Chromosome Architecture and Evolution in Bacteria

by

Heather Lyn Hendrickson

B.S. Biology, University of Utah, 2000

Submitted to the Graduate Faculty of
School of Arts and Sciences in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

College of Arts and Sciences

This dissertation was presented

by

Heather Lyn Hendrickson

It was defended on

August, 28, 2007

and approved by

Karen Arndt, Associate Professor, University of Pittsburgh, Biological Sciences

Graham Hatfull, Professor, University of Pittsburgh, Biological Scienecs

Roger Hendrix, Professor, University of Pittsburgh, Biological Sciences

Nancy Trun, Assistant Professor, Duquesne University, Biological Sciences

Advisor: Jeffrey Lawrence, Associate Professor, University of Pittsburgh, Biological Sciences

CHROMOSOME ARCHITECTURE AND EVOLUTION IN BACTERIA

Heather Lyn Hendrickson PhD

University of Pittsburgh, 2007

Inferences of organismal molecular evolution have been dominated by comparisons of their constituent genes. Yet the evolutionary histories of genes within Bacterial genomes are not necessarily congruent. Here, Horizontal Gene Transfer (HGT) of sequences across species boundaries can confound these analyses. There does appear to be phylogenetic cohesion, where members of higher taxonomic groups share genotypic similarity despite gene transfer. Herein I examine the rules for governing HGT to determine the impact this process has played in the evolution of Bacteria and Archaea. Bacterial chromosomes are more than simple lists of genes. Genomes must maintain information beyond component genes to direct efficient replication and segregation of their chromosomes. I propose that this structure constrains the process of HGT so that transfer among certain pairs of donors and recipients is favored. I present methods to detect this structure and new theories of bacterial cell biology and evolution based on what this structure reveals. I present evidence that bacterial chromosomes are structured by repetitive sequences termed Architecture IMparting Sequences (AIMS). AIMS are found primarily on leading strands and increase in abundance towards the replication terminus. Bacteria with robustly-identified replication origins and termini all have AIMS, and related AIMS are conserved amongst families of bacteria. We propose that AIMS are under selection to provide DNA binding proteins with polarity information, facilitating identification of the location of the replication terminus. Although AIMS evolved to direct the biology of cell division and replication, the conservation of AIMS among related taxa leads to a secondary effect. Because AIMS are counterselected when in nonpermissive orientations, AIMS constrain both intragenomic and intergenomic rearrangements. Thus HGT frequency will depend on AIMS compatibility between different species. We predict that HGT is most common between bacterial genomes which are more closely related and will impede transfer between species which have dissimilar genome architecture. The additional level of selection reflected by AIMS has resulted in cohesive bacterial groups that reflect common gene pools as a result of biased rates of gene transfer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

“Heather, not everything has to have a reason. Like, leaves being green. Oh wait. That’s not a good example. That’s chlorophyll.” ~Jeanie Hendrickson (2005)

“Heather, stop wasting my time.” ~Jeffrey Lawrence (2007)

“That’s so good it’ll… well, I am not even going to say, but it’s that good.”
~John Roth (2007)

The Buffalo Theory “…the human brain can only operate as fast as the slowest brain cells. Excessive intake of alcohol, as we all know, kills brain cells, but naturally it attacks the slowest and weakest brain cells first. In this way, regular consumption of beer eliminates the weaker brain cells, making the brain a faster and more efficient machine. That's why you always feel smarter after a few beers." ~Cliff Claven –Cheers (for the regulars at Dee’z)

My dissertation is the result of more than just the amount of time it took me to obtain a degree from the University of Pittsburgh. This document is the result of the sum of my experiences in graduate school and the influences I had during that time as well as the influences that gave me that opportunity. For that reason I am dedicating this text to the people who put me on the track towards becoming a graduate student and a scientist.

John and Shery Roth

Hal Hendrickson

Jeanie Hendrickson

Kathleen Redd-Jordan

There is no way I could have done this without my incredible mentor Jeffrey Lawrence. He and Elizabeth were like a second family to me through the 6 years I was in Pittsburgh. Jeffrey allowed me the freedom to grow as a scientist. He has opened up his home to me and allowed me to develop my talents both in and out of the laboratory. I will never be able to re-pay him. Luckily I don't need to because it is, after all, a pay it forward system. Thank you.

I was fortunate to have great colleagues in the laboratory with me while I was at the University of Pittsburgh. These included Kristen Butela, Adam Retchless and Hans Wildschutte as my fellow graduate students all of whom were great to talk to and inspired me in their own ways. Rajeev Azad brought a rigor to the bioinformatics that we were able to do and helped me see biology in a different and helpful way. I was happy to have to opportunity to get to know our many undergraduates, especially Sarah Hainer and Jenny Senge. You guys were all great to work with and I hope to see you in the future.

My experiences in graduate school were marked with many colorful characters and collaborations. I will never forget the great scientists and musicians I have known while I was here.

Many thanks to the members of my class and department, especially the Red Meat Mondays Crew, Maggie Braun, Mamie Carlson, Megan Dietz, Stephen Hancock, Jason Hoverman, Laura Marinelli & Lisa Sproul. I could not have made it through graduate school without you and your support during the best and worst times. I know you will all influence other scientists in the positive way that you have encouraged and motivated me. Thanks for all of those memories and I look forward to watching you and your families grow in the future. It is worth noting that Megan bound us together and influenced our friendships and I will always be grateful to her memory for that gift. Karen Hecht was a late arrival to Pittsburgh but she was a great companion during my last year. Thanks for the music and the muffins.

I have had the pleasure of finding great collaborators and companions through Carnegie Mellon University including but not limited to Daniel Spoonhower, Kathy Copic, Sarah Pressman, Jason Reed, Frank Broz, Chris Twigg, Moira Burke, Marek Michalowski, Daniel Lee, Phil Michel, Jenney and Matt McNaughton, Thomas Walter Murphy VII & Matt Rosencrantz.You have contributed immeasurably to the richness of my life in Pittsburgh and I am grateful to all of you for your friendship. I owe you. Rick and Enid Wood have also been a great influence on my life in Pittsburgh, they have generously shared with me their time and talents and I am happy to call them my friends.

1.0 INTRODUCTION

“The time will come,” “…though I may not live to see it, when we shall have very fairly true genealogical trees of each great kingdom of nature”. ~ Charles Darwin

To undertake a study of molecular evolution has historically meant to study the evolution of genes and by inference, organisms. Single nucleotide changes incrementally alter the function of genes over time. These alterations can eventually develop into some change at the organismal level. One of the very early insights in molecular evolution was that the magnitude of change in genes recapitulates the magnitude of change between the organisms in which those genes reside.

There is a largely ignored level at which evolution is taking place however between the gene and the organism, the evolution of the DNA molecule itself. Selection can act at the level of the chromosome for features which enable that chromosome to be managed by the cell. The result of selection at the level of the chromosome is that DNA is not merely a molecule which undergoes evolutionary processes. This thesis describes how the DNA molecule limits the rearrangements it undergoes and by extension, shapes the evolution of the organism for which it encodes.

