This is to certify that the
dissertation entitled

OPEN READING FRAME COMPOSITION AND ORGANIZATION AS INDICATORS OF PHENOTYPIC DIVERSITY IN BACTERIA AND ARCHAEA

presented by

SCOTT HENRY HARRISON

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Microbiology and Molecular Genetics

Julius P. Jackson
Major Professor's Signature

7.14.2006
Date

MSU is an Affirmative Action/Equal Opportunity Institution

OPEN READING FRAME COMPOSITION AND ORGANIZATION AS INDICATORS OF PHENOTYPIC DIVERSITY IN BACTERIA AND ARCHAEA

By

SCOTT HENRY HARRISON

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Microbiology and Molecular Genetics

2006

ABSTRACT

OPEN READING FRAME COMPOSITION AND ORGANIZATION AS INDICATORS OF PHENOTYPIC DIVERSITY IN BACTERIA AND ARCHAEA

By

SCOTT HENRY HARRISON

Phenotypically, intragenomic recombination enables prokaryotic organisms to respond to dramatic changes in environmental conditions by restructuring the genome. The relationship between adaptation and alterations to genome structure over time impacts phylogeny and relates to factors regarding the optimal physiological configuration of genome structure. This study provides a quantitative treatment of open reading frame (ORF) organization based on aspects of functional conservation and DNA mobility. An analytical software system was built to facilitate randomizations, subsamplings, and comparative treatments of calculated and organized measures of open reading frame (ORF) attributes encompassing 447,551 annotated ORFs from 155 fully sequenced prokaryotic genomes. An operational subset of ORFs (O-ORFs) of putative phenotypic importance was selected based on a simple heuristic of similar length and content in comparison to five or more other ORFs. The proportion of total, annotated ORFs represented by O-ORFs strongly correlated with a predicted 3:1 signal-to-noise ratio of O-ORFs, likely associated with some phenotype, to putatively silent ORFs (S-ORFs) of unknown and undefined phenotype. The O-ORF subset had a significant degree of clustered chromosomal organization across a broad phylogenetic range. Additional study of ORF organization was conducted by developing quantitative measures of ORF clustering based on segmentation of the chromosomal sequence into consecutive regions of specified scalings. Properties associated with performance of non-parametric measures were partly characterized by simulation using an extended model of an abstract expansion modification system. Measures of ORF organization were evaluated as potential signatures of the recombinational history of an organism. As predicted by a postulated relationship between genomic organization and phylogenetic relatedness, the measurements had significant correspondence with times of divergence from last common ancestors. The presence of mobile elements predictably correlated with greater deviations from organizational symmetries of ORFs.

In Memory of Richard John Harrison

ACKNOWLEDGEMENTS

I would foremostly like to acknowledge the many in my family, especially my brothers, Richard and Timothy, who for me have always set an example of conscientious ethic and genuine dialogue whatever the season or circumstance. I am grateful to my wife, Tara, and my children for their love, good humor, and kindness. My beloved wife Tara provided an extraordinary degree of comforting support, especially when I was working at greatest intensity.

Special thanks goes to my mentor, Julius H. Jackson, for the many thoughtful conversations we have shared over the years, and his openness and encouragment of innovative effort. Both in terms of magnitude and consistency, Dr. Jackson's assessments of the literature, mentoring of students, demands for rigor with systems-based approaches, and extensive knowledge concerning microbial biochemistry, physiology, and dynamic analysis, reflect the type of scholarly devotion I most hope to emulate in the years ahead. I deeply valued the opportunity to work with and assist the many undergraduate students who were members of Dr. Jackson's laboratory. On both a personal and professional level, Dr. Jackson's efforts towards me, a student and aspiring scholar, will always be for me an example of Socratic virtue in its finest form.

I am grateful to my committee for demanding detail and excellence, for being pithy when I was being protracted, and for their strong, unmitigating sense of scholarship when I was a beginning student of research and science. They are Brian Feeny, Michele Fluck, Patricia Herring, Richard Lenski, and Larry Snyder.

There are also colleagues, friends and faculty to whom I am indebted for their encouragement and demonstrations of professionalism. They include Janet Batzli, Cedric Buckley, James Cole, Tony D'Angelo, Les Dethlefsen, Diane Ebert-May, George Garrity, Feng Han, Helen Keefe, Gerd Kortemeyer, Niels Larsen, Douglas Luckie, Desmond Stephens, and Bill Uicker. I am a better person for having known them. I would like to also make special mention of my high school biology teacher, Phil Browne, who always made biology an exciting and intriguing subject of study.

There are not the words nor the pages to fully describe the contributions of scholars and

others who have played an invaluable role in my development as an aspiring researcher of the natural sciences. I am especially grateful to those, such as Harvey and Pagel (1991) and Torretti (1984), who have each put together illuminating texts on how to assess pattern in the natural world. Also, the bioinformatics aspect to my effort greatly benefitted from the availability of powerful, reliable and highly flexible free and open source software tools.

I am deeply grateful to those who have invested and sacrificed their time, energies, lives and fortunes to safeguard and champion how it is that ongoing generations of persons can so freely and openly pursue the discovery and advancement of knowledge in a democratic society. I would be sorely remiss if I were not to thank my parents for all of their many unheralded acts of love that I believe define the essence of such a profoundly meaningful purpose. Thanks, mom and dad, for emphasizing integrity and giving me a start in life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Images in this dissertation are presented in color.

Chapter 1: Introduction

1.1 Overview

Recombinations are one of two major forms of heritable change in prokaryotic genomes, and the consequence of a recombination event is to restructure the organization and composition of genomic elements such as open reading frames (ORFs) (Brown, 2002). The other form of heritable change, sequence-level mutations, has been well studied in terms of evolutionary models and comparative treatments (Zuckerkandl & Pauling, 1965; Woese, 1987; Ochman *et al.*, 1999). There exists a database of functionally grouped clusters of orthologous genes that characterizes those genes with conserved sequences implicating a directly vertical last common ancestor (Tatusov *et al.*, 1997a, 2003). There is not however a database with a function-based cataloguing of the formations and disruptions of genomic structure. Characterizing prokaryotic diversity in terms of the functional aspects of recombinative change may help develop current proposals as to how expansions and modifications of genomic structure relate to specific phenotypes of an organism (Bentley & Parkhill, 2004; Cohan, 2004; Moran & Plague, 2004; Ochman & Davalos, 2006).

Significant phenotypic change does not necessarily correspond with conventional evaluations of conserved sequence and genomic structure. *Mycobacterium avium* subspecies *avium* is frequently encountered in the environment and causes infections in certain animals and immunocompromised patients. *Mycobacterium avium* subspecies *paratuberculosis* has significantly different behavior than *M. avium* in terms of a much slower growth rate and different pathogenicity. Yet, when *M. avium* subspecies *paratuberculosis* is compared to *M. avium* subspecies *avium*, it is almost identical for 16S rRNA sequence, conserved genomic organization, and the region surrounding *oriC* (Bannantine *et al.*, 2003). Furthermore, it is a complex question to consider how phenotypic categorizations of diversity may efficiently account for the variable lifestyles, ecologies, ancestral lineages, appearances and behaviors of prokaryotic organisms. Taxonomically, for prokaryotes, there is not yet "a consensus for defining the fundamental unit of biological diversity, the species" (Cohan, 2002).

Reconstructions of the prokaryotic phylogeny have often been based upon patterns of

sequence similarity as a primary means for asserting homology (inferred origin of DNA sequence from the same ancestral sequence) (Morell, 1997; Zhaxybayeva *et al.*, 2004; Patterson, 1988). Differences between sequences help reveal branch points in the phylogeny (Pearson & Lipman, 1988) as well as the rates at which mutating random events occur from various branch points to the present (Ochman & Wilson, 1987; Ochman *et al.*, 1999). Sequence-level changes themselves do not fully resolve how surrounding changes in evolutionary rate may occur. For higher taxa, it is especially difficult to see how a procession of changes in ribosomal sequence may closely follow significant alterations in proteomic content. β-proteobacteria differ from α-proteobacteria not just by ribosomal sequence but also by photoreaction centers, cytochrome *c* type, and having cytochromes of the small type compared to the medium-large type (Woese, 1987). More comprehensive assays of genome structure and content may provide the empirical information needed to directly characterize the emergence of such properties within the proteome.

Inspection of synteny (conserved arrangement of genome structure) (Horimoto *et al.*, 2001; Kalman *et al.*, 1999; Rocap *et al.*, 2003) can be a means for inferring branch points based on lineages experiencing different series of rearranging recombinations. This has been especially the case when comparing closely-related strains (Naas *et al.*, 1994; Dalevi *et al.*, 2002; Rocap *et al.*, 2003). While ancestral genome structure may partially dissipate due to both horizontal transfer events (Doolittle, 1999a; Xie *et al.*, 2004) and a high level of intragenomic recombination (Wolf *et al.*, 2001), patterns of vertical ancestry are not completely removed. There are various cases of vertically inherited recombinations being stable despite presumed competition from ongoing emergent recombinants. For example, pulsed-field gel electrophoresis measurements have identified stable recombinant strains of *Campylobacter jejuni* coming from poultry processing batches (Wassenaar *et al.*, 1998). The stability in *C. jejuni* genome structure cannot be attributed to horizontal gene transfer because natural transformation between *Campylobacter jejuni* strains does not occur *in vivo* (Wassenaar *et al.*, 1998). Higher clade investigation in the *Enterobacteriaceae* family presents another case of tractable vertical evolution. Free and random intragenomic shuffling and horizontal replication of genomic structure between species does not occur (Sanderson, 1976; Souza & Eguiarte, 1997). There may also be gross structural conservation between members

from differing phyla or from each of the two domains, Archaea and Bacteria. Horimoto *et al.* (2001) finds a statistically significant portion of orthologs to be constrained in chromosomal position across a phylogenetic span represented by nineteen archaeal and bacterial genomes.

While a comprehensive, functional decomposition cannot yet be accomplished by analysis of a fully sequenced genome, genome sequences do help to delineate practical distinctions between prokaryotic organisms. Genomovars has been a term coined to characterize the capability of genome sequence to chart taxonomic boundaries independent of direct biochemical assessments of phenotype (Ursing *et al.*, 1995). A variety of studies have illustrated how genomovars work to help categorize diverse sets of strains from the genera *Pseudomonas* (Cladera *et al.*, 2004), *Burkholderia* (Vandamme, 2001), and *Sinorhizobium* (Young, 2003). In terms of chromosomal organization, functional accountings of non-random symmetries or periodicities have been proposed to involve past duplications of the chromosome (Kunisawa & Otsuka, 1988), the effects of supercoiling (Jeong *et al.*, 2004), and aspects of gene dosage (Jurka & Savageau, 1985). There is mounting empirical and comparative evidence as to how the physical organization of proteins as they are encoded by open reading frames on the physical length of a chromosome corresponds with both expression (Deng *et al.*, 2005; Higgins *et al.*, 1990) and function (Li *et al.*, 2005; Wolf *et al.*, 2001). This evidence suggests that an analysis of chromosomal ORF organization on the expanding data set of fully sequenced genomes may aid in a greater characterization of the functional and phylogenetic nature of prokaryotic genomes.

1.2 Genomic Variation and Phylogeny

A major goal in biology has been to determine the "universal tree of life" (Philipe & Forterre, 1999; Doolittle, 1999b; Kennedy & Norman, 2005) where various lineages of organisms generate progeny that either survive and reproduce, or do not. Survival and reproduction is influenced by competition with other organisms, ecological conditions, the inherited genetics of the organism, and chance events (Darwin, 1859; Kutschera & Niklas, 2004). For a time period of over several billion years (Schidlowski, 1988), organisms have been affected by many large-scale ecological changes (Nisbet & Sleep, 2001; Battistuzzi *et al.*,

2004), so it is difficult to re-enact *in vivo* the entire formation of life's history. Yet, the evolutionary history of known organisms can be inferred from comparisons of data to produce a phylogeny, a tree of life based on ancestral lineages (Harvey & Pagel, 1991). Comparisons between different ancestral lineages involves measurement of difference, and attempting to characterize when and why differences emerge. On a phylogenetic tree, there are large branches from which smaller branches emerge, eventually leading to known contemporary organisms which are placed at the leaves of the tree. The Woesian tree's largest branches represent three superkingdoms: *Archaea*, *Bacteria*, and *Eukarya* (Woese *et al.*, 1990). Two of these superkingdoms, *Archaea* and *Bacteria*, are prokaryotic; they contain unicellular organisms lacking organelles. The asexual form of prokaryotic reproduction is advantageous for evolutionary studies due to the non-reticulating pattern of vertical ancestry associated with asexual reproduction.

Although prokaryotes reproduce asexually, not all heritable characteristics follow a tree-shaped vertical ancestry. Through a variety of mechanisms, different strains can share genetic material with one another through a process called lateral, or horizontal, gene transfer (LGT or HGT) (Doolittle, 1999a; Battistuzzi *et al.*, 2004). Evidence for LGT challenges the idea of an immutably core set of monophyletic genes; LGT appears to have been a process that extends back billions of years (Rivera & Lake, 2004) with an effect ranging across most, if not all, functional categories of genes (Battistuzzi *et al.*, 2004). Currently, available data and the number of putatively conserved core genes is small enough so as to preclude estimation of the last common ancestor branchpoint (Battistuzzi *et al.*, 2004). Other phenomena that contravene or confound tests for vertical ancestry among the prokaryotes include paralogous origination of new genes (Patterson, 1988) and phenotypic switching (Balaban *et al.*, 2004).

To meaningfully predict behavior as a result of phylogenetic history, changes in the environment must be evaluated in addition to vertical (or horizontal) changes to the genome (Ridley, 1993; Lande, 1985, 1982). Calibrating an inferred history of mutational events against historical changes in the environment enables inferrence of an evolutionary clock's relationship to physical time (Ochman *et al.*, 1999; Battistuzzi *et al.*, 2004). Informationally, such a strategy of analysis has theoretical justification (Zuckerkandl & Pauling, 1965; Woese,

1987). When such calibration has occurred for sequence-level phylogenetic reconstructions however, there is significant variability of molecular clock rates between lineages. Contemporary efforts have sought resolution with either explicitly parametric models (Gillooly *et al.*, 2005), or semiparametric methods that help compensate for the complex "interplay between estimates of divergence times and rates" (Sanderson, 2002). To approach a molecular clock characterization of the dynamics between recombination events and phylogenetic branch points, there may be additional sources of complexity to the available information. Lineage-related diversity of recombinative mechanisms is extensive (Craig *et al.*, 2002). Furthermore, in the DNA sequence, historical evidence of DNA mobility gradually disappears through amelioration (Campbell, 2002). Visualizing and comparing different sets of recombinative events necessarily would involve a degree of inference for reconstructions of past history, especially as might be applicable to the testing of hypotheses involving historical changes in the environment.

While the theory underpinning a molecular clock is informational (Zuckerkandl & Pauling, 1965), hypothesis testing to characterize how lifestyle (evolutionary mode) causes variation in neutral changes (evolutionary tempo) requires identification of the "molecular counterpart of that ill-defined quality, evolutionary mode" (Woese, 1987). Recent observations of recombinative systems under experimental conditions have been consistent with recombinative behavior producing mutations under selective conditions that "cannot readily be produced by point mutations" (Schneider & Lenski, 2004). Treatment of recombinative changes in addition to sequence-level changes may therefore increase the amount of molecular data that can characterize evolutionary tempo and mode. With additional molecular data, wide-ranging credibility intervals for times of divergence in prokaryotic phylogeny (Battistuzzi *et al.*, 2004) may be to some degree shortened. An alternate possibility is that changes in lifestyle may be characterized more in terms of how environmental factors intersect with altered functional compositions produced by recombination (Konstantinidis & Tiedje, 2004).

Evolutionary mode and tempo have been characterized as being quasi-independent (Woese, 1987). Contrasting the effects of mode and tempo would require distinguishing these dynamics of change based on concepts corresponding to nature. Woese (1987) proposes three

distinguishing characteristics for evolutionary tempo: chronic ongoing change ("clocklike behavior"), action over a long period of time ("range"), and "loosely coupled domains" over which chronic changes are averaged (or, as called by Woese: "size"). Recombinations are being associated with an increased number of functional consequences (Schneider & Lenski, 2004). The increasing number of different functional consequences may elevate the possibility of there being "loosely coupled domains" as to how and where different recombinations occur throughout the genome. As would correspond to the chronic-like property of evolutionary tempo, recombinative activity has also been observed to be ongoing throughout both stressful and non-stressful conditions.

In general, gene order is poorly conserved in bacteria, even among closely related bacteria such as *Escherichia coli* and *Pseudomonas aeruginosa* (Nolling *et al.*, 2001). While gene order is more strongly conserved for other lineages such as the *Clostridia*, there is still prevalent disruption of low-level structures such as operons. Yet, although recombination can greatly disrupt genomic structure and, correspondingly, open reading frame (ORF) arrangement (Suerbaum *et al.*, 1998), comparisons of regions larger than operons show remarkably wide-ranging proximal similarity between orthologous pairings among genomes across phylogeny (Horimoto *et al.*, 2001). Such a finding suggests that strong forces of conservation prohibit dissipation of large scale ancestral ORF arrangement. If the pattern of ORF arrangement is retained over lengthy evolutionary ranges, and if changes occur in a chronic ongoing process that each independently influence disparate parts of the genome, then there is theoretical support for some of the variation in ORF structure to reflect evolutionary tempo. Testing of a proposed molecular clock can compare branch lengths so as to evaluate likelihood ratios (Shimodaira & Hasegawa, 1999). Furthermore, to independently compare the robustness of how recombinational history covaries with ribosomal phylogeny, different subtrees in the phylogeny can be identified and a nested analysis, such as that described by Bell (1989), performed.

The process of robustly measured ORF arrangement for clocklike properties may also facilitate baseline comparisons for how supercoiling arrangements contrast with optimal adaptation to variation in the environment. Water, virulence, salt and temperature have all been proposed as environmental factors associated with supercoiling (Higgins *et al.*, 1990;

Luttinger, 1995; Mojica *et al.*, 1994). The absence of a formula to precisely model the biochemistry between supercoiling and an external environment makes comparative tests of optimality implicit in that they rest upon preliminary expectations of maximized levels of "Darwinian fitness" (Harvey & Pagel, 1991). The construction of more explicit assessments could foreseeably involve dynamics of how the regulatory role of supercoiling controls DNA condensation and the transcriptional availability of genomic regions (Worcel & Burgi, 1972; Aki & Adhya, 1997; Reznikoff *et al.*, 1985). Based on current knowledge, there is difficulty with arriving at an explicit assessment. For example, there are varying estimates of nucleoid structure with supercoiling domains being conflictingly characterized as 10 kb per domain (Postow *et al.*, 2004) versus 50 kb - 100 kb per domain (Miller & Simons, 1993). Informational analyses may still help elucidate general evolutionary dynamics such as selection against deleterious mutants (Kimura, 1983). While initial characterization of various evolutionary consequences to recombinative change may be implicit for environment-based optima, implicit functional assessments are "a reasonable first step" (Harvey & Pagel, 1991).

While the fitness of genomic restructuring cannot yet be accounted for by an explicit formula, evolutionary comparisons can infer aspects of fitness and their phylogenetic range. A variety of "functional barriers" have been proposed to the fitness consequences of recombination (Mahan *et al.*, 1990). For example, a phenomenon of "replichore balancing" occurs where evolutionary fit recombinations act to keep the origin of replication at a position halfway (180°) from the termini. This phylogenetically widespread phenomenon can be inferred from various comparisons of closely related genomes belonging to different lineages (Dalevi *et al.*, 2002; Ren *et al.*, 2003; Andersson, 2000; Leblond & Decaris, 1998; Deng *et al.*, 2002). Replichore balancing has also been confirmed experimentally in both Gram negative (Hill & Gray, 1988) and Gram positive bacteria (Campo *et al.*, 2004). The presence of functional barriers such as replichore balancing implies possibilities where selection against definitively deleterious mutants would occur, consistent with neutral theory (Kimura, 1983). Yet, replichore balancing is not mandatory. The high frequency of IS-element recombinations in *Bordetella* spp. appears to overwhelm any selective pressure associated with replichore balancing (Preston *et al.*, 2004). Also, *Chlamydophila pneumoniae* strains J138 and CWL029 have 16 kb hot spots of rearrangements that are not near the chromosomal origin or terminus

(Shirai *et al.*, 2000). In controlled experiments, recombinational events have been observed to cause a wide range of variation without necessarily lethal effect (Mahan *et al.*, 1990). Other proposed "functional barriers" to recombination have included gene dosage effects, and conservation of structure around chromosomal termini (Mahan *et al.*, 1990). *Shigella flexneri* is thought to deviate significantly from *Escherichia coli* based on reoptimized placement of its transcriptional units in respect to the gene dosage gradient relative to *oriC* (Jin *et al.*, 2002). Further evaluation as to the strength of selection, measurement of fitness, and long-term competitiveness of recombinants may be helpful to characterize the evolutionary dynamics of recombinative events.

For purposes of inference, the amount of divergence associated with recombinative change between lineages must be considered. While "most sequence evolution is predominantly divergent" (Harvey & Pagel, 1991), several aspects of recombination confound a scenario of divergent heritable changes. Phenomena include reciprocal events occurring to balance the replichore (Deng *et al.*, 2002), balanced influx and loss of genome segments through horizontal transfer (Lawrence *et al.*, 2001; Parkhill *et al.*, 2001a), biphasic rearrangements (Barbour, 2002; Nanassy & Hughes, 2003), and duplication amplifications (Sonti & Roth, 1989; Read *et al.*, 2000). Promisingly, recombination does not appear to be convergent in scenarios where that might otherwise be expected (Schneider *et al.*, 2000). Recombinative divergence *per se* can be essential for driving rapid evolution of new traits (Sanderson & Liu, 1998). An overall divergent phenomenon of interest is where there is extensive gross-level conservation of genome structure compared to mosaic-like differences in smaller-scale structures. An instance of this phenomena can be observed with the 3 species of *Mycobacterium*: *M. leprae*, *M. tuberculosis*, and *M. bovis* (Philipp *et al.*, 1998).

The available genomes and their ORFs, having arised from various evolutionary lineages, may present challenges with causal and population inferences. Whether concerning sequence-level changes, non-vertical inheritance, or recombinative changes (Gillooly *et al.*, 2005; Zhaxybayeva *et al.*, 2004; Craig *et al.*, 2002), each lineage considered as a treatment is not a random allocation, so causal inference is not directly achievable (Lunneborg, 2000). Population inference requires random sampling (Lunneborg, 2000). The Gammaproteobacteria are likely to be over-represented as evident from larger compilations

structured from rRNA analyses (Garrity *et al.*, 2004). Additionally, beyond just the population of genomes, the ORF population is over-annotated and contains many false positives (Snyder & Gerstein, 2003). Beyond considerations of randomized treatments and randomized samples (or a rich, well-curated data set from which random resampling could be extensively performed), further difficulty with an analysis may come from the imperfectly resolved phylogeny (Kennedy & Norman, 2005) as well as the inavailability of explicit models to relate recombinative change to fitness and speciation. These aspects of observational noise (e.g., hypothetical ORFs), estimation error (e.g., over-representation of certain taxa), and dynamic noise (e.g., the consequences of a given recombination in an organismal population and the surrounding environment) are real-world complexities that make it difficult to characterize system dynamics (Casdagli *et al.*, 1991). A comparative method requires an evolutionary model (Harvey & Pagel, 1991), and a model would ideally have one data point per uniform taxon (Grafen & Ridley, 1997). Aspects of recombinative constancy to the genome is not something yet established for uniform taxonomic classifications.

It may be practically significant to address notions that explore populations of genes from a paradigm of behavioral ecology (Kurland, 2005; Dawkins, 1976). ORFs, mobile elements, and chromosomes have each been characterized as interdependent "populations" with aspects of competitive growth, fitness, and function (Lawrence & Roth, 1996; Schneider & Lenski, 2004; Terzaghi & O'Hara, 1990). Improved measures of associated patterns may better quantify both the observed population of ORFs and the consequences of different rearrangements. There have recently been advances that address the distribution of ORFs as informational units (Azad *et al.*, 2002), as well as advances in how informational signatures of interactions between populations can be detected (Sandvik *et al.*, 2004).

Azad *et al.* (2002) presents an investigation for ORF traits and their relationship to coding sequence versus non-coding sequence. By looking at dynamics of information, Azad *et al.* (2002) claim to "go beyond an analysis of the functional parts of the DNA." The informational analysis of Azad *et al.* (2002) proceeds with measuring information present inside various segments (successive regions of genomic DNA of a specific length in base pairs). Segmentation studies of genomic DNA serve to "break up a complex object into its 'constituent' parts...to understand how the organization comes about in the first place"

(Azad *et al.*, 2002). Azad *et al.* (2002) abstractly evaluate region lengths of potential coding space against a breakage process also known as the Kolmogorov theory of physical fragmentation. Essentially, the breakage of units at random points along their length leads to a log-normal distribution (Li, 1991; Azad *et al.*, 2002). Such a mode of ORF fragmentation could be attributable to nonsense codon mutations; in a study that separates actual genes from annotated genes, the length-based effects of randomly occurring start-stop codon pairs is utilized as a chief and phylogenetically widespread criterion to separate "real" from "non-real" ORFs (Skovgaard *et al.*, 2001).

The diversity and cryptic evidence of past recombinative histories (Campbell, 2002; Craig *et al.*, 2002) makes an exact parameteric model difficult to achieve since such noise must be evaluated to defensibly reconstruct changes in state (Casdagli *et al.*, 1991). Linear relationships cannot be fully assumed for how recombinative changes pass from ancestor to progeny. Ongoing debate and dialogue concerns, for example, the reticulating role of horizontal transfer and paralogous duplication (Kurland, 2000) and the implication of circular evolutionary pathways (Rivera & Lake, 2004). Strategies for direct manipulation of the interactions, parameterization of mechanical models, or direct simplifying assumptions that help characterize "linear relationships between response and predictor values" (Sandvik *et al.*, 2004) may all be significantly limited by current inferrence when applied to charting recombinational history. The field of ecology is producing new approaches that "do not make any *a priori* assumptions about dynamic properties" (Sandvik *et al.*, 2004). Sandvik *et al.* (2004) measures robust signatures of ecological interaction between multiple populations. Sandvik *et al.* (2004) demonstrate the usage of an approach for rigorously characterizing signals of interaction"that avoids these [mechanical models and linear relationships] difficulties."

Aside from bibliographic references to the literature item(s) characterizing particular submitted genome sequences, ecological and phenotypic information is generally absent from submitted genome sequence data files. A curatorial challenge has been to comparatively qualify the different lifestyles and ecologies associated with each genome-sequenced strain. Distinctions implicating varying schemes of genomic expansions, modifications, and contractions can involve the organism's intracellular or extracellular setting as well as

metabolic activity (Bentley & Parkhill, 2004; Ochman & Davalos, 2006). A limiting aspect to the analysis that may bias the set of 155 genomes, is that only 1% of all estimated microbes can be cultivated in artificial laboratory conditions, and the biochemical and metabolic properties of culturable organisms become, by default, "key characteristics" (Santos & Ochman, 2004). Conventional morphological and nutritional criteria used to describe microbes do not lead to a natural taxonomy (Pace, 1997).

A full characterization of phenotypic diversity across the phylogenetic range represented by fully sequenced prokaryotic genomes is a significant enterprise involving hereditary information in addition to the behavior and environments inhabited by prokaryotic strains. The current phylogenetic estimates are quite variable. As calculated from nucleotide sequence changes, the divergence of *Yersinia* from *E. coli* is estimated to be 375 Ma ±145 Ma (Deng *et al.*, 2002). Even the comparably richer historical record concerning *Y. pestis* and *Y. pseudotuberculosis* leads to an estimated time of divergence 1,500-20,000 years ago (Achtman *et al.*, 1999). For time spans involving billions of years, the range of variation for credibility intervals is approximately ± 10-20% (Battistuzzi *et al.*, 2004).

At minimum, for most of the fully sequenced prokaryotes, the genomic DNA is present in the form of at least one distinct chromosome. Variation between species can extend to multiple copies of the same chromosome, multiple different chromosomes, and other replicons such as plasmids. A functional definition of a plasmid is that it is unnecessary for the viability of a particular organism (Bentley & Parkhill, 2004). Yet, such a distinction may not be perfect for current classifications of replicons. Larger plasmids may have especially high maintenance costs and there would need to be some offsetting selective advantage to promote their presence within a prokaryotic organism. *Halobacterium* has a 200 kb plasmid, *pNRC100*, that has "properties of resistance to curing suggest that this replicon may be evolving into a new chromosome" (Ng *et al.*, 1998). Other large plasmids associated with fully sequenced genomes include a 2 million base pair plasmid in *Ralstonia solanacearum*, and a 1.6 million base pair plasmid in *Sinorhizobium meliloti*. Conversely, there are various chromosomes that, based on size and horizontal ancestry, might otherwise be considered plasmids except for having some degree of "essentiality" to the life of the organism. *Vibrio cholerae* has a chromosome that appears to be a captured megaplasmid from a non-Proteobacterial origin

(Heidelberg *et al.*, 2000). It has also been suggested that unknown, novel chromosomal structures may yet be identified. For instance, the conventional PFGE approach misses what new methods, such as optical mapping, can find (Lin *et al.*, 1999; Zhou & Schwartz, 2004).

The number of distinct chromosomes is not necessarily fixed between closely related strains. Different biovars in *Brucella suis* can have either a single 3.3 Mb chromosome, or 2 chromosomes of smaller sizes (Jumas-Bilak *et al.*, 1998; Paulsen *et al.*, 2002). In the sense of a cellular stoichiometry, there can be multiple copies of the same chromosome per cell. Stoichiometric measurements have some relationship to growth, but do not follow an exact formula across all taxa. *Methanocaldococcus jannaschii* has an incremental L-shaped distribution from 1 to 5 chromosome equivalents for stationary growth, and an L-shaped curve ranging from 1 to 15 chromosome equivalents for exponential growth (Malandrin *et al.*, 1999). This is in contrast to the "multiple of 2" distribution of chromosome copy numbers in *Escherichia coli* where the copy numbers of chromosome equivalents ascend in the sequence: 1,2,4,8 (Malandrin *et al.*, 1999). At what may be an upper extreme, *Buchnera* can have 100 genomic copies per cell (Shigenobu *et al.*, 2000).

Association with metabolism and lifestyle is sometimes explicable from recombinative dynamics and conservation. For example, a recombination deletion event can be inferred when observing that *Buchnera aphidicola* has many *fli* and *flg* orthologs to *Escherichia coli*, yet it is missing a *fliC* gene (Tamas *et al.*, 2002). This is evidence for non-motile behavior and corresponds to how the endosymbiotic lifestyle of *B. aphidicola* contrasts with the free-living *Escherichia coli* (Tamas *et al.*, 2002). Intracellular bacteria such as *B. aphidicola* generally represent strains with stable genomes where deletions of repeated sequences are irreversible and mobility has been reduced (Andersson & Kurland, 1998). By contrast, non-intracellular pathogens and commensals that face greater competition and more fluctuation of available resources in their host environments rely on genomic rearrangement to facilitate frequent and revertible phenotypic changes (Hallet, 2001).

In whatever degree of detail the data set is evaluated, there remain additional obstacles to inferring exact molecular changes over a lengthy periods of time. Both the amelioration of DNA composition (Campbell, 2002) and the highly composite, dynamic interaction between interleaving IS elements (Campbell, 2002; Gray, 2000) introduce substantial complexity as to

how the history of the internal genomic structure may be retrospectively untangled. The challenge for a comparative analysis is to identify, measure, and account for the variance of common properties across the phylogenetic range being evaluated. There is not yet however a broadly prescribed method for inferring a historical series of recombinative events so as to evaluate diverse hypotheses about how recombinative changes impact fitness. Explanations as to how strategies of recombinative expansions and modifications relate to ecological adaptation are presently anecdotal (Bentley & Parkhill, 2004). It is difficult to envisage an explicitly parametric model that can directly evaluate how recombinative change relates to the correspondence of phenotypic diversity with genomic structure. The fact that an informational approach does not rely on assuming the constraints of one particular model versus another may be advantageous, especially given the uncertainty as to how recombinative changes in genomic structure relate to changes in fitness for an organism and its lineage.

1.3 Genomic Mobility

Within sets of closely related strains, change in chromosome size is largely due to recombination events. Evolutionary experiments by Bergthorsson & Ochman (1999) show that such changes in the size of chromosomes occur more often than base pair mutations altering restriction sites. Recombination also alters the internal structure of a replicon such as a chromosome (Andersson, 2000). Chromosomal variation is often measured in terms of length differences between ribosomal sequences as assayed by restrictive digests (Ge & Taylor, 1998; Ralyea *et al.*, 1998). This variation is called "ribotype diversity", and is generally attributed to recombinations between rrn operons. In strains of *Salmonella typhi*, ribotype diversity is much greater than corresponding base pair diversity (Ng *et al.*, 1999).

Recombinations can cause deletions, duplications, inversions, and translocations (Andersson, 2000), and recombination frequently involves double strand separation of the double helix to reveal single strands. These strands can either interact with macromolecules to facilitate recombinative mechanisms or, by homology, complementarily bind directly to a single strand of DNA at another site on the double-stranded DNA molecule (Brown, 2002).

Recombinations can sometimes be attributed to mechanisms of replication slippage at the replication fork and duplication events (Liò *et al.*, 1996; Tillier & Collins, 2000). Mechanisms of recombination can be categorized as follows: site-specific, homologous, and illegitimate (Brown, 2002; Ikeda *et al.*, 1982; Bachellier *et al.*, 1996; Nair *et al.*, 2004). The frequency of a recombination event can be dependent on the mobilized length of DNA (Bi & Liu, 1994), and there are also "hot spots" of recombinative activity as well as highly conserved regions (Watanabe *et al.*, 1997).

Recombination frequency is significantly dependent on the mechanism. *RecA*-independent recombination between large repeats (> 100 base pairs) happens at a rate of 10^{-5} to 10^{-4} recombinations per large repeat per generation; when occurring due to a slippage mechanism, this requires that repeats be less than 10 kb apart (Lovett, 2004). *RecA*-dependent tandem duplications between IS elements occurs at a frequency from 10^{-4} to 10^{-2} per IS element per generation (Haack & Roth, 1995). Per hour, this rate has been observed experimentally per IS element as being $2 * 10^{-6}$ to $9 * 10^{-6}$ per cell per hour. (Schneider & Lenski, 2004). Estimates of sequence-level mutational rates range from 10^{-8} (Lovett, 2004) to 10^{-11} (Ochman *et al.*, 1999) changes per genomic base pair per generation. With an estimated 100-300 successful generations per year, Ochman *et al.* (1999) calculate there to be 0.0045 mutations per genome base pair per million years. For a 3 million base pair genome, this corresponds to 1,350 mutations per genome per million years. Contrastingly, without negative selection or reversible changes, tandem duplications attributable to IS element-based changes would be expected to introduce a staggering number of about 200,000 changes per genome per million years.

DNA mobility may relate to evolutionary dynamics in a number of ways. Mobile elements may either be conserved in a mutualistic sense to promote heterogeneous offspring or, alternatively, persist based on their own "selfish" parasite-like behavior (Schneider & Lenski, 2004). The frequency of DNA mobility may impact general diversity of a species-like taxa. *Staphylococcus aureus* has a recombination rate 3 times lower than mutation compared to *Neisseria meningitidis* which has a recombination rate 3.6 times more frequent than mutation (Cohan, 2004). *Staphylococcus aureus* may be thus expected to exhibit greater population clonality in comparison to *Neisseria meningitidis*, where clonality is the stable

transmission of multiple sets of alleles (Wisplinghoff *et al.*, 2003). Intriguingly, *Neisseria meningitidis* can still be very clonal in nature due to a few highly successful strains (Souza & Eguiarte, 1997). In this sense, externally-influenced dynamics of selection can act to filter the retrospectively calculated stochastic dynamics of occurrence.

Recombinant changes between generations may be evolutionarily unstable. Stable, vertically divergent recombinations may be a different type of evolutionary dynamic than genomic plasticity, a variation-producing feature of frequently generated, unstable changes. Genomic plasticity can involve reciprocating changes that occur in response to alternating environmental conditions. One example of genomic plasticity involves the amplifying expression of the *his* operon in the *Salmonella* genome. *RecA* dependent tandem duplications of this operon occur at a frequency of 0.01 to 1 percent of progeny and can be preserved under selected conditions (Haack & Roth, 1995). The rate of deletion that removes these duplicated operons is 1 to 30 percent of progeny. Tandem duplications are often deleted since their duplication produces direct repeats that can subsequently undergo a D-shaped recombination event (Romero & Palacios, 1997). Another example of genomic plasticity involves a site-specific inversion system in *Salmonella* (Nanassy & Hughes, 2003). A *hin* recombinase mediates inversion of 1,000 bp in order to biphasically vary an antigen protein so as to "outsmart" the immune system. None of these examples, however, suggest a basis for the type of long-term trajectory of divergent, conserved change that could correspond to the recombinative dynamics proposed by Lathe *et al.* (2000) or Horimoto *et al.* (2001).

One way to estimate the influence of stable recombinations, is to assess the rules that may apply to how recombinations proceed in nature. There are a variety of parsimonious criteria that, if applicable, can act to compile and summarize the most likely phylogenetic tree (Harvey & Pagel, 1991). Dollo's law "states that complex characters will not have evolved more than once" (Harvey & Pagel, 1991). Yet, since recombinations are frequently produced by specific recombinations involving IS elements on the genome, it is possible that evolution may be somewhat parallel. In the case of 18 replicate populations that were each separately propagated for 1,000 generations, patterns of both parallel and divergent evolution were observed for conditions related to 2,4 dichlorophenoxyacetic acid as a sole carbon source (Nakatsu *et al.*, 1998). Multiple composite recombinations can lead to a wide range of

combinations (Gray, 2000) so, over time, it is plausible that many steps of recombination would be divergent enough to produce distinct signatures for various lineages. Some additional, alternative parsimonious criteria to consider are: "the smallest number of character trait transitions," and "derived characters being lost on fewest occasions" (Harvey & Pagel, 1991). Yet, recombinations can readily violate some of the above assumptions governing vertical ancestry (Patterson, 1988; Snel *et al.*, 2002), so it is difficult to know if there are consistent levels at which rules of vertical ancestry can be considered reliable versus relaxed.

An alternative to parsimonious reconstruction is to approach efforts at phylogenetic reconstruction as a statistical problem. In a parametric fashion however, degrees of freedom may be difficult to characterize in more sophisticated statistical models related to DNA mobility. As mobile DNA and other changes act to both expand and otherwise alter a genome, it is comparable to the, albeit simpler, expansion-modification systems proposed by Li (1991). These systems are a type of "probabilistic context-free Lindenmayer systems" that, as open dynamical systems, have changing degrees of freedom. The fact that these changes occur on a nested hierarchical phylogeny also leads to variable precision as to how degrees of freedom might be characterized (Harvey & Pagel, 1991). Species belonging to the same genus generally have fewer degrees of freedom than species coming from different genera (Harvey & Pagel, 1991). In a biological sense, a hierarchy of recombinational differences may be variable in how they constitute adaptive differences, and such a distinction may be difficult to model (Harvey & Pagel, 1991).

There is also natural variation in how DNA mobility does not fully reflect an intragenomic dynamic proceeding along a vertical hierarchy. The estimated fraction of a genome that has been laterally transferred from other species is 5-10% (Cohan, 2004). Lateral transfer does not always readily occur between species though, and bacterial "sexuality" can be limited to closely related strains within a species such as for *Sinorhizobium meliloti* or occur with significantly fewer constraints of close relationship such as for *Neisseria gonorrhoeae* (Souza & Eguiarte, 1997). Overall, the non-vertical dynamic of intraspecies genomic exchange can be quite frequent. Lawrence (2002) estimate that less than 10 LGT events successfully occur per million years with *Escherichia coli.* Zhaxybayeva

et al. (2004) estimate that "several hundred [genes] every four million years" are transferred among some sets of closely related strains.

A further complication for modelling recombinative change involves the dynamic of illegitimate recombination. At the lower end of recombination frequencies (10^{-12} to 10^{-15} per genome base pair per generation), illegitimate recombinations were first proposed to involve 12 base pairs or less in the asymmetric pairing of complementary sequence (Franklin, 1971). As is the case with bacteriophage λ, these can be site-specific and require extra factors and enzymes like the integration host factor (IHF) and viral integrase (*int*) in order to facilitate the illegitimate recombination (Franklin, 1971). These can also, rather than requiring extra factors, be facilitated directly by hairpin structures (palindromic repeats) surrounded by direct repeats. Hairpin structures like these have been seen in a recombining 96bp *Borrelia* segment that generates genomic diversity in such a way as "to avoid host immune elimination" (Wang *et al.*, 1997). A more updated definition of illegitimate recombination is that it "involves junctions of nonhomologous or very short homologous DNA sequences (often less than 3 bp) which are not recognized by site-specific enzymes" (Nair *et al.*, 2004).

Despite their regulatory importance, operon structures are not conserved and are widely disrupted across various lineages by both intragenomic and intergenomic dynamics (Watanabe *et al.*, 1997; Nolling *et al.*, 2001). Yet, in the form of positive selection, operons can be selected targets of duplication such as can be seen with the multiple copies of ammonia monooxygenase (amo) operons in ammonia-oxidizing autotrophic bacteria (Klotz & Norton, 1998). Such a duplicated operon corresponds to an analysis from Snel *et al.* (2002) suggesting that gene addition is under positive selection. Despite disruption at a localized operon level, there appear to be larger "uber-operonic" aspects to conserved ORF location (Horimoto *et al.*, 2001; Lathe *et al.*, 2000). The fact that laterally transferred, functionally related genes do not reassociate with a corresponding uber-operonic functional complex suggests some limitation as to the frequency or fitness characteristics associated with localized rearrangement events (Lathe *et al.*, 2000).

From the standpoint of altered expression and host immune evasion, DNA rearrangement has been equated to the network motif of a noise amplifier–contributing to population heterogeneity and antigenic variation (Wolf & Arkin, 2003). This noise is

proposed as a way to spread risk over multiple phenotypes and, in abstract engineering terms, may also enhance signal by "stochastic resonance" (Wolf & Arkin, 2003) where possible negative side-effects of an otherwise successful change are balanced out. The spreading of risk may correspond to a lottery model described by (Smith, 1975). In this case of augmented population heterogeneity, those strains with a greater chance of introducing diverse progeny are more likely to hit a metaphorical "jackpot." Another related scenario is the "arms race." This scenario involves those species that can react more quickly to the environment by adaptively changing first with respect to fitness, thereby succeeding over those who are diversifying without direct relationship to fitness (Williams, 1971).

Recombinations associated with speciation do not necessarily relate solely to considerations of stochastic frequency and external conditions. The evolutionarily stable changes may possibly be those that best conserve characteristics of expression or regulation associated with the large scale topology of the entire supercoiled prokaryotic genome (Deng *et al.*, 2005). In addition to specific hot spots on a chromosome influencing the incidence and impact of recombinations such as *oriC*, there may also be other aspects governing the overall genomic distribution on a replicon's topology. A more sophisticated molecular model may be proposed that characterizes how the superstructure to the genome may influence regulation based on topology. The location of functional promoter domains near HU-mediated supercoiling (Tanaka *et al.*, 1993) sterically hinders expression (Kohno *et al.*, 1994). Yet, if ORFs are positioned far away from HU-sites, the degree of expression, looking at 14 different sigma factors, is independent of which supercoiled loop a regulated open reading frame is present upon (Reznikoff *et al.*, 1985). This independence of location is confirmed in a broader survey of other prokaryotes (Wolffe & Drew, 1995). Regulatory dynamics occur between distant chromosomal regions. For example, xylene/toluene metabolism can have four different operon/transcriptional control regions with interactive regulation (Ramos *et al.*, 1997). If recombination repositioned an open reading frame near an HU-site, this could have an impact that may cascade across large functional networks such as described by Ramos *et al.* (1997) and Li *et al.* (2005).

Mechanistically, *RecA* and *HU* are some of the many macromolecules that bind to DNA that may potentially effect genomic structure and subsequent expression. Macromolecular

binding is sequence-dependent, frequently involving DNA recognition of a specific sequence by proteins with the helix-turn-helix motif (Harrison & Aggarwal, 1990). In the case of DNase I, these sequences have been found to be about 8 nucleotides, corresponding to groove width and stiffness associated with the helically wound double-stranded DNA (Lahm & Suck, 1991). Another mechanism involves illegitimate recombinations that are facilitated by DNA gyrase (Ikeda *et al.*, 1982). If gyrase-stimulated recombinations correspond to producing functionally competitive progeny, the archaealogy of genome structure would show how locations of gyrase activity correspond to optimal characteristics of genome organization. Indeed, DNA gyrase activity correlates positionally with restraints on spatial patterns of transcriptional activity (Jeong *et al.*, 2004). It is conceivable that there is a framework of recombinational mechanisms and consequences in fitness that may be corroborated by measures of optimal genome arrangement. It is unknown, however, as to how precisely an analysis of recombinational mechanisms and fitness dynamics will map to the many different possibilities for such a framework. It is also unknown as to how complex the framework would have to be to account for a wide view of both *Archaea* and *Bacteria*.

ORF arrangement and clustering may exhibit some invariance based on patterns of content, size, and distances of ORFs as they occur between diiffering chromosomal regions. While genome structure may change to some extent, various assays provide a basis for relating measures of ORF clustering to evolutionary range. Sequence similarity among ORFs is abundant; "50% of prokaryotic genes emerge from duplication" (Li *et al.*, 2005) where duplicate sequence pattern has been produced from past gene duplications and conserved amongst various domain rearrangements. There is also evidence that the evolutionary heritage of a DNA segment containing multiple ORFs relates to the evolutionary heritage of encoded ORFs. In *Thermoplasma acidophilum*, 32% (484) of the ORFs are found in 139 conserved gene clusters (Ruepp *et al.*, 2000). Cluster-related conservation is ascertained by comparison with 13 other prokaryotic genomes where pairs of potentially orthologous ORF sets were separated by at most three other ORFs (Ruepp *et al.*, 2000). In another approach of conserved orthologous proximity, Horimoto *et al.* (2001) find that, while ORFs may wind up on separate locations between two circular replicons from two different species, the ORFs significantly trend to remaining within a 20° (e.g., 600 kb on a 3 Mb chromosome) region on

a chromosomal circle relative in position to to other ORFs. Horimoto *et al.* (2001) note that regional constraints of an ORF are influenced by the functional role of the ORF as evident from functional categories for COG. These inferred regional constraints suggest some interdependence between content of a chromosomal segment and dynamics of alteration to ORF clustering. From the standpoint of function, the *Escherichia coli* K-12 genome may possibly include a 600 kb "supercluster" periodicity that appears to associate with coordinated gene expression (Allen *et al.*, 2003). (Kunisawa & Otsuka, 1988) claim to have found a "7 minute periodicity" (i.e., 350 kb) on the *E. coli* K-12 genome to the clustering arrangement of ORFs. A more recent evaluation characterizes *E. coli* K-12's large-scale periodicity as being "weak" and, in summary, Koonin *et al.* (1996) offer two explanations for large-scale periodic arrangement of ORFs: 1) duplication of large segments of the chromosome early in evolution; and 2) "the periodicity relates to nucleoid superstructure." Yet, a well-parameterized model that makes a defensible account of causative dynamics for large-scale periodicity has not yet been proposed.

1.4 Annotated Open Reading Frames

"Many genomes are over-annotated" in the sense that real genes are not discriminated from random ORFs (Larsen & Krogh, 2003). There exist false positives in the form of annotated ORFs that are not transcribed into functional units such as enzymes (Frishman *et al.*, 1998). A variety of studies have either indicated or predicted that the fraction of annotated ORFs with low, "unreal", or non-functional importance to the organism is $\approx 25\%$ of the total set of annotated ORFs for a given genome (Williamson *et al.*, 1993; Jackson *et al.*, 2002; Skovgaard *et al.*, 2001; Tatusov *et al.*, 2003). An exact, prescribed characterization of every ORF has not yet been achieved (Roberts *et al.*, 2004) and "the boundary between living and dead genes is often not sharp" (Snyder & Gerstein, 2003). This may in part be due to a complex diversity of characteristics and categorizations that may be used to consider each ORF. One set of groupings for ORFs (originally proposed for yeast) has been proposed as: "eORF (essential ORF), kORF (known ORF with a well-characterized function), hORF (ORF validated by homology only), shORF (short ORF), tORF

(transposon identified ORF), qORF (questionable ORF), and dORF (disabled ORF or pseudogene)" (Snyder & Gerstein, 2003).

Analytical criteria that help weigh "the likelihood that a gene encodes a functional product" are: sequence features, evidence for transcription, sequence conservation, patterns of gene inactivation, and functional genomics information (Snyder & Gerstein, 2003). Sequence conservation analyses work to compare an individual DNA sequence from one organism to the sequences of other known sequences, and is "an excellent method to gauge the importance of the gene product" (Snyder & Gerstein, 2003). Sequence features can involve detailed measurement of mutational effects such as codon bias, since there are dynamics underyling the nonrandom use of codons compared to non-coding regions and distinguishing associations between genes involving aspects of expression (Duret & Mouchiroud, 1999), gene length (Eyre-Walker, 1996), and horizontal gene transfer (Garcia-Valivé *et al.*, 2000).

A sequence conservation approach is, however, strongly influenced by the phylogenetic proximities of relationship between the associated organisms (Snyder & Gerstein, 2003). Strains that are phylogenetically close have had, over time, less opportunity for phenotypic deviation due to a recent shared ancestry (Harvey & Pagel, 1991). Strains that are phylogenetically far apart may have conserved sequences due to LGT, or strong evolutionary forces of conservation. In order to utilize sequence conservation as a criterion for separating "real" ORFs from ORFs of little functional or evolutionary importance, there must be some account for phyletic pattern (Glazko & Mushegian, 2004). A monophyletic distribution of similar ORFs saturates a phylogenetic subtree where a last common ancestor can be inferred as having vertically transferred specific ORFs to its descendants. Other phyletic distributions include polyphyletic (occurring among various disparate lineages in a way to suggest non-vertical evolution) and paraphyletic (a subtree with a sub-subtree removed) distributions. As modelled by Snel *et al.* (2002), LGT may account for polyphyletic distributions of ORFs among prokaryotic organisms, and gene loss may account for paraphyletic distributions.

Sequence conservation is often used as a basis for making functional annotations to ORFs whose activity and function have not been directly assayed. Yet, functional genomics information, as recorded in ORF annotations, is significantly incomplete: "all prokaryotic genomes sequenced to date have a fairly high fraction (between 20 and 40%) of genes for

which no function has been assigned" (Van Sluys *et al.*, 2002). Furthermore, 5-10% of functional annotations are wrong (Roberts *et al.*, 2004). The present situation with prokaryotic genomes is that curatorial efforts for improved annotations of ORFs have been "sluggish", and the blurry boundary between living and dead genes may be partly a function of insufficient curatorial effort as well as a lack of more exacting assays of the transcriptome and proteome (Roberts *et al.*, 2004).

Evidence for transcription involves measurement of RNA or protein expression that comes from a given DNA sequence. From the vantage point of transcriptional evidence, a "conceptually straightforward" approach may be to utilize a whole-genome DNA microarray designed to study a fully sequenced microbe (Cummings & Relman, 2000). In *Escherichia coli*, the number of annotated genes to express above background levels is 3,496 (81%) out of a total possible 4,290 ORFs (Tao *et al.*, 1999). Assessments of DNA expression can be unreliable however due to the frequency at which a probe for a falsely annotated gene may associate with an untranslated region of an expressed gene (Skovgaard *et al.*, 2001).

Gene inactivation assays involve measuring the effect of how artificially-induced mutations have a phenotypic consequence due to an inactivated, though still expressed, gene or set of genes. There are currently limits to the availability of data. For *Bacillus subtilis* (Biaudet *et al.*, 1997), only 13% of annotated ORFs have been assessed for patterns of phenotypic inactivation. In general, experimental assessments of annotated ORF operation and function have not been comprehensively performed for the larger set of publicly available, fully sequenced genomes.

Over-annotated false positives (random, "unreal" ORFS) occur predominantly for ORFs that trend toward shortness in length (Larsen & Krogh, 2003; Skovgaard *et al.*, 2001). Such a trend may occur by truncating nonsense mutations (Skovgaard *et al.*, 2001), although there may also be physiological differences to ORF lengths that are accounted for by the multidomain structures of the encoded proteins (Liang *et al.*, 2002). A direct structural classification of evolutionarily divergent proteins and their internal modules is not easily performed. Within the Structural Classification of Proteins database (SCOP) (Murzin *et al.*, 1995), folds (structural similarities) from divergent sequences of common origin lead to superfamily predictions that are only 29% accurate (Lindahl & Elofsson, 2000). Assessment

of sequence similarity on conserved domains, with divergent sequence, are on the level of 75% accuracy (Lindahl & Elofsson, 2000). There are other approaches, such as BLASTCLUST, that address the issue of common evolutionary origins with a variety of default choices for percent identical residues, comparison of length, and BLAST score density which is the proportional amount of length covered by a high scoring segment pair (Altschul *et al.*, 1990). Additional refinements to a sequence conservation analyses can filter out common motifs, such as coiled coil regions, which by themselves do not add much evolutionary signal (Tatusov *et al.*, 1997b). Any comprehensive handling of structural protein features and data involves some "curatorial pain" (Chung & Yona, 2004), and more automated refinements, such as practical adjustment of the expectation score in terms of repetitive low complexity protein structure–especially for smaller proteins (Birkland *et al.*, 2005)–are still not fully usable. Whatever the profile (domain structure) diversity of an ORF, it is generally recommended to evaluate as many sequence homologs as possible to assert meaningful ancestral membership within a protein family (Sadreyev & Grishin, 2004). For example, the detection of remote homologies is three times more likely when more than 2 sequences are used to assess for homology (Park *et al.*, 1998), and there are sequences with less than 30% pairwise identities to other sequences that, when analyzed in groups of several or more, significantly cluster together as homologs. Overall, for purposes of asserting some vertical origin, e-value cutoffs appear to range from 10^{-2} (Altschul & Koonin) to 10^{-8} (Pagni & Jongeneel, 2001; Sadreyev, 2003). Even for strict expectation score cutoffs like 10^{-14}, false positives have still been observed (Sadreyev, 2003). For the purposes of evaluating a sampling of ORFs, there is a way to estimate the number of false positive hits based on a given expectation score cutoff. Expectation scores less than 0.01 are equivalent to the expected percentage of random (false positive) hits within a population of sequences (Koonin & Galperin, 2003). In this regard, surveying 10,000 sequences for a match to a sequence based on an expectation score threshold of 10^{-3} would amount to approximately 10 random hits.

Evolutionary dynamics other than stop codon truncations can also be inferred from ORF length characteristics. For example, Teichmann *et al.* (1998) report, beyond the approximate quarter of *Mycoplasma genitalium* ORFs that contain just one conserved domain, that the "large majority of proteins in the MG genome have involved rearrangement

of domains." This qualification is based on a characteristic distribution of ORFs with distinct composite domains. Wheelan *et al.* (2000), however, find that, whatever the underlying dynamics of gene rearrangements are, the domain size distributions lead to discontinuous frequencies of various ORF lengths. Savageau (1986) makes a case for proteins in *Escherichia coli* generally occurring in structural subunits of 14 kDa which is about 127 amino acids (aa). While *E. coli* protein modules (single domains) have an average length of 219 aa, the normative "bulk" of evaluated modules range in length from 100 to 150 aa (Liang *et al.*, 2002). In an informational sense then, based on relationships between distributions, the impact of recombinative processes of change can be sometimes revealed. Future resolution may involve case-by-case assessments of proteomic structure and function. This is however dependent upon a mixture of curatorial effort and biochemical detail that may be difficult to uniformly apply to each fully sequenced genome.

For operons, a predictive genome-wide algorithm and database was recently established for *Staphylococcus aureus* Mu50 (Wang *et al.*, 2004) which represents a significant innovation beyond databases that have been limited to evaluating *Escherichia coli* K-12 (Huerta *et al.*, 1997). Predictive algorithms are important; even in the well-studied *E. coli* K-12 genome, the RegulonDB database shows that just 869 operons are known compared to 2325 operons that are predicted (Huerta *et al.*, 1997). Operons vary in the number of ORFs that they transcriptionally co-express. In *E. coli* K-12, up to 70% of the transcriptional units are "monocistronic," having just one ORF (Blattner *et al.*, 1997). *S. aureus* is calculated to have 62% of its transcriptional units as monocistronic with an average operon size of 3.47. About 90% of operons have 5 or less ORFs, and only a marginal amount have any more than 10 ORFs (Wang *et al.*, 2004; Huerta *et al.*, 1997). The largest predicted operon in *S. aureus* Mu50 contains 29 ORFs and encodes ribosomal proteins. The two largest predicted operons in *E. coli* K-12 contain 11 ORFs each, and encode phenylacetic acid degradation and sugar transport functions (Huerta *et al.*, 1997). Algorithms for operon (or transcriptional unit) detection have been extended to analyze a variety of other Bacteria and Archaea (Stormo & Tan, 2002; Liu *et al.*, 2003), yet there does not yet appear to be a well-curated database with predicted operon structures on all of the fully sequenced genomes. Other algorithmic efforts are being developed to better quantify the accuracy of operon predictions compared to

evidence from sequence and expressional data (Bockhorst *et al.*, 2003).

It is possible to arrive at some correspondence between an organism's set of ORFs versus metabolic capabilities necessary for the organism's lifestyle. Tamas *et al.* (2002) identify *B. aphidicola* APS as requiring a set of ORFs active in sulphur assimilation since it is endosymbiotic to an aphid that, eating legumes, does not ingest as much cysteine as the grass-eating aphid host of *B. aphidicola* Sg. Evidence suggests that sulphur assimilation genes became inactive in response to cysteine-rich conditions of *B. aphidicola* Sg (Tamas *et al.*, 2002).

While there are existing systematic catalogues of taxa, phenotypes, and some corresponding metabolic and physiologic characteristics (Garrity, 2001), there is not yet an up-to-date synthesis that equates the ORF complement to the phenotype. Analysis of clusters of orthologous groups (COGs) has been one effort in this direction where functional categories such as "RNA processing and modification," "extracellular structures," and "cell motility" are identified (Tatusov *et al.*, 1997b). The link, however, between unique ORFs and speciation (Konstantinidis & Tiedje, 2004), as well as the restriction of important orthologous sets to taxonomic boundaries (Kurland, 2000), suggests that ORF similarity alone cannot fully map the metabolism and physiology.

While functional assessments of ORFs partly rely on anecdotal approaches, characterization of ORFs may be a meaningful step in the accelerating rise of available sequence, ecological, and evolutionary information. Schilling *et al.* (1999) describe a cascading succession of various knowledge domains that are rising up to characterize the genome, transcriptome, proteome, metabolome, and beyond. This succession may be currently evident from the increasing number of tools available to access and characterize the content and metadata surrounding the growing numbers of strains, chromosomes, and genomic structures such as ORFs (Murzin *et al.*, 1995; Koonin & Galperin, 2003; Chung & Yona, 2004; Kent *et al.*, 2005; Wang *et al.*, 2004).

1.5 Summary and Objectives

Aspects of genomic stability have been found to relate to the ecological lifestyle of a prokaryotic organism (Ochman & Davalos, 2006), and various underlying factors of chromosomal topology and expression suggest that the organization of ORFs may have a functional role in the physiology of the organism (Deng *et al.*, 2005; Képés, 2004; Kunisawa & Otsuka, 1988; Svetic *et al.*, 2004; Lathe *et al.*, 2000). Prokaryotic diversity relates to overall genome content, and recombinative expansions and modifications can allow for a faster tempo of change than possibilities attributable to single point mutations (Bentley & Parkhill, 2004). Mechanisms, such as those involving mobile elements, are providing some ability to account for structural changes in genome organization and the density of insertion sequence (IS) elements on the genome that can be an indicator of lifestyle (Moran & Plague, 2004). Many of these studies have drawn their observations and results from the recent increase of publicly available, fully sequenced genomes. There remain, however, a variety of past hypothesis-driven approaches to genomic organization that have not yet been carried forward to the present set of fully sequenced genomes. In particular, there is a set of studies that have sought to account for whether gene density is non-random on the *Escherichia coli* chromosome (Bachmann *et al.*, 1976; Jurka & Savageau, 1985; Kunisawa & Otsuka, 1988; Williamson *et al.*, 1993). In the past, based on the predicted locations of protein-coding sequence, gene density has been evaluated as the number of ORFs per equal-sized segments of a replicon (Jurka & Savageau, 1985). Yet, some ORFs may be more important or "real" than other ORFs (Snyder & Gerstein, 2003; Larsen & Krogh, 2003). Conserved orthology has been an initial approach to characterizing functional roles of ORFs (Bentley & Parkhill, 2004; Tatusov *et al.*, 1997b), and the genomic context can be predictive of gene function (Wolf *et al.*, 2001).

While operon structures and genomic landmarks such as *oriC* may play a role in the functional expression of an ORF (Jin *et al.*, 2002; Wolf *et al.*, 2001), they are not the only factors underlying the conserved positioning of conserved ORFs. The relative locations of multiple sets of orthologs show evidence of conservation across the entire stretch of a genome despite extensive, localized rearrangement and fluid-like alteration of operons (Horimoto *et al.*, 2001; Lathe *et al.*, 2000; Wolf *et al.*, 2001). The functional consequences of

recombinative change have been a topic of substantial interest and modelling (Terzaghi & O'Hara, 1990; Wolf & Arkin, 2003; Snel *et al.*, 2002), and a question arises as to what kind of physiological limits may exist for a prokaryotic organism in terms of radical alterations to ORF organization. For instance, mobile elements are thought to disrupt functional barriers like replichore balancing (Preston *et al.*, 2004) and cotranscriptional association with the direction of replication (Andersson *et al.*, 1998; Brüggemann *et al.*, 2003).

