This is to certify that the
dissertation entitled

GENOMIC INSIGHTS INTO ECOLOGICALLY IMPORTANT QUESTIONS FOR SOIL BACTERIA

presented by

Konstantinos T. Konstantinidis

has been accepted towards fulfillment
of the requirements for the

Ph. D degree in **Crop and Soil Sciences**

Major Professor's Signature

12/14/04

Date

GENOMIC INSIGHTS INTO ECOLOGICALLY IMPORTANT QUESTIONS FOR SOIL BACTERIA

By

Konstantinos T. Konstantinidis

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Crop and Soil Sciences

2004

ABSTRACT

GENOMIC INSIGHTS INTO ECOLOGICALLY IMPORTANT QUESTIONS FOR SOIL BACTERIA

By

Konstantinos T. Konstantinidis

Diverse prokaryotic species are the principal catalysts of the biogeochemical cycles that sustain life on Earth, however, fundamental issues that define that diversity are unresolved such as what are the extent and patterns of that diversity at the genome level, what is the interplay between genome evolution and ecological niche, and what defines a prokaryotic species in a manner that is predictive of phenotype. I compared 175 prokaryotic whole-genome sequences to address these questions. While prokaryotic genomes vary from 0.6 Mb to over 10 Mb and show enormous gene sequence diversity, several universal trends were noted that reflect the cellular and ecological strategies used by this simple but successful life form. For instance, large prokaryotic genomes, contrary to their eukaryotic counterparts, do not accumulate non-coding DNA or hypothetical genes and they are disproportionately enriched in regulation and secondary metabolism genes compared to medium and small-sized genomes. These trends suggest that larger genome-sized species may dominate in environments where resources are scarce but diverse and where there is little penalty for slow growth, such as soil. The genetic and functional diversity among highly related genomes was examined more closely to better understand the breadth and origin of biodiversity within a species and to use this insight to advance the current definition of species for Prokaryotes. Strains of the same species vary up to 30% in gene content while a large fraction, e.g. up to 65%, of these gene differences is frequently associated with bacteriophage and transposase elements,

indicating a much more important role of these elements during bacterial speciation than previously thought. Additional analysis suggests that a more stringent definition for species, which should also consider the ecology of the strain, is both more appropriate and plausible. Expansion of the approach to include the higher than species ranks of the prokaryotic taxonomy revealed that there are many irregularities in the current classification schema for the 175 genomes used in this study. Consequently, the predictive power of the higher ranks of taxonomy in terms of genetic relatedness among the grouped organisms is currently rather poor.

To provide for more extensive, robust, genome-based analysis of nature's vast microbial diversity, I explored a high-through-put approach of using microarrays for comparative genome hybridizations. Therefore, the performance of different microarray platforms was evaluated by comparing the expected (*in-silico*) microrray results to results from comparisons of whole-genome sequences (control). Oligo-arrays, i.e., one 50 base probe per gene, were found to perform comparably to whole-ORF arrays as long as the evaluated strains reside in the same or highly related species whereas whole-ORF arrays should perform better for more distantly related strains. In all cases tested, non-specific hybridization signal was found to be substantial and could lead to misleading results if not taken into account, but can also be used to indicate potential gene duplications.

My results have important implications for our understanding of the basis for and value of prokaryotic biodiversity and has broader impacts such as for reliable diagnosis of plant and animal disease agents, intellectual property rights, quarantine and (inter-) national regulations for transport and possession of microbes.

This dissertation is dedicated to the memory of my mother Sophia, who passed away in the middle of my studies but never stop inspiring my character, personality and hence, my work.

ACKNOWLEDGMENTS

I would like to thank several individuals for helping me during the course of my graduate studies. Most importantly, I thank my advisor, Professor James Tiedje, for providing me with the creative freedom and intellectual guidance that fostered my growth as a scientist. In addition, I thank him for exposing me in the academic world through writing of grants, participating in conferences, improving my communication skills and for giving me the chance to meet and learn from the experts in my field. Last, special thanks for being always supportive in my difficult, personal moments.

I wish to extent my thanks to the remaining members of my PhD guidance committee, Professors Terence Marsh, Syed Hashsham, and Michael Thomashow, as well as, Professor Tomas Schmidt of MSU, for their long-term commitment to my academic progress and contributions towards the development of critical thinking. I also feel indebted to Dr. John Heidelberg of The Institute for Genomic Research (TIGR) for stimulating my interest and training me in the area of comparative genomics, which is the subject of this dissertation.

I began this journey alone five years ago, a long way from my home, Thessaloniki, Greece. It was only the love, encouragement, and patience of my sister Maria, brother Yiannis, and many friends that converted this arduous journey into a wonderful experience. I would like to thank all of them for this, and particularly, Kiki, Demertis, Alban, Joel, Hector, and Elias.

Last, I want to acknowledge the funding I received from the Bouyoukos Fellowship Program, an endowment of the late Professor George Bouyoukos. Without this support, it would not have been possible for me to begin and complete this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

AN INTRODUCTION INTO PROKARYOTIC DIVERSITY AND GENOMICS

I have authored parts of this chapter in the book chapter: K. T. Konstantinidis, and J. M. Tiedje. Microbial diversity and genomics. *In* Microbial Functional Genomics. J. Zhou, D. K. Thompson, Y. Xu, and J. M. Tiedje (eds.) John Wiley & Sons. Hoboken, New Jersey, 2004, pg. 21-46.

INTRODUCTION

The prokaryotic biomass has been estimated to (at least) equal that of terrestrial and marine plants (70) while the prokaryotic biodiversity (genetic or biochemical) is presumably the largest reservoir of biodiversity on Earth based on recent estimates of 6,300 distinct prokaryotic species in a single gram of soil (15). Yet, several aspects of this biodiversity remain unexplored. For instance, the full extent of the total genetic diversity of Prokaryotes or even the genetic diversity within a single prokaryotic species remains unknown, and there is inadequate understanding of the interactions between ecological niche and biodiversity, or how important biodiversity is for (specific) ecosystem function and stability. Gaining information into such issues is at the heart of understanding the basis for and value of biodiversity, and for understanding the diverse environmental microbes that catalyze much of the biogeochemical cycles that sustain life on Earth. Such information can then be used to successfully apply microbes or control their *in-situ* activities for specific purposes such as bioremediation, plant protection or nitrogen fixation etc.

Genomics[1] offer a great opportunity to explore (at least) parts of the immense prokaryotic biodiversity, and in fact, they have already succeeded in significantly broadening our knowledge of it. For example, genomic comparisons have shown that prokaryotic genomes are much more "fluid" than previously thought e.g., mobile elements and lateral gene transfer events play a major role in the evolution and shaping of the genome space. This "fluidity" drives, by and large, the extensive genetic diversity revealed within the currently named species. Another major contribution of genomic

[1] Study that aims to reach a genome-level understanding of the molecular basis of the structure, functions, and evolution of biological systems using whole-genome sequence information and high-throughput technologies.

BACKGROUND

Prokaryotic biochemical and genetic diversity.

Prokaryotic life emerged about 3.7 billion years ago, or about 2 billion years before eukaryotic life arose (reviewed in (30). Thus, prokaryotic organisms had a long time to evolve and this accounts for the high biochemical and genetic diversity that characterizes prokaryotes. The extent of prokaryotic metabolic and enzymatic diversity is such that it is believed that a handful of prokaryotic species can live on almost any carbon source or redox couple available on Earth.

Prokaryotic organisms occupy two-thirds of the biodiversity on Earth, namely the Bacteria and the Archaea (71). What characterizes Bacteria compared to the other two domains of life is that its species that are closely related by molecular criteria (e.g., ribosomal RNA gene identities) can display strikingly different carbon and energy metabolisms. For instance, in the relatively closely related γ-Proteobacteria subgroup one can find very phenotypically different organisms such as the *E. coli* (organotroph), *Chromatium vinosum* (hydrogen sulfide-based phototroph), and the symbiont of *R. pachyptila*, the tubeworm (hydrogen sulfide-based symbiont). The situation is even more profound when specific biochemical traits, e.g. functional proteins, are considered. For instance, the ability to denitrify (making use of N-oxides as terminal electron acceptors) occurs sporadically among the cultivated bacterial species of coherent 16S rRNA clusters (73). The 16S rRNA sequence information is commonly used for the construction of phylogenetic trees to infer the ancestry and relatedness of organisms. As apparent from the denitrification example, even organisms that are identical or cluster tightly by the 16S rRNA criterion may not share most essential physiological similarities. Furthermore, the

functional genes involved in the denitrification pathway (e.g. nitrite reductase, nitrous oxide reductase) exhibit substantial sequence diversity in the cultivated representatives from a single gram of marine sediment or forest soil (73). The lack of general correspondence between metabolism and evolutionary relatedness is attributed to lateral gene transfer, large-scale symbiotic fusions (e.g., between a bacterium and a bacteriophage) and the great ability of bacteria to evolve to exploit available ecological niches.

The other domain of Prokaryotes, the Archaea, shows considerably less metabolic and genetic diversity compared to Bacteria based on the study of the representatives of the two domains that have been isolated to date. However, recent findings from culture-independent surveys for presence of archaeal-specific signatures in the environment suggest that archeal ecological significance and global distribution is much higher than represented in the currently cultured species (4, 8, 17, 18). For instance, several phyla- and order- level lineages and a new kingdom of Archaea, the Korarchaeota (4), have been proposed based on cloned 16S rRNA sequences from different environment sources. The lack of cultivable species representative of these lineages or information about the physiology these species severely limits our ability to summarize archaeal metabolic diversity.

As far as eukaryotic organisms are concerned, they also appear to be much less metabolically versatile than Bacteria in terms of range of substrates for growth and electron terminal acceptors they can utilize. For example, organotrophy, in which reduced organic compounds are used for energy and carbon, appears to be the main mode of nutrition for most non-photosynthetic Eukaryotes, and even in the case of organotrophy,

the number of different metabolic processes carried out by Bacteria far exceeds the ones carried out by Eukaryotes. Further, photosynthesis was also a bacterial innovation, and is ecologically and physiologically more diverse in the bacteria. Most bacterial photosynthesis is anaerobic and widely distributed among different bacterial phyla in contrast to a single kind of photosynthesis in Eukaryotes, i.e., the oxygenic photosynthesis of plants, and Archaeal, i.e., the photosynthesis of the *Halobacterium* genus. There is now conclusive genomic evidence that the eukaryotic photosynthetic machines, i.e. the chloroplasts of plants, originated from a symbiotic event between a eukaryotic cell and a cyanobacterium (37). Last, Eukaryotes that exploit other modes of nutrition such as lithotrophy, in which energy is derived from the oxidation of reduced inorganic compounds by a chemical oxidant, do so only in close association (symbiosis) with prokaryotic organisms.

Although there is relatively little information about the metabolic breadth of a major lineage of Eukaryotes, the amitochondriate Eukaryotes, the indisputable conclusion from reviewing the current knowledge on the metabolic and biochemical repertoire of prokaryotic species is: the versatility of Bacteria makes the metabolic machineries of Archaea and Eukarya seem comparatively monotonous.

How many prokaryotic species are there and what is a "species"?

While the previous discussion points out the immense metabolic diversity of the prokaryotic organisms, what makes prokaryotic species and consequently the metabolic processes they carry out important to Earth is their huge number of cells and their ubiquity. The most recent estimates suggest that the total number of prokaryotes on earth

to be 4-6 x 10^{30} cells and their cellular carbon to be 350-550 Pg (70). Hence, prokaryotic carbon is 60-100% of the estimated carbon in terrestrial and marine plants while prokaryotic biomass is presumably the largest pool of recyclable nitrogen (N) and phosphorus (P) since the (N+P)/C ratio is higher for the prokaryotic cell. Most of the earth's prokaryotes are found in the open ocean and in soil, where the total number of cells is in the order of 10^{29}-10^{30} (70), and particularly their subsurface, i.e., below 8 m for the terrestrial environment and below 10cm for oceanic sediments (26, 70), although there has been limited sampling of these environments and hence uncertainty in the accuracy of these predictions. The activity of prokaryotes is substantial in surficial marine and soil environments based on cell turnover times, which have been estimated at 6-25 days for the upper 200 m of ocean and 2.5 years for soil (19, 27, 70), whereas prokaryotic activity in the subsurface is orders of magnitude lower, e.g. turnover times of 1-2 x 10^3 years (16).

Although there is no doubt that prokaryotic cells are ubiquitous and far exceed any other type of life in numbers, the enumeration of prokaryotic species is far from being resolved. This is due to both fundamental problems regarding definition of species as well as practical limitations in counting prokaryotic species. The current species concept for prokaryotes (52, 58, 68) despite being pragmatic, operational and applicable (52, 58), remains controversial (9, 12). The controversy stems from the fact that prokaryotic species lack diagnostic morphological characteristics and are asexual organisms that exchange genetic material in their unique and unusual ways compared to eukaryotes. Therefore, none of the 22 species concepts described for Eukaryotes is applicable to Prokaryotes (52). In addition, it is not always feasible, due to technological

limitations and/or poor understanding of the metabolic and physiological properties of prokaryotic cells, to define unique phenotypic characteristics that are required for a species description (66). This has led most prokaryotic taxonomists to agree on a functional species definition for prokaryotes that is rooted in the degree of DNA/DNA reassociation. In this definition, two strains belong to the same species when their purified DNA molecules show at least 70% hybridization (59, 68).

This definition does not translate well to Eukaryotes, however. Application of the same definition to Eukaryotes would lead to the inclusion of members of many taxonomic tribes in the same species (55). For example, all the primates (i.e. humans, orangutans and gibbons) would then belong to the same species (56). Furthermore, gorillas and orangutans would not be considered threatened because they would be the same species as humans, which are numerous and cosmopolitan. Thus, a simple comparison of the number of eukaryotic and prokaryotic species greatly underestimates prokaryotic diversity. Indeed, the prokaryotic species concept is probably comparable to that of animal family or perhaps even an animal order

The other obstacle in enumerating prokaryotic species is the fact that only a small fraction of many microbial communities, typically about 1%, is cultivable. The problem of "cultivating the uncultivable" has been extensively discussed and reviewed elsewhere (1, 49, 60, 62) and won't be further discussed here. To give a illustrative example of this issue, however, about half of the 65,872 16S rRNA sequences in the Ribosomal Database Project (RDP) database as of April 2003 (13) were obtained from environmental clone libraries as opposed to isolated organisms (Figure 1.1). Furthermore, the habitats where prokaryotic species live are sometimes difficult to sample (e.g. deep ocean, subsurface)

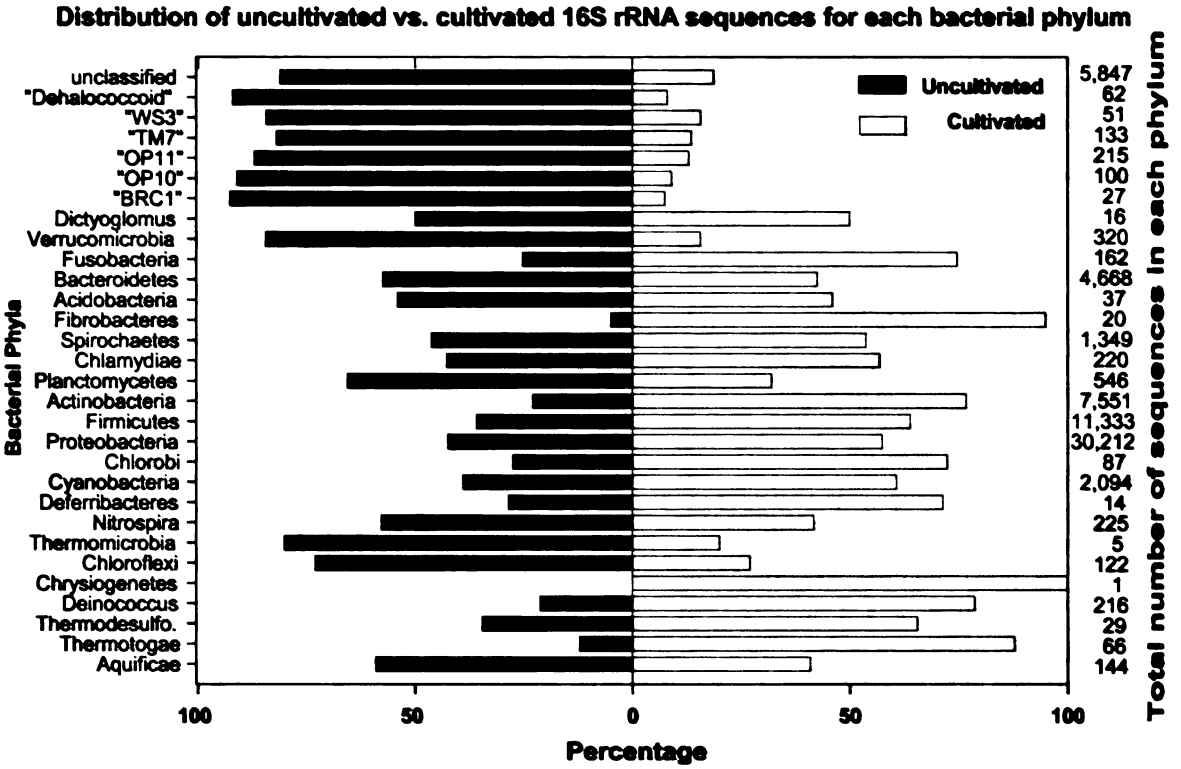


Figure 1.1. Distribution of uncultivated vs. cultivated 16S rRNA gene sequences for each bacterial phylum. Data were collected from 65,872 sequences deposited in the Ribosomal Database Project (RDP) database, as of April 2003. The classification is based on an annotation from GenBank and was provided courtesy of Ryan Farris and James R. Cole of RDP.

or too complex such as soil or sediments for an exhaustive count of prokaryotic species. This has led researches to try to model the total number of prokaryotic species rather than exhaustively count them. In one such classical study, Torsvik and colleges employed whole community DNA-DNA re-association kinetics to estimate the total number of genome equivalents or species considering the 70% DNA-DNA association cut-off as the definition of species (64). Based on this approach 350-1500 and 3500-8800 different prokaryotic species were found in the Norwegian soils sampled (47, 48, 63). Using the same method, the prokaryotic diversity in aquatic environments was found to be orders of magnitude less than that in soil (47). Dykhuizen using data from whole community DNA-

DNA association between related communities estimated that more than a billion (10^9) prokaryotic species exist in soil (20). Several reasons can explain the high soil microbial diversity such as the high diversity of carbon resources; the rather stable, protective, even ancient environment; and what appears to be a high degree of spatial isolation that reduces competition, thereby maintaining less competitive members (61, 65, 72).

Others have used clone libraries of the 16S rRNA gene from environmental samples to estimate prokaryotic diversity. The distribution of unique (representing different species) 16S rRNA gene sequences relatively to the sequences that were observed more than once in these libraries was used for extrapolation to the total number of species in the environment. Assuming a lognormal distribution of species, that is, if species are assigned to log abundance classes, the distribution of species among these classes is normal, Curtis and colleges estimated 6,300 species per gram in two grazed grass-land soils (15). Extrapolating to a larger scale, they estimated the entire bacterial diversity in the oceans to be up to 2×10^6 species, while a ton of soil could contain 4×10^6 different species (15). Hughes and colleges don't share the opinion that species follow a lognormal distribution and thus, they employed a different statistical approach and estimated about 500 species for the same dataset (31).

There are several technical limitations in these approaches to estimated species richness such as the limited sampling and the exhaustiveness of the clone libraries, and uncertainty of the natural species distribution in the environment, the analytical discussion of which is not relevant here. Furthermore, it's uncertain how different species are in different environments, for example different soils, or how much they vary in different geographic locations. In one study to address this question Cho and Tiedje

found that fluorescent *Pseudomonas*, a cosmopolitan heterotroph that is frequently recovered from soil, show a high degree of endemicity at the genotype level (11). If microbial populations have a high degree of endemicity, it greatly expands the earth's total microbial diversity. Second, the description of species based on 16S rRNA gene sequence is problematic mostly because the sequence of this molecule is too conserved to resolve species (59). While the accuracy of estimates of global microbial diversity is in question, it is beyond question that the number of prokaryotic species is large, most probably much larger than the most diverse eukaryotic phylum, the insects (with greater than 10^6 species). Currently only 4,500 prokaryotic species are described (25), which appears to be less than the number of species in a few grams of soil.

Whole-genome sequencing and diversity of prokaryotic genomes.

Whole-genome sequencing[2] was initially employed to advance understanding of species physiology and metabolism but it was soon realized that it could revolutionize the study of other major microbiology disciplines, including functional and genetic diversity. For these reasons and following the major improvements in sequencing technology, capacity, and cost reduction, prokaryotic genome sequencing projects have grown rapidly (Figure 1.2*B*) such that over 115 genomes have been classified as of the end of 2003 and more than 300 other projects are underway. This set of genomic information is now large enough to reveal some major trends in and impressions about prokaryotic genomes and is consistent with the very high prokaryotic diversity discussed above.

Whole-genome sequencing revealed much higher genetic diversity within species than originally anticipated. An example is the *E. coli* case, where whole-genome

[2] deciphering the sequence of all nucleotide bases in the genome

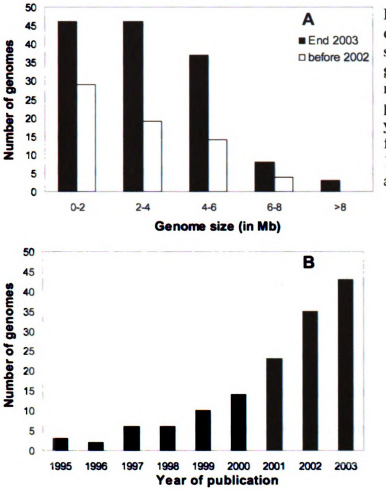


Figure 1.2. Genome size distribution of the fully sequenced prokaryotic genomes (A) and the number of published prokaryotic genomes per year (B). Data were retrieved from NCBI and included all 115 prokaryotic genomes available at the end of 2003

sequences of four strains are now available. Comparative genomic analysis of these sequences revealed that the pathogenic Sakai strain has a genome 1 Mb larger than that of the laboratory strain K12 and about 25% of its genes are not conserved in strain K12 (29, 50). If one considers that prior to whole-genome sequencing, strains of the same species were believed to harbor minimum genotypic differences because they only rarely could be differentiated based on phenotypic characteristics, the genetic heterogeneity revealed between *E. coli* strains was surprisingly high for its time (3 years ago!). Furthermore, the

annotation of the strain specific gene set offered novel insight into the pathogenic lifestyle of strain Sakai relative to the innocuous K12, revealing that our knowledge of even the best-studied pathogen was impeded by the incompleteness of the available conventional methods. Most of the strain Sakai-specific genes are now believed to have been acquired through lateral transfer events based on atypical sequence characteristics (35) and the enrichment of mobile elements such as phage, prophages and insertion sequences in the Sakai genome (29, 50). These findings also suggested that an environmental and benign strain could evolve relatively easily into a devastating pathogen in only about 4.5 million years (rather short in evolutionary time) since the last common ancestor between the two strains (51).

The availability of two additional genomic sequences of *E. coli* strains (strains CFT073 and EDL933) revealed further surprises with regard to the extent of genetic diversity within *E. coli*. Only about 3,000 genes are shared among all four *E. coli* available genomes (69), compared to about 4,000 genes shared between Sakai and K12 strains (50). The 3,000 genes conserved in all *E. coli* strains show, however, a remarkable synteny (interrupted by strain-specific islands) suggesting a vertical transmitted backbone gene set for *E. coli* (69). In summary, the genomic sequencing of *E. coli* strains has revealed not only an enormous genetic diversity at the sub-species level but also the presence of very different selection forces that has led to the accumulation or deletion of genetic material. This subspecies diversity appears not to be unusual since preliminary evidence suggests that *Streptococcus pneumoniae,* based on comparative microarray hybridizations (28), and *Burkholderia cepacia,* based on genome size estimations (38), seem also to have high diversity.

On the other hand, species such as *Mycobacterium tuberculosis* do not appear to share the genetic diversity observe in *E. coli*. Based on both comparative analysis of the sequenced strains (22) and comparative microarray hybridization analysis of several strains (5), *M. tuberculosis* strains are unlikely to be more than 1-2% different in terms of gene content, although the current analysis might be biased due to the study of exclusively clinical isolates. These findings raise another fundamental issue as well. The current species definition based upon 70% DNA-DNA association values poorly correlates with gene content differences within species, as is apparent from the comparison among the *E. coli* and *M. tuberculosis* strains.

Genome structure and its relation to the ecological niche.

Genome sizes vary by more than an order of magnitude among the known prokaryotes (e.g. 0.5-10 Mb). However, the genome size distribution does not appear to be random, for example, it correlates with the ecological niche of the organisms. The smaller genomes are found in endocellular parasites or symbionts (0.5-1.2 Mb), because these organisms occupy a very narrow niche and hence have undergone reductive evolution. For instance, the endosymbiont of aphids, *Buchnera sp.* has a genome size of only 650 Kb compared to 4 Mb of its ancestors from which *Buchnera* diverged 150-250 million years ago (53). For free-living bacteria, genome size correlates with the species metabolism and the width of its ecological niche. Pathogenic species with a narrow range of hosts (or, more generally, species with a narrow ecological niche) also have small genomes, for example, *Helicobacter sp.* and *Streptococcus sp.* Anaerobic bacteria with a restricted metabolism, such as methanogens, typically have small genome sizes, ranging

from 1.5-2.5 Mb. In contrast, aerobic organisms and opportunistic pathogens show higher diversity in genome sizes with some species such as *Pseudomonas* having genomes as large as 6Mb. The largest genomes are found in species that have complex life styles, including myxobacteria and actinobacteria (8-9 Mb). All these observations lead to the conclusion that the interaction between an organism and its particular habitat(s), for example, resource availability and diversity, stable or fluctuating environmental conditions, selects the genome size of species. Nonetheless, what controls the upper genome size in prokaryotes remains poorly understood. Several hypotheses exist, such as the decreased fidelity of replication in large genomes and energy cost to successfully control excessive metabolic repertoire, but none has been experimentally proven.

The variation of genome sizes within species is believed to be rather limited (54), which has been supported by recent genomic sequence data and by pulse field gel electrophoresis of genomes (41). Some of the better-documented exceptions to this are the *E. coli* and *Burkholderia cepacia* species mentioned above, where different strains can vary up to 25% or up to 50% in genome size, respectively (7, 38). On the other hand, genome size can vary up to 3 fold for different species of the same genera! At one end of the spectrum there are species like *Borrelia sp.*, whose chromosomes vary by less than 15 Kb in size (10), whereas species like spirochete *Treponema sp.* (40, 67), and *Mycoplasma sp.* (3) show a variation in genome sizes up to 3- and 2.3-fold, respectively. Perhaps more typical are genera like *Streptomyces* and *Rickettsia*, which vary from 6.4 to 8.2 Mb (36) and 1.2 to 1.7 Mb (24), respectively. It is should be pointed out, however, that too few strains within species and within genera have been studied to give us a complete understanding of the natural variation in the size of prokaryotic genomes.

Although Bacteria are believed to have a single, circular chromosome, an increasing number of exceptions to this are being identified. For example, several species of the α- and β- Proteobacteria have multiple rather than single chromosomes (differentiated from large plasmids by harboring housekeeping genes like ribosomal or tRNA genes), and in at least two, the *Brucella* and *Burkholderia*, the multiple chromosomes are a stable property of the genus. In the proteobacterial phyllum, the multiple chromosomes correlate with a free-living, opportunistic lifestyle, whereas species that are obligatorily associated with animal host or vectors contain no plasmid and, with a few exceptions, single chromosomes (44). Based on these observations, multiple chromosomes are postulated to confer increased genome plasticity and potential for diversification but this has not been proven experimentally yet. Several species with linear, instead of circular, chromosomes and/or plasmids have also recently been described such as the *Streptomyces, Rhodococcus, Borrelia,* and *Agrobacterium* species. Linearity, at least in *Streptomyces* and *Borrelia,* is believed to enhance genomic plasticity, because linear chromosomes (or plasmids) are very unstable and undergo, at high frequency, amplifications and large deletions, often removing the telomeres. This was confirmed with the whole-genome sequencing of *S. coelicolor*, which showed that the secondary metabolite-related genes (*Streptomyces* is notorious for its secondary metabolites like antibiotics) are more frequently encountered in the arms of the chromosome than in its center; the center is biased toward housekeeping genes (6). However, whether linearity offers a selective advantage and why it is phylogenetically constrained to a limited number of species remain unclear.

While it is clear from the above discussion that there is considerable functional and sequence diversity among and within prokaryotic species, whole-genome sequencing has also revealed some universal functional trends as well. For instance, the genomes of endosymbionts have preferentially lost genes involved in metabolism, biosynthesis and regulation while retaining most of the informational genes compared to their free-living relatives (2, 42, 43). Interestingly, although there is a strong deletion bias toward the former major functional categories in the symbiotic genomes, the specific pathways lost appear to be lineage specific, e.g., *Buchnera sp.*, an obligate symbiont of aphids, contrary to other endosymbionts, retains the genes for the biosynthesis of all amino acids (53).

Further, in almost every genome sequenced to date there is a constant percent (about 20-30%) of the predicted protein-coding genes (CDS) that show no homology to any known protein (23). Although it has been suggested that the majority of these are non-coding DNA based on *in-silico* analysis (32, 46, 57), more recent proteomic analyses suggest that at least a portion are translated into proteins, i.e., they presumably represent functional proteins (14, 33, 39). The significant number of "function unknown" genes in every genome also suggests that novel processes are still likely to function in every prokaryotic cell and await characterization. Alternatively, some of these genes might function in well-studied cell processes but their sequences have diverged too much to resemble any of the known annotated sequences.

Biases in the collection of sequenced species.

The current collection of sequenced species is rather limited (compared to the extant of species richness) and there are several issues that should be pursued in the

future for a comprehensive understanding of prokaryotic genetic and functional diversity. A major limitation is that several major phylogenetic lineages remain under- or over-represented with sequenced representatives. For instance, 61 (39.1%) of the 156 completely sequenced strains as of December of 2003 belong to the phylum Proetobacteria while some of the most dominant phyla in nature still have a limited number of sequenced representatives. For example, the Acidobacteria, which appear to be numerically dominant forming up to 52% of 16S rRNA gene sequences in clone libraries from different soils (21, 34, 45) have no sequenced species. Archaea have sixteen completely sequenced species, but this collection is limited to methanogenic or thermophilic species and does not include mesophilic species that are widespread in the ocean and soil environments.

Another limitation of the current collection of sequenced species is that the collection is heavily biased towards organisms with smaller genomes, often from strains living in simpler, resource-rich environments such as endocellular parasites or pathogens (Figure 1.2.*A*). About 70% of the bacterial strains fully sequenced are of clinical importance. A representative example is the Actinobacteria phylum, a dominant group in soil based on culture-independent methods, which has nine sequenced species but all of them are of clinical origin. This picture appears to be changing, however, based on the fact that about half of the ~400 prokaryotic genomic projects that are under-way at the end of 2004 worldwide involve non-pathogenic strains. In conclusion, our current knowledge of prokaryotic physiology and metabolism based on genomic approaches might be still limited and novel findings are anticipated in the near future, particularly among the environmental microbiology.

THESIS OUTLINE

The previous discussion has pointed out the wealth of information laying in the whole-genome sequences and the power of comparative genomic analysis in providing novel insights into (previously) tantalizing scientific questions. Although substantial progress has been made in several areas of microbiology based on analyses of whole-genome sequences, there are still major questions unanswered. I have undertaken several different, and sometimes novel, comparative genomic approaches in order to address several such questions related to the ecology of prokaryotic organisms. Recognizing, at first place, that the collection of currently available genomic sequences is rather limited compared to the extent of prokaryotic diversity, my approaches were designed mostly towards "methodology development" to test larger datasets when these will become available rather than reaching definite conclusions at present. Nonetheless, several reliable trends in and impressions about the interrelationship between ecology and genomic diversity were revealed through my research.

In particular, chapter 2 describes an effort to functionally characterize 115 genomes and to comprehensively evaluate how the relative usage of the genome for specific functions changes with genome size. Such analysis should be informative of what drives genome expansion, provide (further) insight into the interaction between organism's genome and its particular habitat(s) (discussed previously), and suggest what ecological benefits accrue for large genome-sized species. The latter species are believed to be (more) ecologically successful in the soil environment but there is currently limited understanding of why this is the case and the relation of ecology to genome evolution of these species. This work was inspired by and expanded over previous, analogous, studies

on the small genomes of endosymbiotic parasites (summarized earlier), which has offered novel insights into the ecology and evolution of these species. Chapter 2 ends with the analysis of a close phylogenetic group of species, the *Burkholderia cepacia* complex (Bcc), to investigate differences between short (Bcc genomes that recently became available) and long evolutionary scales (previous comparison in the chapter).

The species concept remains a highly controversial and unsettled issue, which has broader impacts such as for reliable diagnosis of infectious disease agents, (inter-) national regulations for transport and possession of pathogens, intellectual property rights, and applications of microorganisms for bioremediation or agriculture purposes. Chapter 3 summarizes my attempts to assess the species-level genetic and functional differences between 81 closely related genomes representing several of the major phylogenetic lineages of Bacteria and thus, help to refine the species concept for Prokaryotes. My approach employed whole-genome sequence comparisons to determine whether species-specific genetic signatures are identifiable (and thus, it is meaningful to have a species concept) as well as the role of the organism's ecology on its common gene content. This information together with information from other approaches, e.g., population-based or gene-expression studies, should eventually converge to a more soundly based species definition for Prokaryotes.

Taxonomic ranks higher than the species rank for Prokaryotes are primarily based on the phylogenetic analysis of the small subunit ribosomal RNA gene (16S rRNA) and secondarily on old microscopic and/or biochemical observations about the relatedness of the organisms. Chapter 4 describes a genome-based approach, expanding from the one undertaken in Chapter 3, to better inform the higher ranks of taxonomy

based on the genetic relatedness of the organisms. Further, the relatedness (between two organisms) estimated by my genome-based approach, which presumably represent a very reliable measurement since it is derived from thousands of independent data points (i.e., genes), was compared to the relatedness estimated by traditional genetic markers such as the 16S rRNA gene sequence to evaluate the robustness and accuracy of the later.