1.1 MOLECULAR EVOLUTION AND BACTERIA

1.1.1 A Brief History of Evolutionary Study

In 1831 Charles Darwin joined the company of the H.M.S. Beagle as the ship's naturalist. In his travels he collected organisms with the intention of classifying them upon his return to England. This was a common pursuit at the time. Naturalists were interested in observing organisms to examine the diversity of life and to classify living things into hierarchical groups. In observing the affinities between organisms as well as their geographic relationships, a young Darwin was forced to turn his attention to the question of the origin of species (DARWIN 1859). What had caused the patterns of similar species that he could observe? What was the consequence of the differences between individuals that appeared to be members of the same species? Through his studies he came to the principles of natural variation as the raw material for change and natural selection as the propagating force for evolutionary change over time. Species were similar because they had descended from common ancestors through the action of natural selection on incipient variation.

Naturalists eventually embraced Darwin's theory of evolutionary change and with this philosophy in mind continued for 100 years after the publication of his 'Origin of Species (1859) to classify organisms in hopes of revealing their evolutionary relationships. Evolutionary study became a pursuit that involved collecting samples of extant organisms, establishing "types" (representatives) for each species and classifying these with an eye towards discovering how organisms were related to one another. Taxonomists used morphological, behavioral and physiological traits to infer phylogenies but these pursuits were often met with difficulty. The traits used to sort organisms were sometimes the result of environmental effects or phenotypic plasticity; organisms could look more or less similar depending on the environmental conditions they had experienced. Phenotypic variation could therefore be misleading as to the evolutionary relatedness of the organisms in question.

Darwin and his fellow naturalists were of course missing one piece of his puzzle: the mechanism behind the natural variation observed. While Darwin worked and studied in England, the basis of inheritance was being discovered quietly in a monk's garden in France by Gregor Mendel.

In 1866 Gregor Mendel first published on the laws of segregation and the independent assortment of alleles. The term 'gene' was not applied to his observations however until the 1900's when the value of his work was recognized. From this sprang the discipline of genetics including the idea that there was some biological element in every organism that established its 'genotype', and that this affected the outward appearance or 'phenotype'. By the 1940's the transformative property of DNA was recognized, making this molecule (rather than proteins or RNA) the probable molecule of genetic information (Macleod, McCarty and Hershey).

As the nature of the genetic material was being revealed contention arose between the fields of genetics and evolution. How were these fields to relate to one another? Darwin had been unaware of the mechanism of heredity and had even posited that 'blending inheritance', or the combining of parental traits, like the blending of paints, explained the source of an offspring's characteristics.

Ultimately these two schools of thought were synthesized into one. Evolution and genetics were fused and the resulting framework for the study of evolution was termed 'neo-darwinism'. By 1958, according to the account of Sir Julian Huxley, the majority of evolutionary biologists had fully embraced this way of thinking (ED. APPLEMAN 1979). Dobzhanksky put it well when he rephrased natural selection in this way, "natural selection means differential reproduction of carriers of different genetic endowments…" (ED. APPLEMAN 1979).

Neo-darwinism was a new model by which the phenotypes that had previously been used to establish phylogenies could be tied to genetic differences between types. Establishing phylogenetic relationships could be based on utilization of information directly from the genes themselves instead of using potentially flawed apparent phenotypes to determine relatedness. As molecular data became available, including DNA sequences, a new field of study emerged, molecular evolution. Molecular evolution focuses on using the sequence of biological polymers to determine how things have evolved.

1.1.2 The Synthesis: Molecular Evolution

Modern molecular evolution includes molecular phylogenetics, the use of molecular data to determine how genes are related to one another. This is a powerful tool which avoids many of the problems that plagued early evolutionary study. DNA sequences for the same gene, but from multiple organisms, can be compared and the differences can be used to infer how those genes have evolved since they were present in a last common ancestor (ZUCKERKANDL 1965;

ZUCKERKANDL and PAULING 1965).This idea relied upon the conception of a molecular clock, where by mutations would occur in DNA sequences at a uniform rate. The evolutionary history of a gene in an organism can therefore be taken as the evolutionary history of the organism itself. Since DNA is the molecule which bears genetic information among all cellular life the relationships between all such life can be addressed by molecular analysis. On a smaller scale, within species variability can be examined through its differences at the DNA level. This circumvents the necessity for scrutiny of tiny differences in morphology within the same species. In addition, the comparison of genes that encode for a particular trait can distinguish between traits that are homologous from those that are analogous. Homologous traits are those that have been derived from a common ancestor, like fore limbs in cats and dogs. Analogous traits are those that appear similar but arose independently, like eyes in mammals and the eyes of cephalopods.

The pursuit of molecular evolution has been preoccupied with the study of the gene. Consider one of the most commonly used molecular evolution techniques, nucleotide sequence alignment of a collection of homologous genes from different organisms. Such a comparison may reveal how these genes have changed or stayed the same at particular positions along their lengths. The observed differences between extant genes represent how those genes have changed since they were present in some last common ancestor of the organisms they are present in today. These differences can be seen as describing the relationship between the organisms in which they reside. Two genes with a large number of differences can indicate that they have evolved separately for longer than two more similar genes. This idea was crystallized in a scientifically rigorous manner by Zukerkandl and Pauling in 1965 in their molecular chronometers paper (ZUCKERKANDL 1965; ZUCKERKANDL and PAULING 1965). This work established that by studying the evolutionary relationships of genes one could make inferences about the organisms from which they come.

A debate that had yet to be resolved at this time was why the variation existed. Early on it was believed that the observed differences amongst members of a population represented adaptation and adaptation only. Variation in a population was thought by some to be the result of adaptation and selection for various forms. Kimura developed the idea that the differences that arose and persisted in populations were the result of neutral mutational processes. These nearly identical variants would then remain or be lost to the population by stochastic processes and random genetic drift. This was the birth of the Neutral Theory of molecular evolution, the critical recognition that the vast majority of changes that arise during the course of evolution are neutral. This was an important distinction for the resolution and final synthesis of Darwinian evolutionary principles and genetics (KIMURA 1980; KIMURA 1981; KIMURA 1983).

1.1.3 Molecular Evolution and the Bacteria

In the early 1900's as progress was being made by taxonomists using morphological differences in the multicellular world to classify organisms there was not a similar amount of success found by microbiologists. Though aware of Darwinian concepts of evolution, microbiologists did not have a rich morphological milieu from which to choose traits for classification. Physiological traits were sometimes used, but close relatives could easily lack a trait (sugar utilization, for example) and therefore be misclassified leading to confusion (GEVERS *et al.* 2006; STALEY 2006; WOESE 1987). The changes brought about by the invention of molecular evolution were therefore particularly significant for the study of bacterial evolution. In the beginning DNA-

DNA hybridization was used to infer how similar the chromosomes might be. Even today, for a new species to be described it must be grown in pure culture and an *in vitro* analysis of DNA-DNA hybridization to known type species must be performed. The threshold for species definition has been set at 70% hybridization. Anything that cannot hybridize to this degree with something previously known is defined as a new species (GEVERS *et al.* 2006).