My hypotheses of physical clustering initially approach the question of ORF density and organization by segmenting the physical chromosome into spatial regions. There are three basic hypotheses: 1) ORF density is random; 2) there is periodicity to the distribution of ORF densities on the chromosome; and 3) ORF densities form localized shapes that are non-random and interdependent with other regions on the chromosome. A controlling parameter to the evaluation of these hypotheses is the actual segmentation size (region length in base pairs) used to count up numbers of ORFs per segment. A parallel hypothesis relates to some ORFs being more important than other ORFs, and my fourth hypothesis is that only a limited subset of 75% of annotated ORFs are truly coding for function (Jackson *et al.*, 2002; Tatusov *et al.*, 2003).

An additional set of hypotheses is based on the notion that varying arrangements of open reading frames would, in part, reflect different sets of recombinative events occurring in the midst of evolutionary dynamics. In this set of hypotheses, I seek to evaluate whether there is any intragenomic aspect of ORF clustering that occurs robustly as a uniform property of each prokaryotic organism. These hypotheses are: 1) there is cotranscriptional association with the direction of replication for all prokaryotic organisms; and 2) the physical clustering of ORFs within COGs is non-random. I also seek to revisit my three spatial hypotheses based on evaluation of a 75% subset of ORFs constructed by filtering out those annotated ORFs that are putative false positives. As a testable outcome to the study, I would postulate that a meaningful measure of the internal physical clustering of ORFs would show some characteristic of vertical ancestry, and that outliers from a trend of vertically conserved ORF organization are attributable to the activity of mobile elements.

The recent increase in the number of publicly available, fully sequenced genomes has led to an opportunity for revisiting questions concerning the nature and organization of ORFs.

By investigating relationships between distributions of ORFs on fully sequenced prokaryotic chromosomes, this study measures the internal physical clustering of open reading frames. The performance of ORF organization as an indicator of vertical evolution can be assessed from estimated times of divergence from a last common ancestor as they are available in published studies (Battistuzzi *et al.*, 2004) and there are initial summaries of mobile element densities that may account for variation within the data set (Moran & Plague, 2004).

Chapter 2: Methods and Developed Methodology

2.1 Analytical Design

There is difficulty with establishing parameters for how genomic organization influences the phenotype of a strain, and I did not find previously developed parametric methods for fully characterizing ORF organization on the genome as a product of evolution. In order to arrive at legitimate statistical inferences, I structured the analysis to take into account the limited sample size and, where possible, avoid a priori assumptions.

Figure 1 is a synthesized view of how this study navigates between approaches to random and non-random resamplings and inference based on a structured approach to data analysis (Lunneborg, 2000). The data set provides ORF annotations that contain a potentially separable mixture of both real ORFs and putative false positives. I sought to establish a significant filtering between real and false ORFs and further compare the results of this distinction to random assignment of "realness." To examine this distinction over the chromosome, I conducted segmentation analyses to examine ORF clustering by delineating sections of the chromosome and, as a negative control, shuffling the $x_1, x_2, x_3, ..., x_n$ series of ORF regions of segmentation size r (Figure 2). Systems of simulation and comparisons against the likely phylogenetic tree are two approaches for evaluating the potential types of causes that might be associated with given chromosomal organizations of ORFs. In the event that random resampling may not allow for testing inferences of causality or population, I sought to perform basic subsampling to see how the data may be robustly described and effectively interpreted. A robust description that has significant coverage across either the phylogeny or functional grouping may lead to more confident assessments of constraints associated with the underlying natural system.

It is in the form of a rough confirmatory analysis (Behrens, 1997; Darlington, 1990) that I moved beyond a merely correlative approach to evaluate what underlying physiological and phenotypic relationships may relate the arrangement and mobility of genomic structure to

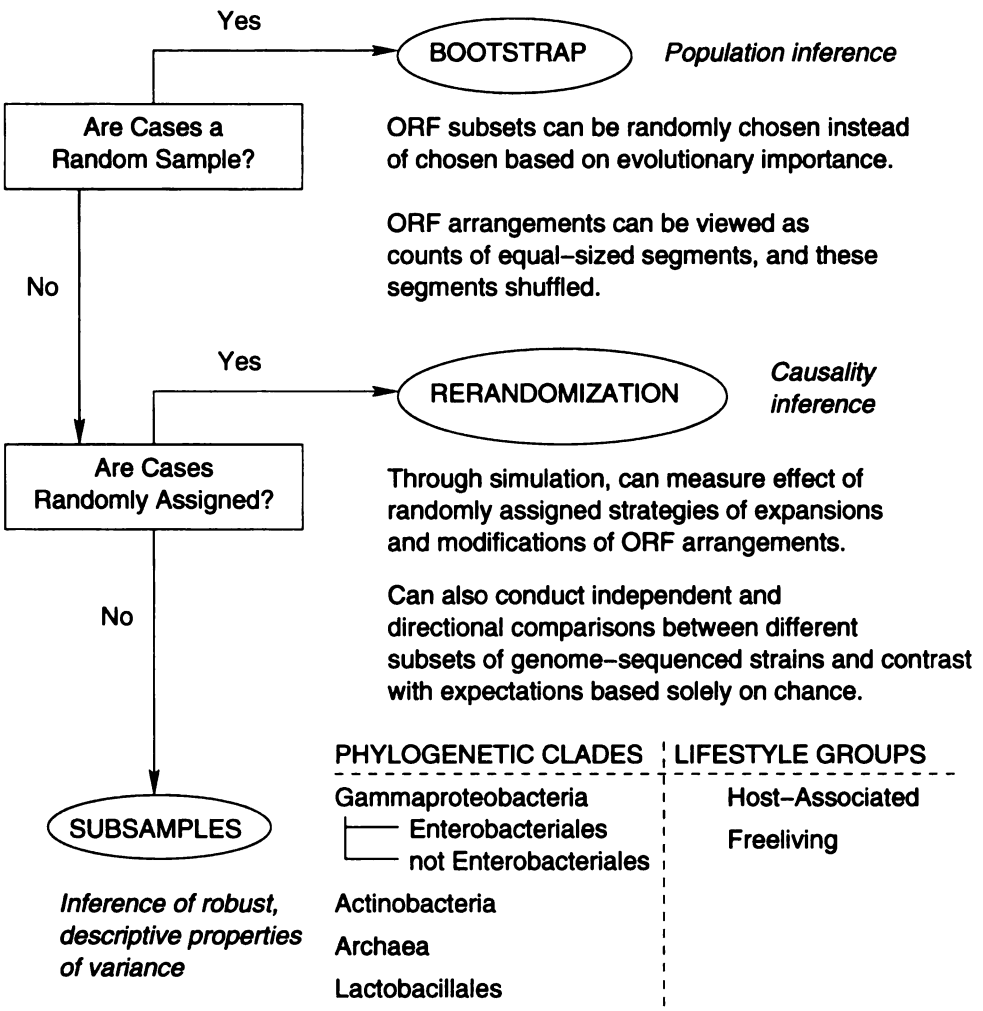

Figure 1: Resampling strategies and randomization design. Differing assumptions (shown in boxes) underlying a data set control the different ways (shown in ovals) there are for describing and resampling the data. My specific techniques for assaying the data are described in the text outside of each oval.

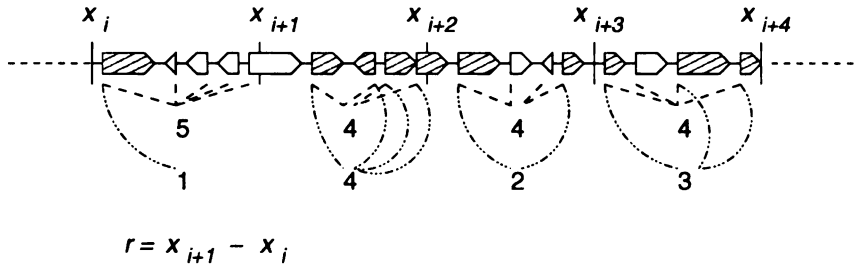

Figure 2: Calculation of regional ORF counts. The DNA sequence of a chromosome is subdivided into regions of equal physical length in base pairs. The locations of region boundaries are symbolically represented by the series $..., x_i, x_i + 1, x_i + 2, x_i + 3, ...$ The translational start point of each ORF, shown as a straight vertical edge, is used as the reference point for counting within each region. Two region-based ORF counting series are presented. The upper series is a count of all ORFs occurring within each region. The bottom series is a count corresponding to a filtered subset of ORFs (indicated by slashed shading). The process of counting is illustrated by the curved lines descending from the ORFs in of each chromosomal region to the associated ORF count.

the optimal function of the prokaryotic organism. With hypotheses concerning ORF density (Jackson *et al.*, 2002), organization (Kunisawa & Otsuka, 1988; Jurka & Savageau, 1985), and the role of mobile elements (Bentley & Parkhill, 2004; Ochman & Davalos, 2006), I sought to investigate general correspondences and detailed variation.

2.2 Data Assembly

2.2.1 Collection of Genome Data

The data set of fully sequenced prokaryotic genomes was accessed from the National Center for Biotechnology Information (NCBI) public archives (Wheeler *et al.*, 2000) in March, 2004. Data set files in these archives are distributed per chromosome and plasmid replicons. For the 155 fully sequenced genomes, there were 234 sets of files corresponding to 165 chromosomes and 69 plasmids available from the FTP address `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/`. There were four file formats for each of the replicons (Table 1). Specific versions of genomes corresponding to the original time of download can be accessed by visiting `http://www.ncbi.nlm.nih.gov/entrez/sutils/girevhist.cgi`.

The species of Archaea and their associated chromosomal accession numbers are

Aeropyrum pernix (NC_000854), *Archaeoglobus fulgidus* DSM 4304 (NC_000917), *Halobacterium* sp. NRC-1 (NC_002607), *Methanocaldococcus jannaschii* (NC_000909), *Methanopyrus kandleri* AV19 (NC_003551), *Methanosarcina acetivorans* C2A (NC_003552), *Methanosarcina mazei* Goe1 (NC_003901), *Methanothermobacter thermautotrophicus* str. Delta H (NC_000916), *Nanoarchaeum equitans* Kin4-M (NC_005213), *Pyrobaculum aerophilum* str. IM2 (NC_003364), *Pyrococcus abyssi* (NC_000868), *Pyrococcus furiosus* DSM 3638 (NC_003413), *Pyrococcus horikoshii* (NC_000961), *Sulfolobus solfataricus* (NC_002754), *Sulfolobus tokodaii* (NC_003106), *Thermoplasma acidophilum* (NC_002578), and *Thermoplasma volcanium* (NC_002689).

The species of Bacteria and their associated chromosomal accession numbers are *Agrobacterium tumefaciens* str. C58 (Cereon) (NC_003062, NC_003063), *Agrobacterium tumefaciens* str. C58 (U. Washington) (NC_003304, NC_003305), *Aquifex aeolicus* VF5 (NC_000918), *Bacillus anthracis* str. A2012 (NC_003995), *Bacillus anthracis* str. Ames (NC_003997), *Bacillus cereus* ATCC 10987 (NC_003909), *Bacillus cereus* ATCC 14579 (NC_004722), *Bacillus halodurans* (NC_002570), *Bacillus subtilis* subsp. *subtilis* str. 168 (NC_000964), *Bacteroides thetaiotaomicron* VPI-5482 (NC_004663), *Bdellovibrio bacteriovorus* (NC_005363), *Bifidobacterium longum* NCC2705 (NC_004307), *Bordetella bronchiseptica* (NC_002927), *Bordetella parapertussis* (NC_002928), *Bordetella pertussis* (NC_002929), *Borrelia burgdorferi* B31 (NC_001318), *Bradyrhizobium japonicum* USDA 110 (NC_004463), *Brucella melitensis* 16M (NC_003317, NC_003318), *Brucella suis* 1330 (NC_004310, NC_004311), *Buchnera aphidicola* str. APS (*Acyrthosiphon pisum*) (NC_002528), *Buchnera aphidicola* str. Bp (*Baizongia pistaciae*) (NC_004545), *Buchnera aphidicola* str. Sg (*Schizaphis graminum*) (NC_004061), *Campylobacter jejuni* subsp. *jejuni*

Table 1: Descriptions of four file formats for the NCBI prokaryotic genome FTP repository.

Format	**Description**
`.asn`	ASN stands for abstract syntax notion
`.gbk`	a readable plain text version of the `.asn` files
`.faa`	FASTA-formatted listing of amino acid sequences
`.ffn`	FASTA-formatted listing of coding strand nucleotide sequences

NCTC 11168 (NC_002163), *Candidatus Blochmannia floridanus* (NC_005061), *Caulobacter crescentus* CB15 (NC_002696), *Chlamydia muridarum* (NC_002620), *Chlamydia trachomatis* (NC_000117), *Chlamydophila caviae* GPIC (NC_003361), *Chlamydophila pneumoniae* AR39 (NC_002179), *Chlamydophila pneumoniae* CWL029 (NC_000922), *Chlamydophila pneumoniae* J138 (NC_002491), *Chlamydophila pneumoniae* TW-183 (NC_005043), *Chlorobium tepidum* TLS (NC_002932), *Chromobacterium violaceum* ATCC 12472 (NC_005085), *Clostridium acetobutylicum* (NC_003030), *Clostridium perfringens* str. 13 (NC_003366), *Clostridium tetani* E88 (NC_004557), *Corynebacterium diphtheriae* (NC_002935), *Corynebacterium efficiens* YS-314 (NC_004369), *Corynebacterium glutamicum* ATCC 13032 (NC_003450), *Coxiella burnetii* RSA 493 (NC_002971), *Deinococcus radiodurans* R1 (NC_001263, NC_001264), *Enterococcus faecalis* V583 (NC_004668), *Escherichia coli* CFT073 (NC_004431), *Escherichia coli* K-12 (NC_000913), *Escherichia coli* O157:H7 (NC_002695), *Escherichia coli* O157:H7 EDL933 (NC_002655), *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586 (NC_003454), *Geobacter sulfurreducens* PCA (NC_002939), *Gloeobacter violaceus* (NC_005125), *Haemophilus ducreyi* 35000HP (NC_002940), *Haemophilus influenzae* Rd KW20 (NC_000907), *Helicobacter hepaticus* ATCC 51449 (NC_004917), *Helicobacter pylori* 26695 (NC_000915), *Helicobacter pylori* J99 (NC_000921), *Lactobacillus johnsonii* NCC 533 (NC_005362), *Lactobacillus plantarum* WCFS1 (NC_004567), *Lactococcus lactis* subsp. *lactis* (NC_002662), *Leptospira interrogans* serovar *lai* str. 56601 (NC_004342, NC_004343), *Listeria innocua* (NC_003212), *Listeria monocytogenes* EGD-e (NC_003210), *Mesorhizobium loti* (NC_002678), *Mycobacterium avium* subsp. *paratuberculosis* str. k10 (NC_002944), *Mycobacterium bovis* subsp. *bovis* AF2122/97 (NC_002945), *Mycobacterium leprae* (NC_002677), *Mycobacterium tuberculosis* CDC1551 (NC_002755), *Mycobacterium tuberculosis* H37Rv (NC_000962), *Mycoplasma gallisepticum* R (NC_004829), *Mycoplasma genitalium* (NC_000908), *Mycoplasma mycoides* subsp. *mycoides* SC (NC_005364), *Mycoplasma penetrans* (NC_004432), *Mycoplasma pneumoniae* (NC_000912), *Mycoplasma pulmonis* (NC_002771), *Neisseria meningitidis* MC58 (NC_003112), *Neisseria meningitidis* Z2491 (NC_003116), *Nitrosomonas europaea* ATCC 19718 (NC_004757), *Nostoc* sp. PCC 7120 (NC_003272), *Oceanobacillus iheyensis* HTE831 (NC_004193), *Onion yellows phytoplasma* (NC_005303), *Pasteurella multocida* (NC_002663), *Photorhabdus luminescens*

subsp. *laumondii* TTO1 (NC_005126), *Pirellula* sp. 1 (NC_005027), *Porphyromonas gingivalis* W83 (NC_002950), *Prochlorococcus marinus* str. MIT 9313 (NC_005071), *Prochlorococcus marinus* subsp. *marinus* str. CCMP1375 (NC_005042), *Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986 (NC_005072), *Pseudomonas aeruginosa* PAO1 (NC_002516), *Pseudomonas putida* KT2440 (NC_002947), *Pseudomonas syringae* pv. tomato str. DC3000 (NC_004578), *Ralstonia solanacearum* (NC_003295), *Rhodopseudomonas palustris* CGA009 (NC_005296), *Rickettsia conorii* (NC_003103), *Rickettsia prowazekii* (NC_000963), *Salmonella enterica* subsp. *enterica* serovar Typhi (NC_003198), *Salmonella enterica* subsp. *enterica* serovar Typhi Ty2 (NC_004631), *Salmonella typhimurium* LT2 (NC_003197), *Shewanella oneidensis* MR-1 (NC_004347), *Shigella flexneri* 2a str. 2457T (NC_004741), *Shigella flexneri* 2a str. 301 (NC_004337), *Sinorhizobium meliloti* (NC_003047), *Staphylococcus aureus* subsp. *aureus* MW2 (NC_003923), *Staphylococcus aureus* subsp. *aureus* Mu50 (NC_002758), *Staphylococcus aureus* subsp. *aureus* N315 (NC_002745), *Staphylococcus epidermidis* ATCC 12228 (NC_004461), *Streptococcus agalactiae* 2603V/R (NC_004116), *Streptococcus agalactiae* NEM316 (NC_004368), *Streptococcus mutans* UA159 (NC_004350), *Streptococcus pneumoniae* R6 (NC_003098), *Streptococcus pneumoniae* TIGR4 (NC_003028), *Streptococcus pyogenes* M1 GAS (NC_002737), *Streptococcus pyogenes* MGAS315 (NC_004070), *Streptococcus pyogenes* MGAS8232 (NC_003485), *Streptococcus pyogenes* SSI-1 (NC_004606), *Streptomyces avermitilis* MA-4680 (NC_003155), *Streptomyces coelicolor* A3(2) (NC_003888), *Synechococcus* sp. WH 8102 (NC_005070), *Synechocystis* sp. PCC 6803 (NC_000911), *Thermoanaerobacter tengcongensis* (NC_003869), *Thermosynechococcus elongatus* BP-1 (NC_004113), *Thermotoga maritima* (NC_000853), *Treponema denticola* ATCC 35405 (NC_002967), *Treponema pallidum* (NC_000919), *Tropheryma whipplei* TW08/27 (NC_004551), *Tropheryma whipplei* str. Twist (NC_004572), *Ureaplasma urealyticum* (NC_002162), *Vibrio cholerae* (NC_002505, NC_002506), *Vibrio parahaemolyticus* RIMD 2210633 (NC_004603, NC_004605), *Vibrio vulnificus* CMCP6 (NC_004459, NC_004460), *Vibrio vulnificus* YJ016 (NC_005139, NC_005140), *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis* (NC_004344), *Wolbachia* endosymbiont of *Drosophila melanogaster* (NC_002978), *Wolinella succinogenes* (NC_005090), *Xanthomonas axonopodis* pv. *citri* str. 306 (NC_003919), *Xanthomonas campestris* pv. *campestris* str.

ATCC 33913 (NC_003902), *Xylella fastidiosa* 9a5c (NC_002488), *Xylella fastidiosa* Temecula1 (NC_004556), *Yersinia pestis* CO92 (NC_003143), and *Yersinia pestis* KIM (NC_004088).

I inspected the data of these chromosomes for annotated circular or linear topologies and also identified those genomes with multiple, distinct chromosomes based on information in the NCBI data files and in the literature. Other key attributes to the chromosomes were their physical lengths and NCBI taxonomy.

I evaluated other general qualities to the data set such as associated plasmids as well as changes to ORF lengths and ORF annotations over time. I identified the sequenced plasmids associated with each of the 155 genomes based on the NCBI data files and compared this to what was characterized in the literature. I evaluated the number of fully sequenced genomes, ORF counts, and distribution values for ORF lengths at three different dates: December 2002, March 2004, and June 2005. As a prelude to intensive analysis of ORF data, I identified changes to the ORF accession versions as they occurred between the three different dates of December 2002, March 2004, and June 2005. Based on changes in ORF accession version numbers, I counted up and characterized the changes to ORFs and the number of associated genomes.

2.2.2 Management of ORF Data

The scope of the data set involved 447,551 ORFs present on 165 chromosomes as well as other associated chromosomal attributes that were originally sourced from the NCBI (Wheeler *et al.*, 2000). I did the initial parsing of NCBI data files with various small perl scripts and manual investigations of resultant output. I managed BLASTP calculations and statistical analyses among the ORFs with a controllable analytical software pipeline I constructed across a set of five computers. To make a clear division between the initial parsing and subsequent analyses of specific ORF subsets, I developed a web application named MYCROW (Matrix-Yanking Coding Region Objects Workbench) as a front-end to my software pipeline. The central feature of MYCROW is to allow users to retrieve specific chromosomal sets of ORF records that provide fields containing various descriptive and quantitative characteristics for each ORF. I developed a simple XML approach to manage the packaging, installation, upgrading, and running of service scripts and application files

Table 2: Abbreviated labels for chromosomal accession numbers.

Accession	Label	NCBI Name
NC_000854	A.pnx.	*Aeropyrum pernix*
NC_000917	Arch.ful.	*Archaeoglobus fulgidus* DSM 4304
NC_002696	Caul.cre.	*Caulobacter crescentus* CB15
NC_003450	Cor.glut.	*Corynebacterium glutamicum* ATCC 13032
NC_002662	Lac.lact.	*Lactococcus lactis* subsp. *lactis*
NC_003552	Mt.acet.	*Methanosarcina acetivorans* C2A
NC_003901	Mt.maz.	*Methanosarcina mazei* Goe1
NC_002755	tb1551	*Mycobacterium tuberculosis* CDC1551
NC_000962	tbH37Rv	*Mycobacterium tuberculosis* H37Rv
NC_000908	M.gen.	*Mycoplasma genitalium*
NC_000912	M.pnm.	*Mycoplasma pneumoniae*
NC_002771	M.pulm.	*Mycoplasma pulmonis*
NC_003272	Nostoc	*Nostoc sp.* PCC 7120
NC_000868	Py.aby.	*Pyrococcus abyssi*
NC_000961	Py.hor.	*Pyrococcus horikoshii*
NC_003413	Py.fur.	*Pyrococcus furiosus* DSM 3638
NC_003103	R.con.	*Rickettsia conorii*
NC_000963	R.pro.	*Rickettsia prowazekii*
NC_002737	S.pyog.	*Streptococcus pyogenes* M1 GAS
NC_003028	Str.pnm.	*Streptococcus pneumoniae* TIGR4
NC_002754	Sulf.solf.	*Sulfolobus solfataricus*
NC_003106	Sulf.tok.	*Sulfolobus tokodaii*
NC_000911	Synec.	*Synechocystis sp.* PCC 6803
NC_004113	Th.elon.	*Thermosynechococcus elongatus* BP-1
NC_003919	Xn.axon.	*Xanthomonas axonopodis* pv. *citri* str. 306
NC_003902	Xn.cmp.	*Xanthomonas campestris* pv. campestris str. ATCC 33913
NC_004556	X.fas.	*Xylella fastidiosa* Temecula1

associated with the MYCROW system as it emerged from a prototype into a reliable laboratory solution.

Two of the five computers were used for archiving and curating the data set. The front-end archive computer was used to dynamically compile specific data sets for further analysis, and was set up as a web server to provide an interface for specifying and retrieving various matrices of ORF-specific data. The back-end archive computer handled and processed a pipeline of information coming from external sources such as NCBI. Much of the parsing of NCBI genome files and sequence-level BLAST calculations occurred on the back-end archive computer. Both archive computers ran FreeBSD 4.9 or higher.

I used the other three of the five computers for data analysis of the curated, retrievable data set. One of these computers was used as a relational database and archive of statistical methods. The database was run with MySQL version 3.23 or higher. The interactive data analysis was distributed on 2 other computers. All three of these computers ran Mandriva Linux version 10.0 or higher. The "S" statistics language (Chambers & Hastie, 1992) was used to perform most of the statistical calculations, and was run on the "R" statistical environment (R Development Core Team, 2005). Perl and R were the primary software languages used to write necessary algorithms on both the curatorial and data analysis computers.

The front-end web server computer ran with an Intel Pentium 4 ® CPU 1.60 GHz computer chip, and 750 MB RAM memory. The back-end archive server ran with an Intel Pentium 4 ® CPU 2.40 GHz, and 750 MB RAM memory. The statistics archive and MySQL database server ran on a Pentium II ® 400 MHz computer chip, and 500 MB RAM memory. One of the computers for interactive data analysis ran on a AMD Athlon ® 1.67 Hz computer chip, and 500 MB RAM memory. The other computer used for interactive data analysis ran on a Pentium III ® 931 MHz computer chip, and 500 MB RAM memory.

For the purposes of a final, expedited run of the bootstrap residue calculations, a 128-processor computer was used, courtesy of the MSU High Performance Computing Center (`http://www.hpc.msu.edu/`).

Screenshots of the MYCROW user interface are shown in Figures 3 and 4. The options for the retrieval fields associated with ORF records are: similarity score (based on

expectation score $< 10^{-6}$), GenBank accession number (the sequence-unique "geninfo" number was used for this), annotation, function, product, amino acid length, nucleotide length, chromosome location of the translational start point, chromosome location of the translational end point (excluding stop codon), polarity (orientation on chromosome), chromosome shape, chromosome size, a series of numbers quantifying the period-three signal in the DNA, organism identifier (NCBI taxon id), organism name (the NCBI epithet-like name), translation table (the value is based on descriptions at EMBL), statistical profile (% GC, codon counts), DNA sequence, and amino acid sequence.

2.2.3 Taxonomic and Phylogenetic Categorizations

I applied a taxonomic hierarchy to the data set of fully sequenced genomes by using the taxonomic rankings and nomenclature from the National Center for Biotechnology Information (NCBI). The chromosomal accession numbers were used as a querying list for the NCBI Taxonomy Common Tree application (`http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi`). The data was parsed, and an outline of taxonomic groups built, by tallying up those taxonomic groupings that contained less than 20 representative strains.

For subsampling, I generally used five taxonomy-based groupings of genomes. These taxonomic groupings are: Archaea - 17 genomes; Actinobacteria - 13 genomes; Enterobacteriales - 17 genomes; Gammaproteobacteria without Enterobacteriales - 20 genomes; and Lactobacillales - 13 genomes.

I built four phylogenetic trees with times of divergence based on estimates from Battistuzzi *et al.* (2004) to characterize samplings of Archaea, Gammaproteobacteria, Bacilli, and Actinobacteria. I evaluated the growth of phylogenetic range by looking at the number of available chromosomes, species with more than one representative strain, and number of representative phyla for each year since the beginning of 1995 based on associated date of publication in the literature or submission to GenBank.

I also sought to evaluate the relationship of the set of 155 genomes with natural diversity of the prokaryotic biota based on an historical reference on infectious disease (Hoeprich, 1972) and my own classifications of ecological and lifestyle categories based largely on

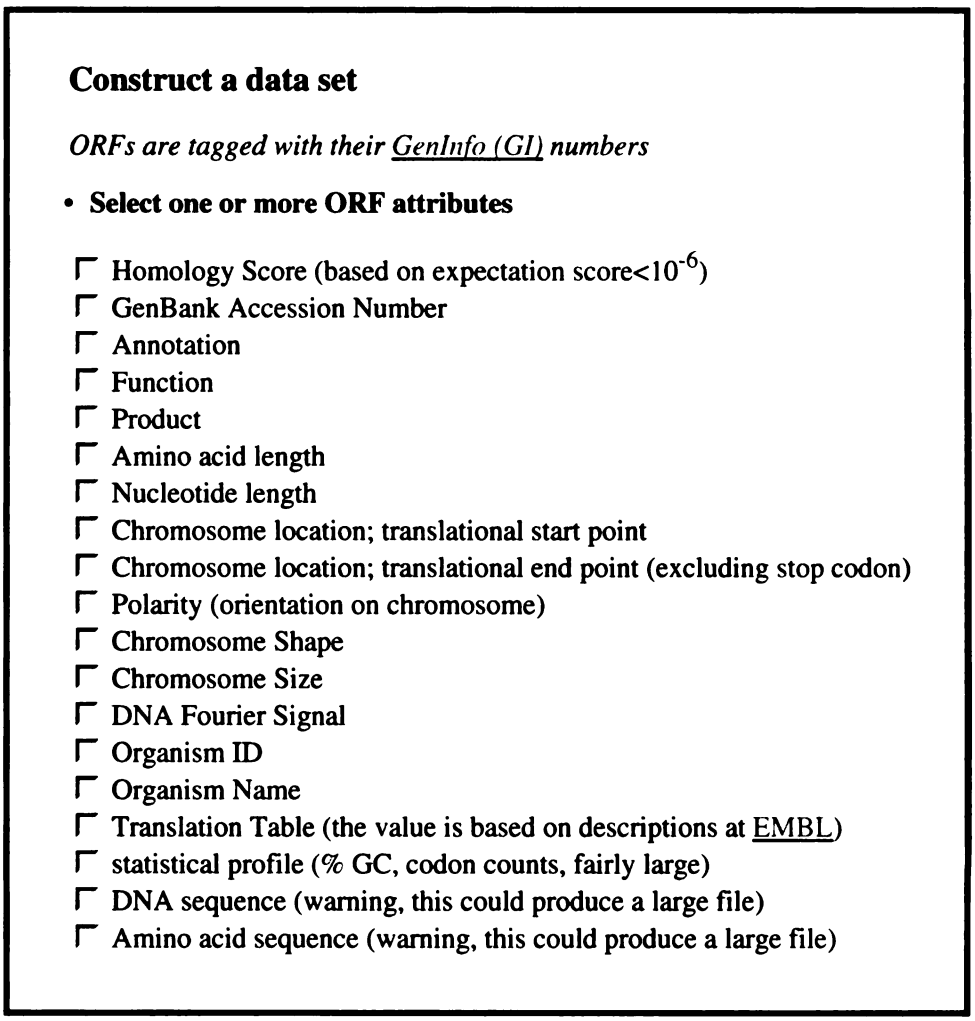

Figure 3: Selecting ORF attributes for retrieval from the online MYCROW information retrieval web page. The user constructs the columns of a data set by selecting checkboxes corresponding to ORF features of interest.

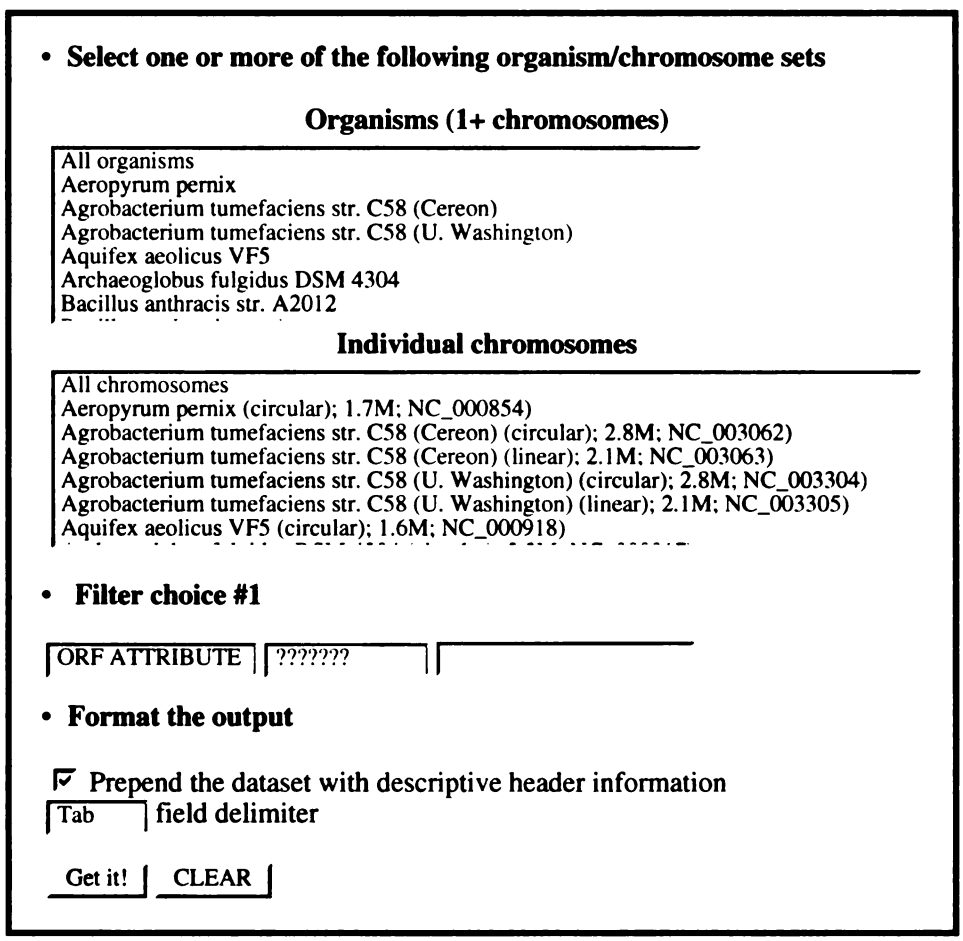

Figure 4: Selecting a set of chromosomes or organisms from the online MYCROW information retrieval web page. Options for conditionally filtering the set of ORFs based on ORF attributes, and options for formatting the output are provided in addition to the chromosome and organism selection windows.

descriptions in the original genome journal publications.

2.2.4 External ORF-Based Data Sets

A study from Covert *et al.* (2004) focuses on *Escherichia coli* K-12, and uses functional predictions and microarray assays of gene expression to better characterize bacterial networks. Their "Supplementary Data 6" data file provides raw microarray data organized into treatments of one wild-type strain and six knockout mutant strains growing under aerobic versus anaerobic conditions. Of the 3699 ORF regions listed in the microarray data, 3309 of these regions (89%) had identifiers that mapped to ORFs within the MYCROW database records for *Escherichia coli* K-12. There were 3 strain replicate trials for each type of strain under both aerobic and anaerobic conditions. The resulting 42 characterizations of gene expression (present, absent, marginal) were evaluated to assess the general transcriptional expression of each ORF.

The analysis of *Bacillus subtilis* by Biaudet *et al.* (1997) characterizes 19 phenotypic consequences of mutation associated with 554 genes. In that study, mutations are found to range in effect from single phenotypic changes to six phenotypic changes. The phenotypic categories are: osmotic stress, oxidative stress, temperature and pH stress, electron transfer, general stress, stress by metals, starvation stress, N or C sources, glucose effect, amino acid induction and repression, amino-acids and translation, macromolecules, protein/secretion, envelope/lysogeny, envelope/AP/BG, cell cycle, competence, sporulation, and germination. There were 533 of the 554 genes that had unambiguous matches to *B. subtilis* ORFs inside the MYCROW database based on identifier information.

I compared the degree of paralogous representation of my ORF similarity clusters (as calculated by the criteria of section 2.5) to paralogy sets as calculated by (Pushker *et al.*, 2004) for 4 strains of *Escherichia coli*, 3 species of *Pseudomonas*, 4 strains of *Streptococcus pyogenes*, and 3 strains of *Staphylococcus aureus*.

2.2.5 Mobile Elements

Data on IS element density for 50 genomes came from a study by Moran & Plague (2004). From their graphical plot of IS density, I assayed density values by intervals of 1 IS

element per 350,018 bp to construct boundaried bins of $i/350018$ to $(i+1)/350018$ where $i = \{1, 2, 3, ..., 34, 35\}$. The characteristic IS element density for each genome was set to the value of i. The data for IS element served to comparatively characterize underlying molecular factors of disruption to chromosomal organization.

2.3 Dot Matrix Evaluation of Conserved Chromosomal Organization

To build dot matrices, I used data from the NCBI GenePlot application (Wheeler *et al.*, 2000) to gather the symmetrical best hits of ORFs as they occur between pairs of genomes. I translated the identifiers of the bidirectional best hits, as catalogued by the NCBI GenePlot application, to base pair coordinate information from ORF objects stored inside the MYCROW data files. Each dot matrix represents a pairwise comparison upon which a greater-ranging phylogenetic analysis can be conducted. I constructed a set of nine phylogenetic comparisons where each comparison involved three strains with characterized times of divergence from Battistuzzi *et al.* (2004). Among these sets of three strains, two of the strains were more closely related to each other than to a third more "distant cousin."

2.4 Measuring Mutual Information

Mutual information (Weaver & Shannon, 1949; Feeny & Lin, 2004; Church & Hanks, 1990) was a measurement approach for two contexts. As applied to the dot matrix plots, I divided the plots into square windows $s_{x,y}$ of a given size w (see Figure 5). The remaining, sometimes rectangular, windows on the mth row and nth column were also included in the analysis. These square (and remaining) windows were summed up by columns $(c_y = \sum_{i=1}^{m}(s_{i,y}))$ and rows $(r_x = \sum_{i=1}^{n}(s_{x,i}))$. I calculated the total sum as $t = \sum_{x=1}^{m}\sum_{y=1}^{n}(s_{x,y})$. For each square window $s_{x,y}$ containing plotted points $(s_{x,y} \neq 0)$, I calculated the pointwise mutual information by Equation 1. The average pointwise mutual information was the arithmetic mean of all values of $I_G(x,y)$ where $s_{x,y} \neq 0$. I evaluated the average pointwise mutual information for various window sizes,

w = 10 kb, 20 kb, 30 kb, ... , 150 kb. I wrote a function that ran in the "R" statistics package, version 1.9.1, to perform this calculation.

$$I_G(x,y) = \log_2 \left(\frac{\frac{s_{x,y}}{t}}{\frac{r_x}{t} \cdot \frac{c_y}{t}} \right) \quad (1)$$

My second context of usage was to evaluate the lag average mutual information (AMI) calculations for neighboring window-sized calculations of intrachromosomal organization (see section 2.8). I calculated lag mutual information (in this case based on natural logarithms) with the "mutual" function of the "tseriesChaos" package, version 0.1-6 with the "R" statistics package, version 2.3.0. For a given window-sized calculation of intrachromosomal organization $G_w(h)$ on a chromosome h and the ORF densities based on that window w_i (see the P and Q measures of Section 2.8.3), the AMI was calculated for various lag comparisons b based on the following procedure. At most, 16 bins u_i on the ORF densities series d_i were determined. The mutual information $I_L(x,y)_b$ between each pair of differing bins, u_x and u_y, was calculated as shown in Equation 2. Let $H(w,i,h) = G_{w_i}(h)$. The b value is the amount of lag between the two series, $j = H(w,1,h), H(w,2,h), H(w,3,h), ..., H(w,N,h)$ and $k = H(w,b+1,h), H(w,b+2,h), H(w,b+3,h), ..., H(w,b+N,h)$. A circular boundary condition was applied where $H(w,i+N,h) = H(w,i,h)$. Two counting functions were used. There was a counting function that determined the number of j or k values falling within a given respective bin $q(u_x)$ or $q(u_y)$. There was also a counting function that determined the number of j and k values jointly falling within two bins $q(u_x,u_y)$ at the specified lag b. I verified performance of the "tseriesChaos" package's "mutual" function by custom-writing a separate function that produced the same results over various test cases.

$$I_L(x,y)_b = \log \left(\frac{\frac{q(u_x,u_y)}{N}}{\frac{q(u_x)}{N} \cdot \frac{q(u_y)}{N}} \right) \quad (2)$$

I compared the lag AMI series for the different measures of intrachromosomal organization described in Section 2.8.3.

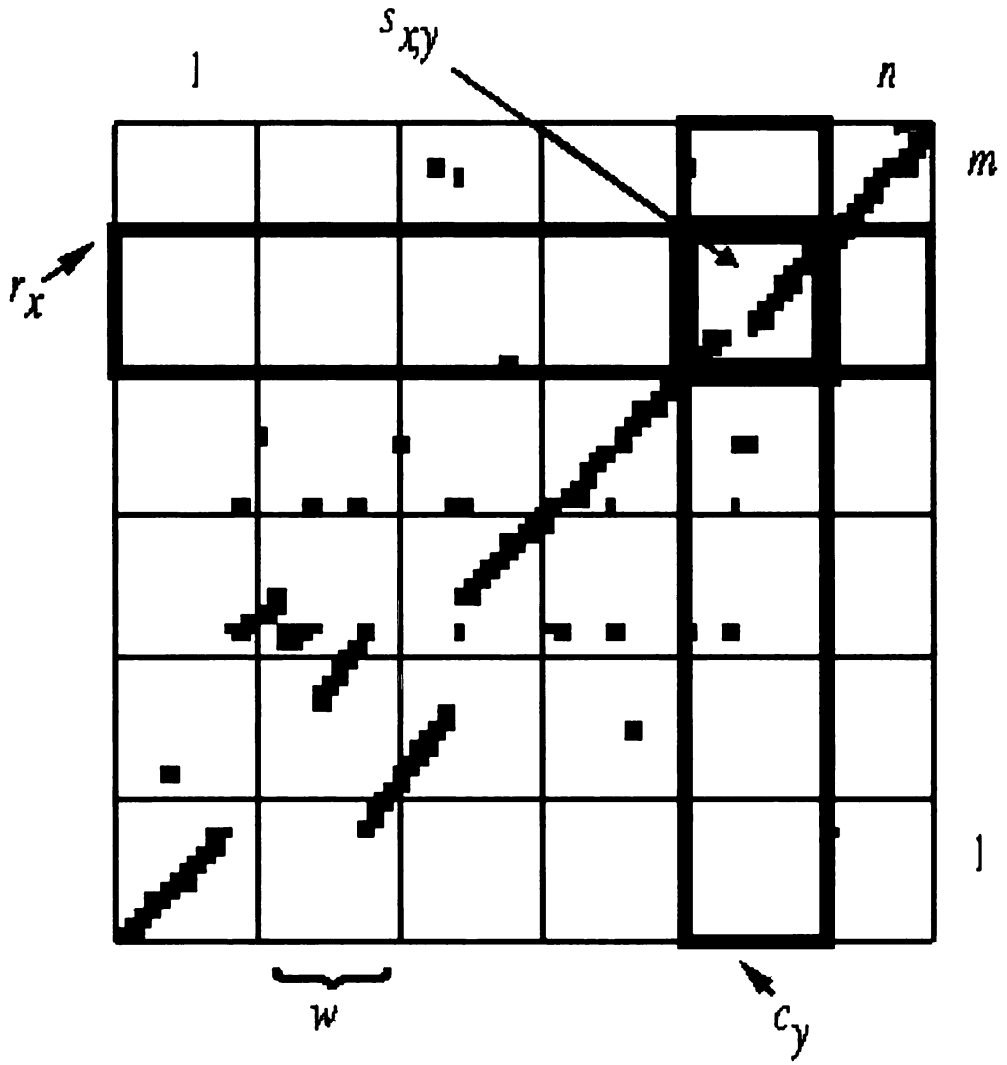

Figure 5: Calculating the pointwise mutual information on a dot matrix of conserved ORF organization. A grid of m rows and n columns is applied to a dot matrix based on a given window size w. Rows (r_x), columns (c_y) and squares on the grid (r_x) are regions used for counting up relative densities to the overall number of dots on the dot matrix.

2.5 Specification of Operational ORF Subset

Similarity counts were evaluated for each of the 447,551 ORFs. The similarity count was the inclusive number of ORFs in the set of 447,551 ORFs that matched a similarity filter for a given ORF based on characteristics of the amino acid sequences. The similarity filter involved two requisite criteria: a BLASTP expectation score $\leq 10^{-6}$, and an amino acid length difference of at most $\pm$ 10%. To avoid computational times exceeding one month, I did not calculate 447,551 consecutive one-to-many BLASTP comparisons, nor did I run a BLASTCLUST computation on the entire set of 447,551 ORFs. Rather, I developed a speedier approach that took 8 days on the archive computer. This approach is done by calculating sets of ORF that range in length by 10%. The parameters for length variation are incremented stepwise by 5%. BLASTCLUST is then calculated on each length-based set of ORFs. Then, for each ORF, I re-tallied the computed similarity clusters to identify the putatively matching ORFs that were within a length range of $\pm$ 10% of the particular ORF. I performed one iteration of resolving transitive relationships.

I then categorized each ORF by the number of similar ORFs in the data set of 447,551 ORFs. An ORF that was not "similar" (according to the criteria of 10% length and $\leq 10^{-6}$ expectation score similarity) to any other ORF was in a cluster of size 1. A pair of ORFs that were only similar to each other were in a similarity cluster of size 2. I characterized O-ORFs as belonging to similarity clusters of size ≥ 6, and the putatively false set of ORFs (S-ORFs) as belonging to similarity clusters of size ≤ 5.

2.6 Running Tally

I developed an approach I call a "running tally" to measure the nonrandomness of chromosomal clustering and ORFs. Running tallies contrast the spatial chromosomal clustering attributable to original assignments of ORF properties with the clustering effect due to randomized assignments of ORF properties. Figure 6 illustrates how the running tally approach can visually present both the magnitude and shape of the natural invariance compared to a randomized control. The running tally approach is similar to that of plotting and measuring a random walk (Pearson, 1905).

I initially used running tallies on the ORFs of each chromosome to assess both the inclusion and exclusion of ORFs within COGs as well as the associated coding strand (polarity). Polarity data came from the MYCROW web application. COG data was accessed from the COG database, `ftp://ftp.ncbi.nih.gov/pub/COG/COG` (Tatusov *et al.*, 1997b, 2003), and represents data updated on March 2, 2003. The scope of NCBI's COG database limited the evaluation to just 67 (41%) of the 165 chromosomes. Of these 67 chromosomes, there were 4 Actinobacteria, 13 Archaea, 12 Gammaproteobacteria, 3 Lactobacillales and 32 bacteria belonging to other taxonomic classes. I also used running tallies to evaluate the clustering of O-ORF and S-ORF assignments.

I used a bootstrap to calculate the z-score difference between running tally measures related to two negative controls versus those running tally measures involving original assignments. I measured the difference between each pairwise comparison by calculating the integral area between the two running tallies. I characterized the z-score value by the standard deviation σ of the distribution calculated by measured differences among pairs of negative controls. Significance was evaluated by bootstrap where, for multiple times, the area between running tallies of 50 randomized assignments versus the non-randomized assignment was calculated. The mean of these 50 measurements was calculated. This step was repeated 100 times so that there were 100 means from which a bootstrap estimate m_t was calculated. This estimate was contrasted with 50×100 measurements between the running tallies from pairs of two randomized assignments from which an estimate m_f was calculated and a standard deviation σ_f. The z-score difference between m_t and m_f was characterized as $(m_t - m_f)/\sigma_f$.

2.7 Simulation of Informational Expansion and Modification

I built a simulated model of recombination similar to an expansion-modification system (EMS) (Li, 1991), yet my simulation model is a probabilistic context-sensitive grammar as opposed to a probabilistic context-free grammar. My test was to see whether various initially set parameters of the simulation model can be inferred retrospectively from a measurement of

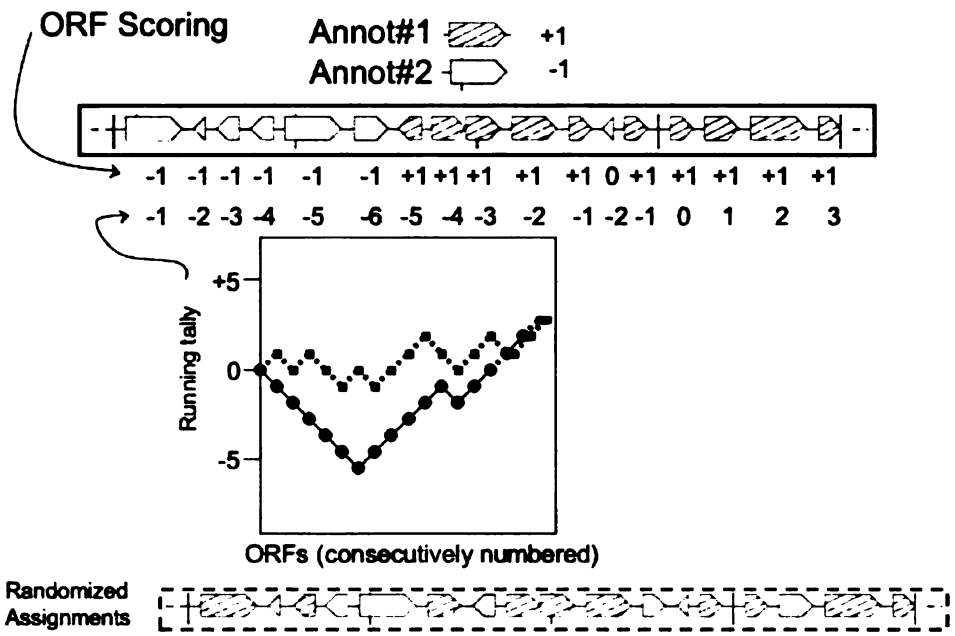

Figure 6: My method for counting up a running tally for original and randomized ORF annotations. Two annotation states are evaluated and scored with either a +1 or a -1. A deliberately constructed non-random pattern of ORFs is shown in the uppermost solid rectangle. A pattern of ORFs based on randomized assignments of ORF annotations is shown at the bottom of the figure in a dashed boundaried rectangle. A running tally series for the upper pattern of ORFs is plotted with closed circles and solid lines. The running tally series for the lower pattern of ORFs is plotted with small squares and dotted lines.

the constructed pattern built with simulated expansions and movement of information.

The symbolic structure of my EMS-like model consists of a series of letters from the set {A,B,C,D,E,F,G,H} as described in Equation 5. Various functions $T_x(S)$ and $D_x(S)$ are abstractions of cut-and-paste ("translocation") and tandem copy operations on the model replicon. While my symbolic model is not applicable to the physicochemical detail of recombinative change, it serves to 1) validate the idea that different sizes and stochastics associated with mobile and duplicating segments of a sequence can produce an interpretable signature and 2) test assumptions about how a given measure corresponds to underlying model parameters. To investigate segmentation patterns within the final output series of alphabetic symbols produced by each simulation ($s_p \in S$) I calculated densities of "H" characters within windowed subseries of each generated s_p.

Equations 3 and 4 show a sequence of 8 letters that is triplicated to form a sequence that is 24 letters in length. Depending on the value for n in Equation 4, other replicate structures of octets can be generated (e.g., $n = 2$, duplicated octets; $n = 4$, quadruplicated octets; or $n = 5$, quintuplicated octets). The n parameter acts to both set the length of the starting sequence and establish an initial, non-stochastic pattern.

The $D_x(S)$ rule system (Equation 9) is to tandemly duplicate a randomly selected internal 6-letter sequence. For example, the sequence ABCDEFGHABCDEFGH can have a randomly selected 6-letter subsequence - AB(CDEFGH)ABCDEFGH - that, when duplicated, creates ABCDEFGHCDEFGHABCDEFGH. The $N(Y) = 6$ condition is an adjustable, initial parameter for the simulation, and I evaluated this condition over a range of conditions $N(Y) = 2$ to $N(Y) = 12$.

The $T_x(S)$ rule system (Equations 6 - 8) involves identifying two locations of a "HA" subsequence. A target location for this translocational event is then identified and the translocation event performed. A more detailed illustration of this system is shown in Figure 63 in Chapter 5.

My stochastic rule system is shown in Equation 11 where q is a uniformly distributed random variable on the interval $[0, 1]$.

In a rough sense, I meant for the original simulation design to have one letter corresponding to 10,000 bases. Based on this relationship, 500 letters equals 5,000,000 bases,

a value that is loosely representative of a prokaryotic chromosome's size. The simulation ends once the sequence expands to more than 500 letters.

$$M \rightarrow \{A, B, C, D, E, F, G, H\} \tag{3}$$

$$J \rightarrow ABCDEFGH$$

(4)

$$L \rightarrow \{x^n \mid x = J\}$$

$$n = 3$$

$$L \rightarrow \{xxx \mid x = J\}$$

$$S \mid \ S \in \ M^* \tag{5}$$

$$T_1(S): \text{ If } \exists(W, X, Y, Z) \left| \begin{array}{l} (W \in M^*) \\ \wedge\ (X \in M^*) \\ \wedge\ (Y \in HAM^*HA) \\ \wedge\ (Z \in M^*) \\ \wedge\ (S \equiv WHAXYZ) \end{array} \right. , S \rightarrow WYXZ \tag{6}$$

$$T_2(S): \text{ If } \exists(W, X, Y, Z) \left| \begin{array}{l} (W \in M^*) \\ \wedge\ (X \in HAM^*HA) \\ \wedge\ (Y \in M^*) \\ \wedge\ (Z \in M^*) \\ \wedge\ (S \equiv WXYHAZ) \end{array} \right. , S \rightarrow WYXZ \tag{7}$$

$$T_3(S):\ \text{If } \neg\exists(W,X,Y,Z) \left| \begin{array}{l} (W \in M^*) \\ \wedge\ (X \in M^*) \\ \wedge\ (Y \in HAM^*HA) \\ \wedge\ (Z \in M^*) \\ \wedge\ ((S \equiv WHAXYZ) \vee (S \equiv WYXHAZ)) \end{array} \right. , S \rightarrow S \quad (8)$$

$$D_1(S):\ \text{If } \exists(X,Y,Z) \left| \begin{array}{l} (X \in M^*) \\ \wedge\ (Y \in M^*) \\ \wedge\ (Z \in M^*) \\ \wedge\ (N(Y) = 6) \end{array} \right. , S \rightarrow XYYZ \quad (9)$$

$$D_2(S):\ \text{If } \neg\exists(X,Y,Z) \mid \begin{array}{l} (X \in M^*) \\ \wedge\ (Y \in M^*) \\ \wedge\ (Z \in M^*) \\ \wedge\ (N(Y) = 6) \end{array} , S \rightarrow S \quad (10)$$

$$S \rightarrow \begin{cases} T_x(S)\ \{1-q\} \\ D_x(S)\ \{q\} \end{cases} \quad (11)$$

There is a total of three parameters that can be specified for each simulation trial: 1) the stochastic incidence q of tandem copy events ($D_x(S)$) versus cut-and-paste events ($T_x(S)$), 2) the size of tandem copy events specified by the $N(Y)$ condition in Equations 9 and 10, and 3) the number of consecutive octets representing the starting sequence (Equation 4).

2.8 Measures of Internal Physical Clustering

2.8.1 ORF Density Calculation and Randomization

Let C be the set of chromosomes where $C = \{c_1, c_2, c_3, ..., c_{164}, c_{165}\}$. Let A_i be the set of ORFs on each chromosome c_i as they are annotated from the NCBI microbial genomes database (Wheeler *et al.*, 2000). Based on the O-ORF subset definition arrived at in Chapter 4, define $R_i \subset A_i$ as the set of O-ORFs for a given chromosome c_i, and let $r_x \in R_i$. As

measured from a somewhat arbitrary zero-point on a chromosome c_i,[1] let $P(r_x)$ represent the translational start point of each O-ORF r_x. When dividing the chromosome length L into δ-sized segments, let $\Delta(a, b)$ represent the number of ORFs for which $a \leq P(r_x) < (a + b)$. $\Delta(\delta n, \delta)$ is the number of ORFs for which $\delta n \leq P(r_x) < \delta(n + 1)$. Let F be the O-ORF density series $\{\Delta(0\delta, \delta), \Delta(1\delta, \delta), \Delta(2\delta, \delta), ..., \Delta((n-1)\delta, \delta), \Delta(n\delta, \delta)\}$.

I evaluated chromosomal segmentation sizes δ ranging up to 150,000 bp. Shuffling involved rearrangement of δ-sized segments. For each segmentation size δ, I constructed 10 shuffled chromosomes for each of the 165 chromosomes to generate a set of 1,650 shuffled versions of chromosomes. Shuffling was done by randomizing the ordering of $\Delta(\delta n, \delta)$ observations to produce the χ set. Let $\chi = X_1, X_2, ..., X_{10}$ be independent, identically distributed shuffled samples from the ordered sequence of translational start point counts F. For example, if $F = \{\Delta(0\delta, \delta), \Delta(1\delta, \delta), \Delta(2\delta, \delta), \Delta(3\delta, \delta), \Delta(4\delta, \delta)\}$, then a random reassignment of order could be $X_1 = \{\Delta(3\delta, \delta), \Delta(2\delta, \delta), \Delta(4\delta, \delta), \Delta(1\delta, \delta), \Delta(0\delta, \delta)\}$.

To contrast F with χ, I used the bootstrap procedure by resampling shuffled versions of the 165 chromosomes. For both unshuffled and shuffled versions of chromosomes, I calculated various measures of internal physical ORF clustering (Section 2.8.3). My objective was to use the distribution of measures on the χ set as a basis for assessing measures of F (the unshuffled chromosome) versus any single X_i (a shuffled chromosome). A more detailed description of how bootstrap calculations were organized is in Section 2.8.4.

2.8.2 Lag k Autocorrelation

For a series of ORF densities of length N, I calculated kth neighbor product-moment autocorrelations by lagging the series by k and dividing a covariance by the product of deviations as shown in Equation 12 (Box & Jenkins, 1976). B represents either $N - k$ or N for linear or circular chromosomes respectively. Let $d_i = F_i$. With respect to circular chromosomes, a circular boundary condition applies where $d_{i+N} = d_i$. For each analyzed chromosome, a series of Pearson product moment autocorrelation r values

[1]The zero location on a chromosome was based on NCBI's data files and does not definitively correlate with any natural landmarks on the chromosome such as *oriC*. For example, annotated locations of non-zero *oriC* on chromosomes include locations 915,732 (on a sequence of 2,841,490 bp), 4,788,169 (on a sequence of 5,528,445 bp), and 3,840,051 (on a sequence of 4,599,354 bp).

$(r_1, r_2, r_3, ..., r_{B-1})$ is generated.

$$r_k = \frac{\sum_{i=1}^{B}(d_i - \bar{d})(d_{i+k} - \bar{d})}{(B-1)\sqrt{\sum(d_i - \bar{d})^2 \sum(d_{i+k} - \bar{d})^2}} \tag{12}$$

2.8.3 Scalar Residue Measures of Internal Clustering

To quantify the interdependence among consecutive values in a lag k autocorrelation series $(r_k, r_{k+1}, r_{k+2}, ...)$, I calculated a sum of squared differences as shown in Equation 13. I compared the $E(F)$ value to similarly calculated values based on shuffled versions of the ORF count series $E(X_i)$, and the deviation from the bootstrapped distribution of shuffled-based values calculated.

$$E(F) = \sum_{k=3}^{k<(N-1)} (r_k - r_{k-1})^2 \tag{13}$$

The calculated deviation was relative in that I compared the bootstrapped distribution of abs$(E(F) - E(X_i)$ to the bootstrapped distribution of abs$(E(X_i) - E(X_j))$ as described in Section 2.8.4.

I also developed an alternate measure of ORF arrangement that is similar to that described for the Angular Frequency Transform of Sandvik *et al.* (2004). This alternate measure treats the ORF count series as a pseudophase space. The trajectory angles θ_i and rotations w are measured on the pseudophase space, and frequency of occurrence evaluated. The transform of F_i to Θ_i was done through the plotting of x and y coordinates as described in Equations 14 - 16.

$$(x_i, y_i) = (F_i, F_{i+1}) \tag{14}$$

$$(x_i, x_{i+1}, x_{i+2}) = (F_i, F_{i+1}, F_{i+2}) \tag{15}$$

$$(y_i, y_{i+1}, y_{i+2}) = (F_{i+1}, F_{i+2}, F_{i+3}) \tag{16}$$

The angle θ_i and its rotational direction w is calculated with the three points (x_i, y_i),

$(x_i + 1, y_i + 1)$, $(x_i + 2, y_i + 2)$. Clockwise rotations are represented by $w = 1$. Counter-clockwise rotations are represented by $w = -1$. The rotational angle Θ_i is the product of w and θ_i; $\Theta_i = \theta_i * w$.

I assessed the heteroscedasticity of the angular change distribution of θ_i values on the pseudophase space based on shuffled versions of the ORF count series (see Section 2.8.4). I evaluated $D(F, X_i)$ as the average Kolmogorov-Smirnov (KS) statistic (Young, 1977) between the distribution of Θ_i values from F versus the Θ_i distribution from a randomly selected (with replacement) X_i. I compared $D(F, X_i)$ to measures of the KS statistic between two shuffle-based distributions, $D(X_i, X_j)$. The bootstrapped difference between the mean characteristic value of $D(F, X_i)$ and the mean characteristic value of $D(X_i, X_j)$ was an angular frequency residue $P(c_i, \delta)$ of a given chromosome c_i and segmentation size δ.

Both $P(c_i, \delta)$ and $Q(c_i, \delta)$ were evaluated for multiple O-ORF density series based on segmentation sizes ranging from 500 bp to 150,000 bp.

2.8.4 Bootstrapping

For each segment size and chromosome, the differences were resampled 10 times based on measures of the unshuffled version versus a randomly selected (with replacement) shuffled version from a set χ of 10 shuffled versions. The mean of these 10 values was computed, $v = \overline{\chi}$. The process for computing v scores was repeated 20 times to produce the values $V = \{v_1, ..., v_{20}\}$. The mean of these 20 values was computed, $b = \overline{V}$. The process of computing b scores was repeated 10 times to produce the values $B = \{b_1, ..., b_{10}\}$. The process for computing $\overline{B}$ was repeated 10 times for the pseudophase angular assay (related to the D function) and 20 times for the lag k-based assay (related to the E function of Equation 13). The means for each of the $\overline{B}$ assays were selected as the characteristic scores for the evaluated segment size and chromosome. As a random control, the characteristic scores were recalculated with a pool of ten random shuffled versions substituting (at random) for the unshuffled version for each of the mean(B) calculations. The scheme of calculating processes repeating other sub-processes led to an initial sampling iteration count of 10 and resampling iteration counts of 4,000 for $Q(c_i, \delta)$ scores (based on the E comparisons, Equation 13) and 2,000 $P(c_i, \delta)$ scores (based on $D(F, X_i) - D(X_i, X_j)$ comparisons) for each segment length δ

and chromosome c_i. The bootstrapping iterations on the averaged v values were respectively 200 and 100. Overall, I ran 2,970,000,000 calculations of these scalar measures for internal physical clustering. Multiple trials of this entire process led to characteristic scores that were generally at most ±5% different from repetitious calculations of characteristic scores for the same segment size and chromosome. The Q scores were termed as symmetry scores, and the P scores were termed as symmetrical shape scores.