It will become evident from the discussions in Chapters 2 and 3 that the number of available genomes is still rather limited to allow for robust interpretations and conclusions. Therefore, a better sampling with genome-scale information of more species and particularly closely related species is needed. However, it is currently economically unrealistic to do this based on genomic sequencing and thus alternative, high-throughput, methods must be developed. Whole-genome DNA microarray technology appears to be such a promising alternative because it can reveal exact, genome-level, genetic differences between closely related strains based on Competative Genome Hybridization (CGH) of the strains. However, the potential of the microarray technology for CGH has not yet been fully explored. Chapter 5 describes an attempt to simulated microarray CGH experiments *in-silico* by comparing the expected (*in-silico*) microrray results to results from comparisons of whole-genome sequences (control), to evaluate microarray performance for genetic comparisons between strains. Several technical aspects were evaluated, including the resolution level of microarrays, the extent of false positives or negatives and the influence of non-specific signal on the microarray results.

REFERENCES

1. **Amann, R. I., W. Ludwig, and K. H. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev **59:**143-69.

2. **Andersson, S. G., and C. G. Kurland.** 1998. Reductive evolution of resident genomes. Trends Microbiol **6:**263-8.

3. **Barlev, N. A., and S. N. Borchsenius.** 1991. Continuous distribution of *Mycoplasma* genome sizes. Biomed Sci **2:**641-5.

4. **Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace.** 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci U S A **93:**9188-93.

5. **Behr, M. A.** 2002. BCG--different strains, different vaccines? Lancet Infect Dis **2:**86-92.

6. **Bentley, S. D., K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood.** 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417:**141-7.

7. **Bergthorsson, U., and H. Ochman.** 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol Biol Evol **15:**6-16.

8. **Bintrim, S. B., T. J. Donohue, J. Handelsman, G. P. Roberts, and R. M. Goodman.** 1997. Molecular phylogeny of Archaea from soil. Proc Natl Acad Sci U S A **94:**277-82.

9. **Brenner, D., J. Staley, and N. Krieg.** 2000. Bergey's manual of systematic bacteriology, 2nd ed, vol. 1. Springer-Verlag, New York.

10. **Casjens, S., M. Delange, H. L. Ley, 3rd, P. Rosa, and W. M. Huang.** 1995. Linear chromosomes of Lyme disease agent spirochetes: genetic diversity and conservation of gene order. J Bacteriol **177:**2769-80.

11. **Cho, J. C., and J. M. Tiedje.** 2001. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol **67:**3677-82.

12. **Cohan, F. M.** 2002. What are bacterial species? Annu Rev Microbiol **56:**457-87.

13. **Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje.** 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res **31:**442-3.

14. **Corbin, R. W., O. Paliy, F. Yang, J. Shabanowitz, M. Platt, C. E. Lyons, Jr., K. Root, J. McAuliffe, M. I. Jordan, S. Kustu, E. Soupene, and D. F. Hunt.** 2003. Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. Proc Natl Acad Sci U S A **100:**9232-7.

15. **Curtis, T. P., W. T. Sloan, and J. W. Scannell.** 2002. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A **99:**10494-9.

16. **D'Hondt, S., S. Rutherford, and A. J. Spivack.** 2002. Metabolic activity of subsurface life in deep-sea sediments. Science **295:**2067-70.

17. **DeLong, E. F., and N. R. Pace.** 2001. Environmental diversity of bacteria and archaea. Syst Biol **50:**470-8.

18. **DeLong, E. F., L. T. Taylor, T. L. Marsh, and C. M. Preston.** 1999. Visualization and enumeration of marine planktonic archaea and bacteria by using polyribonucleotide probes and fluorescent in situ hybridization. Appl Environ Microbiol **65:**5554-63.

19. **Ducklow, H., and C. Carlson.** 1992. Oceanic bacterial production. Adv. Microb. Ecol. **12:**113-181.

20. **Dykhuizen, D. E.** 1998. Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek **73:**25-33.

21. **Felske, A., A. Wolterink, R. Van Lis, W. M. De Vos, and A. D. Akkermans.** 2000. Response of a soil bacterial community to grassland succession as monitored by 16S rRNA levels of the predominant ribotypes. Appl Environ Microbiol **66:**3998-4003.

22. **Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jacobs Jr, Jr., J. C. Venter, and C. M. Fraser.** 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol **184:**5479-90.

23. **Fraser, C. M., J. Eisen, R. D. Fleischmann, K. A. Ketchum, and S. Peterson.** 2000. Comparative genomics and understanding of microbial biology. Emerg Infect Dis **6:**505-12.

24. **Frutos, R., M. Pages, M. Bellis, G. Roizes, and M. Bergoin.** 1989. Pulsed-field gel electrophoresis determination of the genome size of obligate intracellular bacteria belonging to the genera *Chlamydia*, *Rickettsiella*, and *Porochlamydia*. J Bacteriol **171:**4511-3.

25. **Garrity, G., J. Bell, and T. Lilburn.** Bergey's manual of systematic bacteriology, 2 ed, vol. Release 5.0. Springer-Verlag, New York.

26. **Gold, T.** 1992. The deep, hot biosphere. Proc Natl Acad Sci U S A **89:**6045-9.

27. **Grey, T., and S. Willimas.** 1971. Microbial productivity in soil. Symposia of the Society for General Microbiology **21:**255-286.

28. **Hakenbeck, R., N. Balmelle, B. Weber, C. Gardes, W. Keck, and A. de Saizieu.** 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of Streptococcus pneumoniae. Infect Immun **69:**2477-86.

29. **Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C. G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa.** 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res **8:**11-22.

30. **Hedges, S. B.** 2002. The origin and evolution of model organisms. Nat Rev Genet **3:**838-49.

31. **Hughes, J. B., J. J. Hellmann, T. H. Ricketts, and B. J. Bohannan.** 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol **67:**4399-406.

32. **Jackson, J. H., S. H. Harrison, and P. A. Herring.** 2002. A theoretical limit to coding space in chromosomes of bacteria. Omics **6:**115-21.

33. **Kolker, E., S. Purvine, M. Y. Galperin, S. Stolyar, D. R. Goodlett, A. I. Nesvizhskii, A. Keller, T. Xie, J. K. Eng, E. Yi, L. Hood, A. F. Picone, T. Cherny, B. C. Tjaden, A. F. Siegel, T. J. Reilly, K. S. Makarova, B. O. Palsson, and A. L. Smith.** 2003. Initial proteome analysis of model microorganism *Haemophilus influenzae* strain Rd KW20. J Bacteriol **185:**4593-602.

34. **Kuske, C. R., S. M. Barns, and J. D. Busch.** 1997. Diverse uncultivated bacterial groups from soils of the arid southwestern United States that are present in many geographic regions. Appl Environ Microbiol **63:**3614-21.

35. **Lawrence, J. G., and H. Ochman.** 1998. Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A **95:**9413-7.

36. **Leblond, P., F. X. Francou, J. M. Simonet, and B. Decaris.** 1990. Pulsed-field gel electrophoresis analysis of the genome of Streptomyces ambofaciens strains. FEMS Microbiol Lett **60:**79-88.

37. **Leister, D.** 2003. Chloroplast research in the genomic age. Trends Genet **19:**47-56.

38. **Lessie, T. G., W. Hendrickson, B. D. Manning, and R. Devereux.** 1996. Genomic complexity and plasticity of *Burkholderia cepacia*. FEMS Microbiol Lett **144:**117-28.

39. **Liu, Y., J. Zhou, M. V. Omelchenko, A. S. Beliaev, A. Venkateswaran, J. Stair, L. Wu, D. K. Thompson, D. Xu, I. B. Rogozin, E. K. Gaidamakova, M. Zhai, K. S. Makarova, E. V. Koonin, and M. J. Daly.** 2003. Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. Proc Natl Acad Sci U S A **100:**4191-6.

40. **MacDougall, J., and I. Saint Girons.** 1995. Physical map of the *Treponema denticola* circular chromosome. J Bacteriol **177:**1805-11.

41. **Maule, J.** 1998. Pulsed-field gel electrophoresis. Mol Biotechnol **9:**107-26.

42. **Mira, A., H. Ochman, and N. A. Moran.** 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet **17:**589-96.

43. **Moran, N. A.** 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell **108:**583-6.

44. **Moreno, E.** 1998. Genome evolution within the alpha Proteobacteria: why do some bacteria not possess plasmids and others exhibit more than one different chromosome? FEMS Microbiol Rev **22:**255-75.

45. **Nogales, B., E. R. Moore, W. R. Abraham, and K. N. Timmis.** 1999. Identification of the metabolically active members of a bacterial community in a polychlorinated biphenyl-polluted moorland soil. Environ Microbiol **1:**199-212.

46. **Ochman, H.** 2002. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. Trends Genet **18:**335-7.

47. **Ovreas, L., F. Daae, M. Heldal, F. Rodríguez-Valera, and V. Torsvik.** 2001. Presented at the 9th International Symposium on Microbial Ecology: Interaction in the Microbial World, Amsterdam, 26 to 31 August.

48. **Ovreas, L., and V. V. Torsvik.** 1998. Microbial Diversity and Community Structure in Two Different Agricultural Soil Communities. Microb Ecol **36:**303-315.

49. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734-40.

50. **Perna, N. T., G. Plunkett, 3rd, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner.** 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409:**529-33.

51. **Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam.** 2000. Parallel evolution of virulence in pathogenic Escherichia coli. Nature **406:**64-7.

52. **Rossello-Mora, R., and R. Amann.** 2001. The species concept for prokaryotes. **25:**39.

53. **Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa.** 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp.* APS. Nature **407:**81-6.

54. **Shimkets, L. J.** 1998. Bacterial genomes. Physical structure and analysis. Chapman and Hall, New York.

55. **Sibley, C. G., and J. E. Ahlquist.** 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J Mol Evol **26:**99-121.

56. **Sibley, C. G., J. A. Comstock, and J. E. Ahlquist.** 1990. DNA hybridization evidence of hominoid phylogeny: a reanalysis of the data. J Mol Evol **30:**202-36.

57. **Skovgaard, M., L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh.** 2001. On the total number of genes and their length distribution in complete microbial genomes. Trends Genet **17:**425-8.

58. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kampfer, M. C. J. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52:**1043-1047.

59. **Stackebrandt, E., and B. M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol **44:**846-849.

60. **Staley, J. T., and A. Konopka.** 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol **39:**321-46.

61. **Tiedje, J. M.** 2000. Presented at the Proceedings of the 8th International symposium on microbial ecology, Halifax, Canada.

62. **Tiedje, J. M.** 1995. Approaches to the comprehensive evaluation of prokaryotic diversity of a habitat. CAB International.

63. **Torsvik, V., F. L. Daae, R. A. Sandaa, and L. Ovreas.** 1998. Novel techniques for analysing microbial diversity in natural and perturbed environments. J Biotechnol **64:**53-62.

64. **Torsvik, V., J. Goksoyr, and F. L. Daae.** 1990. High diversity in DNA of soil bacteria. Appl Environ Microbiol **56:**782-7.

65. **Treves, D. S., B. Xia, J. Zhou, and J. M. Tiedje.** 2003. A two-species test of the hypothesis that spatial isolation influences microbial diversity in soil. Microb Ecol **45:**20-8.

66. **Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings.** 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev **60:**407-38.

67. **Walker, E. M., J. K. Howell, Y. You, A. R. Hoffmaster, J. D. Heath, G. M. Weinstock, and S. J. Norris.** 1995. Physical map of the genome of *Treponema pallidum* subsp. *pallidum* (Nichols). J Bacteriol **177:**1797-804.

68. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and T. H. G.** 1987. Report of the Ad Hoc Committee on reconciliation of approaches to Bacterial Systematics. Int. J. Syst. Bacteriol. **37:** 463-464.

69. **Welch, R. A., V. Burland, G. Plunkett, III, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. PNAS **99:**17020-17024.

70. **Whitman, W. B., D. C. Coleman, and W. J. Wiebe.** 1998. Prokaryotes: The unseen majority. PNAS **95:**6578-6583.

71. **Woese, C. R.** 1987. Bacterial evolution. Microbiol Rev **51:**221-71.

72. **Zhou, J., B. Xia, D. S. Treves, L. Y. Wu, T. L. Marsh, R. V. O'Neill, A. V. Palumbo, and J. M. Tiedje.** 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl Environ Microbiol **68:**326-34.

73. **Zumft, W. G.** 1997. Cell biology and molecular basis of denitrification. Microbiol Mol Biol Rev **61:**533-616.

CHAPTER 2

TRENDS BEWTEEN GENE CONTENT AND GENOME SIZE IN PROKARYOTIC SPECIES

Parts of this chapter have been published in the article: K. T. Konstantinidis, and J. M. Tiedje. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. U S A. 2004, 101(9):3160-5.

INTRODUCTION

The genome sequences of the smallest genome-sized prokaryotic species, the obligate endocellular parasites, have provided insight into the interrelationship between the ecology and genome evolution of these species (2, 12, 24). For instance, when compared their free-living relatives, these reduced genomes have preferentially lost genes underlying the biosynthesis of compounds that can be easily taken up from the host, such as amino acids, nucleotides, and vitamins. Furthermore, regulatory elements, including σ factors, have commonly been eliminated from such symbiotic bacteria, presumably due to the rather stable environment inside host cells, which renders extensive gene regulation useless (3, 11, 31). It is not yet clear whether there may also be trends in gene allocation for the larger genome-sized free-living bacteria. If such trends do exist, they could reveal strategies of genome expansion, provide insight into the upper limit of genome size, reveal whether there is more centrally coordinated regulation, and most important, suggest what ecological benefits accrue for such species.

There is currently an increasing amount of evidence that favors the existence of universal trends between functional gene content and genome size. For instance, Jordan *et al.*'s (16) analysis of 21 genomes showed that lineage-specific gene expansion is positively correlated with genome size and may account for up to 33% of the coding capacities in the genome. Furthermore, comparative genomic studies of *Pseudomonas aeruginosa* PAO1 and *Streptomyces coelicolor* A3, two larger genome species, noted a disproportionate increase relative to smaller genome-sized species in regulatory and transport genes and in genes involved in secondary metabolism, respectively (5, 33). However, only a limited number of species were analyzed in both of these studies, and

the analysis was restricted to specific functional processes. Furthermore, in the former study, no other species in the panel of strains evaluated had a genome size comparable to strain PA01, a moderately large (6.3-Mb) genome-sized strain; thus, the significance of these findings for other large prokaryotic genomes is unknown.

We sought to more comprehensively evaluate how the relative usage of the genome changes with genome size, using all sequenced genomes and evaluating all functional classes of genes.

MATERIAL AND METHODS

Functional annotation of all sequenced prokaryotic genomes.

We undertook the functional characterization of 115 completed genomes deposited in the GenBank database as of May 2003 using the Clusters of Orthologous Groups (COG) database (34, 35). The list of genomes used in this study as well as the genome size, the GC% content, the total number of predicted protein-coding sequences (CDS) and what fraction of the CDS was assignable to the COG database (see below) for each genome is available in Table 2.1 of Appendix. At the time of this study, the COG database was comprised of 144,320 protein sequences from 66 completed genomes forming 4,873 groups of orthologous proteins (COG). Individual COG are clustered in 20 individual functional categories, which are further grouped in four major classes (see Table 2.1). All possible CDS from the 115 genomes were assigned to a functional category according to the category where their best COG homolog is classified. Homologs were identified by using the BLAST local alignment algorithm (1) and a cut-off of at least 30% identity at the amino acid level over 70% of the length of the query protein in pair-wise sequence comparisons. This cut-off is above the twilight zone of similarity searches where inference of homology is error-prone due to low similarity between aligned sequences; thus query proteins were presumably homologous to their COG match (28, 30). Homologous proteins can be either orthologs (homology through speciation) or paralogs (homology through lineage specific gene duplication), and both paralogs and orthologs are assumed to retain the same biochemical function, whereas paralogs have usually diverged in specificity (9, 13). Therefore, CDS are expected to share at least the same general function with their COG matches. PERL scripts were used

to edit CDS assignments where necessary; formatting databases for BLAST searches and automatically parsing BLAST outputs.

We further tested our findings from the COG database by using the publicly available data from the ortholog group table database at the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the Comprehensive Microbial Resource database (CMR) supported by The Institute for Genomic Research (TIGR). The KEGG database classifies orthologous genes from all sequenced species into 24 functional categories (17). An identical strategy as previously mentioned for COG was used to assign each CDS from 75 fully sequenced genomes (the same genomes used for TIGR data below) to a KEGG functional category. TIGR performs an automated whole-genome annotation on any published microbial genome, which classifies genes in 19 redundant Role Categories (or functional categories), i.e., a single protein can be assigned in more than one category (27). The number of proteins devoted to a Role Category for each of the 75 genomes incorporated in CMR as of July 2002 was obtained from the Multi Genome Query Tool at the CMR web site (www.spacetransportation.org_Detailed_44108.html).

The amount of noncoding DNA in any genome was calculated by subtracting the sum of the lengths of the coding sequences annotated in the GenBank files from the estimated size of the genome.

RESULTS AND DISCUSSION.

With the previously described strategy, we were able to assign, on average, 70.3% of the CDS in any genome to a COG functional category. If one considers that a significant amount of predicted genes (~15–20%) is species-specific in every genome sequenced so far (25), we have characterized the large majority of the repertoire of each cell.

Data Normalization.

Our main objective was to study the relationship between the total CDS in the genome and the genomic fraction devoted to a functional category. To normalize the effect of the different degrees of representation in the database, genomes with too many or too few genes homologous to the database were not included in inferring patterns with genome size, i.e., genomes in which the percentage of genes homologous to the database fell within one standard deviation from the mean ($\bar{x}$ 70.3%, SD 11.2) are represented by solid squares (87 of the 115 genomes), whereas the rest are represented by open squares (Figure 2.1). Functional categories showed similar trends with total CDS in the genome both when the normalized set and all genomes were considered (Table 2.1). However, trends with the normalized set should be more accurate because this set minimizes the bias in database representation. The power correlation gave among the highest R^2 values from the types of correlations tested for most functional categories. It should be mentioned however, that there were, typically, very small differences between different models (e.g. linear, power, logarithmic etc) in their ability to describe the trends with

Table 2.1. COG functional categories and category correlation with total number of CDS.

Functional Class	*Individual functional categories*	*Correlation**	*Normal. Species†*	*>2000 CDS*	*All Species‡*
Information	**J:** Translation, ribosomal structure and biogenesis	-	0.99, <0.001	0.95, <0.001	0.98, 0.001
	K: Transcription	+	0.44, <0.001	0.18, 0.001	0.37, <0.001
	L: DNA replication, recombination and repair	-	0.21, <0.001	0.19, 0.002	0.24, 0.004
Cellular processes	**D:** Cell division and chromosome partitioning	-	0.41, <0.001	0.55, <0.001	0.37, <0.001
	V: Defense mechanisms	No	- 0.096	0.20 0.001	- 0.38
	O: Posttranslational modification, protein turnover	No	0.13, 0.002	- 0.66	0.29 <0.001
	M: Cell envelope biogenesis, outer membrane	No	- 0.19	- 0.26	- 0.40
	P: Inorganic ion transport and metabolism	No	0.18, <0.001	- 0.60	0.1 0.001
	U: Intracellular trafficking, secretion	No	0.15, 0.001	- 0.36	0.27 <0.001
	N: Cell motility	+	0.1, 0.004	0.16, 0.001	- 0.64
	T: Signal transduction mechanisms	+	0.55, <0.001	0.20, <0.001	0.46, <0.001
Metabolism	**F:** Nucleotide transport and metabolism	-	0.44, <0.001	0.57, <0.001	0.53, <0.001
	G: Carbohydrate transport and metabolism	No	- 0.015	- 0.44	- 0.023
	E: Amino acid transport and metabolism	No	0.29, <0.001	- 0.09	0.07 0.005
	H: Coenzyme metabolism	No	- 0.23	- 0.05	0.11 0.0006
	I: Lipid metabolism	No	- 0.04	0.15, 0.002	- 0.14
	C: Energy production and conversion	+	0.1, 0.004	0.15 0.002	- 0.29
	Q: Secondary metabolites transport and metabolism	+	0.30, <0.001	0.12, 0.005	0.31, 0.001
Poorly characterized	**R:** General function prediction only	No	- 0.012	- 0.48	0.06, 0.008
	S: Function unknown	No	0.4, <0.001	- <0.86	0.24 <0.001

The eighteen functional categories (2nd column) are grouped in four major classes (1st column). Adapted from COGs' web site. *The genomic fraction attributable to a functional category showed universal positive (+), negative (-), or no (No) correlation with total CDS in the genome when both correlations for normalized genomes (4th column) and normalized genomes with more than 2000 CDS (5th column) were significant at a p-value threshold of 0.01 (p-value denotes the confidence level that the correlation observed is significantly different from the null hypothesis e.g. no correlation). †Power correlation R^2 and p-values for each set are shown. ‡The 6th column shows correlations for all 99 bacterial genomes used in this study. The 16 archaeal genomes were not included in the analysis because Archaea had significantly different genomic fractions from Bacteria in many functional categories.

total CDS in the genome (data not shown). Thus, no assumptions can be made about the mechanisms underlying the relationship between functional gene content and total CDS in the genome. Last, the use of genome size instead of total CDS in the genome gave identical results due to the high correlation ($R^2 = 0.98$) between these two parameters of the genome (Figure 2.3*A*). Therefore, total CDS in the genome and genome size are used interchangeably in the following text.

Major Trends with Genome Size.

To identify major universal trends, as opposed to ones that are attributable to the preferential gene loss in the reduced genomes, the analysis was repeated including only normalized genomes that had at least 2,000 CDS annotated in their genomic sequences. COG functional categories that showed correlation with genome size for both sets tested (i.e., all solid squares and solid squares with 2,000 CDS) were considered cases of major trends, and these categories are shown in Figure 2.1. Categories that showed correlation with genome size (at a *P* value threshold of 0.01) for only one of the two sets of genomes tested were considered cases of minor trends and are shown in Figure 2.2. All findings are summarized in Table 2.1.

The COG functional categories that showed universal correlation with genome size were: informational categories of translation, ribosomal structure and biogenesis, and DNA replication recombination and repair. These categories showed a strong negative correlation with genome size, whereas transcription (transcription apparatus and transcription control genes) showed a strong positive correlation (Figure 2.1. *Left*). Of the cellular processing categories, the percent of genes related to cell division and

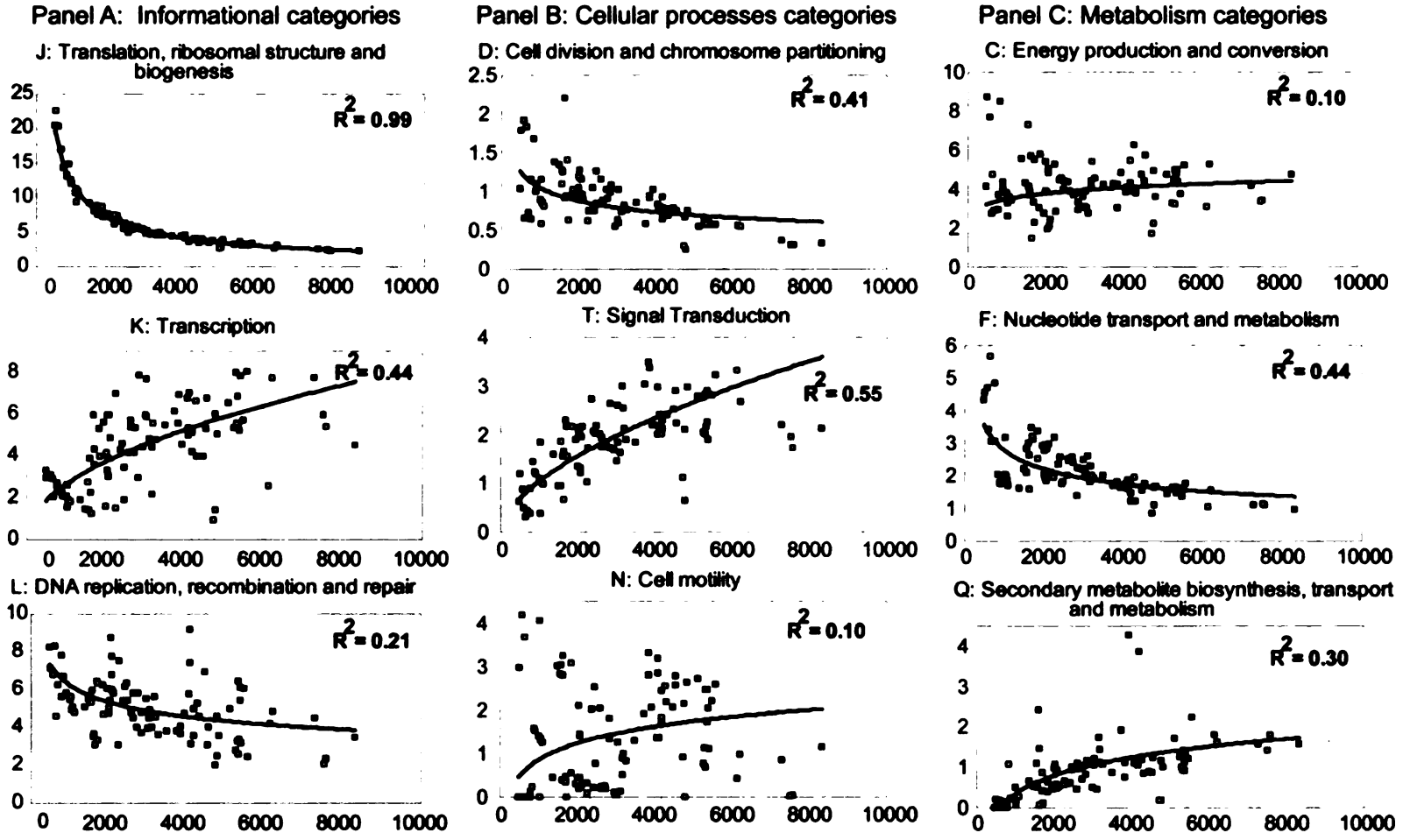


Figure 2.1. COG functional categories that showed universal correlation with total CDS in the genome. Y-axes are the percent of CDS in the genome attributable to a specific COGs category (graph title) and X-axes are the total CDS in the genome for each of the 99 sequenced bacterial genomes. Solid squares represent genomes that had a reasonable number of genes with homologs in the COG database whereas open squares represent genomes that had either too many or too few genes with homologs in the database (outliers). Trendlines and R^2 shown are for the solid squares. Archaeal genomes were not included because Archaea had significantly different genomic fractions from Bacteria in many functional categories.

chromosome partitioning category showed a small decrease with genome size (~1–2%), whereas the percent of genes related to signal transduction mechanisms and cell motility strongly and moderately increased with genome size, respectively (Figure 2.1. *Center*). Among the individual metabolism categories, nucleotide transport and metabolism showed a strong negative correlation with genome size, whereas energy production and conversion and secondary metabolite biosynthesis, transport, and catabolism showed a moderate and strong positive correlation with genome size, respectively (Figure 2.1. *Right*). Notably, genomes with <2,000 CDS have almost no secondary metabolism related genes (Figure 2.1. *Right*).

Minor Trends with Genome Size.

Categories of posttranslational modification and protein turnover, inorganic ion transport and metabolism, intracellular trafficking and secretion, amino acid transport and metabolism, and function unknown categories showed correlation only when all solid squares were considered, i.e., no correlation for solid squares with >2,000 CDS (Table 2.1 and Figure 2.2). Therefore, these trends are attributable to the preferential gene loss in the reduced genomes. Furthermore, several categories that were universally correlated with total CDS in the genome showed stronger correlation with all solid squares compared to solid squares with >2,000 CDS. Thus, such categories like transcription, signal transduction, and secondary metabolite biosynthesis are also affected by preferential gene loss in thereduced genomes. These results are in good agreement with the current knowledge of which functional categories are more likely to have been reduced in the symbiotic genomes.

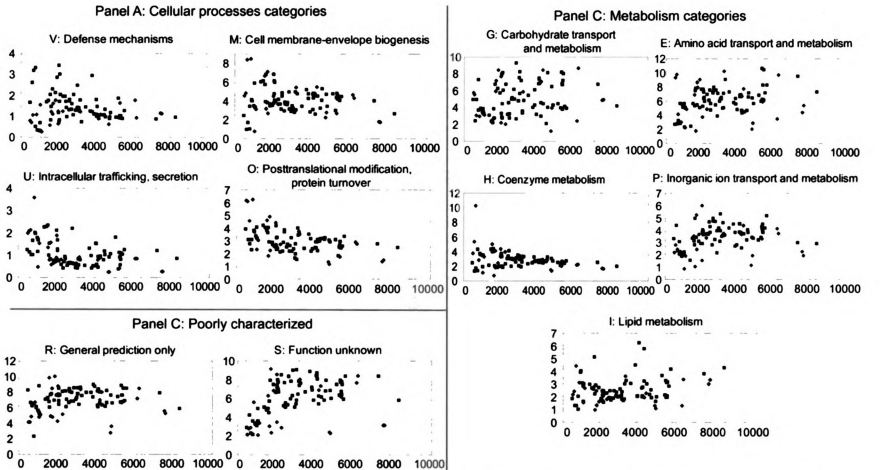


Figure 2.2. COG functional categories that showed no correlation with genome size. These categories showed no correlation with genome size (at a P value threshold of 0.01) for one or both of the sets of species tested (i.e., all solid squares and solid squares with > 2,000 CDS). Only datapoints representing bacterial genomes are shown, because Archaea had significantly different genomic fractions in many of the categories shown.

On the other hand, categories of defense mechanisms and lipid metabolism showed correlation only when solid squares with >2,000 CDS were considered (Table 2.1 and Figure 2.2). These trends, however, are more likely a database artifact due to the under-representation of large genomes than a real preferential accumulation of such genes by the large genomes. The fact that there were several small genomes with high percentages of CDS devoted to these categories (which accounted for the lack of correlation when all solid squares were considered) supports the former interpretation. Last, it should be mentioned that most minor trends involved weak correlations and small changes (~1–2%) in the fraction of the genome devoted to the corresponding functional categories.

Non-coding DNA and Hypothetical CDS.

Interestingly, the genomic fraction assigned to hypothetical CDS (i.e., poorly characterized categories) remained constant for genomes with >2,000 CDS. Moreover, the fraction of non-coding DNA was also invariable (at ~12–14% of the genome) for all 115 genomes evaluated (Figure 2.3.*B*), which confirmed previous results that analyzed a smaller set of species (22). Therefore, the large prokaryotic genomes overall are not explained by disproportionate accumulation of junk DNA, i.e., hypothetical genes or non-coding sequence.

In contrast, genomes with <2,000 CDS have a smaller percent of function unknown (or conserved hypothetical) CDS compared to larger genome-sized species. This suggests that some of these genes, if they indeed code for proteins, have dispensable functions in the larger genome-sized bacteria. If these genes follow the trends of the other

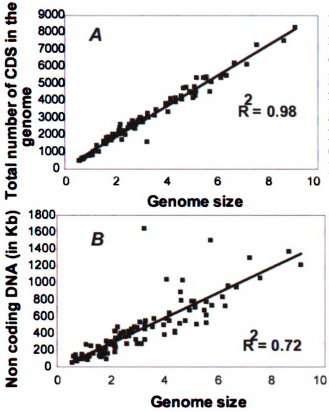


Figure 2.3. Correlation among total number of CDS in the genome, non-coding DNA, and genome size for prokaryotic genomes. (**A**) The total number of CDS in the genome vs. the genome size for 115 completed prokaryotic genomes. (**B**) The total amount of non-coding DNA in the genome vs. genome size.

functional categories, then these unknown genes may be involved in regulation or secondary metabolism rather than in informational processes. Nonetheless, a significant fraction (~3%) of the genes in the reduced genomes remains attributable to the function unknown category. Their retention suggests that at least some of the conserved hypothetical genes encode for functional proteins.

Factors Other than Genome Size.

The correlation R^2 values indicate that genome size can only partially explain some of the shifts in gene content. Strain-specific traits are assumed to be responsible for datapoint dispersion around the mean, which is pronounced for several functional

categories. For example, by examining individual COG, we conclude that the number of the prevalent ABC transporter genes (and transport genes in general) was proportionately increased (i.e., the genomic fraction devoted to them remained constant) with genome size, and there was little dispersion around the mean suggesting a universal relationship with genome size (Figure 2.4). However, specific bacterial groups like the ecologically versatile α-Proteobacteria *Agrobacterium* and *Mesorhizobium* sp. had a disproportionately increased number of ABC transporters, whereas the more habitat-specific bacteria like the γ-Proteobacteria *Xanthomonas* sp. had fewer than the average ABC transporters.

As far as traits other than total CDS in the genome are concerned, we evaluated whether the ribosomal rRNA (*rrn*) copy number could explain some of the shifts in functional gene content. The *rrn* copy number had, typically, a small effect on functional gene content compared to the total CDS in the genome. However, in the case of carbohydrate transport and metabolism, the correlation was stronger for *rrn* copy number

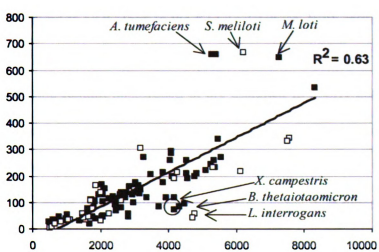


Figure 2.4. ABC transporter genes proportionately increase with genome size. *Y*-axis is the number of genes attributable to ABC transporter functions, and *x*-axis is the total CDS in the genome for each of the 99 fully sequenced bacterial genomes. Genomes that have disproportionately increased or decreased their number of ABC transporter genes are denoted on the graph.

($R^2 = 0.4$, $P < 0.001$) than for total CDS in the genome (correlation not significant at $P = 0.01$). The *rrn* copy number is positively associated with the rate at which phylogenetically diverse bacteria respond to resource availability (18), thus the strong correlation between carbohydrate metabolism and transport and *rrn* copy number is not surprising.

Last, the higher variability observed for data points representing small genomes is partially attributable to the fact that a small genome will show a dramatic change in functional patterns with a small change in the number of genes for a cellular process. Thus, while analyzing the percent of genes in a functional category can reveal major changes, it is less sensitive for detecting changes among large genome-sized prokaryotes.

Results from KEGG and TIGR Annotation Databases.