Perhaps the biggest single contribution of molecular evolution to bacterial phylogenetics was the use of the 16s rRNA sequence to construct the universal tree of life. This sequence was chosen for two primary properties; first, it's ubiquity in the biological world and second, a slow rate of substitution, the latter owed to the necessity for conservation of the folded RNA structure. A universal sequence-based comparison for all life was first established by Fox and Woese who determined that in order to be called a species there should be no more than 97% sequence identity with published 16s RNA sequences (WOESE and FOX 1977). Establishing a universal tree allowed for taxonomic classification on a large scale and made the real scope of prokaryotic evolutionary history (the lengths and depths of the many prokaryotic branches) clear for the first time. The focus was however, still on utilizing the changes taking place in individual genes (a single gene in this case) and using these changes to make inferences about organisms.

The next revolution in molecular evolution is being brought about by the genomic era. Completely sequenced bacterial genomes represent molecular data on an entirely different scale. It is the study of the evolution of molecules as a whole, the complete chromosome that is still underappreciated in modern molecular evolution.

1.1.4 Molecular Evolution and the Chromosome

In 1995 the first genome sequence, *Haemophilus influenzae* was published (FLEISCHMANN *et al.* 1995). Since then there has been an explosion of genomic sequence data and there is no end in sight. Every year the technologies advance. It currently takes only a few weeks to completely sequence a bacterial genome using the latest technology (SMITH *et al.* 2007). The genomic era has allowed us to describe the complete genome of a bacterium. We can know all of the genes that a particular bacterium has. This provides abundant data for considering questions of what it takes to make a minimal organism (ARIGONI *et al.* 1998; KOONIN 2000; LAWRENCE 1999). We are discovering organisms living in environments we might not otherwise have imagined, and it is their genome sequences which can illuminate their strategies for survival. For example, the genome of an organism isolated from a hot spring on a Russian volcanic island, *Carboxydothermus hydrogenoformans* was found to contain five different versions of proteins which appear to function as carbon monoxide dehydrogenases, rendering this otherwise poisonous gas, carbon monoxide plus water into hydrogen and organic carbon for catabolism (WU *et al.* 2005). The catalog of genes that an organism has allows us to conjecture about environments or selective pressures experienced in the microbe's elusive lifestyle (ANDERSSON and DEHIO 2000; KLENK *et al.* 1997; SMITH *et al.* 1997).

Completely sequenced bacterial genomes are however, much more than simple lists of the genes that these organisms contain. Genomes also contain the genomic contexts in which those genes are found. This contextual information allows us to explore expression indirectly, to infer co-regulation and related function from operon structure and to design microarray experiments to examine genome wide expression change under different circumstances. Even this local genetic context however misses another level at which bacteria evolve: the level of the molecule itself.

The chromosome is a massive molecule which must be carefully handled during the life cycle in order to be accurately passed on to future generations. For a chromosome to be properly replicated and to segregate the molecule must carry information above the level of the gene and it is this additional level of complexity that shapes the evolutionary processes which can take place.

1.2 CELL DIVISION IN BACTERIA

1.2.1 Handling of Chromosomes in Cell Division

Bacteria reproduce by binary fission. Cell division has been formally studied on an individual cell level since at least 1911 (KELLY 1931). Division involves the replication of the DNA and subsequent segregation of the newly replicated DNA to two daughter cells. In Bacteria the DNA molecule is, on average 4,000,000 base pairs long. When laid out straight alongside a bacterial cell this has been estimated at about 1,000 times as long as the cell itself (KRAWIEC and RILEY 1990). The methods by which the DNA is kept within the cell are complex and unclear (BOCCARD *et al.* 2005; THANBICHLER *et al.* 2005). It is clear however, that this is the single most important molecule that a bacterium has. It is the only single molecule that, if lost or irretrievably damaged, cannot be replaced.

Therefore it is of critical importance that the DNA molecule itself be managed carefully. In bacteria the replication and subsequent segregation of the DNA molecule into two daughter cells is the most important task an individual cell has to perform in order to be evolutionarily successful. The critical nature of this task has led to the evolution of conserved systems for treatment of the DNA molecule during these processes. These systems are just starting to be elucidated through the use of Green Fluorescent Protein (GFP), single-molecule experiments and complete genome sequencing (GORDON and WRIGHT 2000; TELEMAN *et al.* 1998). These technologies have revealed that, much like eukaryotes, bacteria actively separate their DNA following the process of replication. By understanding the details of the processes of replication and segregation we can begin to grasp the important role of the chromosome itself in directing its own maintenance. Not only does the bacterial chromosome encode for the proteins that ensure it is properly segregated into two daughter cells, but the chromosome carries the signals to tell the proteins how to do this.

1.2.2 DNA Replication and Segregation in Bacteria

Undergraduate biology majors are taught about the processes of replication and segregation in eukaryotic cells. Mitosis and its phases (interphase, prophase, metaphase, anaphase, telophase and finally cytokinesis) are featured in every beginning biology text book. The reason for this is probably two fold; 1) Replication and segregation of DNA are again, the most important things a cell has to do in its lifetime. 2) This process is also trivial to visualize in the eukaryotes and therefore readily examined, even with relatively simple instruments.

The same processes in bacterial cells have been overlooked, in part because the latter is simply not true in these organisms. Bacteria are small, on average they have a diameter of about

2 μm. and a volume of roughly 4. μm^3 Eukaryotic cells, by comparison, have a diameter of approximately 20 μm, and a cell volume of 4000 μm^3 (BLACK 1996). Our ability to explore these important processes in the bacteria have been seriously limited until recently by their tiny size.

For a long time the lack of information might have led to assumptions in some circles that the bacteria did not have equivalently complex machinery for dividing their genetic material into daughter cells. However, we now know that these are dynamic and well choreographed procedures in the bacteria. The mechanisms that have evolved to handle the large scale processing of the chromosome during replication have distinct impacts on chromosome level evolution.