2.8.5 Harmonic Symmetry of ORF Density and the Windowed Asymmetric Deviation

My goal was to 1) identify those segmentation sizes δ (500 bp, 1,000 bp, 1,500 bp, ..., 149,500 bp, 150,000 bp) most closely associated with non-shuffled series of O-ORF densities (F) and 2) characterize and compare the degrees of non-randomness attributable to measures of internal physical clustering. To accommodate the influence of neighboring segmentation sizes, a further objective was to inspect windows of $Q(c_i, \delta)$ values covering multiple segmentation sizes δ.

I first evaluated the simulated outputs $s_p \in S$ (Section 2.7) to investigate whether $Q(s_x, \delta_1)$ was related to $Q(s_x, \delta_2)$ when the difference in segmentation sizes $(\delta_1 - \delta_2)$ for computing density of "H" letters was predictive of the initial model parameter T. To evaluate how differences in Q values relate to underlying segmentation related factors of $\delta_1 - \delta_2$ and T, I applied a fast Fourier transform (FFT) to each T-based series $\{Q(s_x,1), Q(s_x,2), Q(s_x,3), ..., Q(s_x,29), Q(s_x,30)\}$. My theory for this is that insertions of predictable sizes T should produce similar values of $Q(s_x, \delta_1)$ and $Q(s_x, \delta_2)$, and the frequencies associated with the higher FFT-computed amplitudes should inversely relate to the periodicity-generating effect of a particular T value.

With the assumption that mobile elements guide the insertion of new DNA into a replicon of a restricted or non-restricted range of insertion size (comparable to a potentially heritable T-like parameter characteristic of a lineage), I sought to determine the range of segmentation sizes that captured a significant overall rise and fall of Q values. I ran windows of 51 values (25,000 kb) on series of 300 $Q(c_i, \delta)$ values where δ ranged from 500, 1,000, 1,500,..., 149,500, 150,000. For a given window start point x, the 51 values corresponded to a

δ-based series of $x, x+500, x+1,000, x+1,500, ..., x+25,000$. To filter out small-scale effects and characterize the relative amplitude of the overall rise and fall for this range, I calculated the first spectral modulus from an FFT on the series of δ-based Q values for a given c_i and x (Figure 7). The first spectral modulus is the square root of the sum of squared sine and cosine coefficients associated with a frequency value of 1, and I termed this value to be the windowed asymmetric deviation.

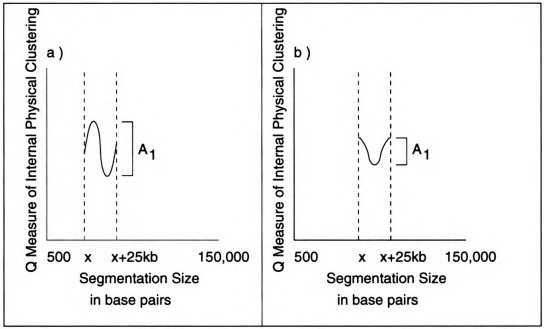

Figure 7: Illustration of the windowed asymmetric deviation measure. An amplitude A_1 is measured for a period-1 wave on a windowed series of Q values. Subfigures a and b show how the characteristic A_1 value can change based on a different window start point x. A_1 is calculated as the first spectral modulus of the FFT on a window of the 300 Q values corresponding to segmentation sizes δ ranging from 500 bp to 150,000 bp.

Chapter 3: Diversity and Stability of Chromosomal Organization and Content

3.1 Taxonomy of Chromosomal Data

3.1.1 Patchiness of Taxonomic Representation

Based on the NCBI taxonomy (Wheeler *et al.*, 2000), the scope of analysis for the 155 strains with fully sequenced genomes involves 16 phyla, 82 genera, and 126 species. These taxonomic groupings are consistent with an externally developed phylogeny (Battistuzzi *et al.*, 2004). A visual outline of this taxonomy (Fig. 8) is a key to the relative representation of various taxonomic groups. Some groups are well-represented while others are not. The two most prominent phyla in Fig. 8 are the Proteobacteria and Firmicutes, each containing over several dozen strains with fully sequenced genomes. At the class level, the Gammaproteobacteria are disproportionately well-represented, representing 24% of the 155 genomes in this study. Seven phyla have only one representative genome. These seven, sparsely represented phyla are Nanoarchaeota, Thermotogae, Aquificae, Deinococcus-Thermus, Plantomycetes, Chlorobi, and Fusobacteria.

A bias for certain types of organisms exists in the data set of 155 strains with fully sequenced genomes. In particular, by using names of species as listed in a widely-cited, historical reference on infectious diseases (Hoeprich, 1972), I calculated a significant bias for pathogenic bacteria in the set of 155 strains. 50 (32%) of the 155 strains were implicated, by their epithet, as belonging to one of the 99 pathogenic bacterial species indexed by Hoeprich (1972). Of the 52 genera I found indexed by Hoeprich (1972), 25 (48%) genera were present in the set of 155 strains. Of the 99 infectious disease species I found listed in Hoeprich (1972), 35 (35%) of these correspond to species in the set of 155 strains. Also, from the

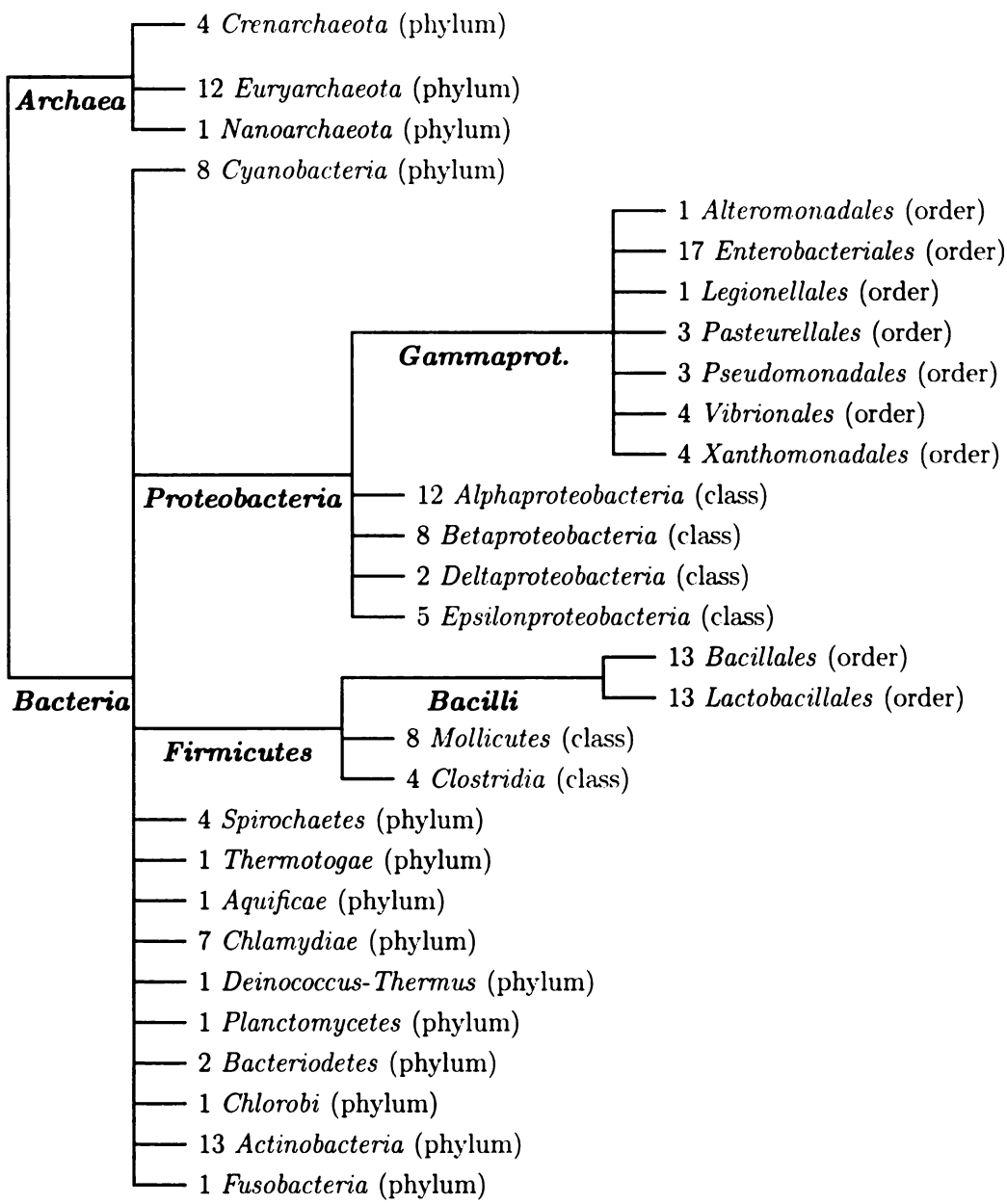

Figure 8: Taxonomic scope of 155 fully sequenced genomes. The patchiness of taxonomic branch representation for 155 genomes is shown by an outline of groupings where each branch contains less than 20 distinct genomes. The hierarchical structure and naming of taxonomic units is based on the NCBI taxonomy. Higher level taxa are labelled underneath their corresponding branch line. Gammaprot. = Gammaproteobacteria.

vantage point of making closely related strain-to-strain comparisons, the 35 infectious disease species in the set of 155 strains corresponded to 50 strains (32%). By contrast, there are 91 species in the set of 155 strains that are not present in Hoeprich (1972). These 91 species correspond to 105 of the 155 strains.

The bias of pathogenic bacteria representing approximately one third of the 155 strains is further characterized by sets of closely related strains. For the 21 sets of strains having the same species name (Fig. 13b), 19 of these sets associated with pathogenic bacteria compared to only 2 sets associated with non-pathogenic bacteria.

3.1.2 Evolutionary Times of Divergence

Fig. 9 - 12 show how the 155 genomes of this study relate to a reconstructed timescale of prokaryotic evolution based on a universal last common ancestor of 4,250 million years ago (Ma) (Battistuzzi *et al.*, 2004). Timescale reconstruction for the Archaea involved a 1,200 Ma fossil calibration (Battistuzzi *et al.*, 2004) (Fig. 9). Timescale reconstruction for the Bacteria involved a 2,300 Ma minimum geological calibration (Fig. 10 - 12). Based on these timescale reconstructions, I found that membership within the same genus corresponds to a time range of 6 Ma - 1,300 Ma.

3.1.3 Comparative Power of Data Set

The breadth and depth of the 155 genomes (165 chromosome sequences) has accumulated over time with increasing comparative power and phylogenetic coverage as shown in Fig. 13. 21 species are present for which there was more than one representative strain and corresponding genomic sequence. 48 fully sequenced genomes had at least one other closely related genome sharing the same species name. Overall, the data set of 155 genomes contains well over a dozen different sampling points for studying broad, phylum-independent patterns as well as for evaluating distinctions among strain-to-strain comparisons.

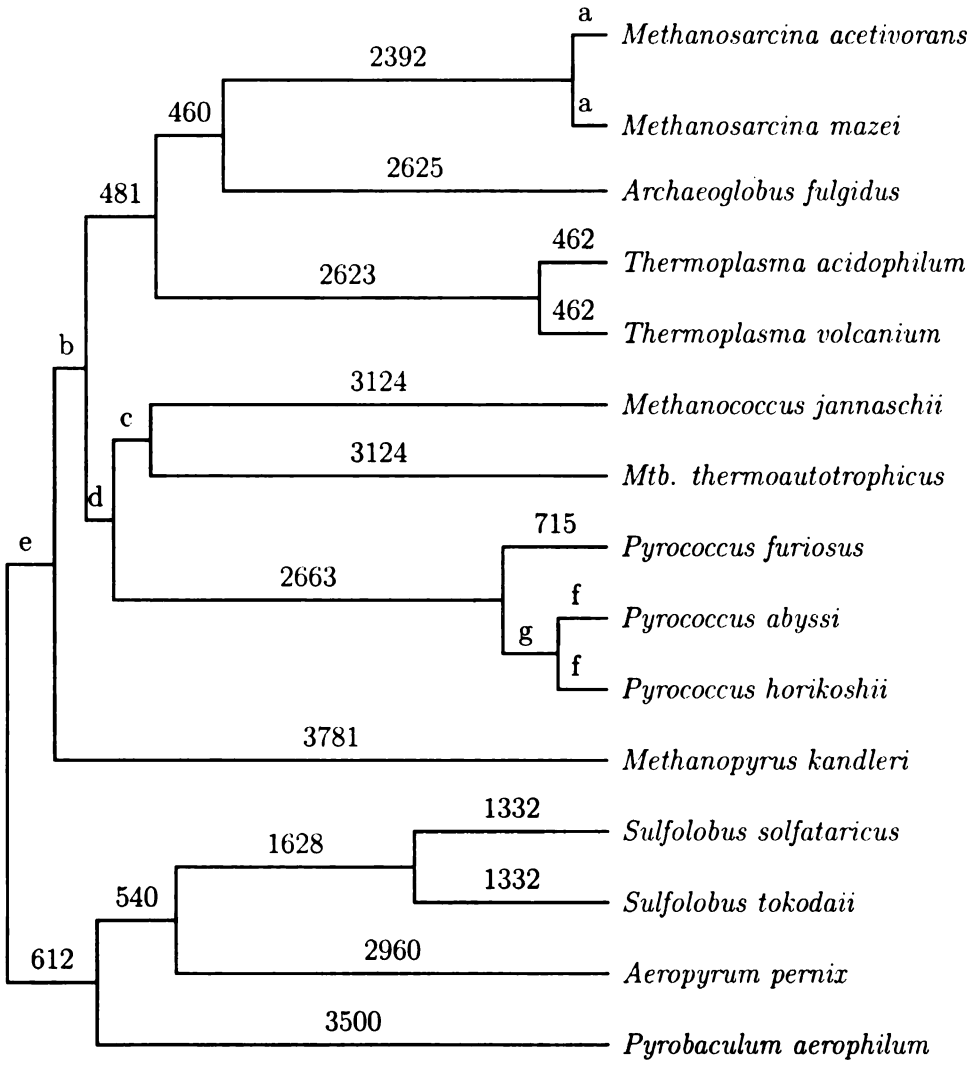

Figure 9: Times of divergence for 15 Archaea. Branch length units are in millions of years (Ma). a=233 Ma. b=215 Ma. c=254 Ma. d=188 Ma. e=323 Ma. f=338 Ma. g=377 Ma.

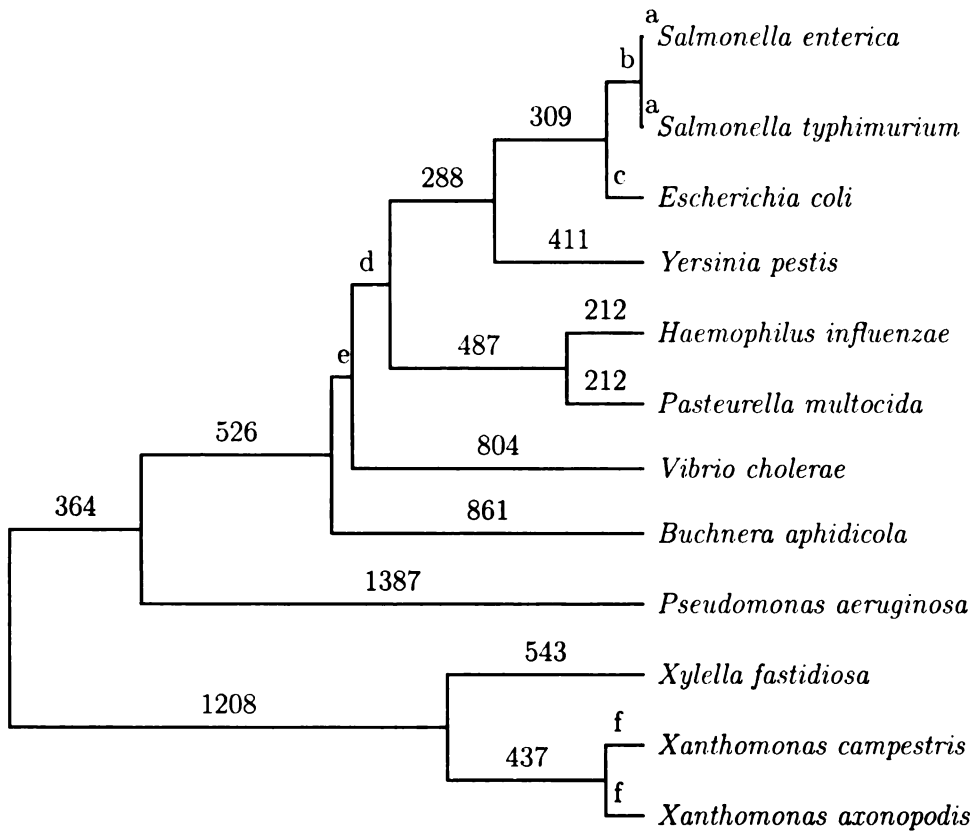

Figure 10: Times of divergence for 12 Gammaproteobacteria. Branch length units are in millions of years (Ma). a=6 Ma. b=102 Ma. c=96 Ma. d=105 Ma. e=57 Ma. f=106 Ma.

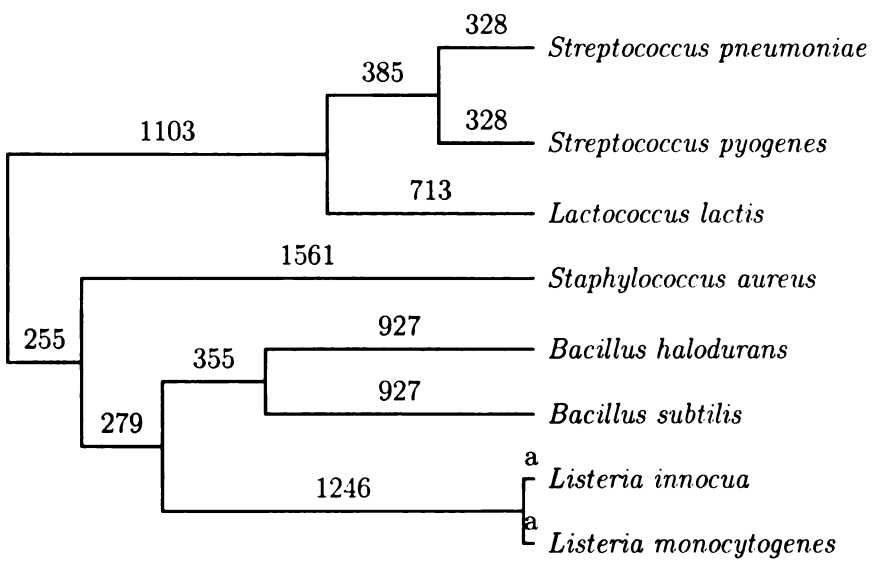

Figure 11: Times of divergence for 8 Bacilli. Branch length units are in millions of years (Ma). a=36 Ma.

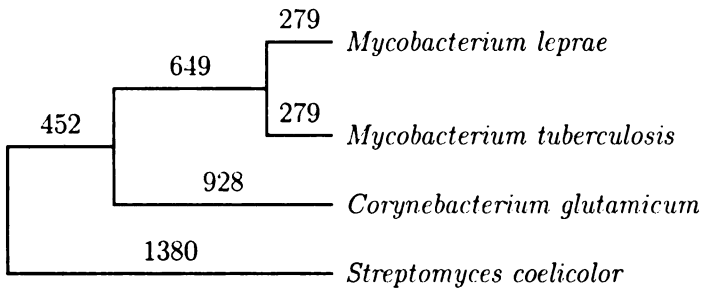

Figure 12: Times of divergence for 4 Actinobacteria. Branch length units are in millions of years (Ma).

3.1.4 Replicon Topology, Size, and Composition

The number and variety of distinct chromosomes constituting each overall genome varied between one and two. A majority (145, 94%) of the 155 genomes contained only a single distinct chromosome. Two distinct chromosomes appeared in the following 10 of the 155 genomes: (*Agrobacterium tumefaciens* C58 U. Washington and C58 Cereon; *Brucella melitensis* 16M; *Brucella suis* 1330; *Vibrio vulnificans* CMCP6 and YJ016; *Vibrio cholerae*; *Vibrio parahaemolyticus*; *Deinococcus radiodurans*; and *Leptospira interrogans*).

While the replicon topology for most of the chromosomes was circular, five of the chromosomes were linear. The genomes with linear chromosomes are *Borrelia burgdorferi* B31, *Agrobacterium tumefaciens* (2 strains, C58 U-Washington and C58 Cereon), *Streptomyces coelicolor* A3(2), and *Streptomyces avermitilis* MA-4680. The *A. tumefaciens* genomes have two topologically distinct chromosomes where one chromosome is circular and the other is linear.

The sizes of the 165 chromosomes range from 360 kb (one of the two distinct chromosomes present inside *Leptospira interrogans* serovar lai str. 56601) to 9,100 kb (*Bradyrhizobium japonicum* USDA 110). Tables 3 and 4 show genus-level and species-level variation in genome sizes based solely on DNA associated with distinct chromosomes. Even with this restricted consideration of genomic content, variation within a genus can be almost three-fold such as with fully sequenced strains of *Mycoplasma* and *Treponema*. As characterized by Tables 3 and 4, the range in median genome size differences among members of the same genus is 256 kb compared to a 52 kb difference among members of the same species. Differences among members of the same species are generally quite small in

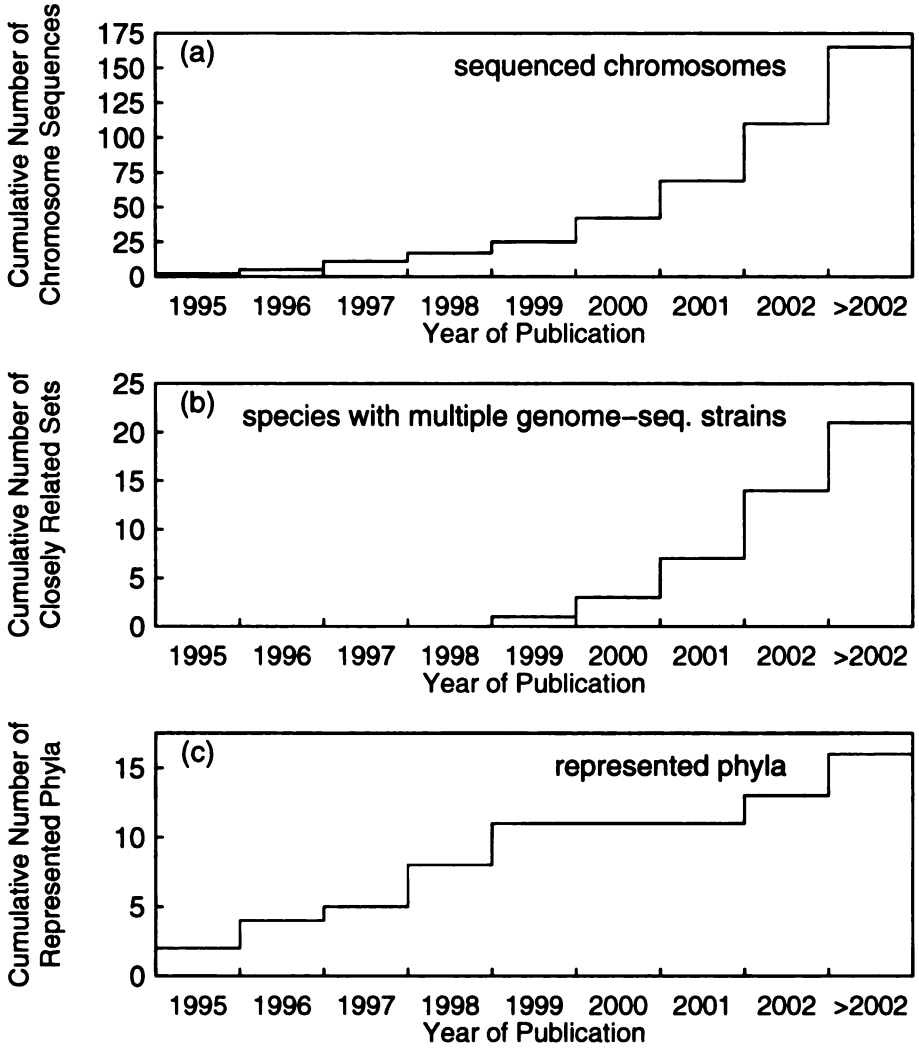

Figure 13: Comparative scope of available chromosomes and genome-sequenced over time. Annual trends showing the cumulative total of (a) number of sequenced chromosomes, (b) sets of two or more genome-sequenced strains belonging to the same species, and (c) number of phyla with one or more sequenced genome-sequenced strains. Start and stop dates are 1995/7/28 (*Haemophilus influenzae*) to 2004/3/20. Years are based on date of cited publication for the genome (or corresponding species set or phyla). When there is not a regular publication, the date of online publication (i.e., "epub") or time of initial full sequence submission to GenBank, was used.

Table 3: Conservation of total genome size for various genera.

Genus[a]	No. of species	Size Range (kb)	Genome Sizes (Mb)
Brucella	2	20	3.3, 3.3
Thermoplasma	2	20	1.6, 1.6
Chlamydia	2	30	1.0, 1.0
Listeria	2	67	2.9, 3.0
Xanthomonas	2	99	5.1, 5.2
Haemophilus	2	131	1.7, 1.8
Rickettsia	2	157	1.1, 1.3
Pyrococcus	3	170	1.7, 1.8, 1.9
Pseudomonas	3	215	6.2, 6.3, 6.4
Sulfolobus	2	297	2.7, 3.0
Streptomyces	2	358	8.7, 9.0
Mycoplasma	6	779	0.58, 0.82, 1.0, 1.0, 1.2, 1.4
Corynebacterium	3	820	2.5, 3.1, 3.3
Clostridium	3	1,141	2.8, 3.0, 3.9
Bordetella	3	1,253	4.1, 4.8, 5.3
Lactobacillus	2	1,316	2.0, 3.3
Methanosarcina	2	1,655	4.1, 5.8
Treponema	2	1,705	2.8, 1.1

[a]The listed comparisons involve strains that are of different species, but belong to the same genus.

comparison to most genus-level comparisons, except for the differences between *Prochlorococcus marinus* strains and *Escherichia coli* that each approach a one million base pair (Mb) difference in genome size.

Based on the available data for fully sequenced genomes, I found genomes to be variable in their number of corresponding plasmids. There were 69 sequenced plasmids of variable topology, and these belonged to just 30 of the 155 genomes. 51 of the plasmids are annotated as having a circular topology, and the data files for 18 other plasmids do not have an annotated topology. I found that many of the plasmids without an annotated topology were reportedly linear (Ikeda *et al.*, 2003; Casjens *et al.*, 2000; Ivanova *et al.*, 2003). As characterized by available data files, 11 of the fully sequenced genomes have just 1 plasmid, and 12 of the fully sequenced genomes have 2 plasmids. The *Yersinia pestis* CO92 genome has 3 plasmids. *Nostoc sp.* PCC 7120 has 6 plasmids. *Borrelia burgdorferi* B31 has 21 plasmids. I found the range of plasmid size to be 1,286 base pairs to 2,095,000 base pairs. The first to third quartile range of plasmid size is 25,110 to 161,600 base pairs. The median

Table 4: Conservation of total genome size for various species.

Species	No. of strains	Size Range (kb)	Genome Sizes (Mb)
A. tumefaciens	2	0.7	4.9, 4.9
Tropheryma whipplei	2	1	0.9, 0.9
Chl. pneumoniae	4	4	1.2, 1.2, 1.2, 1.2
Myco. tuberculosis	2	8	4.4, 4.4
Shigella flexneri	2	8	4.6, 4.6
S. enterica	2	17	4.8, 4.8
Helicobacter pylori	2	24	1.6, 1.7
Buchnera aphidicola	3	25	0.641, 0.641, 0.641
Streptococcus pyogenes	4	48	1.9, 1.9, 1.9, 1.9
Streptococcus agalactiae	2	51	2.2, 2.2
Yersinia pestis	2	53	4.6, 4.7
Staphylococcus aureus	3	63	2.8, 2.8, 2.9
Vibrio vulnificus	2	85	5.1, 5.2
Neisseria meningitidis	2	88	2.2, 2.3
Streptococcus pneumoniae	2	122	2.0, 2.2
Bacillus anthracis	2	134	5.1, 5.2
Xylella fastidiosa	2	160	2.5, 2.7
Bacillus cereus	2	188	5.2, 5.4
Prochlorococcus marinus	3	753	1.7, 1.8, 2.4
Escherichia coli	4	889	4.6, 5.2, 5.5, 5.5

plasmid size is 40,340 base pairs. I found instances where plasmids were not included as part of the fully sequenced genome data, such as for the three plasmids of *Yersinia pestis* KIM (Deng *et al.*, 2002).

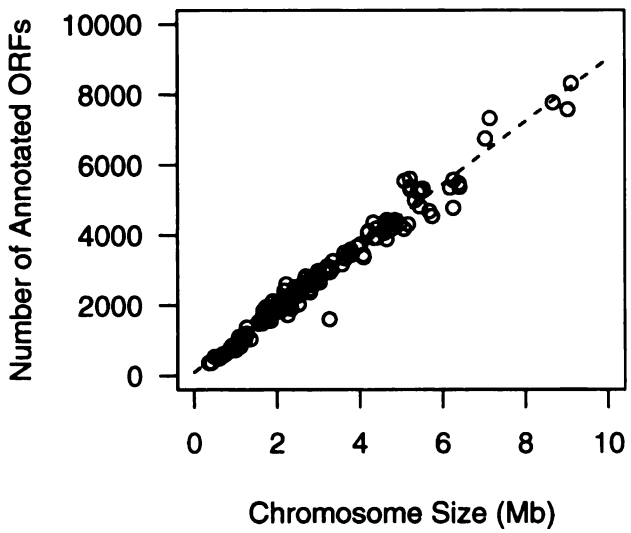

Figure 14: Number of annotated ORFs versus genome size for 155 genomes. The slope is 893 ORFs per Mb of chromosomal DNA (intercept = 94). $r^2 = 0.97$.

3.2 Structural Constraints of Chromosomal Organization and Content

3.2.1 Open Reading Frames

NCBI data files for fully sequenced genomes present a total of 447,551 annotated open reading frames (ORFs) for 155 genomes. Based on these ORF annotations, I found that a total amount of 415,890,648 base pairs (bp) of 483,773,411 bp (86.0%) encodes for amino acids from chromosomal DNA. Per organism, this ratio of total ORF content to chromosome size varied from 49.5% (*Mycobacterium leprae*) to 96.8% (*Pirellula* sp. 1) and encompassed a first-to-third quartile range of 84.1% to 89.5%. I calculated there to be, on average, a density of one ORF for every 1,086 bp for the set of 165 chromosomes. The first quartile value is one ORF for every 1,140 bp, and the third quartile value is one ORF for every 1,020 bp. The lowest density is one ORF for every 2,036 bp (*M. leprae*), and the highest density is one ORF for every 853 bp (*Pyrobaculum aerophilum* str. IM2). Fig. 14 shows a strong linear correlation of annotated ORFs versus total chromosomal content and corresponds to a density of one ORF for every 1,112 bp. I found that the annotated locations and lengths of ORFs remains relatively constant across various versions of data in the NCBI database.

I characterized ORF lengths by the number of encoded amino acids (aa) typically

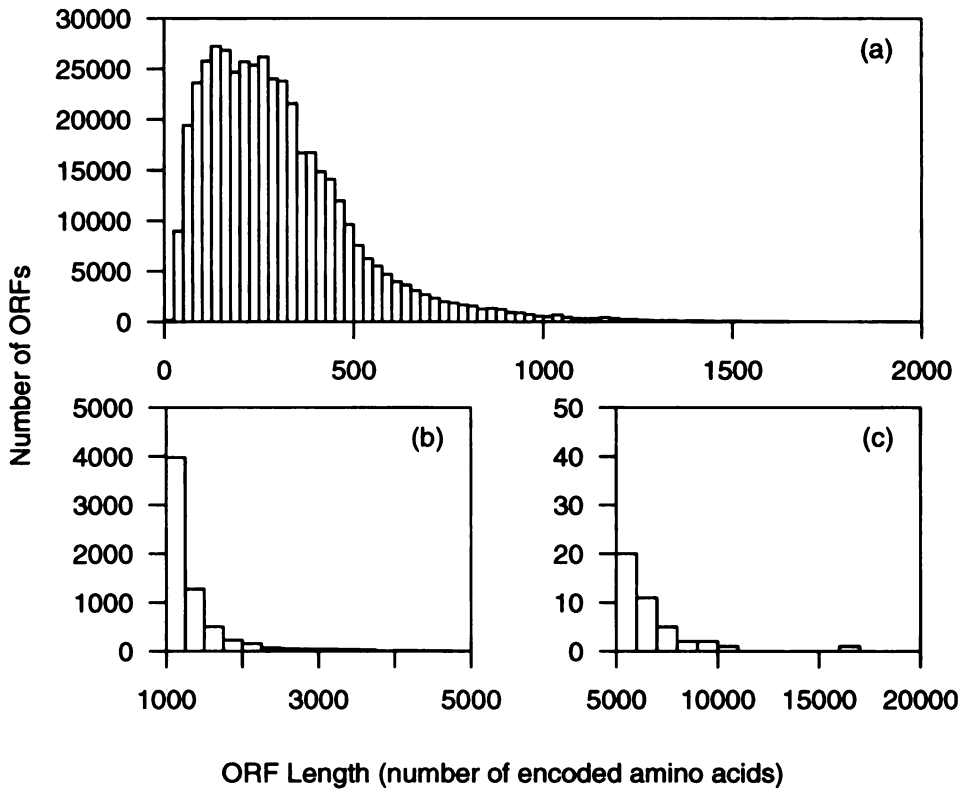

Figure 15: Distribution of 447,551 ORF lengths for 155 genomes. ORF lengths are shown as the number of encoded amino acids (aa) in each ORF. ORF length histograms are shown for three different scalings: (a) 0 to 2,000 aa, bin size = 25 aa; (b) 1,000 to 5,000 aa, bin size = 250 aa; and (c) 5,000 to 20,000 aa, bin size = 1,000 aa.

associated with the translated protein product. Fig. 15a shows an L-shaped distribution of ORF lengths. Most ORFs (> 95%) range in size from 0 to 705 aa. Only 1.4% of ORFs are greater than 1,000 bp (Fig. 15b and 15c). The average ORF length is 310 aa with a standard deviation of 237 aa. The median ORF length is 265 aa.

Fig. 16 shows the L-shaped distribution to be a robust property that occurs across various taxa. The plotted 127 aa line is an indicator of a common protein domain size of 14 kDa (Savageau, 1986). The plotted first quartile mark q_1 ranged from 235 aa to 267 aa. The first quartile mark was about twice that associated with the common protein domain size of 127 aa. These markings on the frequency peak structures of Fig. 16 visually confirm a scenario of modular protein structure where proteins are composed of one or multiple domains. Fig. 17 portrays all 5 distributions together, with a cubic-spline smoothing of each frequency distribution from Fig. 16. The smoothing function fails to produce lengthy,

monotonic regions of increases or decreases in ORF length frequency for ORF lengths > 127 aa. There does appear to be a plateau between 127 aa and 254 aa that has an internal range of variation to be at least 10% in relative frequency.

A lognormal transform of the ORF length distribution is shown in Fig. 18. The fit of Fig. 18 to a normal distribution is $p < 0.01$ based on a Shapiro-Wilk test (Royston, 1982). The skewness value (Joanes & Gill, 1998) on the lognormal transform was -0.32. Distributions with a longer than normal left side have negative skewness values.

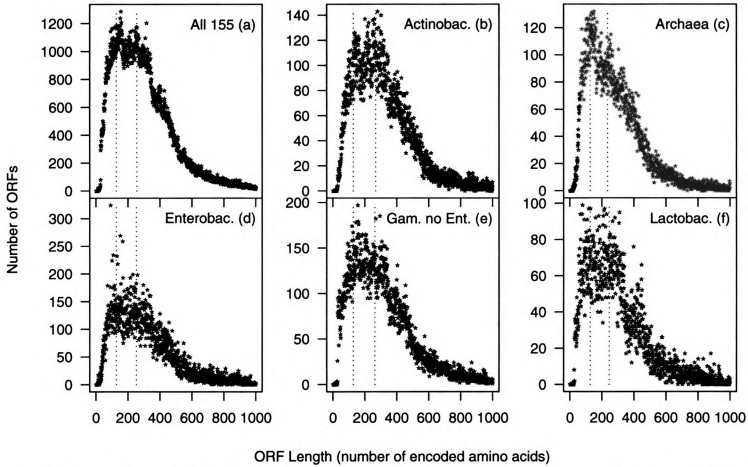

Figure 16: Subsampled distributions of ORF lengths. Two vertical dotted lines are on every plot. The histograms are for ORFs $\leq$ 1,000 aa. Bin size is 1 aa. The leftmost line corresponds to 127 amino acid residues. The rightmost line corresponds to the first quartile q_1 of cumulative summed ORF lengths (25% of the area underneath the distribution). (a) All ORFs, $n = 441,040$, $q_1 = 255$; (b) Actinobacteria, $n = 44,059$, $q_1 = 267$; (c) Archaea, $n = 37,067$, $q_1 = 235$; (d) Enterobacteriales, $n = 56,342$, $q_1 = 252$; (e) Gammaproteobacteria without Enterobacteriales, $n = 58,574$, $q_1 = 262$; (f) Lactobacillales, $n = 27,484$, $q_1 = 245$.

3.2.2 ORF Arrangement

For a pairwise comparison of conserved ORF organization, I examined bidirectional best hits of ORFs by constructing dot matrix plots with the physical coordinates of each ORF's translational start point. Fig. 19 shows the conservation of ORF arrangement among two sets of strains sharing the same species name. A high level of conservation was indicated by a generally non-interrupted line proceeding from the bottom left to the upper right of each dot matrix, and membership within the same species correlated well with this pattern. Fig. 20 shows two dot matrix comparisons between genomes that belong to the same family or genus, but do not share the same species name. There was substantially less disruption of conserved ORF organization in Fig. 19 compared to Fig. 20. The degree of conservation was quantified for various segmentation sizes δ through the use of average pointwise mutual information (APMI), a property I evaluated for δ = 10 kb, 20 kb, 30 kb, ..., 150 kb. Fig. 21a shows the average APMI for species-level comparisons and genus-level comparisons. A smaller segmentation size δ corresponded to a higher amount of measured information common to the relative ORF locations from each pairwise comparison. Higher values of APMI indicate a greater degree of information between the paired genomes compared to lower values of APMI. Mutual information is generally interpreted in units of bits, and APMI values can be meaningfully compared across different chromosomes of different lengths and also for different segmentation sizes δ. Visually, it appears that Fig. 19a and 20a each respectively outperform Fig. 19b and 20b, and the APMI measures are consistent with this. For Fig. 20, the APMI based on δ = 40 kb has a 28% reduction in value compared to an 8% reduction in value for APMI based on δ = 10 kb. I generally found the 40 kb-based APMI values to have a proportionately greater decline in value compared to the 10 kb-based APMI values across various phylogenetic comparisons.

To investigate pairwise comparisons of genomic organization with an estimated time of divergence from a last common ancestor, I evaluated conservation for the comparisons listed in Table 5, and used the times of divergence characterized in Fig. 25 - 33. Fig. 21b shows the average APMI $\overline{I_a}$ for comparisons from column 1 of Table 5 and the averaged APMI $\overline{I_d} = \frac{\overline{I_b} + \overline{I_c}}{2}$ for comparisons from columns 2 and 3 of Table 5.

Table 6 lists the the correlations and slopes of how various δ-based APMI values relate to

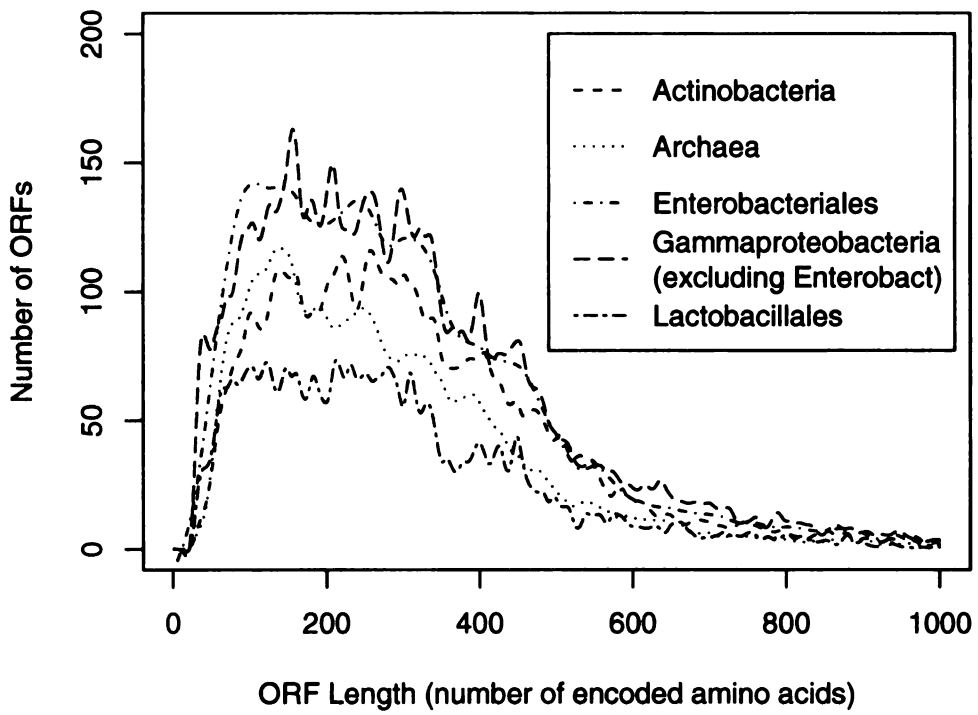

Figure 17: ORF length frequency distributions among 5 taxonomic subgroupings. The series of frequency values for a bin size of 1 aa is smoothed with a cubic spline. Only those ORFs $\leq$ 1,000 encoded amino acids in length are shown. Enterobact = Enterobacteriales.

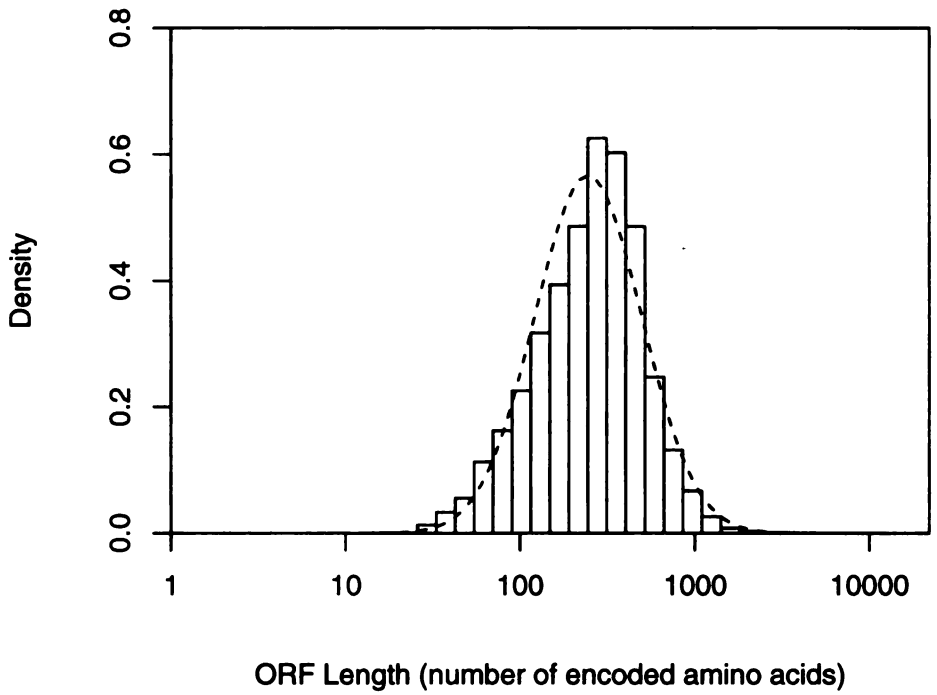

Figure 18: Log-transformed distribution of ORF lengths for 155 genomes. A natural logarithm was calculated for each of the ORF lengths, and the log-transformed values of ORF length were aggregated into a histogram with bin sizes of 0.25. The x-axis is labelled with a transformed power of 10 scaling. A fitted bell curve with a mean at 245 aa is shown with a dotted line.

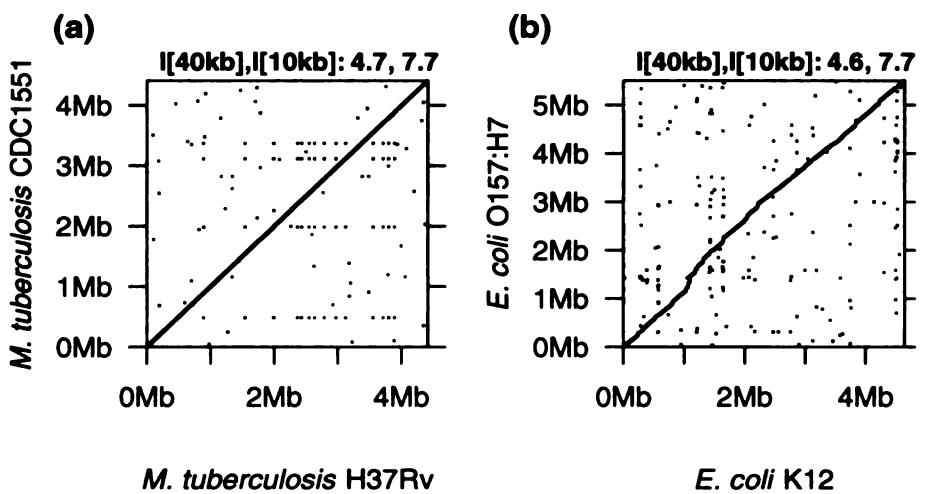

Figure 19: Comparison of ORF organization between two *Mycobacterium tuberculosis* strains and two *Escherichia coli* strains. **(a)** *M. tuberculosis* CDC1551 versus H37R. **(b)** *E. coli* K-12 versus O157:H7. The APMI values for segmentation sizes δ of 40 kb and 10 kb are indicated at the top of each dot matrix.

Table 5: Diverging set of phylogenetic comparisons.[a]

Pairwise Comparisons	Closest Pair	Closest & Distant Ancestor	Closest & Distant Ancestor
A	Py.aby+Py.hor.	Py.aby.+Py.fur.	Py.hor.+Py.fur.
B	Mt.acet.+Mt.maz.	Mt.acet.+Arch.ful.	Mt.maz.+Arch.ful.
C	Sulf.solf.+Sulf.tok.	Sulf.solf.+A.pnx.	Sulf.tok.+A.pnx.
D	tb1551+tbH37Rv	tb1551+Cor.glut.	tbH37Rv+Cor.glut.
E	Nostoc+Synec.	Nostoc+Th.elon.	Synec.+Th.elon.
F	S.pyog.+Str.pnm.	S.pyog.+Lac.lact.	Str.pnm.+Lac.lact.
G	M.gen.+M.pnm.	M.gen.+M.pulm.	M.pnm.+M.pulm.
H	R.pro.+R.con.	R.pro.+Caul.cre.	R.con.+Caul.cre.
I	Xn.cmp.+Xn.axon.	Xn.cmp.+X.fas.	Xn.axon. + X.fas.

[a]The closest pair column is a comparison between the two most closely related strains relative to comparisons involving a more distant last common ancestor (two rightmost columns). The abbreviations used are defined in Table 2.

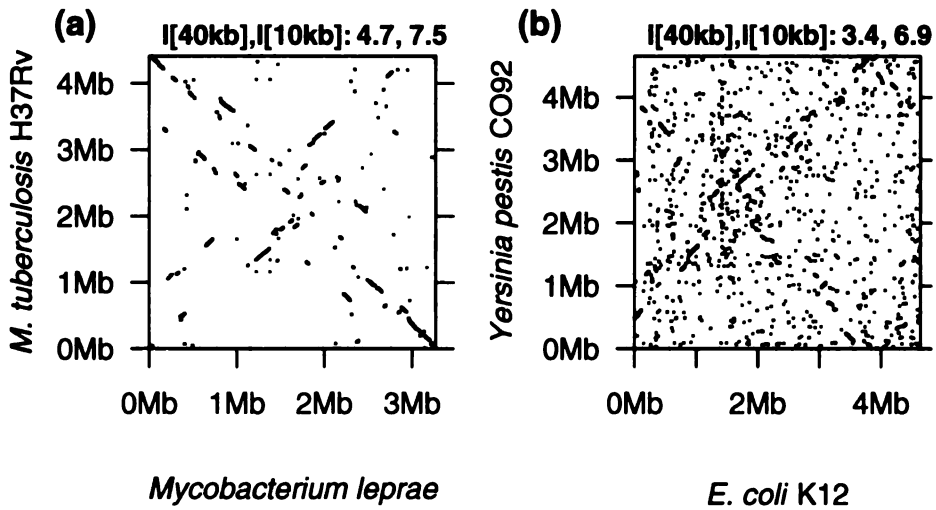

Figure 20: Comparisons of ORF organization for two *Mycobacterium* species and two species from the Enterobacteriales. **(a)** *M. tuberculosis* H37R and *M. leprae*. **(b)** *Escherichia coli* K-12 and *Yersinia pestis* CO92. The APMI values for segmentation size δ of 40 kb and 10 kb are indicated at the top of each dot matrix.

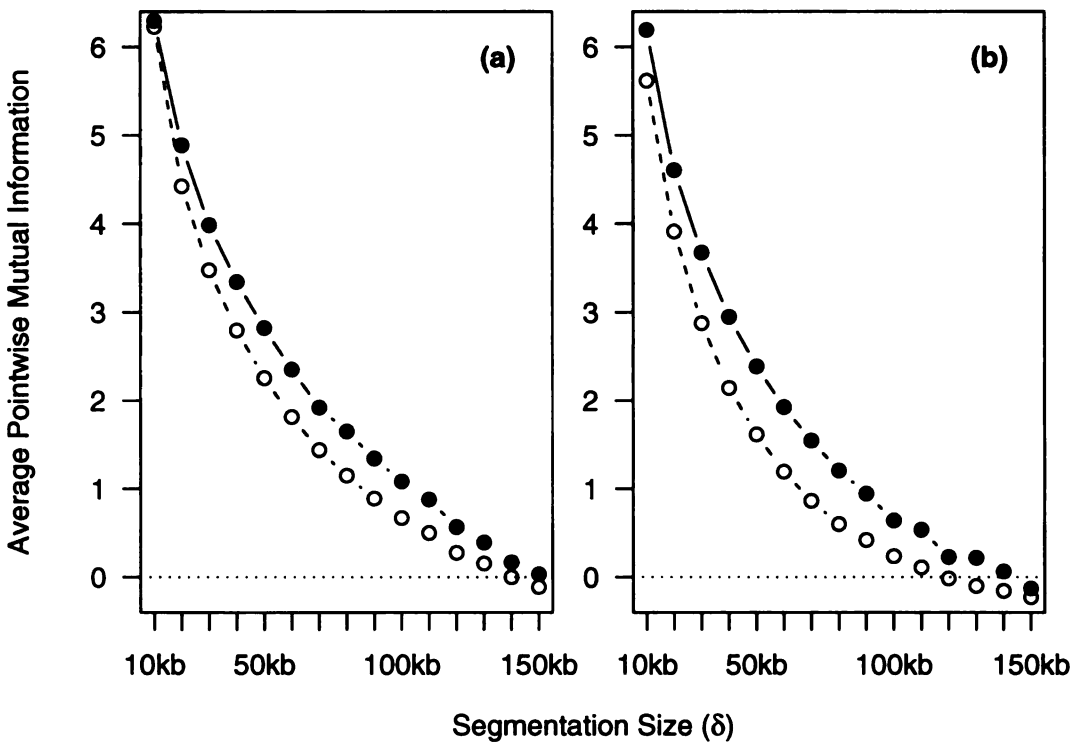

Figure 21: APMI values on dot matrix plots for various taxonomy-based comparisons of chromosomes. **(a)** The values of averaged APMI species-level comparisons are shown by solid lines and closed circles. The values of averaged APMI genus-level comparisons are shown by dashed lines and open circles. **(b)** Averaged APMI values $\overline{I_a}$ (solid lines and closed circles) and $\overline{I_d}$ (dashed lines and open circles) are shown. I_a is the average APMI between the two more closely related strains listed in the first column of Table 5. $\overline{I_d}$ is the average of the averaged APMI values, $\overline{I_b}$ and $\overline{I_c}$, corresponding to those comparisons listed in the second and third column of Table 5. The horizontal dotted lines indicate the zero-point, at or beneath which information is independent or disassociated. The APMI values are calculated for segmentation sizes δ of 10 kb, 20 kb, 30 kb, ..., 150 kb.

Table 6: Correlation of APMI with time of divergence.[a]

δ	m	r
10 kb	-0.421	0.434
20 kb	-0.479	0.472
30 kb	-0.517	0.508
40 kb	-0.509	0.512
50 kb	-0.477	0.490
60 kb	-0.456	0.494
70 kb	-0.398	0.459
80 kb	-0.350	0.438
90 kb	-0.297	0.401
100 kb	-0.216	0.318
110 kb	-0.222	0.342
120 kb	-0.102	0.192
130 kb	-0.141	0.241
140 kb	-0.077	0.145
150 kb	-0.009	0.000

[a]Average pointwise mutual information values are calculated for various segmentation sizes δ from the pairwise comparisons described in Table 5 and associated times of divergence. The slope m and correlation coefficient r are shown for a linear relationship.

an estimated time of divergence. In particular, for the highest correlating segmentation size, δ = 40 kb, Fig. 22 shows how APMI pairwise comparisons from Table 5 relate to estimated times of divergence. For all δ, the r values are very weak or statistically insignificant. The best performing range of δ values in terms of $r > 0.4$ appears to be for 10 kb to 90 kb. For δ = 40 kb, the comparisons among the Archaea (A, B, and C) have the slopes $m_A = 0.77$, $m_B = 0.35$ and $m_C = 0.21$, and the comparisons among the Bacteria (D, E, F, G, H, and I) have the slopes $m_D = 1.8$, $m_E = 1.2$, $m_F = 0.11$, $m_G = 1.4$, $m_H = 0.58$, and $m_I = 1.5$.

The $I_a - I_d$ difference between averaged APMI values for each of the nine sets of comparisons is shown in Fig. 23a and 23b. The highest differences are seen for the D, G, and H series that reach their highest respective values at δ = 30 kb, δ = 40 kb, and δ = 60 kb. The average expectation for these lineage-based comparisons is shown in Fig. 23c. The differences between Fig. 23c and generalized species-versus-genus comparisons (Fig. 23d) are shown in Fig. 24, and have the highest values for segmentation sizes δ of 10 kb and 30 kb.

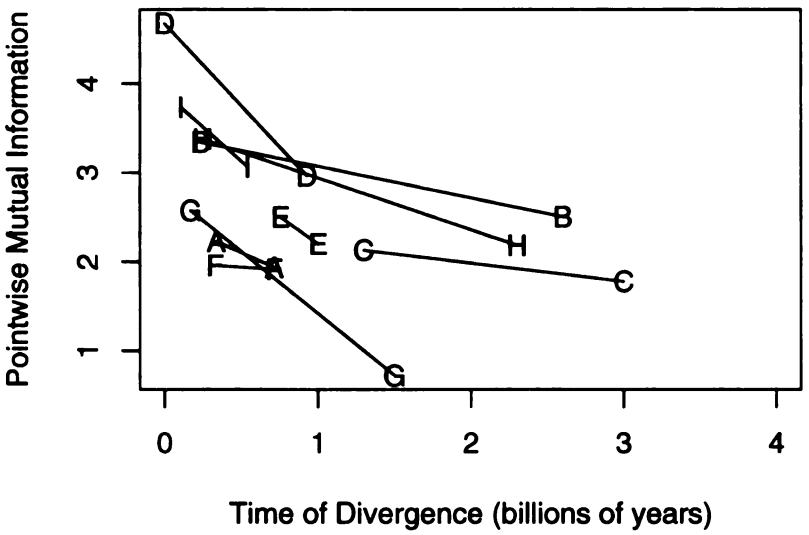

Figure 22: APMI values on dot matrix plots for various times of divergence. The letters represent pairwise comparisons described in Table 5. For each of the 9 letter pairs (18 plotted points), the leftmost letter is the APMI value for the corresponding pairwise comparison from column 1 in Table 5. The rightmost letter is the APMI value for the average of pairwise comparisons from columns 2 and 3. APMI values are calculated for a segmentation size of $\delta =$ 40 kb. For a fitted line with a slope of 0.51 Ga^{-1}, the r correlation coefficient is 0.51.

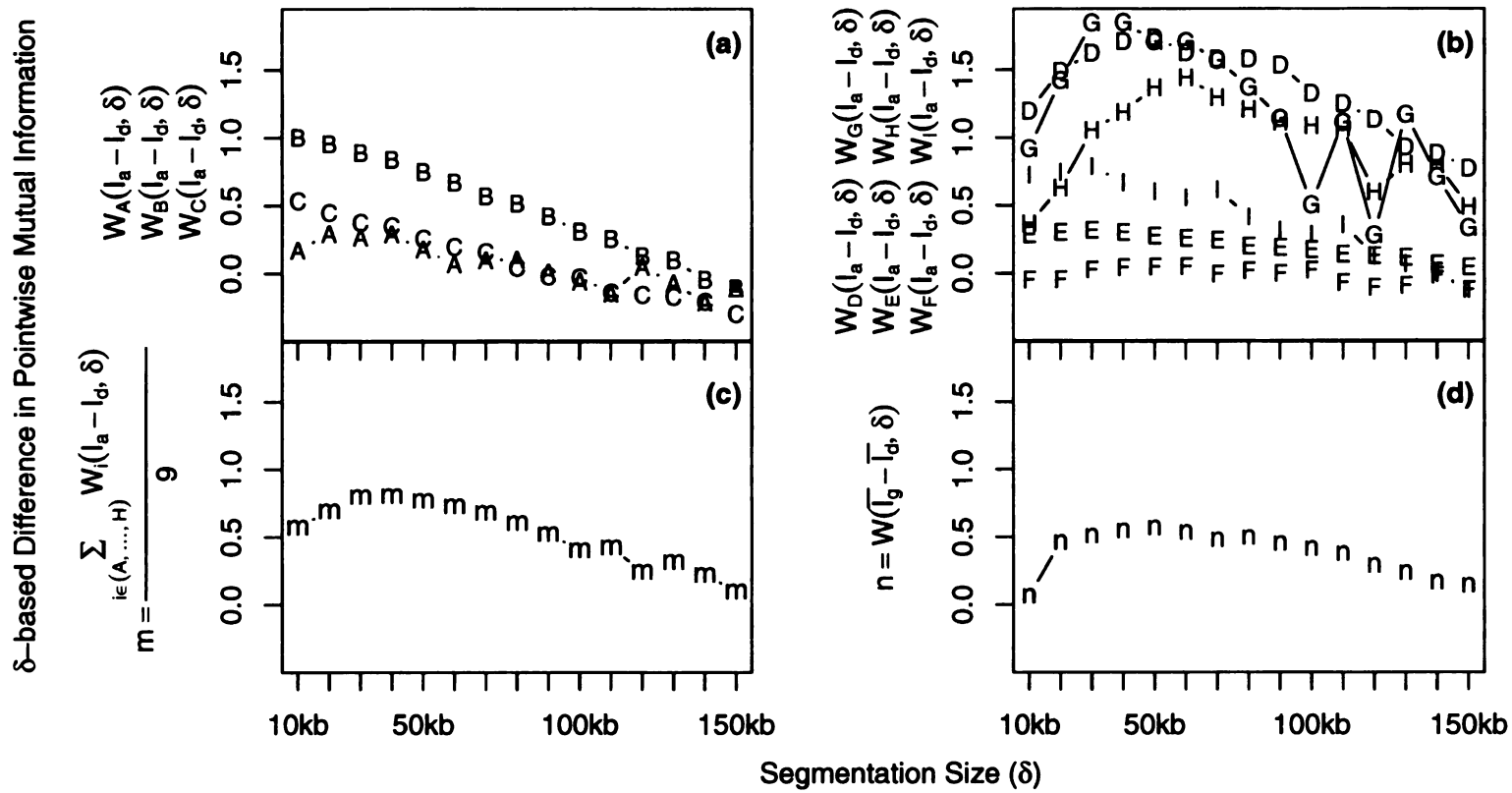

Figure 23: Taxonomy-based differences in average pointwise mutual information (APMI) for various segmentation sizes δ. Each letter, A-I, represents a pairwise comparison described in Table 5. I_a is the APMI between the two more closely related strains listed in the first column of Table 5. I_d is the averaged value of the APMIs, I_b and I_c, listed in the second and third column of Table 5. **(a)** Archaea. **(b)** Bacteria. **(c)** Average difference in pointwise mutual information for all assayed Archaea and Bacteria. **(d)** Difference between APMI for a set of species-level comparisons, $\overline{I_s}$ (Table 4), and APMI for a set of genus-level comparisons, $\overline{I_g}$ (Table 3).