Results using COG, KEGG, and TIGR databases are not always directly comparable because of database-specific characteristics. Although the KEGG Orthology database performs high-quality annotation, it has incorporated a limited (only the well-described) number of pathways and processes (17). Thus, more orthologous groups can be found in COG than in the KEGG database. With respect to TIGR annotation, although assignment of correct function is usually satisfactory (~90%), ~50% of the genes in a genome remain unassigned or are assigned to poorly characterized categories (vs. ~40% for COG) (27). Moreover, as noted on the CMR web site, all Role Category data were generated at the time each genome was entered into the CMR; thus newer genomes may have more genes assigned to Role Categories than older ones. Despite these limitations, there are several categories that are comparable among the three databases and hence can

be used to test the validity of the trends revealed with COG. Our results for these categories were congruent (a selected set of KEGG and TIGR's functional categories is presented in Figure 2.5). For example, KEGG and TIGR informational categories of protein translation and DNA replication were negatively correlated with genome size (R^2 > 0.4 for all categories), whereas regulation category was positively correlated with genome size (R^2 > 0.5), similar to the COG data.

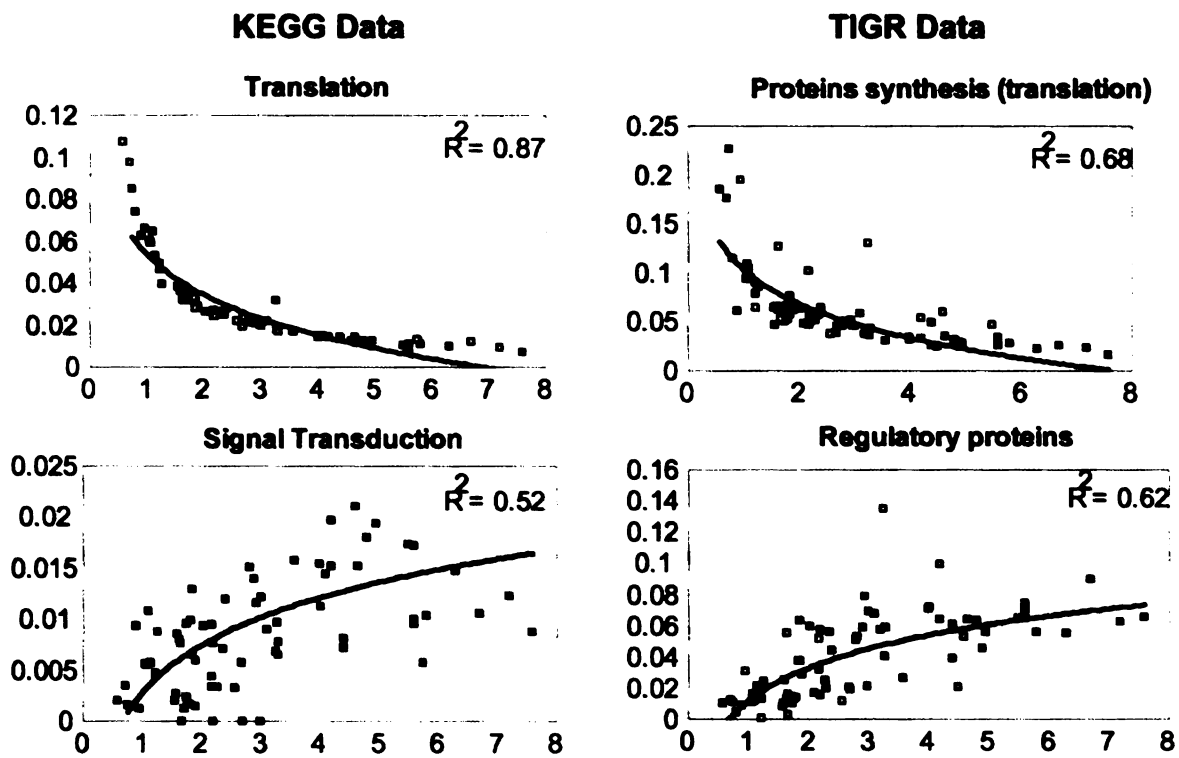


Figure 2.5. Evidence for functional biases with genome size from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and The Institute for Genomic Research (TIGR) annotation databases. Y-axes are the genome portions (CDS attributable to a functional category divided by total number of CDS in the genome) devoted to the specific functional category, and x-axes are the corresponding microbial genome sizes. Solid and open squares are used as previously for COGs data (Figure 2.1). Corresponding functional categories between the two databases are placed next to each other.

We also analyzed the 39 partially sequenced genomes in the JGI database in the same way. This is a collection of exclusively environmental strains, which includes seven strains with genome sizes >6Mb (average genome size, 3.83 vs. 3.23 Mb in the closed set). Although trends between gene functional categories and total CDS in the genome for JGI genomes were very similar to those for the fully sequenced genomes (data not shown), only 59.8% (vs. 70.3% for the closed set) of the CDS in the JGI set were assignable to a COG category. This may indicate that this genome set samples more of the uncharacterized genes in nature, although some of the difference is likely due to the lack of manual curation of the annotation.

Bacteria vs. Archaea.

Our analysis also revealed that there were some notable but small differences between Bacteria and Archaea in the relative usage of the genome for the different cell functions (Figure 2.6). Archaea appeared to have a higher genomic portion devoted to energy production and conversion, coenzyme metabolism, and poorly characterized categories than their bacterial counterparts of the same genome size. On the other hand, Archaea had relatively fewer genes involved in carbohydrate transport and metabolism, cell envelope and membrane biogenesis, and inorganic ion transport and metabolism. Some of the differences, like those concerning energy production, cell envelope, and general prediction-only categories were more strongly supported by the data (compare errors bars in Figure 2.6).

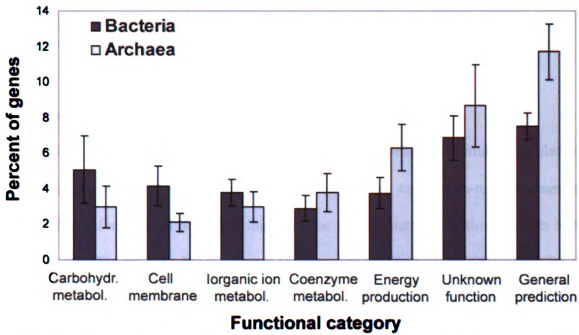


Figure 2.6. Differences between Archaea and Bacteria in the relative usage of the genome. Bars represent the average from 34 bacterial and 12 archaeal genomes, which have between 1,500 and 3,500 CDS (to avoid any genome size effect on the data). Only normalized genomes have been included (see text). Average are statistically different by two-tailed *t* test, assuming unequal variances and 0.05 confidence level. Functional categories that had <2% of the genes in the genome are not shown.

A set of archaeal specific proteins in addition to the standard proteins encountered in a typical prokaryotic cell would explain the higher genomic fraction in the above categories for Archaea. In agreement with this hypothesis, Graham *et al.* (14), in an attempt to define an archaeal genomic signature, concluded that genes with no detectable bacterial or eukaryotic homologs mostly involve energetic systems and cofactor biosynthesis, e.g., genes involved in methanogenesis. On the other hand, the fewer genes for cell-wall biogenesis are probably attributable to the fact that Archaea possess a different cell wall from Bacteria. Archaea lack peptidoglycan in their cell wall, and peptidoglycan biosynthesis requires a battery of enzymes in bacteria (19). Furthermore,

the archaeal cell wall components and metabolism have not been studied to the same extent as those for Bacteria and hence are missing from the database.

What Is Gained in a Large Genome?

Our analysis showed that larger genomes preferentially accumulate regulation, secondary metabolism, and, to a smaller degree, energy conversion-related genes as opposed to informational ones, judging from the inverse pattern for these classes with genome size (Figure. 2.7). We performed the same analysis in May of 2002, using the 75 genomes available at that time and a database of 3,852 COG groups (vs. 4,873 COG currently). The results between this set and the expanded set of 115 genomes presented herein were very consistent, and correlations were often more significant in the latter set. Secondary metabolism and energy conversion rather than general metabolism are disproportionately expanded in larger genomes and thus should explain a large part of the broad metabolic diversity that characterizes large genome-sized species. The expansion involved both expansions of specific COG and *de novo* acquisitions of new COG (or pathways), with the latter case being roughly twice as frequent as the former one (data not shown). On the other hand, the genes assignable to the remaining metabolism, except nucleotide metabolism, and several cellular processes categories are only proportionally increased with genome size (similar to the example of ABC transporter genes mentioned previously).

Regardless of a proportional or disproportional increase in metabolic or cellular pathways, large genome-sized species would need increased regulation to successfully control the extensive metabolic repertoire they apparently possess under different growth

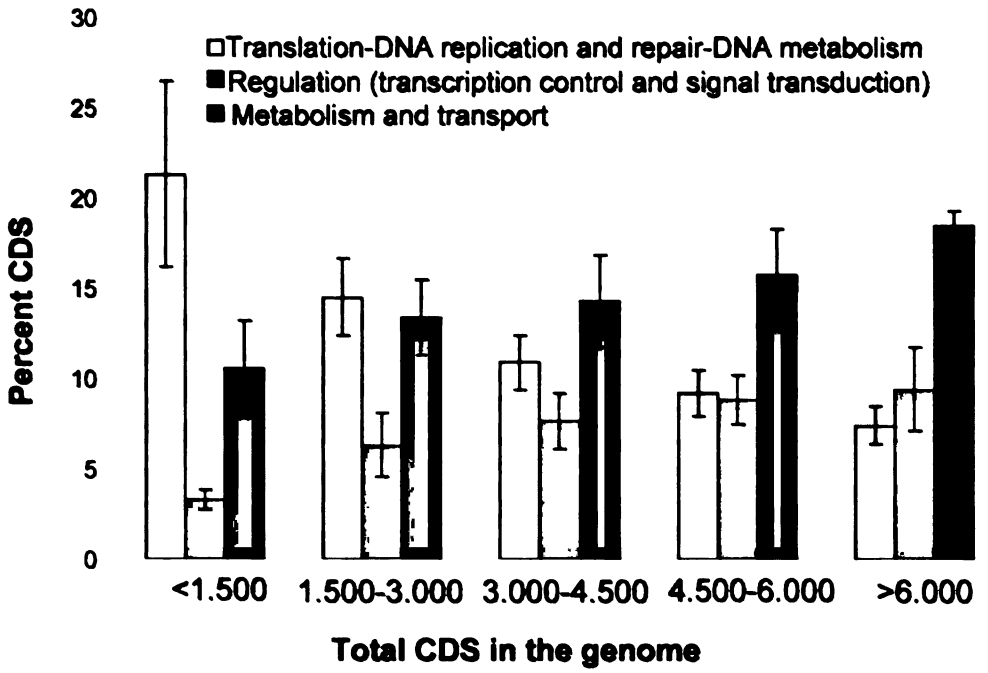


Figure 2.7. Summary of the shifts in gene content with genome size in prokaryotic genomes. The bars represent the sum of the COG functional categories, which showed strong correlation with genome size and are involved in the same major cellular processes. Only normalized genomes (represented by solid squares in Figure 2.1) have been included. Errors bars represent the standard deviation from the mean except for the last genome size class, where error bars represent data range due to a small number of normalized genomes in this class (three genomes).
This consistency gives higher confidence in the trends reported. These data suggest that

conditions. Thus, it is not surprising that regulatory genes, i.e., transcription control, and signal transduction, dominated the genes that are disproportionately increased in larger genomes. In addition, many regulation systems are expected to cross talk, because their genes share high sequence similarity (paralogous genes of expanded gene families), which suggests increased complexity in regulation as well. In agreement with these interpretations, all species with genome sizes >6 Mb in our set are free-living bacteria that can grow in very diverse environments, several using alternative electron acceptors and a great range of substrates for energy production (Table 2.2).

Table 2.2. Genomic information and ecological niche(s) of species with a genome size larger than 6Mb.

*Species**	*Gen. size*	*% in COGs*	*Ecological niche*
Bacteroides thetaiotaomicron	**6.26**	**33.5**	Human gut, metabolically versatile
Bradyrhizobium japonicum	9.11	60.4	Soil, rhizosphere. N_2 fixing symbiont of legumes
Mesorhizobium loti	7.59	69	Soil, rhizosphere. N_2 fixing symbiont of legumes
Nostoc sp.	**7.2**	**58.2**	Cyanobacteria, ubiquitous in nature. Photosynthetic
Pseudomonas sp. (aver. of 3)	6.2-6.4	69-80	Soil, water. Opportunistic pathogen of plants, humans
Sinorhizobium meliloti	6.7	63	Soil, rhizosphere. N_2 fixing symbiont of legumes
Streptomyces avermitilis	**9.03**	**48.8**	Ubiquitous in soil. Very versatile metabolically
Streptomyces coelicolor	**8.67**	**40**	Ubiquitous in soil. Very versatile metabolically

*All environmental and non-proteobacteria strains (bold) have <58.2% (vs. an average of 70.3%) of their genes homologous to COG proteins (3rd column). This indicates that the over-representation of specific lineages (e.g., proteobacteria) and clinical strains in the database has possibly biased our knowledge of microbial functional gene content.

The negative correlation with genome size of informational and DNA metabolism categories is equally interesting (Figures 2.1 and 2.7). This trend suggests that a similar number of informational and DNA metabolism related proteins is able to cope with an increased number of genes. For instance, there is a relatively small increase in the absolute number of genes (of ~20%) in the translation category between 2- and 8-Mb-sized genomes. This may be attributable to there being sufficient informational processes present and active at any time in the cell. Thus, when there is an unusual demand for informational proteins because of a larger genome, their transcription or posttranslational modification can be regulated accordingly to yield sufficient active proteins.

A Hypothesis for Large Genomes.

Presumably the interactions between the organism and particular habitat(s) have selected for genome expansion. Large genomes do not appear to be uncommon in nature

(Table 2.2 and JGI genomes), and hence they must have value. As noted above, all over-amplified gene families are associated directly or indirectly (regulation) with metabolism. However, the lack of knowledge of the population sizes and activities of such species in natural environments does not allow specific inferences about which environmental factors may have fostered genome expansion. In contrast, the genome evolution in endosymbiotic bacteria is much better understood. The relief from selection for specific pathways and regulation systems along with population bottlenecks that allow more rapid fixation of mutations are proposed to determine their genome evolution (2, 10, 22). Also, the higher number of bacterial generations in these nonnutrient-limiting environments probably facilitates loss of DNA through spontaneous recombination events at repeated or mobile sequences (2, 10).

One hypothesis for large genomes consistent with the above data is that Bacteria with such genomes are more dominant, population-wise, in environments where resources are scarce but diverse and where there is little penalty for slow growth. These are characteristics of soil. In support of this, Mitsui *et al.* (23) and Klappenbach *et al.* (18) found slow-growing oligotrophic α-Proteobacteria to be more dominant in soil. In the former study, many of these isolates were nonsymbiotic members of the Rhizobiaceae and Bradyrhizobiaceae (23, 29), families that have genomes >6–8 Mb. Generation times in soil are thought to be low, with mean generations measured at three per year (15).

Although this study shows some clear trends between gene content and genome size, the dispersion around the mean for many categories suggests that features other than genome size likely explain what is gained in larger genomes. These traits need to be explored for a fuller understanding of the interactions between ecology and genome

evolution. This study also draws attention to the limited number of large genomes sequenced to date. The possibility that large genomes represent a significant fraction of the extant microbial world and that they may possess unique traits missed in the current annotation knowledge is a major challenge for microbiologists.

A CASE STUDY: THE *BURKHOLDERIA CEPACIA* COMPEX.

The previously described work clearly indicates what is gained in a large genome and suggests that the interactions between the organism and particular habitat(s) select the organism's genome size and gene content. In order to expand understanding of the latter, we have performed a similar genomic analysis on a model bacterial group, the *Burkholderia cepacia* complex (Bcc) (α-Proteobacteria). The Bcc was chosen because its members are phylogenetically very close, as opposed to previous work that included comparisons between distantly related organisms. This facilitates comparative analysis and could be informative differences between short vs. long evolutionary scales. Furthermore, a substantial body of information on the ecological and physiological differences of its members is available.

Background on *Burkholderia cepacia* complex.

The Bcc consists of ten closely related species (Figure 2.8), which share a high degree of 16S rRNA and *recA* sequence similarity 98-100% and 94-95%, respectively, and moderate levels of DNA-DNA reassociation homology (30-50%) (8). Members of the Bcc are successful in very different ecological niches ranging from rhizosphere colonization, biodegradation of pollutants, plant pathogenesis, and chronically infectioning Cystic Fibrosis (CF) patients, which frequently results in narcotizing pneumonia (known as the "*B. cepacia* syndrome") (4, 7, 26). Moreover, Bcc species are among the most versatile bacterial species known, e.g., the type strain of *Burkholderia cepacia* species (formely *Pseudomonas cepacia*) has been shown to catabolize more than 200 organic sources of carbon (20). While Bcc species have among the largest

prokaryotic genomes, the genome size distribution of the group is very wide, ranging from 6 to 9 Mb (21). Interestingly, the genome is typically organized in 3-4 replicons, which is thought to give Bcc strains genomic plasticity and ecological versatility.

To help understand how the group as a whole has adapted to the very different environments, three Bcc genomes have recently been sequenced. These genomes are: *B. cenocepacia* J2315, an enhanced virulent pathogen in CF, *B. cepacia* ATCC 17760, one of the classical Stanier's collection of strains isolated from Trinidad forest soil (32), and

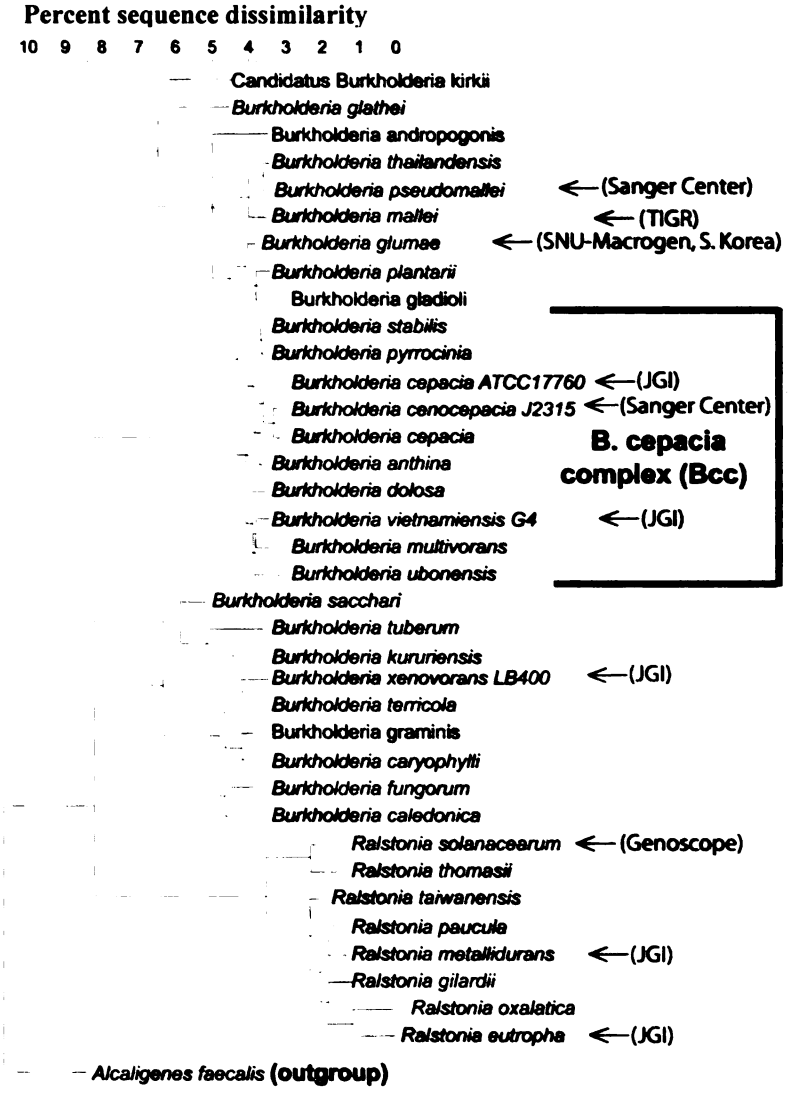


Figure 2.8. The *Burkolderia cepacia* complex and its relationship to other *Burkholderia* spp. 16S rRNA phylogenetic tree (based on the neighbour-joining method) showing the phylogenetic relationships of Bcc and other *Burkholderia* and *Ralstonia* species. Arrows indicate species that are sequenced or are currently being sequenced.

B. vietnamiensis strain G4 (ATCC 53617), a rhizosphere colonizing strain that also oxidizes the groundwater pollutant trichloroethene. The J2314's genome is now fully sequenced by the Sanger Center and consists of three chromosomes, 3.9, 3.2 and 0.9 Mb in size whereas G4 and ATCC 17660 are currently at high draft status e.g., the available sequence covers >95% of the strain's genomic DNA (6). The estimated genome sizes of the sequenced strains are: J2315 8 Mb, ATCC 17660 8.7 Mb, and G4 8.5 Mb.

Genomic comparisons among the Bcc genomes.

Comparative whole-genome analysis of the three available Bcc genomes reveals about 4,200 predicted protein-coding sequences (CDS) that are conserved in all three genomes (Figure 2.9). The distribution of gene functions in this Bcc conserved gene core follows closely the trends with genome size reported in the previous section of chapter 2, e.g., regulation and metabolism functions are disproportionably increased relative to

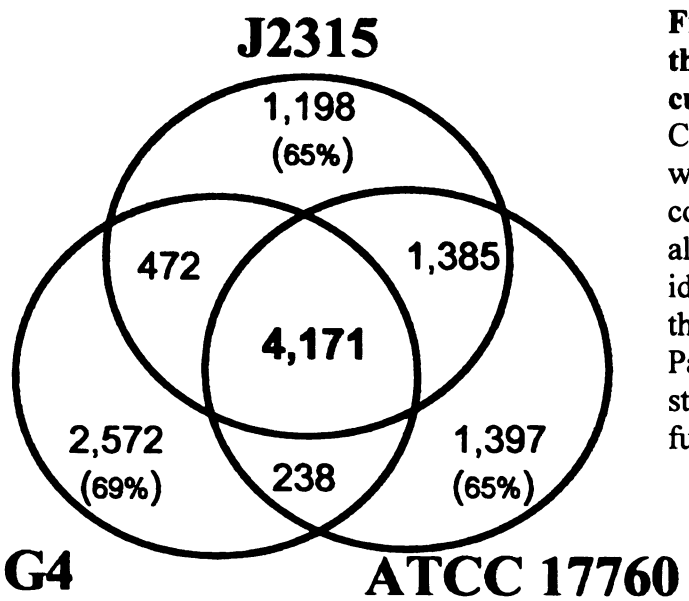


Figure 2.9. Venn diagram showing the gene complements of the currently available Bcc genomes. Conserved genes were defined by whole-genome pairwise sequence comparisons, using the BLAST algorithm (1) using a cut-off of 30% identity (a.a. level) over at least 70% of the length of the query CDS. Parentheses denote the fraction of the strain-specific genes that has unknown function.

information functions according to the correlations described previously (analytical data not shown). In addition, when compared to an average of all closed genomes with a comparable number of CDS in the genomes (i.e., 4-5,000 CDS; average from 12 genomes), the Bcc conserved gene core reveals an excess of metabolism genes, particularly genes involved in metabolism and transport of amino-acids, carbohydrates and ions, and regulation genes (Figure 2.10). These results are in good agreement with the exceptional metabolic and ecological versatility that characterizes Bcc relative to other bacterial species and reveal universal trends in genome expansion for Bcc species.

The genomic comparisons also revealed that the pool of genes unique to each strain (strain-specific) is significantly large, accounting for ~1,200 genes in the clinical strain J2315 and reaching 1,400 to 2,500 genes in the two environmental strains G4 and ATCC 17760, respectively (Figure 2.9). These results reveal a surprising level of genetic diversity within the Bcc given that these species are so closely related that their distinction is frequently difficult by conventional means. The majority of these strain-specific genes have hypothetical or poorly characterized function (i.e. with very low similarity to genes in public databases), which indicates that many functions in Bcc remain undiscovered (Figure 2.9).

Nonetheless, a substantial fraction of the strain-specific genes can be assigned to a well-characterized biological function and we have further investigated this set of genes in order to get insight into what drives genome expansion within each strain and identify traits that are important in different ecological niches. Our results show that these strain-specific genes are closely associated with the known ecological properties for each strain. For example, G4 is a successful root colonizer and degrader of pollutants and the G4-

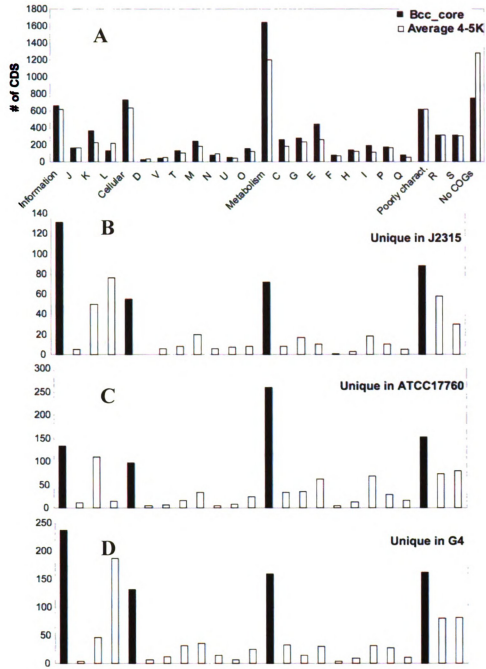


Figure 2.10. Functional annotation of the conserved gene core and the strain-specific genes for the three sequenced Bcc genomes. Bars represent the number of genes assignable to the four major classes (full description on x-axis) and the individual categories of COG database (single-letter description on x-axis; for annotation of the letters see Table 2.1). **(A)** Solid bars represent the conserved gene core between the three available Bcc genomes, while open bars represent the average from all genomes available in GenBank, which have a comparable number of protein coding genes (4-5,000) to the conserved core of the sequenced Bcc genomes. Panels **B**, **C**, and **D** show the annotation of the strain-specific genes for J2315, ATCC 17760 and G4's genomes, respectively. Designations for each functional category have been omitted from x-axes for simplicity.

specific set mostly involves metabolism genes such as oxidoreductases, oxygenases, cytochrome-flavoproteins and transport genes, which are presumably related to aromatic and poly-chlorinated compound degradation (Figure 2.10). The majority of the ATCC-17760-specific genes are also involved in metabolism but the specific functions enriched are rather different from the ones identified for G4. For instance, ATCC 17760 has many unique genes for sugar and carbohydrate metabolism and transport such as acetyl-transferases, oxidoreductases, and lyases, several large gene clusters for polyketide (antibiotics) such as phenazine production, and excreted Fe (III) binding proteins. These genes may explain ATCC 17760 as a successful soil colonizer.

The G4-specific gene set also includes a plethora of mobile elements, e.g., transposase and prophage-like elements. Interestingly, the only other *Burkholderia* strain that includes a comparable number of mobile elements is *B. xenovorans* str. LB400, which is the best-known Poly Chlorinated Biphenyl (PCB) degrader. In fact, many of the G4-mobile elements are conserved in LB400 and not conserved in any of the more closely related Bcc strains. It follows that these mobile elements may be an important trait in biodegradation settings. In such settings, bacteria typically encounter a variety of different pollutant compounds (rather than a single substrate) and hence genomic plasticity and potential for diversification may be more important traits than cell stability and fitness since some of these mobile elements consume resources and could be lethal for the cell when activated.

Chromosomal biases in terms of genetic diversity.

We further examined the set of strain-specific genes to gain a better understanding of how genetic diversity is created in Bcc species. Analysis based on the J2315 genome, which is closed and facilitates analysis, reveals that the amount of genes of unknown function is biased towards the smaller chromosomes. For instance, 21,1%, 31,7% and 34.4% percent of the genes in the largest, medium and smallest chromosome, respectively, can not be assigned to the COGs database and hence, have a hypothetical function (Figure 2.11). When we examined how conserved the genes of each chromosome are in the other two Bcc genomes we noted a similar trend, i.e., the smaller chromosomes harbor more of the J2315's specific-genes. For example, only 50% of the genes in the smallest chromosome have homologs in ATCC 17760 or G4 as opposed to >70% for the large chromosome (Figure 2.11).

Further, about half of the J2315-specific genes have a GC% content that is >5% different from the average of the J2315's genome, suggesting a horizontal acquisition of a large fraction of the strain-specific CDS. Interestingly, the majority of the J2315's CDS with a GC% <5% than the average of the J2315's genome are also J2315-specific whereas this is less pronounce for J2315's CDS with a GC% >5 of the average (compare gray with white bars in Figure 2.11), indicating that horizontal transfer is more frequent from low GC than high GC donors. Comparable results were noted when ATCC17760 or G4 were used as the reference genome instead of J2315 (data not shown). These findings suggest that each chromosome in Bcc species has a different evolutionary history and perhaps origin and may indicate that the Bcc species may have a mechanism to control where the diversity is created in the genome. Further, these findings show that substantial genome evolution and gene turnover take place within very short evolutionary scales,

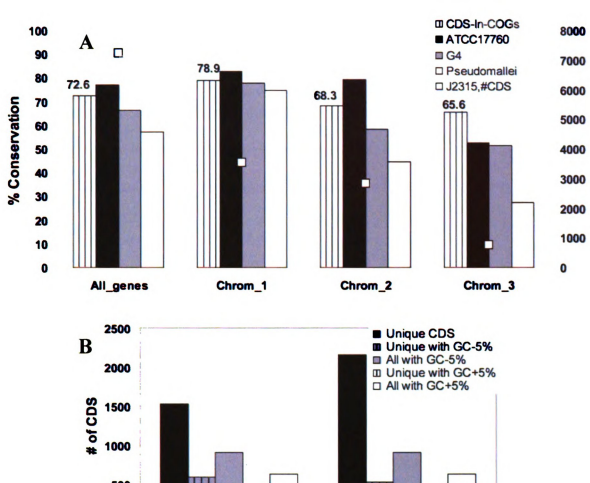


Figure 2.11. Biased in the amount a genetic diversity carried by each chromosome and the GC% composition of the genes that are different between Bcc genomes. (A) Striped bars represent the percent of genes in each chromosome of strain J2315, which are assignable to the COG databases, while the remaining bars show what fraction of the genes in each chromosome is conserved in the other available *Burkholderia* genomes (graph legend). Centered open squares show the number of genes in each chromosome (right y-axis) while the leftmost bars show the same values as above for all genes in the genome (i.e., the average). **(B)** Black bars represent the total number of J2315's CDS that, based on pair-wise whole-genome comparisons, do not have homologs (i.e., they are J2315-specific) in the other Bcc strain (x-axis), while gray and open striped bars represent the fraction of these J2315-specific CDS that has a GC content <5% and >5% than the average of the J2315's genome, respectively. Gray and white bars represent the total number of CDS in J2315's genome that have a GC content <5% or >5% than the average of J2315's genome, respectively.

presumably, as a result of the interaction between the organism and particular habitat(s), and similarly to results reported previously for all bacterial genomes and longer evolutionary scales.

ACKNOWLEDGMENTS

We thank Tom Schmidt, Rebecca Grumet, Joel Klappenbach, Frank Larimer, and an anonymous reviewer for helpful discussions regarding the manuscript. This work was supported by the Bouyoukos Fellowship Program (K.T.K.), the U.S. Department of Energy's Microbial Genome Program, and the Center for Microbial Ecology.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-402.

2. **Andersson, S. G., and C. G. Kurland.** 1998. Reductive evolution of resident genomes. Trends Microbiol **6:**263-8.

3. **Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland.** 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature **396:**133-40.

4. **Balandreau, J., V. Viallard, B. Cournoyer, T. Coenye, S. Laevens, and P. Vandamme.** 2001. *Burkholderia cepacia* genomovar III Is a common plant-associated bacterium. Appl Environ Microbiol **67:**982-5.

5. **Bentley, S. D., K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood.** 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417:**141-7.

6. **Branscomb, E., and P. Predki.** 2002. On the high value of low standards. J Bacteriol **184:**6406-9; discussion 6409.

7. **Coenye, T., and P. Vandamme.** 2003. Diversity and significance of *Burkholderia* species occupying diverse ecological niches. Environ Microbiol **5:**719-29.

8. **Coenye, T., P. Vandamme, J. R. Govan, and J. J. LiPuma.** 2001. Taxonomy and identification of the *Burkholderia cepacia* complex. J Clin Microbiol **39:**3427-36.

9. **Eisen, J. A.** 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res **8:**163-7.

10. **Frank, A. C., H. Amiri, and S. G. Andersson.** 2002. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. Genetica **115:**1-12.

11. **Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, and et al.** 1995. The minimal gene complement of *Mycoplasma genitalium*. Science **270:**397-403.

12. **Galperin, M. Y., and E. V. Koonin.** 1999. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. Genetica **106:**159-70.

13. **Gerlt, J. A., and P. C. Babbitt.** 2001. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem **70:**209-46.

14. **Graham, D. E., R. Overbeek, G. J. Olsen, and C. R. Woese.** 2000. An archaeal genomic signature. Proc Natl Acad Sci U S A **97:**3304-8.

15. **Grey, T., and S. Willimas.** 1971. Microbial productivity in soil. Symposia of the Society for General Microbiology **21:**255-286.

16. **Jordan, I. K., K. S. Makarova, J. L. Spouge, Y. I. Wolf, and E. V. Koonin.** 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res **11:**555-65.

17. **Kanehisa, M., and S. Goto.** 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28:**27-30.

18. **Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt.** 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol **66:**1328-33.

19. **Konig, H.** 1988. Archaeobacterial cell envelope. Can. J. Microbiol. **34:**395-406.

20. **Lessie, T., and T. Gaffney.** 1986. The Bacteria: A Treatise on Structure and Function. Academic, New York.

21. **Lessie, T. G., W. Hendrickson, B. D. Manning, and R. Devereux.** 1996. Genomic complexity and plasticity of *Burkholderia cepacia*. FEMS Microbiol Lett **144:**117-28.

22. **Mira, A., H. Ochman, and N. A. Moran.** 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet **17:**589-96.

23. **Mitsui, H., K. Gorlach, H. Lee, R. Hattori, and T. Hattori.** 1997. Incubation time and media requirements of culturable bacteria from different phylogenetic groups. J. Microbiolog. Methods. **30:**103-110.

24. **Moran, N. A.** 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell **108:**583-6.

25. **Nelson, K. E., I. T. Paulsen, J. F. Heidelberg, and C. M. Fraser.** 2000. Status of genome projects for nonpathogenic bacteria and archaea. Nat Biotechnol **18:**1049-54.

26. **Parke, J. L., and D. Gurian-Sherman.** 2001. Diversity of the *Burkholderia cepacia* complex and implications for risk assessment of biological control strains. Annu Rev Phytopathol **39:**225-58.