1.2.2.1 Chromosomes vs. Genomes

A completely sequenced 'genome' includes all DNA that is consistently replicated in a cell. In contrast, a 'chromosome' is defined as a single DNA polymer which replicates and upon which there are genes that are necessary for the life of the organism. This definition is a poor one at best but it allows for a distinction between chromosomes and plasmids, the latter of which can be quite large and still considered to be accessory. A disadvantage of this definition is that transfer events can place essential genes on very small pieces of replicating DNA and then these elements must be included in the category of chromosomes (CARLSON and KOLSTO 1994). Most chromosomes are made up of large proportions of the total genome and appear to have certain rules which guide their replication. The majority of bacterial chromosomes are circular and singular (BENTLEY and PARKHILL 2004; CASJENS 1998). There are exceptions where chromosomes are linear or paired with other circular or linear chromosomes to make up the bulk of the DNA present in the cell (BERGTHORSSON and OCHMAN 1995; BERGTHORSSON and OCHMAN 1998). I will use the word chromosome to refer to the major necessary replicon present in the bacterial genome, be it linear or circular. For the most part I will be ignoring the secondary chromosomes in the bacterial genomes that I have examined, unless otherwise noted. It has not escaped my attention that there are specific questions to ask about the evolution and maintenance of these secondary chromosomes with respect to the topic at hand, but a rigorous analysis of these effects has not been undertaken at this time (see chapter 6 for more on the subject of plasmid evolution).

1.2.2.2 Comparative Genomics

In many ways we did not have a complete understanding of the ways that chromosomes evolve until we had the powerful tool of comparative genomics. At the time of this writing there are 528 completely sequenced bacterial genomes (NIH 2007). With nearly every newly sequenced bacterial genome we add to our understanding of the diversity of life on this planet. It is rare to find similar genomes (READ *et al.* 2002). This rich resource of sequence information is the input for the study of comparative genomics. Comparative genomics is a discipline which includes contrasting the genes present or absent between sequenced bacterial genomes. This allows us to deduce ancestral states and evolutionary relationships. The dynamic quality of bacterial genome content has become clear through these studies. Comparative genomics bred hypotheses on a whole new scale. For example Losick proposed, based on the tendency for genes to be oriented such that RNA polymerases and DNA polymerases move in the same direction in many genomes, that RNA polymerases might be the driving force behind DNA segregation (DWORKIN and LOSICK 2002; EISEN *et al.* 2000). In addition, pair wise genome comparisons led to the observation that the majority of inversions appear to include the origin or the terminus (EISEN *et al.* 2000).

Comparative genomics, contrasting the genes present or absent between sequenced bacterial genomes, allows us to deduce ancestral states and evolutionary relationships. Comparisons between genomes provide insight into genomic synteny, the conservation of gene order. This comparative approach reveals genomic rearrangements at the molecular level. These include inversions, transpositions, deletions and duplications that have taken place during divergence between related bacteria.

1.2.2.3 Replication Initiation and Polymerase Action Models

Bacterial chromosomes have single origins of replication from which replication forks proceed in each direction until they either meet in the terminus region or come to linear ends. Replication initiation in *E. coli* begins at the well characterized *oriC* chromosomal location. This region contains a number of sequence motifs called DnaA boxes and I sites that are bound by DnaA and DnaA complexed with ATP, respectively. DnaA is the replication initiation protein. When bound to ATP it unwinds nearby AT–rich regions after which DnaA recruits the replicative helicase, DnaB and primase to this origin to pre-prime the way for the DNA polymerase III holoenzymes which assemble at each nascent replication fork (KAGUNI 2006). Immediately following replication in this region, hemimethylated GATC sites (also enriched in the region), are bound by

SeqA, a protein which sequesters the sites and ensures that re-initiation does not immediately occur (NORDMAN *et al.* 2007).

The location of the origin on the chromosome itself has been agreed upon for many years. A topic which is still debated is how the replication forks move though the cytoplasm and how they relate to one another. Are they coordinated in their movements or independent? In 1998 Lemon and Grossman proposed the Factory Model of DNA replication for bacteria, using a major model system, *Bacillus subtilis.* This model posits that the DNA polymerase complexes involved in replication are positionally constrained somewhere near the middle of the cell and the DNA is fed through them and then moved out from that central location. This model also suggested that if the two polymerase complexes were in close proximity then they were most likely coordinating their actions in some way. Though subsequent work has continued to support the notion that newly replicated DNA is moved away from the central site of replication, at times the claim has been to the cell poles (FEKETE and CHATTORAJ 2005; NIKI *et al.* 2000) and at times to the ¼and ¾ positions (NIELSEN 2006). The factory model has gradually lost favor to see a return the "train on a track" model of DNA polymerase motion, which prevails today. By this latter model the polymerase complex moves along the DNA to some degree and the replication forks move independently. This model has prevailed as time lapse observations of labeled DNA have become more detailed. As the intervals of observation shorten, a sometimes cyclical movement of the polymerases away from the cell center is observed (Rodrigo, personal communication). We are far from understanding the dynamics that are present at the replication forks. There does appear to be some constraint to the motion of the polymerase during replication. It is a very large complex and there is evidence of many copies of it in the vicinity of the replication fork during the replication cycle. It has been shown conclusively that the forks are not coordinated in their individual progress during chromosome replication and that one fork may continue replicating while the other sits at a lesion or replication block (BREIER *et al.* 2005; POSSOZ *et al.* 2006).

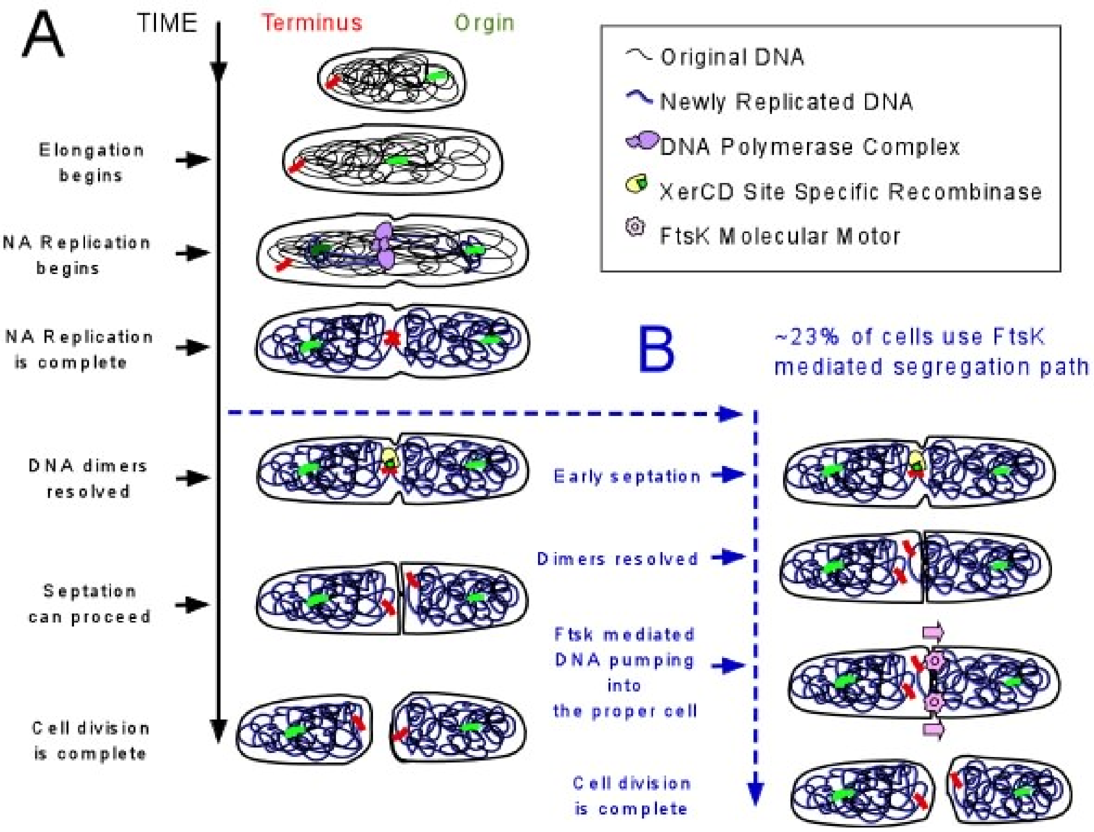
A
TIME
Terminus
Orgin
Original DNA
Newly Replicated DNA
DNA Polymerase Complex
XerCD Site Specific Recombinase
FtsK Molecular Motor
Elongation begins
DNA Replication begins
DNA Replication is complete
B
~23% of cells use FtsK mediated segregation path
DNA dimers resolved
Early septation
Dimers resolved
Septation can proceed
Ftsk mediated DNA pumping into the proper cell
Cell division is complete
Cell division is complete