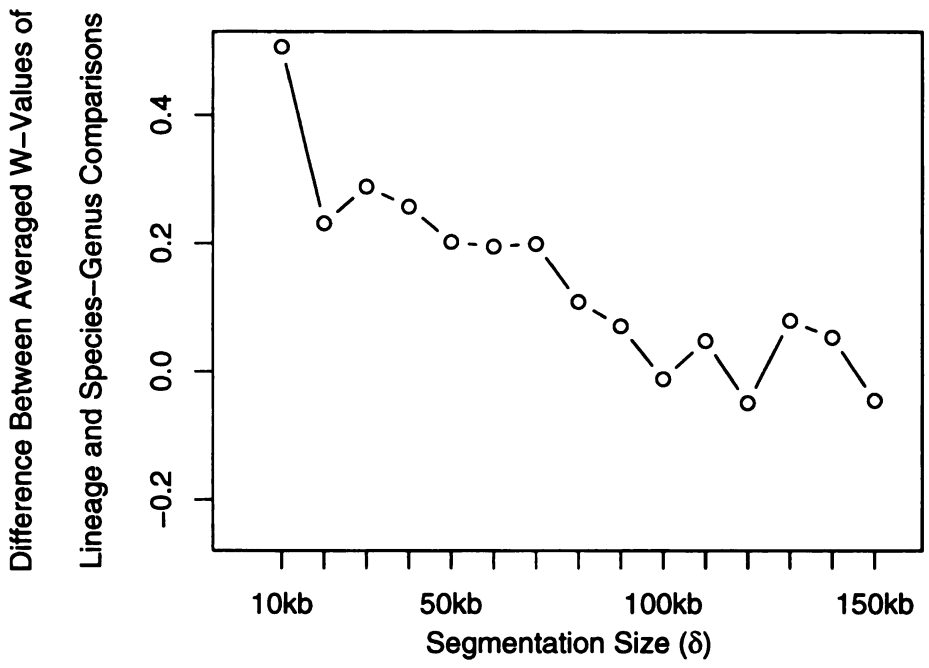

Figure 24: Difference between averaged W-values of lineage and species-genus comparisons. The ordinate value represents, for a given δ, the difference between m and n values from the Fig. 23c and 23d respectively.

To more closely inspect the relationship of estimated times of divergence from a common ancestor to the loss of conserved genomic organization, I visually assayed 9 dot matrix comparisons among a total of 3 Archaea (Fig. 25-27). I further assayed 18 dot matrix comparisons among a total of 18 Bacteria (Fig. 28-33). Overall, each figure (Fig. 25 - 33) contrasts the similarity of genomic organization of two most closely related genomes to the similarity seen with a third, more distant "cousin." Shown in Fig. 25 is a set of three paired comparisons among *Pyrococcus furiosus*, *Pyrococcus abyssi*, and *Pyrococcus horikoshii*. This comparative set of *Pyrococcus* species presents a case of how the loss of ORF organization may be directly related to longer times of estimated divergence. The two most closely related strains, *P. abyssi* and *P. horikoshii*, appear to have longer regions of successively matching ORFs than comparisons with *P. furiosus*.

I found that the Archaea appear visually to relate the time of divergence to the loss of conserved genomic organization (Table 7). The ordering based on times of divergence was fully consistent with my visual orderings of the observed loss of conserved genomic organization, prior to consideration of APMI values. The APMI values for δ = 40 kb (I_{40})

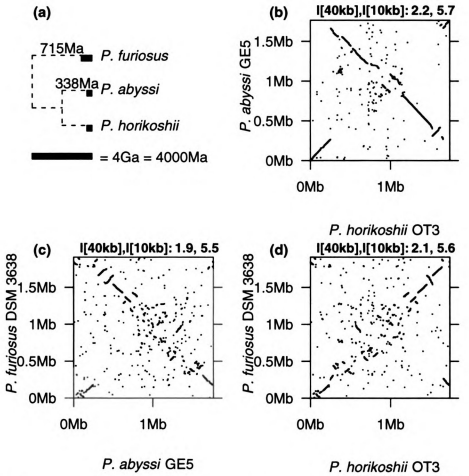

Figure 25: Comparisons of conserved ORF organization among three *Pyrococcus* strains. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

and $\delta = 10$ kb (I_{10}) did not however perfectly equate to the times of divergence. Inconsistencies of APMI values in a divergence time-based ranking generally involved Fig. 25c-d and 26c-d. The sets of comparisons involving genomes of strains that diverged > 2.5 Ga (Fig. 26c-d, and 27c-d) are essentially negligible in terms of any observable conservation. For times of divergence < 1 Ga (Fig. 25, 26, and 27) conserved regions are present across the chromosomes, and there are regions that are at least 100 kb in length for each pairwise comparison.

In addition to the three subtrees of Archaea, I evaluated six subtrees of the Bacteria for conserved patterns of ORF organization and these are shown in Fig. 28 - 33. The phyla that

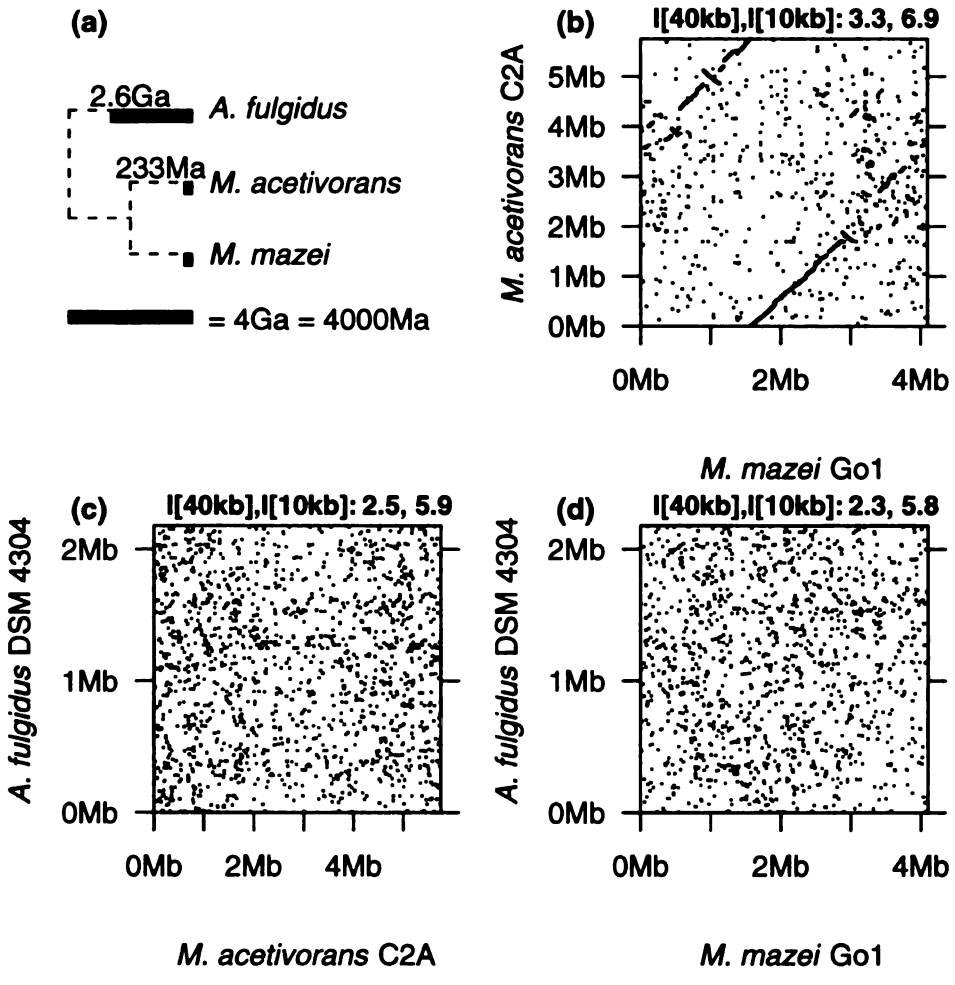

Figure 26: Comparisons of conserved ORF organization among two *Methanosarcina* strains with *A. fulgidus*. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

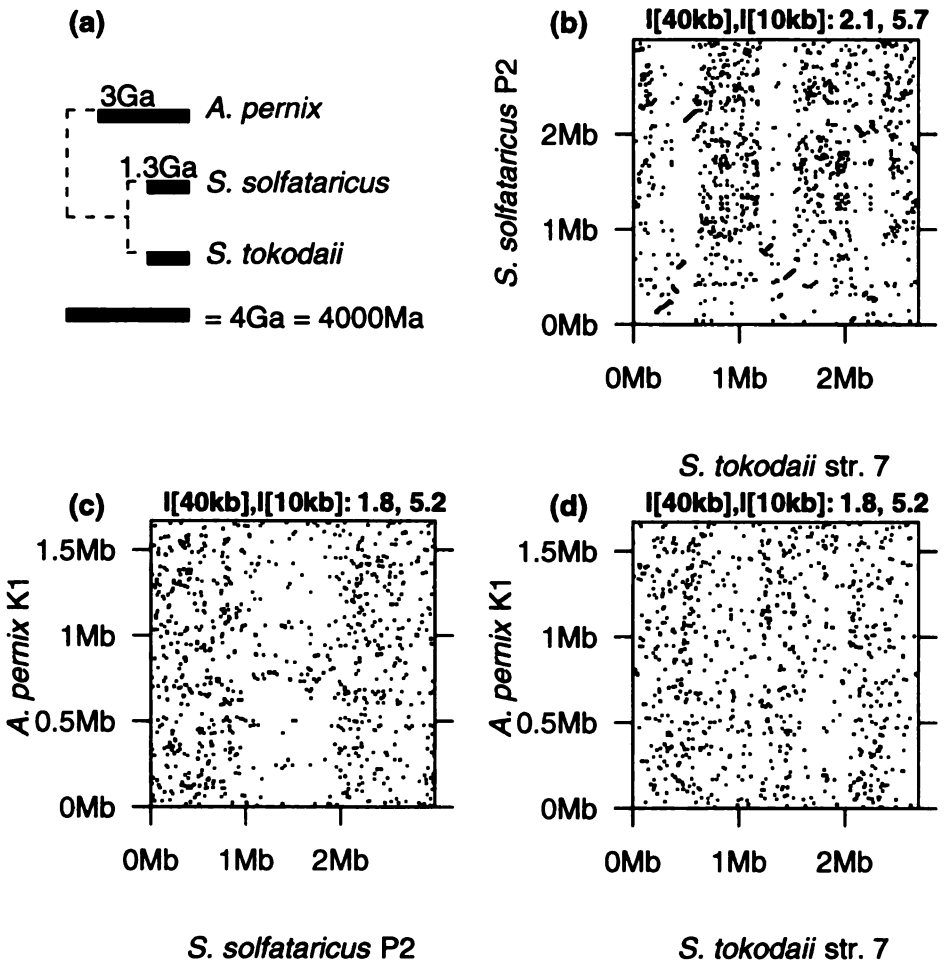

Figure 27: Comparison of three Crenarchaeota strains. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

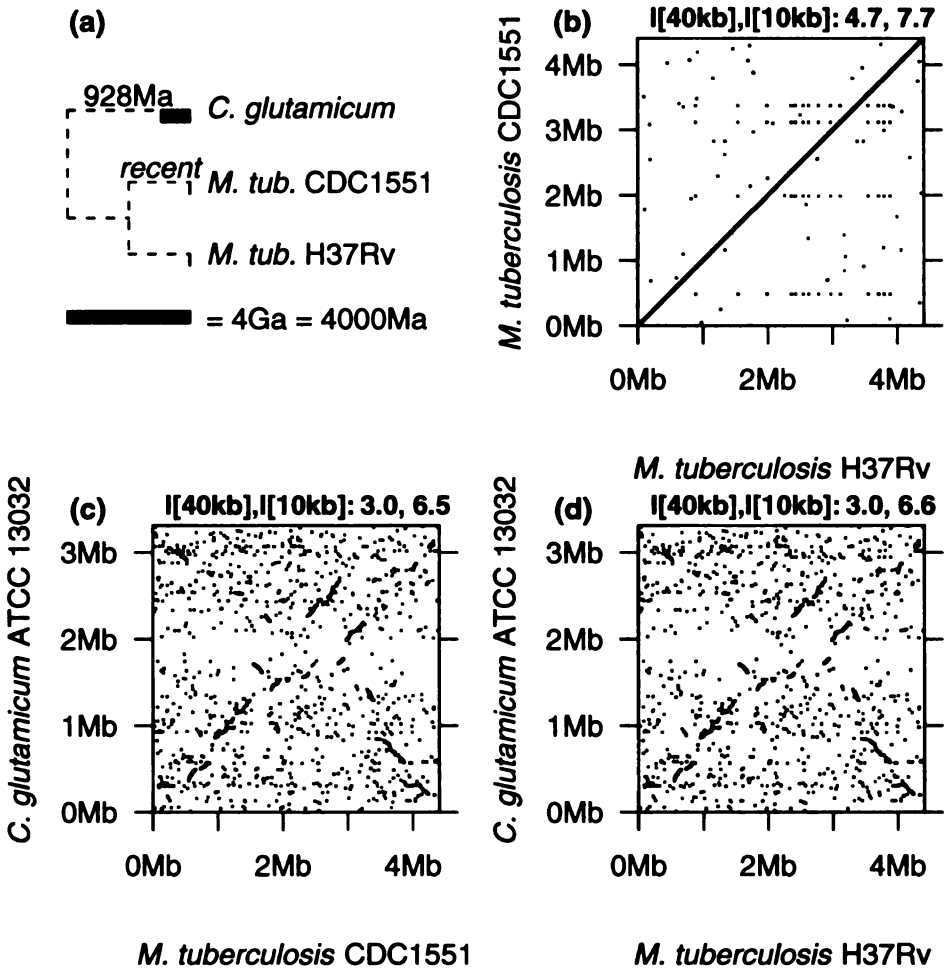

Figure 28: Comparisons of conserved ORF organization among two *Mycobacterium tuberculosis* strains and *Corynebacterium glutamicum*. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

are represented by these bacterial subtrees are the Actinobacteria, the Cyanobacteria, the Firmicutes, and the Proteobacteria.

Bacteria show various visual trends of synteny that do not necessarily correspond with estimated times of divergence as seen with the analyses involving *Xylella fastidiosa* and *Streptococci*. Overall however, there appears to be a general trend of inverse correspondence where a greater time of divergence corresponds to a lower amount of conserved ORF arrangement. For times of divergence (ToD) much greater than 1 Ga, synteny appears to be essentially lost.

My visual groupings of the dot matrix comparisons for a ranking of conserved ORF

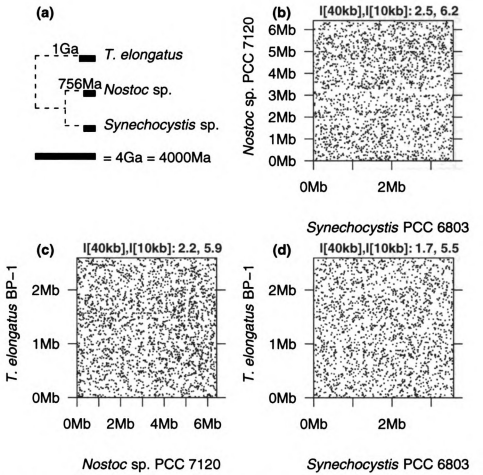

Figure 29: Comparisons of conserved ORF organization among *Nostoc* sp. PCC 7120, *Synechocystis* sp. PCC 6803, and *Thermosynechococcus elongatus* BP-1. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

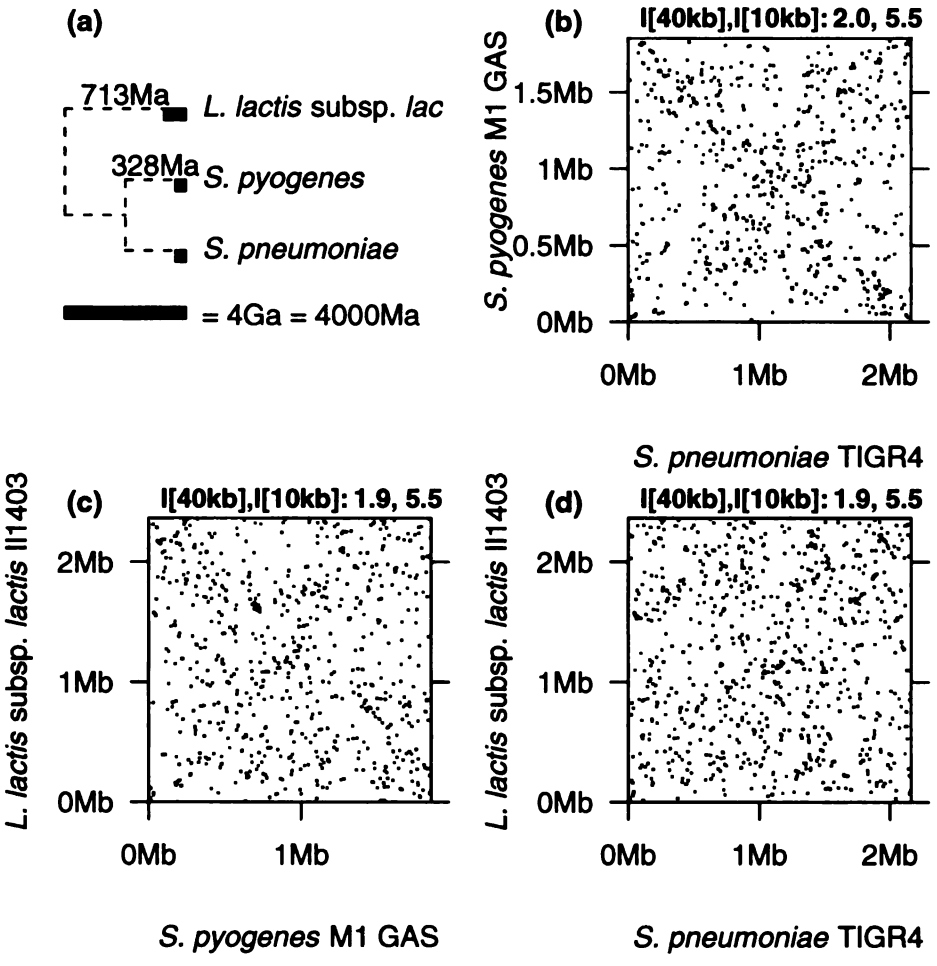

Figure 30: Comparisons of conserved ORF organization among *S. pyogenes*, *S. pneumoniae*, and *L. lactis*. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

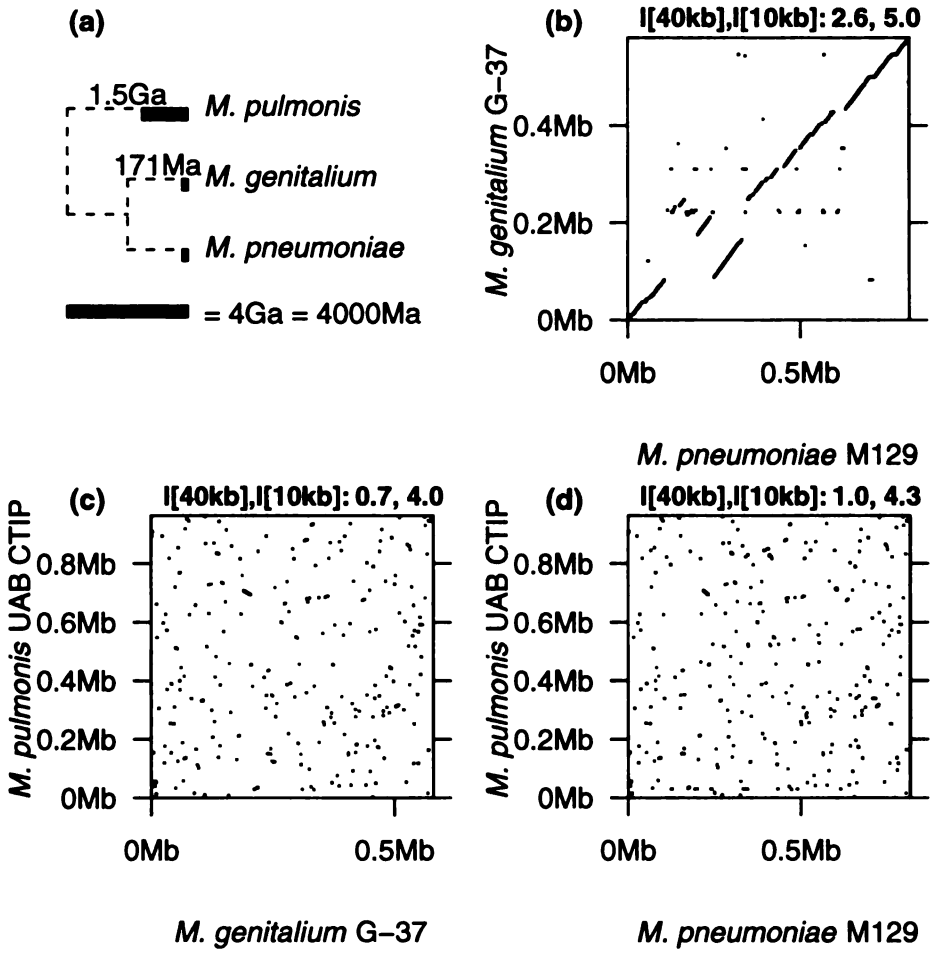

Figure 31: Comparisons of conserved ORF organization among three *Mycoplasma* strains. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

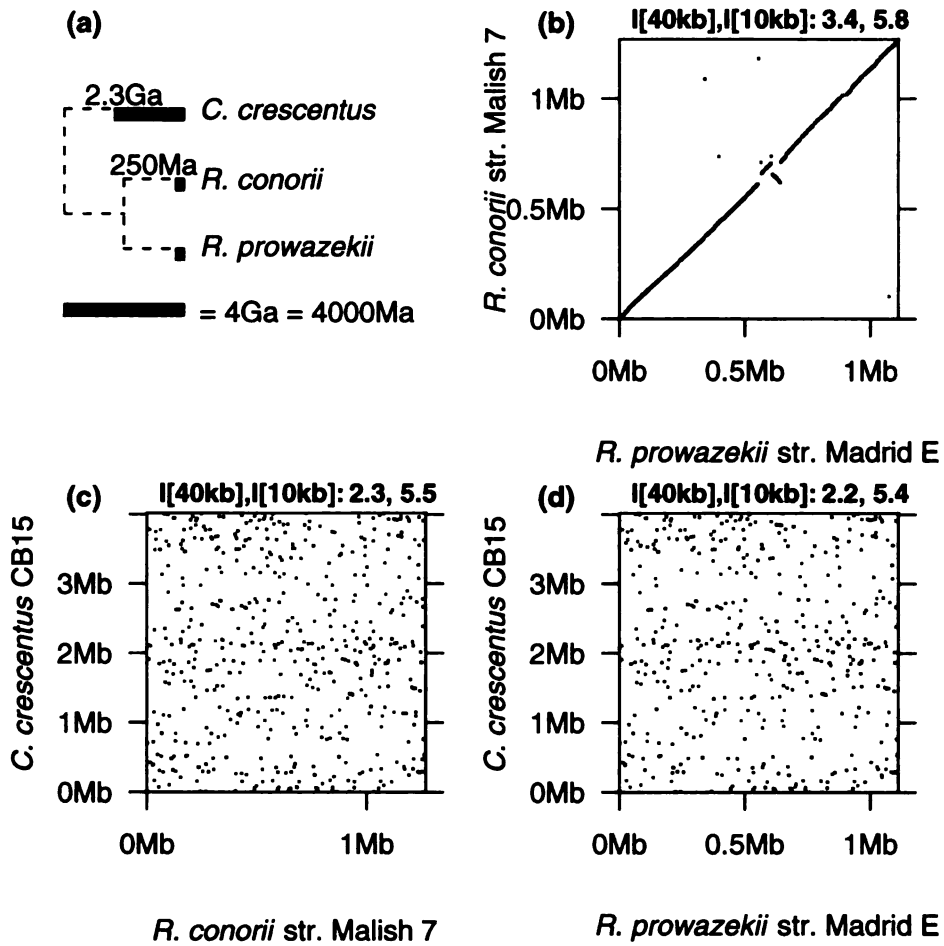

Figure 32: Comparisons of conserved ORF organization among two *Rickettsia* strains and *Caulobacter crescentus*. **(a)** times of divergence from a last common ancestor. **(b-d)** relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

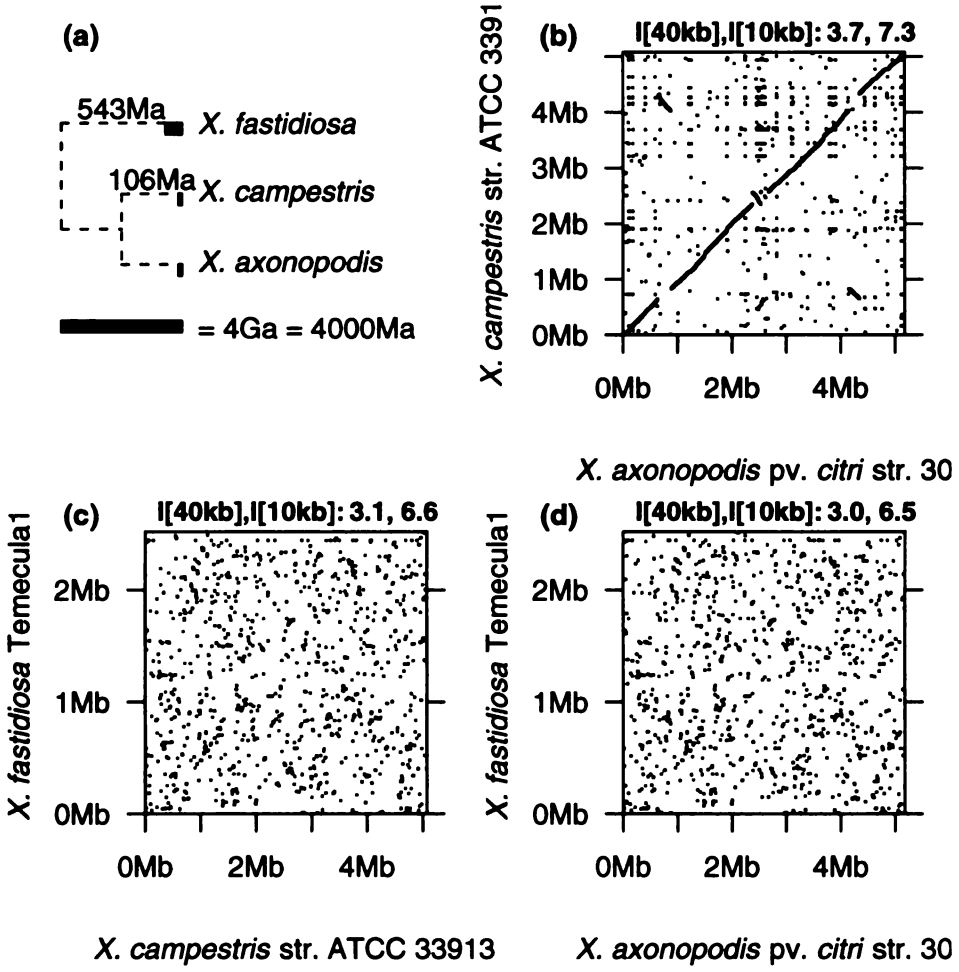

Figure 33: Comparisons of conserved ORF organization among two *Xanthomonas* strains and *Xylella fastidiosa* Temecula1. (a) times of divergence from a last common ancestor. (b-d) relative locations of bidirectional best hits between ORFs are plotted based on pairwise comparisons among three chromosomes. The averaged pointwise mutual information for each dot plot is calculated and shown for segmentation sizes 40 kb and 10 kb.

Table 7: Relationship of time of divergence to quantitative indicators of conserved ORF organization for pairwise comparisons of archaeal chromosomes.[a]

Comparison	Fig.	ToD	I_{40}	I_{10}
Mt.acet. vs Mt.maz.	26b	223 Ma	3.3	6.9
Py.aby. vs Py.hor.	25b	338 Ma	2.2	5.7
Py.aby. vs Py.fur.	25c	715 Ma	1.9	5.5
Py.hor. vs Py.fur.	25d	715 Ma	2.1	5.6
Sulf.solf. vs Sulf.tok.	27b	1.3 Ga	2.1	5.7
Mt.acet. vs Arch.ful.	26c	2.6 Ga	2.5	5.9
Mt.maz. vs Arch.ful.	26d	2.6 Ga	2.3	5.8
Sulf.solf. vs A.pnx.	27c	3.0 Ga	1.8	5.2
Sulf.tok. vs A.pnx	27d	3.0 Ga	1.8	5.2

[a]APMI values I_{40} and I_{10} are calculated for $\delta = 40$ kb and $\delta = 10$ kb for the pairwise comparisons in Fig. 28-33. ToD is the time of divergence from a last common ancestor. When multiple figures are listed in the same row, the averaged I_{40} and I_{10} values are presented. The abbreviations used for organism names are defined in Table 2.

Table 8: Relationship of time of divergence to visual and quantitative indicators of conserved ORF organization for pairwise comparisons of bacterial chromosomes.[a]

Comparison(s)	Fig.	ToD	V.R.	I_{40}	I_{10}
tb1551 vs tbH37Rv	28b	recent	1	4.7	7.7
M.gen. vs M.pnm.	31b	171 Ma	1	2.6	5.0
R.pro. vs R.con.	32b	250 Ma	1	3.4	5.8
Xn.cmp. vs Xn.axon.	33b	106 Ma	2	3.7	7.3
S.pyog. vs Str.pnm.	30b	328 Ma	4	2.0	5.5
(Xn.cmp.& Xn.axon.) vs X.fas.	33c-d	543 Ma	4	3.1	6.6
(S.pyog.&Str.pnm.) vs Lac.lact.	30c-d	713 Ma	4	1.9	5.5
Nostoc vs Synec.	29b	756 Ma	4	2.5	6.2
(tb1551&tbH37Rv) vs Cor.glut.	28c-d	928 Ma	3	3.0	6.6
(Nostoc&Synec) vs Th.elon.	29c-d	1 Ga	4	2.0	5.7
(M.gen.&M.pnm.) vs M.pulm.	31c-d	1.5 Ga	4	0.9	4.2
(R.pro.&R.con.) vs Caul.cre.	32c-d	2.3 Ga	4	2.3	5.4

[a]APMI values I_{40} and I_{10} are calculated for δ = 40 kb and δ = 10 kb for the pairwise comparisons in Fig. 28-33. V.R. is the visual ranking (1: "strong diagonal"; 2: "diagonal plus scattering"; 3: "vestigial diagonal"; and 4: "noise"). ToD is the time of divergence from a last common ancestor. When multiple figures are listed in the same row, the averaged I_{40} and I_{10} values are presented. The abbreviations used for organism names are defined in Table 2.

organization among bacteria as listed with times of divergence from a last common ancestor (ToD) are: #1, "strong diagonal", ToD are recent to 250 Ma – Fig. 28b, 31b, 32b; #2, "diagonal plus scattering", ToD is 106 Ma – Fig. 33; #3, "vestigial diagonal", 928 Ma – Fig. 28c-d; and #4, "noise", ToD are 328 Ma to 2,300 Ma – Fig. 29b-d, 30b-d, 31c-d, 32c-d, and 33c-d. Table 8 shows the times of divergence, visual rank, and APMI values for δ = 40 kb and 10 kb.

As a preliminary assessment of intragenomic structure, Fig. 34 and 35 help characterize the invariant arrangement of ORFs based on both polarity and COG membership. For my population of 165 chromosomes, I inspected the z-score values of significance (number of

Table 9: Normalized ranges of polarity tallies.[a]

Characteristic	All ORFs		1...100		201...300		401...500	
Lifestyle	Mdn	Mean	Mdn	Mean	Mdn	Mean	Mdn	Mean
Oblig. ancient	0.14	0.15	0.29	0.29	0.26	0.26	0.28	0.32
Oblig. recent	0.08	0.09	0.26	0.29	0.26	0.32	0.29	0.28
Freeliving repl.	0.09	0.10	0.24	0.26	0.24	0.28	0.29	0.28

[a]Bacterial chromosomes are grouped into three characteristic lifestyle categories: obligate ancient host associated, obligate recent host associated, and free-living replicative stage. The normalized range of polarity tally is the difference between the highest and lowest points of the tally divided by the number of ORFs. The second column evaluates the entire stretch of the chromosome for each bacteria. The third, fourth, and fifth columns look at selected sets of ORFs where ORFs are numbered consecutively from the start point of the chromosomal annotation.

sigma σ units separating original and randomized assignments of polarity and COG membership) from my running tally methodology. The average polarity running tally z-score difference was 79.7σ ($p < 0.0001$). Only 10 of the 165 chromosomes (6%) had a z-score $< 1.64\sigma$. Of these 10 chromosomes, the two most predominant phyla were Cyanobacteria ($n = 3$) and Euryarchaeota ($n = 2$). The average COG running tally z-score difference was 14.38σ ($p < 0.0001$), yet 14 of the 67 chromosomes (21%) had z $< 1.64\sigma$. Of these 14 chromosomes, the three most predominant phyla were Proteobacteria ($n = 4$), Euryarchaeota ($n = 2$), and Firmicutes ($n = 2$). 49 of 67 chromosomes were significant (z-score ≥ 1.64) for both polarity and COG membership.

Based on the approximate lifestyle boundaries of chromosome size shown in Figure 36, there was an almost two-fold steeper descent and ascent of the polarity-based running tally for obligate, ancient host-associated genome-sequenced strains compared to the genomes of recently host-associated and freeliving strains (Table 9). This overall range of the polarity tally did not uniformly correspond to changes for localized regions of the chromosome. The start point of chromosomal annotations (most likely near to the origin of replication) did not manifest a steeper descent or ascent of the polarity-based running tally for the genomes of obligate, ancient host-associated strains.

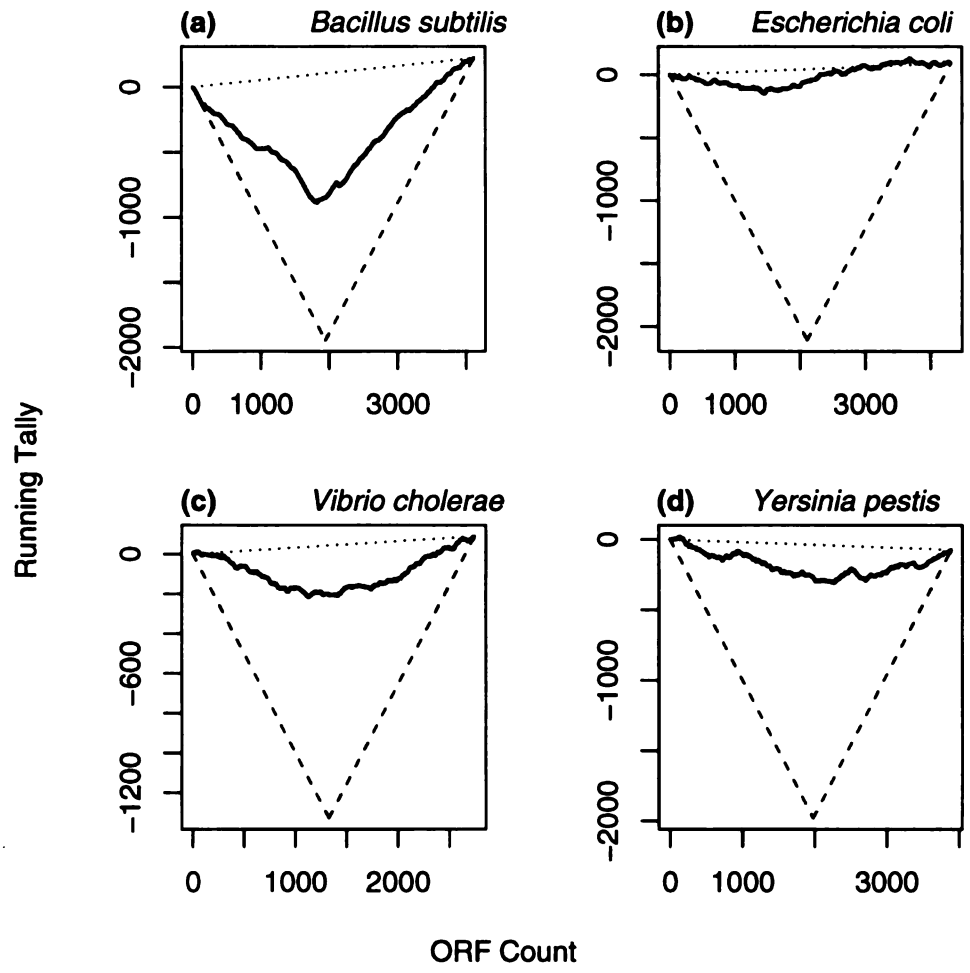

Figure 34: Running tally graphs of polarity along four chromosomes. The thick line represents increments and decrements based on whether an ORF has an assigned polarity value of 1 or not. The dotted diagonal represents random expectation where polarity values are randomly assigned to a chromosomal set of ORFs. The dashed lines forming a V-shape represents the pattern if polarity values were not intermingled. **(a)** *Bacillus subtilis* subsp. *subtilis* str. 168. **(b)** *Escherichia coli* K-12. **(c)** *Vibrio cholerae* (large chromosome). **(d)** *Yersinia pestis* CO92.

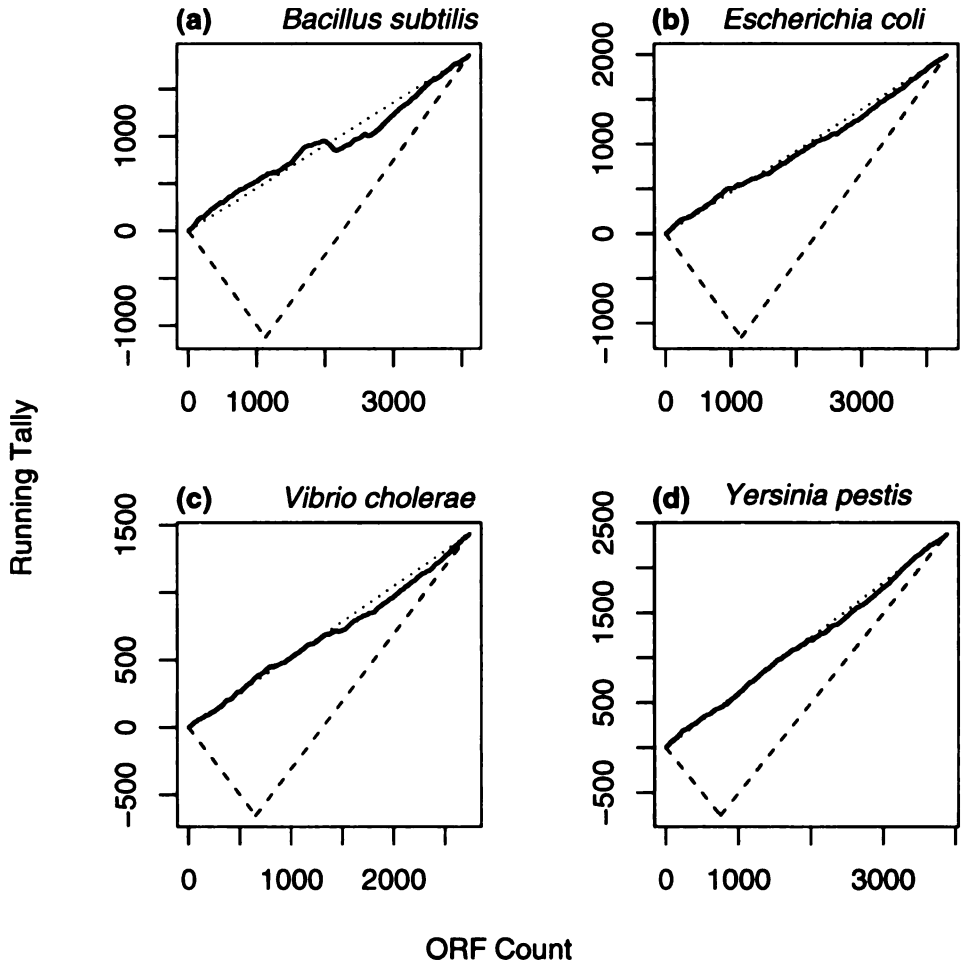

Figure 35: Running tally graphs of COG membership along four chromosomes. The thick line represents increments and decrements based on whether an ORF is a member of a COG or not. The dotted diagonal represents random expectation where COG membership is randomly assigned to a chromosomal set of ORFs. The dashed lines forming a V-shape represents the pattern if all COG non-members were together followed by COG members. **(a)** *Bacillus subtilis* subsp. *subtilis* str. 168. **(b)** *Escherichia coli* K-12. **(c)** *Vibrio cholerae* (large chromosome). **(d)** *Yersinia pestis* CO92.

3.2.3 Chromosome Size

I found that the smallest fully sequenced prokaryotic genome was represented by the *Nanoarchaeum equitans* chromosome (491kb) and the largest prokaryotic genome was represented by the *Bradyrhizobium japonicum* chromosome (9.1Mb). Across the set of 155 genome-sequenced strains, I found that genome size (as represented by chromosomes) was suggestive of ecological boundaries (based on Wilcoxon rank sum calculations) in terms of my own, ad hoc, ecological grouping. I also inspected taxonomic rankings to assess the general degree to which genome size could be a distinguishing characteristic of shared ancestry. Expectation of the difference between median genome sizes is represented by the symbol Δ. The Archaea (genome size, 491kb - 5.8Mb) and Bacteria (genome size, 580kb - 9.1kb) are statistically different based on median genome size ($p < 0.03$; 95% confidence interval (bp): $-1,730,369 < \Delta < -91,633$). The Alphaproteobacteria (1.1Mb - 9.1Mb) versus Gammaproteobacteria (616kb - 6.4Mb) do not however represent a significant difference in the medians of genome sizes ($p < 0.89$; 95% confidence interval (bp): $-1,408,279 < \Delta < 1,595,946$). The Lactobacillales (1.9Mb - 3.3Mb) versus Enterobacteriaceae (616kb - 5.7Mb) represent only a weak significance in median genome size difference, and the interpretation is inconclusive based on the confidence interval ($p < 0.059$; 95% confidence interval (bp): $-2,778,116 < \Delta < 1,154,717$).

Fig. 36 presents the distributions made by the 165 prokaryotic chromosome sizes ($\bar{x} = 3.1$ Mb, $\tilde{x} = 2.7$ Mb, and $s = 1.83$ Mb). The distribution of chromosome sizes appears multimodal. The declining trends at the lower and upper limits of the distribution may reflect a limit to the overall size of a prokaryotic genome. There is evidence for relationships between multimodal ranges of genome size and "regimes" of recombinative change; most intracellular endosymbiotic bacteria have low levels of recombination and smaller genomes (genome size range 640kb to 1.3Mb), compared to free-living bacteria with bigger genomes such as those in the soil (genome size range 4.2Mb to 9.0Mb) (Moran & Plague, 2004). The plotted lifestyles and boundaries of Fig. 36 are based on Ochman & Davalos (2006) and Moran & Plague (2004).

While I did not find that chromosome size directly corresponded to general taxonomic distinctions for taxonomic rankings such as phylum, class, order, and family, I did find

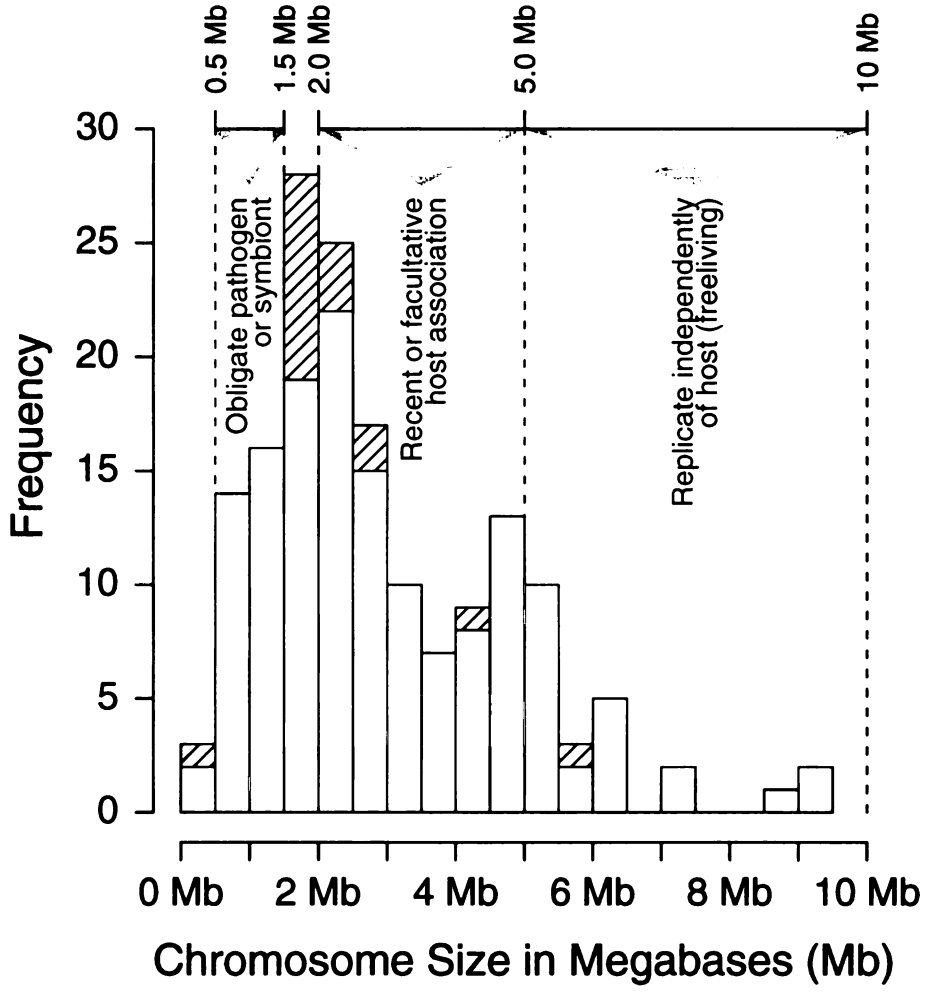

Figure 36: Histogram of 165 chromosome sizes. Bin size is 500,000 base pairs. Chromosome sizes are shown for 165 different chromosomes coming from 155 genome-sequenced strains. Frequencies of archaeal chromosome sizes are indicated by shaded boxes stacked above the frequency counts of bacterial chromosomes shown by unshaded boxes. The approximate boundaries for three lifestyle-based ranges of genome sizes are listed at the top of the figure and are indicated by vertically descending, dashed lines.

evidence for conservation of chromosome size from a last common ancestor. While the overall linear relationship between a change in chromosome size compared to time of divergence from a last common ancestor does not fully account for variation ($r^2 = 0.24$), the slopes in Fig. 37a are uniformly positive and support a general property of conserved genomic size. I found that a change in chromosome size is independent of the shared pointwise mutual information for various segmentation sizes. Fig. 37b characterizes the independence of chromosomal size with respect to APMI for δ =40 kb (I_{40}). While a change in chromosome size (ΔC_s) increases with a greater time of divergence, the trend for I_{40} is to decrease. To evaluate the effect of a joint consideration of both I_{40} and changed chromosomal size (ΔC_s), I evaluated both $\Delta C_s / I_{40}$ (Fig. 37c) and $\Delta C_s - I_{40}$ (Fig. 37d). The joint considerations had stronger linear correlations ($r \geq 0.6$).

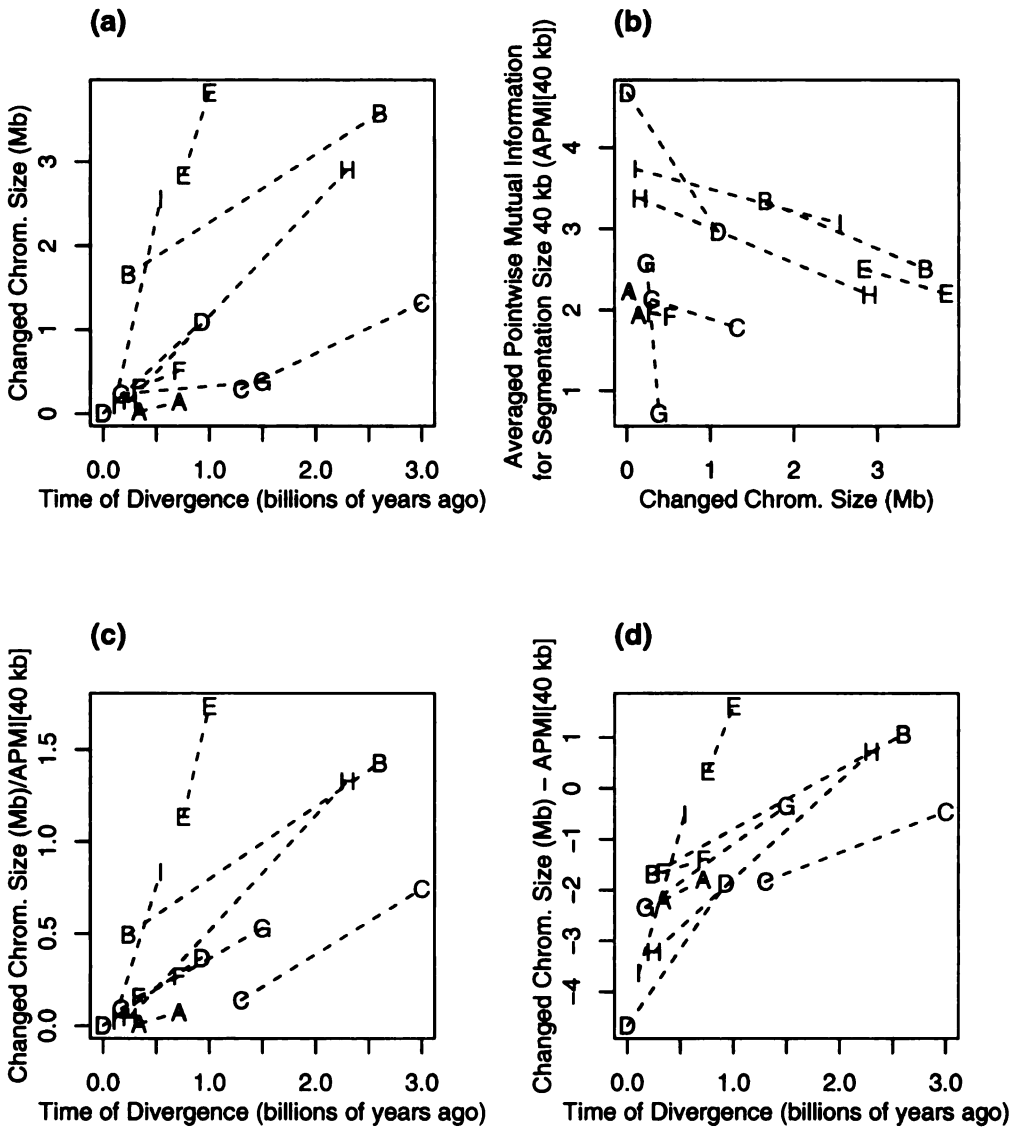

Figure 37: Relationship of divergence time from a last common ancestor to changes in chromosome size and pointwise mutual information. The letters represent pairwise comparisons described in Table 5. **(a)** Absolute difference in chromosome size for various times of divergence, $m = 0.73, r = 0.49$. **(b)** The absolute difference in chromosome size versus the APMI for 40 kb $((\Delta C_S) - I_{40})$, $m = 0.1$ and $r = 0$. **(c-d)** Chromosomal difference for various times of divergence based on joint considerations of APMI and chromosome size. **(c)** The product of absolute difference in chromosome size with the reciprocal of the APMI for 40 kb $(|\Delta C_S|/I_{40})$, $m = 0.37$ and $r = 0.60$. **(d)** The absolute difference in chromosome size with the reciprocal of the APMI for 40 kb $(|\Delta C_S| - I_{40})$; $m = 1.24$ and $r = 0.66$.

3.3 Discussion

The data set of 155 genomes does not reflect an accurate accounting for the overall ecological and phylogenetic diversity of bacteria (Cohan, 2004). The estimated number of prokaryotes on earth is $4 - 6 \times 10^{30}$ with 92 to 94 percent of these prokaryotes being in soil subsurface regions: "marine sediments below about four inches and terrestrial habitats below about 30 feet" (Schloss & Handelsman, 2004). I found ecologies other than soil subsurface regions to be oversampled in the data set of fully sequenced genomes. Despite the patchiness of taxonomic representation, the set of fully sequenced genomes has been a key component in contemporary interpretations and estimations of prokaryotic diversity concerning genomic organization and phylogeny (Moran & Plague, 2004; Ochman & Davalos, 2006; Battistuzzi *et al.*, 2004; Horimoto *et al.*, 2001). Taxonomic estimates based on available ribosomal data are for 35,498 species and 50 phyla, and a total estimation of a planetary species count is 10^5 to 10^7 (Schloss & Handelsman, 2004). While only a paltry 126 species are represented by the set of 155 genomes, the 16 represented phyla provide a broad phylogenetic coverage (32% of 50). In this sense, the set of 155 genomes may reasonably characterize wide-ranging aspects of the prokaryotes. A variety of replicon structures characterizes each of the 155 genomes in terms of linear and circular chromosome topologies and compositions of single or multiple distinct chromosomes. I did not find plasmid sequence data to be consistent between the public NCBI data archive of genomic sequences and the literature. The functional definition of a plasmid as being of a non-essential, and possibly non-stable, association with a viable organism may allow for it to be considered separately from a chromosomal representation of a genome.

Comparing the organization of ORFs can be an effective technique for identifying divergent recombinations (Horimoto *et al.*, 2001; Kalman *et al.*, 1999; Rocap *et al.*, 2003; Zivanovic *et al.*, 2002). Yet, there are large gaps of time in the estimated phylogeny for which the data set may not be large enough to resolve every recombinative event with sufficient statistical power (Fig. 8 - 12), and multiple sets of mobile elements can be expected to produce complex trajectories of altered chromosomal arrangement (Gray, 2000). Based on their representative sample sizes inside my analyzed data set, the Proteobacteria ($n = 68$) and the Firmicutes ($n = 38$) have the greatest comparative power to reconstruct distant

recombinational events. Yet, the Proteobacteria and Firmicutes are well-populated with pathogenic and symbiotic bacteria, and modern analyses of genome size and structure are predisposed to produce categorizations aligned with host-associated lifestyles (Bentley & Parkhill, 2004; Moran & Plague, 2004; Ochman & Davalos, 2006). While divergence from last common ancestors between various prokaryotic strains extends back several billion years or more, a focused perspective on host association only relates to a time span stretching back to the Cambrian age 600 Ma and, more recently, the emergence of mammals 107 Ma (Rokas *et al.*, 2005). The inclusion of Archaea in the analysis helps obviate a limited view of past history since the Archaea appear to strictly exclude pathogenicity as a form of host association. While some Archaea are host-associated commensals, this phenotype may be primarily due to metabolic pathways atypical of the bacteria that do not benefit from mortality of the host organism (e.g., methanogenesis) (Gill *et al.*, 2006). Based on phenotypic descriptions inside genomic sequence publications, I found only one of the 17 Archaea in the data set of 155 genomes to be host associated was *Methanosarcina acetivorans* (Galagan *et al.*, 2002).

An accurate evaluation of character evolution as a consequence of recombination would require treatment of a patchy taxonomy by specification of a uniform taxon (Grafen & Ridley, 1997). As Fig. 17 demonstrates, I used broad, mutually exclusive groupings based on the taxonomy to conduct my subsampling. For a finer-grained treatment, I attempted to contrast the effects of genus membership with effects of species membership, although species and genus definitions are not yet fully defined (Cohan, 2002). Generally, in my analysis, the sets of closely related strains and species (Tables 3 and 4) are distributed among various lineages, and this supports the usage of the available data set to characterize dynamics of chromosomal organization that are common to the prokaryotes. A finer resolution to identify specific selection pressures associated with various lineages in various environments may be achieved by greater numbers of representative strains. I did not arrive at a uniform taxon that was useful for approaching hypotheses of specific recombinative character states. Ideally, the rate of heritable changes produced by chromosomal reorganization could have been analyzed for correspondence with generational or chronological measures (Pagel, 1994).

The scope of the represented taxonomy appeared to support a goal to identify common

limits and general characteristics of prokaryotic change as relates to ORF organization. To accomplish uniformity with the analysis, I sought to inspect prevalent "units" of information on prokaryotic chromosomes, and the most uniformly annotated feature appeared to be that of open reading frames (ORFs). The identification and annotation of these ORFs significantly relies upon automated assessments of ORF regions and similarities to other "known" ORFs (Frishman *et al.*, 1998). A common software tool for identifying ORFs on a DNA sequence is Glimmer (Salzberg *et al.*, 1998), although there are ongoing efforts to better characterize the degree of confidence associated with a computer-generated annotation (Larsen & Krogh, 2003). The need for updating these initial annotations is dire (Roberts *et al.*, 2004). While ORFs may be more uniformly annotated in the NCBI data files than other genomic features, a struggle has been to arrive at a better estimation of real ORFs versus "not real" ORFs and, perhaps, take into account the natural dynamics of gene loss and formation (Skovgaard *et al.*, 2001; Snel *et al.*, 2002). While the number of annotated ORFs on a chromosome appears to correspond strongly to one ORF for every 1,112 base pairs of chromosomal DNA (see Fig. 14), the number of chromosomal base pairs that encode each ORF varies. Consistent with my findings for the data set of 155 genomes (Fig. 15 - 18), a large set of ORF lengths generally follows an L-shaped, lognormal frequency distribution (Skovgaard *et al.*, 2001) that is locally disrupted in a fashion suggestive of underlying, discretely-sized, multidomain protein structures (Wheelan *et al.*, 2000; Savageau, 1986). The lognormal distribution of ORF sizes may be partly explained by a physical model of fragmentation (Azad *et al.*, 2002) where, starting from the right-side of the distribution, there is an exponential growth in the number of ORFs as the ORF length decreases. Potentially then, those ORFs experiencing a higher degree of arbitrary, nonsense mutations will constitute a closer fit to a lognormal distribution than ORFs encoding a protein structure that is strongly conserved in an inviolate form. As seen from Fig. 16 and 17, those ORFs that range in length from 0 to 127 aa are comparatively non-disrupted in their frequency distribution compared to ORFs long enough ($\geq$ 127 aa) to contain a protein domain. The entire distribution of ORFs demonstrates some non-lognormality based on a rightwards shift of the distribution (Fig. 18). For ORFs that are most important to the fitness of an organism, I postulate that their log-scaled distribution of lengths would range

higher in value and have a greater characteristic of non-normality relative to a set of falsely annotated, or putatively noisy ORFs.

ORFs that are members of clusters of orthologous groups (COGs) are a possible lower bound to the total number of annotated ORFs that correspond to real proteins (Skovgaard *et al.*, 2001). In practice, COGs are established by bidirectional best hits, and this strong conservation of sequence is used as a basis to infer vertical divergence from a common ancestral ORF. This characteristic of bidirectional best hits is also useful for assaying conserved ORF organization among related genomes (Zivanovic *et al.*, 2002) as demonstrated by Fig. 19 and 20. Comparisons among genomes belonging to the same species tend to produce a diagonal line from the bottom left to the upper right. The slight deviation observed for the comparison of *Escherichia coli* strains in Fig. 19b is likely attributable to prophage insertions (Hayashi *et al.*, 2001; Canchaya *et al.*, 2003). For more distant times of divergence from a last common ancestor, Fig. 25 - 33 show patterns of dispersal for orthologous ORFs.

Across multiple lineages, I did not find a constant relationship of the time of divergence to the degree of disruption for visual and quantitative assays of conserved ORF organization, although a general trend was evident. There are competing possibilities concerning the interpretation of dot matrix plots of conserved ORF organization. While comparisons with a distant relation *Pyrococcus furiosus* (715 Ma) exhibit less conservation (Fig. 25c-d) than for more closely related *Pyrococcus* species (338 Ma, Fig. 25b), this may relate to more than just a rate of recombinative change over time. There is a set of 23 homologous insertion sequence elements exclusively present in *P. furiosus* (Zivanovic *et al.*, 2002; Lecompte *et al.*, 2001) that is closely associated with putative locations of rearrangement on the *P. furiosus* chromosome, and this greater amount of mobile elements may also account for the greater pattern of disruption. The conservation of ORF organization of *Yersinia pestis* compared to *Escherichia coli* (Fig. 20b) is less than among *Pyrococcus* species (Fig. 25) despite similar times of divergence (Battistuzzi *et al.*, 2004). This may in part correspond to the disproportionately high degree of IS elements in the *Yersinia pestis* genome (3.7%) (Parkhill *et al.*, 2001b) where there are > 100 IS elements (Deng *et al.*, 2002). Other heavily disrupted dot matrix plots with relatively recent ($\leq$ 1 Ga) times of divergence are for the set of Cyanobacteria (Fig. 29) and a set of Lactobacillales (Fig. 30). The dense pattern seen for

the Cyanobacteria may be attributable to the prevalence of transposase genes in freshwater cyanobacteria and the complex adjustments needed to support free-living oxyphototrophy in an unstable aqueous environment (Dufresne *et al.*, 2003). The extensive scattering and rearrangement for *Streptococcus pyogenes* may be due to phage activity (Nakagawa *et al.*, 2003). By comparison, the Mycoplasma lineage relies on homologous recombination due to direct repeats more so than IS elements (Rocha & Blanchard, 2002), and there is less scattering involving solitary ORFs seen on the dot matrix plot in Fig. 31b). While the dot matrices are interpretable from closer analyses oriented for specific genomes, I did not find a simple relationship involving mobile elements, or strong correlation with time, that would uniformly account for the diversity of dot matrix pattern across various lineages. Overall, there is a variety of factors that may underly the patterns of the dot matrix plots, and a comparative study would benefit from a greater sample size to confirm many of the putative relationships with recombinative mechanisms and strategies for genomic plasticity.

Phylogenetically broad patterns of genomic content and reorganization implicate aspects of microbial ecology (Terzaghi & O'Hara, 1990; Moran, 1996). I found evidence for relationships between multimodal ranges of genome size and "regimes" of recombinative change where most intracellular endosymbiotic bacteria have low levels of recombination and smaller genomes (genome size range 640kb to 1.3Mb) compared to free-living bacteria with bigger genomes such as those in the soil (genome size range 4.2Mb to 9.0Mb). Ochman & Davalos (2006) propose a high degree of instability for genomes that are 2 Mb to 5 Mb in size. If this instability means a greater rate of departure from this size range than rate of entry per organism, then the trough in frequency for chromosome sizes between 3.5 Mb - 4 Mb and the chromosome size frequency peaks at approximately 2 Mb and 5 Mb are consistent with such a differential flux in genomic content (Fig. 36).

