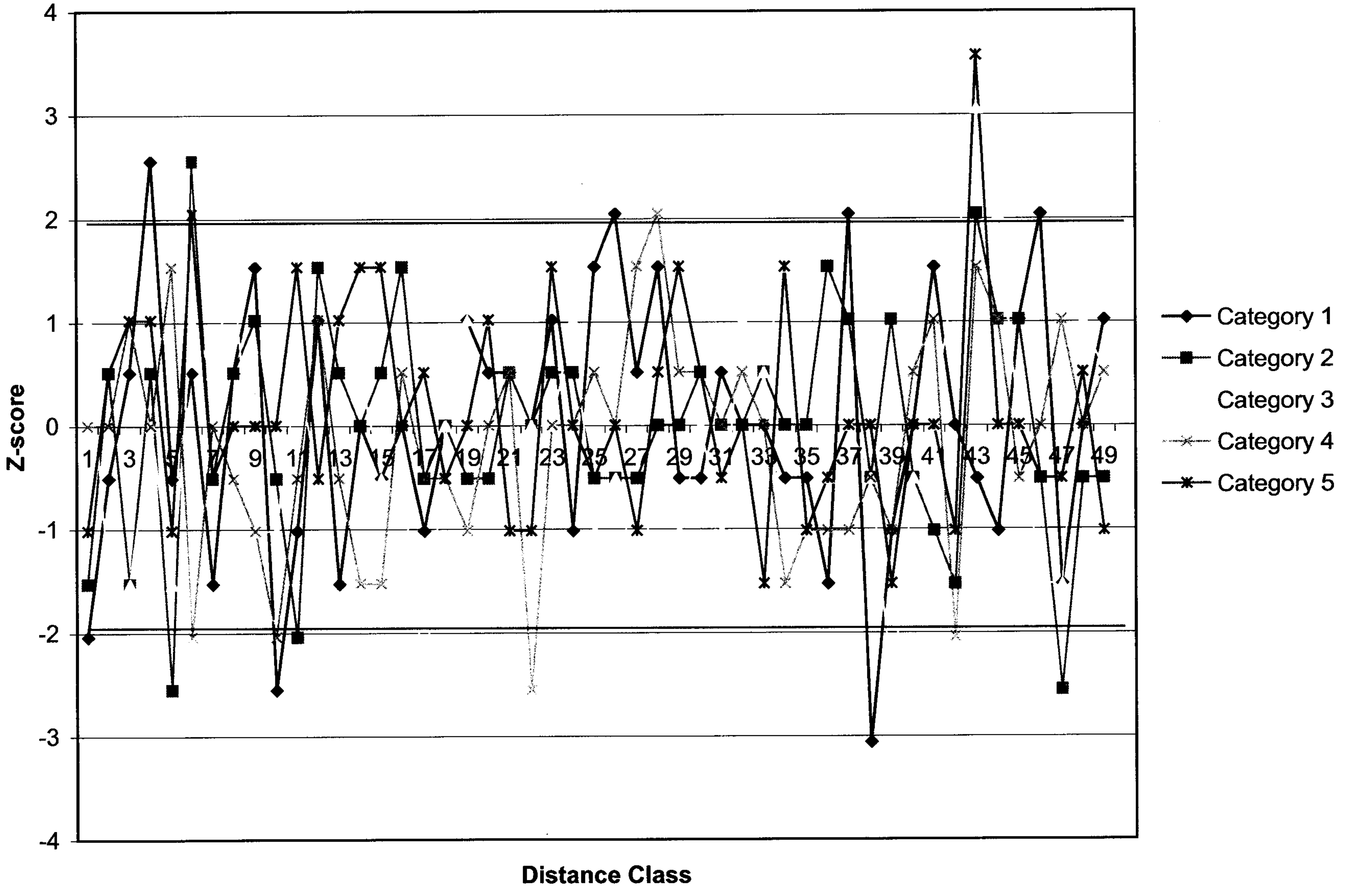
gene position	random genome	hyper-dispersed genome	clustered genome
30	4	5	3
31	2	1	4
32	5	2	4
33	3	3	4
34	3	4	4
35	1	5	4
36	5	1	4
37	2	2	4
38	1	3	4
39	2	4	4
40	4	5	4
41	4	1	5
42	1	2	5
43	5	3	5
44	1	4	5
45	3	5	5
46	3	1	5
47	5	2	5
48	4	3	5
49	5	4	5
50	1	5	5

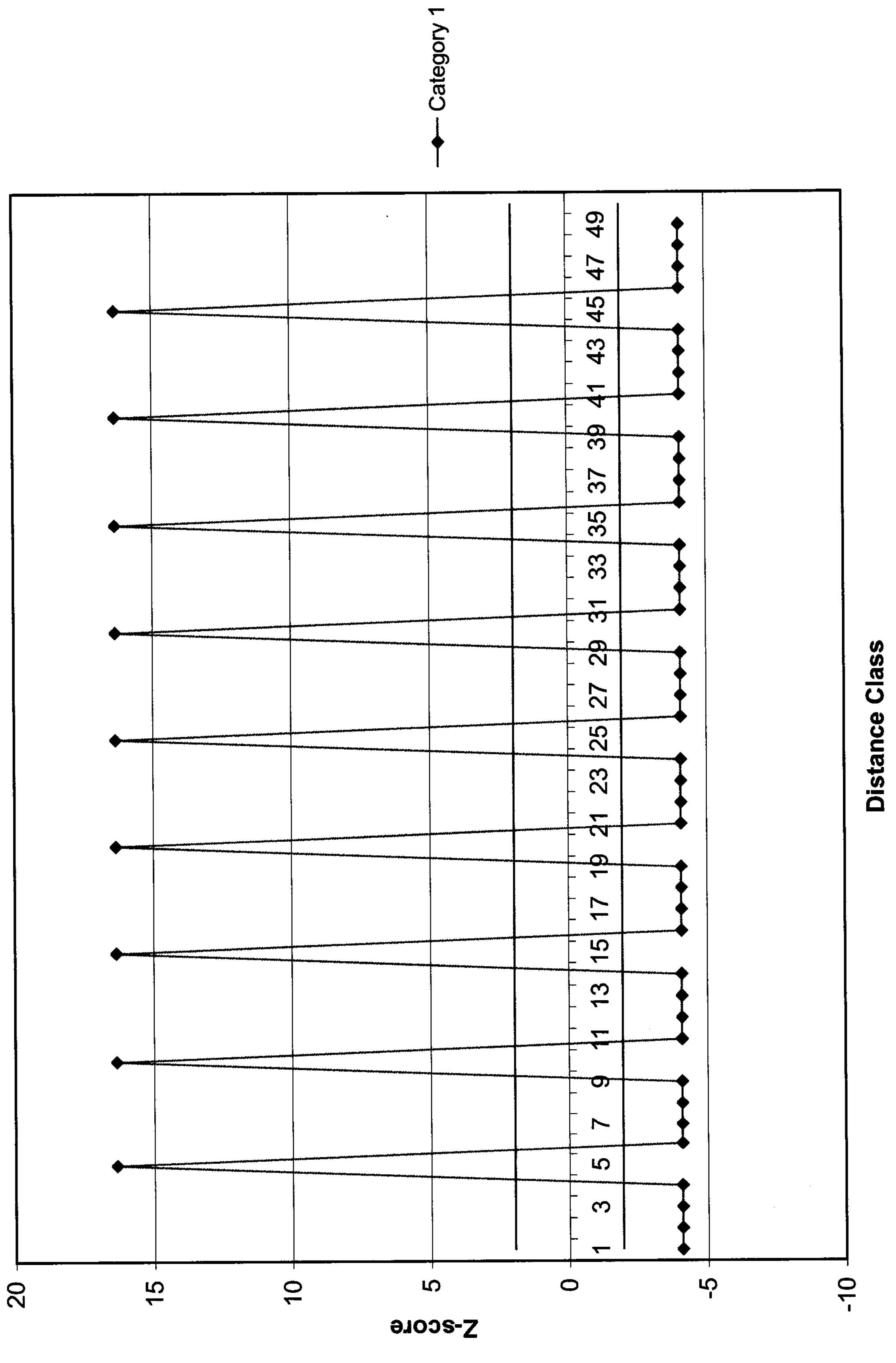
3.3.2: Functional Gene Category Distribution Within *T. maritima* MSB8

Joint count statistics were computed initially for all functional categories present in *T. maritima* MSB8, and are presented in Appendix 1. The graphical representations were used to determine the physical distribution for each category as laid out in Materials and Methods, are summarized in Table 3.4 and are presented in Figure 3.12.

The following categories showed a random distribution in the MSB8 genome: DNA metabolism, Conserved Hypothetical Proteins, Protein fate, Regulatory functions, and Unknown function. The following categories showed significant physical clustering (approximate cluster size given in parentheses): Amino acid biosynthesis (10), Biosynthesis of cofactors, prosthetic groups, and carriers (5), Cell envelope (8), Cellular processes (2), Central intermediary metabolism (3), Energy metabolism (13), Mobile and extrachromosomal element functions (26), Protein synthesis (53*), Purines, pyrimidines, nucleosides, and nucleotides (8), Transcription (3), Transport and binding proteins (23*). Of particular interest was the physical pattern of potential ORFans within the genome.

Figure 3.11 Spatial autocorrelation plots using generated sample datasets. Fictional data were generated, and are presented in Table 3.4. To visualize the distribution for the "gene categories" in the sample datasets, the z-score, as calculated by Genespat v.4, is plotted against distance category. A) *Randomly distributed gene categories* Plots for all five categories cycle around zero, demonstrating that there is no positive or negative association of members within any particular category. B) *Hyperdispersed gene categories* Only one category is shown, as all five are uniformly distributed throughout the "genome". The extreme values, cycling from -4.08 (for distance classes 1 through 4) up to 16.33 for distance class 5, demonstrate that the "genes" within each category are never found adjacent or within two, three or four genes of one another (significant negative value), but genes are extremely likely to be found five genes apart (significant positive value). C) *Clustered gene categories* The significant positive value, decreasing to zero, demonstrates clustering behavior of these gene categories. The X-intercept gives an approximate size of the gene clusters within the genome – here, it crosses at distance class 8, although we know a priori that the clusters all contain ten "genes" Significance cutoff of >|1.96| is indicated on all graphs by thin red lines.


4
3
2
1
0
-1
-2
-3
-4
Z-score
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance Class
Category 1
Category 2
Category 3
Category 4
Category 5


Z-score
20
15
10
5
0
-5
-10
Distance Class
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Category 1

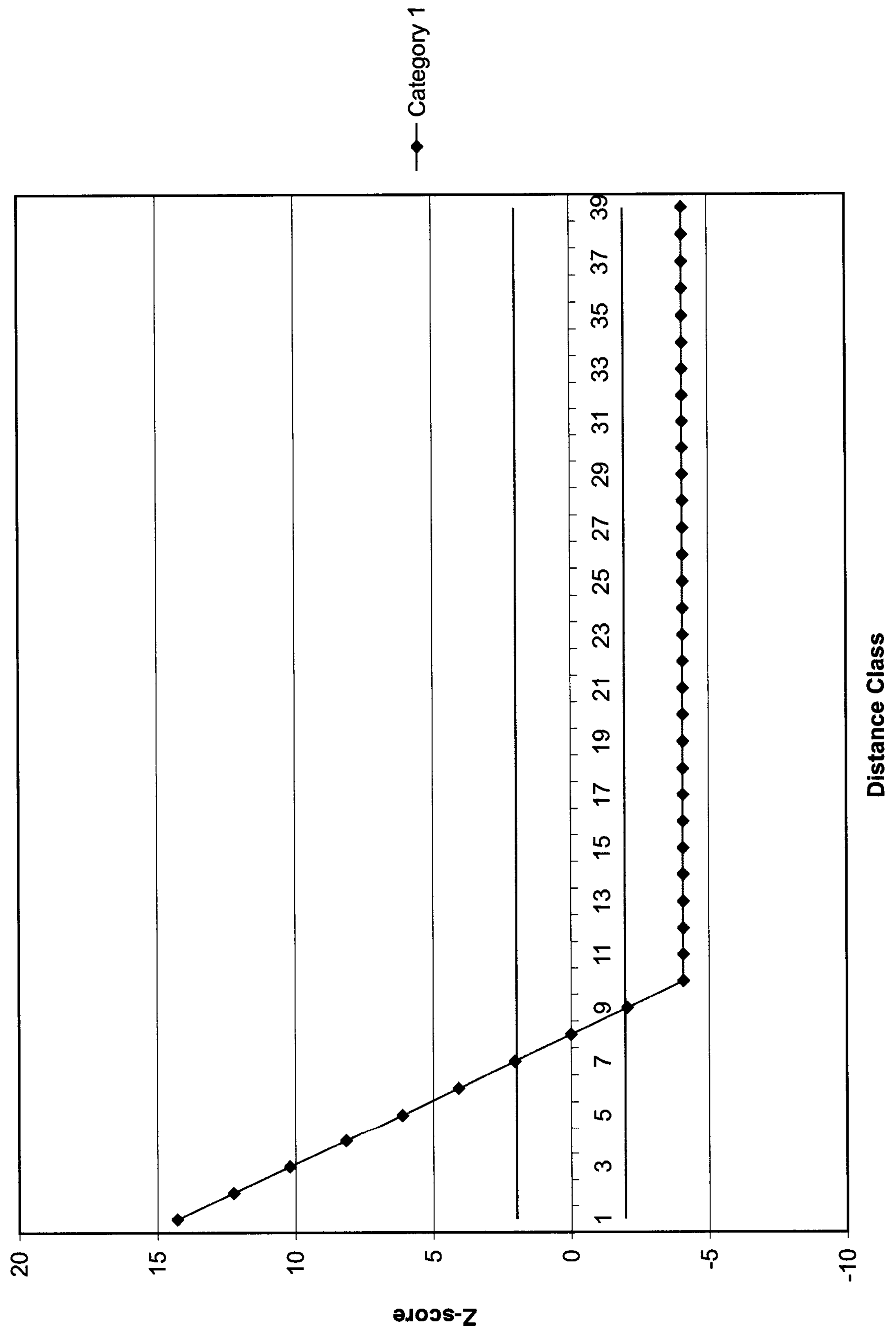

Z-score
20
15
10
5
0
-5
-10
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Distance Class
Category 1

Table 3.4 Summary of the physical distribution of functional gene categories in *Thermotoga maritima* MSB8. The following categories did not have distributions calculated: Disrupted reading frame (only three ORFs); Glimmer rejects, Signal transduction, Viral functions, and Unclassified all were absent from the genome.

Functional category	# of ORFs	% of annotated ORFs[1]	type of distribution[2]	cluster size[3]	lower bound of cluster size
Amino acid biosynthesis	73	0.04138	clustered	10	
Biosynthesis of cofactors, prosthetic groups, and carriers	32	0.01814	clustered	5	
Cell envelope	73	0.04138	clustered	8	
Cellular processes	49	0.02778	clustered	2	
Central intermediary metabolism	45	0.02551	clustered	3	
Disrupted reading frame	3	0.0017	N/A	N/A	
DNA metabolism	55	0.03118	random		
Energy metabolism	197	0.11168	clustered	13	
Fatty acid and phospholipid metabolism	15	0.0085	random		
Glimmer rejects	0	0	N/A	N/A	
Conserved Hypothetical proteins	381	0.21599	random	N/A	
Mobile and extrachromosomal element functions	29	0.01644	clustered	26	
True hypothetical proteins	251	0.14229	hyperdispersed	N/A	
Protein fate	49	0.02778	random	N/A	
Protein synthesis	108	0.06122	clustered	53*	39
Purines, pyrimidines, nucleosides, and nucleotides	45	0.02551	clustered	8	
Regulatory functions	71	0.04025	random	N/A	
Signal transduction	0	0	N/A	N/A	
Transcription	16	0.00907	clustered	3	
Transport and binding proteins	190	0.10771	clustered	23*	3
Unclassified	0	0	N/A	N/A	
Unknown function	82	0.04649	random	N/A	
Viral functions	0	0	N/A	N/A	

Figure 3.12 Physical distribution plots of functional gene categories within *Thermotoga maritima* MSB8. A) Functional gene categories with a random distribution; B) Functional cagegories with a hyper-dispersed distribution; C) Functional gene categories with a clustered distribution. Significance cutoffs of >|1.96| are indicated by thin red lines.

Key: Amino acid biosynthesis: Cat. 1

Biosynthesis of cofactors, prosthetic groups, and carriers: Cat. 2

Cell envelope: Cat. 3

Cellular processes: Cat. 4

Central intermediary metabolism: Cat. 5

DNA metabolism: Cat. 6

Disrupted reading frame: Cat. 7

Energy metabolism: Cat. 8

Fatty acid and phospholipid metabolism: Cat. 9

Glimmer rejects: Cat. 10

Conserved hypothetical proteins: Cat. 11

Mobile and extrachromosomal element functions: Cat. 12

True hypothetical proteins: Cat. 13

Protein fate: Cat. 14

Protein synthesis: Cat. 15

Purines, pyrimidines, nucleosides, and nucleotides: Cat. 16

Regulatory functions: Cat. 17

Signal transduction: Cat. 18

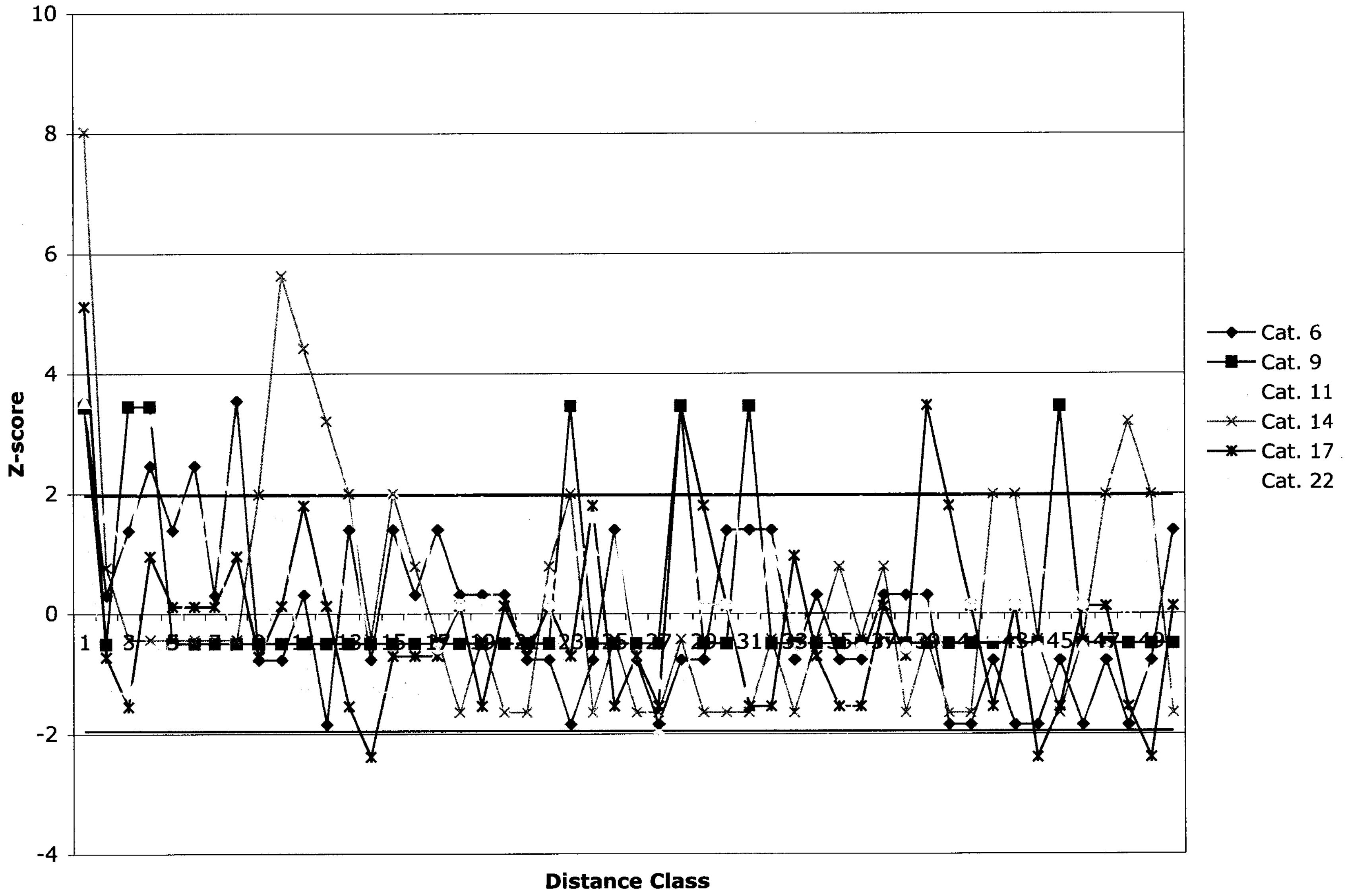
Transcription: Cat. 19

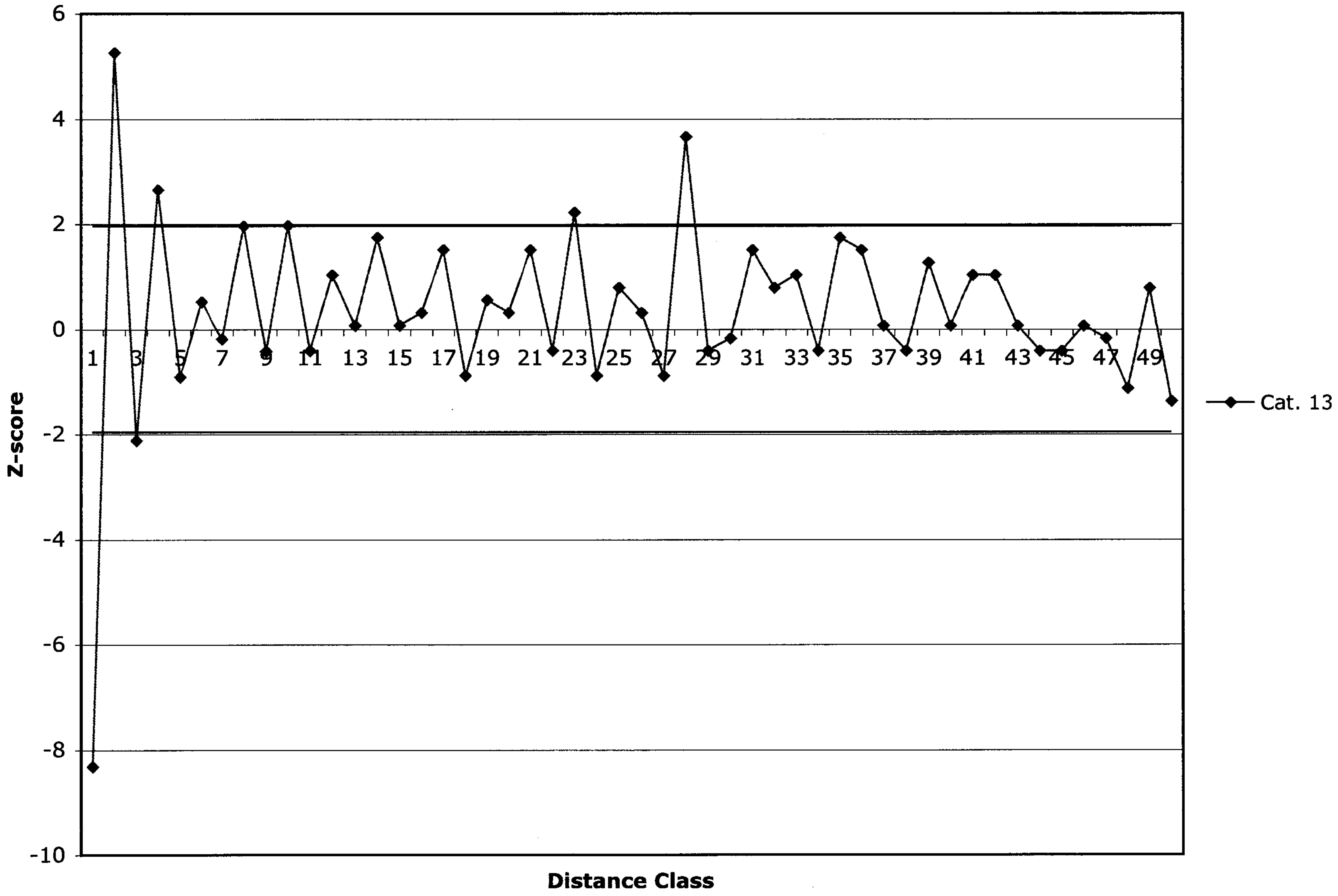
Transport and binding proteins: Cat. 20

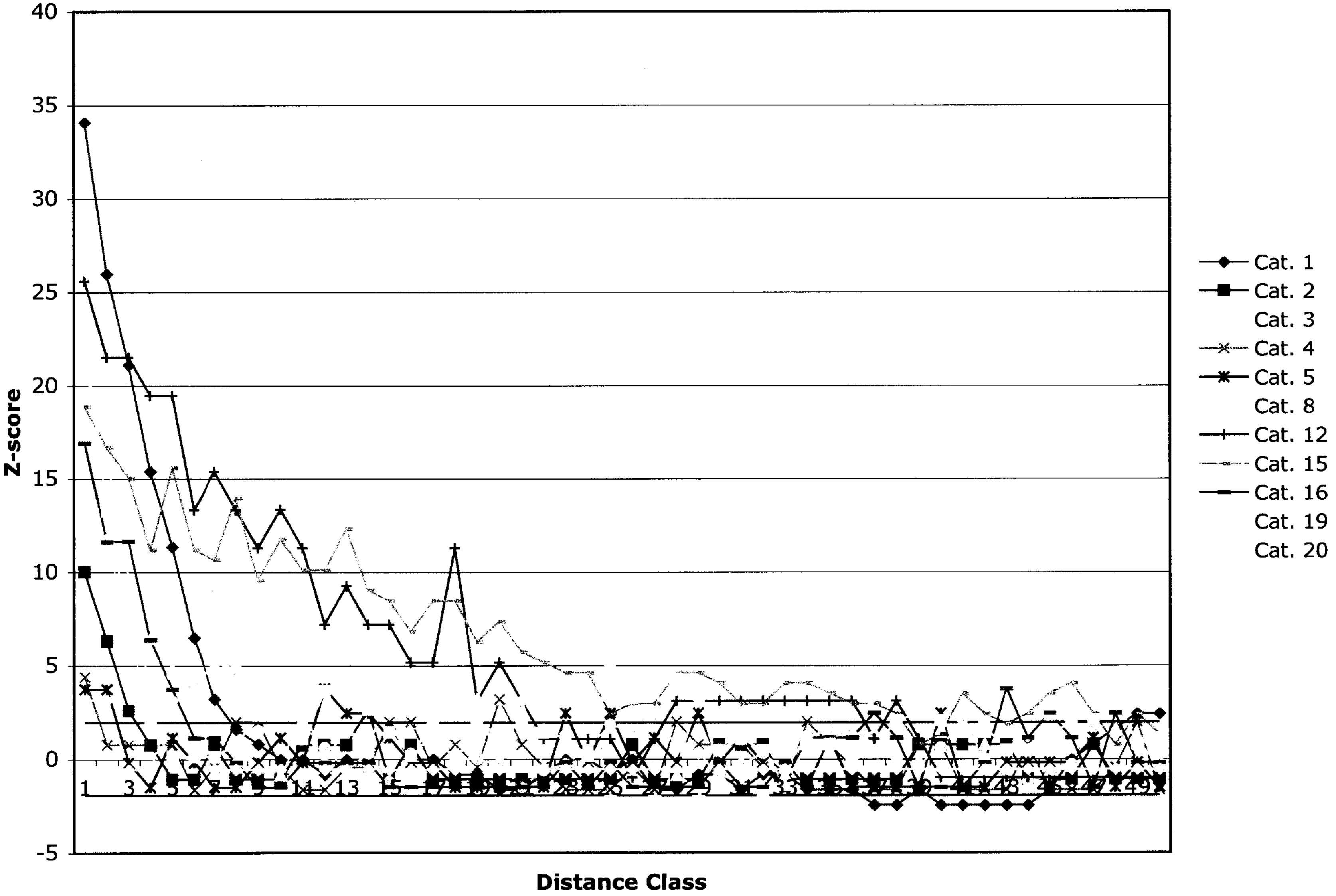
Unclassified: Cat. 21

Unknown function: Cat. 22

Viral functions: Cat. 23


10
8
6
4
2
0
-2
-4
Z-score
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance Class
Cat. 6
Cat. 9
Cat. 11
Cat. 14
Cat. 17
Cat. 22


6
4
2
0
-2
-4
-6
-8
-10
Z-score
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance Class
Cat. 13


40
35
30
25
20
15
10
5
0
-5
Z-score
1
3
5
13
15
17
29
33
43
49
Distance Class
Cat. 1
Cat. 2
Cat. 3
Cat. 4
Cat. 5
Cat. 8
Cat. 12
Cat. 15
Cat. 16
Cat. 19
Cat. 20

The two categories – conserved hypothetical proteins and true hypothetical proteins – showed different distributions.

The conserved hypothetical proteins, which comprise 21.6% of the genome, showed patterning similar to categories with a random distribution (Figure 3.11A), indicative of proteins with as yet unknown functions that likely operate within anestablished cluster or operon, performing or substituting for a known function. This is as opposed to being significantly clustered, which might indicate groups of CHPs operating as a single gene cluster or operon of unknown or unique function.

The true hypothetical proteins, or singleton ORFans, show a hyperdispersed distribution (Figure 3.11B). In this case, ORFs showed a significant negative association at a distance class of 1, indicating that they are rarely, if ever, found adjacent to one another, followed by a significant positive association at distance class 2, demonstrating a tendency to be found with one gene intervening. This up-and-down cycling continues up until distance class 15, at which point the precise cycling degrades. In actual genome data (as compared to generated datasets) signal degrades in both clustering and hyperdispersed categories because positive and negative association tend to happen in close proximity. To maintain the strong cycling as shown in Figure 3.11B, the members of this category would need to be functioning as intervening sequence throughout the whole genome.

3.3.3: Functional Gene Category Distribution Within Other Bacteria

Because there is presently only one strain from the order Thermotogales present in the sequence database, this analysis was performed on an additional 25 genomes. By choosing groups with more than one sequenced strain, more closely related strains could be compared, as well as group trends.

Analysis was performed on the following groups of organisms, all downloaded from the TIGR Comprehensive Microbial Resource: *Bacillus anthracis* (8 strains), *Campylobacter jejuni* (2 strains), *Chlamydia pneumoniae* (4 strains), *Escherichia coli* (4 strains), *Legionella pneumophila* (3 strains), and *Prochlorococcus marinus* (5 strains).

The output from Genespat v.4 was generated for each organism, and summarized by species-group in Appendix 2. Gene frequency output for functional categories within each group is presented in Appendix 2, and summarized below.