27. **Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White.** 2001. The Comprehensive Microbial Resource. Nucleic Acids Res **29:**123-5.

28. **Rost, B.** 1999. Twilight zone of protein sequence alignments. Protein Eng **12:**85-94.

29. **Saito, A., H. Mitsui, R. Hattori, K. Minamisawa, and T. Hattori.** 1998. Slow-growing and oligotrophic soil bacteria phylogenetically close to *Bradyrhizobium japonicum*. FEMS Microb. Ecol. **25:**277-286.

30. **Sander, C., and R. Schneider.** 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins **9:**56-68.

31. **Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa.** 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp.* APS. Nature **407:**81-6.

32. **Stanier, R. Y., N. J. Palleroni, and M. Doudoroff.** 1966. The aerobic pseudomonads: a taxonomic study. J Gen Microbiol **43:**159-271.

33. **Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson.** 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature **406:**959-64.

34. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

35. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-7.

CHAPTER 3

GENOMIC INSIGHTS THAT ADVANCE THE SPECIES CONCEPT FOR PROKARYOTES

INTRODUCTION

The species concept for Prokaryotes remains a highly controversial and unsettled issue, and as a result a number of different concepts exist at present. The most popular of these concepts is, by far, the one proposed by Wayne et al in 1987, which considers a bacterial species to be essentially "a collection of strains that are characterized by at least one diagnostic phenotypic trait and whose purified DNA molecules show at least 70% reassociation homology (DNA homology)" (28, 32, 39). This species definition, while pragmatic and universally applicable within the prokaryotic world, remains controversial because it is difficult to implement due to technological limitations in identifying diagnostic traits and in performing the pairwise DNA hybridizations, is based on a 30-year old arbitrary standard, is not encompassed by any of the eukaryotic species concepts, and is too often not adequately predictive of phenotype (6, 7, 38). Indeed, applying this standard to eukaryotic species would lead to the inclusion of members of many taxonomic tribes in the same species, e.g. all the primates should then belong to the same species (29, 30). Accordingly, there are only about 4,500 prokaryotic species described to date (12), which contrasts to well over 1 million eukaryotic species and yet the prokaryotes have been exploring evolutionary adaptations at least 100 times longer. Furthermore, several theoretical (7) and ecological (38) approaches to define prokaryotic species favor a more natural definition as opposed to the current definition. Last, several strains that show higher than 70% DNA homology are classified into different species, even different genera, usually on the basis of pathogenicity or host range, such as strains of *E. coli* and *Shigella* spp. (5), making the application of the 70% DNA homology standard anecdotal.

To gain insight into these issues, we have performed pair-wise, whole genome comparisons between all closely related (showing >94% 16S rRNA identity), sequenced bacterial strains (64 strains) to determine the conserved protein-coding genes (CDS) between the pair of strains as well as the strain-specific genes and study how these parameters correlate with the evolutionary distance between the strains and the strain assignment to species. This analysis is most informative with respect to the species definition because it concerns genes that largely determine the organism's phenotype. Further, our strain set represents several major bacterial lineages, including α and β Proteobacteria, low GC gram-positive Bacilli, Streptococci, and Staphylococci, and high GC gram-positive Mycobacteria, which allows for robust interpretations (see Table 3.1 in Appendix). We found that strains of the same species can vary up to 30% in gene content raising questions as to whether they should belong to the same species, while a more stringent definition for species, which should also consider the ecology of the strain, is both more appropriate and plausible.

MATERIAL AND METHODS

Sixty-four fully sequenced and closely related genomes were used in this study (Table 3.1 in Appendix). The genomic sequences and sequence annotation for 54 of the 64 closed genomes, which were published at the time of this study (May 2003), were obtained from NCBI's ftp site at ftp://ftp.ncbi.nih.gov/. The remaining 11 genomes were closed at the time of this study; however their annotation was not completed (denoted by NA in Table 3.1). These 11 strains were: *S. bognori* 12419, *Y enterocolitica*, *E. carotovora*, *N. meningitidis* FAM, *S. aureus* MSSR476, and *S. aureus* MRSA252, produced by the Sanger Center and were obtained through the Sanger ftp site at ftp://ftp.sanger.ac.uk/pub/; and *M. avium*, *S. epidermitidis* RP62A, and *C. perfrigens* ATCC 13124, produced by The Institute for Genomic Research (TIGR) and obtained through their website at http://www.tigr.org. *N. gonorrhoeae* FA1090 was produced at the Advanced Center for Genome Technology at the University of Oklahoma (available at http://www.genome.ou.edu/gono.html).

Determination of conserved genes and evolutionary relatedness.

The conserved genes between a pair of genomes were determined by whole-genome sequence comparisons using the BLAST algorithm release 2.2.5 (2). For these pair-wise comparisons, all CDS sequences from one genome (hereafter "the reference" genome) were searched against the genomic sequence of a closely related genome (hereafter "the tester" genome). CDS from the reference genome were considered conserved when they that had a BLAST match of at least 60% overall sequence identity (recalculated to an identity along the entire sequence) and an alignable region more than

70% of their length (nucleotide level) in the tester genome, whereas CDS that had no match or a match below this cut-off were considered "unique" (or genome-specific) in the reference genome. A reciprocal best match approach was also employed to determine what fraction of the previously determined conserved genes is orthologous. The BLAST was run with the following settings: X = 150 (drop-off value for gapped alignment), q = -1 (penalty for nucleotide mismatch), and F = F (Filter for repeated sequences), the rest of the parameters were at default settings. These settings give better sensitivity with more distantly related genomes compared to default settings, because the default settings target more highly identical sequences. The genomes that were used as reference genomes, the genome sizes and total number of CDS for all genomes used is this study, as well as the raw data from the pair-wise comparisons are summarized in Table 3.1 of Appendix.

Searching for the gene function (i.e. amino acid level) predicted more conserved genes than the nt. level search only when the evaluated strains show less than 97% 16S rRNA sequence identity. This, however, did not affect anything more than a slight up-shifting on the left part of the regression line in Figure 3.4*A* of the article. Further, the usage of less stringent cut-offs for the determination of conserved sequences did not significantly differentiate our final conclusions (data not shown). Last, the use of a cut-off for match length and identity without manual inspection of the alignments proved highly accurate for the prediction of conserved sequences. For instance, Parkhill and coworkers (26) have identified 4,297 and 3,394 CDS of *B. bronchiseptica* RB50 to have orthologs in *B. parapertussis* and *B. pertussis*, respectively whereas our approach predicted 4,261 and 3,382 CDS for the same comparison, respectively.

The evolutionary distance between a pair of strains was measured by the average nucleotide identity (ANI) of all conserved genes between the strains as computed by the BLAST algorithm. Duplicated genes within a genome were defined as the genes that had a better match within their genome than in another genome during a pair-wise whole-genome comparison, using, in all cases, a minimum cut-off for a match 60% identity over at least 70% of length of the query gene. Despite the use of the rather stringent cut-off in these comparisons, cases of independent acquisition of very similar genes (instead of gene duplication) cannot be excluded.

Determination of DNA homology and 16S rRNA gene sequence identity.

DNA homologies between species were obtained from the literature (5, 16, 18, 34, 37, 41). When the sequenced strains were the same as the ones used in the DNA homology experiments, we directly compared the DNA homology values with the ANI of the sequenced genomes. When the strains were different (the majority of cases), we used the average DNA homology values (or ANI) for several strains of the same species for the comparisons.

The 16S rRNA sequence identity between strains was determined as the average identity between all copies of the 16S rRNA gene the strains possess. 16S rRNAsequence identity was determined using the Phylip package with Kimura 2-parameter method, available online at the tools of the Ribosomal Database Project (http://rdp.cme.msu.edu/cgis/phylip.cgi) (8).

CDS functional annotation and intergenic regions.

We obtained more high-level annotation (compared to the one found in the GenBank files) of the CDS in the reference genome using the twenty functional categories in the recently updated Cluster of Ortholgous Genes (COG) database (33). Each COG functional category represents a major cellular process, like transcription, signal transduction etc. However, because several of our reference genomes were not incorporated into the COG database, we performed our own CDS assignment to the COG database as described previously (20). For the genomes that were already incorporated in the COG database, our assignments were more than 99% consistent with those already in the COG database.

CDS that were assignable to the COG database and were not associated with phage or transposase elements were denoted as well-characterized genes. Hypothetical genes were defined in this study as the genes that were not assignable to COG database and were annotated as hypothetical or unknown function in the primary annotation (GenBank files), including hypothetical genes carried by phages. This category included the majority (>50%) of the genes not assignable to COG database and consisted between 10-20% of the total number of annotated genes in a genome. Genes that were annotated as hypothetical in the primary annotation and were assignable to COG conserved hypothetical or other category were considered "conserved hypothetical" (and well-characterized) and denoted as such in the article.

The non-coding sequences between the annotated protein-coding (CDS) and RNA genes of the reference genomes were extracted from the GenBank files, after removing 100 bases upstream of the start site of the downstream gene to avoid any selection on the promoter of gene. These intergenic sequences, when longer than 100 nt, were searched

against the whole genomic sequence of the tester genomes, as described previously for CDS, to determine whether they are conserved in the tester genomes. Removing a longer fragment than 100 bases upstream of the start site did not significantly affected our conclusions (data not shown).

PERL scripts were used to edit CDS assignments where necessary; extracting sequences from GenBank files; formatting databases for BLAST searches, and automatically parsing BLAST outputs.

RESULTS AND DISCUSSION

For our purposes there was need for precise measurement of the evolutionary distance between closely related strains and particularly between strains of the same species. We noticed that the average nucleotide identity (ANI) of all conserved genes (typically >1,500 genes) between two strains strongly correlated with the reported DNA-DNA reassociation homologies between the same strains (Figure 3.1*B*). Based on these results, the 70% DNA-DNA homology standard corresponds to about 93-94% ANI, which roughly agrees with previous experimental evidence (reviewed in (13). Therefore, strains that show higher than 94% ANI should belong to the same species according to

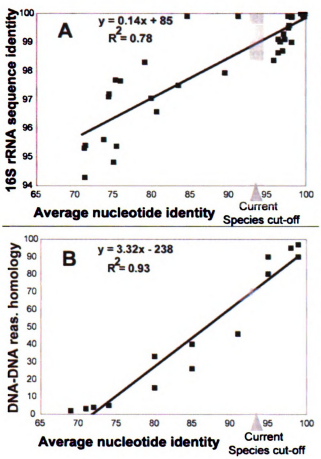


Figure 3.1. Relationships between average nucleotide identity (ANI), 16S rRNA sequence identity and DNA homology. Each dot represents the ANI of all conserved genes between two strains plotted against the 16S rRNA sequence identity (**A**) and the DNA homology (**B**) of the two strains. The shaded bar represents 93-94% ANI, which approximately corresponds to 70% DNA homology, i.e., the species cut-off for prokaryotic species, according to the regression analysis in panel **B**. 16S rRNA identity and DNA homology values were computed as described in methods section.

the DNA homology standard. This was also confirmed by the fact that all strains in our set that reside in the same species or in species that show higher than 70% DNA-DNA homology showed higher than 94% ANI. Furthermore, the ANI strongly correlated with the 16S rRNA sequence identity but gave higher resolution, since a 0-5% 16S rRNA sequence miss-pairing is spread between 0-30% average nucleotide miss-pairing (Figure 3.1*A*). In summary, the strong correlations observed as well as the large number of genes used in the calculations suggest that ANI represents a robust measure of evolutionary distance, which should not be affected by lateral transfer or varied recombination rates of single (or a few) genes and offers resolution at the subspecies level where 16S rRNA gene or other single markers are not useful.

Conserved gene core and genetic diversity within species.

Using the 94% ANI criterion for strain assignment to species, we first attempted to evaluate the extent of genetic diversity within a single bacterial species. Our results for *E. coli*, the best sampled species with genomic sequences, show that when a strain showed less than 98-99% nt. identity to all eight remaining strains, it had a sizeable number of sequences, ranging between 5-15% of the total CDS in the genome, that could not be identified in any of the remaining strains. At the same time and as expected, strains that showed at least 99% nt. identity to any of the remaining eight strains had a small (<1-2%) number of unique sequences such as the two strains of the *E. coli* O157 lineage or the two strains of the *S. flexneri* 2a lineage. Accordingly, the number of unique genes in all nine genomes together clearly exceeds 8,000, with the trendline suggesting that a continued increase is expected with the sequencing of new genomes of the species

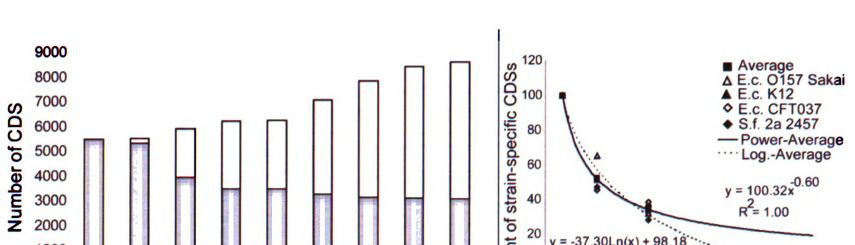


Figure 3.2. Conserved gene core vs. genetic diversity within *E. coli* species. (A) Starting with the 5,447 CDS in the genome of *E. coli* O157 strain Sakai the next bar to the right represents how many unique CDS in total are found in strain EDL and Sakai together (empty bars) and how many of the 5,447 CDS are conserved in EDL (filled bars) etc. Hence, the empty bars represent the total genetic diversity within species whereas the filled ones represent the conserved core for the species. **(B)** All CDS in a strain (graph label) were searched against a database of an increasing number of genomes. The number of strain-specific CDS, expressed as a percentage of the strain-specific CDS when only one genome was used as database, is plotted against the number of genomes used as database. The almost identical genomes of *E. coli* O157 and *S. flexneri* 2a lineages were pooled together so that the seven genomes finally compared showed similar average nucleotide identity between each other. The genomes of *S. sonnei*, *E. coli* str. 042 and str. E2348 were not annotated at the time of this study. For these genomes, the genomic sequence was cut in 1,000nt long consecutive fragments and these fragments were used instead of CDS. Applying this strategy to annotated genomes gave comparable results to the ones obtained using annotated CDS. The logarithmic and power correlations shown are not statistically different from each other.

(Figure 3.2). On the other hand, the conserved gene core between all nine genomes is only 3,050 genes, which is about half of the genes that most strains of the species possess. Results from seven genomes of *S. enterica* and the five genomes of the Gram-positive *Staphylococcus aureus* indicate that other species may show extensive genetic

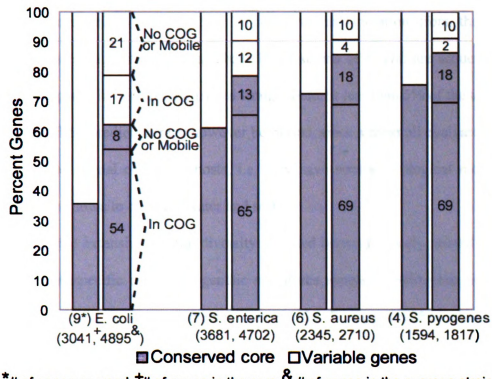


Figure 3.3. Conserved gene core vs. genetic diversity of species. The first column shows what fraction of the total, non-redundant list of genes found in all genomes of the species belongs to the species' conserved core and what fraction is variable (i.e., not in the core). The second column shows the same distribution for the "average" strain of the species. The functional annotation of the genes in the average strain of the species is also shown as exemplified for *E. coli*. *E. coli* shows the greatest and *S. pyogenes* the lowest genetic diversity; note, however, that *E. coli* genomes are generally more distantly related between each other compared to genomes of the other species based on ANI measurements (ANI between *E. coli* genomes ~96-97% vs. >98% for the others).

diversity as well (Figure 3.3).

We also attempted to predict whether the genetic diversity within the *E. coli-Shigella* spp. species would be exhaustible with additional sequenced strains by searching all genes in a strain against a database of an increasing number of genomes. While the number of novel genes in a strain declines with greater coverage of the species with genomic sequences, the number of available genomic sequences is still too limited to predict how many strains would need to be sequenced to discover most of the gene diversity of the species (Figure 3.2*B*). Nonetheless, extrapolation from the current genomic sequences suggests that when about 12-14 strains of *E. coli* are sequenced, the amount of new genes in the next sequenced strain would be less than 5% of the total CDS in the genome. This prediction may however be biased, since almost all evaluated strains are pathogens of animal or human hosts, i.e. they have similar ecological niches, and some *E. coli* are known to colonize water and soil (1).

Despite the extensive genetic diversity revealed between closely related bacteria, however, species-specific diagnostic genetic signatures appear to exist, thus, it appears that it is meaningful to have a species concept for Prokaryotes. For example, by comparing the nine *E. coli-Shigella* spp. genomes against the seven genomes of *Salmonella enterica* (a close relative of *E. coli,* ANI between *E. coli* and *Salmonella spp.* genomes is ~80%), we identified ~300 genes, i.e. ~6% of the total genes, in any *E. coli-Shigella* spp. strain that are not conserved in any *S. enterica* genome whereas, the reverse comparison revealed ~12% of the genes to be *S. enterica*-specific. About half of the genes in these signatures are related to traits that are known to differentiate *E.coli-Shigella* spp. from *S. enterica* species; for instance, the *E.coli/Shigella* contain about 80 genes involved in transport and metabolism of sugars, amino acids and oligopeptides,

which is consistent with this species growth on sucrose and production of indole from tryptophan, whereas *S. enterica* can do neither (5). Likewise, the *S. enterica* signature included genes for growth on hydrogen sulfide, which is not used by *E.coli/Shigella spp.* (5). The other half of the genetic signatures involves genes not assignable to COGs or of general function prediction only, which may yield even more distinguishing phenotypic traits.

The current species definition appears to be too liberal.

We then studied how the amount of conserved genes between two strains correlated with their evolutionary relatedness for all 64 strains compared in this study. Conserved genes were expressed as percentage of the total CDS in the reference genome to normalize for the genome size effect. Our results suggest that there is strong correlation between these two parameters over longer evolutionary distances, i.e. corresponding to 0-5% 16S rRNA miss-pairing, and this correlation appears to be consistent among several major bacterial lineages (Figure 3.4*B*). However, when the analysis was restricted to strains that show >94% ANI, i.e., they should belong to the same species, this correlation collapsed (Figure 3.4*A*). According to this dataset, strains of the same species frequently differ in up to 30% of their total genes, and of these up to 50% are well-characterized genes. Well-characterized denotes genes that are assignable to the Cluster of Orthologous Groups (COG) database and are not associated with phage, or transposase elements whose significance on the cell phenotype remains largely unexplored. When a reciprocal best match approach was employed to determine the orthologous fraction of the conserved genes in an effort for a more conservative

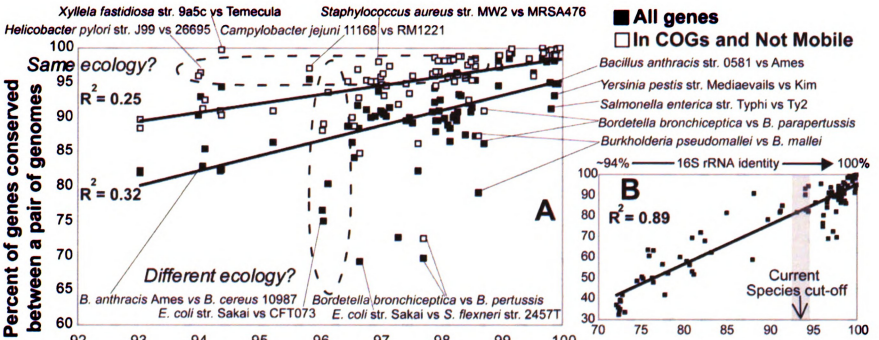


Figure 3.4. Correlation between conserved genes and evolutionary distance for bacterial species. Each datapoint represents the percent of conserved genes between two strains plotted against their evolutionary distance, measured as average nucleotide identity (ANI) of all conserved genes between the strains. Solid squares represent all genes while open squares represent the fraction of all genes that are well-characterized genes (see methods section). Panel **A** includes only pairs of strains that should belong in the same species according to the current species definition standard (see Figure 3.1), whereas panel **B** includes pairs of more distantly related strains.

estimation of functional similarity, then the gene differences were even higher (but generally not considerably higher) by an average of 1.12% (STDEV 1.15, MAX 6.78%). To extend the comparison to higher organisms, only about 25% of the human genes do not have homologs in the distantly related fish genome, *Fugu rubripes* (3), while the ANI between humans and chimpanzees is 98.7% (10) i.e., much higher than the current standard for prokaryotic species. Therefore, the genetic differences we find among several strains of the same bacterial species are extensive when viewed from a eukaryotic perspective.

We also noticed that pairs of strains that presumably have an overlapping ecological niche, like *Xyllela fastidiosa* and *Helicobacter pylori* strains that cause the same disease in closely related plant species and humans, respectively (11, 36), have more genes conserved relative to pairs of strains that show a comparable evolutionary relatedness but presumably have non-overlapping ecological niches, like *E. coli* strains that cause different diseases in humans, i.e., enterohemorrhagic vs. uropathogenic (40) (the dashed circles in Figure 3.4*A* represent graphically this point). The former cases typically involved obligatory pathogens with small genome sizes whereas the latter involve free-living or opportunistic pathogens with large genomes. Species with larger genomes are thought to be more ecologically versatile (20), which is consistent with the previous interpretations. Further, sexual isolation is more pronounced in the former species due to restrictions in their dispersion as is documented by *Helicobacter pylori* biogeography (11), which may explain why strains of these species show substantial nucleotide divergence while sharing a nearly identical gene content.

In summary, our results (Figures 3.3 & 3.4) show that the current species definition results in too much genetic diversity within species and hence a more stringent definition is needed if species should be reasonably predictive of the phenotype and ecological potential of the organism. For example, a species definition, which includes only strains that show at least ~99% ANI or less than 99% ANI but share a common ecology, would be consistent with this goal because such strains should have minimum (i.e., <5%) gene differences (Figure 3.4*A*). Several additional independent lines of evidence support that a species definition based on these principles may be more appropriate than the current one.

First, genetic signatures, like the ones described previously between *E. coli-Shigella* spp. and *S. enterica* genomes, are identifiable among some groups of strains that show between 94% and 99% ANI. For example, the two pathogenic genomes of the *S. enterica* pathovar Typhi share ~325 genes that are not conserved in any of the three pathovar Typhimurium, str. PT2 and *S. gallinarum* str. 287/91 genomes (ANI between the Typhi genomes is >99%, between Typhi genomes and others 97-98.5%) (Figure 3.5*A*). Many of the Typhi-specific genes are potential pathogenicity factors, such as fimbrial and exported polysaccharide gene clusters, further supporting the ecological importance of this genetic signature. These extensive gene differences may also indicate that Typhi strains do not directly compete with the other *S. enterica* strains *in-situ* (i.e., they exploit a different ecological niche) otherwise the genetic differences should be purged by natural selection. The lack of competition between two populations is considered strong evidence towards describing the populations as different species by several prokaryotic taxonomists (7, 38). A similar comparison revealed ~4% of the genes

to be Typhimurium-specific, while comparable results were obtained for other groups with several sequenced representatives, such as the *Listeria monocytogenes* and *Neisseria* spp. Importantly, the *E. coli-Shigella* spp. and *S. enterica* genomes compared previously are much more distantly related (i.e. ~80% ANI) than the genomes compared here, nonetheless, the genetic signatures revealed are comparable in size.

Second, in at least two cases in our dataset we could not identify species-specific genetic signatures when applying the current definition. For instance, there are two strains of *Bacillus cereus* fully sequenced, str. ATCC 10987 and ATCC 14579, with the former showing ~94% ANI to the *B. anthracis* strains (thus, albeit marginally, str. 10987 should belong to the same species with *B. anthracis* according to the DNA homology standard) and the latter only ~91% (ANI between the two *B. cereus* genomes is 91.2%) (Figure 3.5*B*). Str. 14579 however, has more genes conserved with the *B. anthracis* genomes than str. 10987, and no genetic signature is identifiable for the *B. anthracis*-str. 10987 group. Such instances prove that the current standard is rather arbitrary and suggest that any species definition (like the DNA homology) that does not consider the ecology of the strains in addition to their genetic relatedness is problematic. This is also evident by the low correlation observed between conserved gene content and evolutionary distance over a short evolution scale (Figure 3.4*A*).

Last, gene expression, which is another important determinant of organism's phenotype apart from gene presence (10, 25), is likely to be different between strains that show a substantial number of nucleotide substitutions, like between strains that show 94-97% ANI. Notably, about half of the nucleotide substitutions between such strains cause non-synonymous amino acids substitutions in our dataset.

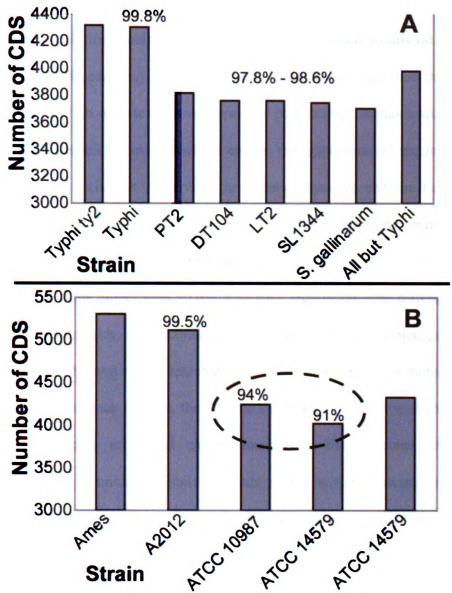


Figure 3.5. Genetic signatures among groups of strains that show higher than 94% average nucleotide identity (ANI). Starting with all CDS in the leftmost strain the next bar to the right represents how many CDS are conserved in the next strain (x-axis) (similarly to Figure 3.2). The ANI to the leftmost strain is also shown on the top of the bars for each strain. **(A)** A genetic signature between the pathovar Typhi strains and the rest *Salmonella* strains is identifiable. **(B)** No genetic signature is evident for the *B. anthracis-B. cereus* ATCC14579 group (dashed circle). The rightmost bar in panel **B** shows how many of the conserved CDS between the two *B. anthracis* strains are also conserved in strain ATCC14579 alone. Strains from left are: (**A**) *S. enterica* ser. Typhi Ty2, *S. enterica* ser. Typhi Typhi, *S. enterica* PT2, *S. enterica* ser. Typhimurium DT104, *S. enterica* ser. Typhimurium LT2, *S. enterica* ser. Typhimurium SL1344, *S. gallinarum,* and a pool of all *Salmonella* but the Typhi strains. **(B)** *B anthracis* Ames, *B anthracis* A2012, *B cereus* ATCC 10987, and *B cereus* ATCC 14579.

What is an ecotype?

If one is to define species as a collection of very similar strains (at the nt. level and/or the number of genes they share) as proposed here, then the question that remains is what is an ecotype? In my view, an ecotype is a population that has acquired a small number of extra genetic elements, which enable the population to exploit a slightly different ecological niche but preserving the genetic signature and the full ecological potential that characterizes its species. Such ecotypes do exist among strains that show higher than 99% ANI. For example, several *Bacillus anthracis* or *S. enterica* pathovar Typhi strains that show higher than 99.6% ANI have significant gene differences, which primarily involve plasmids, and secondary phage and transposase-related genes (Figure 3.4*A*). These plasmids have been connected to a strain's ability to cause increased disease symptoms (see for instance 15), i.e., they enable the strains to exploit a slightly different but highly overlapping ecological niche compared to their species. Such genetic differences borne as plasmids or mobile elements cannot be viewed as genetic signatures that justify a description as a new species because they are not stable properties of the genome. Moreover, otherwise identical populations that acquire a small number of beneficial mutations that enable the population to exploit a new substrate, like the parallel evolving *E. coli* strains founded from the same ancestor (35), can also be viewed as ecotypes of the same species.

There are a few, more complicated cases with respect to speciation in our dataset, which can be exemplified by the three pathogenic *Bordetella* spp. genomes. These organisms, which are colonizers of the respiratory tracts of mammals, show 97.8-98.7% ANI between each other's genomes and it appears that *B. pertussis* and *B. parapertussis*

have evolved by a (considerable) genome reduction from a *B. bronchiseptica*-like ancestor; presumably as a result of population bottlenecks or ecological specialization since these genomes show increased host-specificity compared to *B. bronchoseptica* (26) (see Figure 3.4*A*). However, no clear and ecologically meaningful genetic signature is identifiable for *B. pertussis* or *B. parapertussis* to justify their description as separate species, since the genes specific to these two genomes are limited or of hypothetical and/or transposase function. Viewing these genomes as ecotypes of *B. bronchiseptica* would deviate from the proposed rule that an ecotype should preserve the full potential of its species since *B. bronchiseptica* has at least 600 additional genes compared to *B. pertussis* or *B. parapertussis*. One possibility is that the latter genomes represent snapshots of an active speciation process, which might have not yet reached the stage of a diagnosable species-specific genetic signature. Alternatively, such instances indicate that some species are likely to show a continuum/gradient of genetic diversity rather than defined boundaries diagnosable by species-specific genetic signatures or that one should look for species-specific signatures at a different level e.g., the gene expression level or deletion (instead of acquisition) of specific pathways in order to achieve ecological specialization. Last, the *Bordetella* spp. example indicates that species might be found even among strains that show higher than 99% ANI if the populations have undergone major ecological constrains.

Functional biases in the genome-specific genes.

The functional annotation of the genes that constitute the genome-specific genes in all the pair-wise comparisons between the 64 strains used in this study was also

evaluated to provide insights into the factors that might foster speciation. We found that hypothetical, phage and transposase associated genes comprise 62.4% of the genome-specific genes, with the hypothetical genes comprising the majority, 40.4%; the former percentage becomes even larger, 66.1%, when the analysis is restricted to strains of the same species (Figure 3.6). Hypothetical denotes genes that are not assignable to the COG database and are annotated as hypothetical or unknown function in the primary annotation, while phage genes include all genes (assignable or not to COG) carried by phage genomes (see methods section). The former results contrast with an average of 31.1% of hypothetical, phage and transposase related genes in a typical genome (average from 64 genomes) indicating that hypothetical, phage and transposase related genes might play a more important role in the speciation process than expected based on the frequency at which these genes are encountered in the genome. These genes are, however, largely species- or genome-specific (see also Figure 3.3), which reveals a weak positive selection for these functions and reflects the enormous genetic diversity that characterizes bacteriophages (40-80% of the total genes in a phage genome are

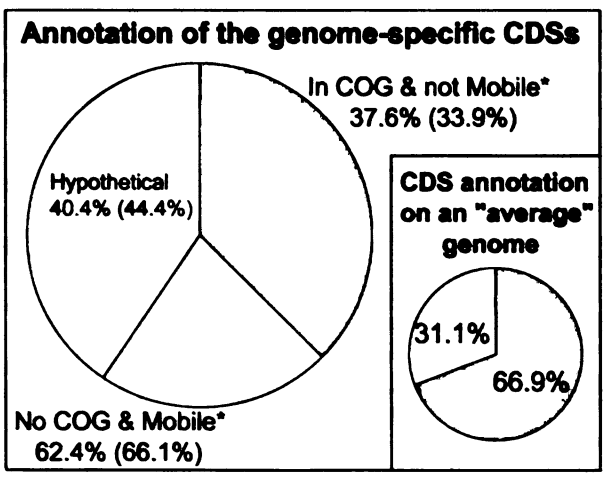


Figure 3.6. Functional distribution of genome-specific CDS from 82 pair-wise, whole-genome comparisons. Results using only strains showing >94 ANI are shown in parentheses. (Inset) Mean functional distribution of annotated CDSs for the 64 genomes deposited in GenBank as of October 2003. *Mobile denotes phage or transposase associated genes.

hypothetical in our dataset) (27) and insertion/transposase elements (23). Collectively, this information is congruent with phage and mobile elements being ephemeral intruders of the genome and have little, if any, value for the cell but occasionally might be important, e.g. when carrying ecologically important genes, and lead to speciation (for examples see (4).

The fraction of the genome-specific genes that is well characterized is, on average, 37.6%, which contrast with an average of 69.9% of such genes in a typical genome (Figure 3.6). Restriction of the analysis to ortholgous genes (i.e. reciprocal best match vs. one-way match approach) did not significantly affect these results. Last, gene duplication appears to play a significant but not major role in the genetic diversity within species. The occurrence of duplicated genes among the genome-specific genes during comparisons of strains of the same species ranged from <1-30% and this variation appeared to be species-dependent.

During the functional annotation of the genome-specific genes, we noted that hypothetical CDS are approximately as conserved as the intergenic sequences, i.e. the fraction of sequences that remain conserved with increasing evolutionary distance is very similar between both classes of sequences. For comparison, the conserved genes that are well characterized (i.e., assignable to COGs, including the conserved hypothetical) are approximately 2.4 times more conserved than the intergenic sequences (Figure 3.7). Furthermore, we could detect very few (<5%) hypothetical or intergenic sequences conserved at the family level and we could not detect any such sequences conserved at the phylum level (data not shown). In contrast, a considerable number of well-characterized genes remain conserved over the same evolutionary scales. This gene set

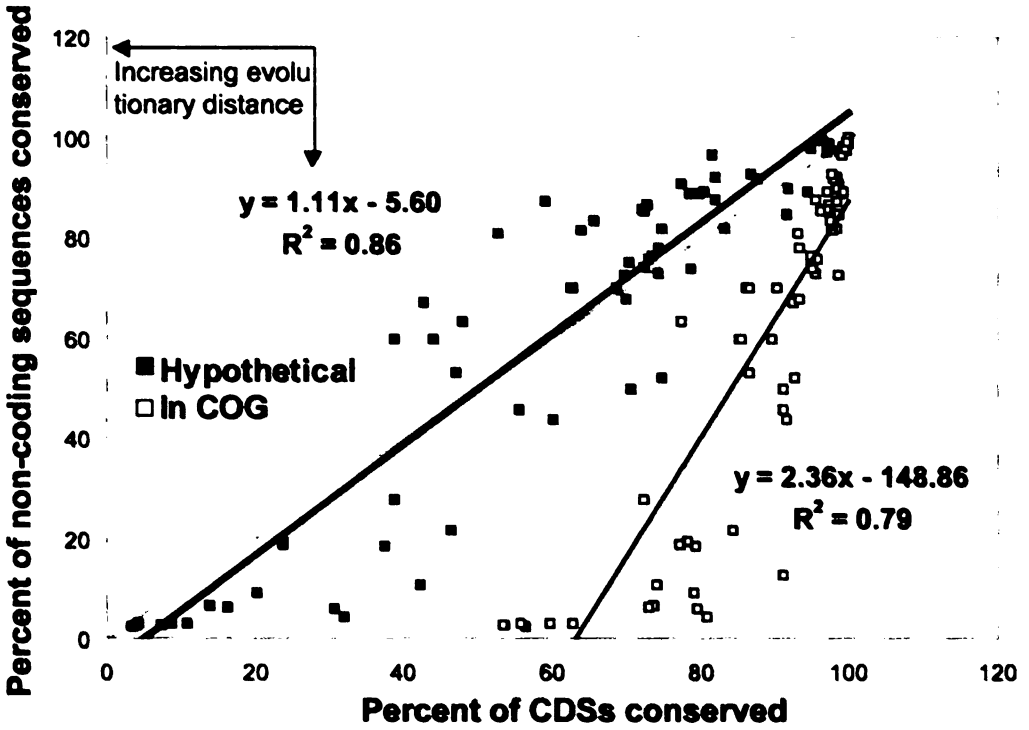


Figure 3.7. Degree of conservation of non-coding and hypothetical sequences vs. well characterized genes. Each datapoint represents the number of non-coding sequences (expressed as a percent of the total sequences to normalize genome size effect) from a reference genome conserved in a tester genome (y-axis) vs. the number of hypothetical genes (solid squares) or well-characterized genes (open squares) from the reference genome conserved in the tester genome (x-axis). The gray diagonal represents the 1:1 regression line.

includes both informational genes, which are highly conserved, as well as non-essential and less evolutionary conserved genes, like secondary metabolism genes, which have presumably been subjected to lateral transfer.