The internals of genomic structure offer some evidence to account for the diversity of ORF organization. A potential consequence for varying degrees of activity of mobile elements may implicate a difference in colinearity of transcription and replication. A reduced correspondence of polarity of ORFs with the replichores is reported for *P. furiosus* where the primary pattern is for only the highly transcribed ORFs to correspond in transcriptional polarity with direction of replication (Zivanovic *et al.*, 2002). My running tally method

implicated astronomically high levels of significance for both ORF polarity and COG membership organization. I found that my measure is inconsistent with the claim by Brüggemann *et al.* (2003) that the *Vibrio cholerae* and *Yersinia pestis* chromosomes do not manifest a cotranscriptional effect. Overall, based on my running tally measure, only 84% of chromosomes were significant for organizational patterns of both ORF polarity and COG membership, and the pattern of ORF polarity was more pronounced than for COG membership. Almost half of the Cyanobacteria genomes were not significant for my polarity-based measure of organization, and there was a steeper descent and ascent of the polarity-based running tally for obligate, ancient host-associated genome-sequenced strains (Table 9). I did not find any further, simple indicators to account for the variation of z-scores in terms of mobile elements or taxonomic groupings. There was some evidence of the the polarity tally being influenced by more than just a cotranscriptional effect; the indistinct change in tally near the origin of replication for obligate, ancient host-associated genome-sequenced strains would be consistent with the origin of replication as a hot spot of localized rearrangements, perhaps due to the greater availability of a single-stranded intermediate at the origin. The pattern of COG member clustering did not correspond to any simple indicators of lifestyle or taxonomy. Broadly considered, the non-random pattern of COG member clustering may be attributable to regulatory (Lathe *et al.*, 2000) and functional (Li *et al.*, 2005; Wolf *et al.*, 2001) constraints on sustainable schemes of genomic reorganization as well as paralog-forming pathways of gene addition (Snel *et al.*, 2002; Liang *et al.*, 2002).

Compared to other types of functional assessments such as operons and promoters, ORF data on the 155 genomes is better annotated and may have greater analytical power both in representative size, and the potential for consistent informational treatment. Also, ORFs appear, as a population, to address meaningful comparisons for metabolic, ecological, and evolutionary questions (Bentley & Parkhill, 2004; Konstantinidis & Tiedje, 2004). Additionally, there are simple quantities that detail an ORF object: start, stop and length are all generally determined by integers corresponding to a zero point on a replicon sequence. There is a reasonable level of accuracy ($> 99\%$) for identification of ORF start and stop points (Delcher *et al.*, 1999). A quantitative approach based on these simple, exactly

described values may be more readily achievable than, for example, estimated kinetics of macromolecular bindings to various consensus-based estimates of promoter sequences. A meaningful evaluation of ORFs as a population across the phylogeny may require some uniform capacity to determine those ORFs that encode for proteins that are important to the physiology of the cell.

Approaches to characterizing ORF sequence conservation have largely involved simplistic ORF comparisons of similarity. A principal criterion of COG assignments is based upon bidirectional best hits (Tatusov *et al.*, 1997b). Bidirectional best hits relate to a pairwise comparison between organisms where an ORF in one organism's genome matches most closely to a particular ORF on the other organism's genome, and vice versa. A COG must contain at least three members from three reasonably separate lineages. Overall, 75% of annotated, prokaryotic ORFs belong to COGs (Tatusov *et al.*, 2003).

The calculation of COGs, as described, has deficiences. The parameters of COG similarity are lax enough to avoid false negatives but, subsequent to the BLASTP search, putative COG groupings have to be manually inspected and sometimes split apart (Tatusov *et al.*, 1997b). The bidirectional best-hit criterion is a pairwise comparison of ORF similarities that ignores meaningful information that can come from comparisons involving several or more ORFs (Park *et al.*, 1998). Pairwise comparison is not just limiting for the assessment of orthology. Over half of the paralogous gene relationships in *Mycoplasma genitalium* are not accounted for when just pairwise sequence comparisons are utilized (Teichmann *et al.*, 1998).

By relaxing or tightening a filter for sequence conservation, various temporal relationships can be investigated. The identification of recent paralogs has involved length similarity of 95% or more and sequence similarity of 95% or more (this implicates about 5% of ORFs) (Kawarabayasi *et al.*, 2001). For larger familial groupings (52% of ORFs), Kawarabayasi *et al.* (2001) considers amino acid identity higher than 30% for over 70% of the entire ORF region. There have been a variety of efforts to better define the meaningful cut-off values for BLAST-like similarity computations and how they relate to structure and function (Chung & Yona, 2004; Bern & Goldberg, 2005; Sadreyev, 2003; Pagni & Jongeneel, 2001; Krasnogor, 2004). A current trend has been for inspecting protein domains (Birkland

et al., 2005; Yang *et al.*, 2005; Service, 2005). While similarity values may sometimes be too restrictive and miss out on larger protein family or functional relationships, being too relaxed can impede discernment of underlying trends associated with conserved domains. Comparisons of biological function is appropriate when there is sequence-level identity of 25% (Krasnogor, 2004). Yet, for sequence-based identities of 20-30%, only one half of the domain repertoire relationships are shown in *Mycoplasma genitalium* (Teichmann *et al.*, 1998).

If the task is to approach a uniform separation of ORF sets that is meaningful across different lineages, a heuristic may utilize patterns of ancestral conservation while accommodating some range of natural divergence. I propose a separation be sought only along generalized objectives to characterize efficiently an approach that removes about 10-30% of the ORFs, separates multiple regimes of variance, and correlates with expected factors of sequence features and functional genomics information. Only after arriving at a plausible distinction of ORFs, can a specific, hypothesized ratio of operational versus silent annotated ORFs be evaluated. Reducing the complex nature of genomic content and organization into comparable events of change across the phylogeny requires some capacity to establish limits and parameters for recombinative units. Prior to evaluating specific hypotheses concerning the nature of ORF clustering, the natural fluctuations of the ORF-ome and the prevalence of putatively false or silent ORFs in the annotated data files presents a challenge for broadly distinguishing those ORFs of functional and evolutionary importance to the composition and organization of a chromosome.

Chapter 4: Subsets of Open Reading Frames

4.1 Comparative Parameters of Sequence Conservation

Ideally, the identification of a generally legitimate, "real" subset of ORFs would reduce observational noise, and further provide some account for the fluctuations and evolutionary pressures that influence the set of ORFs in a given prokaryotic genome. As described in Section 1.4, conservation of sequence is a reasonable basis for inferring the importance of ORFs, and effective approaches have involved measures of sequence similarity (Snyder & Gerstein, 2003), ORF length (Skovgaard *et al.*, 2001), and grouped similarities involving more than just two sequences (Park *et al.*, 1998). I sought to construct a general filter with the parameters L (identity of ORF length), B (identity of ORF sequence), and S, size of a similarity cluster. S is evaluated as an inclusive count of a similarity set. If a sequence is only similar to itself, then $S = 1$. If an ORF sequence is similar to 4 other ORFs (in addition to itself), then $S = 1 + 4 = 5$. As $S > 2$ for a given ORF, information exceeding that of a pairwise comparison is incorporated. S is interdependent with the constraints of similarity specified by L and B.

I evaluated the strength of similarity between any two ORFs as a function of B and L. The basis for a length constraint is that it enforces some conformity for the internal structural integrity of two ORFs with shared ancestry (Wheelan *et al.*, 2000), and further focuses the assessment of similarity to the distribution of "immutable" ORFs, as opposed to those ORFs altered by gene fusions and fissions. Localized alterations to ORF content and structure may have structural and functional consequences in terms of important motifs such as conserved domains. Constraining L so as to achieve this distinction is not a well-modelled proposition, so L was characterized as being, at most, $\pm 10\%$. For example, with $L \rightarrow [-10\%, +10\%]$, an ORF length of 200 aa would match ORFs of lengths 180 aa to 220 aa but not lengths < 180 aa or > 220 aa. By spot checking a few test cases, I found that a L

constraint of $\pm 10\%$ was effective in reducing erroneous homologies that might otherwise be inferred from low-complexity protein domains.

A value of $B \leq 10^{-6}$ is the default for the BLASTCLUST application, where B is the expectation score of a BLASTP comparison among ORFs clustered by similarity. The default parameters of BLASTCLUST reportedly work to "anecdotally" identify closely related protein families from closely related prokaryotes, and "virtually eliminate false positives" (Wolf, 2004). The parameters of "coverage" (`-L 0.0`) and "score density" (`-S 0.0`) were set so as to not be evaluated by the BLASTCLUST algorithm.

If $L \rightarrow \pm 10\%$ and $B \leq 10^{-6}$, the remaining objective for defining an initial sequence conservation threshold for ORFs relates to the value of the similarity cluster size, S. As S increases, the pervasiveness of the ORF as an immutable unit of evolution across phylogeny is justified. Two aspects to the performance of a given S-based constraint involve 1) random similarity matches versus truly homologous matches, and 2) the phylogenetic range of observed matches. In the set of 155 genomes, there were a variety of closely related strains in the data set with up to 5 members of the same species, so an S value of 5 may sometimes only implicate a recently emerged last common ancestor. As ORFs with larger S values are identified, this may encompass a larger phylogenetic range by implicating more distantly related lineages. An S value of 155 would implicate an ORF common to all 155 genomes.

If an ORF matches just one other ORF different than itself ($S = 2$), then there may be a significant chance for the match to be a false positive. A BLASTP expectation score threshold of $B \leq 10^{-2}$, corresponds to an expectation for finding a single ($10^{-2} \times 100 = 1$) false positive match against a set of 100 other sequences (Koonin & Galperin, 2003). In a Bayesian sense, if $B \leq 10^{-6}$, then the percent chance for a false positive match ($S = 2$) to another ORF from the 165 chromosomes is 45% ($447,550 \times 10^{-6} = 0.45$). For an ORF to have two other false positive matches ($S = 3$), the probability is $0.45 \times 0.45 = 0.20$. When $S = 5$ and $S = 6$, the Bayesian-calculated probabilities approach statistically acceptable levels of significance, respectively $0.45^4 = 0.041$ and $0.45^5 = 0.018$. While $S \geq 5$ significantly implicates a true match with at least one other ORF ($p > 0.95$ for a legitimate set of "twins"), $S \geq 6$ implicates the existence of at least two other matches (a legitimate set of "triplets") and a potentially wider range of phylogenetic coverage. If a filtered ORF subset

were to be based on evolutionary information from more than just two ORFs, $S \geq 6$ would be the preferable constraint. Generally considered, ORFs with high S values would more likely belong to an evolutionarily conserved subset compared to ORFs of relatively low S counts such as $S = 1$. ORFs belonging to very large similarity clusters may be pervasively similar due to strong sequence features of evolutionary importance or because of ubiquitous low-complexity subsequences such as those that encode for coiled-coiled regions. For the analysis, ORFs corresponding to $S > 40$ were given an inclusive S value of "40+" so as to not resolve complex transitive relationships of ORF similarity cluster assignments. I termed the subset of ORFs with $S > 40$ as the "ubiquitous" subset of ORFs (U-ORFs). Based on the statistical considerations of length $|L| \leq 10\%$, sequence similarity $B \leq 10^{-6}$ and similarity cluster size $S \geq 6$, I arrived at an initial distinction for operational ORFs (O-ORFs) versus a "silent", putatively false subset of ORFs (S-ORFs) with the expected relationship of U-ORFs $\subset$ O-ORFs. The subset of O-ORFs that does not include U-ORFs ($6 \leq S \leq 40$) is a subset that I termed as the N-ORF subset. A summary of ORF subset terminology is shown in Table 10.

4.2 Paralogous ORFs

I evaluated my similarity clusters for possible instances of paralogy where two or more ORFs in the same similarity cluster belonged to the same chromosome. Objectives for this assay were to 1) evaluate the presence of intragenomic pattern attributable to duplication of content, and 2) analyze and infer taxon-based constraints for paralog formation. The number of recorded paralogs inside the defined similarity clusters for the 165 chromosomes ($2 \leq S \leq 40$) ranged from 1% (*Chlamydia trachomatis* MoPn) to 18% (*Methanosarcina mazei*) of the total number of annotated ORFs for each evaluated chromosome. I compared my putative paralogs to a more expansive effort at characterizing paralogy (Pushker *et al.*, 2004). Pushker *et al.* (2004) characterize a a range of 10% to 50% of ORFs on a given chromosome as belonging to a paralogous cluster of ORFs. My calculation of paralogs for the 165 chromosomes discarded the U-ORF group ($S > 40$), and this may be significant since the average paralogous family size often exceeds 40 (Pushker *et al.*, 2004). A possible

Table 10: Names and descriptions of open reading frame subsets.

Subset Name	Description
ORFs	The full set of ORFs as they are currently annotated in NCBI-based data files of fully sequenced prokaryotic genomes.
O-ORFs	Putatively operational ORFs. This subset consists of those ORFs belonging to similarity clusters of size ≥ 6.
S-ORFs	Putatively silent ORFs (putative false positives in the full, annotated set). This subset consists of those ORFs belonging to similarity clusters of size ≤ 5.
U-ORFs	Putatively ubiquitous ORFs. This subset consists of those ORFs belonging to similarity clusters of size > 40.
N-ORFs	Putatively operational, but not ubiquitous, ORFs. This subset consists of those ORFs belonging to similarity clusters of size ≥ 6 and ≤ 40.
C-ORFs	The subset of ORFs that are members of a COG (cluster of orthologous groups). This subset covers 25 functional classifications of COGs, including the R and S functional classes that are respectively for "general function predictions" and "unknown functions." C-ORFs are only established for the 67 chromosomes for which COG assignments have been conducted.
X-ORFs	The subset of ORFs that are members of one of the 67 chromosomes for which COG assignments have been conducted, yet are not members of a COG.

consequence to inspecting the N-ORF group ($1 \leq S \leq 40$) and not the U-ORF group may be to limit the general evaluation of paralogy to those ORFs that are more recently diverged and are limited to a particular branch on the phylogenetic tree.

Fig. 38 - 41 show distances between paralogous pairs for various sets of closely related genomes based on data from my similarity cluster calculations as well as from Pushker *et al.* (2004). The pattern of paralogy on each chromosome implicates hot spots of duplicated content and distances between similar ORFs in a way that may be visually diagnostic of the species-level taxonomy. Some of the wild-type *Escherichia coli* strains (O157:H7 and CFT073) have high peaks in frequency for long distances between related paralogs that are similar in value to the first bin ($< 5,000$ bp) (Fig. 38). Most of the *Streptococcus pyogenes* strains have distinctly elevated frequency peaks for related paralogs that are more than 100,000 bp apart (Fig. 40). Most of the paralogs for chromosomes of the *Pseudomonas* species appear to be separated by distances less than 5,000 base pairs (Fig. 39). In addition to having high frequency peaks for related paralogs that are 0 to 10,000 bp apart, *Staphylococcus aureus* species have a slight upswing in frequency for distances greater than 10,000 bp approaching 200,000 base pairs (Fig. 41). For many of the low frequency distances between paralogs, Pushker *et al.* (2004) characterize paralogy for well over 10× the number of paralogous ORFs that I place into similarity clusters. By contrast, the distinctly high frequency peaks of paralogs from my similarity clusters (Fig. 38-41[a,c,e,g]) versus the paralogy clusters of Pushker *et al.* (2004) (Fig. 38-41[b,d,f,h]) are generally less than an order of magnitude (10×) different in value for examinations of identical locations on the respective histograms.

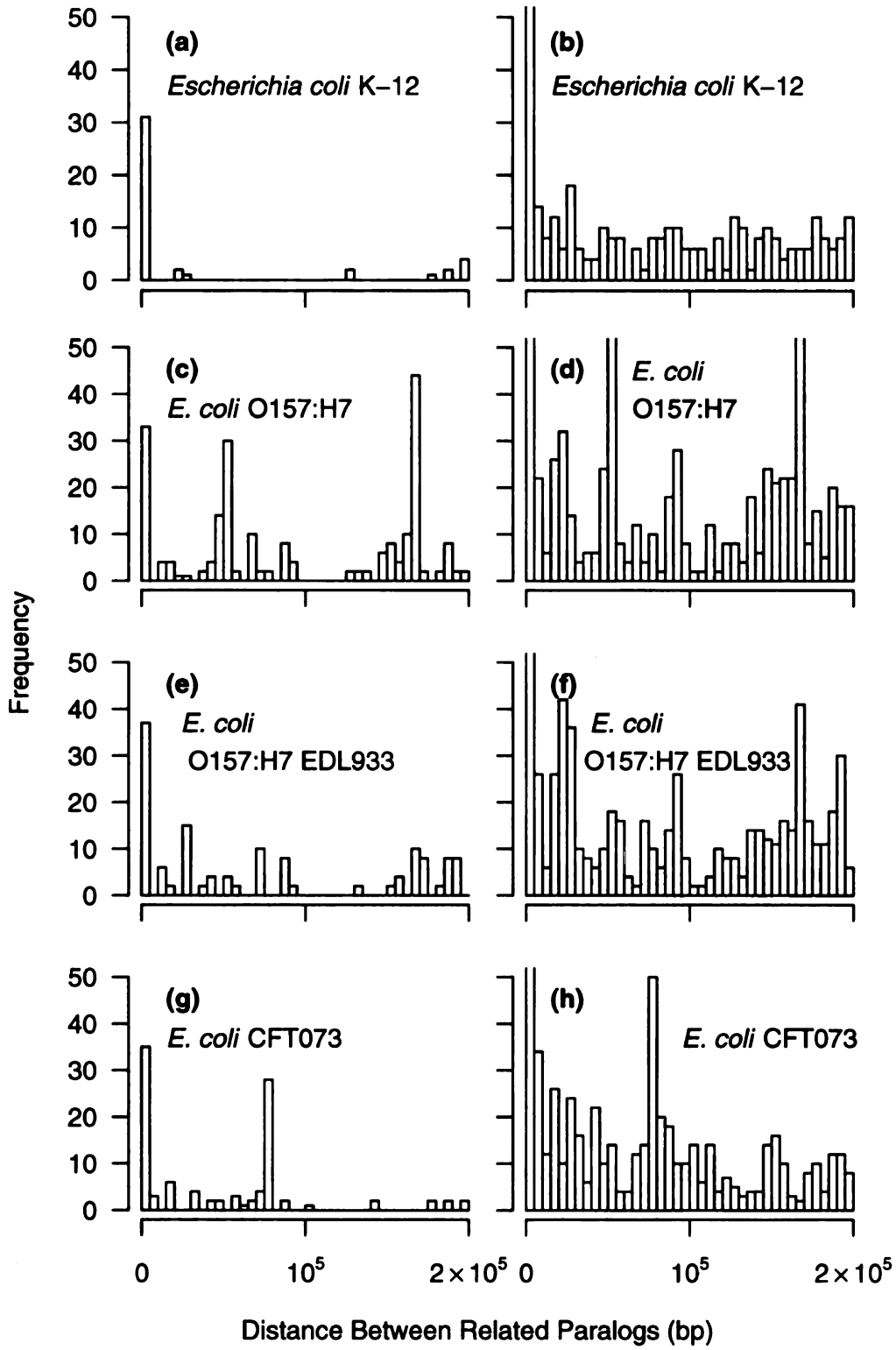

Figure 38: Frequency of distances between related paralogs for four strains of *Escherichia coli*. Only those distances < 200,000 base pairs are shown. Bin sizes are 5,000 bp. **(a,c,e,g)** show the distances between all pairs of paralogs based on similarity clusters involving ORFs on the same chromosome where $6 \leq S \leq 40$. **(b,d,f,h)** are based on data from Pushker *et al.* (2004). Frequency values higher than 50 are truncated on the plot and range from 54-150.

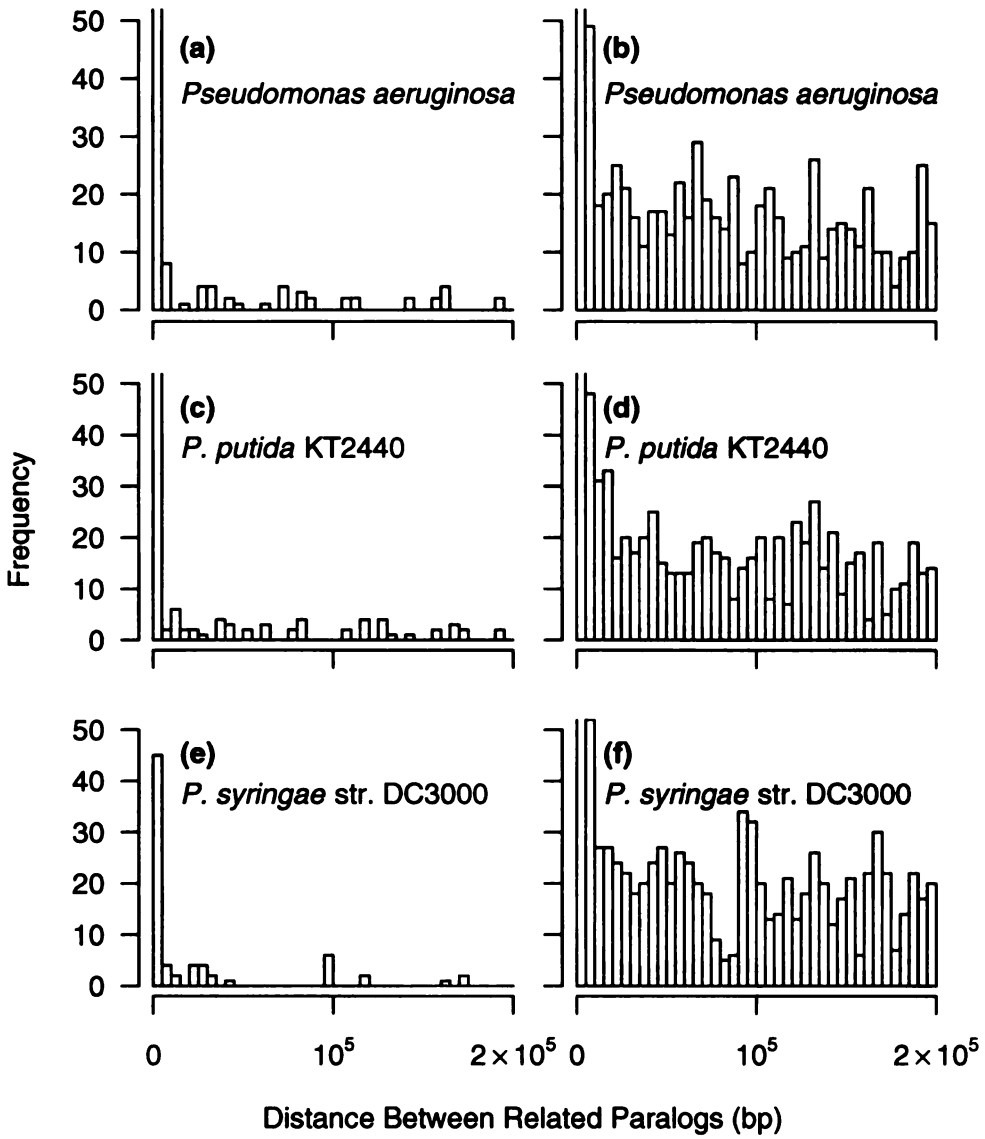

Figure 39: Frequency of distances between related paralogs for three strains of *Pseudomonas* species. Only those distances < 200,000 base pairs are shown. Bin sizes are 5,000 bp. **(a,c,e)** show the distances between all pairs of paralogs based on similarity clusters involving ORFs on the same strain's chromosome where $6 \leq S \leq 40$. **(b,d,f)** are based on data from Pushker *et al.* (2004). Frequency values higher than 50 are truncated on the plot and range from 52-232.

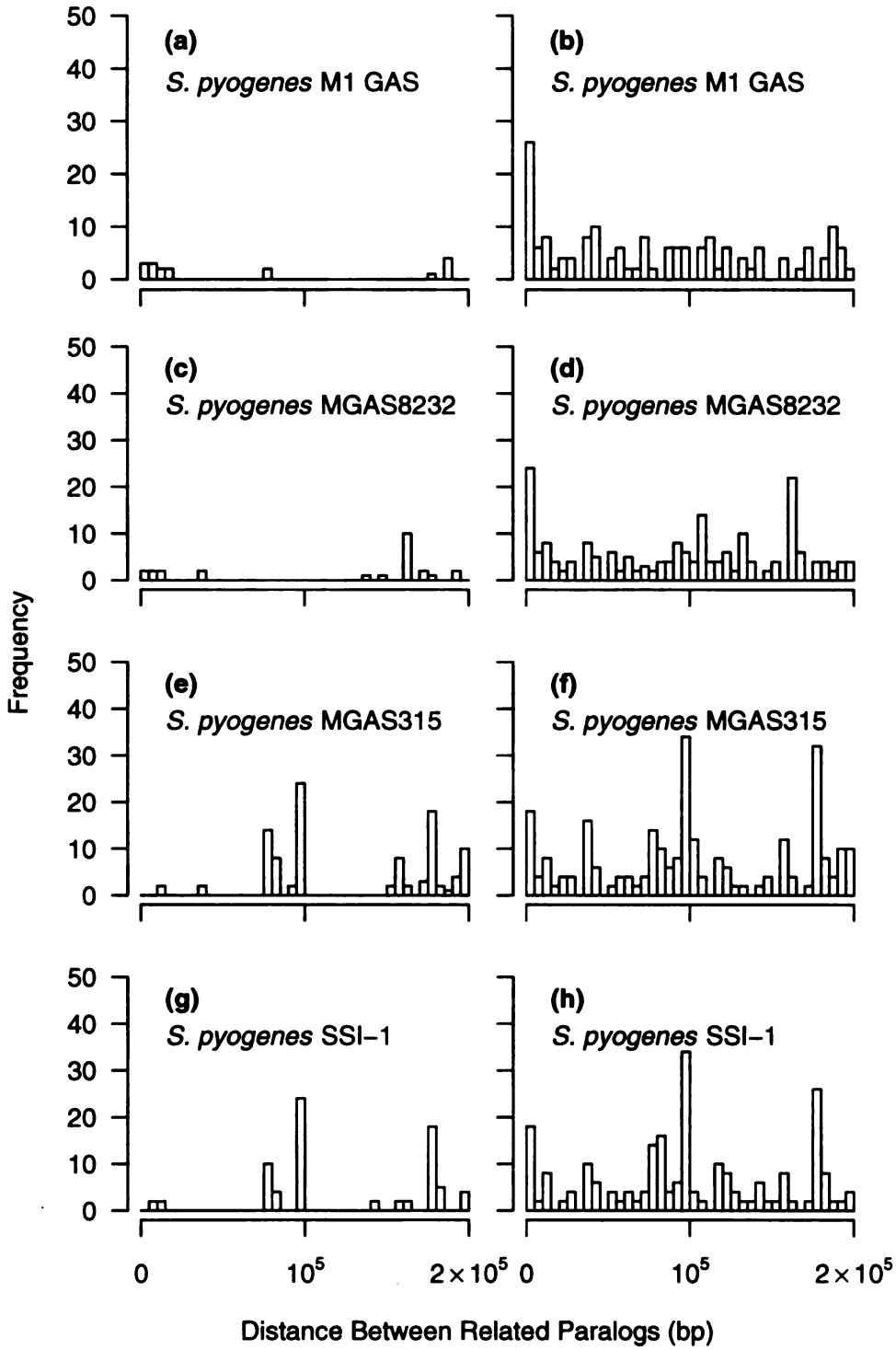

Figure 40: Frequency of distances between related paralogs for four strains of *Streptococcus pyogenes*. Only those distances < 200,000 base pairs are shown. Bin sizes are 5,000 bp. **(a,c,e,g)** show the distances between all pairs of paralogs based on similarity clusters involving ORFs on the same strain's chromosome where $6 \leq S \leq 40$. **(b,d,f,h)** are based on data from Pushker *et al.* (2004).

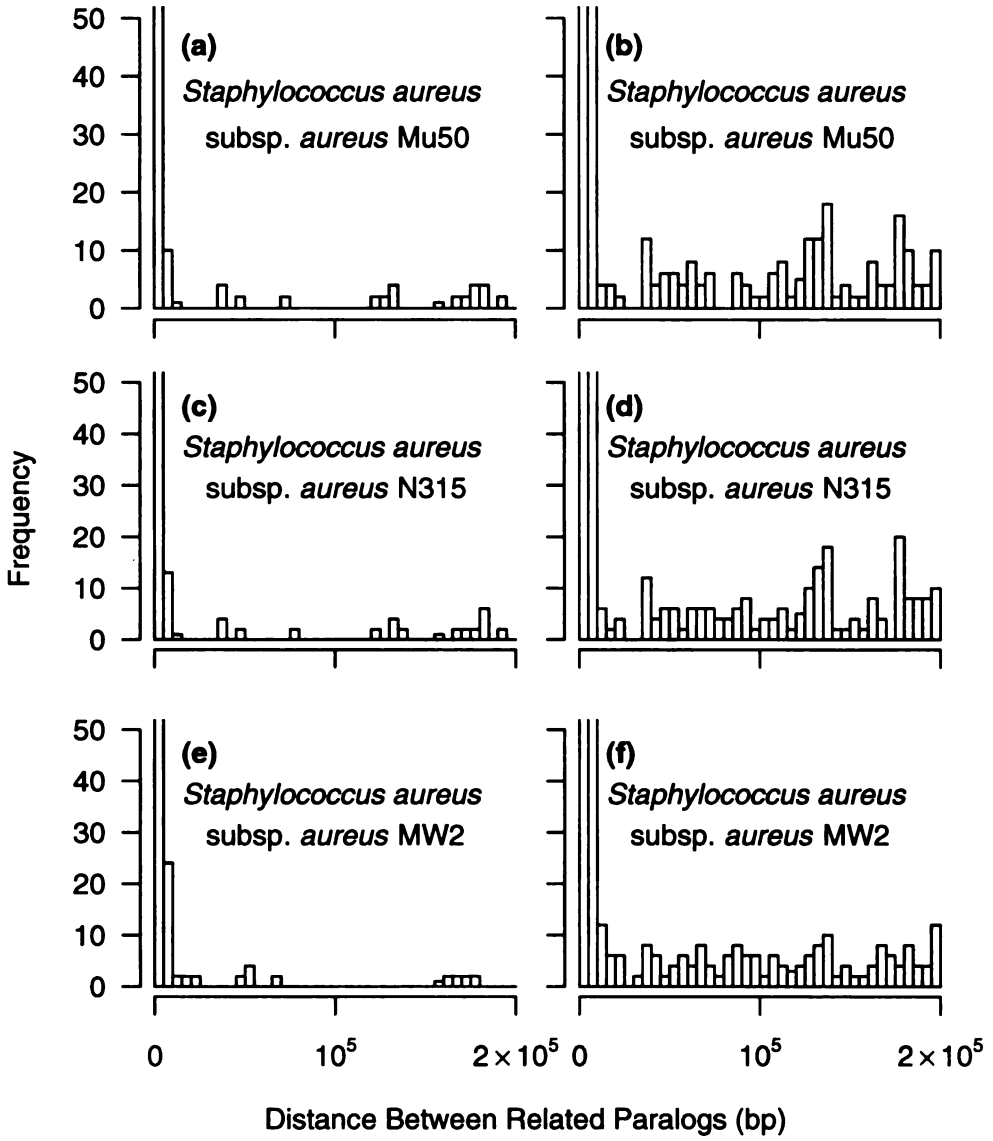

Figure 41: Frequency of distances between related paralogs for three strains of *Staphylococcus aureus*. Only those distances < 200,000 base pairs are shown. Bin sizes are 5,000 bp. **(a,c,e)** show the distances between all pairs of paralogs based on similarity clusters involving ORFs on the same strain's chromosome where $6 \leq S \leq 40$. **(b,d,f)** are based on data from Pushker *et al.* (2004). Frequency values higher than 50 are truncated on the plot and range from 56-336.

4.3 Findings for an Operational ORF Subset

4.3.1 Differences in Length and Similarity Cluster Size

Lognormal, multimodal ORF length distributions have been previously reported (Skovgaard *et al.*, 2001). As would be expected for a theoretical fit to a normal distribution of log-transformed ORF lengths, I investigated how similar the sample mean is to the sample median (Table 11). The S-ORF and O-ORF subsets (n=140,805 and n=306,746) had a stronger theoretical fit to a lognormal model than the entire set of ORFs (n=447,551). The difference in median and mean values for the S-ORF subset was consistent with a distribution slightly skewed to the left, and the difference in median and mean values for the O-ORF subset was consistent with a distribution skewed to the right. The observed ranges of ORF length medians across the taxonomic groupings were 246-286 aa (all ORFs), 126-187 aa (S-ORFs), and 291-329 aa (O-ORFs) Although the median values for O-ORF lengths were almost twice that of S-ORF lengths, the variance of lengths produced a significant overlap between the two ORF length distributions as shown in Fig. 42.

The relative numbers of ORFs, S-ORFs, O-ORFs and U-ORFs for each taxonomic grouping are shown in Table 12. Proportionally, the five subsamplings of O-ORFs ranged between 54.5% (Archaea) to 82.0% (Enterobacteriales). The five subsamplings of S-ORFs ranged between 18.0% (Enterobacteriales) to 45.5% (Archaea). These ranges broadly encompass a predicted 3:1 ratio of O-ORFs to S-ORFs. The Archaea had the lowest relative percentage of U-ORFs. Proportional trends for O-ORFs, S-ORFs, and U-ORFs are further characterized by Fig. 43. The Enterobacteriales set had the second greatest number of representative ORFs in the set of 447,551 ORFs and had the highest proportional amount of O-ORFs (82.0%), possibly due to the large number of Enterobacteriales genomes present in the data set acting by relation to elevate the similarity cluster sizes specifying the O-ORF subset. Yet, the Enterobacteriales O-ORF set shows a similar trend of length perhaps indicating that the B and L thresholds compensate for over-representation of the Enterobacteriales taxon. The taxonomic grouping with the highest proportional number of S-ORFs is the Archaea.

Table 11: O-ORF and S-ORF comparison of ORF length norms for 6 taxonomic groupings.[a]

	All ORFs		S-ORFs		O-ORFs	
Taxonomic Group	Mdn	Lnm.Mean	Mdn	Lnm.Mean	Mdn	Lnm.Mean
All 155	265	245 (-7%)	157	160 (+2%)	310	299 (-4%)
Actinobacteria	286	269 (-6%)	187	189 (+1%)	329	324 (-2%)
Archaea	246	232 (-6%)	169	175 (+4%)	310	294 (-5%)
Enterobacteriales	262	244 (-7%)	134	140 (+4%)	291	275 (-5%)
Gam. no Ent.	271	250 (-8%)	153	152 (0%)	315	307 (-3%)
Lactobacillales	254	234 (-8%)	126	133 (+6%)	289	281 (-3%)

[a]All 155 = all of the 155 genomes. 5 taxonomic subsamplings were taken from this set of 155 genomes. Gam. no Ent. = Gammaproteobacteria without Enterobacteriales. The two statistical norms for ORF lengths are a median (Mdn), and a lognormal mean (Lnm.Mean; the exponential function of the mean of the logarithm-transformed ORF lengths). The percentage increase or decrease from the median to the lognormal mean is indicated. ORF length units are in the number of corresponding amino acids to their translated product.

Table 12: ORF subset comparisons of ORF counts for 6 taxonomic groupings.[a]

	All S		S-ORFs		O-ORFs		U-ORFs	
Tax. Grp.	#ORFs	Rel%	#ORFs	Rel%	#ORFs	Rel%	#ORFs	Rel%
All 155	447,551	100%	140,805	31.5%	306,746	68.5%	202,697	45.3%
Actino.	44,866	100%	15,267	34.0%	29,599	65.9%	19,131	42.5%
Archaea	37,438	100%	17,043	45.5%	20,395	54.5%	10,866	29.0%
Enterobact.	57,115	100%	10,262	18.0%	46,853	82.0%	29,982	52.5%
Gam.no.E.	59,538	100%	17,261	29.1%	42,277	71.1%	29,175	49.1%
Lacto.	27,847	100%	6,857	24.6%	20,990	75.4%	14,333	51.5%

[a]All 155 = all of the 155 genomes. 5 taxonomic subsamplings were taken from this set of 155 genomes. The taxonomic groupings (Tax. Grp.) correspond to those in Table 11. For each taxonomic grouping and ORF subset, both the number of ORFs (# ORFs) and the relative percentage (Rel%) to the total taxonomic grouping ORF count are shown. The U-ORF subset of ORFs, belonging to similarity clusters > 40 ORFs, is also evaluated in addition to the S-ORF and O-ORF subsets.

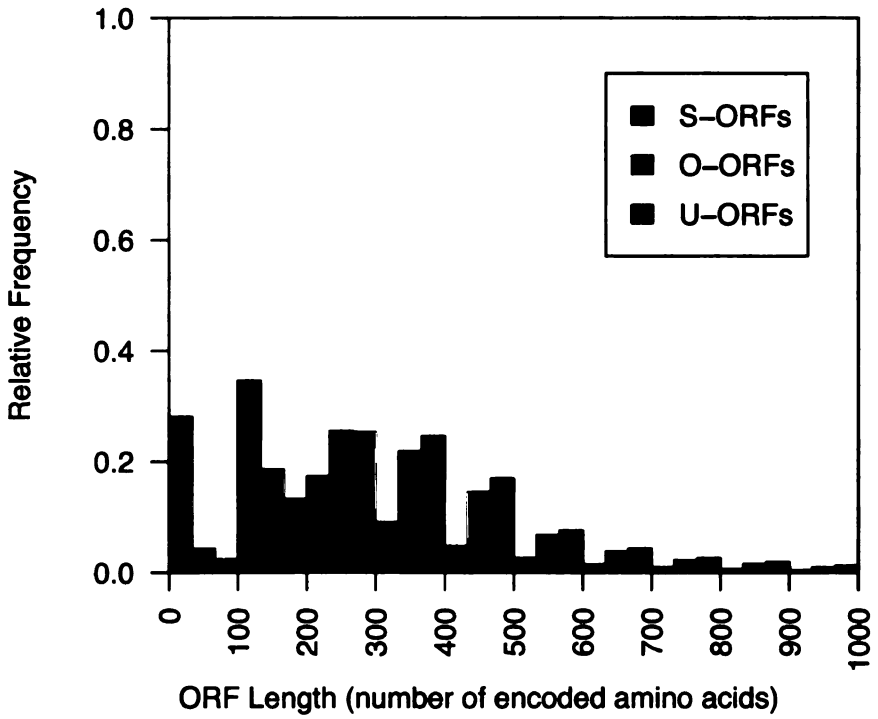

Figure 42: Relative frequency distributions of S-ORF, O-ORF, and U-ORF lengths for 155 genomes. Only the subset of ORFs that are $\leq$ 1000 aa in length is presented (441,040 out of 447,551 ORFs). The x-axis is labelled with the boundaries of each bin (bin size = 100 aa).

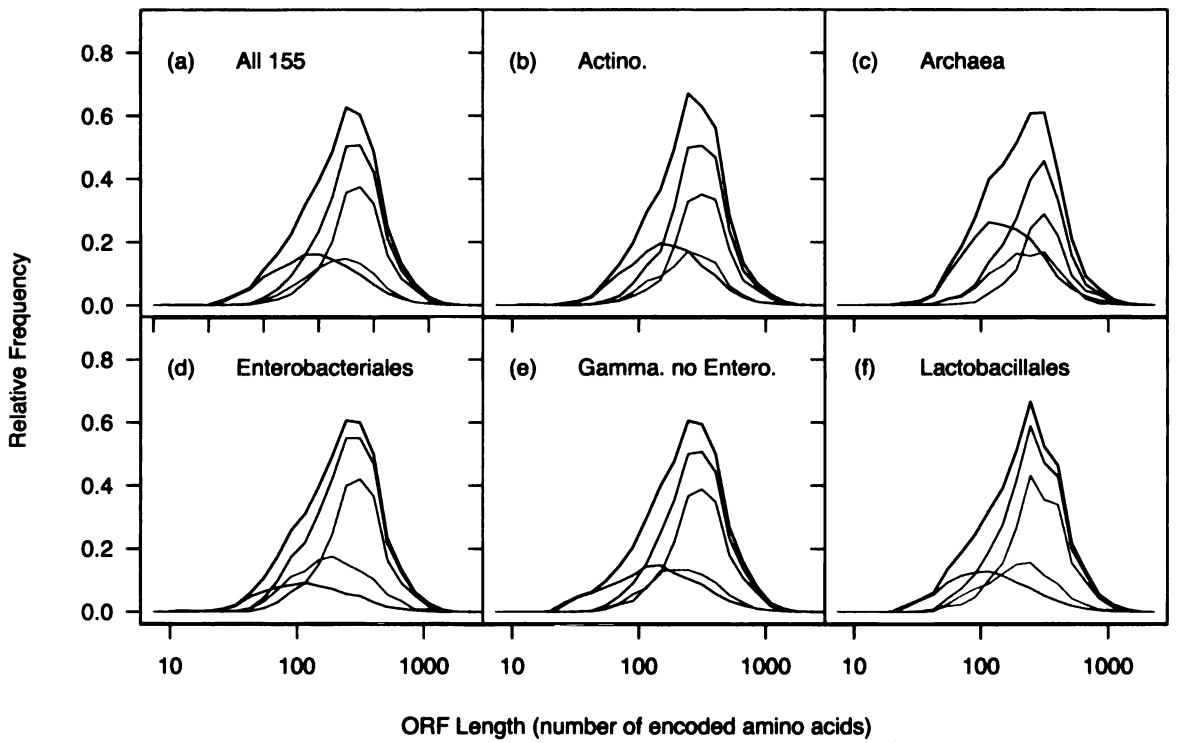

Figure 43: Frequency distributions of logarithm-transformed ORF lengths for various samplings of genomes. (a) All 155 genomes; (b) Actino. = Actinobacteria; (c) Archaea; (d) Enterobacteriales; (e) Gamma. no Entero. = Gammaproteobacteria without Enterobacteriales; (f) Lactobacillales. The x-axis is labelled with a transformed power of 10 scaling. Only the subset of ORFs that are $\leq 2,000$ aa in length is presented (447,009 out of 447,551 ORFs). A natural logarithm was calculated for each of the ORF lengths, and the log-transformed values of ORF length were aggregated into various histogram with bin sizes of 0.25. Black represents the normalized distribution of all ORFs for the given sampling of genomes. Red (rightmost curve) is the set of S-ORFs. Purple is the set of N-ORFs. Green is the set of U-ORFs. Blue is the set of O-ORFs. The red, green, purple, and blue distributions are proportional to the total number of ORFs indicated by the black curve.

ORFs occur within various ranges of similarity cluster sizes more tightly bounded than $S \geq 6$ (Fig. 44). I found that as the ORFs were evaluated for the discrete set of similarity cluster size intervals, $S = \{1, 2, 3, ..., 39, 40\}$, the S-ORF versus O-ORF distinction appeared to separate two regimes of variation (Fig. 45, 46 and 47).

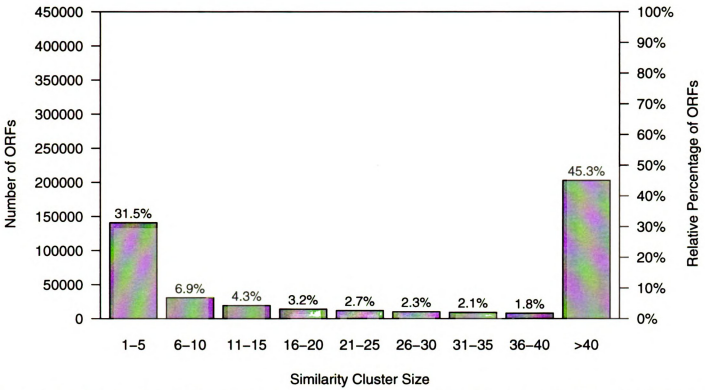

Figure 44: Number of ORFs associated with various ranges of similarity cluster sizes. The number of ORFs were tallied for various non-overlapping ranges of similarity cluster sizes. Similarity cluster size ranges were: 1-5, 6-10, 11-15, 16-20, 21-25, 26-30, 31-35, 36-40, $>$ 40 (40+). The lowest similarity cluster size, 1, means that the ORF does not meet similarity criteria in comparison to any of the other 447,551 ORFs.

The relationship between median ORF lengths and associated similarity cluster size is shown in Fig. 45-46. For the S-ORFs, there was a steady rise of median ORF lengths from a lower bound of 100 aa to values ranging from 160-250 aa. The changes in median ORF lengths were variable between different subsampled sets from the 155 genomes. For S values between 1 and 5, the Actinobacteria rise from 160 aa to 250 aa. The Enterobacteriales rise from about 100 aa to 160 aa. Except for the Actinobacteria, the median ORF lengths appeared to reach 250 aa for $S \geq 20$. For O-ORFs, there is a less steep ascent of median ORF length that proceeds from values greater than 200 aa to values greater than 280 aa. The O-ORF ascent in median ORF length is somewhat continuous, and had various rises and falls of 50 aa to 100 aa in magnitude occurring for differential changes in the similarity cluster size of ≈ 5.

A log-log relationship accounted for how the number of ORFs equates to increasing range intervals of S as shown in Fig. 47 where the range intervals of S are [1,1], [1,2], [1,3], ..., [1,40]. The second value for each interval is the similarity cluster size limit, c. The most inclusive S range of [1,40] ($c = 40$) includes all those ORFs with $1 \leq S \leq 40$, but not $S > 40$. The logarithm of the number of ORFs characterized by $[1, c]$ was directly proportional to the logarithm of c. The slopes and correlations of three linear fits were calculated for various ranges of the similarity cluster size limit ($c \rightarrow [1, 40]$, $c \rightarrow [1, 5]$, and $c \rightarrow [6, 40]$), and in all comparisons, the slope for $c \rightarrow [1, 5]$ is 23% to 100% steeper than the slope for $c \rightarrow [6, 40]$. The largest distinctions between $c \rightarrow [1, 5]$ and $c \rightarrow [6, 40]$ were for the Actinobacteria and the Archaea. For all taxonomic groupings examined, the linear fits associated with $c \rightarrow [1, 5]$ and $c \rightarrow [6, 40]$ had the strongest correlations with a linear fit, although all of the fitted lines significantly accounted for variation ($r^2 > 0.97$).

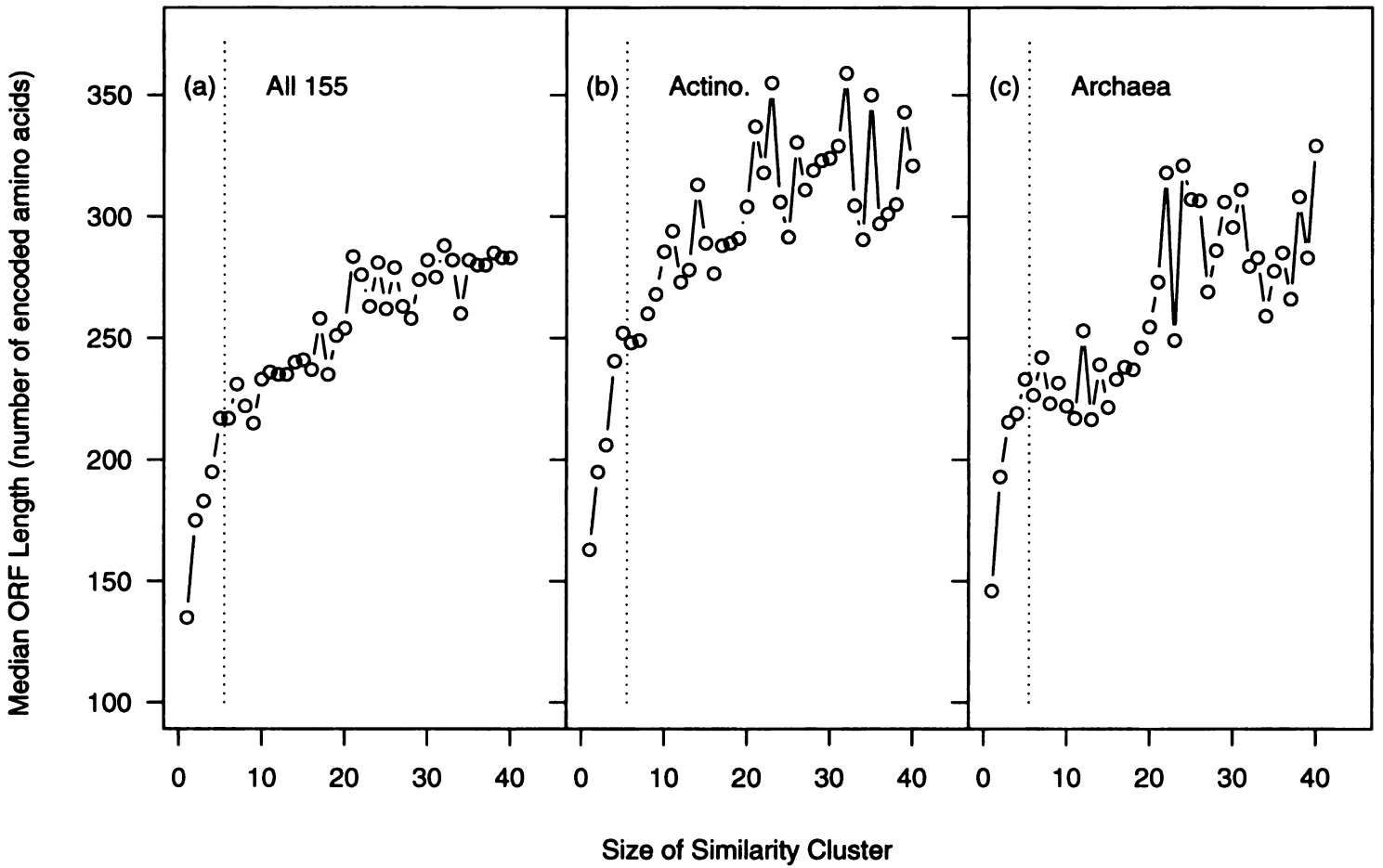

Figure 45: Relationship between ORF length and ORF similarity: (a) All 155 genomes, n=244,854; (b) Actino. = Actinobacteria, n=25,735; (c) Archaea, n=26,572. Range of similarity cluster sizes is 1 to 40 (> 40 not shown). Similarity cluster size is the number of ORFs in the set of 447,551 ORFs with a similar length and sequence. Each dot is the median value for corresponding ORF lengths.

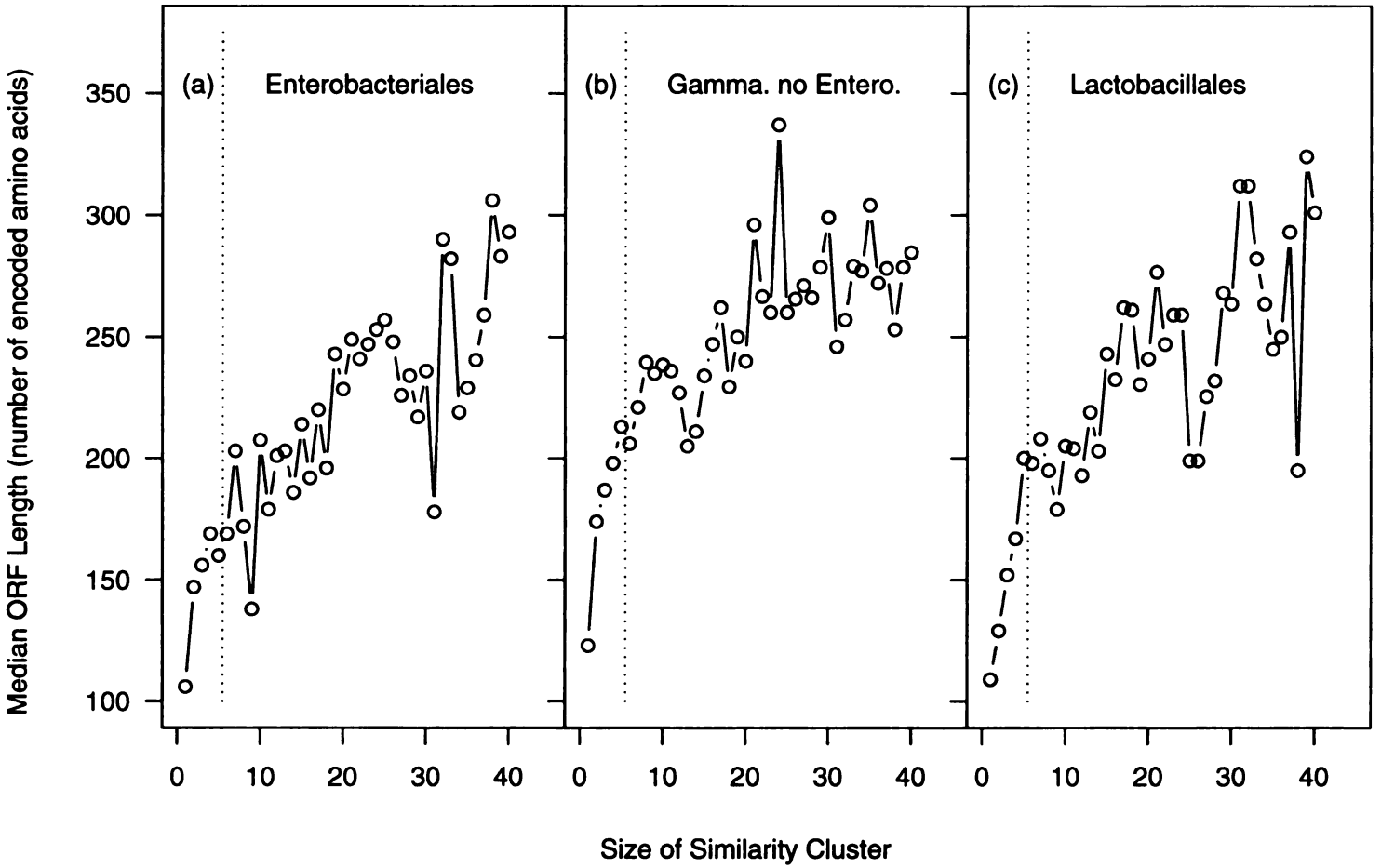

Figure 46: Relationship between ORF length and ORF similarity: (a) Enterobacteriales, n=27,133; (b) Gamma. no Entero. = Gammaproteobacteria without Enterobacteriales, n=30,363; (c) Lactobacillales, n=13,514. Range of similarity cluster sizes is 1 to 40 (> 40 not shown). Similarity cluster size is the number of ORFs in the set of 447,551 ORFs with a similar length and sequence. Each dot is the median value for corresponding ORF lengths.

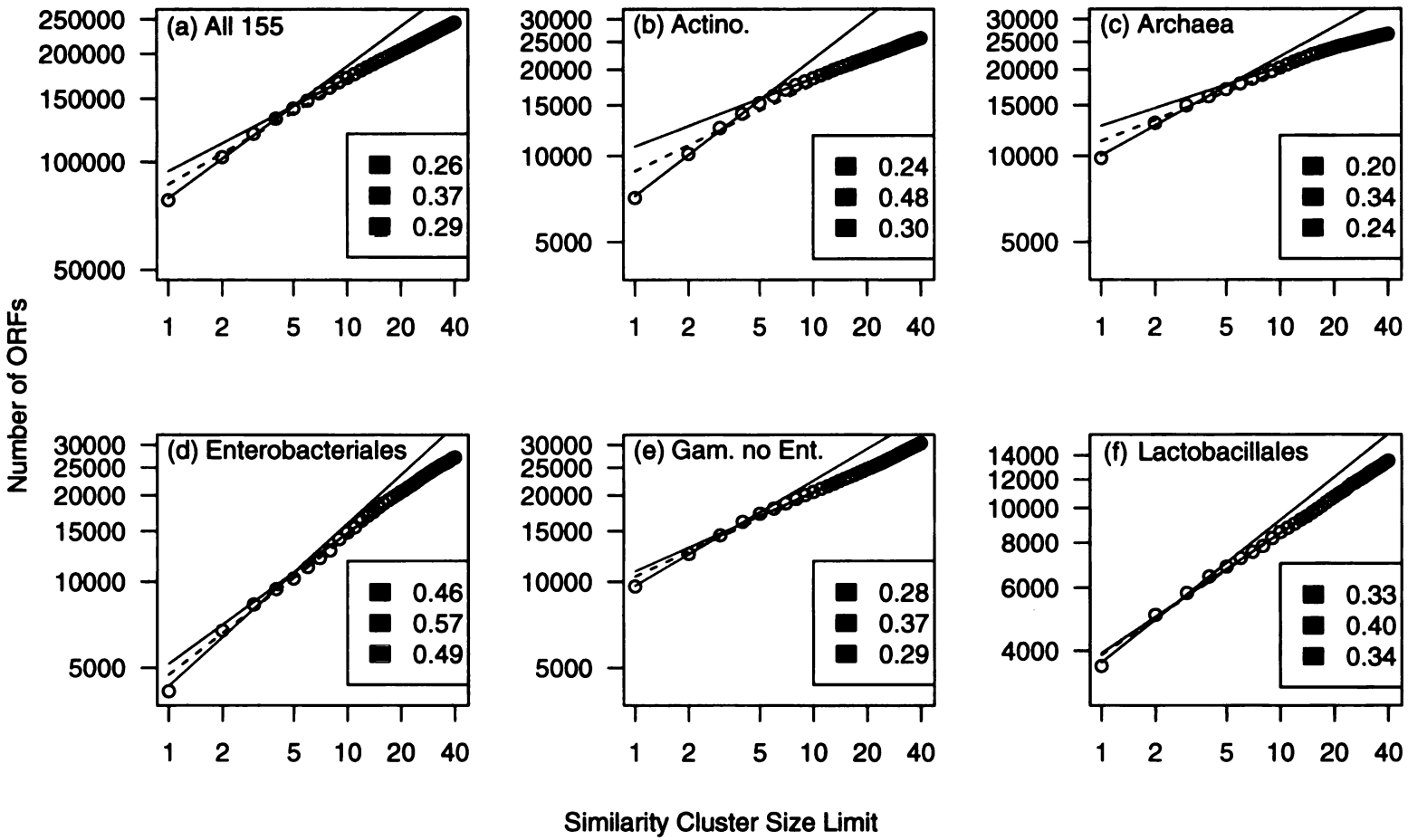

Figure 47: Log-log frequency distributions of number of ORFs for similarity cluster size limits. Each dot corresponds to the number of ORFs that are less than or equal to similarity cluster size limits ranging from 1 to 40. ORFs belonging to similarity cluster sizes greater than 40 are not evaluated. Linear relationships are drawn, and slopes indicated, for cluster size limit ranges: 1-5 (red), 6-40 (blue), and 1-40 (black; dotted line). 6 taxonomic groups are evaluated. (a) All 155 = All 155 genomes; (b) Actino. = Actinobacteria; (c) Archaea; (d) Enterobacteriales; (e) Gam. no Ent. = Gammaproteobacteria without Enterobacteriales; (f) Lactobacillales. The axes are log-transformed.

I inspected the linear relationship between observed distributions and expected distributions (as calculated from sample means and standard deviations) for the log-transformed lengths of various subsets of ORFs (Fig. 48-50). Fig. 49 presents positive skewness values ($w > 0$), implicating a non-normal leftwards shift to the distribution of S-ORF lengths. By contrast, the ORF set and O-ORF subset had negative skewness values ($w < 0$), implicating a non-normal rightwards shift to the distribution of ORF lengths. These directional shifts were consistent with my findings for the arithmetic mean's relationship to the median (Table 11).

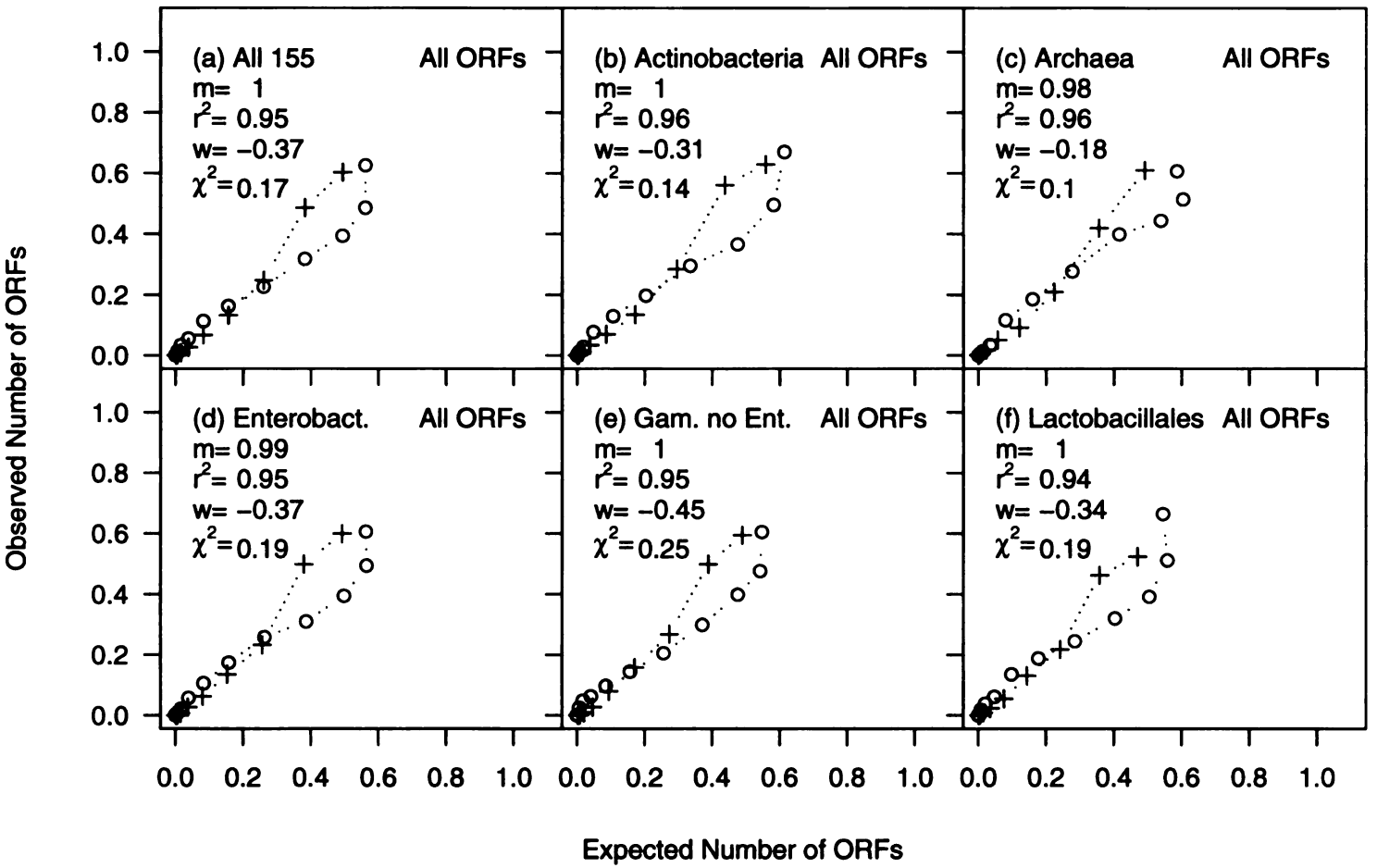

Figure 48: Normality of log-transformed ORF lengths. All 155 = All 155 genomes. Enterobact. = Enterobacteriales. Gam. no Ent. = Gammaproteobacteria without Enterobacteriales. Only those ORFs $\leq$ 2000 aa are evaluated. Log lengths of ORFs are computed, and a relative frequency distribution generated from 2 to 8 with a bin size of 0.25 to be the observed values. The mean and s.d. generate an expected distribution of ORF counts. Listed values are m for slope, r^2, w for skewness, and χ^2. The circles are scatter plot comparisons to the left of the median log ORF length value. The plus-signs are comparisons to the right of the median.