Gene frequencies for the eight sequenced strains of B. *anthracis* are comparable within most role categories, with the exception of central intermediary metabolism, conserved and true hypothetical proteins, and unclassified or proteins of unknown function. C. *pneumoniae* strains show similar patterns in the conserved and true hypothetical proteins, with the frequency of these functional categories showing the highest variation, along with energy metabolism. The four *E. coli* strains examined showed more variability, in the conserved and true hypothetical proteins, as well as cell envelope, mobile and extrachromosomal elements, protein synthesis, regulatory functions, unclassified and unknown functions, as well as viral functions (variability here may be due to the absence of these categories in one or more of the strains analyzed). The three *L. pneumophila* strains showed little to no variability, with the exception of the conserved hypothetical proteins, at a very low level. The five *P. marinus* strains showed variation in cell envelope, DNA metabolism, both conserved and true hypothetical proteins, regulatory functions, transport and binding proteins, as well as unclassified and unknown function ORFs.

Joint count statistics for each of the groups are presented in Appendix 3 and 4. Of particular interest is the distribution of both conserved and true hypothetical proteins, as compared both within and between strain groups. Plots for all strains are presented in Appendix 3 and 4, while characteristics are summarized in Table 3.5, and explained below.

Bacillus anthracis: Both the conserved hypothetical proteins and true hypothetical proteins show significant self-association, or clustering with *B. anthracis* Ames and Ames ancestor standing out slightly. All strains show clustering of CHPs at a size of much greater than 5 ORFs, but the signal for both Ames and Ames ancestor decays faster, indicating a cluster size of 4 ORFs. The opposite pattern is seen in the true hypothetical proteins, with Ames and Ames ancestor showing a cluster size of greater than 7 ORFs, while the rest of the strains show a cluster size of around 4 ORFs. Also, *B. anthracis* Sterne contains no true hypothetical proteins.

Campylobacter jejuni: Conserved hypothetical proteins in both *C. jejuni* strains don't show much clustering beyond one or two ORFs, before the signal degrades into randomness. However, *C. jejuni* NCTC11168 has only three ORFs annotated as THPs, while *C. jejuni* RM1221 has large clusters of THPs, up to 46 genes long.

Chlamydia pneumoniae: Conserved hypothetical proteins in C. *pneumoniae* strains show some clustering, ranging in size from 2 to 10 ORFs. Strain J138 has the largest cluster size at 10 ORFs, strain AR39 has clusters of 6, strain TW183 has clusters of 5, and strain CWL029 has clusters of 2. The true hypothetical proteins show quite distinct patterns

Table 3.5 Summary of the physical distribution of A) conserved hypothetical proteins and B) true hypothetical proteins in six groups of closely related bacterial genomes.
[1]% of ORFs annotated, taking into account that several ORFs may be annotated into more than one functional category
[2]as defined in Materials and Methods - either random, hyperdispersed, or clustered
[3]approximate cluster size based on curve shape and X-intercept
*clusters are likely smaller, since they decay into noise earlier than the x-intercept, so lower bound is given (where plot crosses below significance cutoff of >|1.96|)

A) Conserved Hypothetical Proteins

Strain	# of ORF S	% of annotated ORFs[1]	type of distribution[2]	cluster size[3]	lower bound cluster size
Bacillus anthracis					
B. anthracis A0039	1211	0.21216	clustered	28*	12
B. anthracis Ames	1177	0.20502	clustered	4	
B. anthracis Ames Ancestor	1175	0.20474	clustered	4	
B. anthracis Sterne	931	0.16572	clustered	7	
B. anthracis str. France	1194	0.212	clustered	30*	10
B. anthracis str. Kruger B	1196	0.2103	clustered	30*	10
B. anthracis Vollum	1207	0.2131	clustered	29*	5
B. anthracis Western North America USA6153	1215	0.21256	clustered	28*	9
Campylobacter jejuni					
C. jejuni NCTC 11168	249	0.14343	clustered	20*	3
C. jejuni RM1221	266	0.13441	clustered	2	
Chlamydia pneumoniae					
C. pneumoniae AR39	281	0.27877	clustered	6	
C. pneumoniae CWL029	127	0.12713	clustered	2	
C. pneumoniae J138	259	0.22941	clustered	10	
C. pneumoniae TW-183	298	0.23974	clustered	5	
Escherichia coli					
E. coli CFT073	1097	0.19558	clustered	20	
E. coli K12-MG1655	949	0.21932	clustered	5	
E. coli O157:H7 EDL933	933	0.16075	clustered	41*	19
E. coli O157:H7 VT2-Sakai	1099	0.19848	clustered	16	
Legionella pneumophila					
L. pneumophila Lens	672	0.21684	clustered	45*	9
L. pneumophila Paris	744	0.22921	clustered	56*	10
L. pneumophila Philadelphia 1	633	0.19968	clustered	12*	8
Prochlorococcus marinus					
P. marinus CCMP1375	265	0.12771	random	N/A	
P. marinus CCMP1378 MED4	259	0.13303	random	N/A	
P. marinus MIT 9312	368	0.19167	random	N/A	
P. marinus MIT9313	345	0.13642	random	N/A	
P. marinus NATL2A	61	0.03062	random	N/A	

B) True Hypothetical Proteins

Strain	# of ORFS	% of annotated ORFs[1]	type of distribution[2]	cluster size[3]	lower bound of cluster size
Bacillus anthracis					
B. anthracis A0039	105	0.0184	clustered	9	
B. anthracis Ames	845	0.14719	clustered	28*	7
B. anthracis Ames Ancestor	847	0.14759	clustered	31*	7
B. anthracis Sterne	1	0.00018	N/A	N/A	
B. anthracis str. France	95	0.01687	clustered	10	
B. anthracis str. Kruger B	105	0.01846	clustered	10	
B. anthracis Vollum	99	0.01748	clustered	13	
B. anthracis Western North America USA6153	103	0.01802	clustered	10	
Campylobacter jejuni					
C. jejuni NCTC 11168	3	0.00173	N/A	N/A	
C. jejuni RM1221	286	0.14452	clustered	46	
Chlamydia pneumoniae					
C. pneumoniae AR39	120	0.11905	hyperdist.	N/A	
C. pneumoniae CWL029	362	0.36236	clustered	6	
C. pneumoniae J138	0	0	N/A	N/A	
C. pneumoniae TW-183	3	0.00241	N/A	N/A	
Escherichia coli					
E. coli CFT073	11	0.00196	N/A	N/A	
E. coli K12-MG1655	570	0.13173	clustered	9	
E. coli O157:H7 EDL933	4	0.00069	N/A	N/A	
E. coli O157:H7 VT2-Sakai	267	0.04822	clustered	30*	23
Legionella pneumophila					
L. pneumophila Lens	2	0.00065	N/A	N/A	
L. pneumophila Paris	2	0.00062	N/A	N/A	
L. pneumophila Philadelphia 1	1	0.00032	N/A	N/A	
Prochlorococcus marinus					
P. marinus CCMP1375	404	0.1947	clustered	100*	82
P. marinus CCMP1378 MED4	276	0.14176	clustered	42*	18
P. marinus MIT 9312	71	0.03698	clustered	41*	22
P. marinus MIT9313	433	0.17121	clustered	81*	27
P. marinus NATL2A	88	0.04418	clustered	82*	60

within each strain. Strain J138, which has the largest clusters of CHPs, contains no true hypothetical proteins. Strain TW18 only contains 3 true hypothetical proteins, and they show an association of distance 10; the 3 ORFs are in a similar region of the genome, but fairly evenly spaced. Strain CWL029 has clusters of true hypothetical proteins of 6 ORFs. Perhaps most interestingly, strain AR39 shows a significant dissociation of true hypothetical proteins at distance class 1, and a significant positive association at distance class 2, similar to the true hypothetical proteins of *T. maritima* MSB8.

E. *coli*: Strain K12-MG1655 contains significant clusters of CHPs of 5 ORFs, while all three pathogenic strains have larger clusters. Strain CFT073 has clusters of 20 ORFs, strain O157:H7 VT2-Sakai has clusters of 16 ORFS, and strain O157:H7 EDL933 shows clusters of 41 ORFS (although the plot has quite a bit of noise, and the clusters are likely somewhat smaller). Both strains CFT073 and O157:H7 EDL933 both contain very few true hypothetical proteins (11 and 7), and show no association whatsoever. Strain K12-MG1655, which contains 570 true hypothetical proteins, shows clusters of size 8, while O157:H7 VT2-Sakai, which contains 267 hypothetical proteins, shows much larger clusters of true hypothetical proteins, of size 30.

L. pneumophila: All strains show some clustering of conserved hypothetical proteins, but the signal degrades into noise before an accurate estimation of the cluster size can be made. All strains also have either one or two true hypothetical proteins per genome, and so would have no clustering.

P. marinus: Conserved hypothetical proteins show no clustering in any of the *P. marinus* strains examined. Similar to *L. pneumophila* CHPs, true hypothetical proteins show significant association at close distance classes, but the signal degrades into noise before an accurate estimation of cluster size can be made.

CHAPTER 4: DISCUSSION

4.1: The Proximal Flagellar Cluster (PFC) is Maintained Within a Group of Thermotogales

Using a modified long walk PCR technique, unknown regions downstream of a conserved gene, *prfA*, were amplified successfully. Each member examined from the genus *Thermotoga* was found to have genes known to code for flagellar proteins (see Figure 3.2). Table 4.1 shows the strains that have been examined for both flagellation and motility; those marked with an asterisk (*) are either strains examined in this thesis, or close SSU rRNA relatives.

Table 4.1 Flagellation and motility in the Thermotogales

Organism[1]	Flagella? [2]	Motility? [2]
*Thermotoga maritima**	single, subpolar	+
Thermotoga elfii	peritricious	+
Thermotoga hypogea	lateral	+
Thermotoga neapolitana NS-E*	*none*	-
Thermotoga thermarum LA3*	lateral	+
*Thermotoga subterranea**	ND	ND
Thermotoga petrophila RKU 1T*	multiple; lateral & subpolar	ND
Thermotoga naphthophila RKU10-IT*	multiple; lateral & subpolar	ND
*Petrotoga miotherma**	*none*	-
*Petrotoga mobilis**	ND	+
*Fervidobacterium islandicum**	ND	+
*Fervidobacterium nodosum**	ND	+

[1] Flagellation information for *T. petrophila* and *T. naphthophila* is taken from Takahata *et al.* (Takahata *et al.* 2001); the remainder is from Bergey's Manual (Reysenbach 2001).
[2]ND = not determined

The high level of sequence conservation of the proximal flagellar cluster suggests that all strains examined have had recently functioning flagella. All strains of *Thermotoga neapolitana* examined, for example, contained the full cluster, and it was intact. The type strain, NS-E, has no flagella and is non-motile, but maintains at least the genetic ability to form the proximal structure. Southern blots (data not shown) indicate that these strains also contain at least a partial distal cluster, even though it was not successfully amplified or sequenced.

Some types of strain differences, such as the presence of pathogenicity or ecological islands (Nesbø *et al.* 2002; Welch *et al.* 2002), are easy to see, especially with the availability of completely sequenced genomes. These types of comparisons clearly point out regions of prokaryotic genomes with differing histories, or origins. However, exchange between close relatives is both more likely to happen by virtue of sequence similarity and more likely to be invisible for the same reason. The recombination that has occurred involving *T. maritima* MSB8, *T.* sp. RQ2, *T. petrophila* RKU1 and *T. naphthophila* RKU10 is a visible example of the exchange that can take place between closely related strains, and that may go unnoticed as a direct result of that close relationship. In this case, it is not just gene clusters, or whole genes, but parts of genes that have been involved in recombination. The overall signal of the cluster shows a different history than that of the SSU rRNA of these organisms, while small regions within that cluster show a third relationship (see Figure 3.6).

The maintenance of a single flagellar gene cluster downstream of *prfA*, in both motile and non-motile strains, might simply be a result of in sufficient time for gene shuffling within individual strain genomes. However, the presence of a second flagellar gene cluster immediately downstream of *prfA* in the related strains *Thermotoga thermarum* LA3 and *T. subterranea* SL7 (one of which is known to be flagellated and motile) suggests that the higher order structure of expression of flagellar genes may benefit from the presence of a protein release factor. The inability to amplify the distal flagellar cluster might be attributable to a higher rate of evolution of these protein genes; they have fewer interacting partners within the structure itself, and are further away from the molecular motor and export apparatus. Amino acid changes may be accepted because of the smaller number of protein-protein interactions, making the design of degenerate primers much more difficult.

4.2: Ribosomal Protein Clusters Show High Conservation Amongst Bacteria and Archaea, While Individual Genes are Subject to Recombination

4.2.1: Update of Coenye and Vandamme (2005)

An analysis similar to one completed in 2005 (Coenye and Vandamme 2005) was extended to all available prokaryotic genomes, using an All vs All BLAST utility available on the TIGR CMR website at the time of analysis (http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi). The authors were selective with genomes, choosing what they felt were representative taxa; because we know that even close SSU rRNA relatives can have quite varied genome content and order, no exclusions were made in the present study.

Gene order within operons can be highly unstable (Lathe *et al.* 2000; Tamames 2001), but there is a very high level of conservation of both members and relative genome location for the three ribosomal gene clusters examined. In the majority of cases where a homolog is present in a genome, it retains its status as a cluster member as well as being in the same position relative to the other genes within the cluster. Along with operon instability, lateral gene transfer can also affect gene clusters. Ribosomal proteins are often thought to be resistant to transfer, as members of complex cellular systems; the complexity hypothesis suggests that they would co-evolve with other molecules in the ribosome within a lineage (Jain *et al.* 1999); however, such genes are not completely immune to transfer (Brochier *et al.* 2000). The conservation of both presence and order within these ribosomal clusters over large evolutionary distances suggests there is a force inherent to genome architecture that keeps these features constant.

4.2.2 Conservation of the L21 Ribosomal Protein Gene Cluster in Thermotogales

Sequence analysis of the L21 ribosomal protein cluster reveals an extremely high level of conservation at both the DNA and protein levels. The translated amino acid sequences for L21, L27, and L13 are highly conserved, and because of the small size of the genes, do not lend themselves to tree building. In all cases, the strains examined, along with *T. maritima* MSB8, fall into two distinct groups that can be seen by visual inspection of alignments of the DNA heterogeneities. Some recombination or lateral gene transfer is likely in two cases: 1) when one compares the groupings in Figure 1.1A in *T.* sp. RQ7, a "*neapolitana*" strain by SSU rRNA classification, all five genes from the cluster show near 100% identity to those found in *T. maritima* MSB8 and *T.* sp. RQ2, both considered to me "*maritima* strains"; and 2) *T. thermarum* LA3 and *T. africanus*, which sit outside the *maritima*/*neapolitana* clade, show near 100% identity (*T. thermarum* has one base pair difference over the entire cluster, while *T. africanus* has three). The level of conservation, which far exceeds that of the SSU rRNA sequence of this clade compared with the maritimas (at approximately 96%), implies that the ribosomal genes L21, L27, and L13, as well as the intervening conserved hypothetical protein and true hypothetical protein, are homogenized within this group. Although this could be considered contrary to the complexity hypothesis, such inter-strain and inter-species recombination may not be deemed a true or significant lateral gene transfer, as it would be if found amongst distant SSU rRNA relatives. Even with the recombination, however, the cluster genes, including the CHPs and THPs, are conserved in both presence and order, implying that the location of these ORFs relative to one another is important.

The grouping of the *T.* sp. RQ7 sequences with *T. maritima* MSB8, instead of the other *T. neapolitana* strains examined, is likely the result of a recombination between ancestors of these two groups. Nesbø *et al.* (Nesbø *et al.* 2006) have found one other case of recombination between *T.* sp. RQ7, and *T. maritima* strains, involving homologs of TM0938, a conserved hypothetical protein. The flagellar protein genes amplified from *T.* sp. RQ7, however, agree with the SSU rRNA tree (see Figure 3.4).

4.2.3: Conservation of the L21 Ribosomal Protein Gene Cluster in Other Bacteria

The genomic context of the three ribosomal protein genes from this cluster is similar to that found in both the previous and present studies of the *s10*, *spc* and *alpha* ribosomal protein gene clusters (Coenye and Vandamme 2005) and Results Section 3.2.1. Homologs of L21 and L27 are always found in the same orientation on the DNA circle, and in nearly 70% of cases are adjacent. At most, there are two intervening ORFs, and the ribosomal protein genes themselves never occur further apart. However, L13 is rarely, if ever, found with these two genes, and is likely not historically part of this cluster, or linked to the first two genes.

Sequence analysis of TM1455, a conserved hypothetical protein, and its homologs in Thermotogales suggests that its function is conserved amongst those strains examined, and its position is constant. Homologs of TM1455 are much more rare within the database, having been found in only 7.5% of genomes assessed. No definitive function has been assigned, although TIGR BLAST searches retrieve other conserved hypothetical proteins that have putative membrane domains. Similar to L13, homologs of TM1455 in other prokaryotes are not found associated with other ribosomal proteins, suggesting that these two proteins are not part of this ribosomal protein cluster.

TM1457 has no homologs in the database at present, and the discovery of homologs in other closely related Thermotogales, and subsequent deposition of sequence data, would by definition change the status of this ORF from a true hypothetical protein (singleton ORFan) to a conserved hypothetical protein (orthologous ORFan) (Siew and Fischer 2003a; Siew and Fischer 2003b). However, because of the tendency of other prokaryotic genomes to contain different true hypothetical proteins in the same location (that is, between the genes for L21 and L27), it is more likely that the ORF coded for by TM1457 is misannotated spacer DNA.

4.3: Spatial Autocorrelation of Different Functional Gene Categories in Prokaryotes

4.3.1: Functional Gene Categories Within *Thermotoga maritima* MSB8

Physical distributions of different functional categories within *T. maritima* MSB8 demonstrated that not all categories form clusters within the genome, and those that tend to cluster do so in clusters of different sizes. Cluster size itself does not seem to be correlated with the number of ORFs found in each category, suggesting that genes in any given functional category have their own distinct distribution.

The distribution of potential ORFans within the genome paints an interesting picture, and may help in determining the actual nature of these ORFs. Category 11, which represents the conserved hypothetical proteins (orthologous ORFans), shows a random distribution; that is, the ORFs in this category do not tend to cluster within the genome. If these ORFs code for expressed, functional genes, they are not likely to be part of the small conserved cluster that contains L21 and L27.

The hyperdispersal of Category 13, representing true hypothetical proteins (singleton ORFans), within *T. maritima* MSB8 gives an interesting insight into what function, if any, they may play within the genome. The hyperdispersal of the ORFs in this category, with the strong dissociation at distance classes of 1 and 3, demonstrates that ORFs in this functional category are rarely, if ever, found adjacent to one another. This, combined with the strong association at distance classes 2 and 4 (before signal degradation), indicates that while they are never found adjacent, they are often found separated by a single ORF of a different category. This tendency to be placed between other functional genes leads to the conclusion that, within this genome, ORFs coded as true hypothetical proteins are not in fact proteins, but are misannotated pieces of intervening DNA sequence that happen to have a start and stop codon within close proximity. It is unlikely that all ORFs annotated as true hypothetical proteins are misannotations, however; once a hyperdispersal pattern is uncovered, these ORFs must be more closely examined both within their own genome and, if possible, genomes of closely related strains as they are sequenced (at which point, they would become conserved hypothetical proteins). Assessment of other ORF features, such as codon usage, ORF length, possible conserved domains, and relative position to other established functional clusters and operons will give additional insight into any possible functionality of these ORFans.

4.3.2: Conserved Hypothetical Proteins and True Hypothetical Proteins in Other Prokaryotes

In the majority of bacterial strain groups examined, both Category 11 (conserved hypothetical proteins/orthologous ORFans) and Category 13 (true hypothetical proteins/singleton ORFans) show clustered distributions. In these cases, the ORFs in these categories could represent long stretches of junk DNA, which (like the true hypothetical proteins in *T. maritima* MSB8) happen to have start and stop codons in close proximity. However, in many cases, the clusters are quite large (sometimes well over 50 ORFs), and could represent ecological islands. These could be specific to a group of strains (in the case of the conserved hypothetical proteins) or an island that is specific to one strain (in the case of the true hypothetical proteins).

The few notable exceptions are conserved hypothetical proteins in all strains of *Prochlorococcus marinus*, and true hypothetical proteins in one strain of *Chlamydia pneumoniae* (strain AR39). The *Prochlorococcus* conserved hypothetical proteins all show a random distribution; these genomes tend to be very large (Dufresne *et al.* 2003; Rocap *et al.* 2003), and so could contain a large proportion of misannotated ORFs. However, the randomly distributed CHPs could represent unique, ecologically adapted genes that have become part of existing clusters. Future studies could include a secondary analysis of the location of randomly distributed CHPs, to determine if they tend to interrupt known gene clusters, and thus are contributing to a pathway via a completely novel mechanism.

The true hypothetical proteins in *Chlamydia pneumoniae* AR39 show the same hyperdispersed pattern as those present in *T. maritima* MSB8. In this case, the ORFs in this genome may again be misannotated intervening DNA sequence with start and stop codons, rather than functional genes.

4.4: Maintenance of Functional Gene Clusters and Higher Order Physical Genomic Architecture in Prokaryotes

By examining different types of gene clusters at several levels for recombination and rearrangement, a better picture can be obtained about higher order architecture of prokaryotic genomes. In the Thermotogales, two separate cellular systems – the flagellar apparatus and the ribosome – show recombination amongst closely related strains, while maintaining higher order structure of functional gene clusters. One system is operational, and one informational; however, both are susceptible to recombination, provided it is within certain parameters (i.e. ORFs must remain intact and functional, and genome context must be conserved). In the case of flagellar protein genes, proximity to *prfA* seems to be essential. Ribosomal protein genes L21 and L27 require close proximity, but do not have to be adjacent.

In *T. maritima* MSB8, several functional gene categories show significant clustering within the genome; these categories often comprise genes that are part of larger pathways, such as amino acid biosynthesis, central intermediary metabolism, and transcription. Clustering in *T. maritima* MSB8, which is thought to be a highly mosaic genome (Nelson *et al.* 1999), may also be a result of this feature, as in the case of mobile and extrachromosomal elements.

Clustering of potential ORFans, both conserved hypothetical and true hypothetical proteins, suggests that members of these functional gene categories represent functional, but unknown, gene clusters. Such large stretches of DNA would be susceptible to drift and degeneration without some form of positive selection acting to maintain intact open reading frames. The fact that they are clustered instead of randomly distributed throughout the genome could indicate that they form functional clusters themselves, rather than being part of other, annotated genes.

With complete genome sequences of many strains, higher order functional architecture of prokaryotic genomes is often thought to be non-existent, because of the single or few DNA molecules, the relatively low occurrence of synteny, and the relatively gene-dense nature of the prokaryotic chromosome. By treating each genome like a landscape, this novel application of biogeographical methods can reveal architecture both within closely related strain groups, or between more distantly related organisms. Even after degradation of phylogenetic signal through sequence evolution or gene order shuffling, areas of the genome may be conserved and dedicated to performing certain functions.

4.5: Summary

The complete genome sequence of one or a few members of any given bacterial group can give us some insights into the biology of those organisms, but this information is by no means exhaustive. Until genome sequencing comes down in price, to the point where dozens of strains of any species or group can be sequenced, we must resort to detailed comparative studies at the level of the gene, or smaller segments of the genome. Lateral gene transfer, even if it takes place as often as, or more than, vertical inheritance, cannot completely erase evolutionary relationships. It can, however, introduce a wide variety of functions and capabilities into one or several species that cannot necessarily be predicted from existing genome sequences. Closer examination of two genetic systems within the Thermotogales, the flagellar apparatus and the structural portion of the ribosome, one can see that lateral gene transfer occurs involving existing, complex, and essential systems, but serves to maintain gene clusters at a higher level of architecture. General genome structure can also tell us something about the history of the organisms that are being examined, as well as the likelihood of certain functional groups (of genes) to be transferred or rearranged. When more than one genome sequence is available from a group of closely related organisms, one can better assess the evolution of genome architecture. By applying a biogeographical approach normally reserved for larger, multicellular organisms, the physical distribution of genes, as entities within a genome, can be determined. This demonstrates that, as with examination of smaller gene clusters, there is a higher level of genome architecture that is maintained within closely related strains. By looking at the distribution of both ORFs of unknown function, and established gene clusters and operons, one can evaluate their conservation and importance to the biology of a strain group, as well as assign functions to novel proteins. However, one must still deal with groups of interest in detail, and in depth, in order to continue to gain insight into their complete biologies.

References

Akopyants NS, Fradkov A, Diatchenko L, Hill JE, Siebert PD, Lukyanov SA, Sverdlov ED, Berg DE (1998) PCR-based subtractive hybridization and differences in gene content among strains of *Helicobacter pylori.* Proc Natl Acad Sci U S A 95:13108-13

Amador E, Martin JF, Castro JM (2000) A *Brevibacterium lactofermentum* 16S rRNA gene used as target site for homologous recombination. FEMS Microbiol Lett 185:199-204

Archibald JM, Roger AJ (2002) Gene conversion and the evolution of euryarchaeal chaperonins: a maximum likelihood-based method for detecting conflicting phylogenetic signals. J Mol Evol 55:232-45

Asai T, Condon C, Voulgaris J, Zaporojets D, Shen B, Al-Omar M, Squires C, Squires CL (1999a) Construction and initial characterization of *Escherichia coli* strains with few or no intact chromosomal rRNA operons. J Bacteriol 181:3803-9

Asai T, Zaporojets D, Squires C, Squires CL (1999b) An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc Natl Acad Sci U S A 96:1971-6

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2006) New Screening Software Shows that Most Recent Large 16S rRNA Gene Clone Libraries Contain Chimeras. Appl. Environ. Microbiol. 72:5734-5741

Atomi H, Matsumi R, Imanaka T (2004) Reverse Gyrase Is Not a Prerequisite for Hyperthermophilic Life. J. Bacteriol. 186:4829-4833

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289:905-20

Barloy-Hubler F, Lelaure V, Galibert F (2001) Ribosomal protein gene cluster analysis in eubacterium genomics: homology between *Sinorhizobium meliloti* strain 1021 and Bacillus subtilis. Nucleic Acids Res 29:2747-56

Bingen-Bidois M, Clermont O, Bonacorsi S, Terki M, Brahimi N, Loukil C, Barraud D, Bingen E (2002) Phylogenetic analysis and prevalence of urosepsis strains of *Escherichia coli* bearing pathogenicity island-like domains. Infect Immun 70:3216-26

Blattner FR, Plunkett G, III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The Complete Genome Sequence of *Escherichia coli* K-12. Science 277:1453-1462

Boucher Y, Douady CJ, Sharma AK, Kamekura M, Doolittle WF (2004) Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. J Bacteriol 186:3980-90

Brochier C, Philippe H, Moreira D (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet 16:529-33

Casjens S, Delange M, Ley HL, 3rd, Rosa P, Huang WM (1995) Linear chromosomes of Lyme disease agent spirochetes: genetic diversity and conservation of gene order. J Bacteriol 177:2769-80

Cazalet C, Rusniok C, Bruggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C (2004) Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet 36:1165-1173

Charbonnier F, Forterre P (1994) Comparison of plasmid DNA topology among mesophilic and thermophilic eubacteria and archaebacteria. J Bacteriol 176:1251-9

Chien M, Morozova I, Shi S, Sheng H, Chen J, Gomez SM, Asamani G, Hill K, Nuara J, Feder M, Rineer J, Greenberg JJ, Steshenko V, Park SH, Zhao B, Teplitskaya E, Edwards JR, Pampou S, Georghiou A, Chou I-C, Iannuccilli W, Ulz ME, Kim DH, Geringer-Sameth A, Goldsberry C, Morozov P, Fischer SG, Segal G, Qu X, Rzhetsky A, Zhang P, Cayanis E, De Jong PJ, Ju J, Kalachikov S, Shuman HA, Russo JJ (2004) The Genomic Sequence of the Accidental Pathogen *Legionella pneumophila*. Science 305:1966-1968

Cho JC, Tiedje JM (2000) Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. Appl Environ Microbiol 66:5448-56

Coenye T, Vandamme P (2005) Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. FEMS Microbiol Lett 242:117-26

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23:324-8

Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14:1036-42

Dobrindt U, Blum-Oehler G, Nagy G, Schneider G, Johann A, Gottschalk G, Hacker J (2002) Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536. Infect Immun 70:6365-72

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F, Makarova KS, Ostrowski M, Oztas S, Robert C, Rogozin IB, Scanlan DJ, de Marsac NT, Weissenbach J, Wincker P, Wolf YI, Hess WR (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci USA 100:10020-10025

Epperson BK, Allard RW (1989) Spatial Autocorrelation Analysis of the Distribution of Genotypes Within Populations of Lodgepole Pine. Genetics 121:369-377

Fardeau ML, Ollivier B, Patel BK, Magot M, Thomas P, Rimbault A, Rocchiccioli F, Garcia JL (1997) Thermotoga hypogea sp. nov., a xylanolytic, thermophilic bacterium from an oil-producing well. Int J Syst Bacteriol 47:1013-9

Firneisz G, Zehavi I, Vermes C, Hanyecz A, Frieman JA, Glant TT (2003) Identification and quantification of disease-related gene clusters. Bioinformatics 19:1781-6

Fischer D, Eisenberg D (1999) Finding families for genomic ORFans. Bioinformatics 15:759-62

Forterre P (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet 18:236-7

Fouet A, Smith KL, Keys C, Vaissaire J, Le Doujet C, Levy M, Mock M, Keim P (2002) Diversity among French *Bacillus anthracis* Isolates. J Clin Microbiol 40:4732-4734

Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, Brinkac LM, DeBoy RT, Parker CT, Daugherty SC, Dodson RJ, Durkin AS, Madupu R, Sullivan SA, Shetty JU, Ayodeji MA, Shvartsbeyn A, Schatz MC, Badger JH, Fraser CM, Nelson KE (2005) Major Structural Differences and Novel Potential Virulence Mechanisms from the Genomes of Multiple *Campylobacter* Species. PLoS Biology 3:e15

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. Science 209:457-63

Franklin RB, Blum LK, McComb AC, Mills AL (2002) A geostatistical analysis of small-scale spatial variability in bacterial abundance and community structure in salt marsh creek bank sediments. FEMS Microbiol Ecol 42:71-80

Franklin RB, Mills AL (2003) Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field. FEMS Microbiol Ecol 44:335-46

Gaffney DJ, Keightley PD (2005) The scale of mutational variation in the murid genome. Genome Res 15:1086-94

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44:632-6

Garcia-Vallve S, Simo FX, Montero MA, Arola L, Romeu A (2002) Simultaneous horizontal gene transfer of a gene coding for ribosomal protein l27 and operational genes in *Arthrobacter* sp. J Mol Evol 55:632-7

Hashimoto JG, Stevenson BS, Schmidt TM (2003) Rates and consequences of recombination between rRNA operons. J Bacteriol 185:966-72

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han C-G, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H (2001) Complete Genome Sequence of Enterohemorrhagic *Eschelichia coli* O157:H7 and Genomic Comparison with a Laboratory Strain K-12. DNA Res 8:11-22