There are many inconsistencies between different published genomes with regard to the annotation and nomenclature of hypothetical genes, which impedes robust interpretations. These inconsistencies also explain part of the high dispersion of datapoints around the mean observed in Figure 3.7. Although we have not extensively

evaluated the effect of such inconsistencies, our results from comparisons of hypothetical genes to intergenic sequences clearly suggest that the function of the majority of hypothetical genes, if any, is different from the annotated genes (Figure 3.7). This agrees with conclusions reached by others using fundamentally different approaches, such as synomynous vs. non-synonymous amino acid substitutions (24), gene length distributions (31) and simulations on the coding capacity of the genome (17). Although there are specific caveats in all these methods (21, 24), the emerging picture is consistent with the majority of the hypothetical CDS being indispensable but not protein-coding parts of the prokaryotic genome.

This conclusion seems contradictory to recent proteomic data that show that a significant portion of what is annotated as hypothetical CDS is indeed translated to proteins (9, 19, 22). The discrepancy, however, is at least partially attributed to inconsistencies in nomenclature, e.g. we did not consider conserved hypothetical in our analysis as did Kolker et al. (19) and Corbin et al. (9), or to the study of phylogenetically diverse or not well-studied species where the fraction of annotated CDS as hypothetical genes is higher (22). Furthermore, recent evidence suggests that, in many genomes, a small (but not negligible) number of short protein-coding genes have escape identification (14) and are consequently annotated as non-coding DNA. This may have caused an underestimation of the coding potential of hypothetical CDS in our comparisons. In summary, our results do not contradict that some hypothetical genes are protein-coding, rather they suggest that such genes should constitute a small fraction of the total and their effect on cell phenotype may be uncertain in several cases, such as for the phage-related hypothetical genes. Given, however, the high frequency of hypothetical

CDS among the strain-specific sequences (Figure 3.6), the small number of coding hypothetical CDS may quantitatively contribute significantly to the species functional diversity.

OUTLOOK

Our analysis shows that if species should be reasonably predictive of phenotype and ecological potential then species should comprise a much more uniform suite of strains than provided by the current definition. In practical terms, it appears that such strains may be only the ones that show higher than 99% ANI or are less identical at the nt. level but share at least 95% of their well-characterized genes as a result of having a very overlapping ecological niche. This definition is closer to the eukaryotic standards as well. Such a stringent standard, however, would be impractical to implement, since it would instantaneously increase the number of existing species probably by a factor of 10 (6), and cause considerable confusion in the diagnostic and legal fields. Hence, the existing classification system should be maintained but adopt more stringent standards where needed, like in the case of distinguishing important species for diagnosis, patents, quarantine, transportation and possession. Our analysis clearly shows that strains of the same species according to the current standards may be too different to be considered the same species.

Our analysis also reveals several issues that must be addressed before more robust interpretations are possible. Most importantly, although species-specific genetic signatures appear to exist, this conclusion is based on a limited number of available sequenced strains. Therefore, the alternative hypothesis, i.e., there is a continuum of genetic diversity, which is not supportive of a species concept for Prokaryotes, cannot be currently rejected. It is also likely that a continuum of genetic diversity would be applicable only to specific species and/or ecological niches. Last, the importance of the species' ecology on the conserved genes needs to be more fully evaluated and quantified.

Related to this, the full ecological potential of most (even the sequenced!) species remains largely unknown due to the lack of knowledge on their population sizes and activities in their natural environments. A better coverage with genomic sequences of several closely related species from characterized niches is needed to further advance these cornerstone issues for microbiology and systematics.

AKNOWLEDGMENTS

We thank The Institute for Genomic Research (TIGR) and the Sanger center for permission to use preliminary sequence data. This work was supported by the Bouyoukos Fellowship Program (KTK), the DOE's Microbial Genome Program and the Center for Microbial Ecology.

REFERENCES

1. Report of the Tropical Indicator Workshop, available at: http://www.wrrc.hawaii.edu/tropindworkshop.html.

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

3. **Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.-m. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. S. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. K. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner.** 2002. Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. Science **297:**1301-1310.

4. **Boyd, E. F., and H. Brussow.** 2002. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. Trends Microbiol **10:**521-9.

5. **Brenner, D.** 1984. Bergey's manual of systematic bacteriology, 1st ed, vol. 1. William and Wilkins, Baltimore.

6. **Brenner, D., J. Staley, and N. Krieg.** 2000. Bergey's manual of systematic bacteriology, 2nd ed, vol. 1. Springer-Verlag, New York.

7. **Cohan, F. M.** 2002. What are bacterial species? Annu Rev Microbiol **56:**457-87.

8. **Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje.** 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res **31:**442-3.

9. **Corbin, R. W., O. Paliy, F. Yang, J. Shabanowitz, M. Platt, C. E. Lyons, Jr., K. Root, J. McAuliffe, M. I. Jordan, S. Kustu, E. Soupene, and D. F. Hunt.** 2003. Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. Proc Natl Acad Sci U S A **100:**9232-7.

10. **Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, R. E. Bontrop, and S. Paabo.** 2002. Intra- and Interspecific Variation in Primate Gene Expression Patterns. Science **296:**340-343.

11. **Falush, D., T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, M. J. Blaser, D. Y. Graham, S. Vacher, G. I. Perez-Perez, Y. Yamaoka, F. Megraud, K. Otto, U. Reichard, E. Katzowitsch, X. Wang, M. Achtman, and S. Suerbaum.** 2003. Traces of Human Migrations in *Helicobacter pylori* Populations. Science **299:**1582-1585.

12. **Garrity, G., J. Bell, and T. Lilburn.** Bergey's manual of systematic bacteriology, 2 ed, vol. Release 5.0. Springer-Verlag, New York.

13. **Goodfellow, M., and A. O'Donnell.** 1993. Handbook of New Bacterial Systematics. Academic Press Inc, San Diego.

14. **Harrison, P. M., N. Carriero, Y. Liu, and M. Gerstein.** 2003. A "polyORFomic" analysis of prokaryote genomes using disabled-homology filtering reveals conserved but undiscovered short ORFs. J Mol Biol **333:**885-92.

15. **Hoffmaster, A. R., J. Ravel, D. A. Rasko, G. D. Chapman, M. D. Chute, C. K. Marston, B. K. De, C. T. Sacchi, C. Fitzgerald, L. W. Mayer, M. C. Maiden, F. G. Priest, M. Barker, L. Jiang, R. Z. Cer, J. Rilstone, S. N. Peterson, R. S. Weyant, D. R. Galloway, T. D. Read, T. Popovic, and C. M. Fraser.** 2004. Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. Proc Natl Acad Sci U S A **101:**8449-54.

16. **Imaeda, T.** 1985. Deoxyribonucleic acid relatedness among selected strains of the *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium microti*, and *Mycobacterium africanum*. Int. J. Syst. Bacteriol. **35:**147-150.

17. **Jackson, J. H., S. H. Harrison, and P. A. Herring.** 2002. A theoretical limit to coding space in chromosomes of bacteria. Omics **6:**115-21.

18. **Kawamura, Y., X. G. Hou, F. Sultana, H. Miura, and T. Ezaki.** 1995. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. Int J Syst Bacteriol **45:**406-8.

19. **Kolker, E., S. Purvine, M. Y. Galperin, S. Stolyar, D. R. Goodlett, A. I. Nesvizhskii, A. Keller, T. Xie, J. K. Eng, E. Yi, L. Hood, A. F. Picone, T. Cherny, B. C. Tjaden, A. F. Siegel, T. J. Reilly, K. S. Makarova, B. O. Palsson, and A. L. Smith.** 2003. Initial proteome analysis of model microorganism *Haemophilus influenzae* strain Rd KW20. J Bacteriol **185:**4593-602.

20. **Konstantinidis, K. T., and J. M. Tiedje.** 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. PNAS **101:**3160-3165.

21. **Lawrence, J.** 2003. When ELFs are ORFs, but don't act like them. Trends Genet **19:**131-2.

22. **Liu, Y., J. Zhou, M. V. Omelchenko, A. S. Beliaev, A. Venkateswaran, J. Stair, L. Wu, D. K. Thompson, D. Xu, I. B. Rogozin, E. K. Gaidamakova, M. Zhai, K. S. Makarova, E. V. Koonin, and M. J. Daly.** 2003. Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. Proc Natl Acad Sci U S A **100:**4191-6.

23. **Mahillon, J., and M. Chandler.** 1998. Insertion sequences. Microbiol Mol Biol Rev **62:**725-74.

24. **Ochman, H.** 2002. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. Trends Genet **18:**335-7.

25. **Oleksiak, M. F., G. A. Churchill, and D. L. Crawford.** 2002. Variation in gene expression within and among natural populations. Nat Genet **32:**261-6.

26. **Parkhill, J., M. Sebaihia, A. Preston, L. D. Murphy, N. Thomson, D. E. Harris, M. T. Holden, C. M. Churcher, S. D. Bentley, K. L. Mungall, A. M. Cerdeno-Tarraga, L. Temple, K. James, B. Harris, M. A. Quail, M. Achtman, R. Atkin, S. Baker, D. Basham, N. Bason, I. Cherevach, T. Chillingworth, M. Collins, A. Cronin, P. Davis, J. Doggett, T. Feltwell, A. Goble, N. Hamlin, H. Hauser, S. Holroyd, K. Jagels, S. Leather, S. Moule, H. Norberczak, S. O'Neil, D. Ormond, C. Price, E. Rabbinowitsch, S. Rutter, M. Sanders, D. Saunders, K. Seeger, S. Sharp, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, L. Unwin, S. Whitehead, B. G. Barrell, and D. J. Maskell.** 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet **35:**32-40.

27. **Pedulla, M. L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix, and G. F. Hatfull.** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell **113:**171-82.

28. **Rossello-Mora, R., and R. Amann.** 2001. The species concept for prokaryotes. **25:**39.

29. **Sibley, C. G., and J. E. Ahlquist.** 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. J Mol Evol **26:**99-121.

30. **Sibley, C. G., J. A. Comstock, and J. E. Ahlquist.** 1990. DNA hybridization evidence of hominoid phylogeny: a reanalysis of the data. J Mol Evol **30:**202-36.

31. **Skovgaard, M., L. J. Jensen, S. Brunak, D. Ussery, and A. Krogh.** 2001. On the total number of genes and their length distribution in complete microbial genomes. Trends Genet **17:**425-8.

32. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kampfer, M. C. J. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52:**1043-1047.

33. **Tatusov, R., N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, B. S. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

34. **Tonjum, T., D. B. Welty, E. Jantzen, and P. L. Small.** 1998. Differentiation of *Mycobacterium ulcerans*, *M. marinum*, and *M. haemophilum*: mapping of their relationships to *M. tuberculosis* by fatty acid profile analysis, DNA-DNA hybridization, and 16S rRNA gene sequence analysis. J Clin Microbiol **36:**918-25.

35. **Treves, D. S., S. Manning, and J. Adams.** 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. Mol Biol Evol **15:**789-97.

36. **Van Sluys, M. A., M. C. de Oliveira, C. B. Monteiro-Vitorello, C. Y. Miyaki, L. R. Furlan, L. E. A. Camargo, A. C. R. da Silva, D. H. Moon, M. A. Takita, E. G. M. Lemos, M. A. Machado, M. I. T. Ferro, F. R. da Silva, M. H. S. Goldman, G. H. Goldman, M. V. F. Lemos, H. El-Dorry, S. M. Tsai, H. Carrer, D. M. Carraro, R. C. de Oliveira, L. R. Nunes, W. J. Siqueira, L. L. Coutinho, E. T. Kimura, E. S. Ferro, R. Harakava, E. E. Kuramae, C. L. Marino, E. Giglioti, I. L. Abreu, L. M. C. Alves, A. M. do Amaral, G. S. Baia, S. R. Blanco, M. S. Brito, F. S. Cannavan, A. V. Celestino, A. F. da Cunha, R. C. Fenille, J. A. Ferro, E. F. Formighieri, L. T. Kishi, S. G. Leoni, A. R. Oliveira, V. E. Rosa, Jr., F. T. Sassaki, J. A. D. Sena, A. A. de Souza, D. Truffi, F. Tsukumo, G. M. Yanai, L. G. Zaros, E. L. Civerolo, A. J. G. Simpson, N. F. Almeida, Jr., J. C. Setubal, and J. P. Kitajima.** 2003. Comparative Analyses of the Complete Genome Sequences of Pierce's Disease and Citrus Variegated Chlorosis Strains of *Xylella fastidiosa*. J. Bacteriol. **185:**1018-1026.

37. **Vauterin, L., B. Hoste, K. Kersters, and J. Swings.** 1995. Reclassification of *Xanthomonas*. Int. J. Syst. Bacteriol. **45:**472-489.

38. **Ward, D. M.** 1998. A natural species concept for prokaryotes. Curr Opin Microbiol **1:**271-7.

39. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and T. H. G.** 1987. Report of the Ad Hoc Committee on reconciliation of approaches to Bacterial Systematics. Int. J. Syst. Bacteriol. **37:** 463-464.

40. **Welch, R. A., V. Burland, G. Plunkett, III, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. PNAS **99:**17020-17024.

41. **Yabuuchi, E., Y. Kosako, H. Oyaizu, I. Yano, H. Hotta, Y. Hashimoto, T. Ezaki, and M. Arakawa.** 1992. Proposal of *Burkholderia* gen. nov. and transfer of seven species of the genus *Pseudomonas* homology group II to the new genus, with the type species *Burkholderia cepacia* (Palleroni and Holmes 1981) comb. nov. Microbiol Immunol **36:**1251-75.

CHAPTER 4

TOWARDS A GENOME-BASED TAXONOMY FOR PROKARYOTES.

INTRODUCTION

Prokaryotic taxonomy consists of three separate components: classification (i.e., the arrangement of organisms into groups or taxa), nomenclature and identification. Although there is no official classification for Prokaryotes, the classification system represented by the Bergey's Manual is widely accepted by the community of microbiologists and therefore is currently considered the best approximation to an official classification (3). This classification system is primarily based on the phylogenetic analysis of the small subunit ribosomal RNA gene (16S rRNA) and secondarily on old microscopic and/or biochemical observations about the relatedness of the organisms (3, 16, 18). The current classification system has been valuable in describing and appreciating the breadth of prokaryotic diversity and setting the framework for the study of relationships between taxa. Further, results from new approaches enabled by the availability of whole-genome sequences such as phylogeny based on shared content of orthologous genes (10, 13, 15, 24), indels or signature sequences (8, 14), concatenated alignments of many proteins (4, 11, 26), are generally congruent with the grouping of organisms based on the 16S rRNA gene, which adds further value in the current system.

It is important to realize, however, that the definition or standards for the existing taxonomic ranks are far from being well delineated, particularly for the higher than the species ranks. In fact, considerable subjectivity in designating genera, families etc., has been allowed, which is (at least) partially attributable to the great biochemical and morphological diversity exhibited by Prokaryotes and prevents the employment of the same measuring rules for all groups of organisms (3). The only major prerequisite for designating taxonomic ranks is that clustering by 16S rRNA data should support such

designations but no standards exist about the absolute genetic distance (measured by 16S rRNA gene sequence or other markers) between the different taxonomic ranks (16). Accordingly, the current taxonomy has frequently caused a lot of confusion, e.g., *Shigella spp.* and *E. coli* strains represent different genera (2) although based on their genetic relatedness they should belong to the same species (25), and uncertainty about how comparable the taxonomic ranks between different lineages may be. Most importantly, the relative predictive power of the different taxonomic ranks in terms of phenotype or relatedness of the grouped organisms remains unclear.

Genomic approaches hold great promise to provide insights into these issues because they can accurately reveal the genetic and functional relatedness between organisms at any resolution level. However, genomic studies to date have been mostly focused on assessing the accuracy of phylogenetic reconstruction, particularly in the light of lateral gene transfer (LGT), rather than the differences in the ranks of taxonomy between lineages, and have failed to address these issues systematically for all prokaryotic taxa. Here we have assessed the consistency of the taxonomic ranks for 175 fully sequenced genomes in terms of genetic distance, using as a measure for the latter the average amino acid identity of all conserved genes between any two organisms. Based on this measure, we found that there are many irregularities in the current classification schema for these 175 genomes while there is little, if any, value in the predictive power of the higher taxonomic ranks such as the order, class, or phylum with the exception of the domain rank, as these ranks are currently used. Our approach also provided means to evaluate the robustness of 16S rRNA gene and alternative molecular markers for phylogenetic purposes.

MATERIAL AND METHODS

Determination of conserved genes and genetic relatedness.

The genomic sequences and sequence annotation of the 175 genomes used in this study were obtained from NCBI's ftp site at ftp://ftp.ncbi.nih.gov/. Conserved genes between a pair of genomes were determined by whole-genome, pair-wise, sequence comparisons using the BLAST algorithm release 2.2.5 (1). For these comparisons, all CDS sequences from one genome were searched against the genomic sequence of the other genome (protein query vs. translated database, tBLASTn). CDS were considered conserved when they that had a BLAST match of at least 30% identity at the amino acid level (recalculated to an identity along the entire sequence) and an alignable region more than 70% of the length of the query CDS. This cut-off is above the twilight zone of similarity searches where inference of homology is error-prone due to low similarity between aligned sequences; thus query CDSs were presumably homologous to their match (21, 22) while searching against genomic sequences (as opposed to CDS) circumvented the problem of inconsistencies in the annotation between different genomes. When a reciprocal best match approach was employed to determine the orthologous fraction of the conserved genes in an effort for a more conservative estimation of functional similarity, then the amount of genes conserved between two genomes was smaller (but generally not considerably smaller) by an average of ~1.2%.

The genetic relatedness between a pair of genomes was measured by the average amino acid identity (AAI) of all conserved genes between the genomes as computed by the BLAST algorithm. 16S rRNA gene or other genetic marker identity was calculated in the same way as AAI, i.e., based on BLAST searches (nucleotide level -blastn- for 16S

and 23S rRNA and amino acid level -Blastp- for protein-coding genes), for consistency in comparing the results.

Taxonomic information.

The taxonomic information for each of the 175 genomes was extracted from the Hierarchy browser of the RDP database, release 9 (http://rdp.cme.msu.edu/index.jsp), which implements the newer version of Bergey's taxonomy (9). The taxonomic information included all the officially recognized taxonomic ranks, i.e., domain, phylum, class, order, family, genus, and species, with the exception of the subspecies rank. This information can be viewed in Table 4.1 of Appendix, which also includes the genome size and total number of CDS for each genome.

Phylogenetic analysis and sequence divergence.

Phylogenetic analysis was performed using the Neighbor Joining program of the Phylip package, version 3.62 (12) and the Weighbor (weighted neighbor joining) program (5). Sequence divergence at synonymous (Ks) and nonsynonymous (Ka) sites was calculated with DIVERGE software of the GCG package, which uses the method of Li (17).

RESULTS AND DISCUSSION.

Average amino acid identity is a robust measurement of relatedness.

For our purposes there was need for precise measurement of the genetic relatedness between any two strains. The main limitations in performing this task universally for all prokaryotic taxa are the lack of genes that are widely distributed in all taxa, e.g., recent estimates suggests that there are less than a hundred such genes, the varied evolutionary histories (mutation rate and selection pressures) of different genes and the, yet unclear, effect of LGT on inferred phylogenies. For these reasons and in order to maximize the robustness of our approach we employed the average amino acid identity (AAI) of all conserved genes between two strains to measure their genetic relatedness.

There are several strengths in using AAI for these purposes. First, AAI is a simple, useful, overall descriptor of genetic relatedness. Second, it is derived from lineage-specific genes, in addition, to the widely distributed ones (typically >500 genes in total), which increases the robustness of the phylogenetic signal extracted. Further, due to the large number of genes used in the calculations, AAI should be superior to a single gene, such as 16S rRNA gene sequence, for measuring relatedness and should not be prone to varied evolutionary rates or LGT events of single or a few genes. Even if genes with different evolutionary histories represent a large fraction the genome, their effect on AAI is minimized when some evolve faster but others slower than the average of the genome and hence should not be problematic for AAI (see also Figure 4.1.*A*). AAI also offers higher resolution than 16S rRNA gene sequence since a 0-40% 16S rRNA sequence miss-pairing (40% is the maximum 16S rRNA distance observed, i.e., between

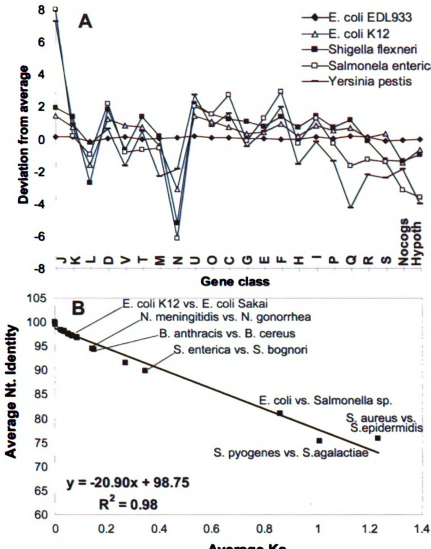


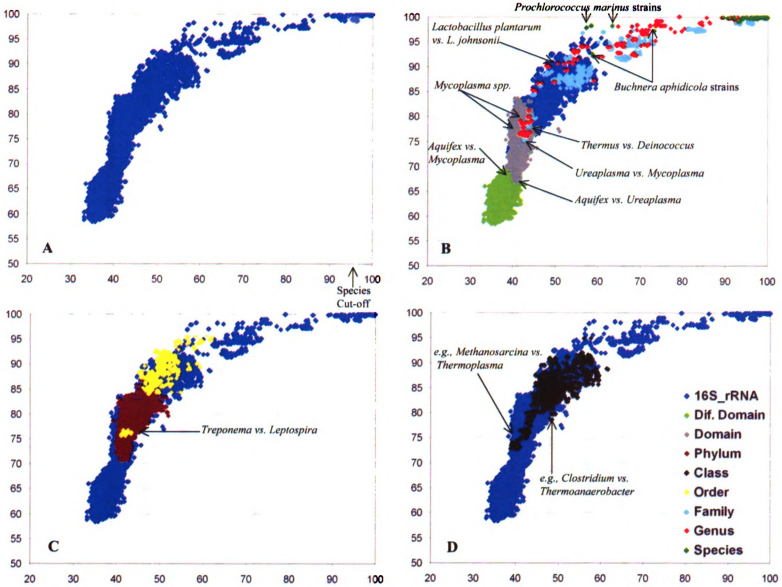
Figure 4.1. Average Nucleotide Identity (ANI) and genetic distance. (*A*) The ANI for all genes in the genome, and all genes in a COG category (designated by a single letter on x-axis; see Table 2.2 for letter designation) between *E. coli* strain Sakai and another genome (graph legend) were determined and the difference of the average identity of the genes in each category from the average identity of all genes in the genome is shown (y-axis). These results reveal that the nucleotide identity of most orthologs between any two genomes is within +/- 6-8% of the ANI between the genomes. A comparable picture was obtained for the Burkholderia, Mycobacteria and Streptococci groups (data not shown). **(B)** The average rate of non-synonymous substitutions (Ks) for all orthologs between two genomes strongly correlates with the ANI between the genomes, suggesting that ANI may be a useful descriptor of the evolutionary distance. Only genomes that show <3% 16S rRNA miss-pairing were included in the analysis to avoid saturation of nucleotide substitutions at non-synonymous sites. ANI correlates strongly with Average Amino acid Identity (AAI) ($R^2 > 0.95$) therefore the previous conclusions are translatable to AAI as well. ANI was preferred to give higher resolution between very closely related genomes.

domains) is spread between 0-70% average amino acid miss-pairing (since 30% identity was the cut-off for calling conserved genes) (Figure 4.2*A*) and can resolve areas where the 16S rRNA gene is inadequate, like the species level (see chapter 3 of this thesis). Last, AAI correlates strongly with the average rate of synonymous substitutions i.e., with the rate of sequence divergence, which suggests that AAI may be a useful descriptor of the evolutionary in addition to just genetic distance between two organisms (Figure 4.1.*B*).

Evaluation of the taxonomic ranks in terms of genetic relatedness.

We first compare the AAI to 16S rRNA identity for all pairs of the 175 genomes used in this study (175 X 175, 30,635 pairs in total) to gain insight into the interrelationship between these two parameters. Our results show that there is a strong correlation between 16S rRNA identity and AAI, and that the logarithmic model best describes this correlation (R^2 = 0.84, P < 0.0001) (Figure 4.2*A*). When the analysis is restricted to pairs of genomes with higher than 87-90% 16S rRNA identity, however, there is no significant difference between the logarithmic (R^2 = 0.834) and the linear model (R^2 = 0.825). These results indicate that influence of additional mutations (presumably in the 16S rRNA gene) is offset by recurrent mutations when 16S rRNA sequences are less than ~85-87% identical. In any case, the strong correlation observed further supports the robustness of 16S rRNA-based phylogeny for Prokaryotes. 16S rRNA appears to have limited resolution between closely related genomes, e.g., showing higher than 80% AAI, whereas it has higher resolution than AAI between (very) distantly

Figure 4.2. Relationships between 16S rRNA, AAI, and taxonomic information for the 175 sequenced genomes. Panel **A** shows the 16S rRNA gene sequence identity (y-axis) plotted against the average amino acid identity (AAI) for each pair of the 175 genomes (30,635 pairs in total). The smallest taxonomic rank that the two genomes of each pair share has been overlaid in panels **B, C,** and **D**. The area corresponding to the current standard for species delineation as well as representative pairs of genomes (discussed in the text) have been annotated. Images in this thesis are presented in color.


A
Species
Cut-off
B
Prochlorococcus marinus strains
Lactobacillus plantarum
vs. L. johnsonii
Mycoplasma spp.
Buchnera aphidicola strains
Aquifex vs.
Mycoplasma
Thermus vs. Deinococcus
Ureaplasma vs. Mycoplasma
Aquifex vs. Ureaplasma
C
Treponema vs. Leptospira
D
e.g., Methanoxarcina vs.
Thermoplasma
e.g., Clostridium vs.
Thermoanaerobacter
16S_rRNA
Dif. Domain
Domain
Phylum
Class
Order
Family
Genus
Species

related genomes, i.e., showing 30-40% AAI, presumably because this area approaches the cut-off used.

We then determined for each pair of genomes their closest taxonomic relationship, i.e., what is the smallest taxonomic rank they share, and overlay this information on the graph of Figure 4.2. It appears that there are many inconsistencies between the different taxonomic ranks since all ranks higher than the species and with the exception of the different domain show extensive overlap (compare for example genus vs. family in panel *B* or same domain vs. phylum between panels *B* and *C*). These results clearly show that the predictive power of current taxonomic ranks in terms of genetic distance between the grouped organisms is rather limited. In few cases the overlap is limited to a few genomes, such as among the *Prochlorococcus marinus* or the *Buchnera aphidicola* genomes (Panel B) and between *Treponema* and *Leptospira* (Panel C) genomes, whose genetic distance does not justify their inclusion in the same species and order, respectively. Such cases are apparently artifacts, e.g., *P. marinus* strains were grouped in the same species based solely on their high 16S rRNA gene sequence similarity (6, 7) and *Treponema* and *Leptospira* were assigned to the same order due to their common spirochete-like morphology (20), which can be corrected.

Another remarkable trend revealed in our data is that the currently named bacterial phyla are approximately as distant from each other in terms of AAI as Bacteria are from Archaea. This becomes more obvious on a neighbor joining tree built based on the full matrix of AAI between the 175 genomes. All bacterial phyla and sometimes classes, such the Mollicutes and Clostridia of the Firmicutes phylum and the α, δ, and ε classes of the Proteobacteria phylum, on this tree are as deeply branching as are Archaea

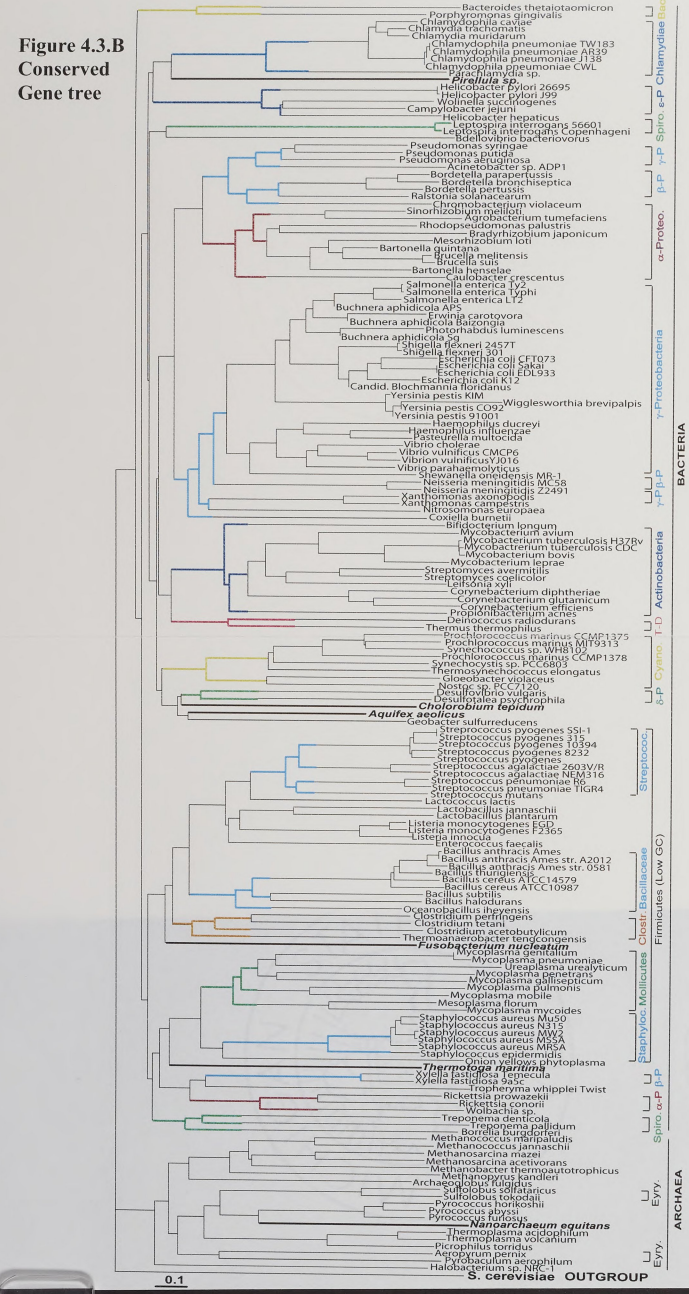
(see colored groups on Figure 4.3*A*). At the same time, clustering at nodes of the tree that correspond to well-defined relationships between groups is as expected, e.g., enterics are clustered together, with *Salmonella* spp. being the closest relative to *E. coli-Shigella* spp. group etc., which adds further support to the results. In addition, we found that there is strong linear correlation between the AAI between two genomes and the amount of genes that these genomes share ($R^2 = 0.70$, $P < 0.0001$), and this correlation becomes even stronger when the 32 reduced genomes of endo-symbiotic species are removed from the analysis ($R^2 = 0.82$, $P < 0.0001$) (Figure 4.4). The stronger correlation in the second case is attributable to the reduced genomes being enriched in highly conserved, housekeeping genes relative to the core of free-living species (the majority in the current dataset) and therefore the amount of conserved genes is overestimated in the former genomes relative to the latter. These results reveal that the genetic distance between the previous phyla/classes corresponds to comparably large functional/biochemical (gene) differences as well.

The previous conclusion is further supported by the distance tree derived from the full matrix of the percent of conserved genes between the 175 genomes. On this tree, one can see that most of the deep branching bacterial groups (phyla or classes) in the AAI tree are similarly deep branching in the conserved gene tree, i.e., genomes of these groups share a comparable amount of genes with genomes of the remaining bacterial phyla to the amount of genes they share with archaeal genomes (Figure 4.3*B*). For instance, the Thermus-Deinococcus, and the Actinobacteria phyla and the Clostridia, δ and ε Proteobacteria classes are deep branching in both trees, whereas the few apparent exceptions such as the Molicutes and α Proteobacteria classes that are not deep branching

Figure 4.3. Phylogenetic relationships between the 175 fully sequenced genomes. Neighbor joining trees derived from the full matrix of AAI (***A***) and percent of conserved genes (***B***) between the 175 genomes used in this study. The percent of conserved genes (instead of absolute number of conserved genes) was used to accommodate for genome size differences (up to 10 fold) among the 175 genomes. Groups that are deep branching on the AAI tree are denoted by colors. Phyla represented by a single genome are in bold. *Saccharomyces cerevisiae* genome was used to root the trees (outgroup). Scale bar represent 10% difference. Note the difference in scale between **A** and **B**, i.e., the underlying differences are about 25% larger in the conserved gene tree for the same branch length. Abbreviations are as follows (top to bottom in panel **A**): T-D -- Termus-Deinococcus phylum, Spiro. -- Spirochaetes phylum, Bact. -- Bacteroidetes phylum, α– β– γ– δ– ε– P. -- α– β– γ– δ– ε– Proteobacteria class respectively, Cyano. -- Cyanobacteria phylum, Streptococ. -- Streptococcaceae family, Staphyloc. -- Staphylococcaceae family, Eury. -- Euryachaeota phylum, Crena. -- Crenarchaeota phylum. Images in this thesis are presented in color.