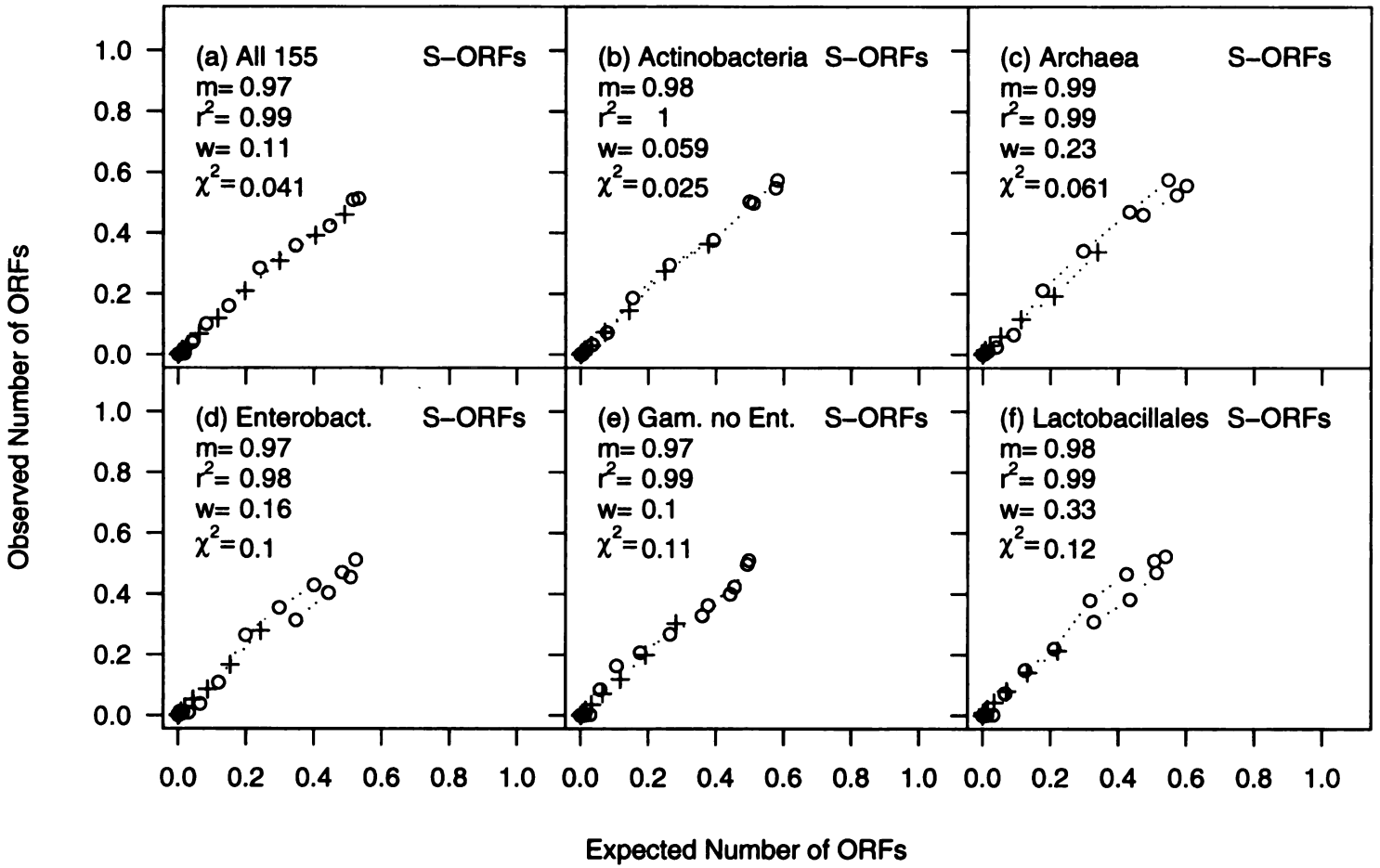

Figure 49: Normality of log-transformed S-ORF lengths. All 155 = All 155 genomes. Enterobact. = Enterobacteriales. Gam. no Ent. = Gammaproteobacteria without Enterobacteriales. Only those ORFs $\leq$ 2000 aa are evaluated. Log lengths of ORFs are computed, and a relative frequency distribution generated from 2 to 8 with a bin size of 0.25 to be the observed values. The mean and s.d. generate an expected distribution of ORF counts. Listed values are m for slope, r^2, w for skewness, and χ^2. The circles are scatter plot comparisons to the left of the median log ORF length value. The plus-signs are comparisons to the right of the median.

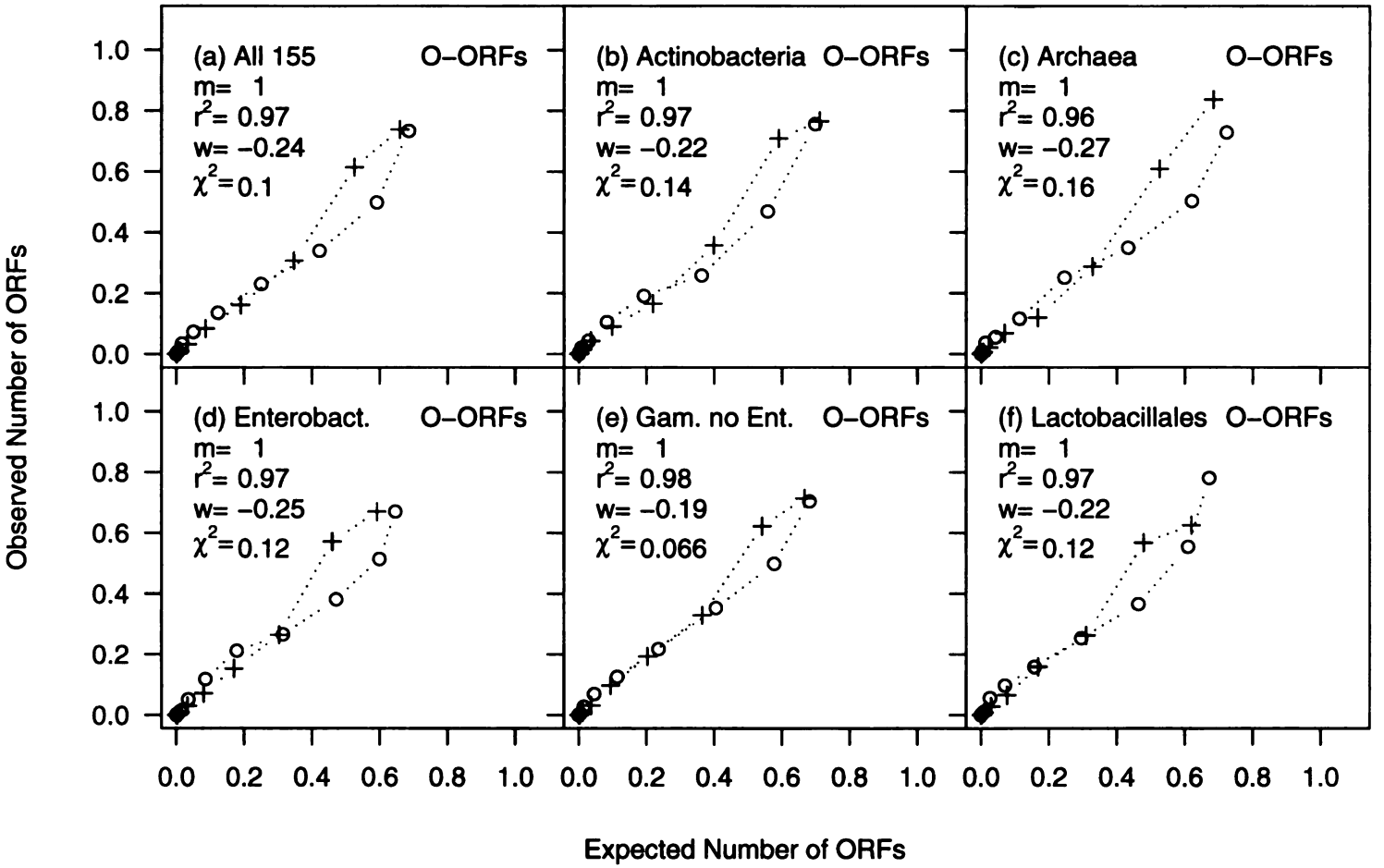

Figure 50: Normality of log-transformed O-ORF lengths. All 155 = All 155 genomes. Enterobact. = Enterobacteriales. Gam. no Ent. = Gammaproteobacteria without Enterobacteriales. Only those ORFs $\leq$ 2000 aa are evaluated. Log lengths of ORFs are computed, and a relative frequency distribution generated from 2 to 8 with a bin size of 0.25 to be the observed values. The mean and s.d. generate an expected distribution of ORF counts. Listed values are m for slope, r^2, w for skewness and χ^2. The circles are scatter plot comparisons to the left of the median log ORF length value. The plus-signs are comparisons to the right of the median.

4.3.2 Composition of Phyla within ORF Similarity Clusters

I investigated the number of ORFs in each similarity cluster belonging to the same lineage based on membership within one or all of 7 different phyla (Fig. 51 and 52). For each analysis, a random selection of 1000 ORFs came from the phylum (or 7 phyla) under evalation. The Proteobacteria and Firmicutes had the highest degree of similarity clusters of sizes 6 to 15 containing members within the phyla. Yet, at least 25% of Proteobacteria and Firmicutes similarity clusters tended to have non-phylum members for $S \geq 6$.

4.3.3 Expressional, Phenotypic, and Functional Aspects of the ORF Subsets

I evaluated expressional, functional, and phenotypic aspects of the O-ORF and S-ORF subsets to more fully assess their empirical correspondence with a distinction of real ORFs versus unreal ORFs.

I assessed published transcriptional data for 3,309 ORFs of *Escherichia coli* K-12 (Covert *et al.*, 2004) and compared patterns of presence or absence of transcriptional expression to the similarity cluster size stored in the MYCROW database. In this set of 3,309 ORFs, there were 2,112 U-ORFs (64%), 290 S-ORFs (9%), 3,019 O-ORFs (91%), and 2,787 C-ORFs (84%). By comparison, for the total set of annotated ORFs for *E. coli* K-12 (n=4,311), there were 2,388 U-ORFs (55%), 513 S-ORFs (12%), 3,798 O-ORFs (88%), and 3,153 C-ORFs (73%).

For the 42 separate assays of transcriptional expression on the set of 3,309 ORFs, 1,992 ORFs were designated "present" for all 42 transcriptional assays of expression, 331 ORFs were designated "absent" for all 42 transcriptional assays of expression, and 780 ORFs had marginal or conflicting expression designations of presence and absence among the 42 transcriptional assays. Of the 2,323 ORFs that are either uniformly present or absent across the 42 assays of transcription, 86% are transcriptionally present compared to 14% that are absent. Of the 1,992 transcriptionally present ORFs, there is a ratio of 22:3 for ORFs belonging to a COG and a 24:1 ratio for ORFs belonging to the O-ORF subset. Of the 331 transcriptionally absent ORFs, there is a 1:3 ratio for ORFs belonging to the O-ORF subset

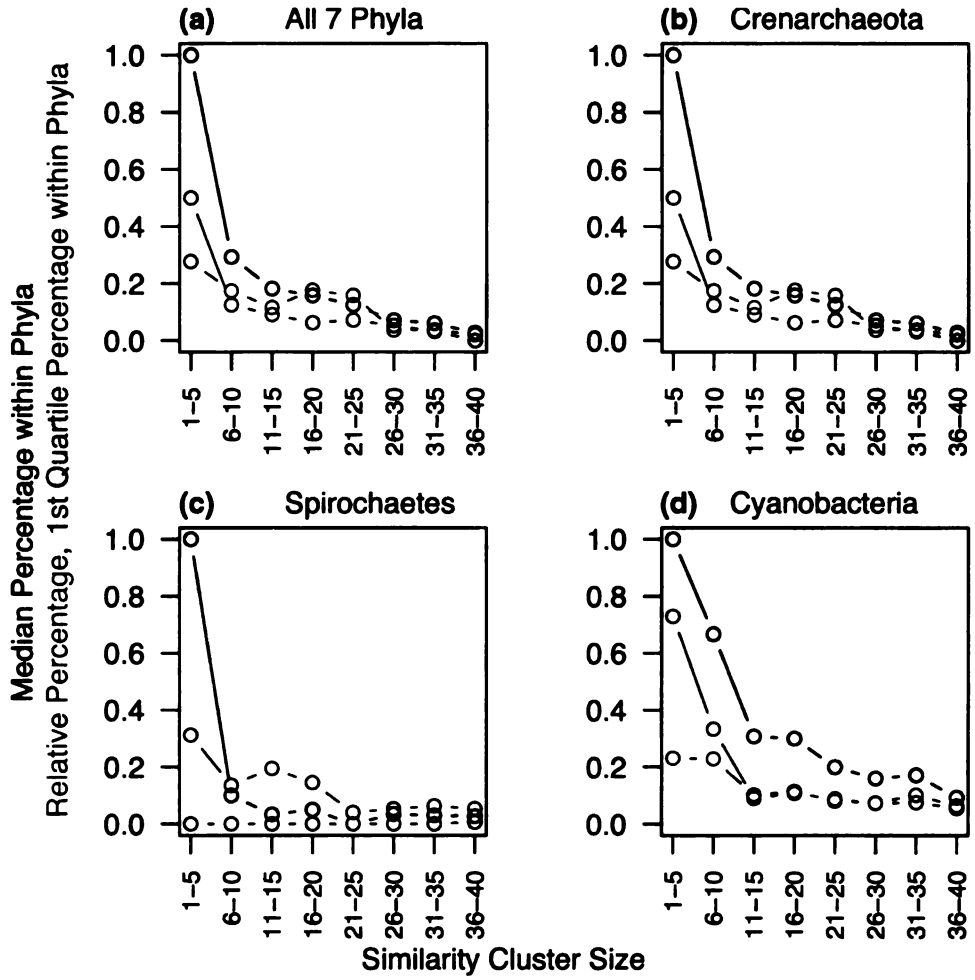

Figure 51: Membership within phyla for similarity clusters. All 7 phyla (Actinobacteria, Crenarchaeota, Cyanobacteria, Euryarchaeota, Firmicutes, Proteobacteria, Spirochaetes; 148 genomes) and, separately, the three phyla, Crenarchaeota (4 genomes), Spirochaetes (5 genomes), and Cyanobacteria (genomes) are examined. Each plot is based on 1000 randomly selected ORFs and their corresponding similarity clusters. Black: median percentage membership of similarity cluster belonging to the same phylum. Blue: the first quartile of percentage memberships for same phylum. Red: the relative percentage of ORFs belonging to each range of similarity cluster sizes.

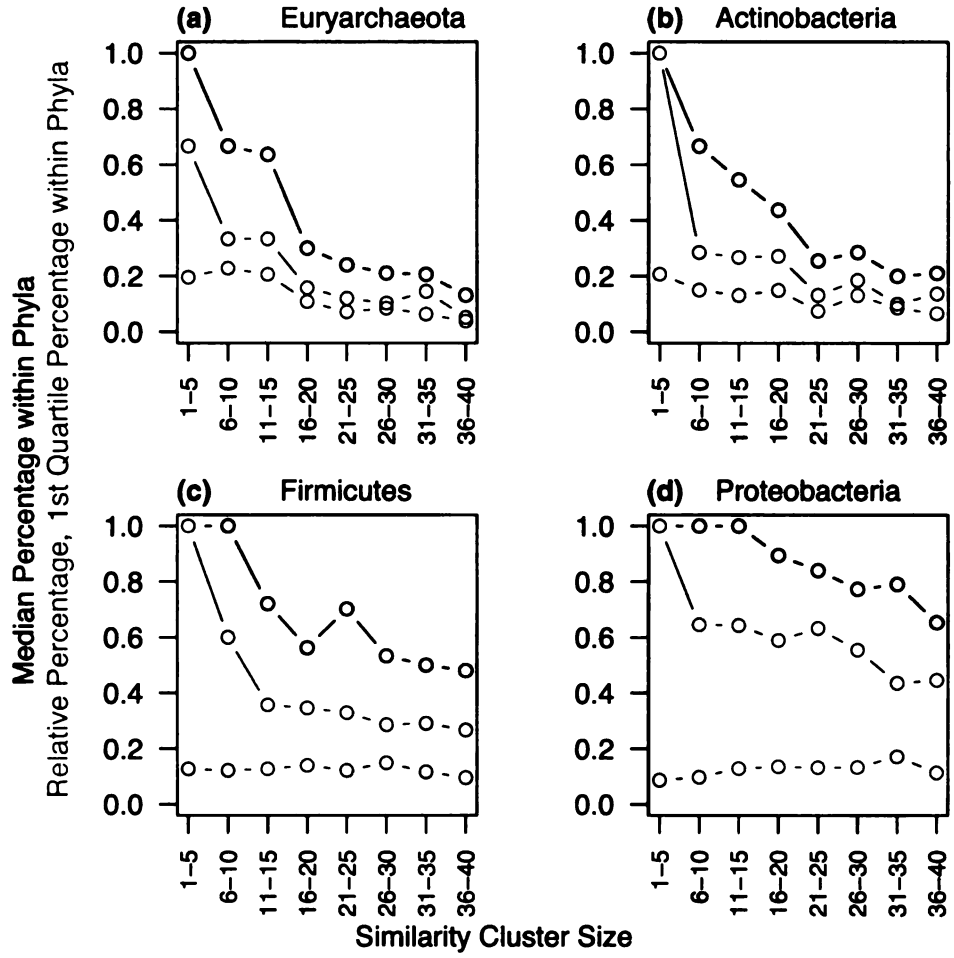

Figure 52: Membership within four other phyla for similarity clusters. Four separately considered phyla: Euryarchaeota (12 genomes), Actinobacteria (13 genomes), Firmicutes (38 genomes), and Proteobacteria (68 genomes). Each plot is based on 1000 randomly selected ORFs and their corresponding similarity clusters. Black: median percentage membership of similarity cluster belonging to the same phylum. Blue: the first quartile of percentage memberships for same phylum. Red: the relative percentage of ORFs belonging to each range of similarity cluster sizes.

and a 3:7 ratio of ORFs belonging to a COG. Both in terms of proportional categorization (86% versus 91%) and ratios of association (24:1 and 1:3), the O-ORF versus S-ORF subsets were more closely aligned with expectations for transcriptional expression for *E. coli* K-12 than a COG-based distinction.

Covert *et al.* (2004) also propose various ORFs as having functional and regulatory importance based on growth and no growth predictions for 143 types of media conditions such as "growth on citric acid", "growth on methionine", and "growth on adenosine." Of the ORFs (99 of 110) that were unambiguously mapped to ORFs within the MYCROW database, 98 (99%) of these ORFs were O-ORFs and 80 (80%) were U-ORFs. Only one ORF was an S-ORF and it belonged to a similarity cluster size of 5.

Table 13 shows the association between number of phenotype effects and similarity cluster size S based on data for *Bacillus subtilis* (Biaudet *et al.*, 1997). Of the 352 genes that only have a single phenotypic effect when mutated (66%), 28% of them were S-ORFs. Of the 181 genes that have two or more phenotypic effects (34%), only 19% of them were S-ORFs. 261 of the 533 ORFs were U-ORFs (49%) and 140 ORFs were N-ORFs (26%). 73% of S-ORFs had a single phenotype effect, whereas only 63% of N-ORFs and 64% of U-ORFs had a single phenotype effect. Quadruple and sextuple phenotype effects occurred exclusively for O-ORFs. Overall, multiple phenotypic effects were found to be more closely associated with O-ORFs than S-ORFs, although inactivation of S-ORFs is generally associated with a phenotypic consequence.

Also, as seen in Table 14, examination of data for *Bacillus subtilis* (Biaudet *et al.*, 1997) showed an increase in "single phenotype" ORFs for S-ORFs compared to ORFs that are not in COGs, and a proportionately greater amount of "multiple phenotype" ORFs for O-ORFs compared to C-ORFs. The evidence, while not exhaustive, is consistent with the O-ORF versus S-ORF distinction relating to whether or not an ORF is expressed and whether or not there is significant operational consequence to the organism's physiology.

For functional grouping of ORFs, the functional classification of COGs was evaluated based on NCBI's COG database that contained data for 67 (41%) of the 165 chromosomes. For this sampling of 67 chromosomes and their total numbers of ORFs, the mean and median levels of O-ORFs were both equal to 68% whereas the mean and median levels of C-ORFs

Table 13: Phenotypic inactivation associated with *Bacillus subtilis* O-ORFs and S-ORFs.[a]

	All ORFs		S-ORFs		O-ORFs	
Number of Affected Phenotypes	#ORFs	Rel.%	#ORFs	Rel.%	#ORFs	Rel.%
1 or more	533	100.0%	132	24.8%	401	75.2%
2 or more	181	100.0%	35	19.3%	146	80.7%
3 or more	47	100.0%	10	21.3%	37	78.7%
4 or more	13	100.0%	0	0.0%	13	100.0%

[a]533 ORFs, when mutated, ranged from single phenotype effects to six phenotypic effects. The number of ORFs (# ORFs) is shown as well as the relative percentage (Rel.%) that number of ORFs to the total sample of ORFs for the given range of phenotypic effects.

Table 14: ORF counts from *Bacillus subtilis* for single and multiple phenotypes based on COG and O-ORF categorizations.

	S-ORF	X-ORF	O-ORF	C-ORF
Single Phenotype	97	84	255	268
Multiple Phenotypes	35	32	146	148

were both equal to 74% (based on 25 functional classes). An expectedly stronger association was observed between O-ORFs and C-ORFs (median, 64%) compared to O-ORFs and X-ORFs (median 4%). An expectedly stronger association was also observed between S-ORFs and X-ORFs (median, 20%) compared to S-ORFs and C-ORFs (median, 11%). The "General function prediction only" and "Function unknown" subsets of COGs amounted to about 11% and 7% respectively of 161,990 ORFs for 67 chromosomes.

Linear modelling, as shown in Fig. 53 and 54, further investigated how categories of COG and O-ORF membership scale with comparison to the total count of ORFs for a given replicon. In these models, the proportional measure of O-ORFs to total ORFs was 73% and, for COGs, 72%. The strongest linear associations were for the subset of ORFs that are jointly both O-ORFs and C-ORFs, and the separately considered O-ORF and C-ORF subsets. The next strongest linear association was for the X-ORFs. Although still accounting for most of the original variability ($r^2 > .5$), both the S-ORF subset and the joint intersection subset of the S-ORF and C-ORF subsets showed markedly reduced linear correlation coefficients,

suggesting that the number of S-ORFs is less likely to relate directly to the total ORF count. The scattering of dots away from the fitted line in Fig. 53c and 53c-d occurs between 2,000 annotated ORFs and 2,500 annotated ORFs. Based on a relationship of one ORF for every 1,100 base pairs of chromosomal DNA, the increased pattern of scattering attributable to S-ORFs likely occurs for chromosome sizes > 2 Mb, corresponding to a a proposed distinction of microbial ecology and genomic stability (Fig. 36) (Ochman & Davalos, 2006). Further measurement showed the canonical correlation, r^2, to be different for ranges of total ORF counts $< 2,000$ compared to $> 2,000$. The r^2 value for the S-ORF subset count where total ORF count $< 2,000$ is 0.58 compared to 0.43 for the X-ORF subset. For total ORF counts $> 2,000$, r^2 for the S-ORF subset count is 0.39 compared to 0.65 for the X-ORF subset.

As ORF sets are examined by 25 different functional COG categories, Fig. 55 shows there to be close correspondence between the number of O-ORFs and overall number of ORFs for a given COG category. The greatest variation appears to be for the COG categories of "general function prediction only" and "function unknown" where the number of corresponding O-ORFs drops, with an inverse rise in the number of S-ORFs.

Table 15 shows how the percentage amounts of S-ORFs can vary for different functional COG categories across different subsets of the genomes. Most ORFs that are not in a COG are S-ORFs (77%). For most all categories, the Archaea have the highest percentage of S-ORFs. The Actinobacteria generally have the second highest percentage association of S-ORFs with functional COG categories except for the categories of cell motility (N) and secondary metabolites biosynthesis, transport and metabolism (Q). If S-ORFs truly mean "not operational", the lower values of Actinobacteria S-ORFs for categories N and Q is consistent with the soil lifestyle (Garrity, 2001).

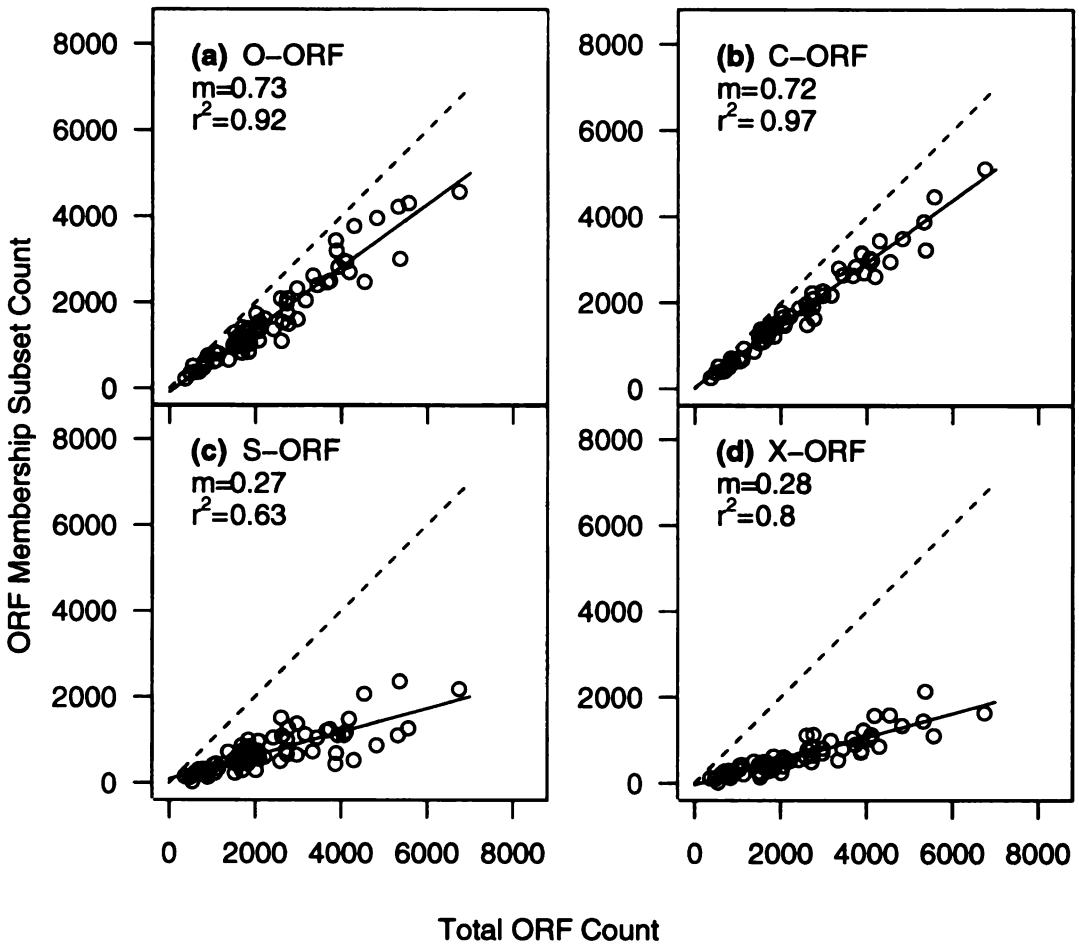

Figure 53: ORF membership subset comparisons with total ORF counts. Each of the four plots shows the relative proportions of the O-ORF, C-ORF, S-ORF, and X-ORF subset ORF counts, where each point corresponds to one of 67 genomes. A fitted line is shown along with slope (m) and r^2.

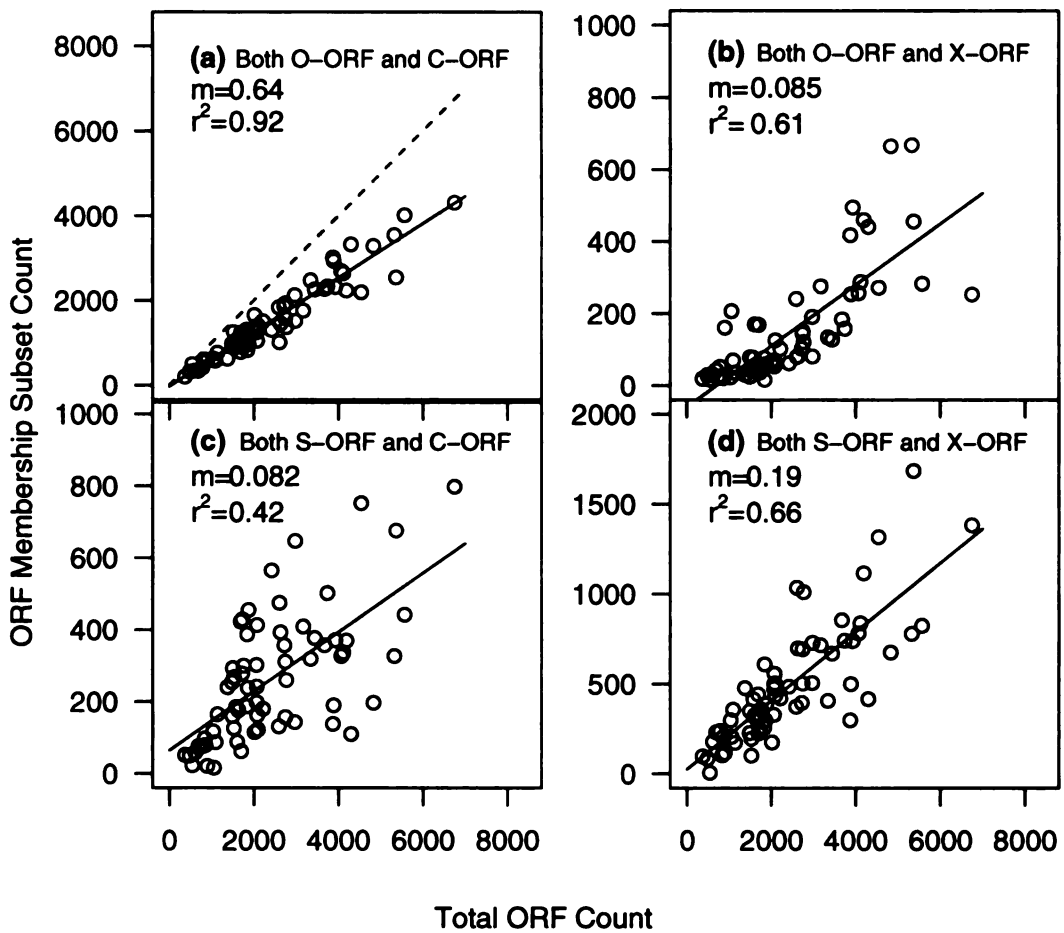

Figure 54: ORF membership intersecting subset comparisons with total ORF counts. Each of the four plots shows the relative proportions of various intersections of ORF subsets, where each point corresponds to one of 67 genomes. A fitted line is shown along with the slope (m) and r^2.

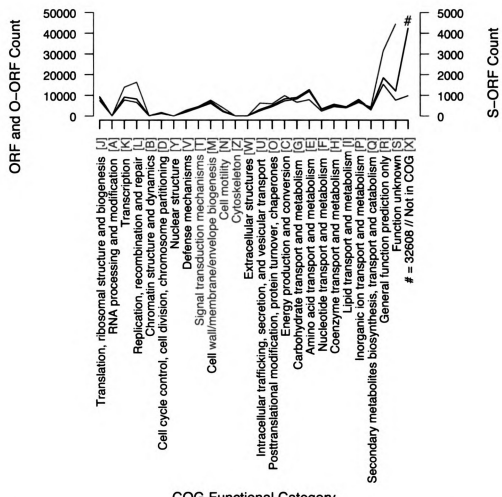

Figure 55: Number of ORFs, O-ORFs, and S-ORFs for various COG-based categories of function. The number of ORFs (black line) and O-ORFs (blue line) are indicated by the axis on the left. The number of S-ORFs are indicated by the axis on the right. The rightmost number of S-ORFs (32,608) for the "Not in COG" category is not plotted to scale, but abbreviated with the pound symbol. 25 different functional categories are shown.

Table 15: S-ORF percentages of ORF counts for various functional COG categories.[a]

Code	All	Acti.	Arch.	Ente.	Gamm.	Lact.	Description of Functional COG Category
J[b]	8	5	21	2	2	3	Translation, ribosomal structure and biogenesis
L[b]	20	17	39	9	12	8	Replication, recombination and repair
D[b]	13	24	15	5	11	15	Cell cycle control, cell division, chromosome partitioning
F[b]	7	7	11	1	6	3	Nucleotide transport and metabolism
K[c]	15	17	38	7	6	13	Transcription
T[c]	11	13	26	4	9	8	Signal transduction mechanisms
N[c]	16	8	33	8	15	27	Cell motility
Q[c]	13	10	17	5	11	8	Secondary metabolites biosynthesis, transport and catabolism
C[c]	12	8	23	2	7	5	Energy production and conversion
A	24	0	52	0	0	0	RNA processing and modification
B	15	0	25	0	0	0	Chromatin structure and dynamics
Y	0	0	0	0	0	0	Nuclear structure
V	13	14	13	7	5	12	Defense mechanisms
M	11	8	19	5	9	9	Cell wall/membrane/envelope biogenesis
Z	100	0	0	0	0	0	Cytoskeleton
W	4	0	100	0	0	0	Extracellular structures
U	20	23	39	10	17	17	Intracellular trafficking, secretion, and vesicular transport
O	12	8	24	2	9	7	Posttranslational modification, protein turnover, chaperones
G	7	10	14	3	4	7	Carbohydrate transport and metabolism
E	6	7	10	2	4	2	Amino acid transport and metabolism
H	8	6	17	1	4	5	Coenzyme transport and metabolism
I	9	8	16	3	8	2	Lipid transport and metabolism
P	8	12	14	3	6	4	Inorganic ion transport and metabolism
R	17	18	27	8	12	10	General function prediction only
S	37	37	61	13	26	18	Function unknown
X	77	69	89	52	80	73	Not in COG

[a]The first column has the letter codes for COG functional categories. X is an additional category for ORFs that are not in a COG. All = All 155 genomes. Acti. = Actinobacteria. Ente. = Enterobacteriales. Gamm. = Gammaproteobacteria without Enterobacteriales. Lact. = Lactobacillales.

[b]Expected to correlate negatively with genome size.

[c]Expected to correlate positively with genome size.

If the COG set includes falsely annotated ORFs, I theorize that the O-ORFs inside functional classifications of COGs should enhance the positive and negative genome size correlations characterized for various functional groupings (Bentley & Parkhill, 2004; Konstantinidis & Tiedje, 2004). Table 16 shows how the O-ORF counts for the J, L, D and F set of COG functional categories enhanced the predicted negative genome size correlation, and even more strongly enhanced the predicted positive genome size correlation associated with the K, T, N, Q, and C set of COG functional categories. The definitions of the R, S, and X groups suggest a gradient of decreasing functional evidence for their respective sets of ORFs, and the decreasing proportion of O-ORFs inside each group corresponds to this gradient.

Table 16: Percent O-ORF representation for groups of COG categories based on bacterial genomes and genome size indicators of organism lifestyle.

Genome Size	Characteristic Lifestyle	**J, L, D, and F COGs**	**K, T, N, Q, and C COGs**	**R COG**	**S COG**	**X COG**
0.5 Mb - 1.5 Mb	Obligate pathogen or symbiont	91%	86%	80%	64%	22%
2.0 Mb - 5.0 Mb	Recent or facultative host association	91%	89%	86%	70%	25%
5.0 Mb - 10 Mb	Replicate independently from host	88%	91%	86%	69%	30%

4.4 Non-Stochastic Clustering of O-ORFs

For my population of 165 chromosomes, I inspected the z-score values of significance (number of sigma σ units separating original and randomized assignments of O-ORF membership) from my running tally methodology. The O-ORF running tally z-score difference was 14.9σ. There were 28 of the 165 chromosomes (17%) that had a z-score $< 1.64\sigma$. Of the 14 of the 67 chromosomes evaluated for COG membership with COG-based z-scores $< 1.64\sigma$, 7 (50%) had O-ORF-based z-scores $< 1.64\sigma$.

4.5 Discussion

I established the parameters for distinguishing real ORFs from putative, false ORFs by general statistical expectations. As shown by Fig. 55, I found the O-ORF subset to follow trends similar to a COG membership subset of ORFs (C-ORFs). C-ORFs have been previously reported by Skovgaard *et al.* (2001) as a lower bound to the total number of annotated ORFs that correspond to real proteins. As shown by Table 17, my O-ORF specification follows a higher threshold parameter of similarity cluster size (S), and has requisite criteria for ORF length similarity (L) and sequence similarity (B). Despite significantly different approaches to threshold parameters, similar percentages of ORFs belong to the subset of O-ORFs (73%) compared to the subset of C-ORFs (72%). While the O-ORF specification involves a variety of more stringent threshold parameters, it neither imposes the orthologous bidirectional best hit criterion of COGs, nor does it require sequence conservation to exist across three distant lineages. The inclusion of paralogs and recently evolved ORFs in the similarity cluster scoring of O-ORF membership may meaningfully account for differing results for the prevalence of O-ORFs compared to C-ORFs.

Based on Fig. 54a and 54d, there are about 64% of ORFs per chromosome inside both the C-ORF and O-ORF subsets compared to 19% that are not inside either of the subsets. I expect the O-ORF specification to be better aligned with a real ORF specification versus the C-ORF specification based on greater pairwise comparison thresholds for homology, and the inclusive scoring of paralogs and recently evolved ORFs that are a likely source of functional and real ORFs (Snel *et al.*, 2002; Kurland *et al.*, 2003; Liang *et al.*, 2002; Konstantinidis &

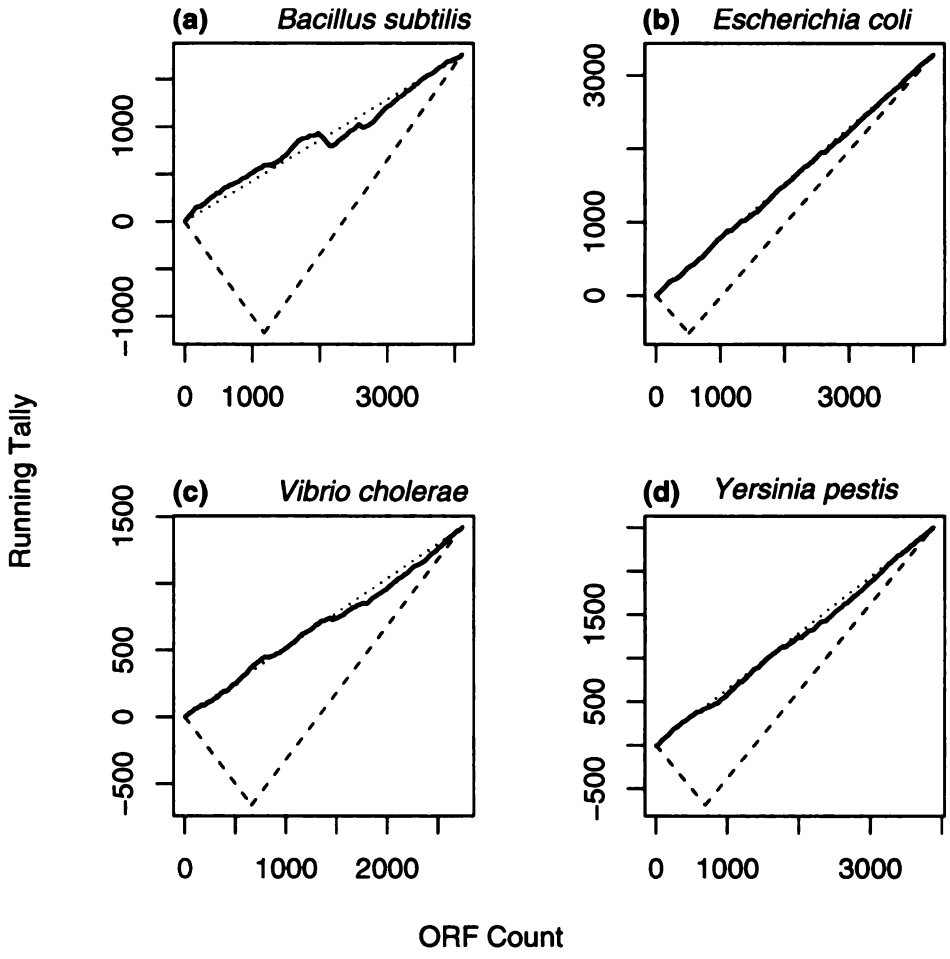

Figure 56: Running tally graphs of O-ORF membership along four chromosomes. The thick line represents increments and decrements based on whether an ORF is an O-ORF or not. The dotted diagonal represents random expectation where O-ORF membership is randomly assigned to a chromosomal set of ORFs. The dashed lines forming a V-shape represents the pattern if all S-ORFs were together followed by O-ORF members. **(a)** *Bacillus subtilis* subsp. *subtilis* str. 168. **(b)** *Escherichia coli* K-12. **(c)** *Vibrio cholerae* (large chromosome). **(d)** *Yersinia pestis* CO92.

Table 17: Threshold parameters of sequence conservation for the operational ORF subset (O-ORFs) and the COG membership subset (C-ORFs).

Threshold	O-ORF	C-ORF[a]
Length Similarity	$\pm 10\%$	$\pm 33\%$
Sequence Similarity[b]	$\leq 10^{-6}$	$< 10^{-3}$
Similarity Cluster Size	≥ 6	≥ 3

[a]Specification criteria for COGs include bidirectional best hits involving three disparate lineages and manual inspection and splitting of tentative clusters. COG analyses are not based on explicit thresholds for sequence similarity. The length similarity and sequence similarity values for COGs characterize the retrospective 90% confidence interval for how any two pairs of ORFs belonging to the same COG correspond in similarity.

[b]BLASTP expectation score for a pairwise comparison.

Tiedje, 2004). By comparison, the C-ORF specification requires distant orthologies. In my study, several analyses provided evidence that the O-ORF set is a more optimal specification for real ORFs compared to the C-ORF set. A greater proportion of O-ORFs are transcribed compared to C-ORFs. The O-ORF and S-ORF specification may also be effective for further characterizing functional groups relevant to fluctuations in genome size and associated prokaryotic lifestyles (Table 16). The S-ORF, O-ORF transition between $S = 5$ and $S = 6$ appears to be an accurate point of separation for different regimes of variation seen for 1) ORF length and similarity cluster size (Fig. 45) and 2) frequency of ORFs associated with various similarity cluster sizes (Fig. 47). Overall, the hypothesis of a coding space limit (Jackson *et al.*, 2002), where 75% of the total set of annotated ORFs would be expected to be real, is supported by two independently developed sets of parameters for O-ORFs and C-ORFs as shown by the linear relationships in Fig. 53.

Transcriptional expression data (Covert *et al.*, 2004) and data from studies of phenotypic inactivation (Biaudet *et al.*, 1997), when applied to the O-ORF and S-ORF subsets, do indicate that some of the S-ORF assignments confer a phenotype or are transcribed. Intriguingly, those ORFs with the highest number of phenotypic effects are all O-ORFs, and this may relate to a high degree of interaction with other proteins (Table 13). Protein evolution is rapid, and only the most highly interactive proteins have a slow rate of evolution (Jordan *et al.*, 2003). The ongoing fluctuation of gene loss, modification, and addition would indicate that there are some ORFs that are in the process of becoming S-ORFs or are in the

process of becoming O-ORFs. Beyond this study, a closer inspection as to the properties of gene loss versus gene addition may further characterize the natural dynamics that account for putative distinctions between real and falsely annotated ORFs. A more refined heuristic might be arrived at by formally characterizing differences between exemplar sets of ORFs with none, some, or all known features of evolutionary and functional importance.

The O-ORF specification allows for paralogy, and Fig. 38-41 help characterize the degree to which paralogy contributes to the O-ORF specification. Precise characterizations of paralogs versus non-paralogs may be difficult to arrive at as evident by conflicting estimations of paralogy for various strains (Nelson *et al.*, 2002; Pushker *et al.*, 2004; Andersson *et al.*, 1998; Simpson *et al.*, 2000), and it may be difficult to comprehensively characterize and compare dynamics of paralogy formation across a broad phylogenetic range. Yet, my more constrained, independently developed filter of sequence conservation effectively characterizes the higher frequencies of distances between related paralogs when compared to data from Pushker *et al.* (2004). These high frequency peaks may represent recent formations of paralogs involving the duplication and translocation of a single region containing a cohort of ORFs, or these peaks may represent two differently located hot spots of tandemly duplicating sets of similar ORFs. The paralogy analysis establishes visual distinctions between four different sets of closely related strains, and this implies different lineage-specific constraints of chromosomal mobility and ORF duplication.

The O-ORF similarity cluster size specifications are inclusive of the effect of paralogs whereas the specification of COGs is designed to exclude paralogs. The presence of paralogy significantly increases as a function of genome size (and, correspondingly, total ORF count) (Pushker *et al.*, 2004). For genome sizes < 2 Mb, the percentage of paralogs ranges from 0 to 20 (Pushker *et al.*, 2004). For genome sizes > 2 Mb, the percentage of paralogs ranges from 10 to 50 (Pushker *et al.*, 2004). The X-ORF subset may more significantly include the paralogs (which are by definition excluded from the C-ORF set) than the S-ORF subset. The presence of paralogs may account for the higher correlation of X-ORFs ($r = 0.81$) versus the correlation of S-ORFs ($r = 0.62$) for total annotation counts exceeding 2,000 ORFs (Fig. 53c).

If S-ORFs are interpreted as trending away from duplicate elements (quasi-independent

of paralogous, lateral, or orthologous originations), then they may represent either newly made ORFs, unique ORFs, or significantly "destroyed" and mutated sequences. Subsets based around COG membership (C-ORFs and X-ORFs) scale in closer association with the annotated ORF count compared to O-ORF and S-ORF membership. While assessments of COG membership may be ideal for characterizing the vertically inherited functionality of a genome, fluctuations such as the recombinative generation of paralogs and attenuation of expression, may be better approached with the O-ORF versus S-ORF criteria. The production of noise has been proposed as a key feature of recombination in pathogenic organisms (Wolf & Arkin, 2003).

The characteristics of organisms as conferred by their O-ORF chromosomal organization may be problematic to compare across lineages. Typically, uniform taxons should be characterized to each contribute single data points to a comparative analysis (Grafen & Ridley, 1997), yet my O-ORF specification is likely to be biased by the over-representation of Proteobacteria and Firmicutes in the set of 155 genomes. Fig. 51-52 shows how the impact of phylum over-representation inflates the similarity cluster size S. While there is an elevating effect on the S score for each ORF, Fig. 43 does establish that sizable populations of S-ORFs ($S \leq 5$) are still characterized for taxonomic classes and orders of the Proteobacteria and Firmicutes. Moreover, the limited inclusion of paralogs is evidence that the B and L thresholds for similarity work to reduce O-ORF membership for ORFs that significantly fluctuate their composition, and phantom similarities among atrophying sequences within an over-represented higher taxonomic rank may in this sense have been somewhat avoided. Lateral gene transfer (LGT) may complicate the inferred ancestries of orthology for various ORFs (Koonin *et al.*, 2001), and phylogenetic trees based on ORFs such as metabolic and environmental genes do not concur with rRNA phylogenies (Pace, 1997). LGT only accounts for $\approx 6\%$ of the ORFs (Kurland *et al.*, 2003) however, and if an ORF is laterally transferred and conserved, that would be a case for inclusion in the O-ORF subset.

Further investigation of meaningful boundaries to ORF subsets could integrate the results of more expansive analyses (Allen *et al.*, 2003; Glasner *et al.*, 2003) with more precise characterizations of similarity based on protein structure (Chung & Yona, 2004) and expectation concerning ORF length (Larsen & Krogh, 2003). From the standpoint of

comparatively characterizing recombinative events as functionally important data points in the context of an evolutionary model, the emergence and role of genes in functional groupings and metabolic pathways may help to more closely establish the consequences associated with associated chromosomal rearrangements along phylogenetic branches. There are a variety of efforts that seek to comprehensively evaluate the functional and metabolic dynamics of ORF populations within each genome (Karp *et al.*, 2005; Caspi *et al.*, 2006) and their relationship to phenotype (Schilling *et al.*, 2006). Yet, from a contemporary standopint, based on the recent, "unprecedented" discoveries of decayed ORFs (Ochman & Davalos, 2006), it is currently a meaningful step to focus upon a broad distinction between an operational subset of ORFs compared to contrasting or randomly selected subsets. While there may be complex dynamics of ORF populations, a more inferrential, prescribed approach may suffer from *a priori* assumptions, estimation error, and also hinder repeatability of an analysis to the expanding data set of fully sequenced genomes.

I evaluated the clustering of O-ORFs by the same running tally methodology used to characterize the statistical significance of C-ORF and polarity-based clustering. The degree of statistical significance for O-ORF clustering is similar to the degree of statistical significance established for C-ORF clustering. The terms "shuffling" and "fluidity" have been used to characterize the relocations of ORFs over time (Zivanovic *et al.*, 2002; Lathe *et al.*, 2000), and the negative control used for establishing the sigma σ unit for the bootstrapped z-score difference in distributions is based on a context of completely random, stochastic resamplings of ORF designations as either S-ORFs or O-ORFs. This style of stochastic assignments may be drastically and predictably different than natural processes of ORF addition and loss (Snel *et al.*, 2002), and may also relate to possible fitness constraints on the recombinative relocation of ORFs (Wolf *et al.*, 2001; Lathe *et al.*, 2000). The degree of non-stochastic positioning of O-ORFs and C-ORFs may 1) better support a proposed model of localized rearrangements of chromosomal organization that does not fully obliterate a global conservation of ORF organization (Horimoto *et al.*, 2001), and 2) act to retain localized positioning of ORFs so as to better optimize regulatory expression or protein-protein interaction (Lathe *et al.*, 2000; Svetic *et al.*, 2004) . A more refined approach to measuring chromosomal organization so as to inductively characterize probable pathways and

limitations of recombinative change would involve a more accurate treatment of underlying factors and dynamics more refined than a negative control of completely shuffled ORFs.

Chapter 5: Measures of Internal Physical O-ORF Clustering

5.1 Lagged Autocorrelations of O-ORF Densities

To investigate periodic invariance of O-ORF density, I evaluated lag k autocorrelations on the series of O-ORF densities (Equation 12). I found a general, δ-dependent, property to unshuffled ORF densities where there appeared to be similarity between neighboring values on lag k autocorrelation series (Fig. 57a, 57c, 57e). This property of similarity between neighboring r_k values contrasted with what I observed for lag k autocorrelation series computed from shuffled series of ORF densities. Fig. 57b, 57d, 57f present extreme cases of neighboring dissimilarities along lag k autocorrelation series calculated from shufflings of ORF densities. Similarity between neighboring r_k and r_{k+1} values generally occurred within the range of $-0.2 < r_k < 0.2$ and did not rely on the first neighbor r_1 autocorrelation value to be greater than 0.3. This weak smoothness property appeared limited to δ values ranging from 20 kb to 80 kb. While the weak smoothness property involving $r_k \approx r_{k+1}$ and $-0.2 < r_k < 0.2$ may be evidence against both a purely random distribution of O-ORF densities and strong periodic effects related to O-ORF organization, it may also evidence for a third hypothesis where O-ORF densities form localized variances or shapes that are non-random and interdependent with other regions on the chromosome. To better assess the potential for such a hypothesis, I sought to further model and characterize by approximation the observed non-random smoothness on the lag k autocorrelation series. Elucidating a possible, underlying rule-based system associated with ORF densities is a prerequisite for hypothesis-driven testing.

I postulate that the smoother series of r_k values in Fig. 57a, 57c, and 57e is an effect of similarly-sized expansions ϵ that act to make a r_k autocorrelation value similar to a r_{k+1} autocorrelation value based on a segmentation size δ where $\epsilon \approx \delta$. When ϵ is generally similar to δ, I describe this as a scenario of constrained sizes of expansion that do not perturb segmentation-based symmetries of chromosomal organization. A more asymmetric variability

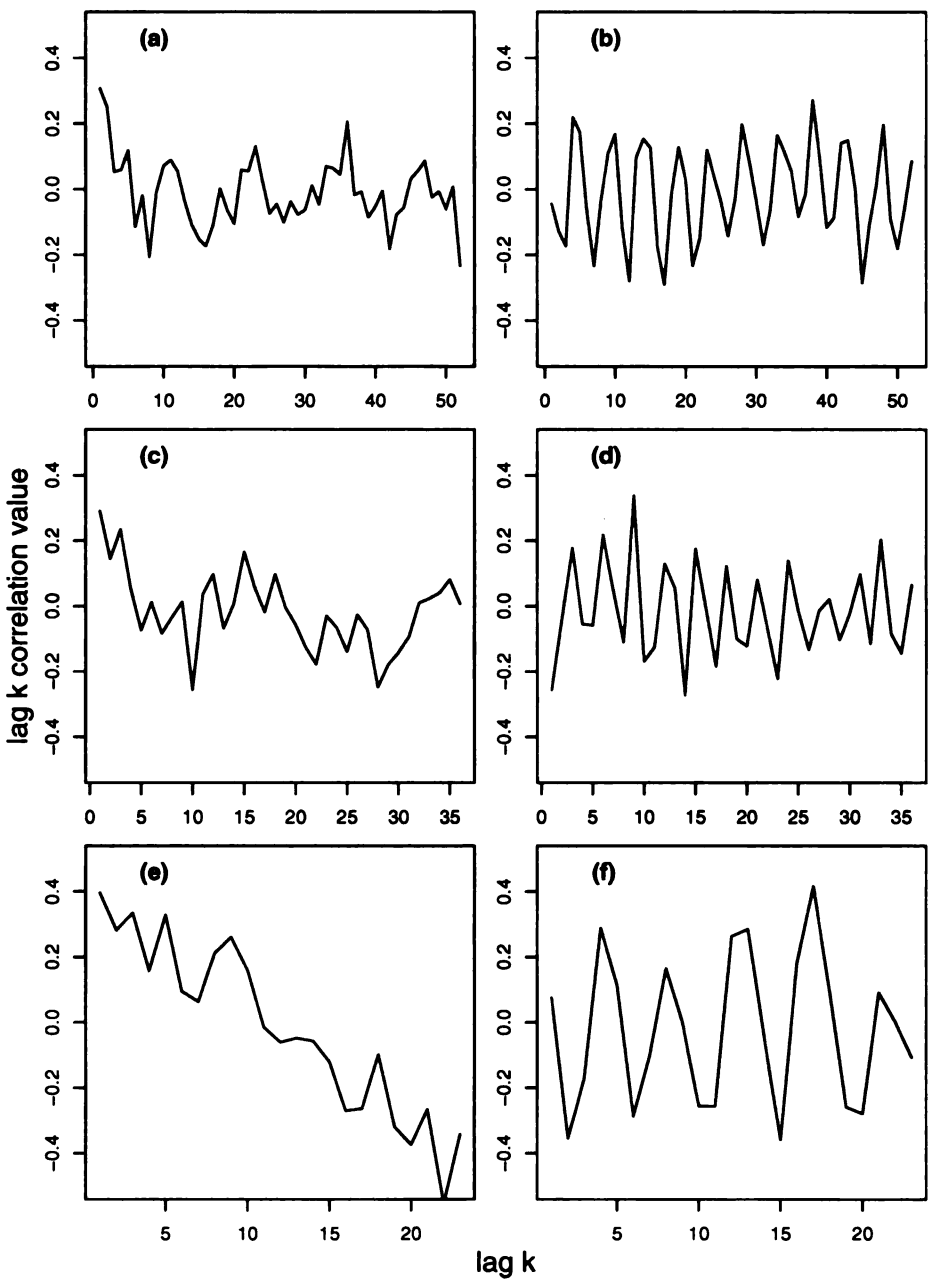

Figure 57: Lag k autocorrelation values calculated on ORF density series built with segmentation size $\delta = 40$ kb. k starts at 1. The right column (b, d, f) represents r values corresponding to shuffled density series. The left column (a, c, e) is for the non-shuffled density series. **(a,b)** *Bacillus subtilis.* **(c,d)** *Vibrio cholerae* (large chromosome). **(e,f)** *Pyrococcus furiosus.*

to O-ORF densities would correspond to ϵ values where $\epsilon - \delta \neq 0$, potentially resulting in larger differences between r_k and r_{k+1} autocorrelation values.

5.2 Scalar Residue Measures

5.2.1 Sum of Squared Differences on Lagged Autocorrelation Series

To non-parametrically measure the degree of symmetry for various segmentation sizes δ, the $Q(c_i, \delta)$ symmetry score is based on a bootstrap comparison of sum of squared differences on the r_k series for shuffled and unshuffled versions of an O-ORF density series F (Equation 13). The first-neighbor autocorrelation value r_1 is not included in the Q scoring. High values in the Q series correspond to reduced squared differences between r_k and r_{k+1}, generally implying a higher degree of symmetry where $r_k \approx r_{k+1}$ for the given segmentation size δ. Fig. 58 shows the Q-based series of values for two groups of related genomes where the main chromosome c_i for each genome is evaluated to produce the series $Q(c_i, \delta = 10 \text{ kb}), Q(c_i, \delta = 20 \text{ kb}), Q(c_i, \delta = 30 \text{ kb}), ..., Q(c_i, \delta = 150 \text{ kb})$. Genus-based relatedness is apparent for both *Sulfolobus* and *Mycobacterium* when compared to the distant relations shown in Fig. 58c and 58f respectively.

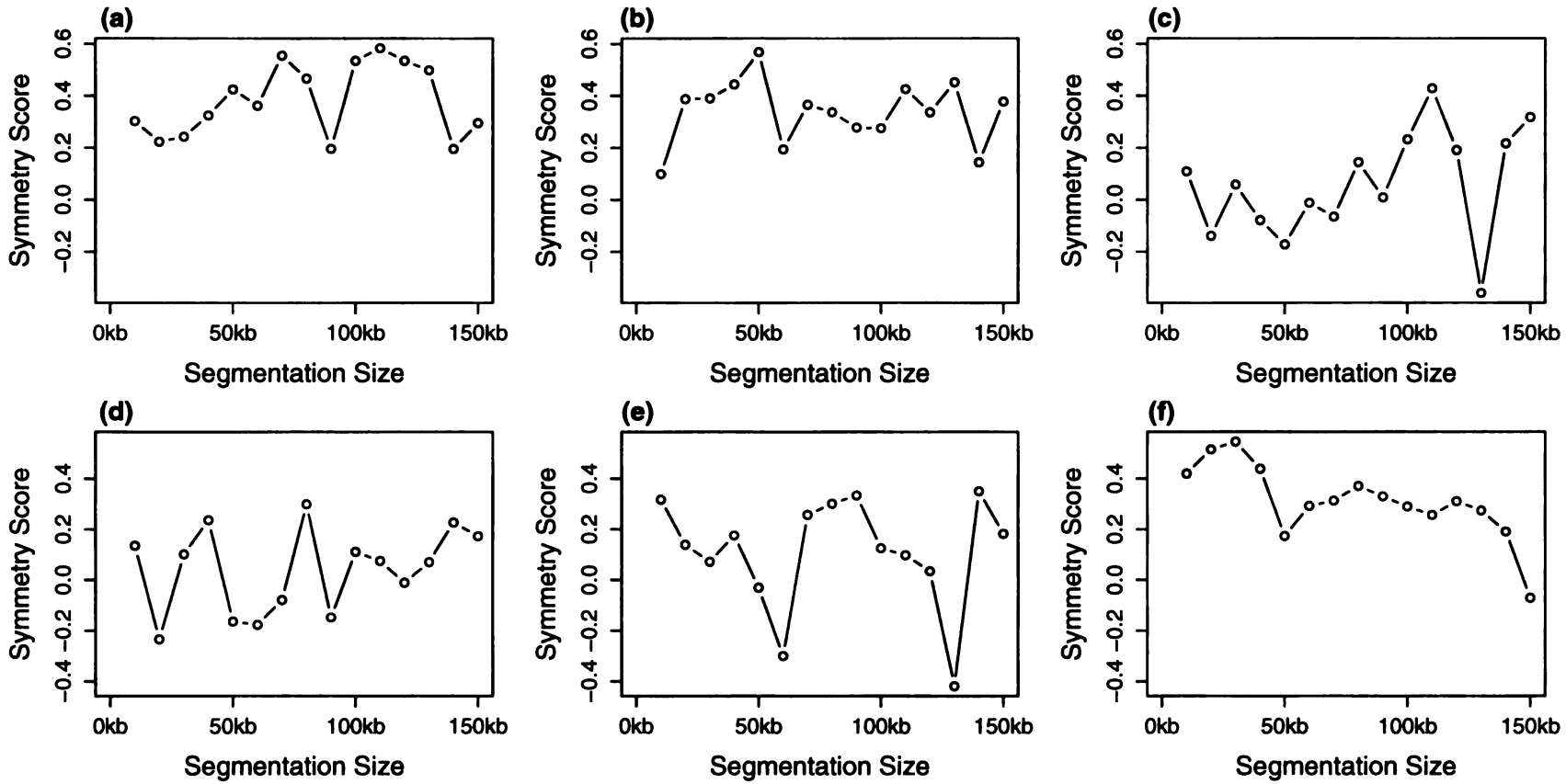

Figure 58: The $Q(c_i, \delta)$ measure of segmentation-based symmetry scores for O-ORF densities among three Crenarchaeota strains and three Actinobacteria strains for $\delta = 10, 20, 30, ..., 150$ kb. **(a)** *Sulfolobus solfataricus.* **(b)** *Sulfolobus tokodaii.* **(c)** *Aeropyrum pernix.* **(d)** *Mycobacterium tuberculosis* CDC1551. **(e)** *Mycobacterium tuberculosis* H37Rv. **(f)** *Corynebacterium glutamicum* ATCC 13032.