Hejnova J, Dobrindt U, Nemcova R, Rusniok C, Bomba A, Frangeul L, Hacker J, Glaser P, Sebo P, Buchrieser C (2005) Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83: K24: H31). Microbiology 151:385-98

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765-74

Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16:332-46

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801-6

Jeanthon C, Reysenbach AL, L'Haridon S, Gambacorta A, Pace NR, Glenat P, Prieur D (1995) *Thermotoga subterranea* sp. nov., a new thermophilic bacterium isolated from a continental oil reservoir. Arch Microbiol 164:91-7

Jones TH, Vaillancourt RE, Potts BM (2007) Detection and visualization of spatial genetic structure in continuous Eucalyptus globulus forest. Molecular Ecology 16:697-707

Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens R (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet 21:385-389

Kao JS, Stucker DM, Warren JW, Mobley HL (1997) Pathogenicity island sequences of pyelonephritogenic *Escherichia coli* CFT073 are associated with virulent uropathogenic strains. Infect Immun 65:2812-20

Katz LA, Curtis EA, Pfunder M, Landweber LF (2000) Characterization of novel sequences from distantly related taxa by walking PCR. Mol Phylogenet Evol 14:318-21

Klein DJ, Moore PB, Steitz TA (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. J Mol Biol 340:141-77

Koski LB, Morton RA, Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18:404-12

Kuehn R, Hindenlang KE, Holzgang O, Senn J, Stoeckle B, Sperisen C (2007) Genetic Effect of Transportation Infrastructure on Roe Deer Populations (Capreolus capreolus). J Hered %R 10.1093/jhered/esl056 98:13-22

Kuroda Y, Kaga A, Tomooka N, Vaughan DA (2006) Population genetic structure of Japanese wild soybean (Glycine soja) based on microsatellite variation. Molecular Ecology 15:959-974

Lan R, Reeves PR (1996) Gene transfer is a major factor in bacterial evolution. Mol Biol Evol 13:47-55

Lan R, Reeves PR (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 8:396-401

Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Marsh TL, Woese CR (1993) The ribosomal database project. Nucleic Acids Res 21:3021-3

Lathe WC, 3rd, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. Trends Biochem Sci 25:474-9

Lu S, Park M, Ro HS, Lee DS, Park W, Jeon CO (2006) Analysis of microbial communities using culture-dependent and culture-independent approaches in an anaerobic/aerobic SBR reactor. J Microbiol 44:155-61

Macnab RM (2003) How bacteria assemble flagella. Annu Rev Microbiol 57:77-100

Makarova KS, Ponomarev VA, Koonin EV (2001) Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome Biol 2:RESEARCH 0033

Matte-Tailliez O, Brochier C, Forterre P, Philippe H (2002) Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol 19:631-9

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399:323-9

Nesbø CL, Dlutek M, Doolittle WF (2006) Recombination in *Thermotoga*: implications for species concepts and biogeography. Genetics 172:759-69

Nesbø CL, L'Haridon S, Stetter KO, Doolittle WF (2001) Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol Biol Evol 18:362-75

Nesbø CL, Nelson KE, Doolittle WF (2002) Suppressive subtractive hybridization detects extensive genomic diversity in *Thermotoga maritima.* J Bacteriol 184:4475-88

Nunan N, Ritz K, Crabb D, Harris K, Wu K, Crawford JW, young IM (2001) Quantification of the in situ distribution of soil bacteria by large-scale imaging of thin sections of undisturbed soil. FEMS Microbiol Ecol 36:67-77

Nunan N, Wu K, Young IM, Crawford JW, Ritz K (2002) In situ spatial patterns of soil bacterial populations, mapped at multiple scales, in an arable soil. Microb Ecol 44:296-305

Ojaimi C, Davidson BE, Saint Girons I, Old IG (1994) Conservation of gene arrangement and an unusual organization of rRNA genes in the linear chromosomes of the Lyme disease spirochaetes *Borrelia burgdorferi, B. garinii* and *B. afzelii.* Microbiology 140 (Pt 11):2931-40

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40:337-65

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

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734-40

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV, Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream M-A, Rutherford KM, van Vliet AHM, Whitehead S, Barrell BG (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665-668

Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529-533

Pope LC, Domingo-Roura X, Erven K, Burke T (2006) Isolation by distance and gene flow in the Eurasian badger (*Meles meles*) at both a local and broad scale. Molecular Ecology 15:371-386

Primmer CR, Veselov AJ, Zubchenko A, Poututkin A, Bakhmet I, Koskinen MT (2006) Isolation by distance within a river system: genetic population structuring of Atlantic salmon, *Salmo salar*, in tributaries of the Varzuga River in northwest Russia. Molecular Ecology 15:653-666

Ravot G, Magot M, Fardeau ML, Patel BK, Prensier G, Egan A, Garcia JL, Ollivier B (1995) *Thermotoga elfii* sp. nov., a novel thermophilic bacterium from an African oil-producing well. Int J Syst Bacteriol 45:308-14

Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, Bass S, Linher K, Weidman J, Khouri H, Craven B, Bowman C, Dodson R, Gwinn M, Nelson W, DeBoy R, Kolonay J, McClarty G, Salzberg SL, Eisen J, Fraser CM (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucl Acids Res 28:1397-1406

Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, Tettelin H, Fouts DE, Eisen JA, Gill SR, Holtzapple EK, Okstad OA, Helgason E, Rilstone J, Wu M, Kolonay JF, Beanan MJ, Dodson RJ, Brinkac LM, Gwinn M, DeBoy RT, Madpu R, Daugherty SC, Durkin AS, Haft DH, Nelson WC, Peterson JD, Pop M, Khouri HM, Radune D, Benton JL, Mahamoud Y, Jiang L, Hance IR, Weidman JF, Berry KJ, Plaut RD, Wolf AM, Watkins KL, Nierman WC, Hazen A, Cline R, Redmond C, Thwaite JE, White O, Salzberg SL, Thomason B, Friedlander AM, Koehler TM, Hanna PC, Kolsto A-B, Fraser CM (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related Bacteria. Nature 423:81-86

Reysenbach A-L (2001) Phylum BIII. Thermotogae *phy. nov.* In: Boone D, Castenholz R (eds) The Archaea and the Deeply Brancing and Phototrophic Bacteria. Springer, New York, p 369-387

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047

Schmuki C, Vorburger C, Runciman D, Maceachern S, Sunnucks P (2006) When log-dwellers meet loggers: impacts of forest fragmentation on two endemic log-dwelling beetles in southeastern Australia. Molecular Ecology 15:1481-1492

Shirai M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shiba T, Ishii K, Hattori M, Kuhara S, Nakazawa T (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucl Acids Res 28:2311-2314

Siew N, Fischer D (2003a) Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins 53:241-51

Siew N, Fischer D (2003b) Twenty thousand ORFan microbial protein families for the biologist? Structure 11:7-9

Straus D, Ausubel FM (1990) Genomic subtraction for cloning DNA corresponding to deletion mutations. Proc Natl Acad Sci U S A 87:1889-93

Suzuki MT, Beja O, Taylor LT, Delong EF (2001) Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. Environ Microbiol 3:323-31

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts

Takahata Y, Nishijima M, Hoaki T, Maruyama T (2001) *Thermotoga petrophila* sp. nov. and *Thermotoga naphthophila* sp. nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata, Japan. Int J Syst Evol Microbiol 51:1901-9

Tamames J (2001) Evolution of gene order conservation in prokaryotes. Genome Biol 2:RESEARCH0020

Tamames J, Casari G, Ouzounis C, Valencia A (1997) Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol 44:66-73

Thoma R, Schwander M, Liebl W, Kirschner K, Sterner R (1998) A histidine gene cluster of the hyperthermophile *Thermotoga maritima*: sequence analysis and evolutionary significance. Extremophiles 2:379-89

Van de Peer Y, Neefs J-M, DeRijk P, DeVos P, DeWachter R (1994) About the Order of Divergence of the Major Bacterial Taxa During Evolution. Systematic and Applied MIcrobiology 17:32-38

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74

von Grunberg HH, Peifer M, Timmer J, Kollmann M (2004) Variations in substitution rate in human and mouse genomes. Phys Rev Lett 93:208102

Watanabe H, Mori H, Itoh T, Gojobori T (1997) Genome plasticity as a paradigm of eubacteria evolution. J Mol Evol 44 Suppl 1:S57-64

Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli.* Proc Natl Acad Sci U S A 99:17020-4

Willumeit R, Diedrich G, Forthmann S, Beckmann J, May RP, Stuhrmann HB, Nierhaus KH (2001) Mapping proteins of the 50S subunit from *Escherichia coli* ribosomes. Biochim Biophys Acta 1520:7-20

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221-71

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088-90

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87:4576-9

Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res 11:356-72

Yu W, Rusterholtz KJ, Krummel AT, Lehman N (2006) Detection of high levels of recombination generated during PCR amplification of RNA templates. Biotechniques 40:499-507

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF (2001) Crystal structure of the ribosome at 5.5 A resolution. Science 292:883-96

APPENDIX 1: JOINT COUNTS FROM *T. MARITIMA* MSB8

Table A1.1 Raw data* from Genespat v.4, from *T. maritima* MSB8.

Category	1	2	3	4	5	6
# of ORFs	73	32	73	49	45	3
Distance	34.0628	10.0249	21.0747	4.3898	3.7457	-0.1013
2	25.945	6.3231	16.204	0.763	3.7456	19.6415
3	21.0901	2.624	9.7187	0.7654	-0.2017	-0.1012
4	15.4103	0.7729	6.4731	0.7665	-1.5184	-0.1012
5	11.3527	-1.0794	4.8509	0.7677	1.1177	-0.1012
6	6.4796	-1.0791	2.4149	-1.6525	-0.1993	-0.1012
7	3.2302	0.7754	1.6038	-0.4411	-1.5172	-0.1011
8	1.6056	-1.0785	-0.0212	1.9825	-1.5167	-0.1011
9	0.7938	-1.0782	-0.0199	1.984	-0.1969	-0.1011
10	-0.0185	-1.0779	-0.8324	-0.4386	1.1236	-0.101
11	-0.0171	0.7787	1.6112	-1.6503	-0.1953	-0.101
12	-0.8312	0.7787	-0.0171	-1.6503	3.7651	-0.101
13	-0.0171	0.7787	1.6112	-0.4378	2.445	-0.101
14	-0.0171	2.635	0.7971	-0.4378	2.445	-0.101
15	1.6112	-1.0775	3.2395	1.9871	1.1248	-0.101
16	-0.0171	0.7787	3.2395	1.9871	-0.1953	-0.101
17	-0.0171	-1.0776	2.4254	-0.4378	-0.1953	-0.101
18	-0.8312	-1.0776	4.0537	0.7746	-1.5155	-0.101
19	-0.8312	-1.0776	3.2395	-0.4378	-1.5155	-0.101
20	-1.6454	-1.0776	3.2395	3.1995	-1.5155	-0.101
21	-1.6454	-1.0776	2.4254	0.7746	-1.5155	-0.101
22	-0.8312	-1.0776	1.6112	-0.4378	-1.5155	-0.101
23	-0.0171	-1.0776	2.4254	-1.6503	2.445	-0.101
24	-0.0185	-1.0778	2.4232	-1.6507	-0.1961	-0.101
25	-0.8312	-1.0776	-1.6454	-1.6503	2.445	-0.101
26	-0.0171	0.7787	-0.8312	-0.4378	-0.1953	-0.101
27	-1.6454	-1.0776	-1.6454	-1.6503	1.1248	-0.101
28	-1.6454	-1.0776	-0.8312	1.9871	-0.1953	-0.101
29	-0.8312	-1.0776	-1.6454	0.7746	2.445	-0.101
30	-0.8324	0.7779	-0.8324	0.7734	-0.1961	-0.101
31	-1.6454	0.7787	-2.4595	0.7746	1.1248	-0.101
32	-0.8312	0.7787	-1.6454	-0.4378	-0.1953	-0.101
33	-0.8312	-1.0776	-0.8312	-1.6503	-1.5155	-0.101
34	-1.6454	-1.0776	0.7971	1.9871	1.1248	-0.101
35	-1.6454	-1.0776	0.7971	0.7746	1.1248	-0.101
36	-1.6454	-1.0776	0.7971	-0.4378	-1.5155	-0.101
37	-2.4595	-1.0776	3.2395	-1.6503	-1.5155	-0.101
38	-2.4596	-1.0776	-0.0171	-1.6503	-1.5155	-0.101
39	-1.6454	0.7787	3.2396	0.7746	-1.5155	-0.101
40	-2.4596	0.7787	1.6112	0.7746	2.445	-0.101
41	-2.4596	0.7787	1.6112	-1.6503	-1.5155	-0.101
42	-2.4596	0.7787	0.7971	-1.6503	-1.5155	-0.101
43	-2.4596	0.7787	1.6112	-1.6503	-0.1953	-0.101
44	-2.4595	-1.0776	3.2395	0.7746	-0.1953	-0.101
45	-1.6454	-1.0776	2.4254	-1.6503	-0.1953	-0.101
46	-0.0171	-1.0776	3.2395	-1.6503	-0.1953	-0.101
47	0.7971	0.7787	1.6112	-1.6503	1.1248	-0.101
48	1.6112	-1.0776	1.6112	-0.4378	-1.5155	-0.101
49	2.4254	-1.0776	-0.0171	1.9871	-0.1953	-0.101
50	2.4254	-1.0776	-0.0171	-1.6503	-1.5155	-0.101

Category	7	8	9	10	11	12
# of ORFs	55	197	15	0	381	29
Distance	3.528	17.1913	3.442	N/A	3.5355	25.5714
2	0.2967	11.7502	-0.5065		1.628	21.4866
3	1.3767	6.6238	3.4445		1.0078	21.5002
4	2.4562	4.5118	3.4458		3.2433	19.4628
5	1.3796	4.5169	-0.5059		2.4559	19.469
6	2.4597	3.9166	-0.5058		0.7127	13.3392
7	0.3035	5.133	-0.5056		2.791	15.3895
8	3.5426	4.8354	-0.5055		1.5247	13.348
9	-0.7739	1.5074	-0.5054		-1.1769	11.3053
10	-0.7731	2.7241	-0.5052		1.2217	13.3568
11	0.308	1.8191	-0.5051		3.1434	11.3129
12	-1.8525	0.6062	-0.5051		1.2295	7.2164
13	1.3883	-0.6066	-0.5051		0.1131	9.2647
14	-0.7722	0.303	-0.5051		0.4321	7.2165
15	1.3883	-1.213	-0.5051		2.0269	7.2164
16	0.308	0.303	-0.5051		-1.3223	5.1682
17	1.3883	1.2127	-0.5051		2.5055	5.1682
18	0.308	-0.0002	-0.5051		1.8675	11.3129
19	0.308	-0.0002	-0.5051		1.8675	3.12
20	0.308	-0.3034	-0.5051		-1.6413	5.1682
21	-0.7722	-1.5162	-0.5051		0.7511	3.12
22	-0.7722	-0.6066	-0.5051		0.4321	1.0717
23	-1.8525	-0.3034	3.4547		1.07	1.0717
24	-0.773	0.6023	-0.5052		1.0621	1.0709
25	1.3883	1.8191	-0.5051		1.708	1.0717
26	-0.7722	1.5159	-0.5051		-0.6843	-0.9765
27	-1.8525	2.4256	-0.5051		-1.1628	1.0717
28	-0.7722	0.6063	3.4547		-0.0464	3.12
29	-0.7722	0.303	-0.5051		0.9106	3.12
30	1.3868	-0.9133	-0.5052		3.4535	3.1185
31	1.3883	-0.3034	3.4547		-0.8438	3.12
32	1.3883	0.303	-0.5051		1.07	3.12
33	-0.7722	-0.6066	-0.5051		0.1131	3.12
34	0.308	1.2127	-0.5051		-1.0033	3.12
35	-0.7722	1.5159	-0.5051		-1.4818	3.12
36	-0.7722	1.2127	-0.5051		1.8675	3.12
37	0.308	1.2127	-0.5051		-0.5249	1.0717
38	0.308	-0.0002	-0.5051		-1.1629	3.12
39	0.308	-0.3034	-0.5051		-0.3654	1.0717
40	-1.8525	0.6063	-0.5051		3.3031	-0.9765
41	-1.8525	2.1224	-0.5051		1.2296	-0.9765
42	-0.7722	-0.0002	-0.5051		-0.3654	-0.9765
43	-1.8525	-0.9098	-0.5051		-0.5249	-0.9765
44	-1.8525	-1.213	-0.5051		0.7511	-0.9765
45	-0.7722	-0.9098	3.4547		0.5916	-0.9765
46	-1.8525	-0.3034	-0.5051		-0.5249	-0.9765
47	-0.7722	1.8191	-0.5051		2.346	-0.9765
48	-1.8525	3.032	-0.5051		0.1131	-0.9765
49	-0.7722	0.9095	-0.5051		-1.1628	-0.9765
50	1.3883	-0.3034	-0.5051		-1.3224	-0.9765

Category	13	14	15	16	17	18
# of ORFs	251	49	108	45	71	0
Distance	-8.3146	8.0166	18.8637	16.909	5.1128	N/A
2	5.2582	0.763	16.6668	11.6436	-0.7294	
3	-2.1152	-0.4443	15.0329	11.6522	-1.5622	
4	2.6561	-0.4435	11.1918	6.3865	0.9447	
5	-0.9147	-0.4427	15.5958	3.7535	0.1108	
6	0.5209	-0.4419	11.2026	1.1189	0.1122	
7	-0.1899	-0.4411	10.6579	1.1201	0.1135	
8	1.9628	-0.4403	13.965	-0.1977	0.9513	
9	-0.4191	1.984	9.5674	-1.5163	-0.7202	
10	1.9738	5.6218	11.7749	-1.5159	0.1177	
11	-0.4097	4.412	10.1281	-0.1953	1.7932	
12	1.0236	3.1995	10.1279	-0.1953	0.1191	
13	0.0681	1.9871	12.3313	-0.1953	-1.555	
14	1.7403	-0.4378	9.0265	-0.1953	-2.3921	
15	0.0681	1.9871	8.4756	-1.5155	-0.7179	
16	0.307	0.7746	6.8233	-1.5155	-0.7179	
17	1.5014	-0.4378	8.4757	-1.5155	-0.7179	
18	-0.8874	-1.6503	8.4757	-1.5155	0.1191	
19	0.5459	-0.4378	6.2725	-1.5155	-1.555	
20	0.307	-1.6503	7.3741	-1.5155	0.1191	
21	1.5014	-1.6503	5.7217	-1.5155	-0.7179	
22	-0.4097	0.7746	5.1708	-1.5155	0.1191	
23	2.2181	1.9871	4.62	-0.1953	-0.7179	
24	-0.8919	-1.6507	4.6164	-1.5159	1.7913	
25	0.7848	-0.4378	2.4168	-0.1953	-1.555	
26	0.307	-1.6503	2.9676	-1.5155	-0.7179	
27	-0.8874	-1.6503	2.9676	-1.5155	-1.555	
28	3.6514	-0.4378	4.62	-1.5155	3.4673	
29	-0.4097	-1.6503	4.62	-1.5155	1.7932	
30	-0.1755	-1.6507	4.0658	-0.1961	0.1177	
31	1.5014	-1.6503	2.9676	-1.5155	-1.555	
32	0.7848	-0.4378	2.9676	-1.5155	-1.555	
33	1.0237	-1.6503	4.0692	-0.1953	0.9562	
34	-0.4097	-0.4378	4.0692	-1.5155	-0.7179	
35	1.7403	0.7746	3.5184	1.1248	-1.555	
36	1.5014	-0.4378	2.9676	1.1248	-1.555	
37	0.0681	0.7746	2.9676	2.445	0.1191	
38	-0.4097	-1.6503	2.4168	1.1248	-0.718	
39	1.2626	-0.4378	0.7644	-1.5155	3.4674	
40	0.0681	-1.6503	1.3152	-1.5155	1.7932	
41	1.0237	-1.6503	3.5185	-1.5155	0.1191	
42	1.0237	1.9871	2.4168	-0.1953	-1.555	
43	0.0681	1.9871	1.866	3.7652	0.1191	
44	-0.4097	-0.4378	2.4168	1.1248	-2.3921	
45	-0.4097	-1.6503	3.5184	2.445	-1.555	
46	0.0681	-0.4378	4.0692	1.1248	0.1191	
47	-0.1708	1.9871	2.4168	-1.5155	0.1191	
48	-1.1263	3.1995	0.7644	2.445	-1.555	
49	0.7848	1.9871	2.4168	-0.1953	-2.3921	
50	-1.3653	-1.6503	1.3152	-0.1953	0.1191	

Category	19	20	21	22	23
Frequency	0.00907	0.10771	0	0.04649	0
# of ORFs	16	190	0	82	0
Distance	3.1614	23.0003	N/A	1.566	N/A
2	3.1614	13.2885		-1.3252	
3	-0.5399	7.3486		-1.3228	
4	-0.5398	1.7102		-0.5982	
5	3.1663	0.146		-0.5968	
6	3.1675	0.1497		0.8523	
7	-0.5393	0.1534		1.5782	
8	-0.5392	0.785		-1.317	
9	-0.539	2.3592		-1.3158	
10	-0.5389	2.6777		1.5843	
11	3.1735	1.1109		-1.3134	
12	-0.5387	3.6249		0.8614	
13	-0.5387	4.2535		2.3112	
14	-0.5387	2.6822		-1.3134	
15	-0.5387	1.4252		-1.3134	
16	-0.5387	0.1682		-1.3134	
17	-0.5387	2.0537		-0.5885	
18	-0.5387	1.4252		0.1364	
19	3.1735	1.7394		0.1364	
20	-0.5387	1.4252		1.5863	
21	3.1735	2.0537		1.5863	
22	-0.5387	1.4252		0.1364	
23	-0.5387	1.1109		-1.3134	
24	-0.5389	0.1645		1.5842	
25	-0.5387	4.882		1.5863	
26	-0.5387	3.3107		-1.3134	
27	-0.5387	4.2535		-2.0384	
28	-0.5387	4.2535		-1.3134	
29	3.1735	3.9392		0.1364	
30	-0.5389	0.4786		0.1348	
31	-0.5387	1.1109		-1.3134	
32	-0.5387	0.4824		2.3112	
33	-0.5387	-1.4031		-0.5885	
34	-0.5387	0.7967		-1.3134	
35	-0.5387	0.7967		-2.7633	
36	-0.5387	-0.1461		-0.5885	
37	-0.5387	0.4824		-1.3134	
38	-0.5387	2.0538		-0.5885	
39	-0.5387	2.368		0.8614	
40	-0.5387	2.0538		-1.3135	
41	-0.5387	1.4252		0.1364	
42	-0.5387	1.7395		0.8614	
43	-0.5387	0.4824		0.1364	
44	-0.5387	0.4824		-0.5885	
45	-0.5387	1.1109		0.8614	
46	-0.5387	2.0537		0.1364	
47	-0.5387	2.0537		-0.5885	
48	-0.5387	3.3107		0.1364	
49	-0.5387	0.7967		0.8614	
50	-0.5387	1.111		3.7612	

*Categories with no members have # of ORFS (0) and Distance (N/A; not applicable).

APPENDIX 2. FUNCTIONAL GENE CATEGORY FREQUENCY OUTPUT

Table A2.1 Functional gene category frequency output for *Bacillus anthracis* strains.

	Bacillus anthracis strain				
Functional Category	*A0039*	*Ames*	*Ames Ancestor*	*Sterne*	*str. France*
Amino acid biosynthesis	0.023	0.016	0.017	0.021	0.023
Biosynthesis of cofactors, prosthetic groups, and carriers	0.029	0.023	0.023	0.028	0.029
Cell envelope	0.073	0.068	0.071	0.055	0.074
Cellular processes	0.062	0.070	0.067	0.072	0.062
Central intermediary metabolism	0.062	0.009	0.010	0.017	0.061
Disrupted reading frame	0.000	0.006	0.006	0.000	0.000
DNA metabolism	0.036	0.020	0.019	0.027	0.037
Energy metabolism	0.068	0.052	0.052	0.073	0.069
Fatty acid and phospholipid metabolism	0.013	0.013	0.013	0.015	0.013
Glimmer rejects	0.000	0.000	0.000	0.000	0.000
Conserved hypothetical proteins	0.212	0.205	0.205	0.166	0.212
Mobile and extrachromosomal element functions	0.008	0.013	0.013	0.018	0.008
True hypothetical proteins	0.018	0.147	0.148	0.000	0.017
Protein fate	0.040	0.024	0.024	0.033	0.040
Protein synthesis	0.028	0.024	0.024	0.031	0.027
Purines, pyrimidines, nucleosides, and nucleotides	0.013	0.012	0.012	0.015	0.013
Regulatory functions	0.068	0.060	0.060	0.048	0.069
Signal transduction	0.000	0.020	0.020	0.007	0.000
Transcription	0.013	0.011	0.010	0.011	0.013
Transport and binding proteins	0.093	0.093	0.092	0.093	0.092
Unclassified	0.108	0.000	0.000	0.170	0.108
Unknown function	0.032	0.113	0.114	0.098	0.032
Viral functions	0.000	0.000	0.000	0.000	0.000

	Bacillus anthracis strain		
Functional Category	*str. Kruger B*	*Vollum*	*Western North America USA6153*
Amino acid biosynthesis	0.023	0.022	0.022
Biosynthesis of cofactors, prosthetic groups, and carriers	0.029	0.029	0.029
Cell envelope	0.072	0.073	0.072
Cellular processes	0.063	0.062	0.062
Central intermediary metabolism	0.062	0.061	0.061
Disrupted reading frame	0.000	0.000	0.000
DNA metabolism	0.037	0.036	0.036
Energy metabolism	0.069	0.069	0.069
Fatty acid and phospholipid metabolism	0.014	0.013	0.013
Glimmer rejects	0.000	0.000	0.000
Conserved hypothetical proteins	0.210	0.213	0.213
Mobile and extrachromosomal element functions	0.008	0.008	0.008
True hypothetical proteins	0.018	0.017	0.018
Protein fate	0.040	0.040	0.040
Protein synthesis	0.027	0.027	0.027
Purines, pyrimidines, nucleosides, and nucleotides	0.013	0.013	0.013
Regulatory functions	0.069	0.069	0.069
Signal transduction	0.000	0.000	0.001
Transcription	0.013	0.013	0.013
Transport and binding proteins	0.092	0.092	0.093
Unclassified	0.108	0.108	0.108
Unknown function	0.032	0.033	0.033
Viral functions	0.000	0.000	0.000

Figure A2.1 Gene frequency bar graph of functional categories within *Bacillus anthracis* strains. Functional Category key is presented in Figure 3.12.

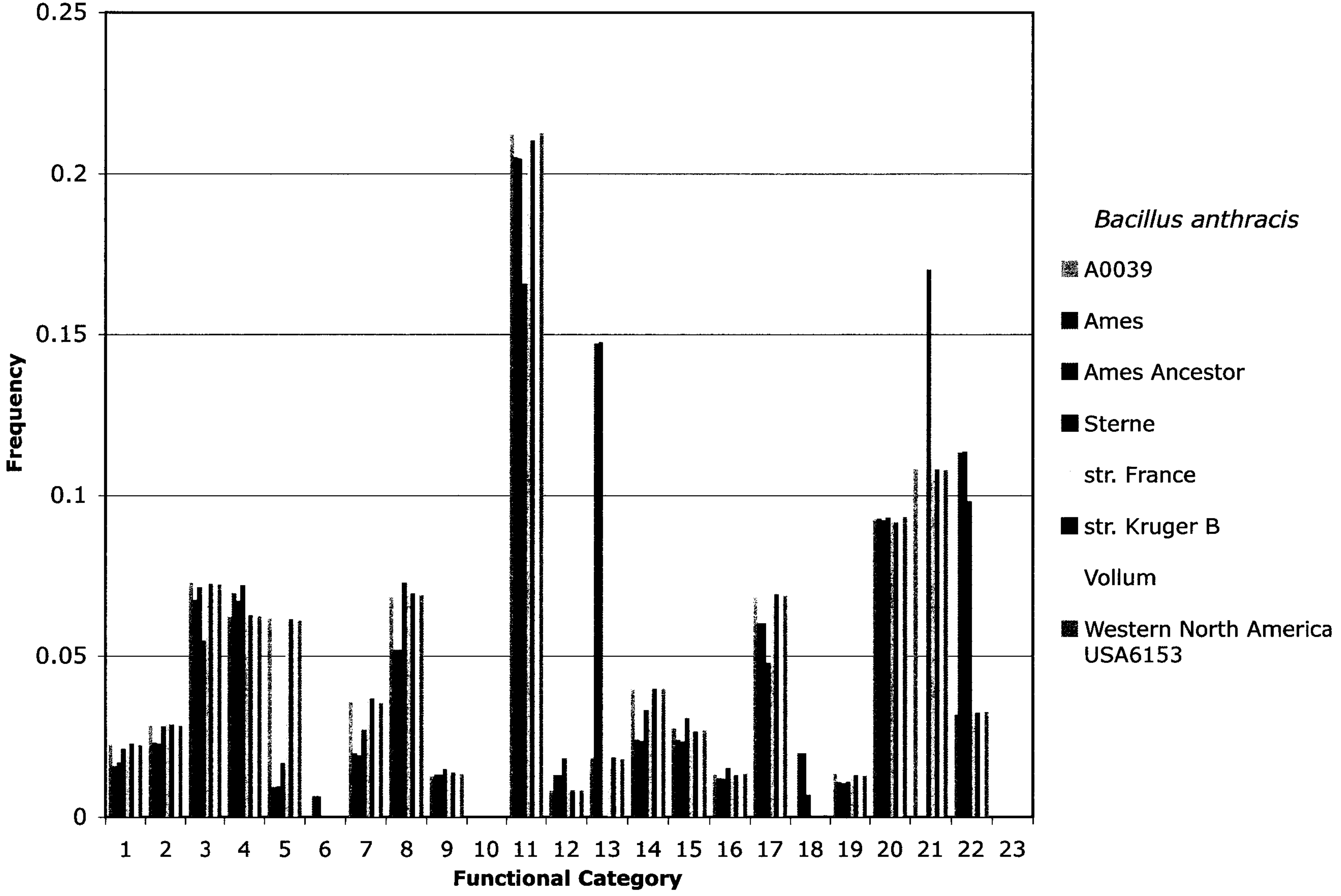

0.25
0.2
0.15
0.1
0.05
0
Frequency
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Functional Category
Bacillus anthracis
▨ A0039
■ Ames
■ Ames Ancestor
■ Sterne
str. France
■ str. Kruger B
Vollum
▨ Western North America USA6153

Table A2.2 Functional gene category frequency output for *Campylobacter jejuni* strains.