Figure 4.3.A
AAI tree

Figure 4.3.B
Conserved
Gene tree



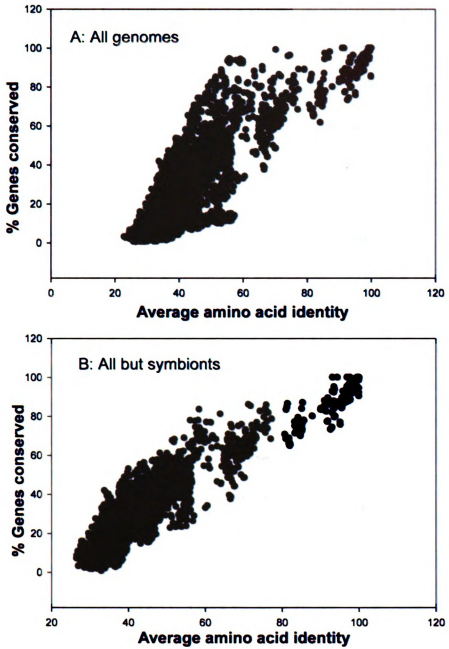


Figure 4.4. Relationship between conserved gene content and genetic distance. Dots represent the percent of conserved genes between a pair of genomes plotted against their genetic distance, measured as the average amino acid identity of the conserved genes. **(A)** All pairs of 175 genomes (30,625 pairs in total) were included, whereas pairs that contain an endosymbiotic genome were removed (32 genomes, 5,600 pairs removed) from the analysis in **(B)**.

in the conserved gene tree (contrary to the AAI tree) are attributable to the bias associated with the reduced genomes (discussed previously) or a shared ecology, e.g., the large genome-sized, free-living α Proteobacteria cluster together with the large genome-sized, free-living β and γ Proteobacteria in less deep nodes of the tree. In summary, these results suggest that there appears to be a much greater genetic and functional diversity in the Prokaryotes than hitherto expected based on the 16S rRNA phylogeny and that organisms of several bacterial phyla appear to be as different (genetically and/or biochemically) from each other as bacteria are different from archaea!

Evaluation of alternative markers to 16S rRNA for phylogenetic purposes.

The robustness of alternative markers to the 16S rRNA gene for phylogenetic purposes was also evaluated using as control in these evaluations the AAI and a similar approach as that used for the 16S rRNA gene. The results show that several of these markers such as RNA-polymerase subunits, t-RNA synthetases, Gyrase, Rec A protein etc. show considerable robustness based on the high correlation ($R^2 > 0.68$, $P < 0.0001$ for all markers tested) observed between the AAI and identity of these proteins for all pairs of the 175 genomes (Table 1 and Figure 4.5). Among the protein-coding genes tested, RNA-polymerase subunit B showed the highest correlation ($R^2 = 0.78$) to AAI and RecA protein the lowest ($R^2 = 0.68$) while all protein-coding genes evaluated showed significantly lower correlation to AAI than 16S rRNA ($R^2 = 0.84$). On the other hand, the large subunit RNA gene (23S rRNA) showed comparable, if not better, correspondence to AAI, suggesting that is a highly reliable marker (Figure 4.5). A similar approach may be used to evaluate the robustness of other markers as well, targeting the full breadth of

prokaryotic diversity or shorter evolutionary scales, e.g. the species level, for specific applications.

Table 4.1. Relationships of different phylogenetic markers to Average Amino acid Identity (AAI).

GENE		R^{2*}
16S rRNA	(Small subunit ribosomal gene)	0.84
23S rRNA	(Large subunit ribosomal gene)	0.84
RecA	(DNA strand exchange and recombination protein)	0.68
RpoB	(RNA polymerase, beta subunit)	0.78
GyrB	(DNA gyrase subunit B)	0.77
IleS	(Isoleucine tRNA synthetase)	0.72
FusA	(GTP-binding protein chain elongation factor EF-G)	0.69

*R^2 is for logarithmic second order correlation. This correlation gave among the highest R^2 values from the types of correlations tested for most genes. It should be mentioned however, that there were, typically, very small differences between different models (e.g. linear, power, logarithmic, sigmoidal etc) in their ability to describe the relationship between individual genes and the average of the genomes. Thus, no assumptions can be made about the underlying mechanisms of this relationship.

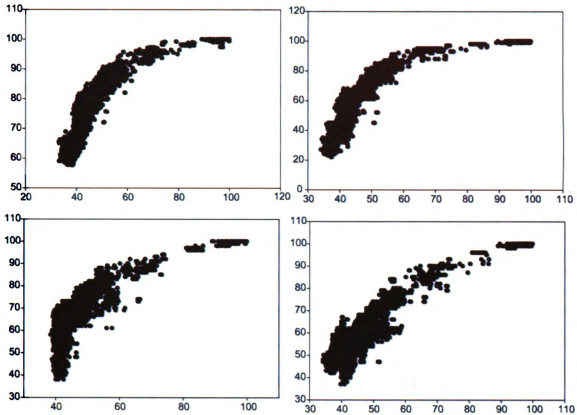


Figure 4.5. Correlation between alternative markers to 16S rRNA and Average Amino acid Identity (AAI). Panels show the correlation between identity of a molecular marker (panel title) and AAI for all pairs of the 175 genomes (at least 20,000 pairs for each gene) used in this study. For the full name description of a marker see Table 4.1.

PERSPECTIVE

The most important contribution of this work is the recognition that the ranks of prokaryotic taxonomy are frequently defined rather arbitrarily with respect to the genetic or biochemical relatedness of the grouped organisms. AAI and conserved gene content represent convenient means to quickly identify such cases and assist in standardizing the definitions of the ranks when these appear problematic. Moreover, it is evident from our analysis that organisms of almost all prokaryotic phyla and sometimes classes (colored groups in Figure 4.3*A*) are very different from each other, similarly to how different Bacteria are from Arhcaea. A number of morphological or physiological traits that characterize these organisms are fundamental and big differences from a prokaryotic perspective and therefore consistent with the vast differences revealed by the genomic comparisons. For example, organisms of the Molicutes class lack a cell wall, spirochetes have unique cell morphology and mode of movement and cyanobacteria are the only prokaryotes able to carry out water-based oxygenic photosynthesis. In addition, these differences are comparable to the morphological or physiological traits that are known to differentiate Archaea from Bacteria, namely, the existence of ether linked branched hydrocarbons in the membrane of the former (vs. ester linked fatty acids for Bacteria) and a few metabolic cofactors that are archaeal-specific such as coenzyme M, tetrahydromethanopterin etc. Last, by comparing the highly branching pattern of the best-represented bacterial phyla, the γ Proteobacteria and Frimicutes (light blue in Figure 4.3), with the deep rooting but not branching pattern of the remaining phyla or classes it becomes obvious that the great majority of the prokaryotic diversity is not yet represented by genomic sequences.

Although averaging across all genes in the genome may miss important, lineage-specific information, AAI (or Average Nucleotide Identity -ANI- for short evolutionary scales) represent a powerful first step towards a genome-based taxonomy because it is simple, robust and pragmatic for all prokaryotic taxa. Moreover, recent reports suggest that it may not be feasible to expand the current (16S rRNA-based) phylogeny by including more genetic markers either due to the shortage of genes widespread in all prokaryotic taxa or the difficulty in designing universal primers for widespread genes (23). Therefore, alternative methods such as the AAI-based method are needed. It may also be feasible to devise a new method or optimize an existing one to indirectly measure AAI i.e., to circumvent the need for whole-genome sequencing. Multi Locus Sequencing Typing (MLST) (19) that employs genes (not necessarily the same genes for all taxa!) that evolve comparably to the genome average may be one such approach, while the methodology described here (Figure 4.5) can assist the identification of good candidate genes for such an MLST-based application. In addition, work in our lab (*J. Goris et al.* in preparation) as well as the 2nd chapter of this thesis show that there is strong correlation between ANI and DNA-DNA reassociation homology values, the classical method for species delineation in Prokaryotes, over a range of relatedness that correspond to 0 to 5% 16S rRNA miss-pairing.

The AAI tree shows the Thermus-Deinococcus, Aquificae, and Thermotogae as the deepest branching bacterial phyla and the closest relative of Archaea similar to previous reports (4, 18, 26) but in conflict with others (10, 14). The differences at the ancestral nodes of the tree are very small (Figure 4.3*A*), however, therefore no definite conclusions can be reached based on these data about the sequence of evolution of the

different bacterial phyla. The same picture was also obtained when a less stringent cut-off for calling conserved genes (i.e., 20% identity instead of 30%) was used, which can pick homologs with weaker similarity at the expense of increasing the rate of false positive homolog recovery (data not shown). These results suggest that homology-based analysis may be inadequate to resolve the early evolutionary events of the prokaryotic life. The 16S rRNA gene might offer better resolution at the deep branches of the tree, however, the relationship between 16S rRNA and AAI (Figure 4.2) as well as the extensive genetic and biochemical distinctiveness of organisms related at this level, which presumably impose varied functional constrains and selection pressures on the 16S gene, raise serious concerns as to how quantifiable are 16S rRNA differences at this level of relatedness.

ACKNOWLEDGES

We thank Pr. George Garrity, James Cole and Dr. Joel Klappenbach for helpful discussions regarding the manuscript. This work was supported by the Bouyoukos Fellowship Program (KTK), the DOE's Microbial Genome Program and the Center for Microbial Ecology.

REFERENCES

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

2. **Brenner, D.** 1984. Bergey's manual of systematic bacteriology, 1st ed, vol. 1. William and Wilkins, Baltimore.

3. **Brenner, D., J. Staley, and N. Krieg.** 2000. Bergey's manual of systematic bacteriology, 2nd ed, vol. 1. Springer-Verlag, New York.

4. **Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope.** 2001. Universal trees based on large combined protein sequence data sets. Nat Genet **28:**281-5.

5. **Bruno, W. J., N. D. Socci, and A. L. Halpern.** 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol **17:**189-97.

6. **Canale-Parola, E.** 1984. Bergey's manual of systematic bacteriology, 1st ed, vol. 1. William and Wilkins, Baltimore.

7. **Chisholm, S., S. Frankel, R. Goericke, R. Olson, B. Palenik, B. Waterbury, L. West-Johnrud, and E. Zettler.** 1992. *Prochlorococcus marinus* nov. gen. sp.: an oxyphototrophic prokaryote containing divinyl chrolophyll *a* and *b*. Arch. Microbiol. **157:**297-300.

8. **Coenye, T., and P. Vandamme.** 2004. Use of the genomic signature in bacterial classification and identification. Syst Appl Microbiol **27:**175-85.

9. **Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje.** 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res **31:**442-3.

10. **Daubin, V., M. Gouy, and G. Perriere.** 2002. A Phylogenomic Approach to Bacterial Phylogeny: Evidence of a Core of Genes Sharing a Common History. Genome Res. **12:**1080-1090.

11. **Daubin, V., N. A. Moran, and H. Ochman.** 2003. Phylogenetics and the cohesion of bacterial genomes. Science **301:**829-32.

12. **Felsenstein, J.** 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

13. **Fitz-Gibbon, S. T., and C. H. House.** 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res **27:**4218-22.

14. **Gupta, R. S., and E. Griffiths.** 2002. Critical issues in bacterial phylogeny. Theor Popul Biol **61:**423-34.

15. **Hong, S. H., T. Y. Kim, and S. Y. Lee.** 2004. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. Appl Microbiol Biotechnol **65:**203-10.

16. **Krieg, N., and G. Garrity.** 2000. Bergey's manual of systematic bacteriology, 2 ed, vol. 1. Springer-Verlag, New York.

17. **Li, W. H.** 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol **36:**96-9.

18. **Ludwig, W., and H.-P. Klenk.** 2000. Bergey's manual of systematic bacteriology, 2nd ed, vol. 1. Springer-Verlag, New York.

19. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A **95:**3140-5.

20. **Munson, M. A., P. Baumann, and M. Kinsey.** 1991. *Buchnera* gen. nov. and *Buchnera aphidicola* sp. nov., a taxon consisting of the mycetocyteassociated, primary endosymbionts of aphids. Int. J. Syst. Bacteriol. **41:**566-568.

21. **Rost, B.** 1999. Twilight zone of protein sequence alignments. Protein Eng **12:**85-94.

22. **Sander, C., and R. Schneider.** 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins **9:**56-68.

23. **Santos, S. R., and H. Ochman.** 2004. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. Environ Microbiol **6:**754-9.

24. **Snel, B., P. Bork, and M. A. Huynen.** 1999. Genome phylogeny based on gene content. Nat Genet **21:**108-10.

25. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kampfer, M. C. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52:**1043-7.

26. **Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin.** 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol **1:**8.

CHAPTER 5

IN-SILICO MODELING OF DNA-MICROARRAY PERFORMANCE FOR GENOMOTYPING BACTERIAL STRAINS.

INTRODUCTION

The recent explosion in genomic sequencing has been accompanied by the development of high throughput technologies for post-sequencing analysis. Microarray technology has been at the cornerstone of this effort and is under continuing development. DNA microarrays were originally used to study gene expression levels between populations of mRNA expressed under different culture conditions or genotype backgrounds. Genes that are differentially expressed are very likely to play an important role in the cell physiology under these conditions and are targeted for further analysis (cf. references (16, 25, 29). More recently, microarrays have been used for genetic (or DNA-DNA) comparisons between different strains. In this case, a microarray is typically built using the available genomic sequence from a particular strain (the reference strain) and is used to competitively hybridize genomic DNA from closely related strains (the tester strains) (2, 6, 12, 17). The objective in this case is to reveal the gene differences between the reference and tester strain(s) that could explain unique characteristics of the strains under study. Last, DNA-DNA studies with specially designed microarray platforms have been proposed as a promising approach for genomo-typing and taxonomic characterization because they offer advantages at the species to genotype level over the 16S rRNA gene sequence analysis and the cumbersome DNA-DNA reassociation (DNA-homology) experiments (4).

A key issue for the successful application of microarrays for DNA-DNA studies is the evolutionary distance, i.e. the degree of nucleotide divergence between the reference and the tester strain(s). A microarray is expected to give false negative signal when the evolutionary distance is such that the nucleotide sequence of genes has diverged but their

amino-acid sequence remains conserved (hence the proteins are conserved). However, the relationship between false negative signal and evolutionary relatedness of the evaluated strains has not yet been investigated. This issue is also problematic when a whole genome microarray based on a reference strain is to be used for expression studies with other than the reference strain. Bioinformatic sequence analysis can potentially offer novel insight into this aspect of microarray technology. For instance, the number of genes from the reference genome conserved (nucleotide level) in the tester genome should approach the number of genes that is expected to cross-hybridize when the tester genome is hybridized on a microarray built from the reference genome. The relationship between sequence identity of the probe-target pair and hybridization kinetics has been extensively studied for different types of probes (14, 17), which allows for a fairly accurate estimation of the number of genes that can cross-hybridize based on their sequence identity.

For DNA-DNA studies, microarrays have been commonly used within species because it is assumed that the rate of false negatives will be minimum at the sub-species level. This assumption is based on the fact that strains that show 70% or greater DNA-homology values (the classical cut-off for species definition) are believed to have at least 95% DNA sequence identity in coding regions (11, 22). At this level of sequence identity (or evolutionary relatedness) no false negatives are expected. However, it is important to realize that DNA-homology values do not reflect the actual degree of sequence identity at the level of the primary structure. For instance, each of the three fully sequenced *E. coli* strains has about 25% if its DNA not shared with the other two sequenced strains (26). It is yet unclear how such differences between strains can affect microarray performance in experiments within species.

Finally, oligonucleotide arrays, which are typically comprised of a short, 30-60 nucleotide long probe per predicted open reading frame (ORF) in the genome, have recently gained popularity over those made from PCR products of ORFs (hereafter termed cDNA arrays) for expression studies due to their higher specificity during hybridizations, flexibility in design and potential for further technological development (14, 19). For DNA-DNA studies within or across species, cDNA arrays are presumably preferable for their higher sensitivity due to the longer probes employed even though the longer probes are prone to more non-specific signal from cross-hybridization of paralogous genes or conserved domains. Whether oligo-arrays can be comparable to cDNA arrays for DNA-DNA studies has not yet been investigated.

Using the available genomic sequences, we attempted to simulate, *in-silico*, the microarray performance and evaluate the previously mentioned issues. Three bacterial groups, namely the enterics, mycobacteria, and streptococci, were used as models in the simulations (Table 5.1) because they include several sequenced representatives (complete or high draft) and these representatives show a gradient of evolutionary relatedness.

MATERIAL AND METHODS

The genomic sequences of the completely sequenced strains of the three groups targeted were obtained from NCBI's ftp site at ftp://ftp.ncbi.nih.gov/. Preliminary sequence data for *M. bovis* and *M. marinum* strains were produced by the Sanger Center and were obtained through the Sanger ftp site at ftp://ftp.sanger.ac.uk/pub/; *M. avium*, *M. smegmatis*, and *S. mitis* were produced by The Institute for Genomic Research (TIGR) and obtained through their website at http://www.tigr.org.

Microarray false negatives.

False negatives for a microarray experiment were defined as the ORFs from the reference genome that were conserved at the amino acid level but were not conserved enough at the nucleotide level in the tester genome to allow cross-hybridization on a hypothetical microarray built based on the reference genome (see how whole genome sequence comparisons were performed below). An ORF was considered to cross-hybridize when it had a match of at least 60% nucleotide identity over more than 70% of its length in the tester genome. ORFs that have 60% or higher nucleotide identity have been shown to give significant cross-hybridization signal on cDNA microarrays in at least two independent studies (6, 17). Murray et al. have also proposed that this level of sequence identity is close to the detection limit on cDNA platforms (6, 17). To determine the number of genes conserved at the amino acid level, two cut-offs in pair-wise sequence comparisons were used: either at least 30% amino acid identity over more than 70% of the length of the query ORF or 60% amino acid similarity over more than 70% of the length of the query ORF. The former cut-off is above the twilight zone of homology

searches thus, the genes that pass this cut-off are expected to be homologous (either orthologs or paralogs) and share at least the same general biochemical function (7, 10, 20). The latter cut-off is comparable to the one used in the nucleotide comparisons (same match length, same degree of similarity) and offers a measure of the different rates of evolution between the nucleotide and the amino acid level. Similarity instead of identity was preferred in this case to make use of the available knowledge on similarities in function between different amino acids. Finally, the cut-off of 70% of the length of the query ORF was used in all cases to ensure that the same gene is involved (not just a conserved domain) but simulations using smaller cut-offs such as 60% of the length did not significantly affected our conclusions (data not shown).

This *in-silico* experiment was performed within each of the three model groups in our study, namely the enterics (10 genomes), mycobacteria (7 genomes) and streptococci (6 genomes). *E. coli* strain O157, *M. tuberculosis* strain H37Rv and *S. pneumoniae* strain TIGR4 were used as the reference genomes in each group, respectively.

Pair-wise whole genome comparisons.

All ORFs annotated as (predicted) protein-coding sequences in the GenBank files at NCBI of the reference genome were searched against the whole genomic sequence of the tester genome using the appropriate versions of the BLAST algorithm release 2.2.4 (1). The blastn (nucleotide level) default settings tend to give shorter alignments compared to blastp or tblastn (amino acid level) with distantly related species (where nucleotide sequences are more diverged) because they are targeting highly identical matches. This caused an underestimation of the number of conserved genes at the

nucleotide level compared to the amino acid level in distantly related species when default settings where applied. In an attempt to make BLAST alignments in the nucleotide search comparable to ones in the amino acid search, we gradually changed several of the blastn parameters until saturation in the total number of matches passing our nucleotide cut-off was reached. When differences were negligible i.e. less than 1% difference in the total number of matches, default settings were used. The same approach was applied for several parameters in the amino acid searches as well. This led to the following parameters used in the study: a) for blastn: X = 150 (drop-off value for gapped alignment), and q = -1 (penalty for nucleotide mismatch), the rest of the parameters were at default settings, b) for blastp or tblastn: default settings c) for the oligo probes (50 mers) blastn search: X = 50, q = -1 and W = 7 (word size); the rest of the parameters were at default settings. Searching against the whole genomic sequences (instead of the annotated ORFs) was preferred to avoid inconsistencies in annotation between two genomes.

cDNA vs. oligo arrays.

Evolutionary relatedness experiment. When microarrays are used to study the evolutionary relationships among strains the following procedure is typically used: The tester strain is labeled with a different dye (e.g Cy3) from the reference genome (e.g. Cy5), the two labeled genomes are then competitively hybridized on a whole genome microarray platform build based on the reference genome and the dye ratios are used to reveal the evolutionary relatedness (in terms of gene content) between the evaluated strains. We attempted to compare cDNA to oligo arrays with this respect by simulating

evolutionary relatedness experiments *in-silico* as follows: We designed hypothetical oligo and cDNA probes (see probe design below) for the reference genomes and use their BLAST matches in the tester genomes to estimate to the expected hybridization signal on a hypothetical microarray experiment. For this, the best blastn match of every query sequence (e.g. oligo probe, cDNA probe or whole ORF sequence) in the tester genome when had an expectation value less than e <0.001 (or e < 10 for oligo-probe sequences) was saved. The length of the match and its identity were transformed to a 0 to 1 scale and the transformed length and identity values were multiplied. In this way the most similar matches were given higher scores, e.g. the perfect matches equaled 1, which was analogous to a hybridization experiment where the genes with a higher degree of similarity are expected to give higher hybridization signals. Thus, the [transformed length X transformed identity] values for each query sequence offered a reliable, qualitative prediction of its expected hybridization signal against the tester genome relative to its expected signal against the reference genome (because the latter equaled 1 for all query sequences), similar to the Cy3/Cy5 ratio used in real microarray experiments.

Phylogenetic trees of the evaluated genomes were subsequently built based on the hierarchical clustering of the predicted hybridization signals (the [transformed length X transformed identity] values) using the Cluster version 3.0 software (8). Final trees were visualized with the TreeView software, available at http://rana.lbl.gov/EisenSoftware.htm (8). Both parametric (Pearson correlation, Euclidean distance) and non-parametric (Spearman correlation) methods were used to calculate distances in the trees. The whole ORF trees presumably represented the expected results and were used as reference for comparisons between the oligo and cDNA trees. Finally, the non-specific hybridization

signal, which is presumably significant in real experiments, was not considered in this simulation since only the best BLAST match for each query sequence was included in the analysis.

Correct gene identification experiment. We also evaluated whether oligo arrays give comparable results to cDNA arrays with respect to the correct gene identification. For this, we determined which oligo or cDNA probes are expected to cross-hybridize (with the tester genome) and check them against the results of the corresponding i.e. the ORF that the probe was designed for, whole ORF sequences. False negatives in this case were defined as the probes that were not predicted to cross-hybridize but the corresponding ORFs were, whereas, the reverse was considered false positive. Oligo probes (50 mers) were expected to cross-hybridize when they had a blastn match better than 80% identity over more than 80% of the length of the oligo probe in the tester genome. Fifty-mer oligonucleotides that share this level of identity with a target sequence have been shown to cross-hybridize to it (14). The same cut-off (i.e. 60% nucleotide identity over more than 70% of the length) as previously used for whole ORF sequences was applied to determine the number of cDNA probes that cross-hybridize. cDNA probes were at least 200nt long and had small differences compared to the corresponding whole ORF sequences (e.g. the average sequence length was 718 vs. 903 nucleotides, respectively), which justified the usage of the same cut-off for cDNA probe sequences. False negatives involved instances where two ORFs had a short non-overlapping region and the probe was designed for this region or the region targeted by the probe had diverged below the probe cut-off but the overall sequence identity of the whole ORFs was still greater than the cut-off used for ORFs. False positives mostly involved cases

where two ORFs have a short overlapping region (less than 70% of the length) and the oligo was designed for this region.

Probe design.

Probes were designed for each reference strain within a group. In the following text, the reference strain for the enterics group, strain O157, is used as a representative example i.e. the same analysis was performed for the remaining two reference strains. cDNA probes were designed as follows: the PRIMEGENS software (27) was used to design primers to amplify unique fragments for every possible ORF in the *E. coli* strain O157's genome. PRIMEGENS was run with default settings except that the amplified region was limited to between 200 to 1000 nucleotides. The sequence between a primer pair (the amplified region) was then extracted from the genomic sequence using PERL scripts and these sequences were used as cDNA probes. With this approach, we were able to design specific primers for 3,994 ORFs in strain O157's genome. The 3,994-cDNA probe sequences were then searched against the remaining genomic sequences in enterics group as previously described for whole ORF sequences. Oligo-probes (50 mers) specific for each ORF in the *E. coli* strain O157 genome were designed using the OligoArray software (21). The OligoArray settings were optimized to ensure probe specificity, avoid secondary structure and poly-nucleotide repeats (> 5 mers, e.g. TTTTT) in the probe sequence. In total, 5,298 oligos were designed for the 5,361 ORFs in strain O157 genome; 63 ORFs failed to give a specific oligo under the selection criteria of our design. The oligo sequences were then searched against the remaining genomic sequences in the enterics group as described previously.

The final comparison between cDNA and oligo probes was performed with the ORF set that had both a cDNA probe and an oligo probe designed (3,992 ORFs in the *E. coli* O157 case).

Non-specific signal.

The influence of non-specific hybridization signal, i.e. signal that is attributable to multiple gene copies, paralogous genes and/or conserved domains rather than the targeted sequence, on microarray results remains a poorly investigated issue. We attempted to evaluate the importance of non-specific signal in DNA-DNA microarray studies by considering all BLAST matches of the whole ORF sequences in the tester genomes. In this case, the [transformed length X transformed identity] values for all matches of an ORF in the tester genome were summed and the result was divided by the sum of the [transformed length X transformed identity] values for all matches of the same ORF within the reference genome. The ratio of the sums was used as a qualitative prediction of the relative hybridization signal between tester and reference genomes; similar to the simulation described previously where only the best match was considered. This approach assumed that two matches of similar identity but of different length (e.g. 10% vs. 100% of the length) would contribute to the overall signal proportionally to their length (e.g. 1/11 vs. 10/11, respectively). Likewise, matches of different levels of identity would contribute proportionally to their identity.

This experiment was not performed for probe sequences because the effect of the position and extent of the miss-pairing on hybridization signal is not easily quantifiable, particularly for short oligo sequences such as 50 mers (14).

Table 1. Pair-wise 16S rRNA gene sequence similarity (upper right) and DNA-DNA reassociation values (lower left) for the species used in this study.

	Species	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Streptococc.	1. *S. pneumoniae*		99.0	95.1	94.3											
	2. *S. mitis*	30-46		93.4	94.5											
	3. *S. mutans*	NS*	NS		94.1											
	4. *S. pyogenes*	NS	NS	NS												
Mycobacteria	5. *M. tuberculosis*						99.9	99.1	99.3	95.6						
	6. *M. bovis*					>90		99.2	99.4	95.5						
	7. *M. marinum*					11	27		99.2	95.4						
	8. *M. avium*					27	25	NA+		95.0						
	9. *M. smegmatis*					NS	NS	NS	NS							
Enterobacteria	10. *E.coli*											98.9	97.8	97.3	93.2	82.7
	11. *S. flexneri*										>70		97.4	96.9	92.9	
	12. *S. enterica*										NA	NA		97.3	92.9	
	13. *K. pneumoniae*										NA	NA	NA		93.8	
	14. *Y. pestis*										NA	NA	NA	NA		
	15. *P. aeruginosa*										NA	NA	NA	NA	NA	

*NS: Not significant (typically below 10% DNA-DNA reassociation). +NA: Not available but presumably, based on the 16S rRNA sequence distance, not significant (22). The DNA-DNA reassociation values were collected from the literature (i.e., for Streptococci see (15), for Mycobacteria see (13, 24), and for Enterobacteria see (3) whereas the 16S rRNA gene identities were computed from the available genomic sequences). The strains used in this study are expected to show little dispersal from the DNA-DNA reassociation values for the corresponding species shown here.

RESULTS

The evolutionary distances, in terms of DNA-homology values and 16S rRNA sequence identity for the species used in this study (Table 5.1) were collected from the literature (3, 13, 15, 24) or computed from the available genomic sequences using the online tool at the Ribosomal Database Project at http://rdp.cme.msu.edu/ (5), respectively. The species evaluated show a gradient of relatedness from highly related pairs such as *E. coli* and *S. flexneri* or *M. tuberculosis* and *M. bovis*, to moderately related ones such as *S. pneumoniae* and *S. mitis* and distantly related ones such as *E. coli* and *Y. pestis* or *Salmonella enterica*. This gradient is reflected in DNA-homology values with >70% (or >99% for 16S rRNA) for the highly related pairs to 50-30% (or >98% for 16S rRNA) for the moderately related ones and <30% (or <98% 16S rRNA) for the distantly related ones. In fact, some of the species we term highly related are considered ecotypes of the same species by many investigators.

Predicted microarray performance.

To comprehensively evaluate microarray performance, we expressed the number of false negatives between any pair of strains (i.e. a reference and a tester strain) as a percent of the total number of ORFs expected to cross-hybridize and plotted it against the DNA-homology and 16S rRNA sequence identity values between the pair of strains (Figure 5.1). The expression of false negatives as a percentage allowed for a genome size independent estimation, since larger genomes (i.e. more ORFs) gave more false negatives (absolute number) compared to smaller genome-sized species that showed similar evolutionary relatedness to the reference strain. Additionally, the usage of number of

ORFs that are expected to cross-hybridize instead of the total number of ORFs in the genome minimizes the effect of the varied levels of genomic diversity (e.g. loss or addition of genetic element) that characterize different species (e.g. the sequenced *E. coli* strains harbor much greater genomic diversity that the *M. tuberculosis* ones).

Our results suggested that false negatives increased with increased evolutionary distance between the reference and the tester strain (Figure 5.1). The two cut-offs used to determine the number of the conserved genes (amino acid level) gave significantly different estimations, with the 30% amino acid identity cut-off giving more false negatives than the 60% amino acid similarity cut-off. For example, a microarray experiment would be expected to miss at least 5% of the conserved genes when the reference and tester strains reside in moderately related species according to the 30% amino acid identity cut-off (Figure 5.1*A*) whereas, the same number of false negatives is expected when the reference and tester strains reside in moderately related species according to the 60% amino acid similarity cut-off (Figure 5.1*B*). Regardless of the cut-off used however, DNA-DNA studies between strains that are less than 97.5-97.0% identical in terms of 16S rRNA sequence are expected to have an unacceptably high number of false negatives (i.e., more than 10%).

With regard to the estimation of the evolutionary distance between reference and tester strain, 16S rRNA sequence identity offered a better measurement than DNA homology values because the latter method gave poor resolution in distantly related species (see DNA homology datapoints below 20% in Figure 5.1). In addition, the 16S rRNA sequence identity values gave a stronger correlation than the DNA-homology values. This is partially explained by the technical limitations in the DNA-homology

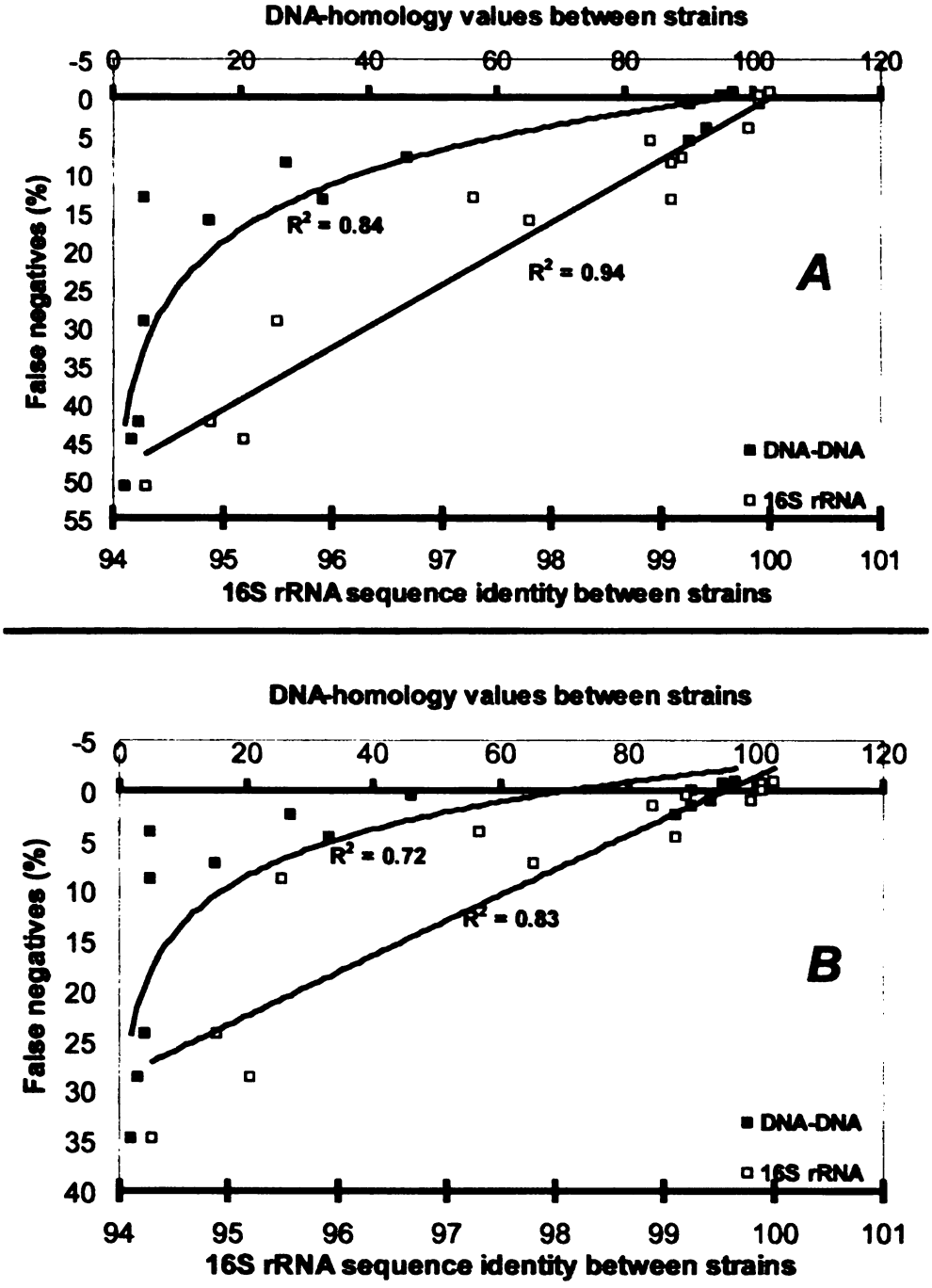


Figure 5.1. Correlation between microarray false negatives and evolutionary distance between reference and tester strain. Each point represents the false negatives, expressed as percentage of the total number of ORFs predicted to cross-hybridize with the tester genome, between a reference and a tester strain plotted against the DNA-homology values (solid squares, upper X-axis) and the 16S rRNA sequence identity (open squares, bottom X-axis) between the reference and tester strain. **(A)**: 30% amino acid identity cut-off. **(B)**: 60% amino acid similarity cut-off.

experiments such as the imprecision of these measures, the varied protocols used and the fact that the strains of the species used in these experiments were different from the strains of the same species sequenced and used in our simulations. The correlation was slightly higher for the 30% identity than for the 60% similarity cut-off (R^2=0.94 vs. R^2=0.83 for the 16S rRNA data and R^2=0.84, vs. R^2=0.72 for the DNA-homology data; all regressions were significant at $P < 0.001$). The strong correlations obtained with the combined data set are indicative of the comparable results obtained within each of the three bacterial groups evaluated (analytical results for each group are not shown).