My postulate is that the Q scoring of symmetry indicates symmetry-conserving expansions where $\delta \approx \epsilon$ and $r_k \approx r_{k+1}$. Yet, there may be the effect of $r_k \approx r_{k+1} \approx 0$, in which case there is minimal signal with which to assert that $\delta = \epsilon$. Yet, if $r_k \approx r_{k+1} \approx r_{k+2}$, I hypothesize that regions on a $Q(c_i, \delta_j)$ series would more powerfully imply an ϵ constrait of symmetry-conserving expansions if a harmonic were observed where $Q(c_i, \delta) \approx Q(c_i, j \times \delta)$ where j is an integer.

Based on my postulate that a measure of the internal physical clustering of ORFs would show some characteristic of vertical ancestry (Section 1.5), I expect significant instances of heritability to be associated with a measure of segmentation-based symmetries among O-ORF densities. To investigate pairwise comparisons of chromosomal organization as characterizaed by the $Q(c_i, \delta)$ scorings of symmetry for δ-based values, I evaluated the cross-correlation between pairs of Q-based scoring series among sets of related chromosomes (Table 18). For each set of three chromosomes evaluated in each row of Table 18, the first column of Table 18 shows the cross-correlation between the two most closely related chromosomes. For the 3 archaeal sets, the cross-correlation identifies the most closely related chromosomes. Yet, of the two most closely related chromosomes from the bacterial sets of comparisons, only the strains of *Mycobacterium tuberculosis* correlate together, and the $r = 0.33$ value is weak.

To broadly assay intrachromosomal symmetries of O-ORF density across multiple phyla, I constructed three series of values by chaining together the Q series of three sets of chromosomes (C_1, C_2, C_3) where the Q series were based on $\delta = 10, 20, 30, ..., 150$ kb. Based on abbreviations listed in Table 2, the three sets of chromosomes were: C_1 = {Py.aby., Mt.acet., Sulf.solf., tb1551, Nostoc, S.pyog., M.gen., R.pro., and Xn.cmp.}; C_2 = {Py.hor., Mt.maz., Sulf.tok., tbH37Rv, Synec., Str.pnm., M.pnm., R.con., and Xn.axon.}; and C_3 = {Py.fur., Arch.ful., A.pnx., Cor.glut., Th.elon., Lac.lact., M.pulm., Caul.cre., and X.fas.}. In this particular analysis, my measurement of dependence between $Q(c_i, \delta)$ and $Q(c_i, \delta + a)$ is limited to $a = 10$ kb. I measured the first neighbor autocorrelation r_1 and also average mutual information values for lags of 1, 2, 3, 4, 5, 6, and 7 on each of the 3 constructed series (Table 19).

Based on Table 19, a putative result to the analysis is that lag 1 or lag 2 measures of average mutual information are generally higher than the lag 3-7 measures of average mutual

Table 18: Cross-correlations r among sets of three chromosomes (c_1, c_2, c_3) between series of $Q(c_i, \delta)$ values for segmentation-based symmetries of O-ORF densities where $\delta = 10, 20, 30, ..., 150$ kb.[a]

Closest Pair	r	Closest & Distant Ancestor	r	Closest & Distant Ancestor	r
Py.aby+Py.hor.	0.37	Py.aby.+Py.fur.	0.017	Py.hor.+Py.fur.	0.17
Mt.acet.+Mt.maz.	0.33	Mt.acet.+Arch.ful.	0.11	Mt.maz.+Arch.ful.	-0.35
Sulf.solf.+Sulf.tok.	0.78	Sulf.solf.+A.pnx.	0.36	Sulf.tok.+A.pnx.	0.32
tb1551+tbH37Rv	0.33	tb1551+Cor.glut.	-0.11	tbH37Rv+Cor.glut.	0.075
Nostoc+Synec.	-0.28	Nostoc+Th.elon.	-0.16	Synec.+Th.elon.	-0.39
S.pyog.+Str.pnm.	0.069	S.pyog.+Lac.lact.	-0.23	Str.pnm.+Lac.lact.	0.37
M.gen.+M.pnm.	0.15	M.gen.+M.pulm.	-0.028	M.pnm.+M.pulm.	-0.32
R.pro.+R.con.	-0.20	R.pro.+Caul.cre.	0.14	R.con.+Caul.cre.	-0.26
Xn.cmp.+Xn.axon.	-0.50	Xn.cmp.+X.fas.	-0.10	Xn.axon.+X.fas.	0.31

[a]The two most closely related chromosomes are c_1 and c_2.

Table 19: Average mutual information of lagged series for $Q(c_i, \delta)$ values of the symmetry score where segmentation size $\delta = 10, 20, 30, ..., 150$ kb.

		Average Mutual Information						
Series	r_1	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7
C_1	0.49	0.66	0.54	0.46	0.44	0.48	0.55	0.49
C_2	0.24	0.50	0.52	0.49	0.49	0.44	0.50	0.47
C_3	0.20	0.59	0.55	0.50	0.53	0.52	0.54	0.52

information.

5.2.2 Frequency of Changes in Angular Variation on the O-ORF Density Series

To further verify the aspect of segmentation-based symmetries implicating an underlying expansionary characteristic of $\epsilon = \delta$, I hypothesize that a similar analysis based on angular change on the pseudophase-spaced series $P(c_i, \delta)$ should demonstrate predictable information between $P(c_i, \delta)$ and $P(c_i, \delta + a)$ where $a = 10$ kb. The $P(c_i, \delta)$ scoring is based on a comparison between angular frequency histograms $D(F)$ for shuffled and unshuffled versions of a pseudophased O-ORF density series F (Equations 14-16). High values of $P(c_i, \delta)$ indicate non-random angular relationships based on a Kolmogorov-Smirnov comparison with a shuffled F series. Fig. 59 shows the P-based series of values for two groups of related genomes where the main chromosome c_i for each genome is evaluated to produce the series $P(c_i, \delta = 10 \text{ kb}), P(c_i, \delta = 20 \text{ kb}), P(c_i, \delta = 30 \text{ kb}), ..., P(c_i, \delta = 150 \text{ kb})$. Putative, genus-based relatedness is apparent for both *Sulfolobus* and *Mycobacterium* compared to distant relations shown in Fig. 59c and 59f respectively. The non-randomness of pseudophased angular change appears to rise with increasing values of δ, and this may be an artificial measurement effect due to the less densely populated histograms of angular frequencies based on fewer densities being generated by larger segmentation sizes.

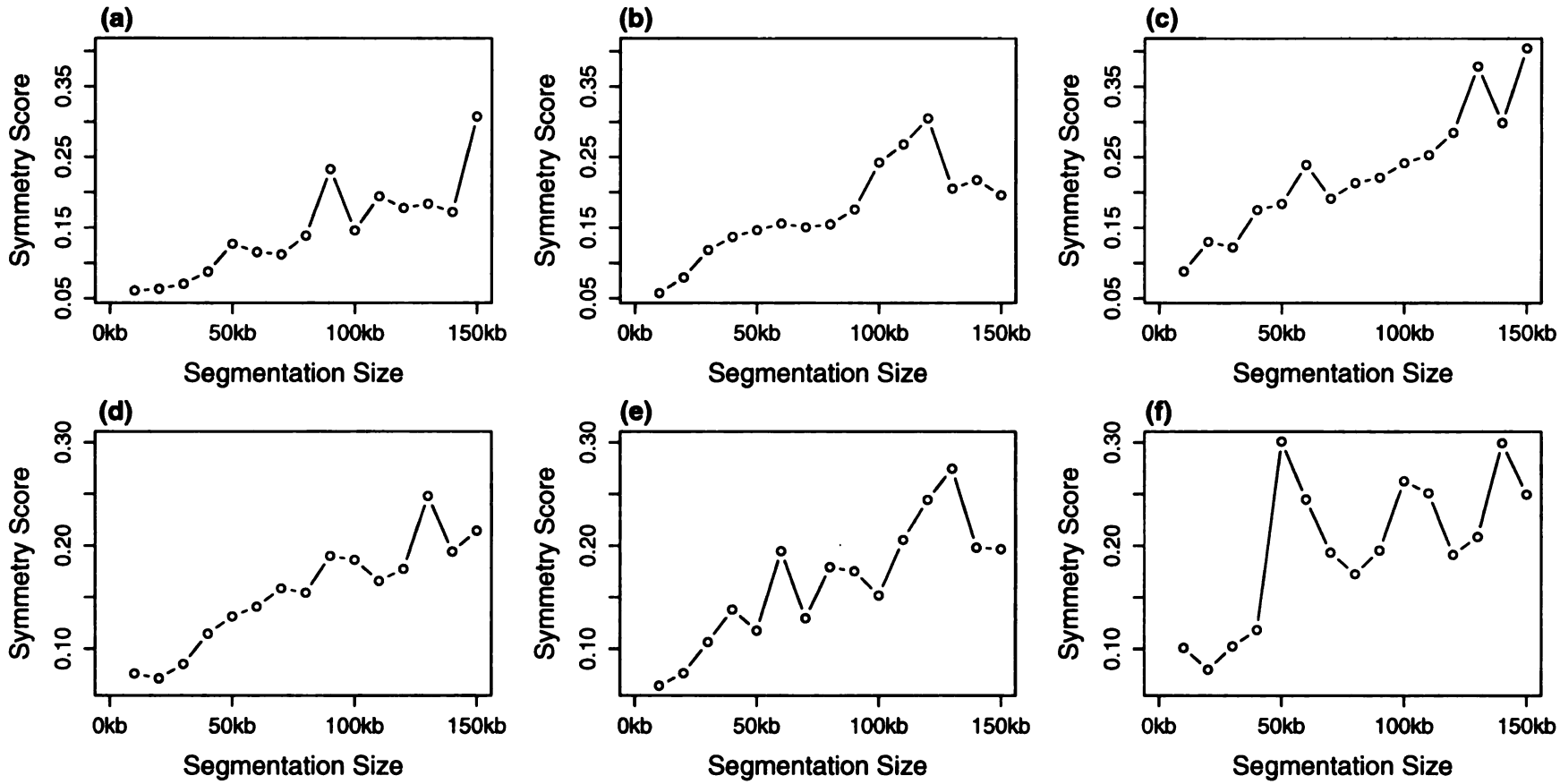

Figure 59: The $P(c_i, \delta)$ measure of symmetric shape among three Crenarchaeota strains and three Actinobacteria strains for $\delta = 10, 20, 30, ..., 150$ kb. **(a)** *Sulfolobus solfataricus.* **(b)** *Sulfolobus tokodaii.* **(c)** *Aeropyrum pernix.* **(d)** *Mycobacterium tuberculosis* CDC1551. **(e)** *Mycobacterium tuberculosis* H37Rv. **(f)** *Corynebacterium glutamicum* ATCC 13032.

Table 20: Average mutual information of lagged series for $P(c_i, \delta)$ values of symmetric shape for segmentation sizes $\delta = 10, 20, 30, ..., 150$ kb.

		Average Mutual Information						
Series	r_1	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7
1	0.70	0.87	0.71	0.70	0.65	0.55	0.53	0.52
2	0.66	1.1	0.96	0.95	0.87	0.79	0.80	0.82
3	0.53	0.74	0.85	0.77	0.78	0.70	0.73	0.64

Based on Table 20, and consistent with Table 19, a putative aspect to the analysis is that lag 1 or lag 2 measures of average mutual information are generally higher than the lag 3-7 measures of average mutual information. The upwards rise in values characterized by Fig. 59 likely accounts for the high r_1 values in Table 20.

5.2.3 Taxonomic Correspondence

While there were some trends evident from the P and Q measures of O-ORF clustering for $\delta = 10, 20, 30, ..., 150$ kb, the chaining together of unrelated chromosomes for Tables 19 and 20 does not allow for an inspection of phylogenetic differences between related organisms. Also, while there was some correlation between the Q measure of segmentation-based symmetries that may relate to phylogeny (Table 18), there was only a minor degree of correlative pattern observed for bacterial chromosomes. To conduct a more detailed analysis, I calculated the Q scoring of O-ORF density symmetries for $\delta = 500, 1,000, 1,500, ..., 150,000$ bp. Fig. 60-62 contrasts the O-ORF density symmetry scores for a pair of two more closely related chromosomes (leftmost columns) with a third more distantly related chromosome (rightmost column). Compared to the low cross-correlation coefficients for bacterial comparisons in Table 18, there is a putative similarity that can be more strikingly observed based on the more detailed examination of segmentation sizes,
$\delta = 500, 1,000, 1,500, ..., 150,000$ bp, as shown in Fig. 61-62. For instance, in Fig. 61, the two strains of *Mycobacterium tuberculosis* appear dramatically different than their distant relation in the Actinobacteria, *Corynebacterium glutamicum* (928 Ma). A notable exception however is for the Xanthomonadaceae set of strains in Fig. 62 where *Xanthomonas campestris* shows more similarity to *Xylella fastidiosa* than to *Xanthomonas axonopodis*. Yet, the various peaks

and troughs associated with specific δ_i segmentation sizes in Fig. 60-62 do not perfectly align from one chromosome to another and, despite the visual strong trend of consistency with taxonomy, the δ-based values did not generally produce significant correlative measures between related chromosomes. Moreover, I did not find a simple smoothing algorithm that would work to generalize the outlying patterns of peaks and troughs.

5.3 Simulation of Informational Change

5.3.1 Development and Characterization of Simulation Model

My hypothesis is that regions on a $Q(c_i, \delta_j)$ series can imply an ϵ constrait of symmetry-conserving expansions. The diagnostic invariance of this symmetry-conserving expansion would be such that $\delta \approx \epsilon$ and $Q(c_i, \delta) \approx Q(c_i, \delta + \epsilon)$. I sought to develop a model of simulated chromosomal expansions and modifications in order to evaluate if $Q(c_i, \delta)$ symmetry scores of the simulation model output meaningful relate, as proposed, to constant-sized expansions. The hypothesized relationship of a harmonic association with ϵ implies that a measure to characterize the amplitudes of the harmonic (e.g., moduli of the fast Fourier transform) should be able to characterize ϵ.

Chromosome simulations were conducted as described in Section 2.7 across various parameters, S, N, and T. The relative stochastic level of occurrence between tandem duplicating expansions and translocating modifications ranged, as a parameter, from 0.3 to 0.7 ($S/10$). The parameter, N, is defined as the number of consecutively stringed octets x where x = ABCDEFGH, $y = N$, and x^y is the starting symbolic sequence. For the running of each simulation trial, N ranged from 2 to 5 and the size of tandem duplications, T, ranged from 3 to 12. A set of examples that helps illustrate the behavior of these parameters is shown in Fig. 63. Segment size and counting was a function of consecutive subsequences of a given length (i.e., segment size or "window" size on the output, simulation-based symbolic sequence), and the number of "H" letters observed in the consecutively windowed subsequences (i.e., a count).

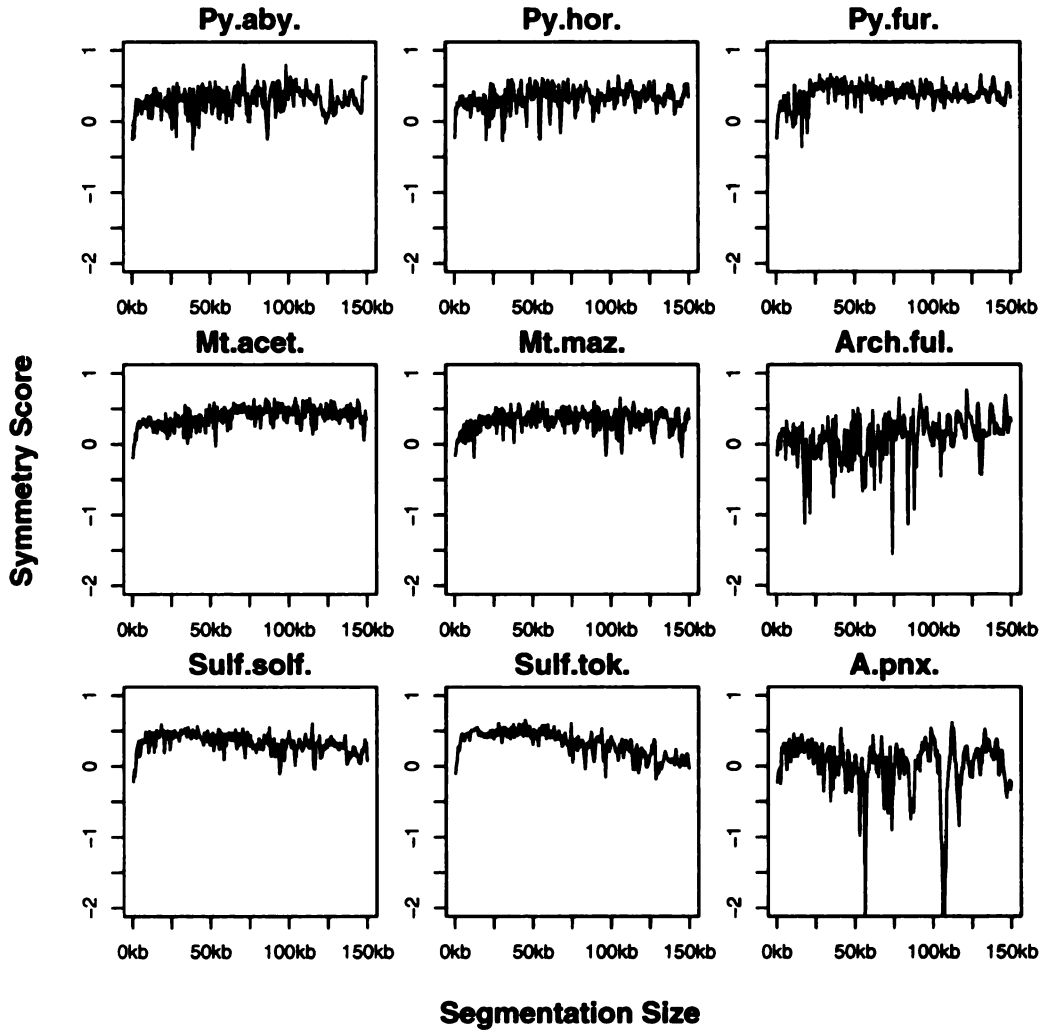

Figure 60: Q-based symmetry scores of O-ORF densities on 9 archaeal chromosomes. Each row corresponds to a set of phylogenetically related strains. The first two columns represent the chromosomes with the most recent common ancestor in comparison to the third column. Abbreviated strain names are defined in Table 2. The $Q(c_i, \delta)$ measure is described in Sections 2.8.3 and 2.8.4.

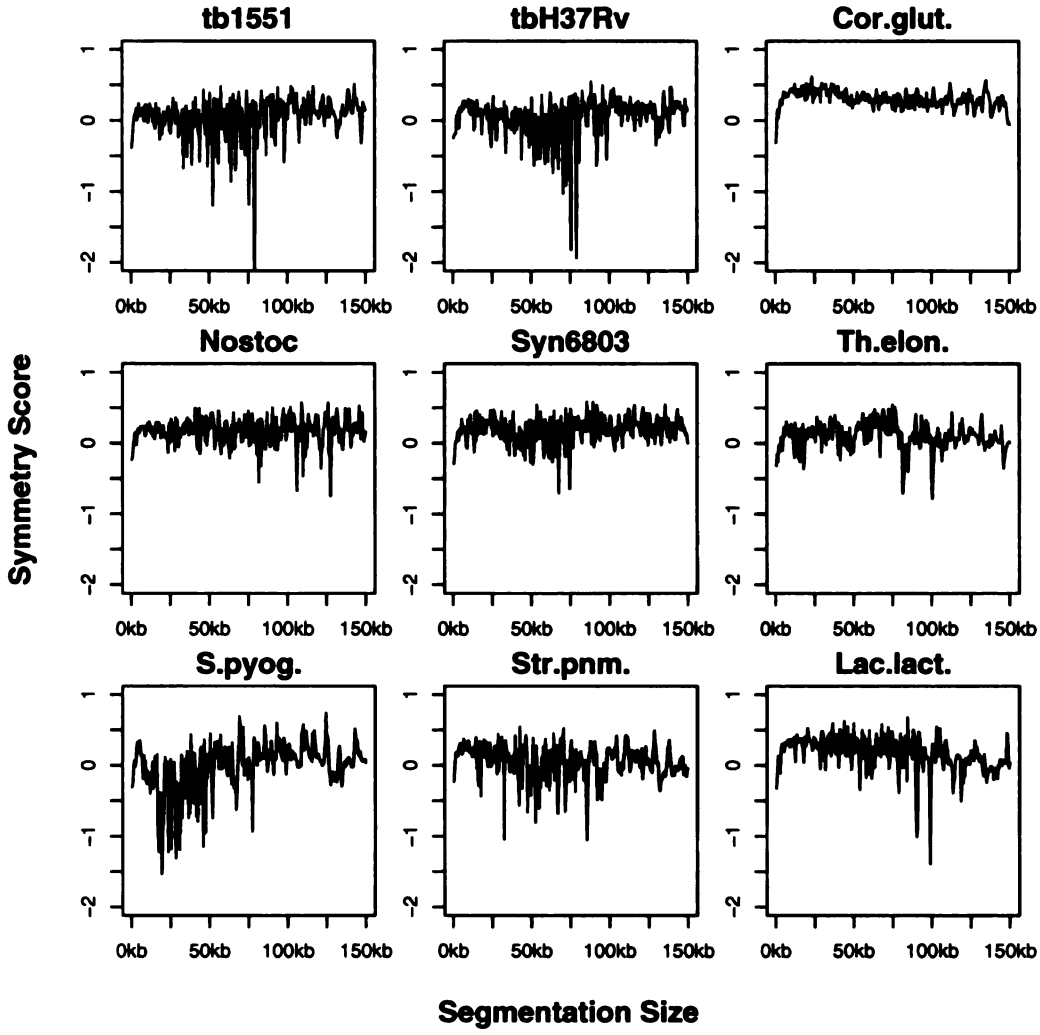

Figure 61: Q-based symmetry scores of O-ORF densities on the chromosomes of 3 Actinobacteria, 3 Cyanobacteria, and 3 Lactobacillales. Each row corresponds to a set of phylogenetically related strains. The first two columns represent the chromosomes with the most recent common ancestor in comparison to the third column. Abbreviated strain names are defined in Table 2. The $Q(c_i, \delta)$ measure is described in Sections 2.8.3 and 2.8.4.

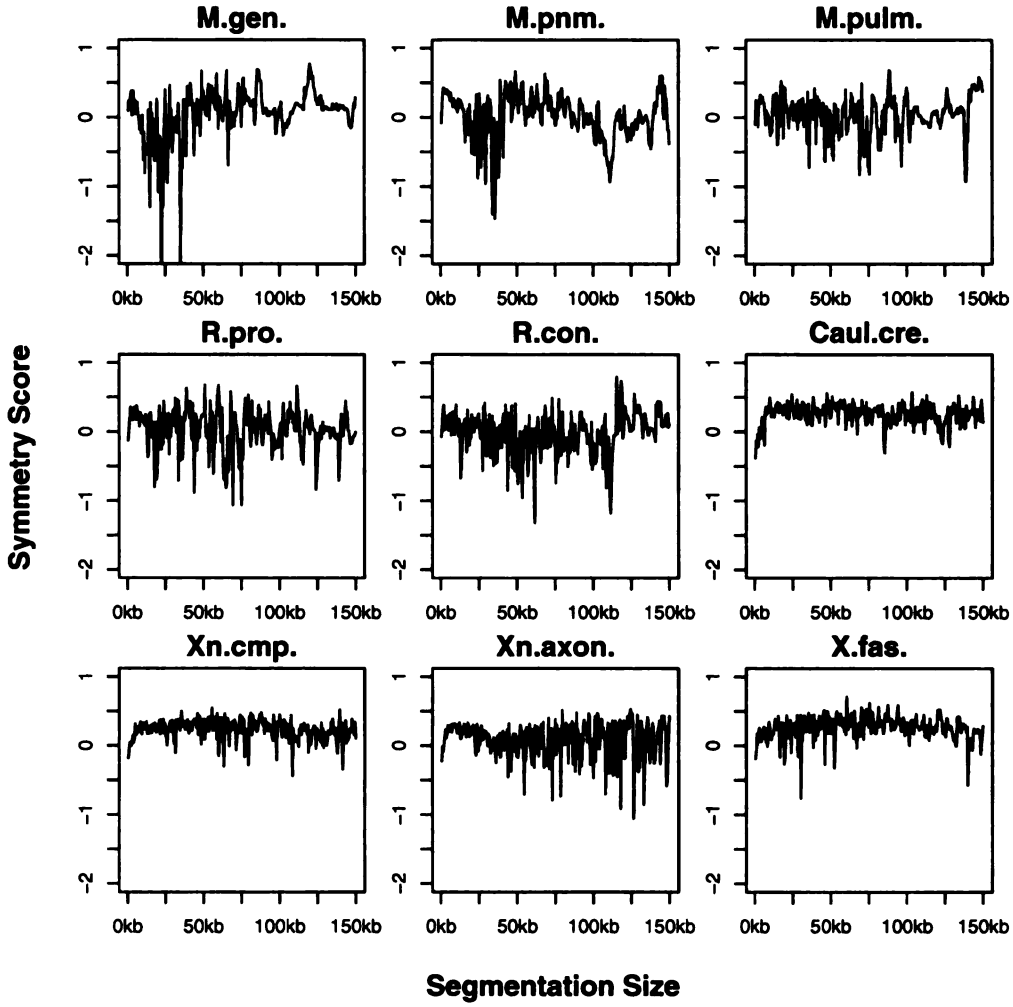

Figure 62: Q-based symmetry scores of O-ORF densities on the chromosomes of 3 Mollicutes and 6 Proteobacteria. Each row corresponds to a set of phylogenetically related strains. The first two columns represent the chromosomes with the most recent common ancestor in comparison to the third column. Abbreviated strain names are defined in Table 2. The $Q(c_i, \delta)$ measure is described in Sections 2.8.3 and 2.8.4.

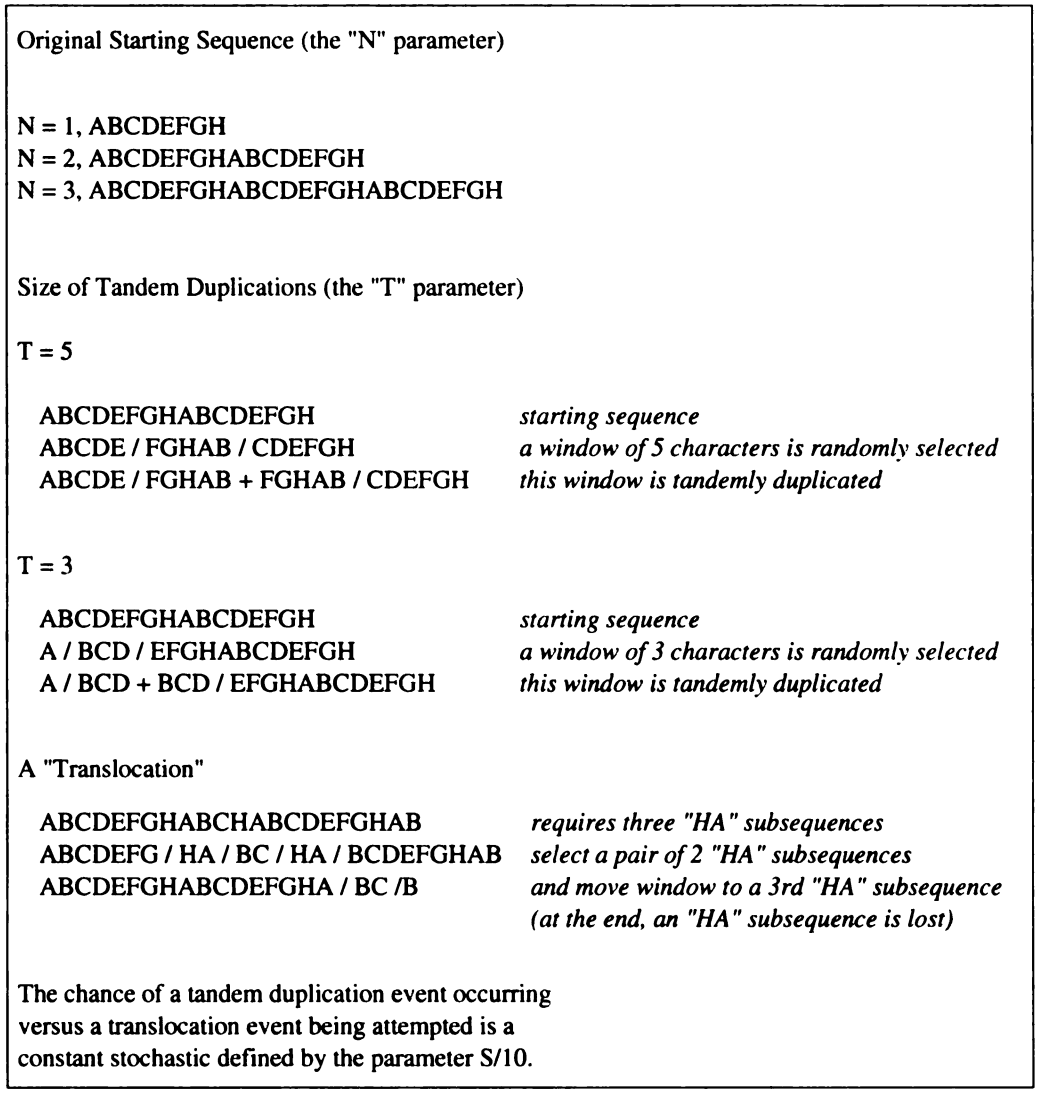

Figure 63: Examples of the abstract simulation for structural duplications and translocations on a symbolic sequence.

5.3.2 Scalar and Spectral Measures of Model Output

Visual examples of symmetry scoring of simulation-produced symbolic output are shown in Fig. 64 - 65. There was some correspondence between symmetry measures for similar, yet non-identical, parameters of S, N, and T as shown in Fig. 66 - 68. The ordinate scale on the simulation-based plots may not directly equate in meaning to the ordinate scale of the Q-based symmetry measures shown in Fig. 58 and Fig. 60-62, yet the ranges are comparable when T is low ($T = 3$) and N is of an intermediate value ($N = 3$ or $N = 5$). An effect of a small T parameter and high N parameter was to reduce the presence of low (< -10) symmetry scores from measurements of small segment sizes. Visually, the T parameter appeared to correspond to a periodicity of the symmetry scoring. S did not have a dramatic impact on the symmetry scores.

Fig. 66 and 67 show how a spectral assessment with the fast Fourier transform (FFT) on the Q series of symmetry scores may help reveal the underlying parameters to the simulation. As T changes (Fig. 66), the moduli of the FFT series form peaks at locations corresponding approximately to 30/(T-1). To illustrate this relationship, a tandem duplication parameter of 6 would potentially result in repetitious measures of density for every 6 characters on the simulated output sequence of characters. The Q series may preferentially measure this effect for segment sizes of 6, 12, 18, 24, and 30 as might be inferred from the behavior of plots in Fig. 64 - 65. This succession of preferential segment sizes (6, 12, 18, 24, and 30) corresponds to a periodicity of 4 on a series from 1 to 30. Fig. 67 shows the visual effect on the FFT modulus series for altering S and N parameters of the underlying simulation model and a mathematical relationship between the structure of the FFT modulus series (Mod(fft(Q)) compared to the S and N parameters is not readily apparent.

As S, N, or T is offset by 1, Fig. 68 shows the degree to which the Q and Mod(fft(Q))-based distributions are altered. Adjusting any simulation parameter by 1 does not radically alter the Q series distributions (Fig. 68a - 68c) and, for alterations of S and N, the Mod(fft(Q)-based distributions (Fig. 68d - 68e). Similarity between distributions is significantly lost for the Mod(fft(Q))-based distributions when T is altered, even by a single increment (Fig. 68f). The Mod(fft(Q))-based assay, in this sense, demonstrates increased sensitivity to relatively small changes in the size of tandemly duplicating expansions.

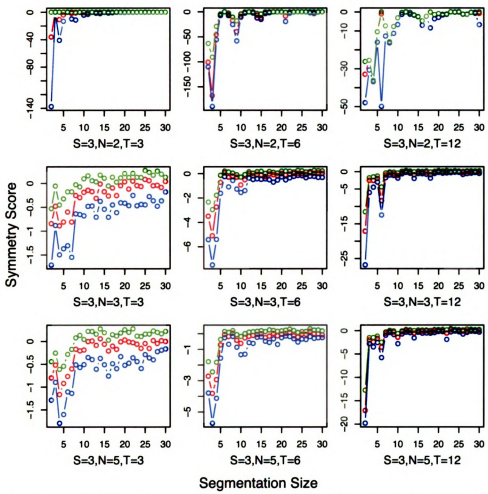

Figure 64: Scorings of segmentation-based symmetries for simulations of informational expansion and modification. 29 segmentation sizes were evaluated $(2, 3, 4, ..., 30)$ for varying parameters. S (relative stochastic) = 3. N (number of originating ABCDEFGH octets): 2, 3, 5. T (size of tandem duplications): 3, 6, 12. Each point characterizes the distribution of 50 replicate simulations on a given segmentation size and set of S, N, and T parameters: first quartile shown in blue; median shown in red; third quartile shown in green.

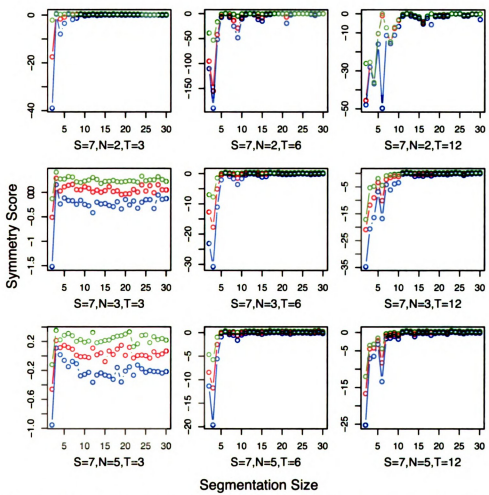

Figure 65: Scorings of segmentation-based symmetries for simulations of informational expansion and modification. 29 segmentation sizes were evaluated $(2, 3, 4, ..., 30)$ for varying parameters. S (relative stochastic) = 7. N (number of originating ABCDEFGH octets): 2, 3, 5. T (size of tandem duplications): 3, 6, 12. Each point characterizes the distribution of 50 replicate simulations on a given segmentation size and set of S, N, and T parameters: first quartile shown in blue; median shown in red; third quartile shown in green.

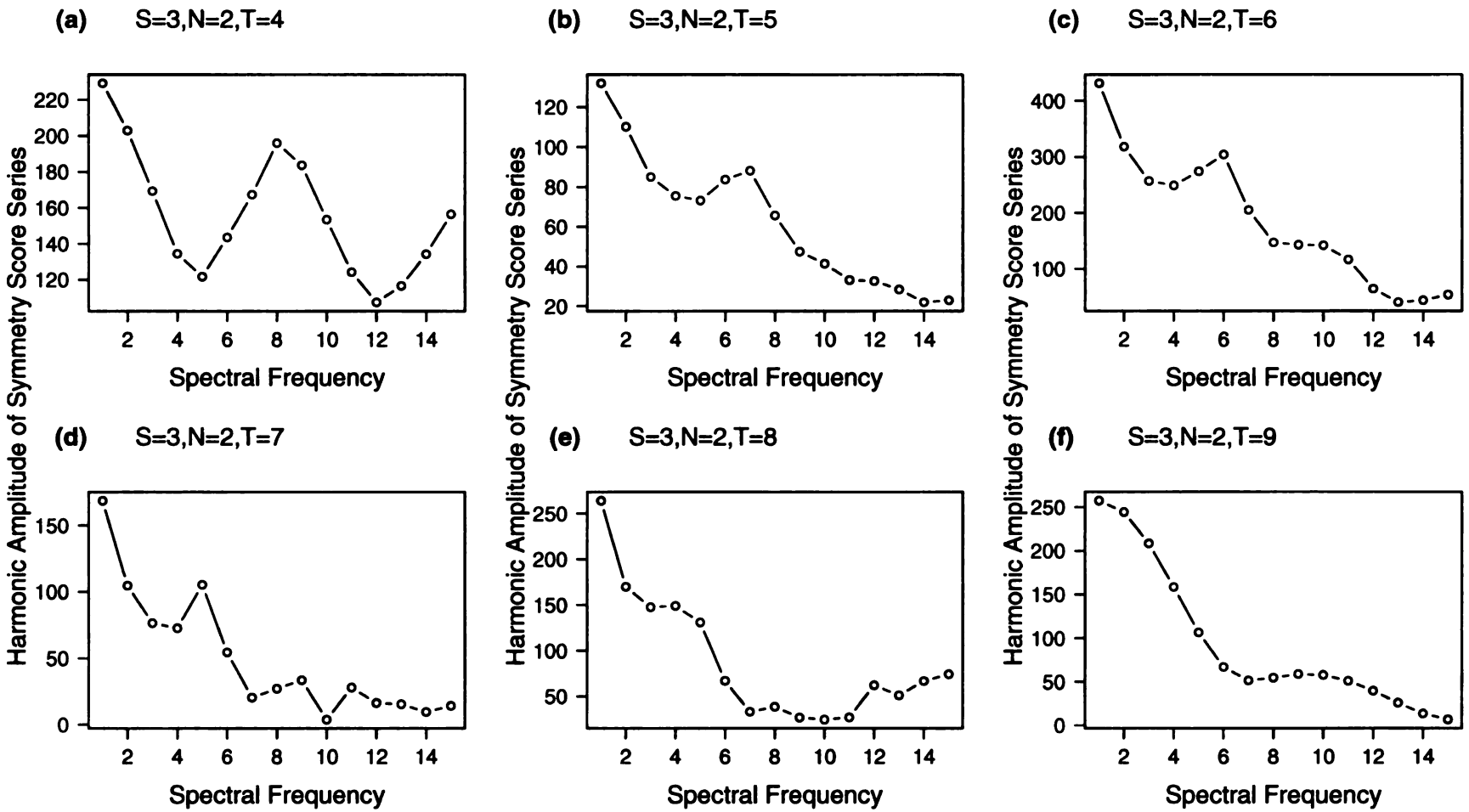

Figure 66: Changes to FFT on the Q series based on adjusting the simulation model T value. T (size of tandem duplications) was set to the values 4 **(a)**, 5 **(b)**, 6 **(c)**, 7 **(d)**, 8 **(e)**, and 9 **(f)**. S (relative stochastic) was set as 3, and N (number of originating ABCDEFGH octets) was set as 2. The harmonic amplitude of symmetry score series is calculated by the formula, Mod(fft(Q)).

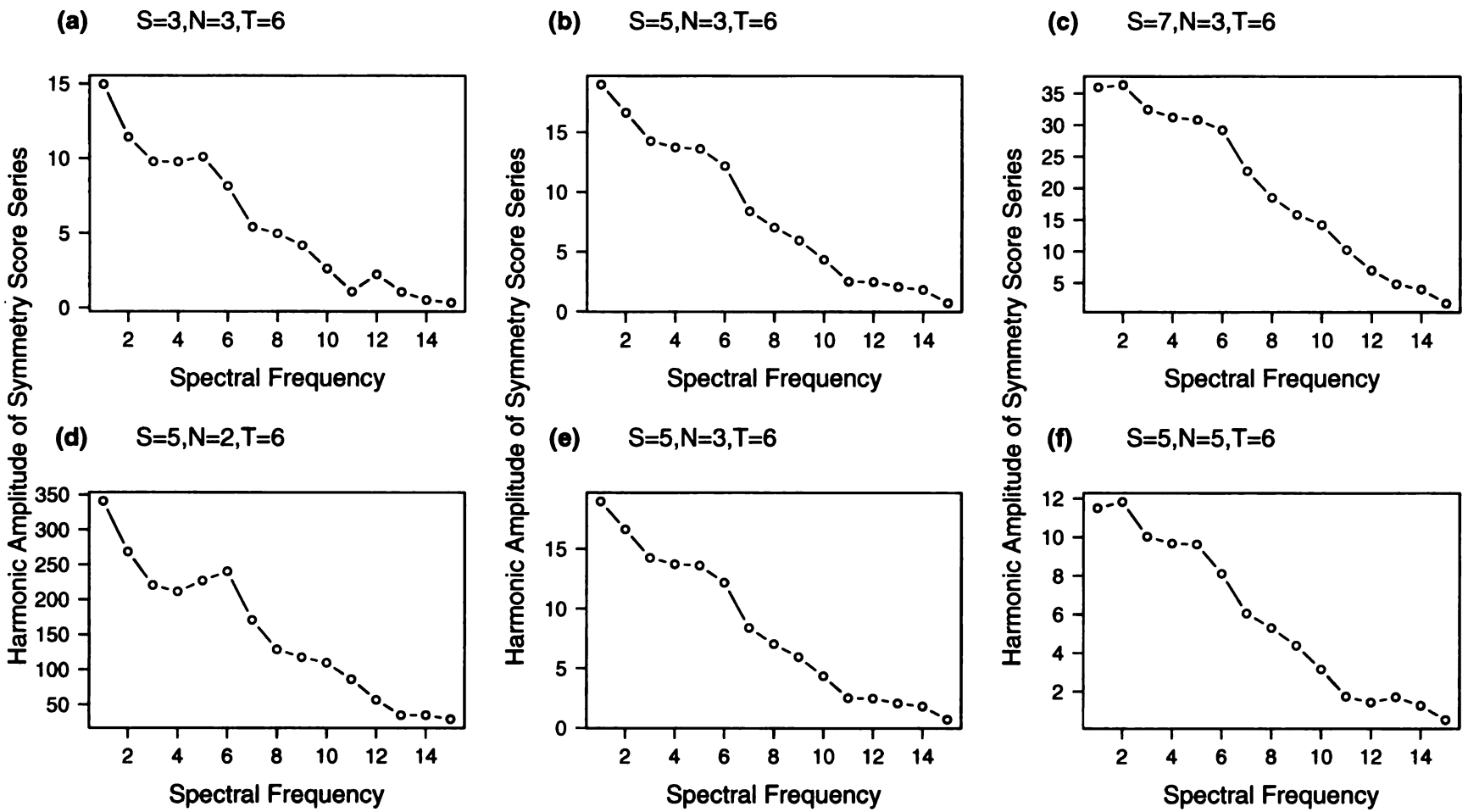

Figure 67: Changes to FFT on the Q series based on adjusting the simulation model S and N values. **(a-c)** show the effect of adjusting S, while holding the N and T parameters constant. **(d-f)** show the effect of adjusting N, while holding the S and T parameters constant. The harmonic amplitude of symmetry score series is calculated by the formula, Mod(fft(Q)).

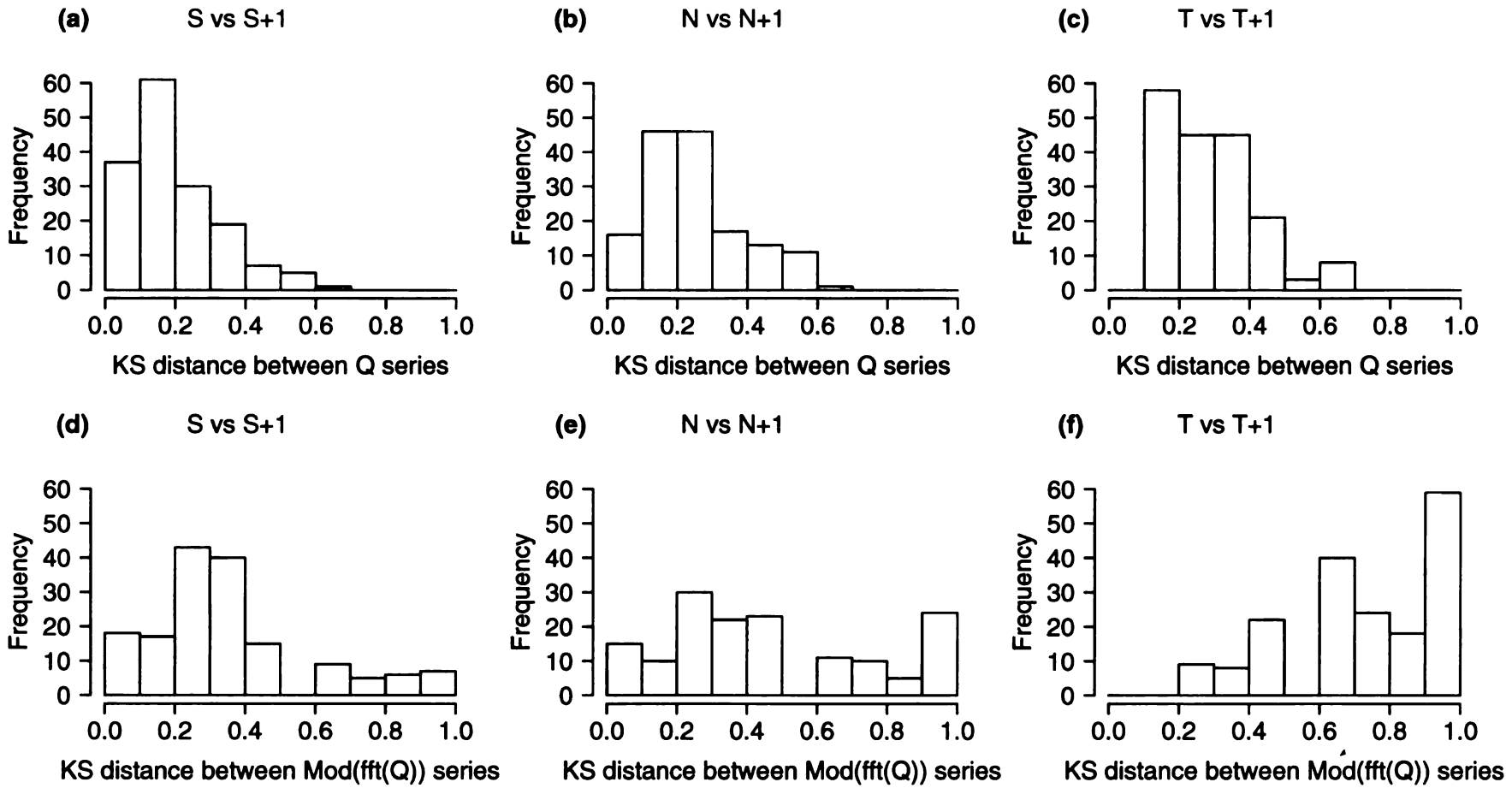

Figure 68: KS distances of Q and Mod(fft(Q))-based measures of simulated rearrangements. The Kolmogorov-Smirnov (KS) distances are based on perturbations involving two sets of simulation parameters where just one of the three integer-based parameters, S (3 to 7), N (2 to 5), or T (3 to 12), is set to differ by a value of 1. The KS distances are presented as cumulative frequency histograms, bin size = 0.1. **(a-c)**, the KS distance based on comparisons of Q series. **(d-f)**, the KS distance based on comparisons of Mod(fft(Q)) series. Differing simulation model parameter is S **(a,b)**, N **(c,d)**, or T **(e,f)**.

5.4 Investigating Phylogeny

Calibration was for a symmetric scoring based on 51 segmentation size values (25,000 bp), and the window for which 4 *Streptococcus pyogenes* chromosomes had the greatest pairwise difference in their characteristic ranges of Q symmetry scores (see Fig. 69). The rationale for this calibrating approach was to approximate the detection of divergence against a lineage with known organizationally divergent properties at the subspecies level. A window size of 25 kb approximates the general fluctation of high and low symmetry scores observable in Fig. 58 and 59, and putatively evident from the lagged average mutual information analyses in Tables 19 and 20.

The first spectral modulus from the 51-value window was used to characterize the range of Q symmetry scores, and was termed the windowed asymmetric deviation. The highest differences of windowed asymmetric deviations among *S. pyogenes* strains were the 75th segmentation size, 37,500 bp, to the 125th segmentation size, 62,500 bp.

There were 24 sets of closely related species. The distribution of closely related species' pairwise differences of windowed asymmetric deviation for 37.5 kb to 62.5 kb is shown in Fig. 70.

Fig. 71 shows the relationship of time of divergence from a last common ancestor to differences in chromosomal structure and organization. I did not find direct cross-correlations between the individual measures of chromosomal structure and organization: chromosome size, windowed asymmetric variation, and average pointwise mutual information, so the covariance of these multiplied measures with times of divergence has added significance. The alphabetic letters in Fig. 71 correspond to pairwise comparisons among sets of three chromosomes (the identity of which are described in Table 5). While linear correlations were significant ($0.66 \leq r \leq 0.87$), the "I" and "E" sets of chromosomes (Xanthomonadaceae and Cyanobacteria) were conflicting in their relationship of difference in chromosomal structure and organization to the estimated time of divergence from a last common ancestor. Incidentally, similar to the analysis in Fig. 69, the highest correlations for the relationship of difference in chromosomal structure and organization to the estimated time of divergence from a last common ancestor occurred for windowed asymmetric deviations for the range of segmentation sizes δ = 37.5 kb to 62.5 kb.

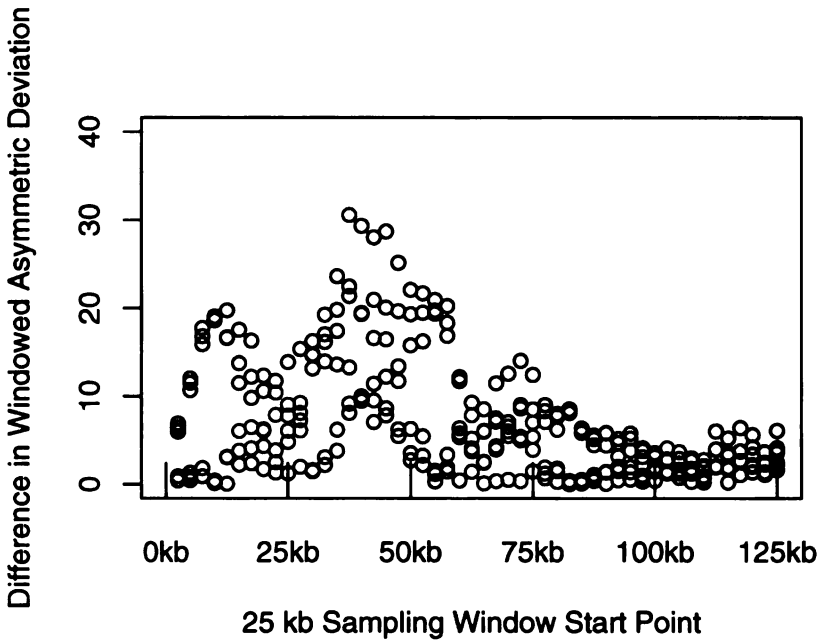

Figure 69: Differences of windowed asymmetric deviations between *S. pyogenes* strains based on a 25 kb window of segmentation sizes. For the 4 strains evaluated, there were 6 pairwise comparisons. The start point for each 25 kb window is the abscissa value.

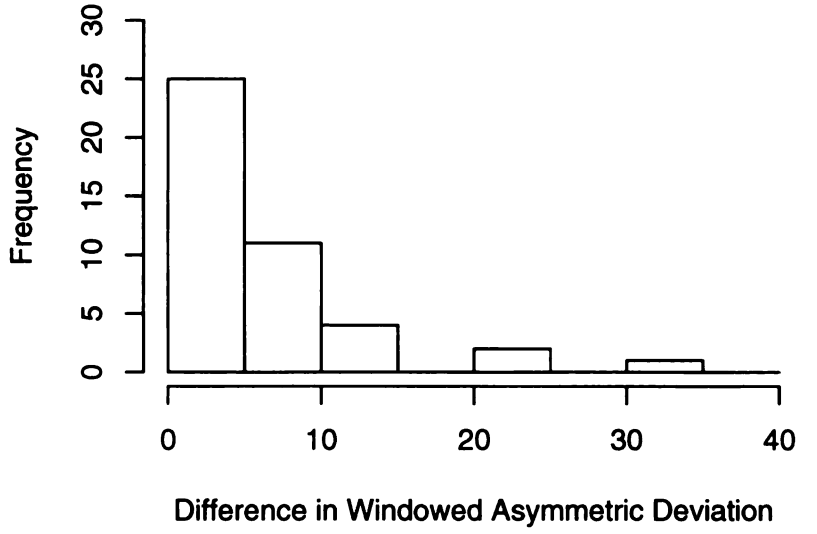

Figure 70: Frequency of differences between windowed asymmetric deviations among closely related strains of the same species. Bin size is 1. As expected, the higher, outlying windowed asymmetric deviation values correspond to comparisons among *S. pyogenes* strains.

Fig. 72 shows the relationship of the windowed asymmetric variation to IS element density. The highest correlation value relates to a first modulus sampling window on the Q series of δ = 39,000 bp, 39,500 bp, 40,000 bp, ..., 64,000 bp. I did not find significant correlation values ($r > 0.3$) with IS element density for comparative measures of average pointwise mutual information or for chromosome size.

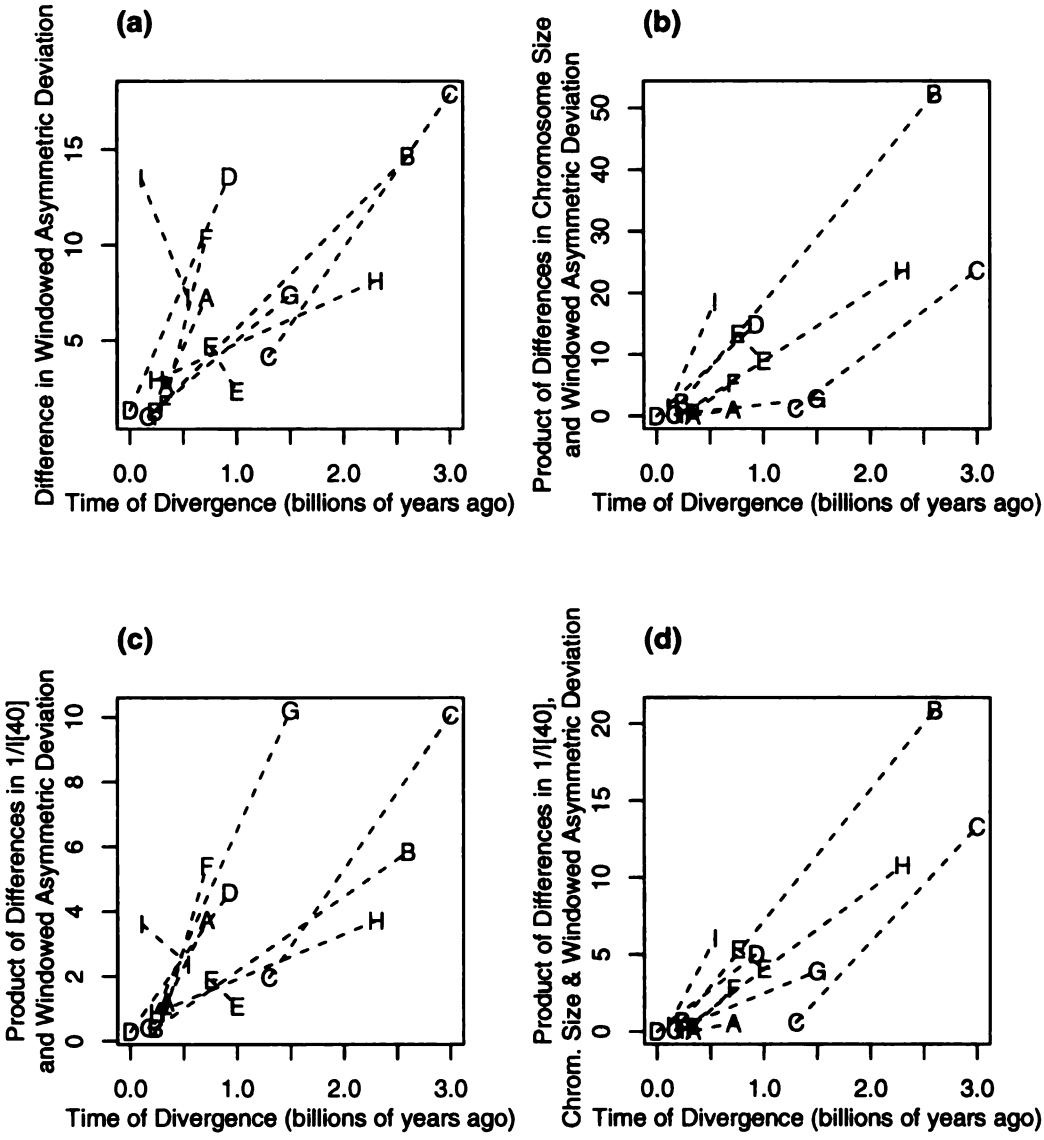

Figure 71: Relationship of divergence time from a last common ancestor to differences in chromosomal structure and organization. The windowed asymmetric deviation is based on a characteristic range of residual summed squared differences on lag k correlation series calculated from ORF densities 37,500 bp, 38,000 bp, ..., 62,500 bp. The letters correspond to comparisons from Table 5. The letter with the smaller x coordinate value is the first column comparison of Table 5. The letter with the higher x coordinate value is the average of the second and third column comparisons. **(a)** Absolute difference in windowed asymmetric deviation for various times of divergence, $m = 3.86, r = 0.66$. **(b)** Product of the absolute differences in chromosome size and windowed asymmetric deviation for various times of divergence, $m = 11.6, r = 0.77$. **(c)** Absolute difference in windowed asymmetric deviation divided by the average pointwise mutual information for 40 kb ($I[40]$), $m = 2.47, r = 0.72$. **(d)** Product of the absolute differences in chromosome size and windowed asymmetric deviation divided by $I[40]$, $m = 5.54, r = 0.87$.

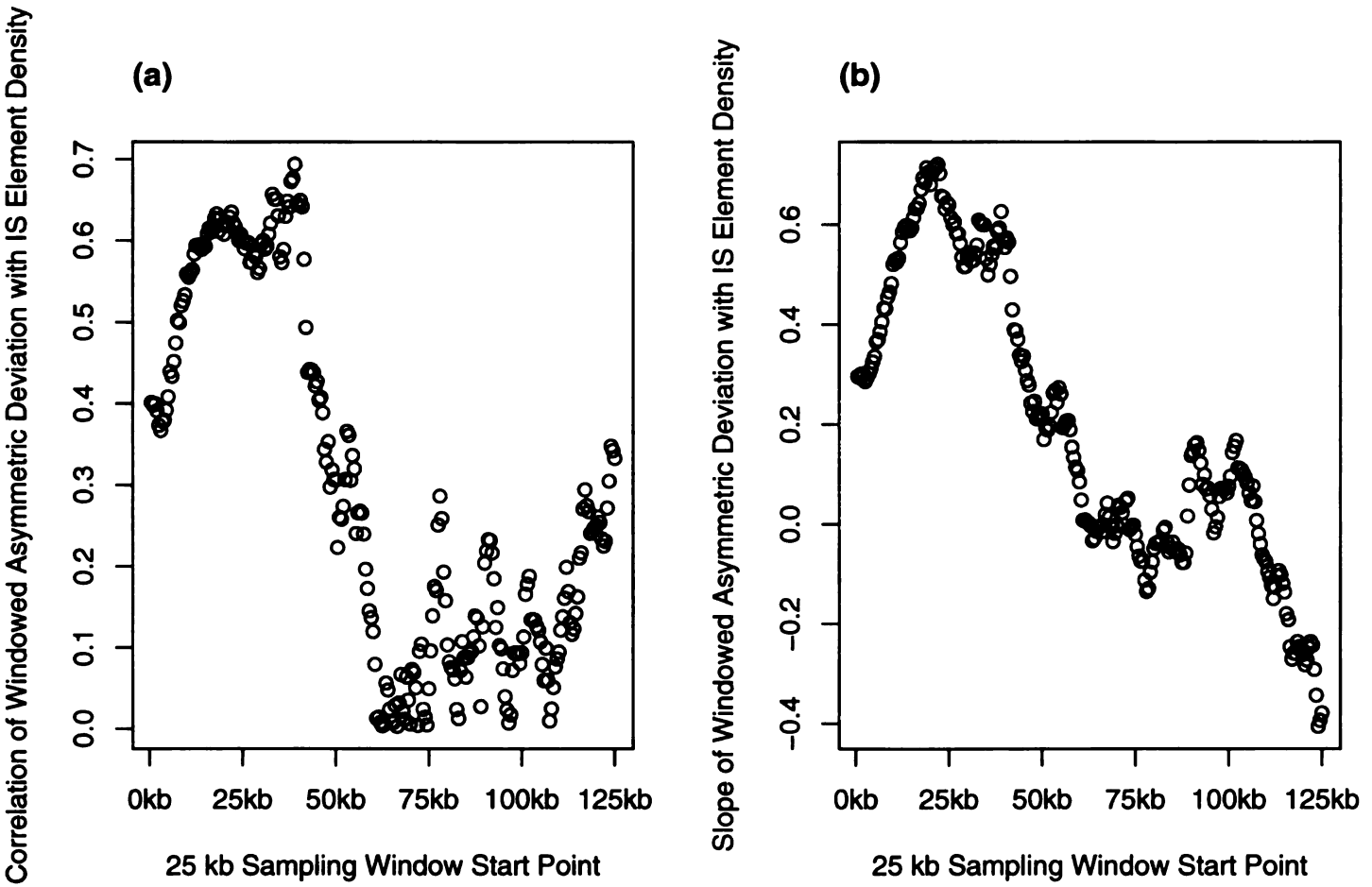

Figure 72: Relationship of windowed asymmetric deviation to IS element density. The windowed asymmetric deviation is based on a characteristic range of residual summed squared differences on lag k correlation series calculated from ORF densities over a 25 kb window of segmentation sizes. **(a)** The correlation of the windowed asymmetric deviation with median IS element density. **(b)** The slope of the windowed asymmetric deviation with median IS element density. Median IS element density is calculated for all those density ranges of IS elements per 350,018 bp for which there was more than one representative genome.