Functional Category	*Campylobacter jejuni strain*	
	RM1221	*NCTC 11168*
Amino acid biosynthesis	0.037	0.038
Biosynthesis of cofactors, prosthetic groups, and carriers	0.038	0.040
Cell envelope	0.095	0.207
Cellular processes	0.077	0.068
Central intermediary metabolism	0.010	0.015
Disrupted reading frame	0.010	0.000
DNA metabolism	0.034	0.036
Energy metabolism	0.057	0.090
Fatty acid and phospholipid metabolism	0.000	0.000
Glimmer rejects	0.014	0.017
Conserved hypothetical proteins	0.134	0.143
Mobile and extrachromosomal element functions	0.018	0.001
True hypothetical proteins	0.145	0.002
Protein fate	0.042	0.037
Protein synthesis	0.060	0.083
Purines, pyrimidines, nucleosides, and nucleotides	0.021	0.024
Regulatory functions	0.019	0.044
Signal transduction	0.008	0.016
Transcription	0.013	0.081
Transport and binding proteins	0.085	0.024
Unclassified	0.000	0.000
Unknown function	0.083	0.033
Viral functions	0.000	0.000

Figure A2.2 Gene frequency bar graph of functional categories within *Campylobacter jejuni* strains. Functional Category key is presented in Figure 3.12.

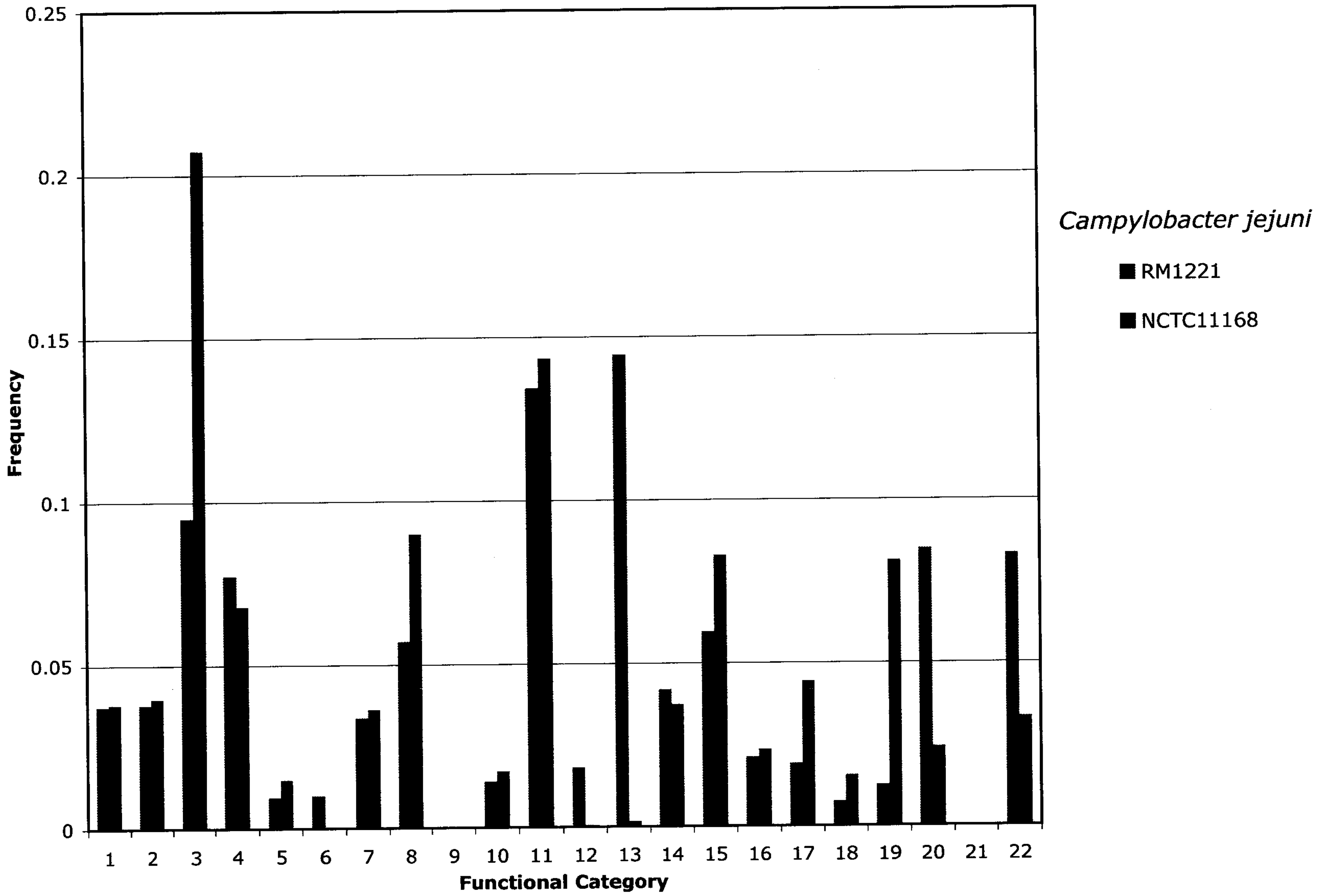

0.25
0.2
0.15
0.1
0.05
0
Frequency
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Functional Category
Campylobacter jejuni
■ RM1221
■ NCTC11168

Table A2.3 Functional gene category frequency output for *Chlamydia pneumophila* strains.

	Chlaymydia pneumophila strain			
Functional Category	*TW-183*	*AR39*	*J138*	*CWL029*
Amino acid biosynthesis	0.015	0.013	0.017	0.013
Biosynthesis of cofactors, prosthetic groups, and carriers	0.033	0.037	0.036	0.028
Cell envelope	0.043	0.061	0.058	0.032
Cellular processes	0.014	0.035	0.033	0.020
Central intermediary metabolism	0.021	0.013	0.023	0.005
Disrupted reading frame	0.000	0.003	0.000	0.000
DNA metabolism	0.047	0.051	0.055	0.041
Energy metabolism	0.115	0.062	0.069	0.057
Fatty acid and phospholipid metabolism	0.018	0.024	0.024	0.017
Glimmer rejects	0.000	0.000	0.000	0.000
Conserved hypothetical proteins	0.240	0.279	0.229	0.127
Mobile and extrachromosomal element functions	0.002	0.000	0.003	0.000
True hypothetical proteins	0.002	0.119	0.000	0.362
Protein fate	0.038	0.062	0.063	0.033
Protein synthesis	0.131	0.100	0.095	0.081
Purines, pyrimidines, nucleosides, and nucleotides	0.017	0.016	0.016	0.016
Regulatory functions	0.070	0.016	0.024	0.014
Signal transduction	0.000	0.000	0.000	0.000
Transcription	0.018	0.023	0.021	0.020
Transport and binding proteins	0.040	0.062	0.058	0.038
Unclassified	0.090	0.000	0.126	0.083
Unknown function	0.046	0.028	0.050	0.011
Viral functions	0.000	0.000	0.000	0.001

Figure A2.3 Gene frequency bar graph of functional categories within *Chlamydia pneumoniae* strains. Functional Category key is presented in Figure 3.12.

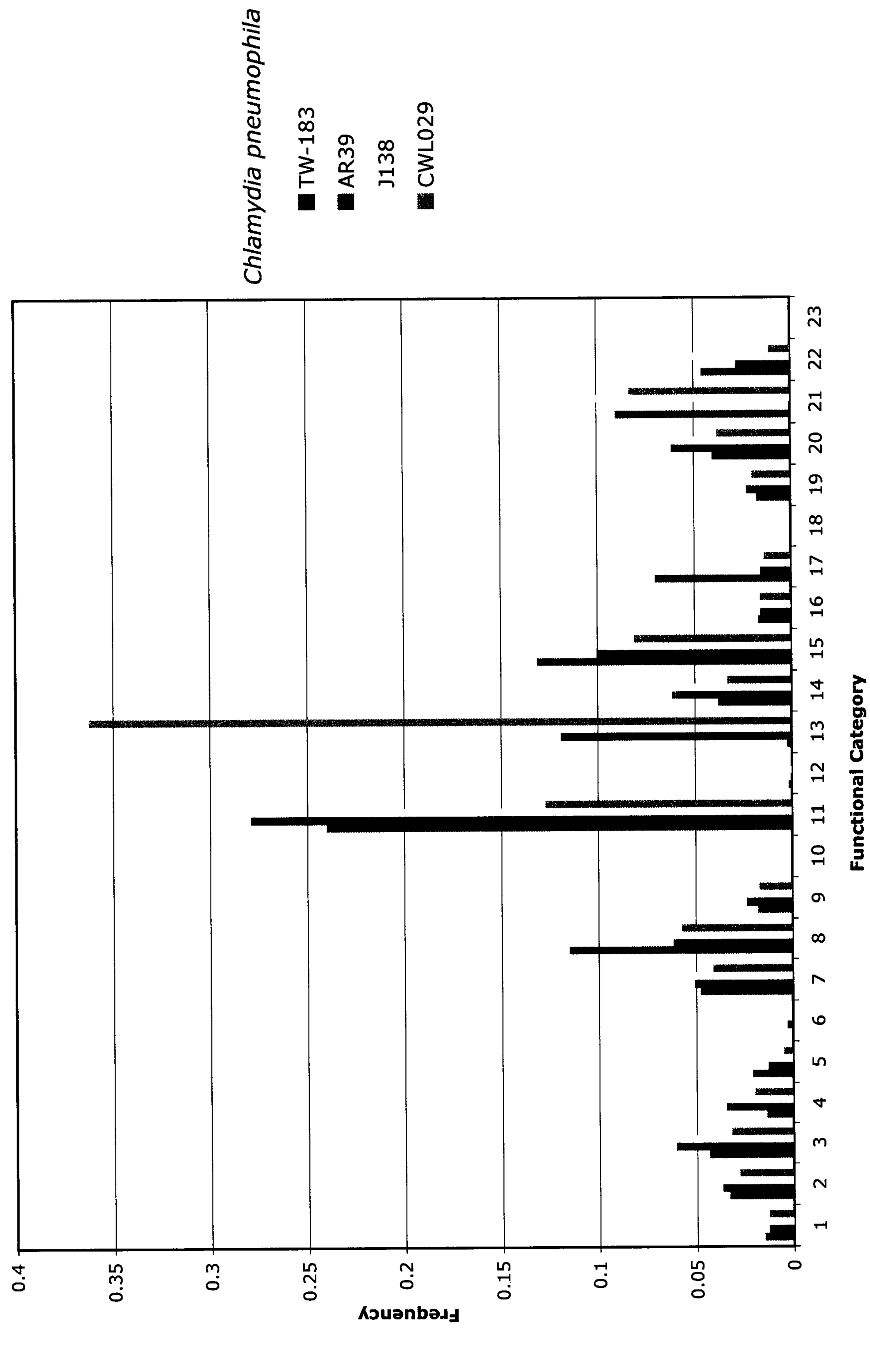

Chlamydia pneumophila
TW-183
AR39
J138
CWL029
Frequency
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Functional Category
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Table A2.4 Functional gene category frequency output for *Escherichia coli* strains.

	Escherichia coli strain			
Functional Category	*CFT073*	*K12-MG1655*	*O157:H7 EDL933*	*O157:H7 VT2-Sakai*
Amino acid biosynthesis	0.019	0.026	0.018	0.020
Biosynthesis of cofactors, prosthetic groups, and carriers	0.028	0.023	0.023	0.025
Cell envelope	0.068	0.040	0.056	0.062
Cellular processes	0.046	0.043	0.040	0.051
Central intermediary metabolism	0.028	0.017	0.030	0.029
Disrupted reading frame	0.026	0.024	0.027	0.034
DNA metabolism	0.000	0.000	0.000	0.000
Energy metabolism	0.112	0.085	0.132	0.100
Fatty acid and phospholipid metabolism	0.014	0.015	0.013	0.014
Glimmer rejects	0.000	0.000	0.000	0.000
Conserved hypothetical proteins	0.196	0.219	0.161	0.198
Mobile and extrachromosomal element functions	0.044	0.011	0.077	0.025
True hypothetical proteins	0.002	0.132	0.001	0.048
Protein fate	0.031	0.027	0.031	0.031
Protein synthesis	0.049	0.028	0.070	0.032
Purines, pyrimidines, nucleosides, and nucleotides	0.014	0.018	0.013	0.016
Regulatory functions	0.071	0.040	0.084	0.062
Signal transduction	0.004	0.000	0.003	0.000
Transcription	0.009	0.009	0.009	0.010
Transport and binding proteins	0.095	0.073	0.085	0.090
Unclassified	0.051	0.154	0.043	0.068
Unknown function	0.091	0.009	0.085	0.036
Viral functions	0.000	0.008	0.000	0.050

Figure A2.4 Gene frequency bar graph of functional categories within *Escherichia coli* strains. Functional Category key is presented in Figure 3.12.

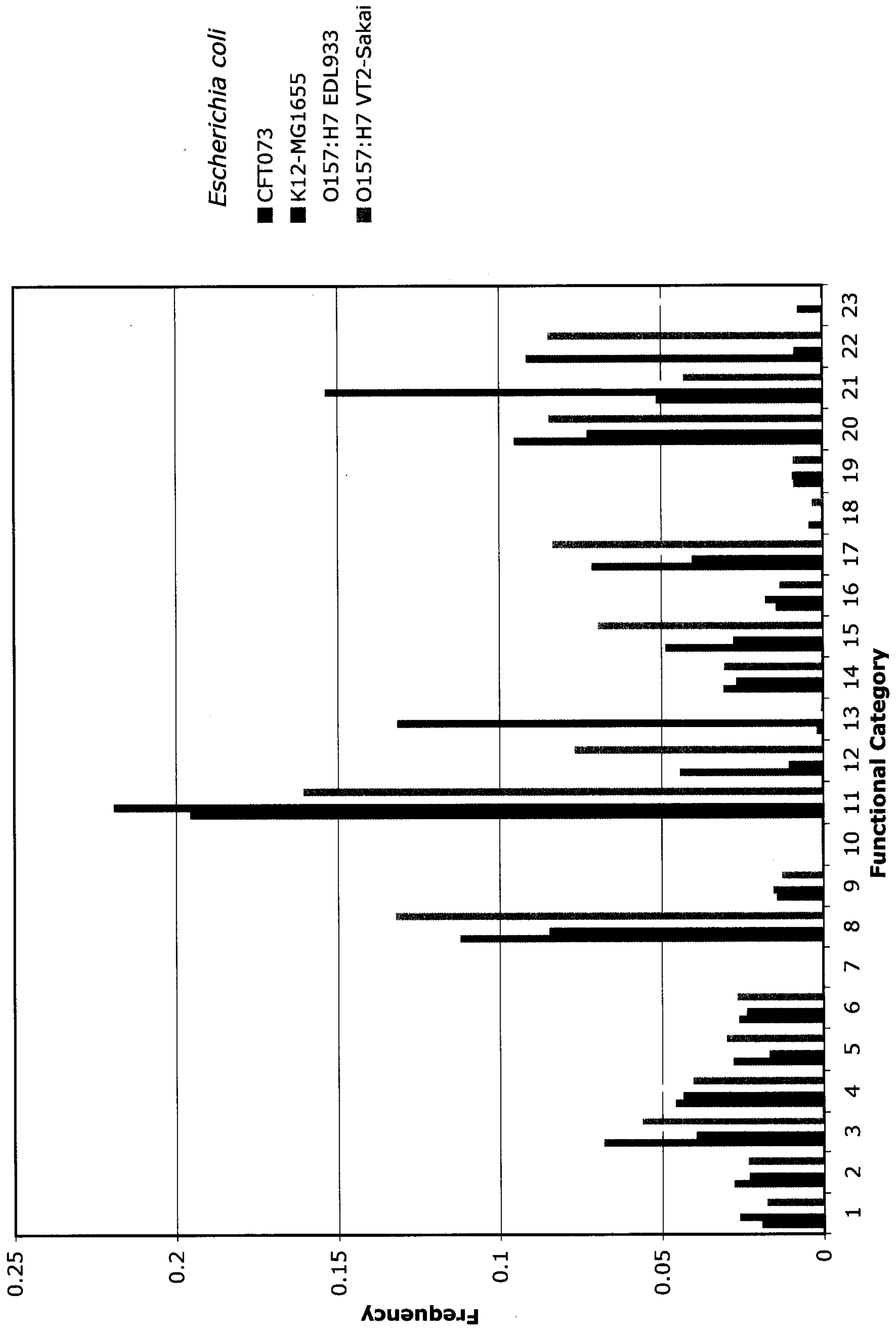

Escherichia coli
■ CFT073
■ K12-MG1655
O157:H7 EDL933
▨ O157:H7 VT2-Sakai
Frequency
0.25
0.2
0.15
0.1
0.05
0
Functional Category
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Table A2.5 Functional gene category frequency output for *Legionella pneumophila* strains.

	Legionella pneumophila strain		
Functional Category	*Lens*	*Paris*	*Philadelphia 1*
Amino acid biosynthesis	0.028	0.026	0.027
Biosynthesis of cofactors, prosthetic groups, and carriers	0.036	0.034	0.035
Cell envelope	0.072	0.070	0.073
Cellular processes	0.060	0.056	0.067
Central intermediary metabolism	0.029	0.029	0.030
Disrupted reading frame	0.034	0.035	0.036
DNA metabolism	0.000	0.000	0.000
Energy metabolism	0.076	0.079	0.077
Fatty acid and phospholipid metabolism	0.025	0.024	0.026
Glimmer rejects	0.000	0.000	0.000
Conserved hypothetical proteins	0.217	0.229	0.200
Mobile and extrachromosomal element functions	0.014	0.015	0.016
True hypothetical proteins	0.001	0.001	0.000
Protein fate	0.041	0.039	0.042
Protein synthesis	0.048	0.046	0.047
Purines, pyrimidines, nucleosides, and nucleotides	0.019	0.019	0.020
Regulatory functions	0.033	0.034	0.033
Signal transduction	0.002	0.002	0.002
Transcription	0.014	0.012	0.013
Transport and binding proteins	0.058	0.056	0.057
Unclassified	0.127	0.130	0.134
Unknown function	0.066	0.064	0.064
Viral functions	0.001	0.000	0.000

Figure A2.5 Gene frequency bar graph of functional categories within *Legionella pneumophila* strains. Functional Category key is presented in Figure 3.12.

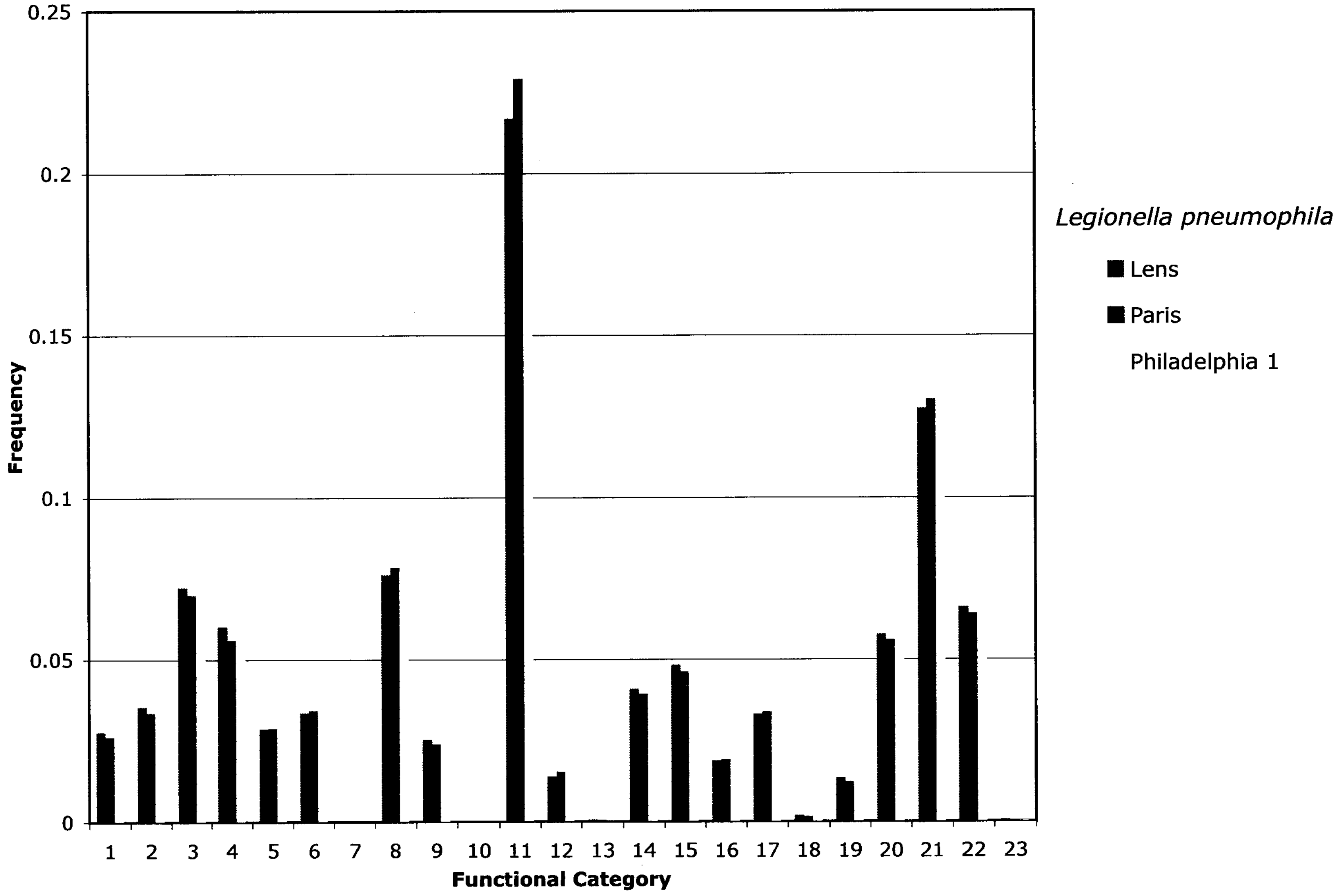
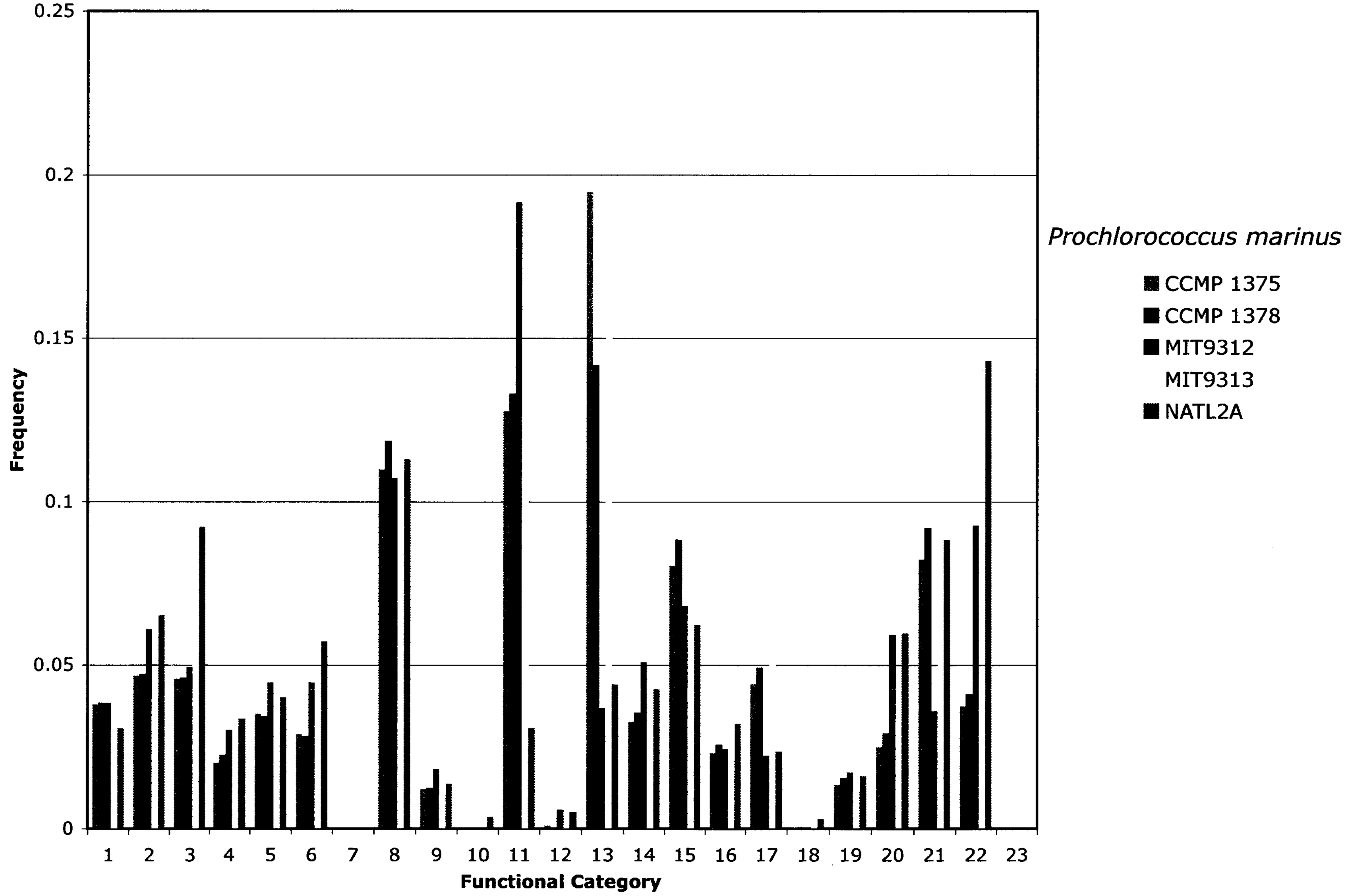

0.25
0.2
0.15
0.1
0.05
0
Frequency
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Functional Category
Legionella pneumophila
■ Lens
■ Paris
Philadelphia 1

Table A2.6 Functional gene category frequency output for *Prochlorococcus marinus* strains.

	Prochlorococcus marinus strain				
Functional Category	*CCMP 1375*	*CCMP 1378 MED4*	*MIT 9312*	*MIT9313*	*NATL2A*
Amino acid biosynthesis	0.038	0.039	0.039	0.032	0.031
Biosynthesis of cofactors, prosthetic groups, and carriers	0.047	0.047	0.061	0.040	0.065
Cell envelope	0.046	0.046	0.049	0.054	0.092
Cellular processes	0.020	0.023	0.030	0.026	0.034
Central intermediary metabolism	0.035	0.034	0.045	0.035	0.040
Disrupted reading frame	0.029	0.028	0.045	0.032	0.057
DNA metabolism	0.000	0.000	0.000	0.000	0.000
Energy metabolism	0.110	0.119	0.107	0.108	0.113
Fatty acid and phospholipid metabolism	0.012	0.012	0.018	0.011	0.014
Glimmer rejects	0.000	0.000	0.000	0.000	0.004
Conserved hypothetical proteins	0.128	0.133	0.192	0.136	0.031
Mobile and extrachromosomal element functions	0.001	0.000	0.006	0.001	0.005
True hypothetical proteins	0.195	0.142	0.037	0.171	0.044
Protein fate	0.033	0.035	0.051	0.033	0.043
Protein synthesis	0.080	0.088	0.068	0.078	0.062
Purines, pyrimidines, nucleosides, and nucleotides	0.023	0.026	0.024	0.018	0.032
Regulatory functions	0.044	0.049	0.022	0.052	0.024
Signal transduction	0.000	0.001	0.000	0.000	0.003
Transcription	0.013	0.015	0.017	0.011	0.016
Transport and binding proteins	0.025	0.029	0.059	0.034	0.060
Unclassified	0.082	0.092	0.036	0.089	0.088
Unknown function	0.038	0.041	0.093	0.039	0.143
Viral functions	0.000	0.000	0.000	0.000	0.000

Figure A2.6 Gene frequency bar graph of functional categories within *Prochlorococcus marinus* strains. Functional Category key is presented in Figure 3.12.


0.25
0.2
0.15
0.1
0.05
0
Frequency
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Functional Category
Prochlorococcus marinus
CCMP 1375
CCMP 1378
MIT9312
MIT9313
NATL2A

Appendix 3. Joint Counts and Distribution of Conserved Hypothetical Proteins

Table A3.1 Conserved hypothetical protein joint counts, generated by Genespat v.4, for *Bacillus anthracis* strains.