Figure 1 DNA replication and segregation. A) Normal process. B) FtsK driven segregation.

1.2.2.4 Dynamic DNA Movement During Replication

As replication is taking place the DNA is not simply diffusing in the cytoplasm. The cell elongates and the origins are actively shuttled outwards (Fig 1a). Studies utilizing GFP have allowed us to examine the rapid, directed movement of DNA in cells during this process. GFP can be directed to particular regions of the bacterial chromosome to visualize genetic loci during the process of replication and subsequent segregation (SHERRATT *et al.* 2001). The origins are moved, soon after replication starts, to the ¼ and ¾ positions and these gradually travel towards their respective cell poles following this initial burst of movement. In addition to these studies, an interesting set of physical modeling experiments have been performed recently which suggest that the initial segregation of the origins of replication are not being driven towards the ¼ and ¾ positions by a combination of entropy of the unconstrained, newly replicated DNA, the compaction of the mother nucleoid during the process of replication and the free space available to the nascent DNA away from the mother nucleoid. The natural repulsive forces of these molecules for one another may be driving some of the dynamics of segregation (JUN and MULDER 2006). This is particularly interesting because chromosome segregation would have to evolve when cellular life was still somewhat simple. It has been suggested, based on the distance that different markers travel immediately after replication, that the cellular addresses of recently replicated DNA are established by the newest DNA pushing on the previously replicated DNA. This outwards motion would explain the slow migration of early markers towards the cell poles.

In addition, recent work by Sherratt and his colleagues have revealed that by the end of the time line shown in Fig 1a there is a strong tendency (¾ of the time) for the left and right halves of the chromosome, each having been replicated by a different DNA polymerase complex, to sort during the replication process such that loci for these two halves are found positioned relative to the origin and terminus in the following way: (O-L-R-T)(T-L-R-O), the other ¼ of the time being found in a non-alternating orientation (O-L-R-T)(T-R-L-O) (WANG *et al.* 2005). The significance of this variation is not known at this time but it is another instance where it appears that chromosomal partitioning during replication and segregation can be thought of as an intricately choreographed dance rather than unregulated bulk movement.

1.2.2.5 Replication Termination and Cell Segregation

In the majority of cases, once chromosome replication is completed (the details of the position of replication termination will be discussed further in chapter 3) a series of proteins assemble at what will become the septum of the bacterium at the point where the cell will become two cells. The positioning of the septum involves the cyclical trafficking of the MinCD and other proteins and the eventual aggregation of a ‘Z-ring’, a three dimensional ring about the mid-cell made up of approximately 20,000 copies of the GTPase, FtsZ. FtsZ is an ancient microtubule homologue (HARRY 2001; LI *et al.* 2003; ROTHFIELD *et al.* 1999; WEISS 2004). Once this location is defined there is recruitment to this location of the other Fts proteins, named for their initially discovered phenotype, filamentation **t**hermo **s**ensitive proteins. The Fts proteins assemble at the mid cell in approximately the order: FtsZ, FtsA, FtsK, FtsQ, FtsL, FtsI, FtsN and FtsW (ROTHFIELD *et al.* 1999). These proteins all appear to be involved in the events that are taking place at the septum, however it is not known if they form a complex. It is worth noting however that there are approximately 50 FtsZ proteins to every one of each of these latter division proteins (ROTHFIELD *et al.* 1999).

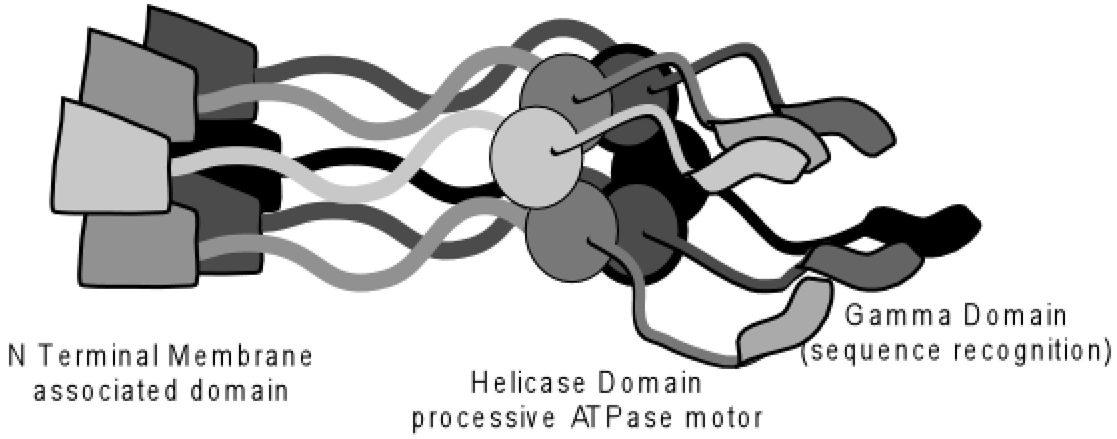

Figure 2 A diagrammatic view of the hexameric FtsK protein complex.