5.5 Discussion

There are a wide variety of approaches for quantifying how a measured complexity of pattern may relate to underlying dynamics (Falconer, 1997; Casdagli *et al.*, 1991; Stearns & Magwene, 2003). My final measure of a windowed asymmetric deviation may represent an advancement beyond a cross-species or cross-strain interchromosomal correlation of gene locations (Horimoto *et al.*, 2001) in that the windowed asymmetric deviation is an intrachromosomal residual value for comparison that may more directly implicate underlying functional optima and mechanistic possibilities for change relating to the clustering of ORFs. The symmetry scoring measure $Q(c_i, \delta)$ upon which the windowed asymmetric deviation is based is the result of a fairly sophisticated algorithm that adds together the squared differences along a lag k autocorrelation series and, by bootstrap, contrasts the outcome for natural, unshuffled chromosomes versus artificially shuffled chromosomes. My $Q(c_i, \delta)$ measure did not emerge through a clear axiomatic procession of analysis upon a well-parameterized model with pre-established properties, but more closely follows an inductive measurement process (Goldfarb & Deshpande, 1997). The Q measure is based on a bootstrapped contrast between sums of squared differences and, in this sense, departs from more conventional approaches involving means of squared differences. By directly measuring the absolute difference of $E(F)$ values with shuffling-based $E(X_i)$ values (Equation 13) prior to any averaging, more of the residual structure may be evaluated separate from any assumption of an interval-strength measurement property (Sarle, 1995) attributed to the $E(F)$ function. This is especially important based on the reported incidence of symmetries in the spatial clustering of gene density being potentially attributable to the skewed frequency distribution of gene densities, and not necessarily a consequence of non-shuffled chromosomal organization (Jurka & Savageau, 1985).

The initial point of empirically based induction involved observations of the lag k series. I developed two measures, $Q(c_i, \delta)$ and $P(c_i, \delta)$, to further quantify the observed invariance where the intent for each of these measures was to independently quantify non-random effects associated with localized variance on the series of ORF densities as opposed to directly correlative assays of density magnitudes involving a defined zero point. Both the $Q(c_i, \delta)$ and $P(c_i, \delta)$ measures are based on approaches frequently used in time series analyses that may

be further developed to characterize an underlying temporal nature to the formation of ORF clustering. The pseudophase space analysis of the $P(c_i, \delta)$ measure was based on an embedding dimension of two. A more refined approach to assaying invariance on a phase space would select an embedding dimension sufficient to accurately characterize nearest neighbors on the dimensional projection (Kennel *et al.*, 1992).

The model usage of my symmetry scoring measure $Q(c_i, \delta)$ was to help quantify the constraint of chromosomal expansions for a segmentation size δ. I hypothesized a harmonic relationship where consecutive chromosomal expansions of δ would implicate chromosomal expansions of $\delta \times j$ where j is an integer. If the symmetry scoring measure $Q(c_i, \delta)$ relates to the likelihood of chromosomal expansion occurring for a given δ, then a reasonable expectation would be for a harmonic effect where $Q(c_i, \delta) \approx Q(c_i, \delta \times j)$. The simulation that I constructed was a meaningful indicator as to the effectiveness of evaluating a harmonic pattern on the Q series in order to infer underyling sizes of organizational expansion. The windowed asymmetric deviation captures a one wave harmonic to characterize rising and falling from high symmetry scores to low symmetry scores.

The final set of $r > 0.6$ values in Fig. 71 demonstrates a relationship between structural and organizational features of compared chromosomes versus time of divergence from a last common ancestor. The windowed asymmetric deviation did not correlate with other measures of chromosomal structure and organization such as chromosome size and average pointwise mutual information. The windowed asymmetric deviation did correlate with time of divergence from a last common ancestor, both by itself and as a jointly considered indicator along with measured differences of chromosome size and average pointwise mutual information. The advancement in methodology represented by the windowed asymmetric deviation presents a novel capability to predict a time of divergence from a last common ancestor independent from analyses of specific conserved sequences. The only sequence analysis necessary to arrive at the windowed asymmetric deviation is to specify the locations of O-ORF translational start points as they occur on a given chromosomal sequence. My novel development of the windowed asymmetric deviation measure may be important to the objectives of a polyphasic taxonomy (Stackebrandt, 2002). Recombinations may conventionally be associated with transitions of evolutionary mode as evident from studies of

genomic plasticity (Romero & Palacios, 1997; Aras *et al.*, 2003; Fuller, 2003; Terzaghi & O'Hara, 1990), as well as the increasingly clear relationship between genome size, genomic instability and lifestyle adaptation (Ochman & Davalos, 2006; Moran & Plague, 2004). The strong trends of chromosome structure and organization for times of divergence in Fig. 71 may contrastingly implicate a fair degree of vertical ancestry possibly aligned with theoretical notions of an evolutionary tempo (Woese, 1987). A direct evaluation of functional conservation would be based on empirical data concerning viable and non-viable reorganizations of the chromosome. An assessment of functional conservation across multiple prokaryotic phyla would likely focus on common molecular factors of chromosome structure. A characterization of chromosomal organization in terms of physical base pair locations, as performed in this study, may aid in the objective evaluation of structure separate from lineage-specific distributions of other chromosomal features.

While an inductive measurement process *per se* is not hypothesis driven, the correlative findings suggest that the stability of chromosomal structure and organization can be characterized over long periods of time. Throughout my analyses, I tried to apply my various measures of chromosomal organization to various evolutionary trait software packages (Pagel, 1994; Huelsenbeck *et al.*, 2001; Ronquist & Huelsenbeck, 2003). Even by relaxing various assumptions, I had difficulty with producing a hierarchy manifesting consensus with current taxonomy. My sample size may be too small or the various measures of chromosomal organization may not yet fully characterize the heritable aspects of the complex recombinational system. Based on the sample of fully sequenced genomes, the most powerful and focused analyses would be for well-represented taxons such as the Proteobacteria and the Firmicutes. An effort for identifying possible metabolic and ecological factors associated with recombinative change, and organisms that transition between differing degrees of genomic stability, may be necessary to more meaningfully characterize branch points of divergence from common ancestors. The optimal relocationing of ORFs may require empirically-driven analyses for effects associated with physical supercoiling and expression (Deng *et al.*, 2005), and optimal expression levels for growth and fitness within the environment (Dekel & Alon, 2006).

The significance to my measures of chromosomal organization may exceed that of simple

correlation with mobile elements and times of divergence spanning billions of years of evolution. It is unexpected by chance for mobile elements and assessments of vertical ancestry to both implicate windows almost exactly the same (about 37,500 bp - 62,500 bp). The repeated implication of high average pointwise mutual information (APMI) for segmentation sizes surrounding $\delta = 40$ kb in Chapter 3 is also generally consistent with the 37,500 bp - 62,500 bp window. The product of differences between the inverse APMI for 40 kb, chromsome size, and windowed asymmetric deviation act to increase correlation with a time of divergence from a last common ancestor, and this may constitute evidence for phylogenetic covariance of these structural and organizational properties.

For the various stages of the approximated model of organizational change to the chromosome, there remain a variety of further empirically-driven treatments and efforts at mathematical modelling that may more rigorously investigate specific molecular pathways of change. The present informational analysis can also be extended by further development of the abstract, simplified system of symbolic translocations and duplications. Presently, for smaller values of the simulation model parameter T (i.e., 3 and 4), the Q symmetry score of the simulated organization is more closely comparable to a natural chromosome based on an ordinate range appearing to be predominantly between -1.0 and $+1.0$. Additional analysis would be required to further ascertain meaningful correspondences between abstract, simulated representations of chromosomal content and symmetry, and aspects of information and noise inside natural chromosomes. A principal question may be to separately account for how any large-scale periodicity to chromosomal organization relates to duplication of large segments of the chromosome versus a relationship with nucleoid superstructure (Koonin *et al.*, 1996).

Chapter 6: Summary and Conclusion

Hypotheses concerning ORF composition and organization of prokaryotic chromosomes were evaluated. Based on prior characterizations of coding content (Jackson *et al.*, 2002; Tatusov *et al.*, 2003; Skovgaard *et al.*, 2001), this study evaluated the hypothesis that 75% of annotated ORFs legitimately encode operational ORFs. This study also proposed and addressed several hypotheses concerning the symmetrical or asymmetrical nature of ORF clustering along prokaryotic chromosomes. The postulated outcome for a symmetrical pattern of ORF clustering was correspondence with vertical ancestry and the effects of mobile elements on detection of organization attributable to vertical ancestry. The findings of this study correlate well with the postulated 75% subset of ORFs that likely have phenotypic activity. In terms of a pattern of non-random clustering across 165 prokaryotic chromosomes, the organization of the operational ORFs was generally non-random in relationship to the contrasting 25% subset of non-operational ORFs. A segmentation analysis of ORF density was conducted where ORFs were counted, based on the locations of their translational start points, within consecutive segments for a given, physical segmentation length in base pairs. For most chromosomes and segmentation sizes, a significant periodic symmetry was not observed on the series of ORF density values. Yet, a pattern of similarity between neighboring lag k autocorrelation values (r_k and r_{k+1}) was evident where the correlation coefficients occurred within the range of $-0.2 < r_k < 0.2$. The weak pattern between r_k and r_{k+1} was hypothetically attributable to segmentary expansions that resulted in more equalized r_k and r_{k+1} values. Development of a model to simulate organizational expansions and modifications supported the efficacy of a proposed, hypothetical, harmonic signal measure to detect constraints on segmentary expansion. When first calibrated to a set of *Streptococcus pyogenes* strains, the harmonic signal successfully correlated with postulated outcomes for lengthy time periods of vertical divergence and the presence of mobile elements.

In the context of a dynamic analysis (Fig. 1), an avenue was explored in Chapter 4 where subpopulations of putatively "noisy" ORFs, likely not to contribute to phenotype, were identified by a basic, heuristic approach that was not lineage-specific. Although the ranges of protein lengths for an operational ORF subset (O-ORFs) and a putatively silent

ORF subset (S-ORFs) were overlapping, trends of non-normality associated with a fragmentation model of protein structure robustly supported a subset distinction of the annotated ORF set across different phyletic groupings. Across the set of 67 chromosomes for which ORFs were assigned to COGs (C-ORFs), the percentage composition of each annotated set of ORFs was analyzed. The annotated set of ORFs for each chromosome generally ($r^2 > 0.9$) consisted of a 72% subset of O-ORFs and a 73% subset of C-ORFs. The O-ORF and C-ORF subsets were not identical and, overall, 9% O-ORFs were did not belong to a COG. Functional, phenotypic, and transcriptional assays resulted in greater empirical support for the O-ORF subset to be operational compared to the C-ORF subset. Examining the underlying nature of annotated ORFs (the principal objects of evaluation) (Chapter 4) was an essential step to take prior to the reconstruction of recombinative and evolutionary dynamics attributable to ORF clustering in Chapter 5. I did not find many of the invariant characteristics of organization observed for O-ORFs to be present for either the total set of ORFs or for randomly selected subsets of ORFs.

The correlative findings of this study for O-ORF organization establish an initial measure for relating differences of chromosomal size and intrachromosomal organization to times of divergence from last common ancestors. Future advancements might jointly estimate times of divergence by a measure constructed with both 16S rRNA sequence analysis along with differences in chromosome size and intrachromosomal organization. Conservation of ORF organization appears to be global across a chromosome and conserved across diverse lineages despite substantially localized disruptions (Horimoto *et al.*, 2001).

Proposed functional barriers of conservation against recombinative change have been supercoiling, replichore balancing, and cotranscriptional effects (Mahan *et al.*, 1990). The relationship of physicochemical chromosomal topology to genomic arrangement is becoming a closely examined phenomenon where the supercoiling structure of the chromosome associates with processes of transcription and gene expression (Deng *et al.*, 2005). Estimates of physical lengths associated with supercoiling domains range from 10 kb to 100 kb (Postow *et al.*, 2004; Miller & Simons, 1993). By contrast, analyses in this study implicate narrower ranges of 40 kb or 37.5-62.5 kb as the physical ranges of segmentation sizes associated with conserved ORF organization. While the dot matrix plots of Fig. 25-33 implicate mobility of

chromosomal segments greater than 10 kb, there also appear to be individual, potentially orthologous ORFs that are distributed away from the main diagonal of conserved ORF organization. There is also a lack of significant periodic signal for long distances along the chromosome (Fig. 57). Overall, the evidence suggests that supercoiling domains do not define rigorous boundaries of ORF clustering, and this may be consistent with recent claims that the supercoiling structure is dynamic and does not represent a fixed scaffold (Deng *et al.*, 2005). A further informational study beyond conventional dot matrices and my own scalar measures of ORF symmetry may evaluate additional features such as the origin of replication for the chromosome, and the directions of transcription for each ORF. The transcriptional orientation of an ORF specific to one of the two intertwined chromosomal strands is a strongly conserved aspect of chromosomal organization, and results from my running tally method stand in direct contrast to a recent report that a significant association of transcriptional directions with replication does not occur for *Vibrio cholerae* and *Yersinia pestis* (Brüggemann *et al.*, 2003). Other types of information spanning the length of chromosomes may also be potentially evaluated; Hallin & Ussery (2004) present an online "genome atlas" where aspects such as intrinsic curvature, stacking energy, position preference, direct repeats, inverted repeats, GC skew, and percent AT are charted in concentric fashion around demarcations of ORFs. A major future objective will be to test the emergent hypotheses from informational analyses of chromosomal structure for correspondence with how lethality (Mahan *et al.*, 1990) and diversification (Vulic *et al.*, 1999) result from alterations to ORF organization.

Beyond the scope of visual atlases, comparative studies of closely related strains, and anecdotal reviews of genomic diversity, a challenge that this study sought to address was the development of a quantitative data analysis that could be efficiently applied to the growing set of fully sequenced prokaryotic genomes (Fig. 13). The rapid, ongoing increase of genomic data is a strong basis for advocating that informational analyses aid in the gathering and processing of observations. The final finding in my study was for an approximated characteristic of chromosomal organization that correlated well with vertical conservation and mobile elements, and more precise characterizations will be likely possible in the future with the greater amount of analytical power provided by a larger data set.

A genome presents not just an extant view of an organism, but may also encode an archaeology corresponding to previous states of adaptation or ancestry. In this study, simulation was used to verify some of the properties associated with calculated, residual values of ORF organization, and a sophisticated treatment based around measuring residual signal led to characterizing prokaryotic diversity to a degree that would not be expected to occur by chance. The properties of both natural and simulated variation provide evidence that the developed measures of ORF organization are not due to artifacts of observational noise or estimation error, and may represent interpretable signatures of past recombinative change. The degree and utility for chromosomal organization to relate to ancestry and divergence was significantly established, and important questions concerning conservation of information, evolutionary mode, tempo, and a legitimate polyphasic taxonomy (Zuckerkandl & Pauling, 1965; Woese, 1987; Stackebrandt, 2002) may now be more addressable.

BIBLIOGRAPHY

BIBLIOGRAPHY

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. & Carniel, E. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA*, 96:14043–14048, November 1999. 11

Aki, T. & Adhya, S. Repressor induced site-specific binding of HU for transcriptional regulation. *The EMBO Journal*, 16(12):3666–3674, 1997. 7

Allen, T. E., Herrgard, M. J., Liu, M., Qiu, Y., Glasner, J. D., Blattner, F. R. & Palsson, B. O. Genome-scale analysis of the uses of the *Escherichia coli* genome: Model-driven analysis of heterogeneous data sets. *J Bacteriol*, 185:6392–6399, 2003. 20, 144

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol*, 215:403–410, 1990. 23

Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends Biochem Sci*, 23:444–447. 23

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H. & G., K. C. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, 396:133–140, 1998. 27, 143

Andersson, S. G. E. The genomics gamble. *Nature Genet*, 26:134–135, October 2000. 7, 13

Andersson, S. G. E. & Kurland, C. G. Reductive evolution of resident genomes. *Trend Microbiol*, 6(7):263–268, 1998. 12

Aras, R. A., Kang, J., Tschumi, A. I., Harasaki, Y. & Blaser, M. J. Extensive repetitive DNA facilitates prokaryotic genome plasticity. *Proc Natl Acad Sci USA*, 100(23): 13579–13584, 2003. 174

Azad, R. K., Bernaola-Galvan, P., Ramaswamy, R. & Rao, J. S. Segmentation of genomic DNA through entropic divergence: power laws and scaling. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65:051909 Epub, 2002. 9, 10, 98

Bachellier, S., Gilson, E., Hofnung, M. & Hill, C. W. *In* Neidhardt, F. C. (ed.), *Escherichia coli and Salmonella typhimurium*, volume 2, chapter 112. Repeated Sequences, pages 2047–2066. American Society for Microbiology, 1996. 14

Bachmann, B. J., Low, K. B. & Taylor, A. L. Recalibrated linkage map of *Escherichia coli* k-12. *Bacteriol Rev*, 40:116–167, 1976. 26

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. & Leibler, S. Bacterial persistence as a phenotypic switch. *Science*, 305:1622–1625, 2004. 4

Bannantine, J. P., Zhang, Q., Li, L. L. & Kapur, V. Genomic homogeneity between *Mycobacterium avium* subsp. *avium* and *Mycobacterium avium* subsp. *paratuberculosis* belies their divergent growth rates. *BMC Microbiol*, 3:10, 2003. 1

Barbour, A. *In* Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds.), *Mobile DNA II*, chapter 41. Antigenic Variation by Relapsing Fever *Borrelia* Species and Other Bacterial Pathogens, pages 972–994. ASM Press, Washington, D. C., 2002. 8

Battistuzzi, F. U., Feijao, A. & Hedges, S. B. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol*, 4:online, 2004. 3, 4, 5, 11, 28, 38, 42, 56, 58, 96, 99

Behrens, J. T. Principles and procedures of exploratory data analysis. *Psychological Methods*, 2:131–160, 1997. 29

Bell, G. A comparative method. *Am Nat*, 133:553–571, 1989. 6

Bentley, S. D. & Parkhill, J. Comparative genomic structure of prokaryotes. *Annu Rev Genet*, 38:771–791, 2004. 1, 11, 13, 26, 31, 97, 101, 138

Bergthorsson, U. & Ochman, H. Chromosomal changes during experimental evolution in laboratory populations of *Escherichia coli*. *J Bacteriol*, 181(4):1360–1363, February 1999. 13

Bern, M. & Goldberg, D. Automatic selection of representative proteins for bacterial phylogeny. *BMC Evol Biol*, 5:34, 2005. 102

Bi, X. & Liu, L. F. *recA*-independent and *recA*-dependent intramolecular plasmid recombination; differential homology requirement and distance effect. *J Mol Biol*, 235: 414–423, 1994. 14

Biaudet, V., Samson, F. & Bessieres, P. Micado–a network-oriented database for microbial genomes. *Comput Appl Biosci*, 13:431–438, 1997. 22, 41, 131, 142

Birkland, A., Chang, K., El-Yaniv, R., Yona, G. & Sharon, I. Correcting blast e-values for low-complexity segments. *J Comp Biol*, 12:980–1003, 2005. 23, 102

Blattner, F. R., Plunkett G 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. & Shao, Y. The complete genome sequence of escherichia coli k-12. *Science*, 277:1453–1474, 1997. 24

Bockhorst, J., Craven, M., Page, D., Shavlik, J. & Glasner, J. A Bayesian network approach to operon prediction. *Bioinformatics*, 19:1227–1235, 2003. 25

Box, G. E. P. & Jenkins, G. *Time Series Analysis: Forecasting and Control*. Holden-Day, New York, 1976. 51

Brown, T. A. *Genomes*. Wiley-Liss, New York, 2002. 1, 13, 14

Brüggemann, H., Bäumer, S., Fricke, W. F., Wiezer, A., Liesegang, H., Decker, I., Herzberg, C., Martìnez-Arias, R., Merkl, R., Henne, A. & Gottschalk, G. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc Natl Acad Sci USA*, 100:1316–1321, 2003. 27, 101, 178

Campbell, A. *In* Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds.), *Mobile DNA II*, chapter 44. Eubacterial Genomes, pages 1024–1039. ASM Press, Washington, D. C., 2002. 5, 10, 12

Campo, N., Dias, M. J., Daveran-Mingot, M. L. Ritzenthaler, P. & Le Bourgeois, P. Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. *Mol Microbiol*, 51:511, 2004. 7

Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brüssow, H. Prophage genomics. *Microbiol Mol Biol Rev*, 67:238–276, 2003. 99

Casdagli, M., Eubank, S., Farmer, J. D. & Gibson, J. State space reconstruction in the presence of noise. *Physica D*, 51:52–98, 1991. 9, 10, 172

Casjens, S., Palmer, N., van Vugt, R., Huang, W., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R., Haft, D., Hickey, E., Gwinn, M., White, O. & Fraser, C. M. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the lyme disease spirochete *Borrelia burgdorferi*. *Mol Microbiol*, 35:490–516, 2000. 63

Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P. & Karp, P. D. Metacyc: A multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34:D511–D516, 2006. 145

Chambers, J. M. & Hastie, T. J. *Statistical Models in S*. Chapman & Hall, London, 1992. 37

Chung, R. & Yona, G. Protein family comparison using statistical models and predicted structural information. *BMC Bioinformatics*, 5:online, 2004. 23, 25, 102, 144

Church, K. W. & Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. 42

Cladera, A. M., Bennasar, A., Barcelo, M., Lalucat, J. & Garcia-Valdes, E. Comparative genetic diversity of *Pseudomonas stutzeri* genomovars, clonal structure, and phylogeny of the species. *J Bacteriol*, 186:5239–5248, 2004. 3

Cohan, F. M. What are bacterial species? *Annu Rev Microbiol*, 56:457–487, 2002. 1, 97

Cohan, F. M. *In* Fraser, C. M., Read, T. & Nelson, K. E. (eds.), *Microbial Genomes*, chapter 11. Concepts of bacterial biodiversity for the age of genomics, pages 175–194. Humana Press, 2004. 1, 14, 16, 96

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. Integrating high-throughput and computational data elucidates bacterial networks. 429:92–96, 2004. 41, 128, 131, 142

Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds.). *Mobile DNA II*. ASM Press, Washington, D. C., 2002. 5, 8, 10

Cummings, C. A. & Relman, D. A. Using dna microarrays to study host-microbe interactions. *Emerg Infect Dis*, 6:513–525, 2000. 22

Dalevi, D. A., Eriksen, N., Eriksson, K. & Andersson, S. G. Measuring genome divergence in bacteria: A case study using chlamydian data. *J Mol Evol*, 55:24–36, 2002. 2, 7

Darlington, R. B. *Regression and Linear Models*. McGraw-Hill, N. Y., 1990. 29

Darwin, C. *On the origin of species. A facsim. of the 1st ed., with an introd. by Ernst Mayr.* Harvard University Press, 1964, Cambridge, 1859. 3

Dawkins, R. *The Selfish Gene.* Oxford University Press, New York, 1976. 9

Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression. *Nature*, 436: 588–592, 2006. 174

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, 1999. 101

Deng, S., Stein, R. A. & Higgins, N. P. Organization of supercoil domains and their reorganization by transcription. *Mol Microbiol*, 57:1511–1521, 2005. 3, 18, 26, 174, 177, 178

Deng, W., Burland, V., Plunkett, G. r., Boutin, A., Mayhew, G. F., Liss, P., Perna, N. T., Rose, D. J., Mau, B., Zhou, S., Schwartz, D. C., Fetherston, J. D., Lindler, L. E., Brubaker, R. R., Plano, G. V., Straley, S. C., McDonough, K. A., Nilles, M. L., Matson, J. S., Blattner, F. R. & Perry, R. D. Genome sequence of *Yersinia pestis* KIM. *J Bacteriol*, 184:4601–4611, 2002. 7, 8, 11, 64, 99

Doolittle, W. F. Lateral genomics. *Trends Cell Biol*, 12:M5–8, 1999a. 2, 4

Doolittle, W. F. Phylogenetic classification and the universal tree. *Science*, 284:2124–2128, 1999b. 3

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y. I. & Hess, W. R. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A*, 100:10020–10025, 2003. 100

Duret, L. & Mouchiroud, D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci USA*, 96: 4482–4487, Apr. 1999. 21

Eyre-Walker, A. Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy? *Mol Biol Evol*, 13:864–872, 1996. 21

Falconer, K. *Techniques in Fractal Geometry.* John Wiley & Sons, New York, 1997. 172

Feeny, B. F. & Lin, G. Fractional derivatives applied to phase-space reconstructions. *Nonlinear Dynamics*, 38:85–99, 2004. 42

Franklin, N. C. *In* Hershey, A. D. (ed.), *The Bacteriophage Lambda*, chapter 8. Illegitimate Recombination, pages 175–194. Cold Spring Harbor Laboratory, 1971. 17

Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*, 26: 2941–2947, 1998. 20, 98

Fuller, T. The integrative biology of phenotypic plasticity. *Biology and Philosophy*, 18: 381–389, 2003. 174

Galagan, J. E., Nusbaum, C., Roy, A., Endrizzi, M. G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W., Zimmer, A., Barber, R. D., Cann, I., Graham, D. E., Grahame, D. A., Guss, A. M., Hedderich, R., Ingram-Smith, C., Kuettner, H. C., Krzycki, J. A., Leigh, J. A., Li, W., Liu, J., Mukhopadhyay, B., Reeve, J. N., Smith, K., Springer, T. A., Umayam, L. A., White, O., White, R. H., Conway de Macario, E., Ferry, J. G., Jarrell, K. F., Jing, H., Macario, A. J., Paulsen, I., Pritchett, M., Sowers, K. R., Swanson, R. V., Zinder, S. H., Lander, E., Metcalf, W. W. & Birren, B. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Res*, 12:532–542, 2002. 97

Garcia-Valivé, S., Romeu, A. & Palau, J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, 10:1719–1725, 2000. 21

Garrity, G. M. (ed.). *Bergey's Manual of Systematic Bacteriology, Second Edition.* Springer-Verlag GmbH, New York, 2001. 25, 133

Garrity, G. M., Bell, J. A. & Lilburn, T. G. *Bergey's Taxonomic Outline, Release 5.0.* Springer, New York, 2004. 9

Ge, Z. & Taylor, D. E. *Helicobacter pylori*: Molecular genetics and diagnostic typing. *Br Med Bull*, 54(1):31–38, 1998. 13

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. & Nelson, K. E. Metagenomic analysis of the human distal gut microbiome. *Science*, 312:1355–1359, 2006. 97

Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc Natl Acad Sci USA*, 102: 140–145, 2005. 5, 8

Glasner, J. D., Liss, P., Plunkett, G. r., Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F. R. & Perna, N. T. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res*, 31:147–151, 2003. 144

Glazko, G. V. & Mushegian, A. R. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol*, 5(5):R32, 2004. 21

Goldfarb, L. & Deshpande, S. What is a symbolic measurement process? *Proc. IEEE Conf. Systems, Man, and Cybernetics*, 5:4139–4145, 1997. 172

Grafen, A. & Ridley, M. A new model for discrete character evolution. *Journal of Theoretical Biology*, 184:7–14, 1997. 9, 97, 144

Gray, Y. H. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet*, 16:461–468, 2000. 12, 16, 96

Haack, K. R. & Roth, J. R. Recombination between chromosomal IS*200* elements supports frequent duplication formation in *Salmonella typhimurium*. *Genetics*, 141:1245–1252, Dec. 1995. 14, 15

Hallet, B. Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Curr Opin Microbiol*, 4:570–581, 2001. 12

Hallin, P. F. & Ussery, D. W. CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics*, 20:3682–3686, 2004. 178

Harrison, S. C. & Aggarwal, A. K. DNA recognition by proteins with the helix-turn-helix motif. *Annual Reviews of Biochemistry*, 59:933–969, 1990. 19

Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology*, chapter 6, pages 171–202. Oxford Series in Ecology and Evolution. Oxford University Press, 1991. vi, 4, 7, 8, 9, 15, 16, 21

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M. & Shinagawa, H. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res*, 8:11–22, 2001. 99

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., Gill, S. R., Nelson, K. E., Read, T. D., Tettelin, H., Richardson, D., Ermolaeva, M. D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R. D., Nierman, W. C., White, O., Salzberg, S. L., Smith, H. O., Colwell, R. R., Mekalanos, J. J., Venter, J. C. & M., F. C. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, 406:477–483, 2000. 12

Higgins, C. F., Dorman, C. J. & Bhriain, N. N. *In* Drlica, K. & Riley, M. (eds.), *The Bacterial Chromosome*, chapter 36, pages 421–432. ASM Press, 1990. 3, 6

Hill, C. W. & Gray, J. A. Effects on chromosomal inversion on cell fitness in *Escherichia coli* K-12. 119:771–778, 1988. 7

Hoeprich, P. D. *Infectious Diseases*. Harper & Row, Hagerstown, MD, 1972. 38, 56, 58

Horimoto, K., Fukuchi, S. & Mori, K. Comprehensive comparison between locations of orthologous genes on archaeal and bacterial genomes. *Bioinformatics*, 17:791–802, 2001. 2, 3, 6, 15, 17, 19, 20, 26, 96, 145, 172, 177

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294:2310–2314, 2001. 174

Huerta, A. M., Salgado, H., Thieffry, D. & Collado-Vides, J. RegulonDB: a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Research*, 26(1):55–60, 1997. 24

Ikeda, H., Aoki, K. & Naito, A. Illegitimate recombination mediated in vitro by DNA Gyrase of *Escherichia coli*: Structure of recombinant DNA molecules. *Proc Natl Acad Sci USA*, 79:3724–3728, 1982. 14, 19

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol*, 21:526–531, 2003. 63

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A., Chu, L., Mazur, M., Goltsman, E., Larsen, N., D'Souza, M., Walunas, T., Grechkin, Y., Pusch, G., Haselkorn, R., Fonstein, M., Ehrlich, S. D., Overbeek, R. & N, K. Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature*, 423:87–91, 2003. 63

Jackson, J. H., Harrison, S. H. & Herring, P. A. A theoretical limit to coding space in chromosomes of bacteria. *OMICS, J Integ Biol*, 6:115–121, 2002. 20, 27, 31, 142, 176

Jeong, K. S., Ahn, J. & Khodursky, A. B. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biology*, 5:online, 2004. 3, 19

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., Zhang, X., Zhang, J., Yang, G., Wu, H., Qu, D., Dong, J., Sun, L., Xue, Y., Zhao, A., Gao, Y., Zhu, J., Kan, B., Ding, K., Chen, S., Cheng, H., Yao, Z., He, B., Chen, R., Ma, D., Qiang, B., Wen, Y., Hou, Y. & Yu, J. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res*, 30:4432–4441, 2002. 8, 26

Joanes, D. N. & Gill, C. A. Comparing measures of sample skewness and kurtosis. *J Royal Stat Soc D: Statistician*, 47:183–189, 1998. 67

Jordan, I. K., Wolf, Y. I. & Koonin, E. V. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol Biol*, 3:1, 2003. 142

Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., O'Callaghan, D. & Ramuz, M. Differences in chromosome number and genome rearrangements in the genus *Brucella*. *Mol Microbiol*, 27:99–106, 1998. 12

Jurka, J. & Savageau, M. A. Gene density over the chromosome of *Escherichia coli* frequency distribution, spatial clustering, and symmetry. *J Bacteriol*, 163:806–811, 1985. 3, 26, 31, 172

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R. W., Olinger, L., Grimwood, J., Davis, R. W. & Stephens, R. S. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genet*, 21:385–389, 1999. 2, 96

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. & Lopez-Bigas, N. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 19:6083–6089, 2005. 145

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Otsuka, R., Nakazawa, H., Takamiya, M., Kato, Y., Yoshizawa, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Masuda, S., Yanagii, M., Nishimura, M., Yamagishi, A., Oshima, T. & Kikuchi, H. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. *DNA Res*, 8:123–140, 2001. 102

Kennedy, D. & Norman, C. What don't we know? *Science*, 75:309, 2005. 3, 9

Kennel, M. B., Brown, R. & Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys Rev A*, 45(6):3403–3411, 1992. 173

Kent, W. J., Hsu, F., Karolchik, D., Kuhn, R. M., Clawson, H., Trumbower, H. & Haussler, D. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res*, 15:737–741, 2005. 25

Képés, F. Periodic transcriptional organization of the e. coli genome. *J Mol Biol*, 340(5): 957–964, 2004. 26

Kimura, M. *The neutral theory of molecular evolution.* Cambridge University Press, New York, 1983. 7

Klotz, M. G. & Norton, J. M. Multiple copies of ammonia monooxygenase (amo) operons have evolved under biased AT/GC mutational pressure in ammonia-oxidizing autotrophic bacteria. *FEMS Microbiology Letters*, 168(2):303–311, 15 1998. 17

Kohno, K., Yasuzawa, K., Hirose, M., Kano, Y., Goshima, N., Tanaka, H. & Imamoto, F. Autoregulation of transcription of the hupA gene in *Escherichia coli*: Evidence for steric hindrance of the functional promoter domains induced by HU. *Journal of Biochemistry*, 115:1113–1118, 1994. 18

Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci USA*, 101(9):3160–3165, 2004. 5, 25, 101, 138, 140

Koonin, E. V. & Galperin, M. Y. *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics.* Kluwer Academic Publishers, Norwell, MA, 2003. 23, 25, 105

Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, 2001. 144

Koonin, E. V., Tatusov, R. L. & Rudd, K. E. *In* Neidhardt, F. C. e. a. (ed.), *Escherichia coli and Salmonella typhimurium*, volume 2, chapter 117. *Escherichia coli* Protein Sequences: Functional and Evolutionary Implications, pages 2047–2066. American Society for Microbiology, 1996. 20, 175

Krasnogor, N. Self generating metaheuristics in bioinformatics: The proteins structure comparison case. *Genetic Programming and Evolvable Machines*, 5:181–201, 2004. 102, 103

Kunisawa, T. & Otsuka, J. Periodic distribution of homologous genes or gene segments on the *Escherichia coli* K12 genome. *Protein Seq Data Anal*, 1:263–267, 1988. 3, 20, 26, 31

Kurland, C. What tangled web: barriers to rampant horizontal gene transfer. 27(7):741–747, 2005. 9

Kurland, C. G. Something for everyone: horizontal gene transfer in evolution. *EMBO Rep*, 11(21):92–95, 2000. 10, 25

Kurland, C. G., Canback, B. & Berg, O. G. Horizontal gene transfer: A critical view. *Proc Natl Acad Sci USA*, 100:9658–9662, 2003. 140, 144

Kutschera, U. & Niklas, K. J. The modern theory of biological evolution: an expanded synthesis. *Naturwissenschaften (online)*, 91:255–276, 2004. 3

Lahm, A. & Suck, D. DNase I-induced DNA conformation 2 Angstrom structure of a dNase I-octamer complex. *J Mol Biol*, 221:645–667, 1991. 19

Lande, R. A quantitative genetic theory of life history evolution. *Ecology*, 63:607–615, 1982. 4

Lande, R. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution*, 39:505–522, 1985. 4

Larsen, T. S. & Krogh, A. EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, 4:online, 2003. 20, 22, 26, 98, 144

Lathe, W. C. r., Snel, B. & Bork, P. Gene context conservation of a higher order than operons. *Trends Biochem Sci*, 25(10):474–479, 2000. 15, 17, 26, 101, 145

Lawrence, J. G. Gene transfer in bacteria: speciation without species? *Theor Popul Biol*, 61: 449–460, 2002. 16

Lawrence, J. G., Hendrix, R. W. & Casjens, S. Where are the pseudogenes in bacterial genomes? *Trends in Microbiol*, 9(11):535–540, 2001. 8

Lawrence, J. G. & Roth, J. R. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143:1843–1860, August 1996. 9

Leblond, P. & Decaris, B. Chromosome geometry and intraspecific genetic polymorphism in Gram-positive bacteria revealed by pulsed-field gel electrophoresis. *Electrophoresis*, 19: 582–588, 1998. 7

Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J. C. & Poch, O. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. *Genome Res*, 11:981–993, 2001. 99

Li, H., Pellegrini, M. & Eisenberg, D. Detection of parallel functional modules by comparative analysis of genome sequences. *Nature Biotechnology*, 23:253–260, 2005. 3, 18, 19, 101

Li, W. Expansion-modification systems: A model for spatial $1/f$ spectra. *Phys Rev A*, 43: 5240–5260, 1991. 10, 16, 46

Liang, P., Labedan, B. & Riley, M. Physiological genomics of *Escherichia coli* protein families. *Physiol Genomics*, 9:15–26, 2002. 22, 24, 101, 140

Lin, J., Qi, R., Aston, C., Jing, J., Anantharaman, T. S., Mishra, B., White, O., Daly, M. J., Minton, K. W., Venter, J. C. & Schwartz, D. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science*, 285:1558–1562, 1999. 12

Lindahl, E. & Elofsson, A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol*, 295:613–625, 2000. 22, 23

Liò, P., Politi, A., Ruffo, S. & Buiatti, M. Analysis of genomic patchiness of *Haemophilus influenzae* and *Saccharomyces cerevisiae* chromosomes. *Journal of Theoretical Biology*, 183:455–469, 1996. 14

Liu, J., Tan, K. & Stormo, G. D. Computational identification of the Spo0A-phosphate regulon that is essential for the cellular differentiation and development in Gram-positive spore-forming bacteria. *Nucleic Acids Res*, 31:6891–6903, 2003. 24

Lovett, S. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol*, 52:1243–1253, 2004. 14

Lunneborg, C. E. *Data Analysis by Resampling: Concepts and Applications.* Duxbury Press, Pacific Grove, CA, 2000. 8, 29

Luttinger, A. The twisted 'life' of DNA in the cell: bacterial topoisomerases. *Mol Microbiol*, 15(4):601–606, 1995. 7

Mahan, M. J., Segall, A. M. & Roth, J. R. *In* Drlica, K. & Riley, M. (eds.), *The Bacterial Chromosome*, chapter 29, pages 341–349. ASM Press, 1990. 7, 8, 177, 178

Malandrin, L., Huber, H. & Bernander, R. Nucleoid structure and partition in *Methanococcus jannaschii*: An archaeon with multiple copies of the chromosome. *Genetics*, 152:1315–1323, 1999. 12

Miller, W. G. & Simons, R. W. Chromosomal supercoiling in *Escherichia coli. Molecular Microbiology*, 10(3):675–684, 1993. 7, 177

Mojica, F. J., Charbonnier, F., Juez, G., Rodrìguez-Valera, F. & Forterre, P. Effects of salt and temperature on plasmid topology in the halophilic archaeon *Haloferax volcanii. J Bacteriol*, 176:4966–4973, 1994. 7

Moran, N. A. Accelerated evolution and muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci*, 93:2873–2878, 1996. 100

Moran, N. A. & Plague, G. R. Genomic changes following host restriction in bacteria. *Curr Opin Gen Dev*, 14(6):627–633, 2004. 1, 26, 28, 41, 92, 96, 97, 174

Morell, V. Microbiology's scarred revolutionary. *Science*, 276:699–702, 2 1997. 2

Murzin, A. G., Brenner, S. E., Hubbard, T. & C., C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247: 536–540, 1995. 22, 25

Naas, T., Blot, M., Fitch, W. M. & Arber, W. Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics*, 136:721–730, Mar. 1994. 2

Nair, S., Alokam, S., Kothapalli, S., Porwollik, S., Proctor, E., Choy, C., McClelland, M., Liu, S. L. & Sanderson, K. E. *Salmonella enterica* serovar typhi strains from which SPI7, a 134-kilobase island with genes for Vi exopolysaccharide and other functions, has been deleted. *J Bacteriol*, 186:3214–3223, 2004. 14, 17

Nakagawa, I., Kurokawa, K., Yamashita, A., Nakata, M., Tomiyasu, Y., Okahashi, N., Kawabata, S., Yamazaki, K., Shiba, T., Yasunaga, T., Hayashi, H., Hattori, M. & S., H. Genome sequence of an M3 strain of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res*, 13: 1042–1055, 2003. 100

Nakatsu, C. H., Korona, R., Lenski, R. E., DeBruijn, F. J., Marsh, T. L. & Forney, L. J. Parallel and divergent genotypic evolution in experimental populations of *Ralstonia* sp. *Journal of Bacteriology*, 180(17):4325–4331, Sept. 1998. 15

Nanassy, O. Z. & Hughes, K. T. In vivo identification of intermediate stages of the DNA inversion reaction catalyzed by the *Salmonella* Hin recombinase. *Genome Biology*, 149: 1649–1663, 2003. 8, 15

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Martins dos Santos, V. A., Fouts, D. E., Gill, S. R., Pop, M., Holmes, M., Brinkac, L., Beanan, M., DeBoy, R. T., Daugherty, S., Kolonay, J., Madupu, R., Nelson, W., White, O., Peterson, J., Khouri, H., Hance, I., Chris Lee, P., Holtzapple, E., Scanlan, D., Tran, K., Moazzez, A., Utterback, T., Rizzo, M., Lee, K., Kosack, D., Moestl, D., Wedler, H., Lauber, J., Stjepandic, D., Hoheisel, J., Straetz, M., Heim, S., Kiewitz, C., Eisen, J. A., Timmis, K. N., Dusterhoft, A., Tummler, B. & Fraser, C. M. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol*, 4:799–808, 2002. 143

Ng, I., Liu, S.-L. & Sanderson, K. E. Role of genomic rearrangements in producing new ribotypes of *Salmonella typhi*. *Journal of Bacteriology*, 181(11):3536–3541, June 1999. 13

Ng, W. V., Ciufo, S., Smith, T., Bumgarner, R., Baskin, D., Faust, J., Hall, B., Loretz, C., Seto, J., Slagel, J., Hood, L. & DasSarma, S. Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? *Genome Res*, 8:1131–1141, 1998. 11

Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature*, 409:1083–1091, 2001. 3

Nolling, J., Breton, G., Omelchenko, M. V., Makarova, K. S., Zeng, Q., Gibson, R., Lee, H. M., Dubois, J., Qiu, D., Hitti, J., Wolf, Y. I., Tatusov, R. L., Sabathe, F., Doucette-Stamm, L., Soucaille, P., Daly, M. J., Bennett, G. N., Koonin, E. V. & Smith, D. R. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J Bacteriol*, 183:4823–4838, 2001. 6, 17

Ochman, H. & Davalos, L. M. The nature and dynamics of bacterial genomes. *Science*, 311 (5768):1730–1733, 2006. 1, 11, 26, 31, 92, 96, 97, 100, 133, 145, 174

Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc Natl Acad Sci U S A*, 96:12638–12643, 1999. 1, 2, 4, 14

Ochman, H. & Wilson, A. C. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J Mol Evol*, 26:74–86, 1987. 2

Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science*, 276:734–740, 1997. 11, 144

Pagel, M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc Royal Soc (B)*, 255:37–45, 1994. 97, 174

Pagni, M. & Jongeneel, C. V. Making sense of score statistics for sequence alignments. *Brief Bioinform*, 2:51–67, 2001. 23, 102

Park, S., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, 284(4):1201–1210, 1998. 23, 102, 104

Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., Churcher, C., Mungall, K. L., Bentley, S. D., Holden, M. T., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R. M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T. T., Holroyd, S., Jagels, K., Krogh, A., Larsen, T. S., Leather, S., Moule, S., O'Gaora, P., Parry, C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S. & Barrell, B. G. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar typhi CT18. *Nature*, 413:848–852, 2001a. 8

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., Baker, S., Basham, D., Bentley, S. D., Brooks, K., Cerdeno-Tarraga, A. M., Chillingworth, T., Cronin, A., Davies, R. M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A. V., Leather, S., Moule, S., Oyston, P. C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S. & G., B. B. Genome sequence of *Yersinia pestis* the causative agent of plague. *Nature*, 413:523–527, 2001b. 99

Patterson, C. Homology in classical and molecular biology. *Mol Biol Evol*, 5:603–625, 1988. 2, 4, 16

Paulsen, I. T., Seshadri, R., Nelson, K. E., Eisen, J. A., Heidelberg, J. F., Read, T. D., Dodson, R. J., Umayam, L., Brinkac, L. M., Beanan, M. J., Daugherty, S. C., Deboy, R. T., Durkin, A. S., Kolonay, J. F., Madupu, R., Nelson, W. C., Ayodeji, B., Kraul, M., Shetty, J., Malek, J., Van Aken, S. E., Riedmuller, S., Tettelin, H., Gill, S. R., White, O., Salzberg, S. L., Hoover, D. L., Lindler, L. E., Halling, S. M., Boyle, S. M. & Fraser, C. M. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc Natl Acad Sci USA*, 99:13148–13153, 2002. 12

Pearson, K. The problem of the random walk. 72:342, 1905. 45

Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85:2444–2448, 1988. 2

Philipe, H. & Forterre, P. The rooting of the universal tree of life is not reliable. *J Mol Evol*, 49:509–523, 1999. 3

Philipp, W. J., Schwartz, D. C., Telenti, A. & Cole, S. T. Mycobacterial genome structure. *Electrophoresis*, 19:573–576, 1998. 8

Postow, L., Hardy, C. D., Arsuaga, J. & Cozzarelli, N. R. Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev*, 18:1766–1779, 2004. 7, 177

Preston, A., Parkhill, J. & Maskell, D. J. The bordetellae: lessons from genomics. *Nat Rev Microbiol*, 2:379–390, 2004. 7, 27

Pushker, R., Mira, A. & Rodriguez-Valera, F. Comparative genomics of gene-family size in closely related bacteria. *Genome Biol*, 5:R27, 2004. 41, 106, 108, 109, 110, 111, 112, 143

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 37

Ralyea, R. D., Wiedmann, M. & Boor, K. J. Bacterial tracking in a dairy production system using phenotypic and ribotyping methods. *J Food Prot*, 61(10):1336–1340, 1998. 13

Ramos, J. L., Marques, S. & Timmis, K. N. Transcriptional control of the *Pseudomonas* TOL plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. *Annual Reviews of Microbiology*, 51:341–373, 1997. 18

Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O., Hickey, E. K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S. L., Eisen, J. & Fraser, C. M. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res*, 28: 1397–1406, 2000. 8

Ren, S., Fu, G., Jiang, X., Zeng, R., Miao, Y., Xu, H., Zhang, Y., Xiong, H., Lu, G., Lu, L., Jiang, H., Jia, J., Tu, Y., Jiang, J., Gu, W., Zhang, Y., Cai, Z., Sheng, H., Yin, H., Zhang, Y., Zhu, G., Wan, Z., Huang, H., Qian, Z., Wang, S., Ma, W., Yao, Z., Shen, Y., Qiang, B., Xia, Q., Guo, X., Danchin, A., Girons, I. S., Somerville, R. L., Wen, Y., Shi, M., Chen, Z., Xu, J. & Zhao, G. Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature*, 422:888=893, 2003. 7

Reznikoff, W. S., Siegele, D. A., Cowing, D. W. & Gross, C. A. The regulation of transcription initiation in bacteria. *Annual Review of Genetics*, 19:355–387, 1985. 7, 18

Ridley, M. *Evolution.* Blackwell Scientific Publications, Inc., Cambridge, MA USA, 1993. 4

Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431:152–155, 2004. 4, 10

Roberts, R. J., Karp, P., Kasif, S., Linn, S. & Buckley, M. S. An experimental approach to genome annotation. In *American Academy of Microbiology*, 2004. 20, 22, 98

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Toloney, A., Webb, E. A., Zinser, E. R. & Chisholm, S. W. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, 424:1042 - 1047, 2003. 2, 96

Rocha, E. P. C. & Blanchard, A. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. *Nucleic Acids Res*, 30(9):2031–2042, 2002. 100

Rokas, A., Krüger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. 310:1933–1938, 2005. 97

Romero, D. & Palacios, R. Gene amplification and genomic plasticity in prokaryotes. *Annu Rev Genet*, 31:91–111, 1997. 15, 174

Ronquist, F. & Huelsenbeck, J. P. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574, 2003. 174

Royston, P. Algorithm AS 181: The W test for normality. *Applied Statistics*, 31:176–180, 1982. 67

Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W., Frishman, D., Stocker, S., Lupas, A. N. & Baumeister, W. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature*, 407:508–513, 2000. 19

Sadreyev, R.and Grishin, N. Compass: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326:317–336, 2003. 23, 102

Sadreyev, R. I. & Grishin, N. V. Quality of alignment comparison by compass improves with inclusion of diverse confident homologs. 20(6):818–828, 2004. 23

Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548, 1998. 98

Sanderson, K. E. Genetic relatedness in the family Enterobacteriaceae. *Ann Rev Microbiol*, 30:327–349, 1976. 2

Sanderson, K. E. & Liu, S.-L. Chromosomal rearrangements in enteric bacteria. *Electrophoresis*, 19:569–572, 1998. 8

Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol Biol Evol*, 19:101–109, 2002. 5

Sandvik, G., Jessup, C. M., Seip, K. L. & Bohannan, B. J. M. Using the angle frequency method to detect signals of competition and predation in experimental time series. *Ecology Letters*, 7:640–652, 2004. 9, 10, 52

Santos, S. R. & Ochman, H. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Env Microbiol*, 6(7):754–759, 2004. 11

Sarle, W. S. *Measurement Theory: Frequently Asked Questions About Measurement*, pages 61–66. Wichita: ACG Press, 1995. 172

Savageau, M. A. Proteins of *Escherichia coli* come in sizes that are multiples of 14 kda: Domain concepts and evolutionary implications. *PNAS*, 83:1198–1202, 1986. 24, 66, 98

Schidlowski, M. A. 3,800 million-year old record of life from carbon in sedimentary rocks. *Nature*, 333:313–318, 1988. 3

Schilling, C., Edwards, J. & Palsson, B. Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol Prog*, 15:288–295, 1999. 25

Schilling, C. H., Mahadevan, R., Park, S., Travnik, E., Palsson, B. O., Maranas, C., Lovley, D. & Bond, D. Simpheny: A computational infrastructure bringing genomes to life. In *Genomics:GTL Contractor Grantee Workshop IV and Metabolic Engineering Working Group Inter-Agency Conference on Metabolic Engineering*, 2006. 145

Schloss, P. D. & Handelsman, J. Status of the microbial census. *Microbiology and Molecular Biology Reviews*, 68:686–691, 2004. 96

Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. Long-term experimental evolution in *Escherichia coli*. IX. characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, 156:477–488, October 2000. 8

Schneider, D. & Lenski, R. E. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res Microbiol*, 155:319–327, 2004. 5, 6, 9, 14

Service, R. A dearth of new folds. *Science*, 307:1555, 2005. 103

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 407:81–86, 2000. 12

Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*, 16:1114–1116, 1999. 6

Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S. & Nakazawa, T. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from japan and CWL029 from USA. *Nucleic Acids Res*, 28 (12):2311–2314, 2000. 8

Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., Alves, L. M., Araya, J. E., Baia, G. S., Baptista, C. S., Barros, M. H., Bonaccorsi, E. D., Bordin, S., Bove, J. M., Briones, M. R., Bueno, M. R., Camargo, A. A., Camargo, L. E., Carraro, D. M., Carrer, H., Colauto, N. B., Colombo, C., Costa, F. F., Costa, M. C., Costa-Neto, C. M., Coutinho, L. L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A. P., Ferreira, A. J., Ferreira, V. C., Ferro, J. A., Fraga, J. S., Franca, S. C., Franco, M. C., Frohme, M., Furlan, L. R., Garnier, M., Goldman, G. H., Goldman, M. H., Gomes, S. L., Gruber, A., Ho, P. L., Hoheisel, J. D., Junqueira, M. L., Kemper, E. L., Kitajima, J. P., Krieger, J. E., Kuramae, E. E., Laigret, F., Lambais, M. R., Leite, L. C., Lemos, E. G., Lemos, M. V., Lopes, S. A., Lopes, C. R., Machado, J. A., Machado, M. A., Madeira, A. M., Madeira, H. M., Marino, C. L., Marques, M. V., Martins, E. A., Martins, E. M., Matsukuma, A. Y., Menck, C. F., Miracca, E. C., Miyaki, C. Y., Monteriro-Vitorello, C. B., Moon, D. H., Nagai, M. A., Nascimento, A. L., Netto, L. E., Nhani, A. J., Nobrega, F. G., Nunes, L. R., Oliveira, M. A., de Oliveira, M. C., de Oliveira, R. C., Palmieri, D. A., Paris, A., Peixoto, B. R., Pereira, G. A., Pereira H. A Jr, Pesquero, J. B., Quaggio, R. B., Roberto, P. G., Rodrigues, V., de M Rosa, A. J., de Rosa V. E Jr, de Sa, R. G., Santelli, R. V., Sawasaki, H. E., da Silva, A. C., da Silva, A. M., da Silva, F. R., da Silva W. A Jr, da Silveira, J. F., Silvestri, M. L., Siqueira, W. J., de Souza, A. A., de Souza, A. P., Terenzi, M. F., Truffi, D., Tsai, S. M., Tsuhako, M. H., Vallada, H., Van Sluys, M. A., Verjovski-Almeida, S., Vettore, A. L., Zago, M. A., Zatz, M., Meidanis, J. & C., S. J. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406:151–157, 2000. 143

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*, 17: 425–428, 2001. 10, 20, 22, 98, 99, 104, 113, 140, 176

Smith, J. M. *The Theory of Evolution, 3rd edition.* Cambridge University Press, Cambridge, 1975. 18

Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: The evolution of archaeal and proteobacterial gene content. 12(1):17–25, 2002. 16, 17, 21, 27, 98, 101, 140, 145

Snyder, M. & Gerstein, M. Defining genes in the genomics era. *Science*, 300:258–260, 2003. 9, 20, 21, 26, 104

Sonti, R. V. & Roth, J. R. Role of gene duplications in the adaptation of *Salmonella typhimurium* top growth on limiting carbon sources. *Genetics*, 123:19–28, September 1989. 8

Souza, V. & Eguiarte, L. E. Bacteria gone native vs. bacteria gone awry: Plasmidic transfer and bacterial evolution. *Proc Natl Acad Sci USA*, 94:5501–5503, 1997. 2, 15, 16

Stackebrandt, E. From species definition to species concept: population genetics is going to influence the systematics of prokaryotes. *WFCC Newsl*, 35:1–4, 2002. 173, 179

Stearns, S. C. & Magwene, P. The naturalist in a world of genomics. *Am Nat*, 161:171–180, 2003. 172

Stormo, G. D. & Tan, K. Mining genome databases to identify and understand new gene regulatory systems. *Curr Opin Microbiol*, 5:149–153, 2002. 24

Suerbaum, S., Smith, J. M., Bapumia, K., Giovanna, M., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci USA*, 95:12619–12624, 1998. 6

Svetic, R. E. Bandelt, H. J., Forster, P. & Röhl, A. A metabolic force for gene clustering. *Bull Math Biol*, 16:37–48, 2004. 26, 145

Tamas, I., Klasson, L., Canbäck, B., Näslund, A. K., Eriksson, A.-S., Wernegreen, J. J., Sandström, J. P., Moran, N. A. & Andersson, S. G. E. 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, 296:2376–2379, 2002. 12, 25

Tanaka, H., Goshima, N., Kohno, K., Kano, Y. & Imamoto, F. Properties of dna-binding of hu heterotypic and homotypic dimers from *Escherichia coli*. *J Biochem*, 113:568–572, 1993. 18

Tao, H., Bausch, C., Richmond, C., Blattner, F. R. & Conway, T. Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media. *J Bacteriol*, 181:6425–6440, 1999. 22

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov1, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. & Natale, D. A. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, 2003. 1, 20, 27, 46, 102, 176

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science*, 278:631–637, 1997a. 1

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science*, 278:631–637, 1997b. 23, 25, 26, 46, 102

Teichmann, S. A., Park, J. & Chothia, C. Structural assignments to the mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci U S A*, 95:14658–14663, 1998. 23, 102, 103

Terzaghi, E. & O'Hara, M. *Advances in Microbial Ecology*, chapter 11. Microbial plasticity: the relevance to microbial ecology (review), pages 431–460. 1990. 9, 27, 100, 174

Tillier, E. R. M. & Collins, R. A. Genome rearrangement by replication-directed translocation. *Nature Genet*, 26:195–197, October 2000. 14

Torretti, R. *Philosophy of geometry from Riemann to Poincaré*. D. Reidel Pub. Co., Dordrecht, Holland, 1984. vi

Ursing, J. B., Rossello-Mora, R. A., Garcia-Valdes, E. & Lalucat, J. Taxonomic note: a pragmatic approach to the nomenclature of phenotypically similar genomic groups. *Int J Syst Bacteriol*, 45:604, 1995. 3

Van Sluys, M. A., Monteiro-Vitorello, C. B., Camargo, L. E., Menck, C. F., Da Silva, A. C., Ferro, J. A., Oliveira, M. C., Setubal, J. C., Kitajima, J. P. & J., S. A. Comparative genomic analysis of plant-associated bacteria. *Annu Rev Phytopathol*, 40:169–189, 2002. 22

Vandamme, P. A. R. Polyphasic taxonomy in practise: the *Burkholderia cepacia* challenge. *WFCC Newsl*, 35:17–24, 2001. 3

Vulic, M., Lenski, R. E. & Radman, M. Mutation, recombination and incipient speciation of bacteria in the laboratory. *Proceedings of the National Academy of Sciences*, 96:7348–7351, 1999. 178

Wang, J., Masuzawa, T., Li, M. & Yanagihara, Y. An unusual illegitimate recombination occurs in the linear-plasmid-encoded outer-surface protein a gene of *Borrelia afzelii*. *Microbiology*, 143:3819–3825, 1997. 17

Wang, L., Trawick, J. D., Yamamoto, R. & Zamudio, C. Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Research*, 32(12):3689–3702, 2004. 24, 25

Wassenaar, T., Geilhausen, B. & Newell, D. Evidence of Genomic Instability in *Campylobacter jejuni* Isolated from Poultry. *Applied and Environmental Microbiology*, 64 (5):1816–1821, May 1998. 2

Watanabe, H., Mori, H., Itoh, T. & Gojobori, T. Genome Plasticity as a Paradigm of Eubacteria Evolution. *Journal of Molecular Evolution*, 44(1):S57–S64, 1997. 14, 17

Weaver, W. & Shannon, C. E. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1949. 42

Wheelan, S. J., Marchler-Bauer, A. & Bryant, S. H. Domain size distributions can predict domain boundaries. *Bioinformatics*, 16:613–619, 2000. 24, 98, 104

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusov, T. A. & Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 28:10–14, 2000. 31, 35, 42, 50, 56

Williams, G. C. (ed.). *Group Selection*. Aldine-Atherton, Chicago, 1971. 18

Williamson, R., Hetherington, J. & Jackson, J. Detection of fundamental principles and a level of order for large-scale gene clustering on the *Escherichia coli* chromosome. *Journal of Molecular Evolution*, 36:347–360, 1993. 20, 26

Wisplinghoff, H., Rosato, A. E., Enright, M. C., Noto, M., Craig, W. & Archer, G. L. Related clones containing SCCmec type IV predominate among clinically significant *Staphylococcus epidermidis* isolates. *Antimicrob Agents Chemother*, 47(11):3574–3579, 2003. 15

Woese, C. Bacterial evolution. *Microbiol Rev*, 51:221–271, 1987. 1, 2, 4, 5, 174, 179

Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*, 87: 4576–4579, 1990. 4

Wolf, D. M. & Arkin, A. P. Motifs, modules, and games in bacteria. *Curr Opin Microbiol*, 6 (2):125–134, 2003. 17, 18, 27, 144

Wolf, Y. private communication, 2004. 105

Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. 11(3):356–372, 2001. 2, 3, 26, 101, 145

Wolffe, A. P. & Drew, H. R. *In* Elgin, S. C. R. (ed.), *Chromatin Structure and Gene Expression*, chapter DNA structure: implications for chromatin structure and function, pages 27–48. Oxford University Press, 1995. 18

Worcel, A. & Burgi, E. On the structure of the folded chromosome of *Escherichia coli*. *J Mol Biol*, 71:127–147, 1972. 7

Xie, G., Bonner, C. A., Song, J., Keyhani, N. O. & Jensen, R. A. Inter-genomic displacement via lateral gene transfer of bacterial trp operons in an overall context of vertical genealogy. *BMC Biol*, 2:online journal, 2004. 2

Yang, S., Doolittle, R. F. & Bourne, P. E. Phylogeny determined by protein domain content. *Proc Natl Acad Sci*, 102:373–378, 2005. 103

Young, I. Proof without prejudice: Use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry and Cytochemistry*, 25(7):935–941, 1977. 53

Young, J. M. The genus name *Ensifer* Casida 1982 takes priority over *Sinorhizobium* Chen et al. 1988, and *Sinorhizobium morelense* Wang et al. 2002 is a later synonym of *Ensifer adhaerens* Casida 1982. is the combination "*Sinorhizobium adhaerens*" (Casida 1982) willems et al. 2003 legitimate? request for an opinion. *Int J Syst Evol Microbiol*, 53: 2107–2110, 2003. 3

Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. Genome mosaicism and organismal lineages. *Trends in Genetics*, 20(5):254–260, 2004. 2, 8, 16

Zhou, S. & Schwartz, D. C. The optical mapping of microbial genomes. *ASM News*, 70: 323–330, 2004. 12

Zivanovic, Y., Lopez, P., Philippe, H. & Forterre, P. *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res*, 30:1902–1910, 2002. 96, 99, 100, 145

Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J Theor Biol*, 8:357–366, 1965. 1, 4, 5, 179