	B. anthracis strain							
Distance Class	A0039	Ames	Ames Ancestor	Sterne	str. France	str. Kruger	Vollum	Western North America USA5163
1	10.7976	6.8823	7.1180	10.3133	10.6264	10.7222	10.8023	10.8496
2	7.3991	4.2640	3.8758	6.3157	7.5509	7.8102	7.3025	7.4440
3	6.0331	0.9217	1.1560	4.4817	5.7267	6.2449	5.8448	6.0082
4	4.1979	-0.0705	-0.7336	2.7690	3.7893	4.4775	4.8809	4.4239
5	2.4091	0.8052	0.7667	2.8870	1.9878	2.2350	1.8728	2.2047
6	2.9350	-0.8943	-1.0278	2.0298	3.2103	4.0858	2.8424	3.4110
7	3.6882	-0.4025	-0.5322	-0.0388	4.1451	3.7733	4.3718	3.4637
8	2.3297	-0.4921	0.0955	3.1738	2.2501	3.4606	2.4024	2.7246
9	2.6503	0.4855	-0.0017	1.7153	2.6605	3.7939	1.4290	1.9196
10	1.3349	-1.6564	-1.6116	2.1216	0.1147	0.6040	1.4126	1.1289
11	2.1107	1.3848	1.8807	-0.4586	2.6394	2.2806	2.2711	2.1591
12	1.8530	2.8382	2.7078	0.8485	1.6152	2.1373	1.7732	2.0971
13	2.8376	1.0188	1.3349	0.7497	3.1126	3.3537	2.4795	2.8844
14	2.6416	1.9502	1.1012	1.5914	2.3181	3.1694	2.5621	2.7148
15	0.9072	2.3125	2.0959	1.1816	0.3781	1.3243	0.6402	1.2238
16	1.7786	0.3330	0.4770	1.3022	3.6523	3.0011	2.9936	2.6931
17	2.5713	-0.0103	0.0322	2.2842	0.6736	1.3445	2.0387	1.4845
18	1.3140	1.9297	1.6313	3.7625	1.4899	0.7319	1.5904	1.7412
19	3.9341	1.2172	1.8922	3.1764	3.9842	3.5508	3.2884	3.6257
20	3.6470	-0.4549	-0.0538	3.7902	5.0850	5.8515	4.0899	4.9214
21	3.4733	1.6312	0.9721	1.8039	3.2360	4.7841	3.4595	3.2798
22	3.4582	0.1727	-0.0425	2.4211	3.5022	3.2112	3.0146	2.5188
23	1.7327	-0.2793	-0.5024	2.2708	0.8777	1.9102	2.6877	3.2751
24	2.0916	0.3673	1.5810	1.2279	2.2946	2.5259	1.4933	2.4785
25	2.5304	1.4513	1.2296	1.1352	1.2882	3.5937	1.6236	0.9324
26	3.1316	1.6266	2.1274	1.4473	2.6316	3.0217	2.7000	2.3951
27	3.4995	2.3684	1.5343	1.7939	1.2416	0.4147	2.3553	2.9441
28	-0.1907	2.0136	2.0658	-0.1838	0.0181	1.4894	0.5850	-0.6427
29	-0.1556	1.5607	1.7762	1.4504	1.4340	0.4989	-0.3275	-0.7784
30	-0.7630	0.1095	-0.6449	0.0592	-1.1333	-0.8049	-0.6886	-0.7181
31	0.3803	0.0274	0.0812	-1.0972	-0.3221	1.4175	0.0894	2.0981
32	-0.3417	-0.6609	-0.4281	0.5406	-0.1618	-0.6623	0.0854	0.2317
33	0.3127	-0.5526	-0.2371	0.1913	-0.0372	0.2343	0.6010	-0.7332
34	0.2969	0.5220	1.0144	-0.1099	1.5047	0.5305	0.7836	0.7071
35	0.2373	-0.3802	-0.9711	1.2993	2.2465	2.6743	-1.2578	-0.3348
36	-0.1203	1.2951	1.3505	1.1442	-0.6508	-0.3211	1.5150	1.3335
37	-0.7586	-1.5696	-1.3416	1.4231	-1.2593	-0.8080	-0.4497	-0.3463

	B. anthracis strain							
Distance Class	A0039	Ames	Ames Ancestor	Sterne	str. France	str. Kruger	Vollum	Western North America USA5163
38	0.3720	-1.3719	-1.5105	0.4412	-0.1695	-0.0185	0.7122	-1.0807
39	0.9041	0.8618	0.3811	2.0742	0.3417	0.0445	0.8217	2.1962
40	0.7598	1.8855	1.4916	0.9986	2.3337	0.6923	0.7753	1.9721
41	2.2354	1.6039	2.1053	0.7848	1.1524	2.7043	1.7377	1.4706
42	1.7226	0.5679	0.4269	2.1762	1.3060	1.6846	1.8556	2.7543
43	1.8592	1.2878	0.4193	2.7044	1.6259	2.4251	2.8857	0.7493
44	0.6238	-1.0017	0.0366	1.0047	-0.1184	0.7307	0.6616	0.7514
45	2.1981	0.1815	1.4132	2.6325	0.4581	-0.2094	0.7856	1.2123
46	2.9815	0.3995	0.3483	2.7540	1.9528	3.1666	0.2565	0.4063
47	0.7104	1.6435	1.4210	0.1908	2.1252	2.7244	1.9359	1.7446
48	1.2732	0.5833	0.5437	1.6462	1.4199	0.5859	2.5577	1.1421
49	2.5175	-0.0391	-0.4398	2.2241	2.6634	1.4917	1.7183	1.8049
50	2.0099	0.2233	-0.2633	2.5460	0.3497	2.0498	1.8520	2.3473
51	1.3102	-0.3991	0.1909	0.8652	2.1291	0.3643	1.4896	0.7875
52	1.7793	-0.2341	0.3559	0.9392	0.8985	0.4368	0.2444	0.7348
53	1.0256	0.6693	1.1585	-0.9391	1.6334	0.2797	1.4894	1.5108
54	1.9892	2.0106	2.0464	0.5857	0.4821	-0.5936	0.0418	0.2914
55	1.4634	1.5570	0.9979	1.2775	0.7589	0.9153	1.8727	1.7895
56	-0.0377	1.2365	0.7469	0.5828	-0.0394	0.5571	1.4309	0.3708
57	1.2049	-0.6468	-0.2483	-0.1739	-0.1158	0.2109	-0.6512	-0.1338
58	0.5078	0.6178	0.3979	1.1720	1.7382	1.6436	0.8423	1.4275
59	-0.1816	0.6178	0.5858	-0.9875	-0.0235	0.5124	0.8342	0.3313
60	0.3405	-0.8121	-0.3122	-0.2439	0.8838	2.1038	0.5041	1.5231
61	0.6991	1.3619	0.9621	-0.4526	0.7629	1.4544	-0.2888	0.5615
62	2.4438	-0.3316	-0.9104	0.3235	1.8554	-1.1867	2.4023	1.3896
63	0.5478	-0.0460	0.6204	0.7402	-0.7448	0.7742	0.3528	0.3352
64	-0.7759	-0.7033	-0.3009	0.6438	-0.6181	0.7135	-1.4205	-0.3570
65	0.7581	1.9145	1.4210	2.0927	-0.3413	-0.6437	-0.4037	0.6915
66	0.1929	2.8410	3.0700	0.0152	0.4500	0.7015	1.4054	-0.0266
67	1.5084	0.6370	0.6012	-1.0845	0.3335	0.7580	0.9804	1.1306
68	1.4250	1.1033	-0.0190	0.7737	0.5505	-0.0116	-0.1853	1.0857
69	-0.6842	2.7585	3.1682	-0.3201	3.4936	1.5795	-0.0698	-0.0108
70	0.3683	0.7542	1.4365	0.9761	-0.3333	0.7175	1.0378	-0.3413
71	-0.5732	0.9540	0.6281	2.4916	-1.1173	0.0888	-0.3409	0.5463
72	0.7271	1.9545	0.3917	1.5478	0.4903	0.5121	0.8788	1.4933
73	0.9229	4.3957	4.7166	0.1120	0.4659	-0.0963	2.7331	0.4184
74	2.7794	0.8988	0.5802	-0.4641	0.8708	1.6842	0.9075	1.9435
75	1.1858	-0.1021	0.1995	2.4553	1.6315	0.1124	0.9866	-0.0982
76	0.5352	1.3385	-0.0563	1.3428	1.1517	1.2287	1.1706	1.6092
77	-0.0071	1.1382	1.4640	1.1125	1.2016	1.7974	1.0286	0.4828
78	0.8632	0.2590	0.6704	2.3592	2.0020	1.3912	1.0868	0.8632
79	-0.0110	-0.4333	-0.8585	2.3292	0.5225	1.3066	0.3293	-0.1188

	B. anthracis strain							
Distance Class	A0039	Ames	Ames Ancestor	Sterne	str. France	str. Kruger	Vollum	Western North America USA5163
80	0.8917	-0.2601	-0.8438	1.4610	0.4697	2.2940	0.7902	1.4423
81	2.3208	2.9920	1.9598	2.3453	1.2306	-0.5932	1.6343	1.2751
82	-0.3506	1.2011	2.2472	1.3802	-0.2448	-0.0454	0.4643	1.0123
83	2.9222	-1.3258	-0.4708	2.7742	0.1647	2.1818	1.1869	1.2959
84	0.5550	-1.0877	-1.4008	1.8514	1.8118	1.1559	0.7726	1.5756
85	2.4387	2.1868	2.0620	2.4672	0.5344	-0.2189	1.6212	1.5631
86	3.8793	2.3019	2.1648	1.6945	1.1708	1.8140	2.1150	1.8668
87	1.5276	-0.0491	-0.5426	-0.4495	2.8145	1.6016	0.8478	1.4753
88	1.5605	-1.1047	-0.6145	1.0501	0.2367	1.2895	1.6675	2.8867
89	2.7408	-0.2337	0.0043	-0.5760	2.1323	0.4260	2.5117	1.6670
90	1.2595	0.6820	0.1739	0.6630	0.5872	1.4026	1.9526	1.8550
91	-0.4149	2.0466	2.0079	2.6086	1.5205	2.5530	1.3954	0.6104
92	0.2363	2.7578	3.1768	3.1777	1.3308	1.3470	0.6600	2.0132
93	0.5124	1.3030	0.7245	4.3297	0.5627	2.5914	0.3099	0.6426
94	2.0882	3.6996	3.5636	2.1181	1.4864	1.4563	0.9301	0.2318
95	2.3309	0.8940	0.9406	0.9154	2.0728	1.0198	0.8215	1.4661
96	0.7887	0.8862	-0.2409	3.1670	1.2242	2.5907	1.1940	2.1564
97	2.5278	0.3587	1.6653	1.9474	2.4700	2.0025	0.8172	2.0681
98	1.4923	0.6147	1.4881	1.2467	2.2001	1.8412	2.1705	0.4509
99	1.7276	3.1603	2.7399	1.5745	0.6155	3.1629	1.4326	2.1598
100	18.1176	18.0600	18.0280	14.7837	17.4198	17.8223	17.7720	18.2154

Table A3.2 Conserved hypothetical protein joint counts, generated by Genespat v.4, for *Campylobacter jejuni* strains.

	C. jejuni strain	
Distance Class	RM1221	NCTC 11168
1	3.1657	6.6657
2	-0.8914	3.8376
3	-3.1803	1.6205
4	-0.1745	3.4424
5	1.4446	2.0994
6	1.4814	2.5972
7	-1.7445	1.8918
8	-0.1251	0.0475
9	0.5759	0.7550
10	1.9663	1.6785
11	1.2595	1.6510
12	-0.1251	0.7496
13	-0.5825	1.4636
14	1.9659	2.3774
15	-1.4798	1.9567
16	-1.9494	0.5399
17	1.5038	2.8243
18	-0.8053	1.5115
19	-0.5584	0.7946
20	-2.4040	-1.0151
21	-1.4756	-1.0283
22	-0.7738	0.3588
23	0.1393	0.1044
24	-0.7658	-0.5673
25	0.6367	-0.3139
26	-1.6920	0.5592
27	-1.9055	-0.5219
28	-1.4380	-1.9206
29	1.3550	0.1715
30	1.8240	1.8166
31	1.3593	1.1162
32	-0.9582	-0.2766
33	-0.2533	0.1811
34	-0.0126	0.6685
35	0.4393	0.1907
36	-0.2533	0.6734
37	1.1404	1.3468
38	2.3164	-0.7029
39	2.7915	-1.3946
40	0.9125	-0.9566

	C. jejuni strain	
Distance Class	RM1221	NCTC 11168
41	1.8655	-0.0025
42	-0.2202	-1.3729
43	0.4477	-0.4387
44	-2.5414	-2.7593
45	-0.4489	-1.3848
46	1.4139	-0.9256
47	0.2496	-0.9210
48	1.6798	0.9708
49	-0.2036	-2.0762
50	-1.1165	-1.6123
51	0.7507	-2.5220
52	1.2314	-0.4061
53	-1.0970	-0.8852
54	0.2969	-2.0203
55	-0.8552	-2.4901
56	0.3228	-0.1541
57	-1.5415	-1.0999
58	-0.3710	-1.7696
59	-0.6132	0.5353
60	-1.0655	-1.2940
61	-0.8430	-0.8310
62	-0.3667	0.7628
63	1.5063	-0.1683
64	-1.7679	-1.3022
65	-1.7607	0.3561
66	1.2768	1.2653
67	-0.5967	1.2916
68	-0.0990	1.5306
69	-1.2686	-1.9734
70	0.3747	2.0039
71	-1.2724	1.2802
72	2.5131	-1.4926
73	-2.1864	0.3364
74	-1.4880	0.8492
75	0.8803	0.1370
76	1.5670	0.1467
77	1.3366	2.2821
78	0.6314	0.4306
79	-0.0652	-0.3260
80	-0.2921	0.8693
81	-0.5274	0.4004
82	-0.0652	-0.5092

	C. jejuni strain	
Distance Class	RM1221	NCTC 11168
83	-1.9285	-0.5227
84	1.3552	-1.4403
85	0.4227	1.1192
86	0.6715	-0.2553
87	-2.3821	1.8337
88	0.4446	-0.0104
89	-0.7305	0.2004
90	-0.9584	0.2003
91	-0.4740	1.8892
92	-0.4822	1.9279
93	0.9480	0.4852
94	-0.0185	0.0382
95	-1.4264	0.4550
96	0.0028	0.7052
97	-0.4821	0.5257
98	2.3907	-0.6884
99	0.0456	0.2892
100	-0.0527	-0.1938

Table A3.3 Conserved hypothetical protein joint counts, generated by Genespat v.4, for *Chlamydia pneumoniae* strains.

	C. pneumoniae strain			
Distance Class	J138	TW-183	CWL029	AR39
1	6.3807	4.7813	2.6965	4.1593
2	1.9466	3.2948	-0.1025	1.8961
3	2.8077	2.7612	1.3145	0.6199
4	3.0511	3.8268	2.7331	1.9525
5	1.9662	0.0191	1.3274	0.3323
6	0.3675	0.1424	0.2736	0.0437
7	1.3608	0.2792	-0.7815	-0.7627
8	2.0361	0.7145	0.6392	-0.9265
9	1.8199	0.5197	0.2912	0.3971
10	-0.7028	0.7366	0.6513	-0.9021
11	0.3058	0.9675	1.3663	-0.3850
12	0.2679	0.1759	-0.4059	1.0793
13	0.2769	0.0222	1.3727	2.0778
14	0.1282	0.0022	0.3148	0.1204
15	0.0837	-0.3324	0.6696	-0.0183
16	-0.9042	0.3642	-0.0400	0.6662
17	0.9459	1.9813	2.4438	2.1337
18	1.4954	2.3595	-0.3949	0.5277
19	0.5186	1.4302	-0.7549	0.8709
20	-0.5260	1.3008	-1.4594	-1.4276
21	0.7429	0.6265	-0.3948	0.2367
22	0.4380	-0.9770	-1.8142	0.7190
23	-0.7272	-1.2596	-1.1045	-1.2383
24	-1.2031	-0.4027	0.3089	-1.8763
25	0.1390	-0.6002	-1.1146	-0.5520
26	0.0101	-1.1268	-0.0514	-0.3865
27	1.1943	0.0864	1.0117	-1.8767
28	2.6218	1.4631	-1.1197	-1.2144
29	0.7014	1.6599	0.2971	0.2758
30	1.2037	0.9418	1.7207	0.7726
31	1.3603	1.3249	1.0118	-0.8955
32	2.4290	1.4725	-1.1147	-0.0554
33	1.1712	0.4584	-0.7602	-0.3991
34	0.6324	1.1916	0.3030	-0.7299
35	1.4168	-0.8572	-0.7602	-1.8767
36	1.9064	0.1011	-0.7602	-1.2144
37	0.3358	1.4431	2.4292	-0.5646
38	1.4606	1.3383	-1.4690	-1.2145
39	0.1970	-0.6819	-1.1146	-1.8658
40	0.0853	-0.0339	-0.0514	-1.3803

	C. pneumoniae strain			
Distance Class	J138	TW-183	CWL029	AR39
41	-1.2383	-0.3092	-1.1196	0.1232
42	1.0061	-0.1479	-0.7655	-1.5344
43	3.8741	-0.1022	-0.4059	-2.3515
44	1.5717	0.2732	0.6574	-4.3422
45	0.2345	-0.2820	-0.0514	-1.8656
46	0.0009	0.4765	-1.4692	-0.0297
47	1.6542	1.2792	-0.7603	-0.8590
48	1.4635	1.6311	-0.0514	0.4547
49	-0.5454	0.8609	-1.1146	-0.3614
50	-0.6695	1.6153	-1.1146	0.6336
51	0.6361	0.4840	-2.5322	-1.5218
52	-0.3290	1.0359	-2.1778	1.1445
53	-1.2075	-0.4315	-0.0514	-0.1827
54	-0.9673	0.0232	-1.8279	-0.1827
55	-0.0655	0.6280	-0.0514	-0.3611
56	0.8252	0.0522	-1.1096	-1.6755
57	0.1243	0.5644	-0.7550	-1.3436
58	-0.8303	-0.4886	1.7273	1.1443
59	0.5463	-0.7971	-0.0457	-0.8584
60	0.3530	-0.0975	0.3089	-0.3358
61	1.6289	0.6432	1.0181	-0.3359
62	0.1914	0.6184	1.0181	-0.5018
63	0.7143	-0.8901	2.4366	-1.9954
64	-0.0181	0.3673	-0.4003	-0.8338
65	0.7338	-0.0547	-2.8827	-1.8298
66	0.9595	0.3879	-0.7602	-0.1699
67	-0.6493	0.9760	3.4926	1.3102
68	-0.1798	-0.0546	0.3030	0.8124
69	-0.7116	-1.0577	-0.4113	0.6466
70	1.0728	0.5018	1.3529	-1.1777
71	-0.0853	-0.9085	1.3595	0.4808
72	0.3621	0.8549	0.6573	-0.3485
73	1.3269	-0.6066	1.0054	0.6467
74	1.2184	-0.4623	2.0679	-0.3359
75	0.1062	-1.1020	0.3030	-1.1536
76	0.8540	0.5530	-0.4058	-0.4893
77	1.0667	-0.2327	-0.0514	-0.3105
78	0.4404	-0.1981	3.1382	0.8531
79	0.5791	-1.0088	1.3595	-0.7969
80	-0.2386	-0.0402	0.3030	0.8530
81	0.0009	-0.0973	-1.4690	0.5339
82	0.0394	2.3402	-0.4058	1.3516

	C. pneumoniae strain			
Distance Class	J138	TW-183	CWL029	AR39
83	-0.7384	1.4721	-1.1096	0.7136
84	-1.0762	0.2071	2.4438	0.2012
85	1.1800	0.0253	0.6696	2.5301
86	-0.1054	1.1935	0.3148	1.5037
87	0.4352	0.5941	-1.4594	0.3543
88	0.3853	0.0077	0.3148	-0.3105
89	-0.1245	-0.7863	-1.1045	0.1882
90	0.0879	0.3414	1.0245	-1.6519
91	-0.1149	0.3864	-0.0400	-1.8066
92	-1.5981	-0.9831	2.0890	-2.8146
93	0.5252	-0.3557	1.3793	-0.6554
94	-0.8208	0.3134	-1.1096	0.3544
95	-0.0282	-1.3318	-0.0514	0.6869
96	-1.9253	-1.4228	2.0750	0.3544
97	-0.4719	-2.6173	2.0750	1.3518
98	0.3305	-1.7607	0.3030	3.1651
99	0.3206	-0.4770	0.6574	1.8222
100	-0.2577	-0.1755	-0.1890	-0.0100

Table A3.4 Conserved hypothetical protein joint counts, generated by Genespat v.4, for *Escherichia coli* strains.

	E. coli strain			
Distance Class	CFT073	K12-MG1655	O157:H7 EDL933	O157:H7 VT2-Sakai
1	9.3387	8.9253	15.3526	11.5111
2	7.2037	4.7498	10.8951	5.8362
3	5.6507	1.5266	8.2525	4.5855
4	5.1910	0.4550	7.3659	2.8873
5	5.6127	-0.0013	5.6502	2.7148
6	4.5161	1.0765	4.8947	1.6067
7	2.4641	0.4933	4.8896	2.2228
8	3.4800	-1.0789	5.3411	2.5989
9	2.7448	-1.0652	6.3536	0.2332
10	3.6783	-1.4573	4.9251	1.3226
11	2.0801	-2.1387	4.4876	1.3642
12	2.9851	-0.8280	5.2639	0.9957
13	0.6262	-1.9209	4.1043	1.2227
14	1.3287	-0.1597	4.6878	0.6823
15	1.6698	0.1440	3.0488	1.1182
16	1.8629	-0.9307	2.5506	-0.1061
17	1.2664	0.5556	2.7875	0.4870
18	2.0408	-1.4255	3.7937	2.8861
19	3.1718	0.6585	1.3465	1.5256
20	-0.3289	0.0742	3.4823	0.3399
21	3.2835	1.1542	2.7484	0.2441
22	2.2755	1.7332	2.0846	-0.5807
23	0.7489	0.8742	1.1232	-0.6942
24	0.4994	1.6500	1.4485	-1.5864
25	0.7419	0.4861	2.1546	0.6641
26	1.2519	-1.1612	1.7085	-0.1693
27	2.2368	0.8049	2.2353	0.1145
28	1.5613	0.7806	0.9830	0.4392
29	1.3319	-0.7540	2.1815	1.7091
30	0.1847	-0.5387	1.6432	2.4716
31	-0.3751	0.4117	1.3050	-0.5983
32	-0.1861	-0.4357	2.5179	-0.3393
33	1.5821	0.3325	0.8746	-0.5591
34	2.3383	-1.5136	1.7192	1.5280
35	1.2225	2.7135	2.6784	2.1054
36	0.9126	0.8388	2.4479	-1.1949
37	0.9381	-0.3418	1.9347	-2.3081
38	0.6377	-0.5293	0.8175	0.5871
39	1.9467	0.7598	1.7038	0.0333
40	1.0544	-0.8015	0.7960	-0.3453

	E. coli strain			
Distance Class	CFT073	K12-MG1655	O157:H7 EDL933	O157:H7 VT2-Sakai
41	-0.8698	0.3804	-0.0228	0.7270
42	-1.2470	-0.1073	-0.2485	0.3058
43	-0.7654	-0.2996	2.0547	-1.8347
44	1.0025	-0.5905	-0.1644	-0.9178
45	-1.3993	0.3058	-0.0942	0.4833
46	0.0375	-0.4733	-1.3420	0.2023
47	0.1180	-0.5626	0.0066	-0.1458
48	-0.1838	-1.6383	0.4886	1.4303
49	-0.5518	0.2263	2.5692	0.3426
50	-0.0402	-0.5532	0.5781	0.5352
51	0.7391	-1.4409	1.8333	0.9282
52	-1.7601	1.3406	-0.8697	-0.2730
53	-0.2161	0.1372	2.1848	0.4608
54	-0.4195	0.0386	0.7534	1.9112
55	-1.7109	-0.2384	0.5750	0.1234
56	0.3990	-0.6332	0.2312	2.2209
57	-0.2625	2.5193	2.0925	-1.2539
58	1.3671	1.9375	-0.3093	-0.2548
59	0.0862	-0.7133	-1.3853	-1.3304
60	-0.4537	3.1266	2.6314	-0.8576
61	0.0759	0.7628	-1.3229	-0.2584
62	-1.7501	1.6764	0.1372	-1.6110
63	-2.2052	0.9897	0.0779	-0.2656
64	-0.2090	0.2136	-0.6652	1.1473
65	-0.0618	0.7176	0.4942	-1.0699
66	-0.1662	0.9105	-0.2350	0.9065
67	-1.8364	2.4037	-1.1565	-0.7041
68	-2.1531	0.6041	1.4532	0.1270
69	-1.3334	0.7224	-0.9945	1.1150
70	0.5089	1.1279	-1.2676	0.9137
71	0.3782	-0.1581	-0.0906	-1.0746
72	-1.2000	-1.5338	0.0779	-2.7584
73	0.1072	-2.8113	-0.6676	-0.5897
74	-1.4617	-1.4302	-0.5119	-1.6650
75	0.3250	0.5585	0.1136	-0.9428
76	0.1353	0.4691	1.0468	-0.0783
77	-0.4658	-1.0022	-0.9761	0.0672
78	-0.1742	-0.7994	0.8828	1.6256
79	0.2373	-0.7050	-0.5145	-1.8424
80	1.1790	-1.1956	2.0002	0.5870
81	-0.0966	0.8945	0.2061	0.2713
82	0.1389	2.1879	-1.4390	-0.3751

	E. coli strain			
Distance Class	CFT073	K12-MG1655	O157:H7 EDL933	O157:H7 VT2-Sakai
83	0.6253	1.0083	-0.1832	-0.8073
84	-0.6786	0.6068	-0.5021	-1.7335
85	-1.4042	1.1074	-2.2146	-0.4211
86	-0.2497	0.8198	-0.9627	-0.6605
87	-0.6338	1.5339	-2.3507	-0.2459
88	0.8636	2.0197	-1.1113	0.6094
89	0.5982	-0.8657	-0.9891	-0.0344
90	0.4882	1.2312	0.6561	1.5469
91	0.3183	1.1418	0.3846	-1.2647
92	-0.4710	-0.1290	-1.0506	-0.6426
93	-0.8751	-0.5449	-0.7518	0.6093
94	1.1819	1.5587	-0.3171	-0.1602
95	-2.5561	0.6652	-1.8328	1.3492
96	-0.6370	1.4693	0.1649	-0.4388
97	0.3600	2.0497	-0.3471	-0.4706
98	-0.0478	0.9726	-0.5616	0.5312
99	-0.6506	0.4665	0.8121	1.0624
100	16.3509	2.0215	15.9237	15.8851

Table A3.5 Conserved hypothetical protein joint counts, generated by Genespat v.4, for *Legionella pneumophila* strains.

	L. pneumophila strain		
Distance Class	Lens	Paris	Philadelphia 1
1	7.6458	8.4987	7.2320
2	5.9509	6.2672	4.2110
3	4.4390	3.7699	4.1623
4	4.2151	3.9562	4.8222
5	4.7784	6.1054	2.8868
6	2.2564	2.6587	2.7185
7	4.3886	3.8469	2.9784
8	2.8672	2.6841	0.1329
9	1.6477	3.8594	1.7661
10	3.4713	1.8326	1.0136
11	2.0331	2.9807	1.6563
12	0.4924	1.5348	-0.0798
13	2.0006	4.2469	-1.1036
14	2.1775	1.8675	-0.2510
15	1.0337	2.4935	0.8012
16	1.3457	1.9145	-0.1323
17	1.1069	0.5664	-0.1979
18	1.4125	3.5693	0.7216
19	0.4072	2.2175	2.6093
20	1.8993	3.9357	1.3098
21	1.1214	1.7182	1.1220
22	2.1330	2.2768	0.9386
23	2.3870	1.7295	1.7890
24	0.1248	0.8845	1.1956
25	1.2618	2.1020	1.0475
26	0.8862	2.9882	-0.0427
27	1.5349	1.4071	2.7471
28	0.7204	1.0161	1.2453
29	1.5129	1.5508	-0.1523
30	1.5795	1.7692	1.6999
31	1.7366	2.4017	1.2306
32	2.0498	3.2513	0.2182
33	1.0142	3.4461	2.9350
34	1.9715	2.1809	2.5511
35	2.0895	2.7800	1.6013
36	0.3590	1.6829	0.5324
37	3.0256	2.0288	2.4296
38	0.8503	0.9234	0.8041
39	2.5848	1.8953	1.7661
40	2.4484	0.6185	0.3476

	L. pneumophila strain		
Distance Class	Lens	Paris	Philadelphia 1
41	2.0105	1.7345	1.2794
42	0.1776	2.1139	1.4142
43	1.4109	0.9910	-0.2333
44	3.2421	2.0964	-0.7785
45	-0.1114	2.9559	-0.1144
46	0.2095	2.9604	1.7907
47	-0.8205	1.4169	-0.2239
48	0.1238	1.1052	0.2669
49	0.3644	1.4110	-0.2239
50	0.4985	1.8846	0.8678
51	-0.2941	0.3362	1.4383
52	-0.5548	2.7762	-0.6133
53	0.0339	0.7493	0.3858
54	1.6454	1.5781	0.0284
55	0.2735	0.5497	-0.0623
56	-0.1027	-0.3646	-0.2098
57	1.2178	1.1967	-0.3289
58	1.6858	-0.5783	-0.1058
59	0.2679	2.2049	-1.4586
60	-0.5901	1.2723	0.7874
61	-1.2082	1.1036	-0.2004
62	2.5033	0.7995	-0.2160
63	1.2226	-0.2657	-1.2235
64	-0.7892	-0.0856	-0.4881
65	-1.4508	0.9991	-0.8101
66	-0.8918	0.5404	1.5384
67	1.2892	-0.1572	-0.8193
68	1.0876	-0.0472	0.8018
69	0.2089	-2.4244	-1.4966
70	-1.5776	-0.0250	-1.0808
71	1.0188	-1.8939	-0.4412
72	1.4579	-0.7058	1.5845
73	1.2837	-1.2011	-0.5749
74	0.5462	-2.2801	-1.9980
75	0.6224	1.2466	-1.0713
76	-1.0406	0.5086	-2.3274
77	0.3321	0.6554	-2.5966
78	1.2147	-0.1571	-2.3353
79	0.4614	0.4692	1.3695
80	-0.0497	0.1344	-0.3307
81	-0.1414	2.1994	0.3256
82	1.0624	1.9185	0.8654

	L. pneumophila strain		
Distance Class	Lens	Paris	Philadelphia 1
83	-1.6644	0.9206	0.2008
84	-0.7165	3.8469	-0.3328
85	0.5690	2.4877	-0.8230
86	-1.7522	0.8616	-0.4455
87	-0.6236	-0.0969	-0.5731
88	-0.0285	0.1454	-0.9980
89	-1.2352	1.0634	-0.1685
90	-1.8753	-0.4342	0.0664
91	0.4016	-0.8288	-0.3798
92	-0.1806	-0.8128	-1.2103
93	0.1271	0.2767	-0.4128
94	0.3804	1.3781	-0.8865
95	1.4050	0.3112	-0.2029
96	0.4609	-0.7078	0.3316
97	2.3398	-1.1625	2.3109
98	1.3988	1.4474	-1.9816
99	2.1357	1.5577	0.7390
100	-0.1666	-0.2036	-0.1009

Table A3.6 Conserved hypothetical protein joint counts, generated by Genespat v.4, for *Prochlorococcus marinus* strains.