Importance of microarray false negatives.

To evaluate their importance, microarray false negatives were checked against the total number of ORFs from the reference strain not conserved at the nucleotide level in the tester strain (i.e., the reference strain-specific ORFs). False negatives comprised, at maximum, one-third and one-fifth of the total number of ORFs not conserved based on the 30% amino acid identity and 60% amino acid similarity cut-off, respectively (Figure 5.2). Furthermore, false negatives became less important, i.e. comprised a smaller fraction of the non-conserved genes, with decreased evolutionary distance between the tester and reference strains. For instance and regardless of the cut-off used, false negatives did not comprise more than 15% of the ORFs not conserved in the tester strain for any tester strain highly or moderately related to the reference strain. These results suggested that although false negatives may occur at significantly high numbers (see 30% amino acid identity cut-off in Figure 5.1), they should represent a small fraction of the genes not shared between highly or moderately related strains (Figure 5.2).

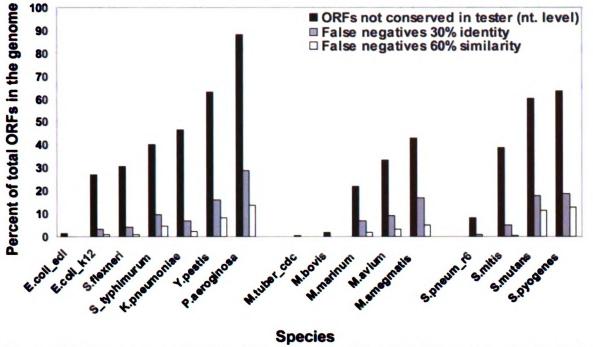


Figure 5.2. Importance of microarray false negatives. Solid bars represent the total number of ORFs from the reference genome not conserved at the nucleotide level in the tester genome (X-axis). Gray and open bars represent the part of these ORFs that are also predicted to be microarray false negatives at the 30% amino acid identity and 60% amino acid cut-offs, respectively.

Non-specific signal.

Non-specific hybridization signal appeared to affect a sizeable number of ORFs in all pairs of reference-tester strains tested. For instance, for the 3,994 whole ORF sequences evaluated between *E. coli* O157 and *S. enterica* pathovar Typhimurium, 1,222 (30.6%) had a different predicted hybridization signal when all matches were considered compared to the best match prediction (Figure 5.3*A*) and 466 (11.7%) of them showed a larger difference than +/- 0.1 from their best match prediction. Of the 466 ORFs, 268 gave higher signal when all matches were considered and 123 of them were predicted to give higher signal with the Typhimurium genome than with strain O157 (datapoints that have values more than 1 on the y-axis). The latter is attributable to the tester strain having

more copies of the gene, paralogous genes, and/or conserved domains than the reference strain for these 123 ORFs. The opposite situation i.e. ORFs showing less hybridization signal when all matches are considered, was true for 198 of the 466 ORFs. Thus, for a significant fraction of ORFs in any competitive hybridization experiment, misleading

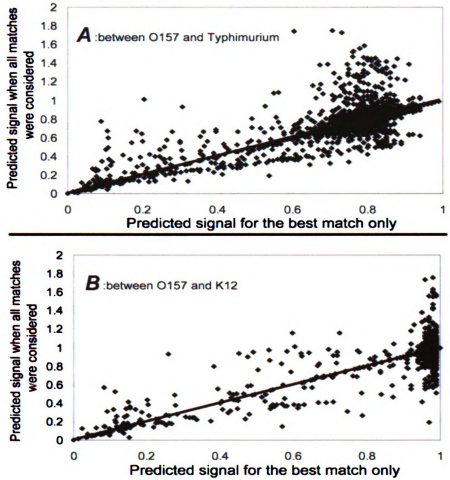


Figure 5.3. Non-specific hybridization signal for whole ORF sequences. Each point represents the predicted signal for an ORF when all its matches in the tester genome were considered (Y-axis) vs. the predicted signal when only the best match was considered (X-axis). Thus, any points that deviate from the diagonal represent ORFs that are predicted to be affected by non-specific hybridization signal. **(A)**: tester strain is *S. enterica* pathovar Typhimurium, **(B)**: tester strain *E. coli* K12. Reference strain is *E. coli* O157.

results, i.e. false positives or false negatives, should be expected as the result of non-specific signal.

E. coli K12 had more highly related matches (datapoints in the 0.8-1 range between Panels *A* & *B*) and fewer ORFs (267) that showed more than +/- 0.1 difference from their best match prediction than Typhimurium reflecting its closer relatedness to the reference strain (Figure 5.3*B*). Nonetheless, strain K12 had a comparable number of ORFs affected by non-specific signal (1232) to Typhimurium. Similar trends were observed for the remaining pairs tested (data not shown).

cDNA vs. Oligo arrays.

The predicted performance of oligo and cDNA arrays was evaluated in terms of: I) the expected results relative to the evolutionary distance among the evaluated strains and II) the correct gene identification.

I) Evolutionary relatedness. Trees based on the hierarchical clustering of the predicted hybridization signal were very similar, both in terms of topology and distances between nodes, between cDNA and whole ORF regardless of the method (parametric vs. non-parametric) used for the calculation of distances (Figure 5.4 *B* & *C*). The high congruence between cDNA and whole ORF trees was probably attributable to the small differences between the cDNA probe sequences and the whole ORF ones (see methods section). On the other hand, the oligo tree tended to overestimate distances in more distantly related strains (relative to the reference strain) compared to the cDNA one. For example, the oligo tree predicted a larger distance between the *E. coli-Shigella* cluster

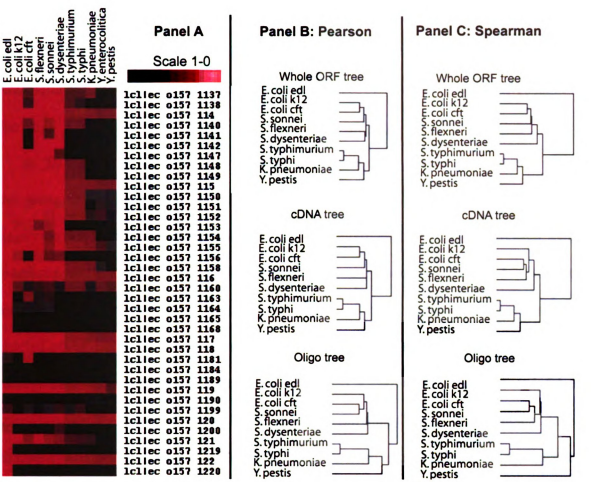


Figure 5.4. cDNA vs. Oligo-arrays: Evolutionary relatedness results. (A): Each spot represents the [transformed length X transformed identity] value for the best BLAST match of an O157 ORF (right) in a tester genome (top). **(B)**: The results from the hierarchical clustering of the [transformed length X transformed identity] values using Pearson correlation for every set of query sequences i.e. whole ORFs, cDNA and oligo probes. **(C)**: Hierarchical clustering using Spearman rank correlation. Images in this thesis are presented in color.

and the *Salmonella* or the *Yersinia* ones than the whole ORF tree. This property of the oligo tree also caused some branching differences in the ancestral nodes when the Spearman correlation was used, e.g. *Yersinia* groups with *E. coli* EDL instead of the *Salmonella-Klebsiella* cluster (Figure 5.4*C*). However there was, overall, high similarity

between the oligo and the cDNA trees as was evident by the identical clustering of strains at the terminal nodes between the two trees. In addition, principal component analysis confirmed the presence of three major clusters (i.e. the *E. coli-Shigella*, the *Salmonella-Klebsiella* and the *Yersinia*) for all three trees e.g. oligo, cDNA and whole ORF (data not shown).

II) Correct gene identification. When the reference and tester genome resided in the same or highly related species, oligos had sufficiently low incidences of false negatives (Figure 5.5). However, oligo-array false negatives dramatically increased with increased evolutionary distance. It appeared that the increase correlated with the transition of the tester strain from highly related to moderately related species and leveled-off when the tester strain is a distantly related species. For instance, all the highly related pairs of reference-tester strains in Figure 5.5 had about 1% predicted false negatives (see *S. pneumoniae* strain R6, *E. coli* K12, *Shigella flexneri* data points) and the moderately related *S. mitis* (46% DNA-DNA reassociation and 99% 16S rRNA sequence identity to the reference *S. pneumoniae* TIGR4) had about 5%. When the tester strain was a distantly related species (e.g. *Salmonella* or *Yersinia* for Panel *A*, or *S. pyogenes* and *S. agalactiae* for Panel *B*), false negatives were between 30-40%. On the other hand, false negatives for the cDNA array were consistently below 5% for all tester strains. Lastly, the predicted false positives for both cDNA and oligo-arrays were consistently below 2-3% regardless of the tester strain used (data not shown). For the oligo-array, this was not surprising inasmuch as the likelihood of getting a 50 nucleotide long exact match in the tester strain by chance alone is $(¼)^{50}$.

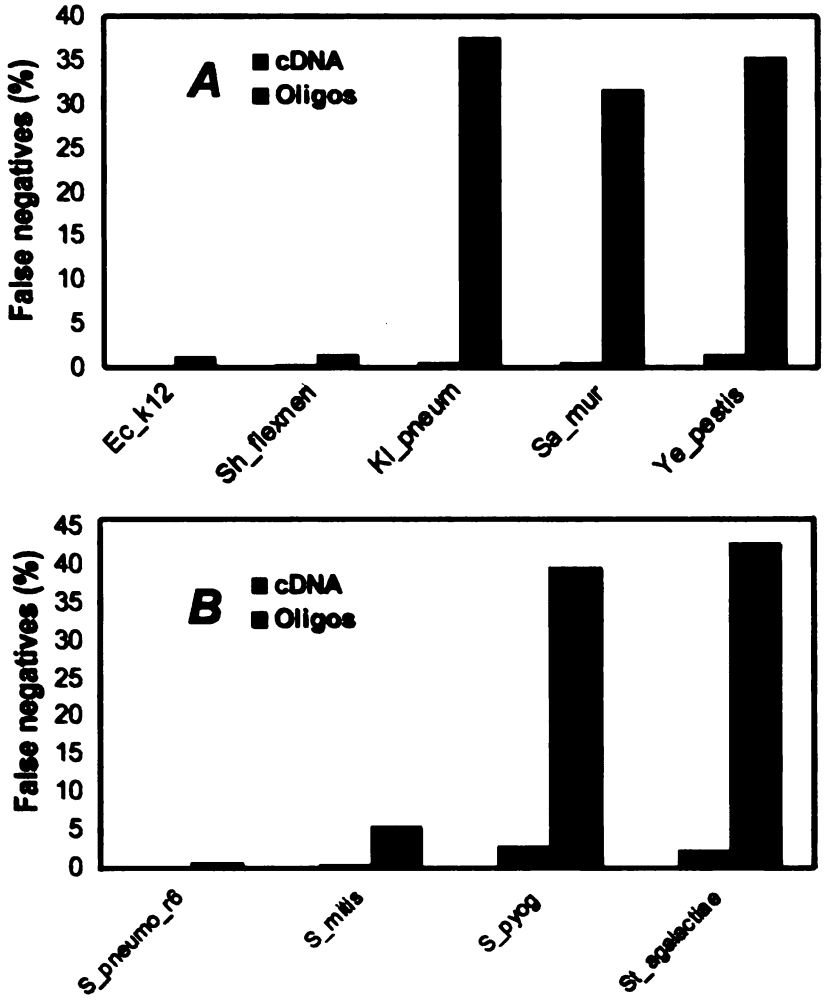


Figure 5.5. cDNA vs Oligo-arrays: Gene identification. Bars represent the predicted false negatives (expressed as percentage of the total number of probes that are expected to hybridize) for the cDNA (open bars) and oligo (solid bars) probes. **(A)**: the enterics. **(B)**: The streptococci. Tester strains (from left to right) are: **(A)**, *E. coli* K12, *Shigella flexneri,* S. Typhimurium, *Klebsiella pneumoniae Y. pestis*; **(B)**, *S. pneumoniae* R6, *S. mitis*, *S. pyogenes*, *S. agalactiae*; reference strains were *E. coli* O157 and *S. pneumoniae* TIGR4, respectively.

DISCUSSION

DNA microarrays have been used for genetic comparisons among strains and have the potential to be used for expression studies with other than the sequenced (reference) strain; but such uses raise potential uncertainties in interpretations. We found that false negatives caused by the different rates of evolution between amino acid and nucleotide sequences comprised a rather small fraction of the total number of ORFs not conserved at the nucleotide level between strains of the same or highly related species (Figure 5.2). This suggests that the total genomic diversity is far more important than false negatives in DNA microarray studies within species. The practical implication of these findings is that, in DNA-DNA studies that attempt to cover a whole species, false negatives should be of secondary importance compared to flexibility in microarray design to accommodate genetic diversity (e.g. more unique sequences). An understanding of the genetic diversity within a species is also required for the successful coverage of the species in such experiments. For example, *M. tuberculosis* and *S. pneumoniae* do not share the genetic diversity of *E. coli* and *Shigella sp.* species, at least based on the available genomic sequences (Figure 5.2).

Experiments with distantly related species are less common and probably involve specialized goals such as taxonomic comparisons. However, microarray false negatives are probably too high to be neglected in this case. The relationship described in this study (Figure 5.1) allows the approximate estimation of the missed genes for a given evolutionary distance between reference and tester strain and a given stringency in the amino acid comparisons. This relationship is probably applicable to bacterial groups besides the ones used in this study because all three groups evaluated gave consistent

results, covered a range of typical bacterial genome sizes (2-5.5 Mb) and included both gram-positive and negative members.

The 30% amino acid identity cut-off gave more false negatives than the 60% amino similarity cut-off due to the lower stringency in sequence comparisons, which selected for more paralogous genes. This was evident in the enterics group where the 30% amino acid identity cut-off predicted a significant number (up to 5%) of false negatives for several *E. coli* or *Shigella sp.* tester strains (Figure 5.2). Indeed, part of the extra DNA in the reference O157 strain compared to these *E. coli* or *Shigella* tester strains involves paralogous genes in expanded gene families and multiple phage copies (26). At the same time and validating its usage, the 30% amino acid identity cut-off predicted almost no false negatives for strain EDL, which is the most closely related, of all strains evaluated, to the reference strain O157 (18); and it predicted very low numbers of false negatives (<1-2%) for the highly related strains of *M. tuberculosis* and *S. pneumoniae*, which is consistent with the decreased genetic diversity within these species compared to *E. coli* (2, 9, 12). On the other hand, the 60% amino acid similarity selected, more frequently than the 30% amino acid identity, the same genes (orthologs) between tester and reference strains and this accounted for the lower numbers of false negatives it typically predicted, particularly within species. Which of the two cut-offs is more suitable depends on the desired stringency in the experiment. It should be mentioned, however, that genes that share 60% or more amino acid similarity are also likely to have diverged in function specificity (although this is less likely than when two genes are related at 30% amino identity) since a few critical amino acid changes could be accompanied by a change in function specificity (10).

The use of the above cut-offs with no manual inspection of the pair-wise alignments proved highly accurate for the prediction of the conserved ORFs between closely related species. For instance, *Fleischmann et al.* (9) identified, based on genomic sequence comparisons, 28 ORFs of *M. tuberculosis* H37Rv not conserved in *M. tuberculosis* strain CDC1551. Our 30% amino acid identity cut-off predicted 29 ORFs for the same comparison (31 for the 60% amino acid similarity cut-off). The low rate of error with closely related species was expected given the low level of sequence divergence between such species. Nonetheless, our approach performed equally satisfactory with distantly related species where the nucleotide divergence is more likely to compromise automated annotations that are based on cut-offs in sequence similarity. For instance, the comparative genomic analysis of the fully sequenced *Streptococcus* species suggested that *S. pneumoniae* TIGR4 shares 1,108 and 1,229 genes with *S. pyogenes* and *S. agalactiae* respectively (23). The 30% amino acid identity cut-off predicted 1,152 and 1,242 conserved ORFs for the same pair of strains, respectively (1,028 and 1,114 ORFs for the 60% amino acid similarity cut-off, respectively).

When DNA microarrays are applied to reveal exact genetic differences, e.g. gene presence or absence, an oligo platform should perform satisfactory with strains of the same or highly related species (Figure 5.5). In this case, DNA-homology value is a better measure of the evolutionary relatedness between tester and reference strains than 16S rRNA identity because it offers better resolution between highly related strains. It should be pointed out, however, that there are too few pairs of moderately related (e.g. DNA-homoology values between 40-60%) strains in the sequenced genome collection for a robust prediction in this critical range.

Oligo-array performance substantially declined (i.e. high rates of false negatives) with distantly related strains however; and this *in-silico* prediction is confirmed by the experimental data to date. For example, oligo-array based genetic comparisons in the streptococci group (12) and in the *Burkholderia* group in our lab (*K. Konstantinidis et al.* unpublished) suggested that strains that are distantly related (e.g. 3-5% 16S rRNA gene miss-pairing) to the reference strain give little hybridization signal relative to the total number of genes conserved based on the genomic sequences. Thus, for experiments with distantly related strains, a cDNA platform should be preferable for its steady performance over this range of evolutionary distance (Figure 5.5). Experimental data with distantly related strains also agree with our predictions for cDNA arrays. *Dong et al.* (6), using a whole genome array that had as probes the whole ORF sequences, have shown that 3,000 ORFs of *E. coli* K12 were conserved (i.e. cross-hybridize) with *K. pneumoniae* strain 342. Our approach predicted 2,890 ORFs of strain K12 to be conserved in *K. pneumoniae* strain M6H 78578 (the sequenced strain). The small difference between our prediction and the experimental results might be due to the different *K. pneumoniae* strain used or ORFs missed in the high draft sequence for strain M9H 78578 or to non-specific signal in the microarray study.

In the case that DNA microarrays are employed to study evolutionary relatedness between species, an oligo array (one 50 mer probe per ORF) will probably give comparable results to a cDNA array (Figure 5.4). This was not surprising inasmuch a 50 mer fragment of an ORF evolves similarly to a larger fragment (e.g. a cDNA probe) or the whole ORF. The oligo platform tended to overestimate distances between distantly related species, however. This is attributable to the difference in information content

between a 50 vs. 718 (on average) nucleotides long sequence for cDNA probes and the lower tolerance of sequence miss-pairing for oligo-probes. Indeed, oligo-probes require, on average, higher sequence identity for cross-hybridization than cDNA probes (>75% vs. 60% identity) (14).

Although our predictions of non-specific signal cannot be absolute because of the complications in quantifying total non-specific signal by adding predicted signal from individual matches, they offered some perspective on this critical issue of microarray technology. And, to the best of our knowledge, no systematic attempt has been ever made to calculate non-specific signal in whole genome DNA-DNA studies. According to our simulation, a significant number of ORFs was affected by non-specific hybridization in any pair of strains evaluated (Figure 5.3). It is also anticipated that any platform, when used for genetic studies, is prone to (at least part of) the non-specific signal revealed for whole ORF sequences in this study. Because, even if the cDNA or oligo-probes are designed to be specific within the reference genome, this does not preclude non-specific hybridization when another genome, which would have different classes of paralogous genes, more copies of genes etc, is used. Such non-specific signal was evident even among strains of the same species (see *E. coli* K12 vs. *E. coli* O157 in Figure 5.3*B*). These findings suggest that misleading conclusions might be reached when non-specific signal is not considered in DNA-DNA microarray studies. On the other hand, if hybridization signal is carefully considered, it has potential to reveal genes and regions that have been duplicated in the tester genome compared to the reference one. Such duplicated regions are likely to play a major role in the unique phenotypic characteristics or ecological niche of the tester strain.

When undertaking microarray approaches both technical and performance issues need to be considered. There are several reviews dealing with technical issues such as flexibility in design, chip technology, probe chemistry, labeling method (cf. references (19, 28). We evaluated the predicted performance of cDNA and oligo arrays as well as false negatives and non-specific hybridization based solely on sequence analysis. Despite certain limitations in the *in-silico* modeling, our results should be a good approximation of reality and can offer useful information in planning appropriate DNA microarray studies. Our results also provide guidance for some experimental tests, which would not only test the validity of our predictions but also enhance predictive ability, especially with moderate and distantly related species.

ACKNOWLEDGEMENTS

We thank TIGR for permission to use preliminary sequence data for *M. avium*, *M. smegmatis*, and *S. mitis*. Sequencing of *M. avium*, *M. smegmatis*, and *S. mitis* was accomplished with support from NIAID and NIH-NIDCR, respectively. We thank Joel Klappenbach and Hector Ayala-del-Rio, for helpful discussions regarding the manuscript. This work was supported by the Bouyoukos Fellowship Program (KTK), the DOE's Microbial Genome Program and the Center for Microbial Ecology.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-402.

2. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520-3.

3. **Brenner, D.** 1984. Bergey's manual of systematic bacteriology, 1st ed, vol. 1. William and Wilkins, Baltimore.

4. **Cho, J. C., and J. M. Tiedje.** 2001. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol **67:**3677-82.

5. **Cole, J. R., B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M. Tiedje.** 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res **31:**442-3.

6. **Dong, Y., J. D. Glasner, F. R. Blattner, and E. W. Triplett.** 2001. Genomic interspecies microarray hybridization: rapid discovery of three thousand genes in the maize endophyte, *Klebsiella pneumoniae* 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. Appl Environ Microbiol **67:**1911-21.

7. **Eisen, J. A.** 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res **8:**163-7.

8. **Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95:**14863-8.

9. **Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jacobs Jr, Jr., J. C. Venter, and C. M. Fraser.** 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol **184:**5479-90.

10. **Gerlt, J. A., and P. C. Babbitt.** 2001. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem **70:**209-46.

11. **Goodfellow, M., and A. O'Donnell.** 1993. Handbook of New Bacterial Systematics. Academic Press Inc, San Diego.

12. **Hakenbeck, R., N. Balmelle, B. Weber, C. Gardes, W. Keck, and A. de Saizieu.** 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. Infect Immun **69:**2477-86.

13. **Imaeda, T.** 1985. Deoxyribonucleic acid relatedness among selected strains of the *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium microti*, and *Mycobacterium africanum*. Int. J. Syst. Bacteriol. **35:**147-150.

14. **Kane, M. D., T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore.** 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res **28:**4552-7.

15. **Kawamura, Y., X. G. Hou, F. Sultana, H. Miura, and T. Ezaki.** 1995. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. Int J Syst Bacteriol **45:**406-8.

16. **Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown.** 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol **14:**1675-80.

17. **Murray, A. E., D. Lies, G. Li, K. Nealson, J. Zhou, and J. M. Tiedje.** 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc Natl Acad Sci U S A **98:**9853-8.

18. **Pupo, G. M., D. K. Karaolis, R. Lan, and P. R. Reeves.** 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect Immun **65:**2685-92.

19. **Relogio, A., C. Schwager, A. Richter, W. Ansorge, and J. Valcarcel.** 2002. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res **30:**e51.

20. **Rost, B.** 1999. Twilight zone of protein sequence alignments. Protein Eng **12:**85-94.

21. **Rouillard, J. M., C. J. Herbert, and M. Zuker.** 2002. OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics **18:**486-7.

22. **Stackebrandt, E., and B. M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol **44:**846-849.

23. **Tettelin, H., V. Masignani, M. J. Cieslewicz, J. A. Eisen, S. Peterson, M. R. Wessels, I. T. Paulsen, K. E. Nelson, I. Margarit, T. D. Read, L. C. Madoff, A. M. Wolf, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, R. T. DeBoy, A. S. Durkin, J. F. Kolonay, R. Madupu, M. R. Lewis, D. Radune, N. B. Fedorova, D. Scanlan, H. Khouri, S. Mulligan, H. A. Carty, R. T. Cline, S. E. Van Aken, J. Gill, M. Scarselli, M. Mora, E. T. Iacobini, C. Brettoni, G. Galli, M. Mariani, F. Vegni, D. Maione, D. Rinaudo, R. Rappuoli, J. L. Telford, D. L. Kasper, G. Grandi, and C. M. Fraser.** 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci U S A **99:**12391-6.

24. **Tonjum, T., D. B. Welty, E. Jantzen, and P. L. Small.** 1998. Differentiation of *Mycobacterium ulcerans*, *M. marinum*, and *M. haemophilum*: mapping of their relationships to *M. tuberculosis* by fatty acid profile analysis, DNA-DNA hybridization, and 16S rRNA gene sequence analysis. J Clin Microbiol **36:**918-25.

25. **Wei, Y., J. M. Lee, C. Richmond, F. R. Blattner, J. A. Rafalski, and R. A. LaRossa.** 2001. High-density microarray-mediated gene expression profiling of *Escherichia coli*. J Bacteriol **183:**545-56.

26. **Welch, R. A., V. Burland, G. Plunkett, III, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. PNAS **99:**17020-17024.

27. **Xu, D., G. Li, L. Wu, J. Zhou, and Y. Xu.** 2002. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics **18:**1432-7.

28. **Yang, Y. H., and T. Speed.** 2002. Design issues for cDNA microarray experiments. Nat Rev Genet **3:**579-88.

29. **Ye, R. W., W. Tao, L. Bedzyk, T. Young, M. Chen, and L. Li.** 2000. Global gene expression profiles of *Bacillus subtilis* grown under anaerobic conditions. J Bacteriol **182:**4458-65.

SUMMARY AND PERSPECTIVES FOR THE FUTURE

Although the evolution process and ecological benefits of symbiotic species with small genomes are well understood, these issues remain poorly elucidated for free-living species with large genomes. Hence, I compared the 115 completed (at the time) prokaryotic genomes to determine whether there are changes with genome size in the proportion of the genome attributable to particular cellular processes since this may reflect both cellular and ecological strategies associated with genome expansion. Large genomes were found to be disproportionately enriched in regulation and secondary metabolism genes and depleted in protein translation, DNA replication, cell division and nucleotide metabolism genes compared to medium and small-sized genomes. Further, large genomes do not accumulate non-coding DNA or hypothetical CDS since the portion of the genome devoted to these functions remained constant with genome size. Traits other than genome size or strain-specific processes are reflected by the dispersion around the average and the current analysis provide means to identify such traits and processes and quantify their importance for every gene functional category or bacterial group of interest. These trends suggest that larger genome-sized species may dominate in environments where resources are scarce but diverse and where there is little penalty for slow growth, such as soil.

Testing this hypothesis is not a trivial task, however. One approach may be to estimate genome sizes of many strains isolated from environmental sources showing different characteristics in terms of resource abundance and availability (for instance soil vs. marine water). The genome size of the isolates can be inferred by the phylogenetic

position of the isolate, e.g., when a closely related isolate with known genome size already exists, or determined by Pulse Field Gel Electrophoresis. Potential pitfalls of this approach are that the evaluated strains must show significant population sizes and activities in their natural environments, i.e., to be ecologically successful as opposed to simply surviving in a dormant or spore state, and represent a phylogenetically unbiased collection. For soil, slow growing isolates should be included in the analysis, when possible, because previous studies suggest that these isolates represent (more) dominant populations in this environment (4, 5). Further, there might be correlation between the time of appearance of an isolate and genome size. One reason for this could be that a large genome-sized species should spend energy to express (at least part of) the increased regulatory proteins it possesses to successfully control its metabolic repertoire. This, all the other equal, might make them to grow slower than smaller genome-sized species.

The species definition for Prokaryotes remains a highly controversial and unsettled issue (2, 8, 9). A comparative analysis -- using gene content derived from genome information -- to identify whether there are species boundaries and determine the role of the organism's ecology on its common gene content was undertaken to better inform the current species definition. It was found that strains of the same species may frequently show too large genetic and functional differences to be considered the same species and that (different) ecology appears to play an important role in these differences. The existence of genetic signatures, i.e., a sizable number of genes of ecological importance, between groups of strains of the same species (current definition) further supports the previous interpretations. The inter-group genetic similarity in several of these cases is as high as 98-99% average nucleotide identity (ANI), indicating that

"species" might be found even among very identical, at the nucleotide level, organisms. Moreover, a large fraction, e.g., up to 65%, of the differences within species (current definition) is associated with bacteriophage and transposase elements, indicating a much more important role of these elements during bacterial speciation than previously expected. The effect of such "mobile" elements on organism's phenotype is currently considered mostly unclear. In conclusion, the results presented in this dissertation support a more stringent and natural definition for prokaryotic species compared to the current one, which should be flexible to accommodate the ecological differences among the organisms.

It is important to realize that the results presented here should be considered as a first step to describe an emerging picture rather than conclusive findings because the available genomes represent only a tiny fraction of the total prokaryotic diversity and are heavily biased towards pathogenic species. Further, in order to obtain a large enough dataset, comparisons between strains of different genera had to be pooled together resulting in clear discontinuities in the results reported. Therefore, a better sampling of species with genomic sequences is still needed to reject, for instance, the hypothesis that there is a continuum of genetic diversity as oppose to species-specific genetic signatures, which is not supportive of a species concept for Prokaryotes. Last, there is inadequate knowledge on the population sizes and activities in the natural environments of most (even the sequenced!) species and hence the quantification of the effect of ecology on the conserved gene content is not currently feasible. Studying natural populations at the genomic level and over time will allow us to more fully evaluate the importance of mobile elements for the process of bacterial speciation as well.

The higher than the species ranks of the prokaryotic taxonomy, i.e., the family, order, class, phylum and domain, are primarily based on phylogenetic analysis of the 16S rRNA gene sequence and secondarily on old observations about the morphological and/or biochemical relatedness of the grouped organisms (2, 6). Phylogenetic clustering based on the genetic distance between two organisms derived from their whole-genome comparison is generally congruent with the clustering based on the 16S rRNA gene, which adds further support to the current classification system. The genomic approach revealed, however, that there is little (if any) predictive power for the currently used higher taxonomic ranks, with the exception of the domain rank, in terms of conserved gene content and genetic distance (measured as the average amino acid identity (AAI) of the conserved genes) of the grouped organisms. Further, organisms of each prokaryotic phylum and several classes may be considered nearly as different from each other as Bacteria are different from Archaea at the genomic level. These findings reveal a much larger genetic and functional diversity for Prokaryotes than previously expected based on the analysis of the 16S rRNA gene.

AAI for longer and ANI for shorter evolutionary scales are simple and highly reliable means to measure relatedness between organisms and evaluate the robustness of genetic markers for phylogenetic purposes. The influence of lateral gene transfer (LGT) on these measures remains to be seen but it is anticipate that it will not be more important than the influence of LGT on measures that are based on single or a few genes such as the 16S rRNA or Multi Locus Sequence Typing (MLST) based approaches. The major limitation in the former measures is that they require the availability of genomic sequences, however, the technological advancements in genomic sequence may render

this less problematic in the near future. Last, AAI and ANI can be uniformly measured for all living organisms, including eukaryotic ones, and therefore contribute towards a uniform taxonomy for all domains of life and provide higher resolution in cases were current methodology is proved inadequate.

DNA microarray technology is currently envisioned as a promising alternative to whole-genome sequencing for genetic (DNA-DNA) comparisons between strains (1, 3), however, several issues regarding the applicability of microarrays for these purposes remain uninvestigated. Using the available genomic sequences (control results) and the existing knowledge on the microarray hybridization kinetics (*in-silico* predicted results), the performance of different microarray platforms for genetic comparisons was first modeled and subsequently evaluated. The number of false negatives, i.e., observing no hybridization signal when the amino acid sequence is conserved but the nucleotide sequence has diverged to a level that does not allow hybridization, were found to be unacceptably high (>10%) between distantly related strains (e.g. <97% 16S rRNA gene identity), but are sufficiently low (<5%) between strains of the same or highly related species (e.g. >98% 16S rRNA gene identity) to not be problematic. Further, oligo-arrays, i.e., one 50mer probe per gene, should give comparable results to whole Open Reading Frame (ORF) arrays as long as the evaluated strains reside in the same or highly related species whereas whole-ORF arrays should perform better with more distantly related strains. Last, a sizeable number of genes (up to 30-35% of the total) in all genomes evaluated appeared to suffer from non-specific hybridization signal from paralogous genes or conserved domains. This non-specific hybridization may lead to significant false positive as well as false negative signal independent of the microarray platform used.

This theoretical analysis assumes that the experimental procedure is ideal, i.e., there are no complications or introduced error during the execution of the experiments. The latter is known to not be true, however, since several technical issues such as probe design and chemistry, labeling method, complications during hybridizations etc. have not yet been fully resolved and thus add complexity to the microarray results. These issues have been extensively reviewed previously (7, 10) and are subjects of ongoing research and continuing improvement. Until the technical aspects of the microarray technology are fully worked out, the results presented here should be a relatively good approximation of reality and can provide guidance for some experimental tests, which would not only test the validity of the predictions described previously but also enhance predictive ability, especially with moderate and distantly related species. Experimental testing of DNA-DNA hybridization (competitive genome hybridization, CGH) among strains by arrays is timely and needed to efficiently advance of understanding of the patterns and order in the high divergent prokaryotic world.

REFERENCES

1. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520-3.

2. **Brenner, D., J. Staley, and N. Krieg.** 2000. Bergey's manual of systematic bacteriology, 2 ed, vol. 1. Springer-Verlag, New York.

3. **Cho, J. C., and J. M. Tiedje.** 2001. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol **67:**3677-82.