During the course of replication it is estimated that as many at 15% of all cells undergo recombination between the sister strands. These events will lead to dimeric chromosomes that must be resolved (CORRE and LOUARN 2005). If these structures are not resolved, the chromosome dimers may be broken at the septum and lead to cell death (CAPIAUX *et al.* 2002). Resolution of these dimers is catalyzed by a pair of enzymes that make up a site specific recombinase; XerCD. These enzymes act at the *dif* site, a 28 base pair sequence which is approximately 50% of the way across the circular chromosome from *oriC*. Effective recombination at the *dif* site requires activation by the hexameric ATPase motor, FtsK. FtsK is a large protein (1329 amino acids in *E. coli)* containing multiple domains which are important for chromosome segregation (Fig 2). The assembled FtsK hexamer has three primary domains, an N-terminal membrane associated domain, thought to mediate a connection to the inner membrane as the closing septa are coming down, a central helicase or motor domain which allows movement along the DNA towards the *dif* site and a C-terminal DNA binding gamma domain thought to recognize polar sequences that direct the motion of the hexamer towards the *dif* site. When the FtsK hexamer reaches the *dif* site it activates the action of the XerCD recombinase, separating the chromosome dimers.

The protein is extremely difficult to study in its entirety but portions of it have been studied *in vivo* to elucidate their functions. One well studied portion of this protein is the FtsK50C, which includes part of the N-terminal domain along with the last 600 amino acids of the C-terminal domain. The FtsK50C monomers are able to form a hexamer that can track directionally along DNA in response to the repeated sequences that will be discussed more in Chapter 4; KOPS or AIMS (BIGOT *et al.* 2004; BIGOT *et al.* 2005; CAPIAUX *et al.* 2002; IP *et al.* 2003; LI *et al.* 2003; MASSEY *et al.* 2006; SIVANATHAN *et al.* 2006; YATES *et al.* 2006). Along with this activation function Ftsk has been implicated as a motor protein which is able to translocate along DNA; when fixed to a membrane, the DNA would move relative to the protein, effectively being pumped into the proper daughter cell in cases where the septa have come down and trapped a portion of the chromosome in the wrong daughter cell (Fig 1B). Cytological studies using FtsK mutants have indicated that such a function might be required in many actively growing cells in culture (get this citation). FtsK also has sequence homology with *Bacillus subtilus* protein SpoIIIE, which is implicated in shuttling DNA from the mother cell to the forespore during sporulation (BARTOSIK and JAGURA-BURDZY 2005; WU and ERRINGTON 1997). It is the conserved repeated sequences, utilized by proteins like FtsK that form the basis for the evolutionary constraints that will be discussed later in this thesis.

When recombination at the *dif* site occurs in a timely manner the terminus region is able to separate and it is likely that proteins like SMC and MukB are involved in condensing the newly replicated chromosomes into their respective daughter cells (BARTOSIK and JAGURA-BURDZY 2005). The septum of the bacterium comes down between the two daughter cells and the membranes are extended between the two opposing walls (Fig 1a, bottom left). The bacterium thus becomes a pair of nearly identical bacteria that are clonally related to one another. Failure to terminate replication properly and subsequently divide leads to long filamentous tracks of cells. Deletions of either FtsK or the *dif* site also result in this phenotype (CAPIAUX *et al.* 2002; MASSEY *et al.* 2006; SIVANATHAN *et al.* 2006).

1.2.3 The Evolutionary Impact of Successful Cell Division

The problem of chromosome replication and segregation is common to all bacteria. In the previous section I have reviewed the state of knowledge of replication and segregation in two of the major model systems, *E. coli* and *B. subtilis*. However, bacteria that have single or multiple chromosomes, whether they are circular or linear, must all find ways to adapt to the problem of properly dividing up their chromosomes and separating without destroying DNA in the process. All bacteria address the problem of chromosome segregation and all bacteria have ways to handle this issue.

Having reviewed the details of chromosome replication and segregation, we now consider the cost of failure. A commonly described phenotype for failure at division is filamentation or long strings of unseparated cells. This can result from failure to properly end DNA replication, an inability to resolve dimers or catemers which have formed, or an improperly formed septum. In any case, left unresolved, this is a disastrous event for the individual cells involved. In addition, since replication had taken place, successful division was the last thing that this cell had to do to achieve evolutionary success during this round of replication.

From an evolutionary perspective, the completion of division (the successful transition from single cell to a pair of daughter cells) is the most important job that this unified group of genes had to complete: make more copies. Owing to the significance of successfully completing this process all bacteria have evolved and conserved mechanisms to handle problems which arise during the course of division. It is therefore reasonable to expect that the ways that all bacterial chromosomes are evolving are being affected by the ways they have solved the problem of chromosome segregation.

1.3 EVOLUTION AND CHROMOSOME REARRANGEMENT

1.3.1 Chromosome Rearrangements

Inversions take place through recombination at interchromosomal locations. The commonly envisaged explanation for chromosomal inversions is depicted in Figure 3. In this genome there are a pair of regions that have homology to one another sufficient that repair machinery in the cell might act at one of these sites to induce recombination with the other. Because these two homologous regions are oriented as inverted repeats the result of this event will be to reverse the order of all of the intervening genes (Fig 3A). This sort of inversions occurs in nature, particularly with naturally homologous sequences such as Ribosomal RNA operons, or tRNAs at the join points to provide the homology. Inversions do not change the content of the bacterium and are therefore more neutral events than deletions or duplications.

Even before the availability of complete genome sequences it was observed that there was a large degree of synteny or conserved gene order between closely related bacteria. Among the enteric bacteria, *Salmonella enterica* Typhimurium LT2 and *E. coli* K12 have long been a standard comparison as they diverged from one anther approximately 150 million years ago and homologous essential genes are only 10 - 20% divergent. Despite the evolution of these genomes that had taken place, it appeared that it was rare for major genome rearrangements such as large inversions to have occurred, or at least if they did occur, it appeared that they rarely survived in competition against un-inverted versions.

This idea, that the DNA molecule can undergo dynamic changes in gene order and content but not all changes are observed, is central to this thesis. There are constraints on chromosome change that are experimentally and evolutionarily observable and it is these constraints that inform us as to processes that are taking place that we would not have anticipated. Genomes must maintain information which is not simply available in the content of genes that they have and it is the preservation of this information which constrains their evolution. There was a single paper which rigorously tested constraints on events transpiring on the DNA molecule and that was Segall et al. 1988.

In order to test what the limits of inversion were in *S. typhimurium* Segall and co-workers built two test constructs to provide ample homology for inversion events across distinct locations in the chromosome to take place along with a selection for inversions which had taken place. Ultimately they found that many inversions could take place, given direct repeats providing homologous sequence for recombination, at a frequency of approximately 10^{-4}. There were however, intervals across the genome which did not naturally invert in either of the test constructs. They found that almost without exception, if the inversion included the origin region or the terminus region it could be detected at the normal frequency, however, the inversion that was closest to the terminus and did not include the terminus was never found to occur. This could apparently be constructed genetically but in culture, in competition with other un-inverted strains, it was never observed (SEGALL *et al.* 1988).