	P.marinus strain				
Distance Class	CCMP 1375	CCMP 1378 MED4	MIT 9312	MIT9313	NATL2A
1	2.0603	0.9460	5.9293	2.1726	-0.0271
2	0.7916	0.2554	2.1413	1.3299	-0.0206
3	0.1062	2.2724	0.5471	1.3915	-2.0039
4	1.6513	-0.0590	2.2969	1.9933	-2.0015
5	0.5110	-0.5317	2.0117	0.3606	-0.0123
6	0.5754	1.0212	2.3231	-1.5500	-2.0017
7	-1.7132	-1.4450	-0.5751	-0.0429	-1.9976
8	1.2178	0.4237	1.0543	-1.0542	1.9849
9	1.2667	1.2120	0.3901	0.1268	0.9813
10	2.1834	1.9632	1.0221	-1.3452	-1.0004
11	-0.1421	-0.5387	0.6024	2.4970	0.9884
12	0.1061	-0.0629	1.5354	-0.2299	-1.9990
13	0.4307	-0.8055	0.2577	1.3591	-0.0021
14	-2.3352	-1.1753	-0.5225	1.2709	1.9950
15	-0.6898	-2.0541	-1.0181	0.5074	0.9895
16	-1.4252	-0.7866	-1.3226	-0.2846	2.9801
17	-1.5894	-0.6511	-0.8204	0.4107	-0.0086
18	-1.0851	-1.3399	-0.0385	-1.3850	0.9931
19	-0.3642	-1.3874	0.1187	-0.5508	1.0013
20	2.6981	0.6779	-0.1898	-0.5069	-0.9997
21	1.7903	1.1089	0.1489	-1.4308	-0.9990
22	-0.6823	1.6334	0.9978	0.8571	-1.9916
23	-0.0444	-2.0230	0.5067	-2.1605	-1.9946
24	-0.4010	-0.6287	0.0035	1.4188	1.0061
25	1.5287	-2.4218	-0.4520	1.6784	5.0100
26	0.1536	-1.7610	1.1998	-1.3440	2.0077
27	0.6509	-1.0885	0.0517	-0.3112	0.0139
28	-0.1893	0.7272	0.1913	-0.3727	1.0167
29	0.2806	1.1294	-2.4024	0.2649	0.0167
30	0.2242	1.1002	0.8904	-0.1089	-0.9873
31	0.2805	-1.3303	-0.1003	-1.8792	0.0073
32	0.5218	-3.1474	-0.3990	2.4452	-0.9879
33	0.2768	-0.6506	1.4157	0.6929	0.0111
34	-0.5773	-0.8653	-0.2346	-0.2041	1.0166
35	0.1868	-1.5215	-0.7219	1.3251	0.0176
36	-0.9236	-1.6695	0.7768	2.1691	1.0251
37	-2.0194	0.5122	-0.8860	-0.1970	1.0214
38	0.3106	0.3097	0.4412	1.6067	3.0286
39	-0.4285	2.5844	1.0991	0.7966	0.0176
40	1.3756	-0.6052	-0.6574	-0.4891	1.0309

	P.marinus strain				
Distance Class	CCMP 1375	CCMP 1378 MED4	MIT 9312	MIT9313	NATL2A
41	0.4903	-1.6462	-0.1987	-0.6170	0.0252
42	0.4678	3.0005	-0.0038	-1.1779	1.0298
43	-0.3138	-1.0081	0.4969	-0.1795	0.0195
44	0.0546	0.1305	0.3506	2.2237	-0.9726
45	-1.0563	-0.6053	0.0143	0.7364	0.0309
46	0.4595	0.7812	0.6932	0.3182	0.0338
47	-0.6042	-1.6940	-0.7933	-0.7392	-0.9711
48	-0.0315	-0.0974	0.8582	0.7855	-0.9684
49	-0.1162	1.0428	1.6564	3.7530	-1.9752
50	3.0310	-0.0856	2.0137	-1.0042	2.0522
51	0.2763	-0.3519	-0.4509	0.4376	-0.9662
52	-0.3605	-0.1325	0.3936	0.2108	3.0673
53	-1.0460	-2.0755	0.5338	0.4012	-0.9711
54	1.0558	-0.2629	2.0676	-0.2212	3.0537
55	1.0727	-0.4762	-0.0842	1.2819	0.0347
56	-1.9237	-0.1558	1.6960	0.3430	-0.9654
57	-1.6468	2.3063	1.3588	0.1752	-0.9696
58	-0.0647	0.9735	1.9280	-1.0333	-1.9732
59	-2.3572	1.2926	0.8130	0.0070	-0.9633
60	-0.2958	0.1945	-0.3913	1.3504	-1.9728
61	1.7278	0.0056	-0.5390	-0.0036	1.0513
62	1.5517	2.1935	-0.0357	-0.7949	2.0636
63	0.8010	0.9072	-0.1775	-0.1233	3.0672
64	0.4094	-0.9372	1.1577	-1.4810	-1.9704
65	-0.0721	-1.1961	1.3303	-1.3289	0.0471
66	0.2230	1.1454	-1.3004	3.0170	-0.9528
67	-2.0238	-0.0460	-0.3193	0.8664	-0.9520
68	0.4668	-0.6401	-1.6214	0.6571	2.0768
69	0.0169	0.6644	0.5497	-1.2589	-0.9563
70	3.3757	1.6088	-0.6047	1.7857	-0.9555
71	-0.2007	0.2594	-2.4068	0.4878	-1.9647
72	0.5221	-1.7372	0.4143	0.0530	-0.9450
73	-1.3182	-0.4754	-0.5748	3.4129	0.0644
74	-0.2812	-0.9207	-0.0924	-1.4753	0.0538
75	-0.4331	-0.2076	0.7670	-0.0950	0.0654
76	-0.2663	0.4930	0.2412	0.1351	-1.9637
77	-0.4726	-0.1565	0.6319	-0.5511	2.0917
78	-0.4259	0.4972	-1.2200	1.5920	0.0625
79	0.6988	1.9407	-0.9434	-0.1193	1.0780
80	0.5169	0.2633	-0.2771	0.3628	-1.9652
81	-2.2656	0.7227	-2.2062	1.8065	-0.9498
82	0.0132	0.1901	-1.3585	-0.4315	-0.9456

	P.marinus strain				
Distance Class	CCMP 1375	CCMP 1378 MED4	MIT 9312	MIT9313	NATL2A
83	0.5518	-0.5244	0.1547	1.2665	2.1124
84	-0.5936	1.5852	1.6689	-1.1825	-0.9399
85	-0.8485	1.7068	0.3096	0.5806	0.0664
86	1.0404	-0.2611	1.8223	-0.6122	0.0635
87	-0.2149	1.2660	-0.6470	0.0116	1.0916
88	0.3141	2.5799	1.8561	0.3553	1.0902
89	0.1808	1.8319	-1.3233	-0.5575	-0.9441
90	1.6708	1.0823	-0.4790	-1.9558	0.0722
91	1.9583	0.7939	-1.2835	1.5306	-0.9398
92	0.7653	1.7103	0.5336	1.0474	-0.9391
93	0.2909	-1.6442	0.1919	0.0654	-1.9553
94	-0.8838	-2.8033	0.2231	0.9514	-0.9397
95	1.9863	0.5926	1.4056	0.8129	-0.9355
96	1.0020	0.8189	0.4035	1.3862	-0.9341
97	0.1033	-0.7206	0.3846	-1.1352	0.0839
98	-1.0252	-0.1284	0.7266	1.2015	-0.9369
99	2.0334	0.1627	-0.9522	-0.5536	1.1074
100	-0.1477	-0.0439	-0.1645	-0.1553	-0.1281

Figure A3.1 Physical distribution plots of conserved hypothetical proteins within *Bacillus anthracis* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

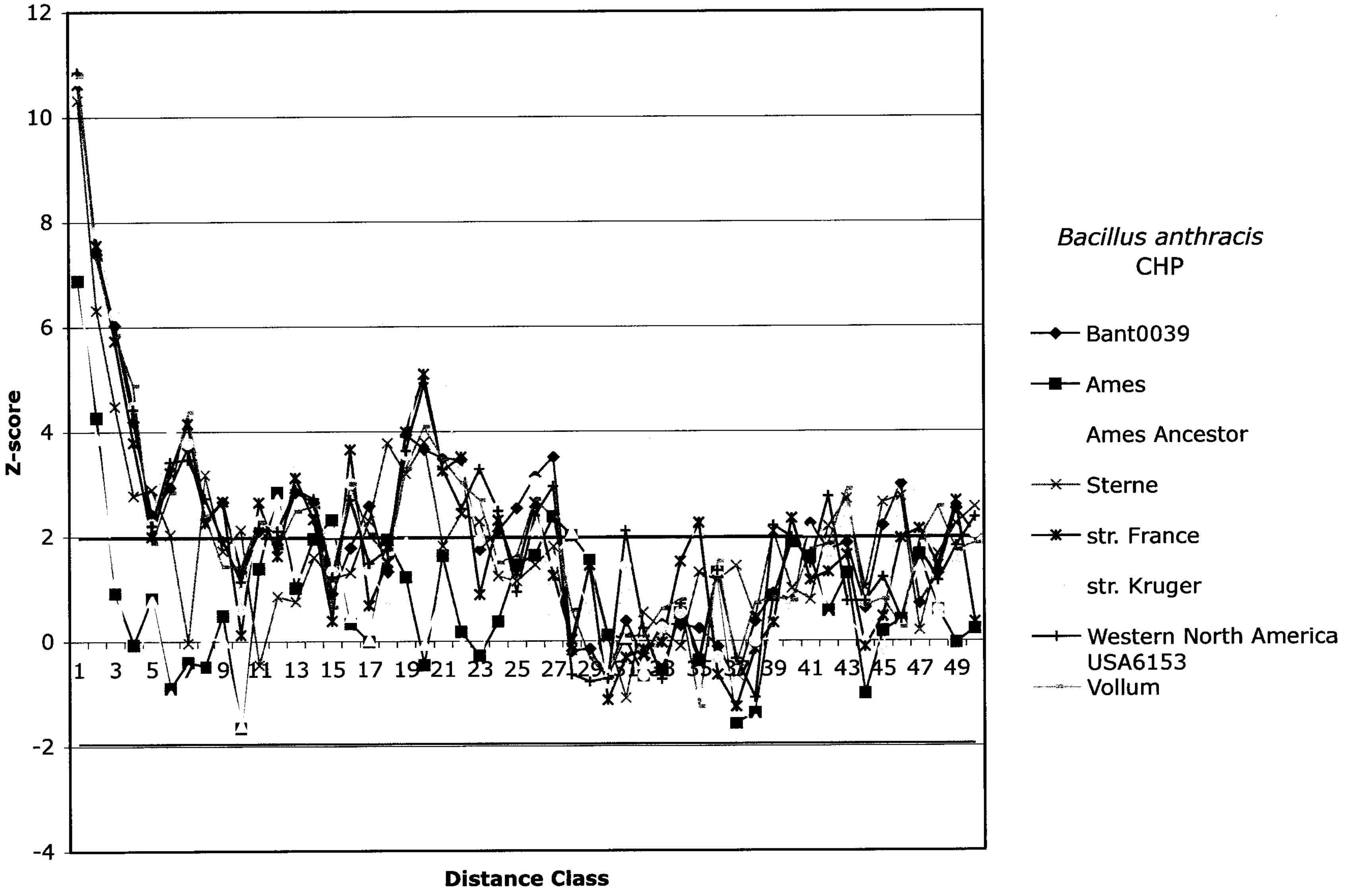

12
10
8
6
4
2
0
-2
-4
Z-score
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Distance Class
Bacillus anthracis
CHP
Bant0039
Ames
Ames Ancestor
Sterne
str. France
str. Kruger
Western North America USA6153
Vollum

Figure A3.2 Physical distribution plots of conserved hypothetical proteins within *Campylobacter jejuni* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

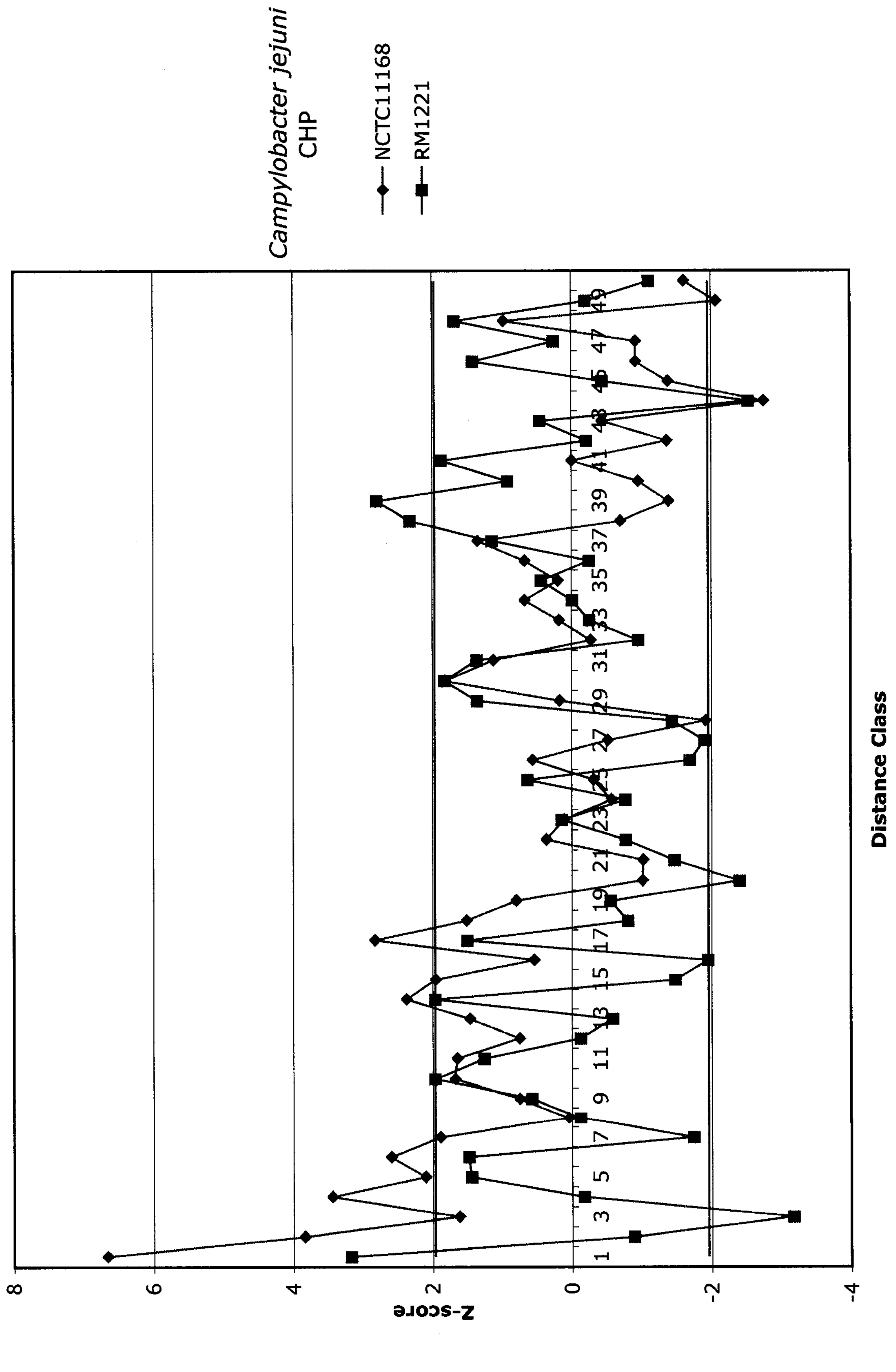

Campylobacter jejuni
CHP
NCTC11168
RM1221
Z-score
8
6
4
2
0
-2
-4
Distance Class
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49

Figure A3.3 Physical distribution plots of conserved hypothetical proteins within *Chlamydia pneumoniae* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

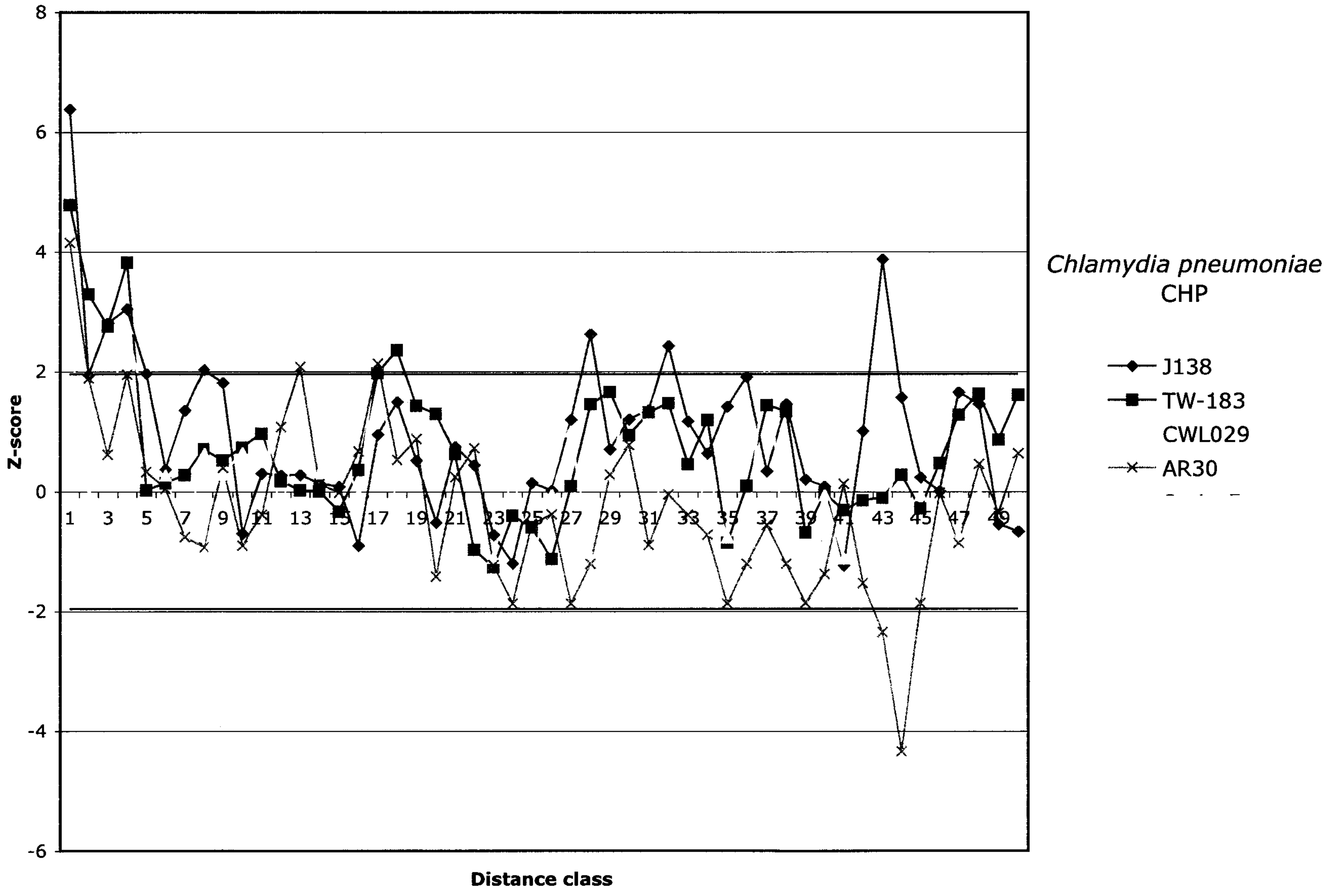

8
6
4
2
0
-2
-4
-6
Z-score
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance class
Chlamydia pneumoniae
CHP
J138
TW-183
CWL029
AR30

Figure A3.4 Physical distribution plots of conserved hypothetical proteins within *Escherichia coli* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

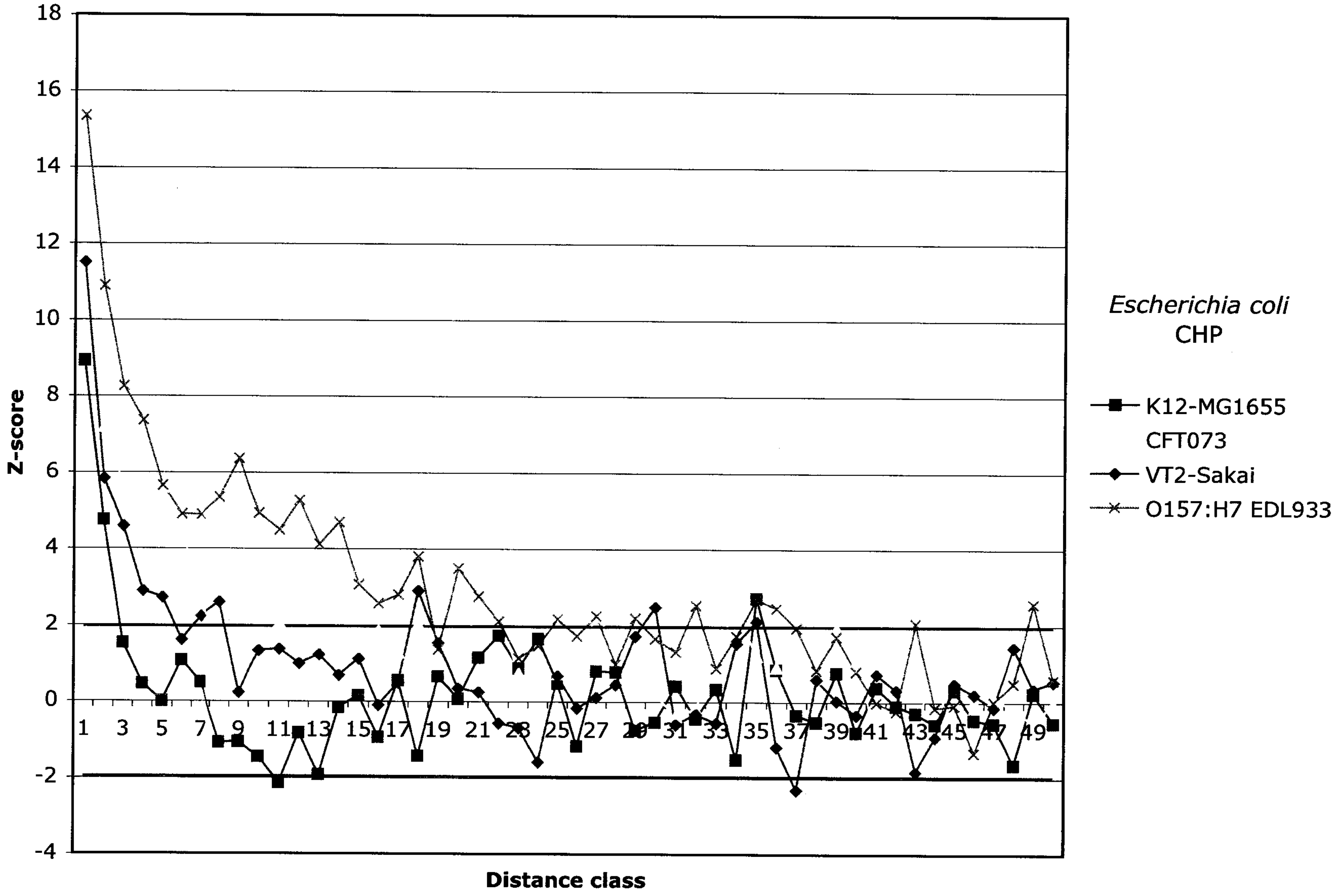

18
16
14
12
10
8
6
4
2
0
-2
-4
Z-score
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance class
Escherichia coli
CHP
K12-MG1655
CFT073
VT2-Sakai
O157:H7 EDL933

Figure A3.5 Physical distribution plots of conserved hypothetical proteins within *Legionella pneumophila* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

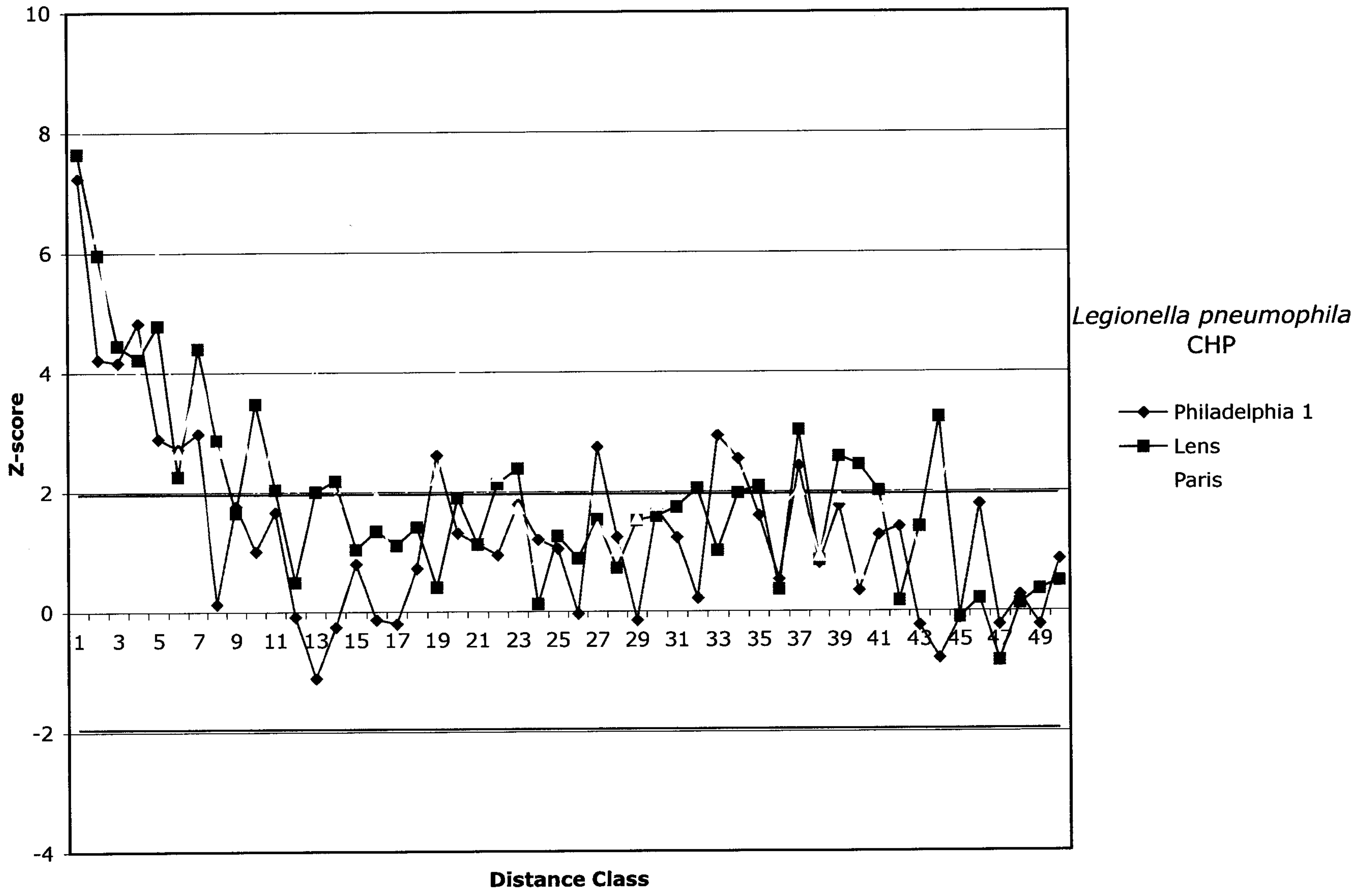
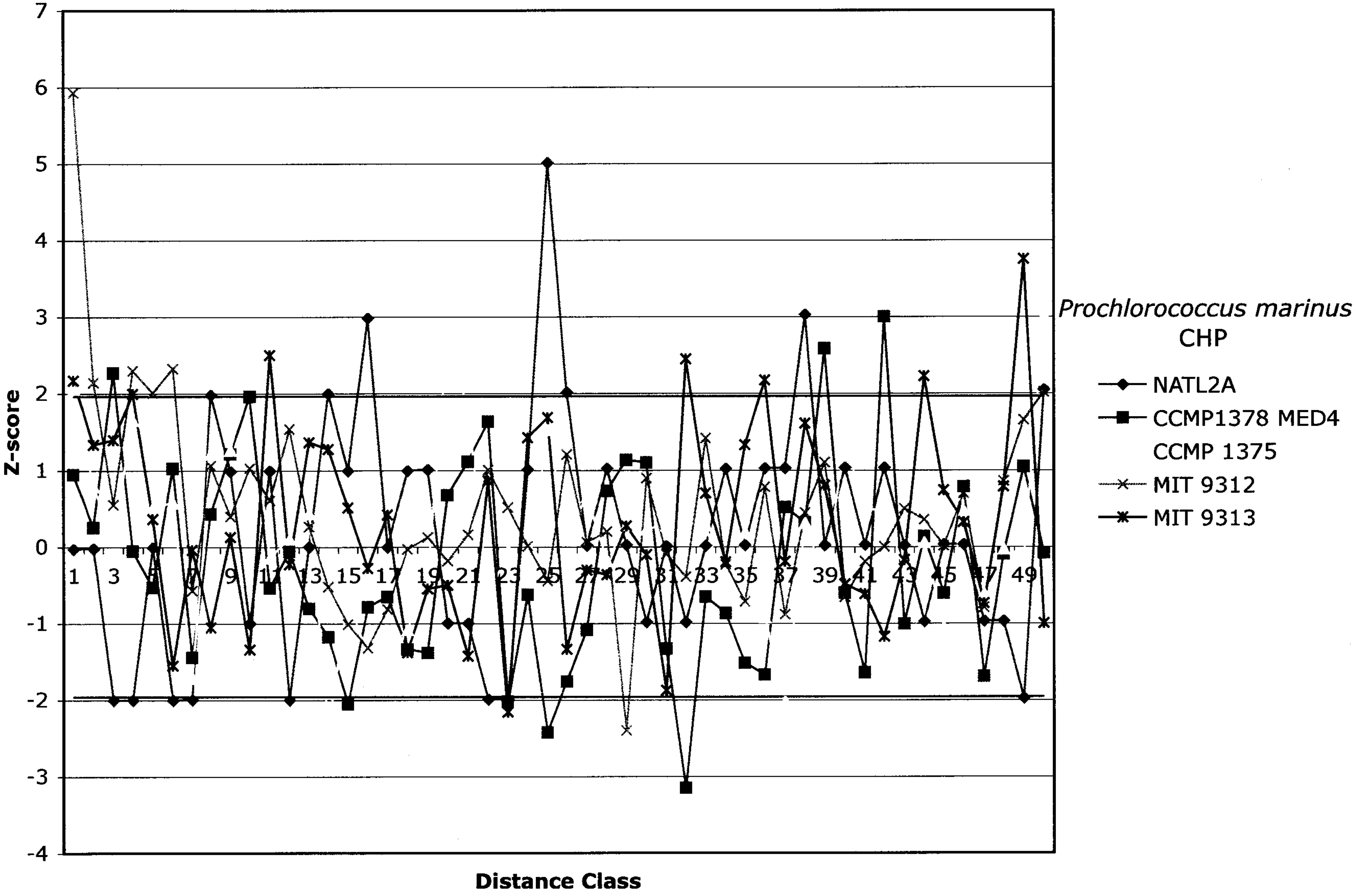

Legionella pneumophila
CHP
Philadelphia 1
Lens
Paris
Z-score
10
8
6
4
2
0
-2
-4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance Class

Figure A3.6 Physical distribution plots of conserved hypothetical proteins within *Prochlorococcus marinus* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.


Z-score
7
6
5
4
3
2
1
0
-1
-2
-3
-4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance Class
Prochlorococcus marinus
CHP
NATL2A
CCMP1378 MED4
CCMP 1375
MIT 9312
MIT 9313

Appendix 4. Joint Counts and Distribution of True Hypothetical Proteins

Table A4.1 True hypothetical protein joint counts, generated by Genespat v.4, for *Bacillus anthracis* strains.