4. **Hattori, T., H. Mitsui, H. Haga, N. Wakao, S. Shikano, K. Gorlach, Y. Kasahara, A. el-Beltagy, and R. Hattori.** 1997. Advances in soil microbial ecology and the biodiversity. Antonie Van Leeuwenhoek **72:**21-8.

5. **Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt.** 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol **66:**1328-33.

6. **Ludwig, W., and H.-P. Klenk.** 2000. Bergey's manual of systematic bacteriology, 2 ed, vol. 1. Springer-Verlag, New York.

7. **Relogio, A., C. Schwager, A. Richter, W. Ansorge, and J. Valcarcel.** 2002. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res **30:**e51.

8. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kampfer, M. C. J. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol **52:**1043-1047.

9. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and T. H. G.** 1987. Report of the Ad Hoc Committee on reconciliation of approaches to Bacterial Systematics. Int. J. Syst. Bacteriol. **37:** 463-464.

10. **Yang, Y. H., and T. Speed.** 2002. Design issues for cDNA microarray experiments. Nat Rev Genet **3:**579-88.

APPENDIX

TABLES OD GENOMES USED IN THIS STUDY AND THEIR GENOMIC INFORMATION

Table 2.1 Prokaryotic species and their genomic information used in this study. The genomic sequences for the 115 microbial genomes used in this study were obtained from NCBI. The percent of CDS from each genome homologous to COG database is shown in 4th column. Species that are within one standard deviation from the average (average 70.3%, standard deviation 11.2) are designated solid whereas the rest are designated open (5th column). The solid and open squares are used in the article figures to represent the corresponding genomes.

Table 2.1

Species	Gen. size	Total # ORFs	% ORFs in COGs	Normalize
Aeropyrum_pernix	1.67	1840	66.5	Solid
Agrobacterium_tumefaciens	5.6	5299	80.7	Solid
Agrobacterium_tumefaciens_UWash	5.6	5402	78.9	Solid
Aquifex_aeolicus	1.6	1560	84.2	Open
Archaeoglobus_fulgidus	2.18	2420	78.3	Solid
Bacillus_anthracis	5.1	5311	57.1	Open
Bacillus_cereus	5.41	5255	59.1	Solid
Bacillus_halodurans	4.2	4066	75.6	Solid
Bacillus_subtilis	4.2	4112	73.8	Solid
Bacteroides_thetaiotaomicron	6.26	4778	33.5	Open
Bifidobacterium_longum	2.26	1729	56.9	Open
Borrelia_burgdorferi	0.9	1638	42.8	Open
Bradyrhizobium_japonicum	9.11	8317	60.4	Solid
Brucella_melitensis	3.3	3198	82.1	Open
Brucella_suis	3.28	3264	73.1	Solid
Buchnera_aphidicola	0.62	504	95.4	Open
Buchnera_sp.	0.71	574	96.5	Open
Campylobacter_jejuni	1.64	1634	78.6	Solid
Caulobacter_crescentus	4.01	3737	77.1	Solid
Chlamydia_muridarum	1.07	916	67.6	Solid
Chlamydia_trachomatis	1.05	895	69.6	Solid
Chlamydophila_caviae	1.17	1005	63.5	Solid
Chlamydophila_pneumoniae_AR39	1.23	1112	57.8	Open
Chlamydophila_pneumoniae_CWL029	1.23	1054	61.4	Solid
Chlamydophila_pneumoniae_J138	1.22	1069	60.7	Solid
Chlorobium_tepidum	2.16	2252	52.2	Open
Clostridium_acetobutylicum	4.1	3848	72.6	Solid
Clostridium_perfringens	3.1	2723	64.2	Solid
Clostridium_tetani	2.8	2373	67.4	Solid
Corynebacterium_efficiens	3.15	2950	64.3	Solid
Corynebacterium_glutamicum	3.3	3040	69.5	Solid
Coxiella_bruneti	2	2009	51.6	Open
Deinococcus_radiodurans	3.28	3182	70.2	Solid
Enterococcus_faecalis	3.35	3113	59.9	Solid
Escherichia_coli_K12	4.6	4279	82.7	Open
Escherichia_coli_O157:H7	5.5	5361	73.2	Solid
Escherichia_coli_O157:H7_EDL933	5.6	5324	74.2	Solid
Escherichia_coli_CFT	5.23	5379	69.0	Solid
Fusobacterium_nucleatum	2.17	2067	73.8	Solid
Haemophilus_influenzae	1.83	1714	91.8	Open
Halobacterium_sp._NRC-1	2.57	2622	67.3	Solid
Helicobacter_pylori_26695	1.66	1576	69.9	Solid
Helicobacter_pylori_J99	1.64	1491	71.9	Solid
Lactobacillus_plantarum	3.31	3009	63.5	Solid
Lactococcus_lactis	2.36	2267	77.6	Solid
Leptospira_interrogans	4.69	4727	27.7	Open
Listeria_innocua	3.01	3043	77.6	Solid
Listeria_monocytogenes	2.94	2846	79.7	Solid
Mesorhizobium_loti	7.59	7275	75.7	Solid
Methanococcus_jannaschii	1.74	1785	79.2	Solid
Methanopyrus_kandleri_AV19	1.7	1687	72.6	Solid
Methanosarcina_acetivorans	5.75	4540	67.3	Solid
Methanosarcina_mazei	4.1	3371	68.4	Solid
Methanothermobacter_thermautotrophicus	1.75	1873	77.7	Solid
Mycobacterium_leprae	3.26	1605	70.5	Solid
Mycobacterium_tuberculosis_CDC1551	4.4	4187	63.7	Solid
Mycobacterium_tuberculosis_H37Rv	4.4	3927	70.1	Solid
Mycoplasma_genitalium	0.58	484	77.9	Solid

Table 2.1 (cont'd)

Species	Gen. size	Total # ORFs	% ORFs in COGs	Normalize
Mycoplasma_penetrans	1.36	1037	44.7	Open
Mycoplasma_pneumoniae	0.81	689	60.5	Solid
Mycoplasma_pulmonis	0.96	782	63.8	Solid
Neisseria_meningitidis_MC58	2.27	2079	73.7	Solid
Neisseria_meningitidis_Z2491	2.18	2065	75.0	Solid
Nitrosomonas_europaea	2.81	2461	66.8	Solid
Nostoc_sp.	7.2	6129	58.2	Open
Oceanobacillus_iheyensis	3.63	3496	69.5	Solid
Pasteurella_multocida	2.4	2015	89.2	Open
Pseudomonas_aeruginosa	6.3	5567	80.8	Solid
Pseudomonas_syringae	6.4	5471	71.5	Solid
Psedomonas_putida	6.18	5350	68.7	Solid
Pyrobaculum_aerophilum	2.2	2605	58.1	Open
Pyrococcus_abyssi	1.76	1769	83.7	Open
Pyrococcus_furiosus	1.9	2065	75.7	Solid
Pyrococcus_horikoshii	1.8	1801	77.6	Solid
Ralstonia_solanacearum	5.8	5116	75.0	Solid
Rickettsia_conorii	1.27	1374	65.8	Solid
Rickettsia_prowazekii	1.1	835	84.9	Open
S.enterica_ser._Typhi	4.8	4767	71.8	Solid
Salmonella_typhimurium_LT2	4.95	4553	79.7	Solid
Shewanella_oneidensis	5.03	4472	62.1	Solid
Shigella_flexneri	4.61	4180	83.2	Open
Sinorhizobium_meliloti	6.7	6205	81.7	Open
Staphylococcus_aureus_Mu50	2.9	2748	73.6	Solid
Staphylococcus_aureus_MW2	2.8	2632	73.7	Solid
Staphylococcus_aureus_N315	2.81	2625	77.1	Solid
Staphylococcus_epidermis	2.5	2419	74.5	Solid
Streptococcus_agalactiae	2.2	2124	69.7	Solid
Streptococcus_mutans	2.04	1960	71.9	Solid
Streptococcus_pneumoniae_R6	2.03	2043	78.8	Solid
Streptococcus_pneumoniae_TIGR4	2.2	2094	74.0	Solid
Streptococcus_pyogenes	1.85	1697	77.7	Solid
Streptococcus_pyogenes_MGAS8232	1.9	1845	72.2	Solid
Streptomyces_avermitilis	9.03	7575	48.8	Open
Strepromyces_coelicolor	8.67	7512	48.3	Open
Sulfolobus_solfataricus	2.99	2977	73.4	Solid
Sulfolobus_tokodaii	2.7	2826	60.4	Solid
Synechocystis_sp._PCC_6803	3.57	3167	70.2	Solid
Thermoanaerobacter_tengcongensis	2.7	2588	64.8	Solid
Thermoplasma_acidophilum	1.56	1482	83.5	Open
Thermoplasma_volcanium	1.58	1499	83.7	Open
Thermosynechococcus_elongatus	2.6	2475	65.0	Solid
Thermotoga_maritima	1.85	1858	82.0	Open
Treponema_pallidum	1.14	1036	69.4	Solid
Tropheryma_whipplei	0.93	783	63.5	Solid
Ureaplasma_urealyticum	0.75	614	66.6	Solid
Vibrio_cholerae	4	3835	73.0	Solid
Vibrio_parahaemolyticus	5.18	4537	68.8	Solid
Vibrio_vulnificus	5.13	4832	64.2	Solid
Wigglesworthia_brevipalpis	0.7	654	89.3	Open
Xanthomonas_campestris	5.08	4181	65.9	Solid
Xanthomonas_pv._citri	5.17	4312	63.9	Solid
Xylella_fastidiosa	2.68	2832	60.0	Solid
Xylella_fastidiosa_pv._temecula	2.52	2036	74.5	Solid
Yersinia_pestis	4.65	4083	79.8	Solid
Yersinia_pestis_pv._kim	4.6	4090	77.3	Solid
AVERAGE	**3.24**	**2987**	**70.3 (STDEV: 11.2)**	**87 Solid**

Table 3.1 Genomic information of 64 genomes used in this study and their relatedness to the reference genomes. The 64 strains used in this study (2nd column), their genome size (3rd column) and total CDSs in the genome (4th column) are shown. Strains in bold were used as reference genomes during the pair-wise comparisons between strains of the same bacterial group (1st column). The groups used are (from top to bottom): Enterics, Pseudomonas, Neisseria, Bordetella, Bacilli, Mycobacteria, Streptococcus, Staphylococcus, Others. NA = Not available, because the genome annotation has not been published. *Average nucleotide identity of the conserved CDSs between the corresponding strain and the reference strain, which is denoted by the superscript number. +Percent of reference strain CDSs that are conserved in the corresponding strain.

Table 3.1.

Group	Strain	Gen. Size	Total CDSs	Average nt identity*	Percent conser. genes +
ENTERICS	**1.E. coli O157:H7 Sakai**	5.50	5361		
	2.E. coli 0157:H7 EDL933	5.60	5324	99.7[1],97.4[3],97.3[5]	98.6[1],90.2[3],89.8[5]
	3.E. coli K12	4.60	4279	97.2[1],97.9[5]	72.9[1],88.1[5]
	4.E. coli CFT073	5.23	5379	95.9[1],96.4[3],96.5[5]	75.5[1],86.8[3],88.8[5]
	5.S. flexneri 2a 2457	4.60	4068	96.5[1],97.5[3]	69.6[1],82.6[3]
	6.S. flexneri 2a 301	4.61	4180	96.4[1],97.5[3],99.8[5]	69.5[1],82.3[3],99[5]
	7.S. typhimurium LT2	4.95	4553	79.9[1],80.7[5]	59.7[1],68.6[5]
	8.S. enterica ser. Typhi Ty2	4.79	4323	80.2[1],	57[1],
	9.S. enterica ser. Typhi	4.80	4767	80.2[1],99.9[8]	57.7[1],99.6[8]
	10.S. bongori 12419	4.46	NA	89.5[8]	78.8[8]
	11.Y. pestis Kim	4.60	4090	71.5[1],71.6[5]	37[1],45.1[5]
	12.Y. pestis CO92	4.65	4083	71.5[1],99.9[11]	37.2[1],99.7[11]
	13.Y. pestis Mediaevails	4.6	4142	99.88[11]	98.56[11]
	14.Y. enterocolitica	4.68	NA	82.1[1]	69.3[1]
	15.E. carotovora	5.06	NA	72.1[1]	38.4[1]
PSEUDOMONAS	**1.P putida KT2440**	6.18	5350		
	2.P aeruginosa PA01	6.30	5567	75.1	56.2
	3.P syringae DC3000	6.40	5471	75.4	50.7
NEISSERIA	**1.N. meningitidis MC58**	2.27	2079	96.7[3]	91.2[3]
	2.N. meningitidis FAM	2.20	NA	97[1],97.1[3]	90.8[1],90.2[3]
	3.N. meningitidis Z2491	2.18	2065	96.9[1]	90.2[1]
	4.N. gonorrhoeae FA1090	2.15	NA	94.3[1],94.3[3]	82.4[1],82.7[3]
BORDETELLA	**1.B pertussis Tohama I**	4.09	3447	96.6[2]	72.1[2]
	2.B. bronchiceptica RB50	5.35	4994	98.4[1]	91[1]
	3.B. parapertussis	4.77	4185	98.2[1],98.3[2]	87.9[1],87.2[2]
BACILLI	**1.B cereus ATCC 14579**	5.41	5255		
	2.B cereus 10987	5.22	NA	91	83
	3.B anthracis A2012	5.10	5311	91.2	85.3
MYCOBACTERIA	**1.M. tuberculosis CDC1551**	4.40	4187		
	2.M. tuberculosis h37Rv	4.40	3927	99.7	99.5
	3.M. bovis	4.35	3920	99.4	98.3
	4.M. avium	5.48	NA	79.1	62.6
STREPTOCOCCI	**1.S. pyogenes SSI-1**	1.89	1861	97.9[2]	92.7[2]
	2.S. pyogenes MGAS8232	1.90	1845	97.8[1]	92.5[1]
	3.S. pyogenes MGAS315	1.90	1865	99.9[1],97.9[2]	100[1],92.7[2]
	4.S. pyogenes M1 GAS	1.85	1697	98[1],97.9[2],71.3[7]	86.8[1],89.7[2],38.5[7]
	5.S. agalactiae 2603	2.20	2124	74.5[1],74.4[2]	56.2[1],56.5[2]
	6.S. agalactiae NEM	2.21	2094	98.5[5]	87.5[5]
	7.S. pneumoniae r6	2.03	2043		
	8.S. pneumoniae TIGR4	2.20	2094	71.4[1],71.3[2],98.4[7]	41.3[1],42.1[2],95.9[7]
	9.S. mutans	2.03	1920	72[5]	46.8[5]

Table 3.1 (cont'd)

Group	Strain	Gen. Size	Total CDSs	Average nt identity*	Percent of conser. genes+
STAPHYLOCOCCI	**1.S aureus Mu50**	2.90	2748	98.2[3]	93.5[3]
	2.S aureus N315	2.81	2625	99.8[1],98.3[3]	94.8[1],93.3[3]
	3.S aureus MW2	2.80	2632	98.2[1]	90.4[1]
	4.S aureus MSSR252	2.80	NA	98.2[1],99.7[3]	89.7[1],98.4[3]
	5.S aureus MRSA476	2.90	NA	97.1[1],97[3]	91.6[1],93.9[3]
	6.S. epidermitidis ATCC12228	2.50	2419		
	7.S. epidermitidis RP62A	2.65	NA	75.9[1],75.3[3],98.9[6]	66.4[1],66.8[3],94.6[6]
OTHERS	**1.X. campestris ATCC33913**	5.08	4181		
	2.X. axonopodis pv. Citri 301	5.17	4312	84.6[1]	82.9[1]
	3.X. fastidiosa temecula	2.52	2036		
	4.X. fastidiosa 9a5c	2.68	2832	95.7[3]	95.8[3]
	5.Vibrio vulnificus CMCP6	5.13	4514		
	6.Vibrio vulnificus YJ016	5.21	5024	97.91[5]	89.5[5]
	7.H. pylori J99	1.64	1491		
	8.H. pylori 26695	1.66	1576	95[7]	93.3[7]
	9.B. melitensis 16M	3.30	3198		
	10.B suis 1330	3.28	3264	99.1[9]	98.4[9]
	11.R. conorri	1.27	1374		
	12.R. prowazekii	1.10	835	87.7[11]	59.4[11]
	13.C perfrigens 13	3.1	2723		
	14.C perfrigens ATCC13124	3.26	NA	98.1[13]	90.5[13]
	15.T. whipplei twist	0.93	808		
	16.T. whipplei TW08/27	0.93	783	99.2[15]	99.3[15]

Table 4.1. Taxonomic information of the 175 genomes used in this study. The taxonomic information for each genome was extracted from the Hierarchy browser of the RDP database, release 9 (http://rdp.cme.msu.edu/index.jsp), which implements the newer version of Bergey's taxonomy. The genome size (in Mb) and the total number of CDS in the genome are also shown in the last two columns. Abbreviations of 4th column: D. -- Domain, B -- Bacteria, A -- Archaea.

Table 4.1.

GENUS	SPECIES	STRAIN	D, PHYLUM	CLASS	ORDER	FAMILY	Size	# CDS
Acinetobacter	sp.	ADP1	B Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	3.6	3325
Aeropyrum	pernix	K1	A Crenarchaeota	Thermophilic cren.	Pyrodictiales	AP pernix subg.	1.67	1840
Agrobacterium	tumefaciens	C58+C4	B Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	5.67	5402
Aquifex	aeolicus	VF5	B Aquificae	Aquificae	Aquificales	Aquificaceae	1.59	1560
Archaeoglobus	fulgidus	DSM_4304	A Euryarchaeota	Methanomicrobacteria	Archaeoglobales	unclassified	2.18	2420
Bacillus	anthracis	Ames_0581	B Firmicutes	Bacilli	Bacillales	Bacillaceae	5.5	5617
Bacillus	anthracis	A2012	B Firmicutes	Bacilli	Bacillales	Bacillaceae	5.37	5544
Bacillus	anthracis	Ames	B Firmicutes	Bacilli	Bacillales	Bacillaceae	5.23	5311
Bacillus	cereus	ATCC_10987	B Firmicutes	Bacilli	Bacillales	Bacillaceae	5.22	5603
Bacillus	cereus	ATCC_14579	B Firmicutes	Bacilli	Bacillales	Bacillaceae	5.43	5255
Bacillus	halodurans	C-125	B Firmicutes	Bacilli	Bacillales	Bacillaceae	4.2	4066
Bacillus	subtilis	str. 168	B Firmicutes	Bacilli	Bacillales	Bacillaceae	4.21	4112
Bacillus	thuringiensis	str. 97-27	B Firmicutes	Bacilli	Bacillales	Bacillaceae	5.24	5117
Bacteroides	thetaiotaomicron	VPI-5482	B Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	6.29	4778
Bartonella	henselae	Houston-1	B Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	1.93	1488
Bartonella	quintana	Toulose	B Proteobacteria	Alphaproteobacteria	Rhizobiales	Bartonellaceae	1.58	1142
Bdellovibrio	bacteriovorus	HD100	B Proteobacteria	Deltaproteobacteria	unclassified	unclassified	3.78	3587
Bifidobacterium	longum	NCC2705	B Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	2.26	1729
Blochmannia	floridanus	Candidatus	B Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	0.71	583
Bordetella	bronchiseptica	RB50	B Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	5.34	4994
Bordetella	parapertussis	str. 12822	B Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	4.77	4185
Bordetella	pertussis	Tohama_I	B Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	4.09	3436
Borrelia	burgdorferi	B31	B Spirochaetes	Spirochaetes	Spirochaetales	Spirochaetaceae	1.52	1638
Bradyrhizobium	japonicum	USDA 110	B Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	9.11	8316
Brucella	melitensis	16M	B Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	3.29	3196
Brucella	suis	str. 1330	B Proteobacteria	Alphaproteobacteria	Rhizobiales	Brucellaceae	3.32	3264
Buchnera	aphidicola	Bp	B Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	0.62	504
Buchnera	aphidicola	Sg	B Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	0.64	546
Buchnera	aphidicola	APS	B Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	0.66	574
Campylobacter	jejuni	NCTC 11168	B Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	1.64	1634
Caulobacter	crescentus	CB15	B Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	4.02	3737
Chlamydia	muridarum	MoPn	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.08	916
Chlamydia	trachomatis	D/UW-3/CX	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.04	895
Chlamydophila	caviae	GPIC	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.18	1005
Chlamydophila	pneumoniae	AR39	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.23	1112
Chlamydophila	pneumoniae	CWL029	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.23	1054
Chlamydophila	pneumoniae	J138	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.23	1069
Chlamydophila	pneumoniae	TW-183	B Chlamydiae	Chlamydiae	Chlamydiales	Chlamydiaceae	1.23	1113
Chlorobium	tepidum	TLS	B Chlorobi	Chlorobia	Chlorobiales	Chlorobiaceae	2.15	2252

Table 4.1 (cont'd)

GENUS	SPECIES	STRAIN	D.	PHYLUM	CLASS	ORDER	FAMILY	Size	# CDS
Chromobacterium	violaceum	ATCC_12472	B	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	4.75	4407
Clostridium	acetobutylicum	ATCC824	B	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	4.13	3848
Clostridium	perfringens	str. 13	B	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	3.09	2723
Clostridium	tetani	E88	B	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	2.87	2373
Corynebacterium	diphtheriae	NCTC_13129	B	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	2.49	2272
Corynebacterium	efficiens	YS-314	B	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	3.15	2950
Corynebacterium	glutamicum	ATCC_13032	B	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	3.31	3040
Coxiella	burnetii	RSA_493	B	Proteobacteria	Gammaproteobacteria	Legionellales	Coxiellaceae	2.03	2009
Deinococcus	radiodurans	R1	B	Deinoc.-Thermus	Deinococci	Deinococcales	Deinococcaceae	3.28	3182
Desulfotalea	psychrophila	LSv54	B	Proteobacteria	Deltaproteobacteria	Desulfobacterales	Desulfobulbaceae	3.66	3236
Desulfovibrio	vulgaris	Hildenborough	B	Proteobacteria	Deltaproteobacteria	Desulfovibrionales	Desulfovibrionaceae	3.77	3531
Enterococcus	faecalis	V583	B	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	3.36	3113
Erwinia	carotovora	SCRI_1043	B	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	5.06	4472
Escherichia	coli	CFT073	B	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	5.23	5379
Escherichia	coli	EDL933	B	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	5.53	5324
Escherichia	coli	K12	B	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	4.64	4279
Escherichia	coli	Sakai	B	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	5.5	5361
Fusobacterium	nucleatum	ATCC_25586	B	Fusobacteria	Fusobacteria	Fusobacteriales	Fusobacteriaceae	2.17	2067
Geobacter	sulfurreducens	PCA	B	Proteobacteria	Deltaproteobacteria	Desulfuromonales	Geobacteraceae	3.81	3445
Gloeobacter	violaceus	PCC 7421	B	Cyanobacteria	Cyanobacteria	Deferribacterales	Family_1.1	4.66	4430
Haemophilus	ducreyi	35000HP	B	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	1.7	1717
Haemophilus	influenzae	Rd KW20	B	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	1.83	1714
Halobacterium	sp.	NRC-1	A	Euryarchaeota	Methanomicrobacteria	Halophilic Archaea	unclassified	2.57	2622
Helicobacter	hepaticus	ATCC_51449	B	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	1.8	1875
Helicobacter	pylori	str. 26695	B	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	1.67	1576
Helicobacter	pylori	J99	B	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	1.64	1491
Lactobacillus	johnsonii	NCC_533	B	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	1.99	1821
Lactobacillus	plantarum	WCFS1	B	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	3.35	3009
Lactococcus	lactis	I_11403	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	2.37	2267
Leifsonia	xyli	CTCB07	B	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	2.58	2030
Leptospira	interrogans	str. 56601	B	Spirochaetes	Spirochaetes	Spirochaetales	Leptospiraceae	4.69	4727
Leptospira	interrogans	Fiocruz L1-130	B	Spirochaetes	Spirochaetes	Spirochaetales	Leptospiraceae	4.63	3658
Listeria	innocua	Clip11262	B	Firmicutes	Bacilli	Bacillales	Listeriaceae	3.09	3043
Listeria	monocytogenes	4b F2365	B	Firmicutes	Bacilli	Bacillales	Listeriaceae	2.91	2821
Listeria	monocytogenes	EGD-e	B	Firmicutes	Bacilli	Bacillales	Listeriaceae	2.94	2846
Mesoplasma	florum	L1	B	Firmicutes	Mollicutes	Entomoplasmatales	Entomoplasmataceae	0.79	683
Mesorhizobium	loti	MAFF_303099	B	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	7.6	7275
Methanobacterium	thermautotr.	Delta H	A	Euryarchaeota	Methanobacteriales	Methanobacterium	MB_Thermoautotr.	1.75	1873
Methanococcus	jannaschii	DSM_2661	A	Euryarchaeota	Methanococcales	MC_jannaschii	unclassified	1.74	1785

Table 4.1 (cont'd)

GENUS	SPECIES	STRAIN	D.	PHYLUM	CLASS	ORDER	FAMILY	Size # CDS
Methanococcus	maripaludis	S2	A	Euryarchaeota	Methanococcales	MC maripaludis	unclassified	1.66 1722
Methanopyrus	kandleri	AV19	A	Euryarchaeota	Methanopyrales	unclassified	unclassified	1.69 1687
Methanosarcina	acetivorans	C2A	A	Euryarchaeota	Methanomicrobacteria	Methanosarcinales	Methanosarcina	5.75 4540
Methanosarcina	mazei	Goe1	A	Euryarchaeota	Methanomicrobacteria	Methanosarcinales	Methanosarcina	4.1 3371
Mycobacterium	avium	k10	B	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	4.83 4350
Mycobacterium	bovis	AF212297	B	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	4.35 3920
Mycobacterium	leprae	TN	B	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	3.27 1605
Mycobacterium	tuberculosis	CDC1551	B	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	4.4 4167
Mycobacterium	tuberculosis	H37Rv	B	Actinobacteria	Actinobacteria	Actinomycetales	Mycobacteriaceae	4.41 3927
Mycoplasma	gallisepticum	R	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	1 726
Mycoplasma	genitalium	G-37	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	0.58 484
Mycoplasma	mobile	163K	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	0.78 633
Mycoplasma	mycoides	PD1	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	1.21 1016
Mycoplasma	penetrans	HG-2	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	1.36 1037
Mycoplasma	pneumoniae	M129	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	0.82 689
Mycoplasma	pulmonis	UAB CTIP	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	0.96 782
Nanoarchaeum	equitans	Kin4-M	A	unclassified	unclassified	unclassified	unclassified	0.49 536
Neisseria	meningitidis	MC58	B	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	2.27 2079
Neisseria	meningitidis	Z2491	B	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	2.18 2065
Nitrosomonas	europaea	ATCC 19718	B	Proteobacteria	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae	2.81 2461
Nostoc	sp.	PCC 7120	B	Cyanobacteria	Cyanobacteria	Bacillales	Family_4.1	7.21 6129
Oceanobacillus	iheyensis	HTE831	B	Firmicutes	Bacilli	Bacillales	Bacillaceae	3.63 3496
Onion	yellows	OY-M	B	Firmicutes	Mollicutes	Acholeplasmatales	Acholeplasmataceae	0.86 754
Parachlamydia	sp.	UWE25	B	Chlamydiae	Chlamydiae	Chlamydiales	Parachlamydiaceae	2.41 2031
Pasteurella	multocida	Pm70	B	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	2.26 2015
Photorhabdus	luminescens	TT01	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	5.69 4683
Picrophilus	torridus	DSM 9790	A	Euryarchaeota	Methanomicrobacteria	Thermoplasmales	Picrophilaceae	1.55 1535
Pirellula	sp.	str. 1	B	Planctomycetes	Planctomycetacia	Planctomycetales	Planctomycetaceae	7.15 7325
Porphyromonas	gingivalis	W83	B	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae	2.34 1909
Prochlorococcus	marinus	MIT_9313	B	Cyanobacteria	Cyanobacteria	Deferribacterales	Family_1.1	2.41 2265
Prochlorococcus	marinus	CCMP1375	B	Cyanobacteria	Cyanobacteria	Deferribacterales	Family_1.1	1.75 1862
Prochlorococcus	marinus	CCMP1986	B	Cyanobacteria	Cyanobacteria	Deferribacterales	Family_1.1	1.66 1712
Propionibacterium	acnes	KPA_171202	B	Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	2.56 2297
Pseudomonas	aeruginosa	PAO1	B	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	6.26 5567
Pseudomonas	putida	KT2440	B	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	6.18 5350
Pseudomonas	syringae	DC3000	B	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	6.54 5471
Pyrobaculum	aerophilum	IM2	A	Crenarchaeota	Thermophilic cren.	Pyb islandicum	unclassified	2.22 2605
Pyrococcus	abyssi	Ge5	A	Euryarchaeota	Thermococcales	PC furiosus group	unclassified	1.77 1769
Pyrococcus	furiosus	DSM 3638	A	Euryarchaeota	Thermococcales	PC furiosus group	unclassified	1.91 2065

Table 4.1 (cont'd)

GENUS	SPECIES	STRAIN	D.	PHYLUM	CLASS	ORDER	FAMILY	Size	# CDS
Pyrococcus	horikoshii	OT3	A	Euryarchaeota	Thermococcales	PC. furiosus group	unclassified	1.74	1801
Ralstonia	solanacearum	GMI_1000	B	Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiaceae	5.81	5116
Rhodopseudom	palustris	CGA_009	B	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	5.46	4814
Rickettsia	conorii	Malish_7	B	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	1.27	1374
Rickettsia	prowazekii	Madrid_E	B	Proteobacteria	Alphaproteobacteria	Rickettsiales	Rickettsiaceae	1.11	835
Salmonella	enterica	Ty2	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.79	4318
Salmonella	enterica	CT18	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	5.13	4767
Salmonella	enterica	LT2	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.95	4553
Shewanella	oneidensis	MR-1	B	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	5.13	4472
Shigella	flexneri	2457T	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.6	4068
Shigella	flexneri	str. 301	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.6	4180
Sinorhizobium	meliloti	str. 1021	B	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	6.69	6205
Staphylococcus	aureus	MRSA_252	B	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	2.9	2656
Staphylococcus	aureus	MSSA_476	B	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	2.82	2579
Staphylococcus	aureus	Mu50	B	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	2.9	2748
Staphylococcus	aureus	MW2	B	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	2.82	2632
Staphylococcus	aureus	N315	B	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	2.64	2625
Staphylococcus	epidermidis	ATCC_12228	B	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	2.5	2419
Streptococcus	agalactiae	2603V/R	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	2.16	2124
Streptococcus	agalactiae	NEM316	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	2.21	2094
Streptococcus	mutans	UA159	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	2.03	1960
Streptococcus	pneumoniae	TIGR4	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	2.16	2094
Streptococcus	pneumoniae	R6	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	2.04	2043
Streptococcus	pyogenes	M1_GAS	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	1.85	1697
Streptococcus	pyogenes	MGAS10394	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	1.9	1866
Streptococcus	pyogenes	MGAS315	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	1.9	1865
Streptococcus	pyogenes	SSI-1	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	1.89	1861
Streptococcus	pyogenes	MGAS8232	B	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	1.9	1845
Streptomyces	avermitilis	MA-4680	B	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	9.12	7575
Streptomyces	coelicolor	A3	B	Actinobacteria	Actinobacteria	Actinomycetales	Streptomycetaceae	9.05	7512
Sulfolobus	solfataricus	P2	A	Crenarchaeota	Thermophilic cren.	Sulfolobus subg. I	SUL sulfotaricus	2.99	2977
Sulfolobus	tokodaii	str. 77	A	Crenarchaeota	Thermophilic cren.	Sulfolobus subg. I	unclassified	2.69	2826
Synechococcus	sp.	WH_8102	B	Cyanobacteria	Cyanobacteria	Deferribacterales	Family_1.1	2.43	2517
Synechocystis	sp.	PCC_6803	B	Cyanobacteria	Cyanobacteria	Deferribacterales	Family_1.1	3.57	3167
Thermoanaerob	tengcongensis	MB4T	B	Firmicutes	Clostridia	Thermoanaerobact.	Thermoanaerobacter.	2.69	2588
Thermoplasma	acidophilum	DSM 1728	A	Euryarchaeota	Methanomicrobacteria	Thermoplasmales	TPL acidophilum	1.56	1462
Thermoplasma	volcanium	GSS1	A	Euryarchaeota	Methanomicrobacteria	Thermoplasmales	unclassified	1.58	1499
Thermosynech.	elongatus	BP-1	B	Cyanobacteria	Cyanobacteria	Deferribacterales	unclassified	2.59	2475
Thermotoga	maritima	MSB8	B	Thermotogae	Thermotogae	Thermotogales	Thermotogaceae	1.86	1858

Table 4.1 (cont'd)

GENUS	SPECIES	STRAIN	D.	PHYLUM	CLASS	ORDER	FAMILY	Size	# CDS
Thermus	thermophilus	HB27	B	Deinoc.-Thermus	Deinococci	Thermales	Thermaceae	2.13	2210
Treponema	denticola	ATCC_35405	B	Spirochaetes	Spirochaetes	Spirochaetales	Spirochaetaceae	2.84	2767
Treponema	pallidum	Nichols	B	Spirochaetes	Spirochaetes	Spirochaetales	Spirochaetaceae	1.14	1036
Tropheryma	whipplei	Twist	B	Actinobacteria	Actinobacteria	Actinomycetales	Cellulomonadaceae	0.93	783
Ureaplasma	urealyticum	ATCC 700970	B	Firmicutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	0.75	614
Vibrio	cholerae	N16961	B	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	4.03	3835
Vibrio	parahaemolyticus	RIMD_2210633	B	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	5.17	4832
Vibrio	vulnificus	CMCP6	B	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	5.13	4537
Vibrio	vulnificus	YJ016	B	Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	5.26	5024
Wigglesworthia	glossinidia	brevipalpis	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	0.7	654
Wolbachia	sp.	melanogaster	B	Proteobacteria	Alphaproteobacteria	Rickettsiales	Anaplasmataceae	1.27	1195
Wolinella	succinogenes	DSM 1740	B	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	2.11	4181
Xanthomonas	axonopodis	str_306	B	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	5.17	2044
Xanthomonas	campestris	ATCC_33913	B	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	5.08	4312
Xylella	fastidiosa	9a5c	B	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	2.73	2832
Xylella	fastidiosa	Temecula1	B	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	2.52	2036
Yersinia	pestis	str_91001	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.8	3895
Yersinia	pestis	CO92	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.83	4083
Yersinia	pestis	KIM	B	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	4.6	4090