These early experiments were the first reliable observation of chromosome structure in bacteria. The authors postulated a number of possible explanations for their results including that some portions of the chromosome arrested replication in a polar manner and could not therefore, be inverted or that some aspect of chromosome maintenance physically prevented the recombination events across certain intervals. These suggestions both appear to be true. This work led to the discovery of strong polar replication terminators, the Ter sites, which are oriented to allow replication forks to move towards the terminus region but not away from it. These have been well characterized for their activity but their function in the chromosome is still debated (COSKUN-ARI and HILL 1997; HILL 1992; HILL and MARIANS 1990; NEYLON *et al.* 2005; VALJAVEC-GRATIAN *et al.* 2005). These sequences become important in my work in Chapter 4 and will be described in more detail there. Recognizing that the Ter sites existed was a direct and important result of this early work but it does not appear that Ter sites are the only information which constrains rearrangements in bacterial chromosomes.

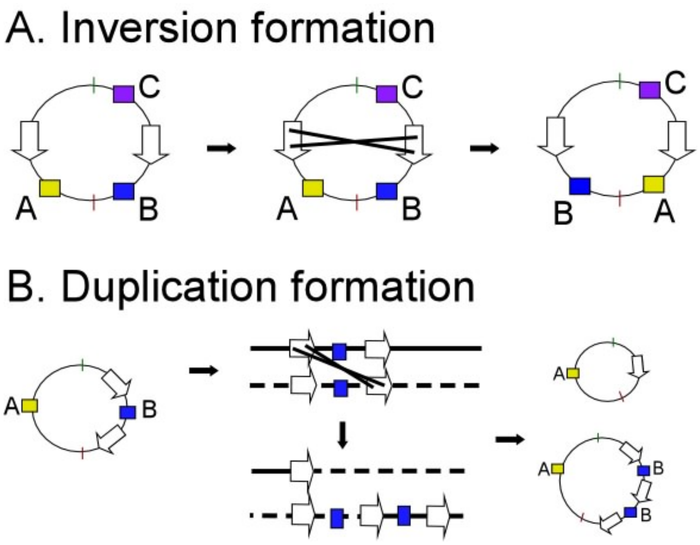

Figure 3 Recombination crates A) inversion at inverted repeats B) duplication at direct repeats.

Inversions are one way in which chromosomes can evolve that does not change the genes that are present in the bacterium. These are not however neutral evolutionary events. These early observations demonstrated that the chromosome has structure which prevents some of these seemingly passive events from taking place.

Duplications can be caused by recombination events in bacterial chromosomes as well. As chromosomes are replicated homologous sister strands are in close proximity to one another. If unequal cross-over takes place between direct repeat elements then the intervening DNA will be present in two copies in one daughter cell and absent in the other (Fig 3b). This leads to an increase in the copy number of the genes that have been duplicated, and can have long term effects if these duplicated genes are maintained and undergo selection for new functions. It has been suggested that long term maintenance of duplication events seems unlikely (LYNCH *et al.* 2001). The reversion of duplication is likely because of the huge amount of identical DNA substrate available for an additional homologous recombination event. This secondary event is therefore even more likely than the first event. Therefore, in order for a duplication to occur and be conserved over time there would have to be selection to maintain the duplication immediately, such as increased dosage of the genes involved. Though this has been observed experimentally in a special case, there is no evidence to date that this is a pathway to gene evolution in the prokaryotes in general (HENDRICKSON *et al.* 2002). The suggestion that increased dosage maintains the copy number increase, even as secondary abilities are being selected all on the same single gene is perhaps asserting too much. However, the possibility that one type of selection is going on at one locus and that a nearby locus is undergoing a transition to a different function seems reasonable. An analysis of the gamma proteobacteria has suggested that horizontal gene transfer and not duplication has contributed primarily to the expansion of gene families (LERAT *et al.* 2005). However this line of speculation runs deep into the realm of the difficult to prove. Chromosomal duplications are, none the less, simple to produce in the laboratory, observed in completely sequenced genomes and taking place spontaneously in culture conditions.

In addition to inversions and duplications another major chromosomal rearrangement that can be observed is deletions. An example of large scale deletions which have taken place during evolution and are clear in genomic comparisons is that of the *Buchnera* genus members. Comparisons between *Buchnera* genomes and that of *E. coli* show a high degree of synteny but with large patches of genes that are entirely missing. Deletions can probably be caused in a number of ways including homologous recombination events that simply take linear pieces of DNA and recombine the ends such that plasmids are formed or the DNA is recombined out and deleted. Many scenarios can be envisaged. Specific deletion machinery has not been discovered and is probably just another result of recombination and repair machinery. If deletions occur at some constant rate in nature then genes that are not under selection for function (or are under weak selection) can be removed with little effect on an organism's fitness. Deletion of extraneous chromosomal material may provide small beneficial changes in the rate of replication and segregation. The bacterial genome is rich in coding regions, much more so than in the eukaryotes. If deletions are a common occurrence then natural selection may act to eliminate wasted genetic material in these organisms.

Transposition is another type of rearrangement which occurs in bacterial chromosomes and can be observed through comparative genomic analysis. Transposons are genetic elements which include, at least, a pair of Insertion Sequences (ISs) flanking a transposase gene that encodes the recombination functions necessary to recognize and recombine the ISs. These elements are able to transport themselves to different locations in the DNA of a bacterium. This can mean physically recombining themselves out of the DNA and recombining back in elsewhere or copying themselves and moving the copy, thereby increasing their copy number as they transpose. These elements have been manipulated as genetic tools since the beginning of bacterial genetics and can either interrupt gene function or drive expression of nearby genes (KLECKNER *et al.* 1991; MULLER-HILL 1996; RAPPLEYE and ROTH 1997; WANG and ROTH 1988). In the grand scheme of recombination in bacterial chromosome rearrangements, the replicative transposons are candidates for long stretches of homology that can lead to duplications, inversions or deletions (BALBINDER 1993).

Chromosome rearrangements do not take place on an empty pallet of bacterial DNA. The study of forbidden inversions led to discoveries regarding chromosome structure. Our understanding of the Ter sites in bacterial chromosomes and the larger domains of chromosome condensation (the latter is still not well understood) have grown immensely in the past 20 years. We are still far from understanding how the chromosome itself is restricting rearrangements. Today, the field of molecular evolution is on a path towards incorporating ideas about constraints on the DNA itself and how these affect the frequency and maintenance of genomic rearrangements during evolution. One class of rearrangements in the chromosome that have been badly neglected in this respect and which have a large impact on the evolution of the bacteria is Horizontal Gene Transfer.

1.3.2 Horizontal Gene Transfer: A Potent Force in Prokaryotic Evolution.

Horizontal gene transfer is another chromosomal rearrangement and one that is such an interesting source of influence on evolution in the bacteria, it is deserving of its own section. For the purposes of this document Horizontal Gene Transfer (HGT) can be defined as the accidental and illegitimate recombination of foreign DNA into a recipient genome followed by selection and maintenance in a population (LAWRENCE 2002). There are three mechanisms by which HGT occurs in nature and they are depicted in Figure 4. Conjugation is the movement of plasmids between bacteria though conjugation tunnels, constructed by the plasmids for their propagation. Transduction involves accidental phage mediated transfer of DNA during phage infection from one bacterium to another. Transformation requires that DNA be taken up by bacteria from the surrounding medium, probably as a food source.