	B. anthracis strain							
Distance class	A0039	Ames	Ames Ancestor	Sterne	str. France	str. Kruger	Vollum	Western North America USA5163
1	8.8872	12.6294	12.9867	-0.0199	9.0491	8.8666	9.6193	9.1527
2	7.944	7.5315	7.5406	-0.0198	8.0078	8.9184	7.5668	7.1713
3	4.9498	6.8376	7.0956	-0.0198	5.8106	5.9261	6.5032	7.1623
4	1.9773	5.8411	5.6082	-0.0198	1.4534	1.9668	1.2593	2.0982
5	0.9773	4.8346	4.8479	-0.0198	2.5403	2.9543	2.3054	1.078
6	0.9777	4.4635	4.5981	-0.0198	1.4481	2.9549	1.2558	2.0915
7	2.9743	1.529	1.7992	-0.0198	0.3608	1.9668	1.2606	2.0977
8	0.9781	2.1482	2.4167	-0.0198	3.6321	0.9685	1.2554	0.0634
9	-0.0116	2.0183	1.5463	-0.0198	0.3615	0.9753	0.2101	2.1002
10	-0.0123	2.2719	2.9074	-0.0198	-0.7317	-1.0138	0.2088	-1.9637
11	0.9824	4.1263	3.892	-0.0198	2.5451	0.9736	1.2597	-0.9487
12	-1.0081	4.142	3.784	-0.0198	0.3622	-0.0173	0.2098	3.1166
13	-0.0106	1.7677	1.2936	-0.0198	-0.7309	-1.0126	-0.8422	-1.963
14	-1.0078	3.5288	3.9139	-0.0198	-0.7312	0.9745	0.2101	0.0694
15	-0.0116	2.5401	2.437	-0.0198	-0.7317	-2.0067	-0.8427	0.0684
16	-2.003	1.7908	1.1984	-0.0197	-1.8239	-1.0126	-1.8952	-1.9636
17	-2.0027	1.5545	1.8191	-0.0197	-1.8236	-1.0116	-1.895	-0.9475
18	2.9712	3.7604	4.154	-0.0198	-0.7334	-1.0168	0.2091	0.0681
19	-0.0133	1.1627	1.0607	-0.0198	-1.8244	0.9774	0.2057	2.0965
20	0.9878	2.549	2.3163	-0.0197	-0.7287	-1.0105	-0.8397	0.0721
21	-0.0086	1.4504	1.9747	-0.0197	-0.7286	-2.0057	0.2122	2.1069
22	0.9865	0.8133	0.8415	-0.0197	0.3656	-0.0159	-0.8415	0.0677
23	-2.0014	2.7937	2.4396	-0.0197	0.3666	-1.0093	-1.8937	-1.9614
24	-2.0041	2.3126	2.3308	-0.0197	-0.7323	-2.0077	-0.8436	-1.9648
25	-0.0073	3.1866	3.4495	-0.0197	-0.7297	-1.0088	0.2118	0.0731
26	-0.0079	3.9273	3.4493	-0.0197	0.3663	-2.0056	-0.8395	-0.945
27	-1.0036	1.3401	1.6203	-0.0197	-0.7272	-0.0139	-1.893	-1.9608
28	0.9878	-0.1463	0.0059	-0.0197	0.3684	0.983	0.2125	1.0894
29	-0.0056	0.4715	0.7345	-0.0197	1.4643	-0.0142	1.271	1.0937
30	-0.0053	0.8326	1.232	-0.0197	0.3677	-0.0125	-0.8385	0.0735
31	-1.9996	-1.1469	-1.3605	-0.0197	0.3694	-1.0071	-1.8928	-1.9599
32	0.9904	0.7363	1.2603	-0.0197	1.4629	1.9737	3.3758	3.1237
33	-0.0103	1.8684	1.6405	-0.0197	1.4571	0.9808	-0.8397	-0.9447
34	0.9933	3.2317	3.3682	-0.0197	1.4625	-1.0073	1.2683	-0.9417
35	-1.0001	2.1107	2.0001	-0.0197	-0.725	-0.0095	0.2193	0.0765
36	0.9955	2.7077	2.4836	-0.0197	1.4674	1.9841	2.3263	2.114
37	1.9898	0.9761	1.2489	-0.0197	1.4682	-0.0129	0.2135	2.1108
38	0.992	-0.7485	-1.7138	-0.0197	-0.7274	-0.0132	-0.8382	-1.9614

	B. anthracis strain							
Distance class	A0039	Ames	Ames Ancestor	Sterne	str. France	str. Kruger	Vollum	Western North America USA5163
39	-1.9995	0.4851	0.2697	-0.0197	-0.7284	-1.0155	-1.8975	-0.9412
40	-0.0093	1.2555	1.5308	-0.0197	-1.8206	-1.0184	1.2486	2.0854
41	-1.0057	-0.368	-0.2148	-0.0197	-0.7372	-1.0089	-1.8929	-1.9661
42	-1	1.651	1.4181	-0.0197	-1.8282	-1.0067	-1.892	-0.9393
43	-0.0187	2.7707	2.7759	-0.0197	2.5616	-0.0081	-0.8387	-0.9444
44	0.9816	2.365	3.1294	-0.0197	-0.7242	0.9867	1.274	1.0987
45	-0.9978	3.495	3.1505	-0.0197	-1.8185	0.9902	0.2221	-0.9375
46	-0.9981	3.5257	2.9147	-0.0197	1.4727	2.9866	-0.834	0.0802
47	0.0012	2.7674	3.0326	-0.0196	-0.7217	-0.0054	-0.833	0.0826
48	-0.0029	2.7822	3.795	-0.0197	2.5577	-1.0037	-0.8423	-0.9417
49	-0.0012	2.0415	0.9474	-0.0197	-0.7262	1.9788	0.2197	-0.9436
50	1.9975	2.16	1.813	-0.0197	0.3742	0.9863	2.3358	3.1401
51	0.0001	1.2948	2.0638	-0.0197	0.3756	2.9907	-1.8894	-0.9367
52	-1.9991	2.4026	1.4321	-0.0197	-0.7212	-0.0044	-0.8347	-0.9399
53	-0.9946	-0.4551	0.6825	-0.0197	-0.7195	-0.0027	-1.8894	-1.9554
54	0.0056	2.0266	1.2937	-0.0197	-0.7198	-0.0017	0.2258	-0.935
55	0.0056	0.5317	0.6964	-0.0196	-0.719	0.997	3.395	-1.9548
56	-0.9953	0.5592	0.9584	-0.0197	0.3787	-0.0017	0.2248	0.0846
57	-0.9943	1.9342	1.5761	-0.0197	0.3797	-0.001	-0.8305	0.0863
58	-0.994	2.1861	2.9534	-0.0196	0.3797	-0.9984	-0.8302	0.0853
59	1.0078	0.4429	0.3516	-0.0196	-0.7182	-0.9972	-0.8307	0.0887
60	-0.9935	2.0703	1.3479	-0.0196	-1.8157	0.0017	0.2279	-0.9345
61	-1.9937	2.2036	2.3501	-0.0196	1.4775	-0.9992	0.2275	-1.9549
62	1.0074	0.7225	1.3735	-0.0196	-0.7168	0.0007	0.2255	1.1107
63	1.0108	1.4817	1.4948	-0.0196	-1.8171	0.0017	0.2299	-0.9323
64	0.009	2.4584	2.1014	-0.0196	-0.7168	-1.0011	-1.8867	-1.9532
65	-1.9947	0.7169	0.3681	-0.0196	-0.716	-0.9974	-0.8314	1.1016
66	0.0093	1.1077	1.2548	-0.0196	0.3828	-0.9964	-0.828	-0.933
67	-1.9928	1.7022	1.4805	-0.0196	-0.718	0.9991	-0.8324	-1.9531
68	0.0096	0.8415	1.7376	-0.0196	2.5827	0.0027	1.2866	1.1064
69	1.0103	1.1132	1.2604	-0.0196	1.486	1.0017	1.2896	0.0893
70	2.0114	0.9744	1.3732	-0.0196	-1.8149	-0.9954	-0.829	1.1128
71	0.0113	2.1108	1.2519	-0.0196	-0.7148	0.0038	1.2887	-0.93
72	3.0136	0.4951	0.531	-0.0196	0.3863	1.0012	1.2878	0.0907
73	0.0117	1.8704	2.3884	-0.0196	0.3842	0.0031	0.231	0.0887
74	0.0127	1.8875	1.5321	-0.0196	0.3811	1.0034	-0.8262	0.0917
75	0.0117	0.6337	0.7714	-0.0197	0.3856	2.0062	-0.8267	0.0893
76	-0.9922	2.378	2.2758	-0.0196	2.5803	1.0021	2.3494	1.1154
77	2.0197	1.9889	1.8906	-0.0196	-0.7135	-0.9995	-1.8866	3.1619
78	-1.9904	2.6303	2.154	-0.0196	-1.8135	-1.9931	1.2913	-0.9319
79	-0.9879	1.5302	1.4132	-0.0196	0.3891	-1.993	-1.8834	-0.9277
80	-0.9864	1.5244	0.9266	-0.0196	-0.7149	2.0018	0.2299	0.0938

	B. anthracis strain							
Distance class	A0039	Ames	Ames Ancestor	Sterne	str. France	str. Kruger	Vollum	Western North America USA5163
81	-1.9916	2.898	3.4169	-0.0196	-0.7117	-1.9943	-0.8247	-1.954
82	1.018	1.6608	1.3115	-0.0196	0.3839	-0.9931	-0.8264	1.1171
83	-1.989	-0.96	-0.8087	-0.0196	-0.7174	-0.9926	-0.8252	-0.927
84	0.0103	0.5143	0.7908	-0.0196	-0.714	-1.9952	-1.8832	0.0975
85	-0.9884	1.6549	1.184	-0.0196	0.3901	0.0079	0.2334	0.0965
86	0.0164	1.6723	2.4385	-0.0196	1.493	0.0072	-0.8222	0.0927
87	1.0215	2.6721	2.1952	-0.0196	-0.7122	-0.9901	-0.8244	0.0965
88	0.0209	0.1944	0.3397	-0.0196	-0.713	-0.9928	1.297	0.0972
89	0.0168	2.0441	2.8211	-0.0196	-1.8119	-0.9913	-0.8247	0.0979
90	1.021	3.3029	2.0645	-0.0196	-0.7077	-0.9913	0.2389	-0.9249
91	0.0175	-0.1916	-0.2863	-0.0196	1.4925	-0.9916	1.2961	0.0975
92	1.0245	1.5501	1.7025	-0.0196	0.3901	0.0126	0.2337	1.1209
93	3.0342	0.1698	0.3259	-0.0196	2.5961	0.0126	3.4243	1.1244
94	4.0369	2.4515	2.0938	-0.0195	2.6011	0.0144	3.425	4.2019
95	0.0192	1.0755	1.7226	-0.0195	1.4948	4.0253	0.2399	2.1528
96	3.036	1.32	0.2232	-0.0196	0.3961	1.0145	2.3638	1.1266
97	1.0249	-0.0449	0.1013	-0.0196	3.705	2.0207	0.2396	-0.9227
98	0.0226	3.1959	2.9727	-0.0195	0.3961	0.0168	-0.8189	-1.9486
99	-0.9836	1.5786	2.0878	-0.0195	1.5002	0.0157	-0.8209	0.1023
100	4.8679	14.665	14.6737	154.5512	4.4889	4.8289	4.6633	4.8467

Table A4.2 True hypothetical protein joint counts, generated by Genespat v.4, for *Campylobacter jejuni* strains.

	C. jejuni strain	
Distance class	RM1221	NCTC 11168
1	-0.1048	18.6854
2	-0.1052	16.0599
3	-0.1049	15.6559
4	-0.1049	14.2045
5	-0.1048	15.5071
6	-0.1046	15.5804
7	-0.1047	14.5156
8	-0.1047	14.0946
9	-0.1046	12.3902
10	-0.1046	13.9043
11	-0.1047	12.8024
12	-0.1046	12.1573
13	-0.1045	10.6592
14	-0.1045	11.7512
15	-0.1043	10.2856
16	-0.1045	10.269
17	-0.1046	9.1707
18	-0.1042	7.8793
19	-0.1044	9.6283
20	-0.1042	9.2046
21	-0.1043	10.2895
22	-0.1042	7.2882
23	-0.1044	5.975
24	-0.1043	5.3615
25	-0.1041	6.4588
26	-0.1044	6.4387
27	-0.104	6.6869
28	-0.1041	6.694
29	-0.104	6.2749
30	-0.1038	3.4712
31	-0.1039	4.3349
32	-0.1039	4.5567
33	-0.1039	4.7842
34	-0.1038	4.5805
35	-0.1039	3.9133
36	-0.1038	5.0006
37	-0.1039	4.5672
38	-0.1037	3.2863
39	-0.1036	3.2973
40	-0.1038	3.7075

	C. jejuni strain	
Distance class	RM1221	NCTC 11168
41	-0.1036	3.7355
42	-0.1035	3.0966
43	-0.1035	1.9769
44	-0.1034	0.9393
45	-0.1036	0.2841
46	-0.1036	-0.1495
47	-0.1036	1.1508
48	-0.1034	1.837
49	-0.1036	0.7317
50	-0.1036	0.3211
51	-0.1034	-0.9824
52	-0.1033	0.335
53	-0.1034	-0.3086
54	-0.1032	-0.5348
55	-0.1032	-1.3883
56	-0.1031	0.145
57	-0.1032	-0.7176
58	-0.103	-0.4998
59	-0.1032	-0.2907
60	-0.103	-1.1452
61	-0.103	-1.5931
62	-0.1032	-2.4568
63	-0.1032	-2.4568
64	-0.103	-0.9268
65	-0.1029	-1.5729
66	-0.103	-0.4908
67	-0.1029	-0.2727
68	-0.1029	-1.1154
69	-0.1028	-0.4554
70	-0.1028	-0.8925
71	-0.1029	-0.4598
72	-0.1027	-1.9659
73	-0.103	-0.4289
74	-0.1027	-1.094
75	-0.1028	0.0182
76	-0.1027	-0.4376
77	-0.1026	0.0045
78	-0.1025	0.2233
79	-0.1028	0.2326
80	-0.1026	1.3374
81	-0.1026	0.461
82	-0.1025	-0.6432

	C. jejuni strain	
Distance class	RM1221	NCTC 11168
83	-0.1025	-0.1827
84	-0.1024	-0.4153
85	-0.1025	-1.2834
86	-0.1023	-1.4903
87	-0.1025	-1.6976
88	-0.1023	-0.3841
89	-0.1024	-1.2581
90	-0.1024	-0.8102
91	-0.1022	-0.7972
92	-0.102	-0.5859
93	-0.1021	0.3072
94	-0.102	-0.8142
95	-0.1023	-1.2453
96	-0.1022	-1.4525
97	-0.1019	-1.4649
98	-0.1021	-0.9913
99	-0.1019	-1.8525
100	-0.3646	-0.8452

Table A4.3 True hypothetical protein joint counts, generated by Genespat v.4, for *Chlamydia pneumoniae* strains.

	C. pneumoniae strain			
Distance class	AR39	CWL029	J138	TW-183
1	-5.0601	8.9434	0	-0.1344
2	3.8327	3.4996	0	-0.1335
3	-0.6031	1.4161	0	-0.1327
4	1.2636	0.7767	0	-0.1325
5	0.1548	0.1365	0	-0.1318
6	0.9153	-0.1095	0	-0.1327
7	2.0384	0.4349	0	-0.1327
8	2.0384	-0.3388	0	-0.1327
9	1.6721	-0.4537	0	-0.1322
10	0.9327	-0.0405	0	15.0041
11	-1.6683	0.6375	0	-0.1322
12	1.3113	0.6374	0	-0.1325
13	2.4369	0.3908	0	-0.1325
14	0.95	0.0116	0	-0.1319
15	-0.5336	0.5408	0	-0.1321
16	2.4705	1.5991	0	-0.1323
17	1.7153	2.2606	0	-0.1322
18	-0.5136	1.3345	0	-0.1311
19	0.9909	0.6551	0	-0.1307
20	-0.8786	0.5408	0	-0.1306
21	1.378	1.0698	0	-0.1302
22	1.7468	-0.6499	0	-0.13
23	2.135	-1.576	0	-0.1308
24	0.2685	0.523	0	-0.1309
25	-0.8593	-0.6836	0	-0.1304
26	-0.4834	0.109	0	-0.1308
27	0.6445	0.7693	0	-0.1314
28	0.6445	0.8835	0	-0.1305
29	1.3964	0.3555	0	-0.1316
30	2.5244	0.5053	0	-0.1301
31	1.3902	1.298	0	-0.1311
32	1.7724	0.3732	0	-0.1304
33	-0.4884	0.2411	0	-0.1311
34	-0.4884	0.3731	0	-0.131
35	1.7724	0.1089	0	-0.131
36	-0.8593	1.0336	0	-0.1306
37	0.2631	0.241	0	-0.1299
38	1.7724	0.6373	0	-0.1303
39	-1.2309	-0.2873	0	-0.13
40	4.0285	-1.4762	0	-0.1298

	C. pneumoniae strain			
Distance class	AR39	CWL029	J138	TW-183
41	1.779	-0.9643	0	-0.1304
42	0.6503	-1.2286	0	-0.1311
43	-0.0969	-0.5517	0	-0.1301
44	0.656	1.166	0	-0.129
45	1.4027	-0.2874	0	-0.1295
46	1.7853	-1.4766	0	-0.1295
47	0.656	-0.6837	0	-0.1306
48	2.5314	1.0336	0	-0.129
49	-1.6027	-0.0231	0	-0.1298
50	2.1616	-0.6836	0	-0.1297
51	0.6559	0.5052	0	-0.1294
52	0.285	0.5052	0	-0.1289
53	0.6616	1.562	0	-0.1295
54	1.4147	0.7514	0	-0.1287
55	-0.0969	0.7693	0	-0.1287
56	0.6616	-1.1956	0	-0.1293
57	1.4146	-0.9312	0	-0.129
58	-0.0916	0.2586	0	-0.1298
59	0.2795	0.2586	0	-0.129
60	0.6673	0.2586	0	-0.1293
61	-0.0863	-0.138	0	-0.1294
62	0.6673	-0.6668	0	-0.1287
63	2.551	-1.0634	0	-0.1287
64	1.4208	0.6551	0	-0.1292
65	-0.4631	0.6552	0	-0.1291
66	-0.0863	-0.1552	0	-0.1276
67	0.285	-0.2873	0	-0.1292
68	-1.5977	-0.8157	0	-0.1291
69	1.7912	0.7514	0	-0.1297
70	-0.4681	0.206	0	-0.1286
71	0.285	0.7514	0	-0.128
72	-1.2213	2.3543	0	-0.1276
73	0.285	0.3555	0	-0.128
74	0.2905	0.4875	0	-0.1289
75	1.05	1.1657	0	-0.1282
76	1.427	-0.4194	0	-0.129
77	1.4331	-0.5515	0	-0.1277
78	-0.8302	-1.3441	0	-0.1275
79	1.4395	-0.0404	0	-0.1269
80	1.4332	0.109	0	-0.1267
81	0.6845	-0.1552	0	-0.1277
82	1.4332	0.3731	0	-0.1277

	C. pneumoniae strain			
Distance class	AR39	CWL029	J138	TW-183
83	1.4457	0.6552	0	-0.1275
84	-1.9578	1.5991	0	-0.1277
85	4.4594	1.7314	0	-0.1271
86	1.05	0.9375	0	-0.1267
87	2.942	1.0699	0	-0.1273
88	0.6788	2.3929	0	-0.1272
89	2.1878	0.0116	0	-0.1265
90	0.296	-0.7822	0	-0.1271
91	1.4333	-0.3853	0	-0.1276
92	1.05	-0.3853	0	-0.1267
93	0.673	-1.0468	0	-0.1259
94	-0.4529	-1.9887	0	-0.1264
95	2.1878	-1.4762	0	-0.1261
96	-1.5846	0.5052	0	-0.1266
97	1.8106	0.9015	0	-0.1274
98	0.673	0.3731	0	-0.1268
99	2.1746	1.0336	0	-0.1265
100	-0.5193	-0.1658	0	-0.6628

Table A4.4 True hypothetical protein joint counts, generated by Genespat v.4, for *Escherichia coli* strains.

	E. coli strain			
Distance class	CFT073	K12-MG1655	O157:H7 EDL933	O157:H7 VT2-Sakai
1	-0.2201	12.5661	-0.08	29.0088
2	-0.2191	6.7994	-0.0798	20.7335
3	-0.2185	3.2644	-0.0798	19.6265
4	-0.2185	3.1066	-0.0797	13.2436
5	-0.2187	1.519	-0.0796	14.7595
6	-0.2183	0.2306	-0.0798	12.8706
7	-0.219	1.5219	-0.0794	9.0827
8	-0.2185	0.2306	-0.0794	9.4639
9	-0.2181	-1.0497	-0.0795	10.231
10	-0.2184	0.7224	-0.0796	8.3369
11	-0.2183	-0.0804	-0.0794	6.4664
12	-0.2179	-0.7033	-0.0793	6.0885
13	-0.2178	0.5785	-0.0793	6.4467
14	-0.2179	1.0592	-0.0794	6.4517
15	-0.2181	2.0328	-0.0793	8.369
16	-0.218	1.3907	-0.0797	3.4454
17	-0.2178	2.3677	-0.0796	4.56
18	-0.2184	1.5487	-0.0793	7.5819
19	-0.2178	0.5957	-0.0793	5.7025
20	-0.2181	0.9214	-0.0793	7.2174
21	-0.2177	0.1145	-0.0793	4.5723
22	-0.2178	1.3994	-0.0793	3.4403
23	-0.2177	3.6745	-0.0794	1.9514
24	-0.218	4.6362	-0.0793	3.4428
25	-0.2177	1.4174	-0.0794	3.078
26	-0.2174	1.1093	-0.0795	2.699
27	-0.2179	1.7554	-0.0795	1.1886
28	-0.2177	-0.1971	-0.0792	1.5802
29	-0.2179	-0.983	-0.0794	1.1956
30	-0.2178	-1.4575	-0.0794	-0.6913
31	-0.2178	-2.4448	-0.0794	1.1956
32	-0.2178	-1.1315	-0.0793	1.1906
33	-0.2176	0.4744	-0.0794	2.3429
34	-0.2179	0.0035	-0.0792	1.1965
35	-0.2176	0.4859	-0.0794	2.3384
36	-0.2178	-0.4872	-0.0792	-0.3059
37	-0.2181	0.1596	-0.0792	1.2085
38	-0.2182	-1.6139	-0.0793	1.5956
39	-0.218	0.0092	-0.0793	1.5697
40	-0.2175	-0.1443	-0.0792	3.1096

	E. coli strain			
Distance class	CFT073	K12-MG1655	O157:H7 EDL933	O157:H7 VT2-Sakai
41	-0.2176	0.0176	-0.0791	1.9617
42	-0.2176	0.8297	-0.0794	1.9735
43	-0.218	-1.11	-0.0792	1.2113
44	-0.2173	1.1594	-0.0792	3.8613
45	-0.2175	0.1936	-0.0795	2.7293
46	-0.2176	2.3053	-0.0792	1.2125
47	-0.2172	0.6911	-0.0793	3.114
48	-0.2173	0.0485	-0.0794	1.5984
49	-0.2173	-0.2812	-0.0794	0.8459
50	-0.217	-2.2205	-0.0792	0.0889
51	-0.2173	-0.4377	-0.0792	0.8537
52	-0.217	-1.7239	-0.0791	1.614
53	-0.217	-1.0859	-0.0793	3.5096
54	-0.2171	-2.5443	-0.0792	2.3619
55	-0.2167	-2.5345	-0.0792	2.0012
56	-0.2171	0.222	-0.079	1.2322
57	-0.2168	-0.2672	-0.0791	1.6254
58	-0.2169	-0.7482	-0.0788	1.6191
59	-0.2169	-1.2268	-0.079	0.0854
60	-0.2173	-0.7454	-0.0791	2.0129
61	-0.2174	-0.2561	-0.0791	3.5171
62	-0.2169	-0.5669	-0.0792	-1.0321
63	-0.2174	0.4099	-0.0791	0.8564
64	-0.217	-0.0687	-0.0791	1.2422
65	-0.2167	-1.5251	-0.0789	0.0809
66	-0.2172	-0.2283	-0.0791	1.6325
67	-0.2169	-0.2227	-0.079	0.8632
68	-0.2171	-0.2339	-0.0789	0.4823
69	-0.2166	-1.6849	-0.0789	-1.025
70	-0.2166	-1.0297	-0.079	-0.6477
71	-0.2168	-0.8672	-0.079	-0.2768
72	-0.2169	-0.8617	-0.0791	-1.7998
73	-0.2168	-0.3685	-0.0791	-0.2716
74	-0.2168	0.7666	-0.0788	-1.4039
75	-0.217	-0.5284	-0.079	-1.028
76	-0.2166	-1.3361	-0.0788	-2.5443
77	-0.2169	-0.6773	-0.0788	-2.171
78	-0.2168	-1.1628	-0.079	-1.7774
79	-0.217	-0.0264	-0.0788	-1.4223
80	-0.2164	-0.3491	-0.0789	-2.1608
81	8.9983	-0.5063	-0.0791	-2.5482
82	-0.2166	-0.6664	-0.0789	-3.6888

	E. coli strain			
Distance class	CFT073	K12-MG1655	O157:H7 EDL933	O157:H7 VT2-Sakai
83	-0.2165	0.6419	-0.0787	-2.5416
84	-0.2166	1.1273	-0.079	-1.7794
85	-0.2165	-0.1724	-0.0787	-2.5392
86	-0.2163	2.2765	-0.0788	-3.309
87	-0.2163	1.3113	-0.0791	-2.541
88	-0.2166	0.8164	-0.0789	-3.2996
89	-0.2161	-0.1667	-0.0789	-1.7743
90	-0.2162	3.1006	-0.079	-1.7742
91	-0.2166	-0.6416	-0.0789	-2.9209
92	-0.2164	-0.7938	-0.0788	-2.5286
93	-0.2162	-1.9455	-0.0788	-2.5376
94	-0.216	-1.1199	-0.0788	-4.826
95	-0.2168	-1.1199	-0.0789	-3.6807
96	-0.2164	0.3535	-0.0787	-4.0645
97	-0.2169	1.323	-0.0789	-4.819
98	-0.2166	-0.6251	-0.0786	-2.916
99	-0.2164	-2.4242	-0.0787	-1.3967
100	5.9949	1.4698	22.1127	6.8994

Table A4.5 True hypothetical protein joint counts, generated by Genespat v.4, for *Legionella pneumophila* strains.

	L. pneumophila strain		
Distance class	Lens	Paris	Philadelphia 1
1	-0.0535	-0.0524	-0.0266
2	-0.0534	-0.0522	-0.0266
3	-0.0533	-0.052	-0.0265
4	-0.0533	-0.0521	-0.0265
5	-0.0533	-0.052	-0.0265
6	-0.0532	-0.052	-0.0264
7	-0.0532	-0.0521	-0.0265
8	-0.0533	-0.0521	-0.0265
9	-0.0533	-0.0521	-0.0265
10	-0.0532	-0.052	-0.0265
11	-0.0532	-0.052	-0.0264
12	-0.0532	-0.052	-0.0265
13	-0.0531	-0.0518	-0.0265
14	-0.0532	-0.052	-0.0264
15	-0.0532	-0.052	-0.0265
16	-0.0531	-0.0519	-0.0264
17	-0.0532	-0.0521	-0.0265
18	-0.0531	-0.0519	-0.0264
19	-0.0531	-0.0519	-0.0264
20	-0.0531	-0.052	-0.0264
21	-0.0531	-0.0519	-0.0264
22	-0.0531	-0.0518	-0.0264
23	-0.053	-0.0519	-0.0264
24	-0.053	-0.0519	-0.0264
25	-0.0532	-0.0519	-0.0264
26	-0.0531	-0.052	-0.0264
27	-0.0531	-0.0519	-0.0264
28	-0.0532	-0.0519	-0.0264
29	-0.053	-0.0518	-0.0264
30	-0.053	-0.0518	-0.0264
31	-0.053	-0.0518	-0.0264
32	-0.053	-0.0518	-0.0264
33	-0.0531	-0.0518	-0.0264
34	-0.053	-0.052	-0.0264
35	-0.053	-0.0519	-0.0264
36	-0.053	-0.0518	-0.0264
37	-0.0531	-0.0518	-0.0263
38	-0.053	-0.0519	-0.0264
39	-0.0529	-0.0518	-0.0264
40	-0.0529	-0.0518	-0.0263

	L. pneumophila strain		
Distance class	Lens	Paris	Philadelphia 1
41	-0.0529	-0.0517	-0.0264
42	-0.053	-0.0518	-0.0264
43	-0.053	-0.0518	-0.0263
44	-0.0531	-0.0518	-0.0264
45	-0.0529	-0.0518	-0.0263
46	-0.0529	-0.0518	-0.0263
47	-0.053	-0.0518	-0.0263
48	-0.0529	-0.0518	-0.0263
49	-0.0529	-0.0516	-0.0263
50	-0.0528	-0.0517	-0.0263
51	-0.0528	-0.0517	-0.0263
52	-0.0529	-0.0517	-0.0264
53	-0.053	-0.0518	-0.0263
54	-0.0528	-0.0517	-0.0263
55	-0.0528	-0.0517	-0.0263
56	-0.0527	-0.0516	-0.0263
57	-0.0528	-0.0516	-0.0263
58	-0.0527	-0.0516	-0.0262
59	-0.0528	-0.0515	-0.0263
60	-0.0528	-0.0517	-0.0263
61	-0.0529	-0.0516	-0.0263
62	-0.0527	-0.0517	-0.0262
63	-0.0528	-0.0516	-0.0263
64	-0.0527	-0.0515	-0.0263
65	-0.0527	-0.0516	-0.0262
66	-0.0527	-0.0515	-0.0263
67	-0.0527	-0.0515	-0.0262
68	-0.0528	-0.0515	-0.0263
69	-0.0528	-0.0516	-0.0262
70	-0.0527	-0.0516	-0.0262
71	-0.0527	-0.0516	-0.0262
72	-0.0527	-0.0515	-0.0262
73	-0.0527	-0.0515	-0.0262
74	-0.0528	-0.0515	-0.0262
75	-0.0526	-0.0514	-0.0262
76	-0.0526	-0.0514	-0.0262
77	-0.0526	-0.0515	-0.0262
78	-0.0526	-0.0515	-0.0262
79	-0.0526	-0.0515	-0.0262
80	-0.0526	-0.0514	-0.0262
81	-0.0526	-0.0514	-0.0262
82	-0.0527	-0.0515	-0.0262

	L. pneumophila strain		
Distance class	Lens	Paris	Philadelphia 1
83	-0.0527	-0.0514	-0.0262
84	-0.0526	-0.0514	-0.0261
85	-0.0526	-0.0514	-0.0261
86	-0.0525	-0.0514	-0.0262
87	-0.0525	-0.0514	-0.0261
88	-0.0526	-0.0514	-0.0262
89	-0.0526	-0.0513	-0.0262
90	-0.0525	-0.0513	-0.0262
91	-0.0526	-0.0514	-0.0262
92	-0.0525	-0.0514	-0.0262
93	-0.0525	-0.0515	-0.0262
94	-0.0525	-0.0514	-0.0262
95	-0.0525	-0.0513	-0.0261
96	-0.0526	-0.0514	-0.0261
97	-0.0525	-0.0513	-0.0261
98	-0.0525	-0.0513	-0.0262
99	-0.0524	-0.0513	-0.0261
100	-0.5278	-0.5301	-0.6945

Table A4.6 True hypothetical protein joint counts, generated by Genespat v.4, for *Prochlorococcus marinus* strains.

	P.marinus strain				
Distance class	CCMP 1375	CCMP 1378 MED4	MIT 9312	MIT9313	NATL2A
1	14.855	11.7233	21.8642	10.9043	14.2397
2	12.2528	6.9297	13.5721	10.1363	11.53
3	11.185	7.8233	10.2358	8.0287	9.465
4	9.387	7.1453	10.2123	5.7113	10.8641
5	10.3566	7.0932	11.0751	6.5455	5.3757
6	10.3576	5.4419	9.3873	5.2811	3.2998
7	8.745	5.9986	9.4062	6.9417	4.0112
8	9.2686	5.9526	6.8841	4.5578	2.6367
9	9.0869	4.9882	7.7177	5.567	8.1264
10	9.1249	3.857	7.7079	2.4865	3.328
11	9.2074	4.3856	5.2175	5.0369	8.1472
12	8.544	3.8134	4.3589	3.6565	6.7534
13	8.3246	4.3076	5.2098	4.3868	6.0945
14	8.1344	3.3879	5.2248	2.9187	8.8654
15	8.4812	4.0118	6.0646	3.3419	4.0148
16	8.0226	2.4789	7.7623	4.6123	5.3932
17	9.3667	3.291	7.7685	3.9297	6.0735
18	8.1608	1.8053	8.5913	4.2016	7.4707
19	7.0116	2.3583	6.9031	3.5949	8.1843
20	8.6489	2.0361	6.0697	2.614	6.7772
21	7.0124	0.7722	3.5455	3.697	5.4008
22	8.3393	0.8529	1.0231	4.0281	8.1877
23	8.1006	0.7368	2.7117	3.1347	6.7836
24	7.4446	2.2785	3.5516	3.7553	5.4315
25	7.9715	1.0074	3.5642	4.4508	6.1349
26	8.0386	0.9979	2.7229	2.6203	2.6689
27	8.3032	1.0078	1.8772	1.8715	4.0647
28	7.1068	2.5102	1.8704	2.6788	5.4557
29	7.4427	1.6266	5.2619	3.3894	4.7652
30	5.9916	1.8028	1.882	3.2399	5.4474
31	6.4039	2.019	3.5725	3.4141	5.4303
32	7.3248	2.1538	1.8888	2.3158	6.1369
33	7.4329	0.5813	1.0431	3.9094	7.5186
34	7.9945	2.0486	1.8888	2.4773	6.8413
35	7.0786	2.0633	4.4326	2.6898	4.7671
36	7.3641	1.7514	3.5916	2.9599	8.2533
37	6.4362	1.0557	1.8903	4.2979	8.9345
38	7.7951	1.2876	4.4368	3.4147	6.1657
39	7.0614	3.2194	1.8936	2.8879	7.5399
40	6.4807	2.722	1.0587	2.9335	4.7925

	P.marinus strain				
Distance class	CCMP 1375	CCMP 1378 MED4	MIT 9312	MIT9313	NATL2A
41	8.1308	0.3057	-0.6467	2.2735	6.1774
42	6.1721	-0.1692	1.0574	1.53	3.4003
43	7.7557	1.7365	2.7582	2.8131	4.771
44	8.4592	2.5995	3.6145	3.4803	3.4157
45	7.6153	2.7221	5.3127	2.7632	4.8024
46	6.4391	0.4768	1.9168	1.3561	2.0249
47	5.9822	1.0824	0.2151	5.0954	4.1163
48	4.7423	0.2843	-0.6358	3.2877	4.8238
49	5.3161	-1.7965	0.2102	4.4031	4.8235
50	5.4275	-1.3845	2.7674	4.2083	4.8236
51	5.6431	-0.7971	1.0687	3.0558	0.6442
52	6.1675	0.0362	1.073	3.1495	5.5338
53	4.6745	0.3319	1.9181	2.8261	4.1161
54	5.5374	0.9815	1.9301	4.4067	4.8151
55	4.6775	-0.06	2.7827	2.5828	2.0265
56	6.8037	-0.6149	0.2175	3.0421	4.1351
57	4.5463	-2.0581	0.2162	2.1686	3.4239
58	6.0132	0.0277	1.9366	4.025	3.4374
59	6.1214	-0.5106	1.0934	3.1036	4.1424
60	5.7112	-0.6966	1.9366	2.8236	1.3453
61	4.9238	-1.7183	1.0874	2.183	4.8358
62	5.4634	-2.1897	1.089	2.88	2.7475
63	5.3346	-2.5635	2.8055	2.8916	4.1374
64	5.7595	-1.5487	1.9517	1.6376	2.0527
65	4.6538	-2.0031	1.0977	2.6326	2.7495
66	4.1882	-2.1295	-1.4655	4.1274	4.1797
67	4.5749	-2.396	-1.4677	3.3969	4.1815
68	5.5667	-1.2638	0.2507	2.4607	3.4663
69	5.431	-1.3089	0.2519	3.8082	4.1669
70	5.0165	-2.3287	1.1124	2.4396	2.7683
71	3.9485	-2.5382	-0.5975	2.4872	2.0776
72	3.2967	-0.5908	-1.4592	2.4142	5.6131
73	5.057	-0.2707	-1.4575	4.0861	3.4893
74	5.4402	-1.1092	-1.4599	3.4495	4.8641
75	3.3395	-0.647	-0.5975	2.4736	3.4915
76	3.9678	-0.4114	-1.4598	3.7487	3.4846
77	2.2186	-0.3826	-1.4529	1.1548	2.7872
78	3.9533	-2.1018	-0.5944	3.2981	1.3796
79	5.2674	-1.1816	1.1108	1.5249	2.085
80	3.2502	-2.1098	0.2531	1.2084	1.3724
81	3.1816	-1.4658	-0.5914	-0.1098	1.3811
82	1.8201	-1.7507	-0.5893	0.8653	-0.7171

	P.marinus strain				
Distance class	CCMP 1375	CCMP 1378 MED4	MIT 9312	MIT9313	NATL2A
83	3.9133	-0.7179	0.2731	0.5333	2.1118
84	4.2165	-0.2359	-0.5873	1.1876	0.7011
85	4.0214	-1.8228	-2.3121	0.9364	1.3862
86	2.3805	-0.8955	-0.5893	0.5744	1.3809
87	2.1138	-1.8115	-2.3078	0.1808	1.4021
88	2.6918	-2.6974	-1.4465	1.8073	2.104
89	1.5659	-1.9026	-0.5831	0.3428	0.6917
90	2.8086	-0.919	-1.4448	-0.3473	2.8033
91	2.4443	-1.3993	-0.576	1.688	4.9267
92	2.2256	-2.0308	0.2818	0.4252	3.5206
93	2.048	-1.3385	0.2806	0.5556	0.0009
94	0.79	-1.7905	-2.3048	1.0725	1.4052
95	1.8814	-1.5897	-0.576	0.8337	1.4158
96	2.3862	-2.2266	-2.3035	2.0537	0.0077
97	2.7233	-0.6843	-2.3051	1.2529	1.4176
98	0.0418	-0.1402	-0.5769	-0.08	1.4122
99	0.8875	-0.0726	1.1506	0.1943	2.837
100	-1.3312	-0.2996	-1.0369	-0.623	-1.7458

Figure A4.1 Physical distribution plots of true hypothetical proteins within *Bacillus anthracis* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

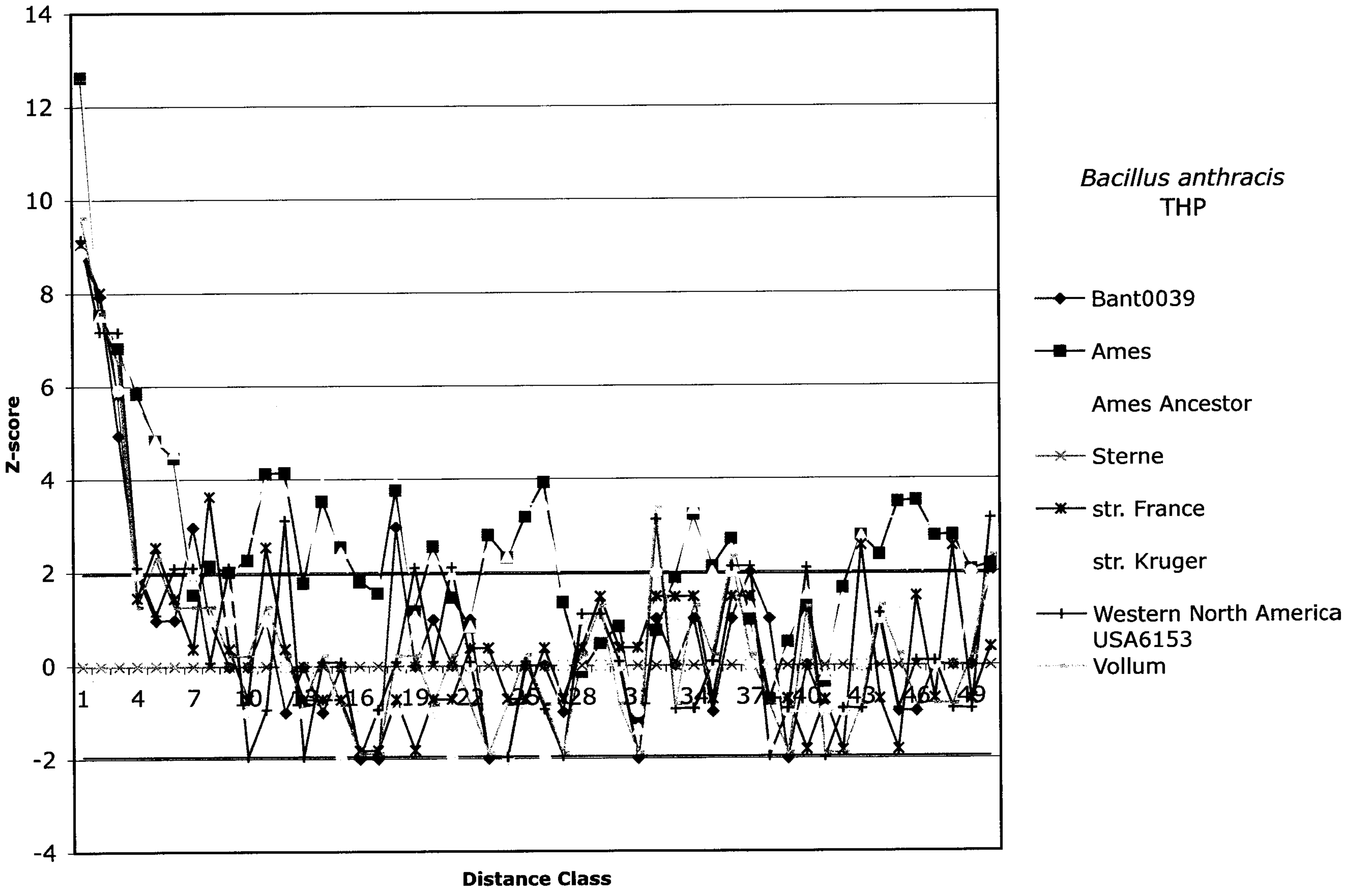

14
12
10
8
6
4
2
0
-2
-4
Z-score
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
Distance Class
Bacillus anthracis
THP
Bant0039
Ames
Ames Ancestor
Sterne
str. France
str. Kruger
Western North America
USA6153
Vollum

Figure A4.2 Physical distribution plots of true hypothetical proteins within *Campylobacter jejuni* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

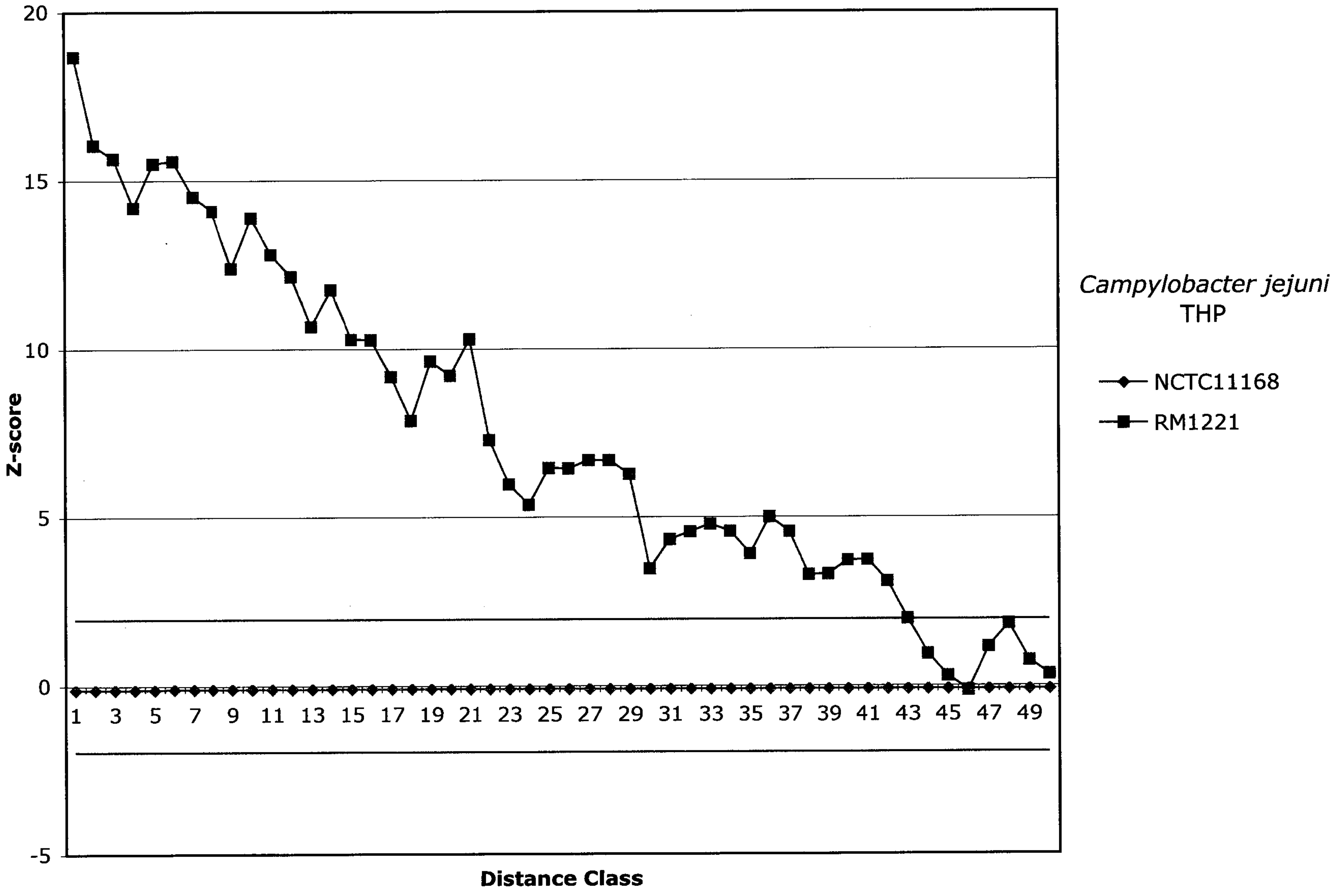

Campylobacter jejuni
THP
NCTC11168
RM1221
Z-score
20
15
10
5
0
-5
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Distance Class

Figure A4.3 Physical distribution plots of true hypothetical proteins within *Chlamydia pneumoniae* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

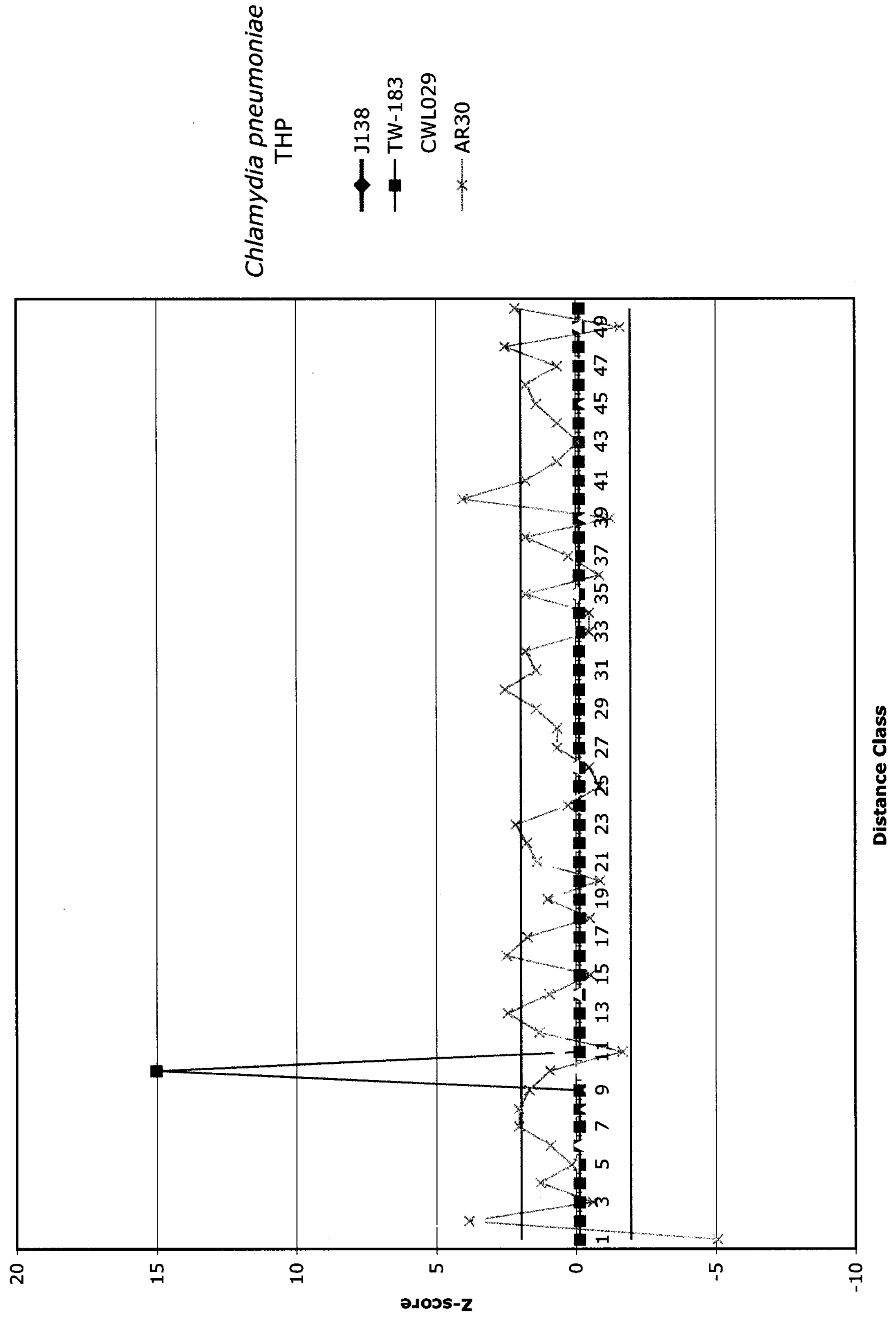

Chlamydia pneumoniae
THP
J138
TW-183
CWL029
AR30
Z-score
20
15
10
5
0
-5
-10
Distance Class
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49

Figure A4.4 Physical distribution plots of true hypothetical proteins within *Escherichia coli* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.

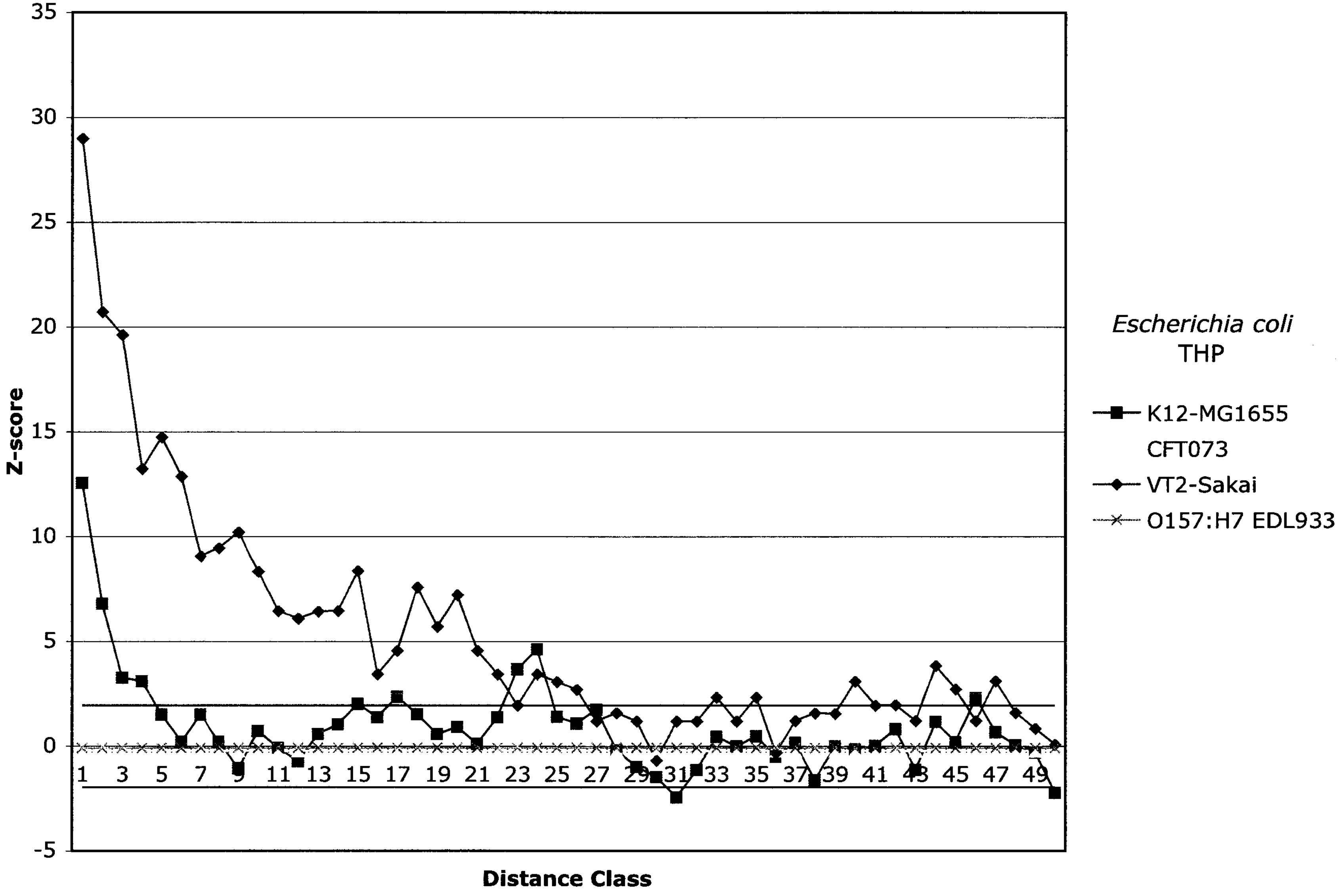
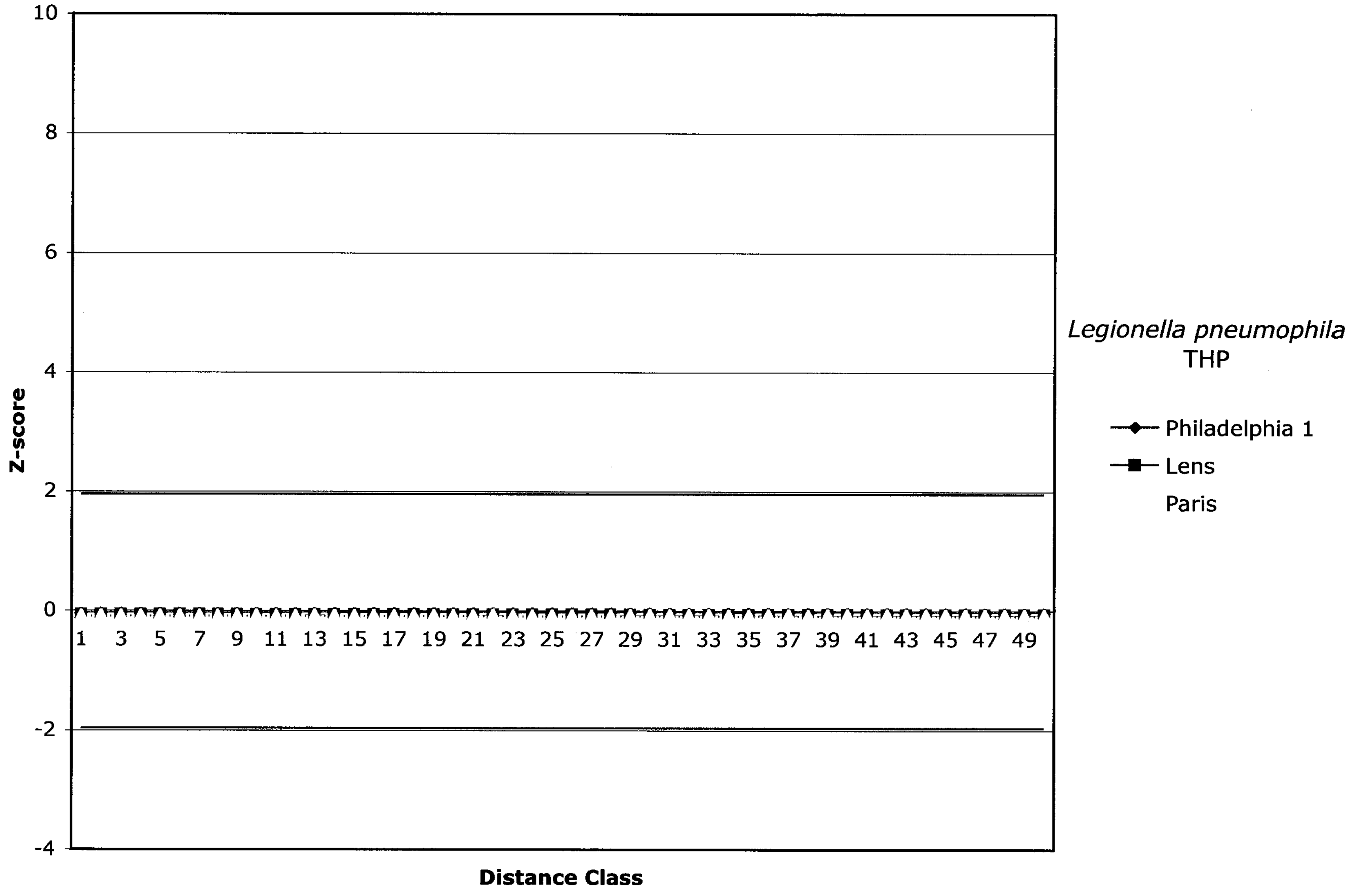
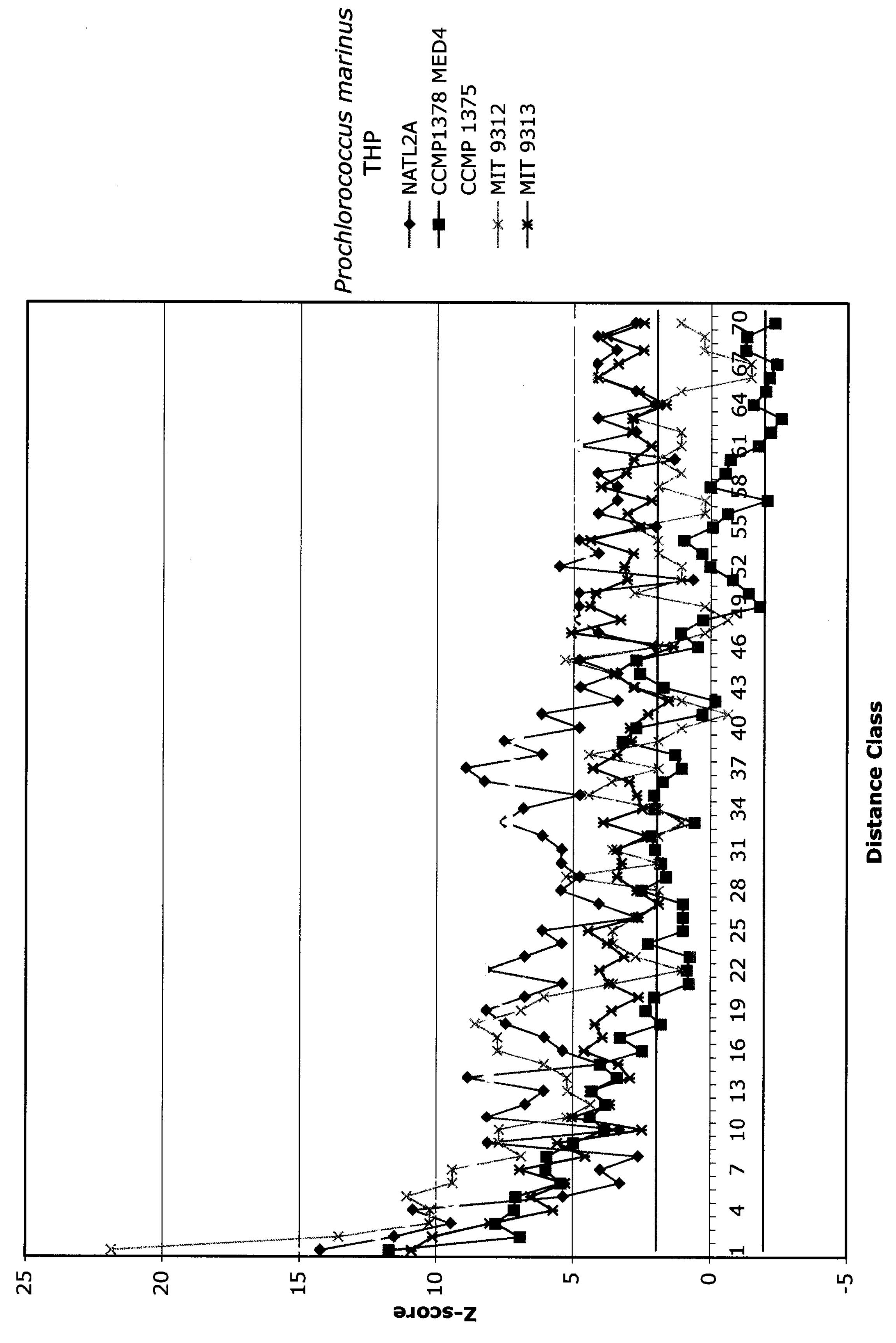

35
30
25
20
15
10
5
0
-5
Z-score
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
Distance Class
Escherichia coli
THP
K12-MG1655
CFT073
VT2-Sakai
O157:H7 EDL933

Figure A4.5 Physical distribution plots of true hypothetical proteins within *Legionella pneumophila* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.



Legionella pneumophila
THP
Philadelphia 1
Lens
Paris
Z-score
10
8
6
4
2
0
-2
-4
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Distance Class

Figure A4.6 Physical distribution plots of true hypothetical proteins within *Prochlorococcus marinus* strains. Significance cutoffs of >|1.96| are indicated by thin red lines.


Prochlorococcus marinus
THP
NATL2A
CCMP1378 MED4
CCMP 1375
MIT 9312
MIT 9313
Z-score
25
20
15
10
5
0
-5
Distance Class
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
67
70