There have been some propositions in the literature that HGT, particularly in the case of transformation, is sometimes an intentional process. This seems to stem from the fact that competence (the ability of bacteria to take up DNA from the habitat) is a regulated process that is conserved in many branches of the bacteria. I do not favor this idea for several reasons 1) newly acquired DNA is far more likely to be hazardous to a bacterium (selfish), or simply useless, than beneficial. 2) During transformation newly acquired DNA is rendered single stranded in most competent bacteria. 3) This newly acquired DNA is guaranteed to be a source of the basic building blocks of life (food) and 4) Competence is most often turned on in limiting media where food is scarce. Given these facts it would seem that DNA uptake or competence, is a food gathering mechanism but in rare cases, accidents happen and DNA can be recombined into the genome.

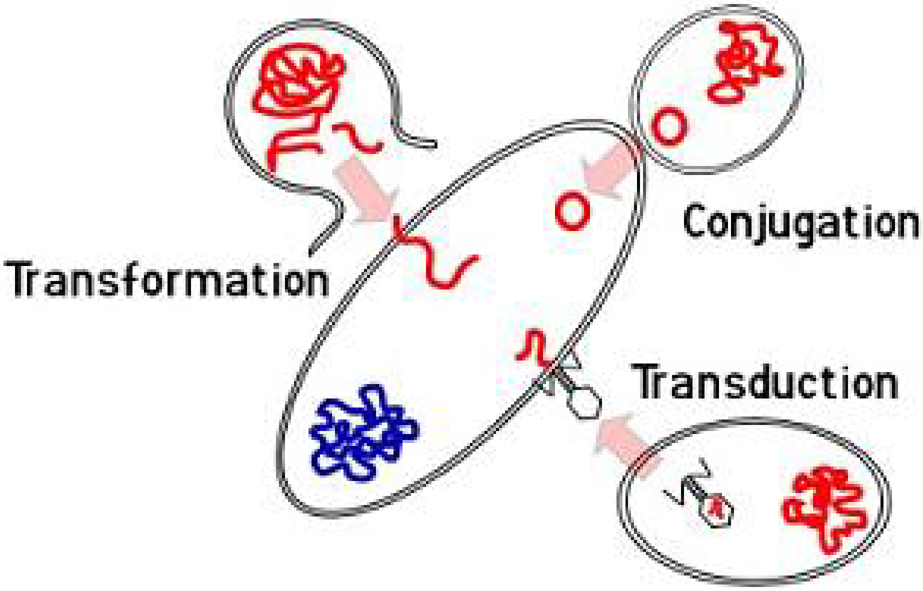

Figure 4 Horizontal gene transfer; transformation, conjugation and transduction.

Once new DNA has been brought into a cell, recombination events must take place to incorporate it into the genome. Homologous recombination leading to gene replacement can take place if the donor and the recipient are sufficiently similar. This sort of transfer is possible but made less likely by recombination limitations which will be discussed later in the introduction. When recombination does occur, it may bring in entirely new genes or copies of genes which already exist elsewhere in the genome. This latter class would appear upon first inspection to be duplications or paralogs in their new context. In either case new genes, if expressed, can lead to entirely new abilities for bacteria in which they are acquired. Niche expansion can occur in a single step instead of waiting for the slow progress of mutation and selection.

It is estimated that as much as 24%of the deadly 'Jack-In-The-Box' strain of *E. coli,* O157-H7 genome has been brought in recently (1,257 Sakai unique, 3963 shared between *E. coli* 015737 Sakai and *E. coli* K12) (WICK *et al.* 2005). Recent reports describe *Acetinobacter baumannii,* a pathogen plaguing American troops in the current Iraq war, as having acquired 17% of its genes recently, in 28 separate islands. 16 of the 28 islands are apparently involved in the devastating virulence of this strain. Though these are likely extreme cases, an average of 6.6% newly acquired genes has been reported in an analysis of 17 distinct genomes (OCHMAN *et al.* 2000).

The amount of HGT that is inferred for a particular genome varies widely with the method used to detect the HGT. There are two primary methods used to detect HGT, parametric and phylogenetic methods. Parametric methods are those that depend on the physical qualities of the DNA in a particular organism to determine if parts of the genome are atypical (Fig. 5A). Every genome has physical features, the result of years of mutation with resident polymerases, repair with native repair machineries, and selection for transcription and translation efficiency. An unbiased model of usage would predict 25% usage for each of the 4 nucleotides in DNA. However, the majority of genomes show unequal usage of bases or asymmetry in composition

(LOBRY and SUEOKA 2002). Nucleotide composition can range from 25% GC in *Mycoplasma* to 75% in *Micrococcus* (LAWRENCE and OCHMAN 1997). Another measure of atypical genes, codon usage bias, is reported in terms of selection strength. The strength of selection on codon usage bias whereby, particular codons are preferentially used to encode amino acids, can also vary widely between genomes (SHARP *et al.* 2005). Other, more subtle parametric measurements of atypical DNA composition have been utilized as well and combining techniques can lead to a refinement in the ability to accurately identify modeled or artificial HGT (AZAD and BORODOVSKY 2004; AZAD and LAWRENCE 2005; HAMADY *et al.* 2006).

Acquisition of the characteristics of the native DNA, or amelioration, through replication and mutation in the new genomic context, will be a gradual and inevitable process once HGT has occurred (LAWRENCE and OCHMAN 1997; OCHMAN and LAWRENCE 1996). This amelioration process limits the effectiveness of parametric methods to detect very anciently transferred DNA. An additional caveat to detection based on atypical qualities is that these methods assume that DNA was detectably atypical when acquired. DNA which has come from more closely related organisms will not have as strong an atypical DNA signal and may be completely lost depending on the thresholds set in these analyses (KOONIN *et al.* 2001; KOSKI *et al.* 2001; RAGAN 2001).

The other major class of methods for detection of HGT in genomes can be called the phylogenetic methods. Sequence data can be used to infer trees of relatedness for genes or genomes. For a single gene, present in many genomes, sequence similarities can be compared in order to determine how genes have evolved (GALTIER *et al.* 1996; THOMPSON *et al.* 1997; THOMPSON *et al.* 1994). However, in these trees for several genes within the same organism, one observes that different genes will produce different phylogenies (CRAWFORD and MILKMAN 1991; MÉDIGUE *et al.* 1991; REEVES 1993; SMITH *et al.* 1992). This observation is also a signifier that HGT has taken place and resulted in genomes that are essentially chimeras, containing genes with disparate evolutionary histories. A majority signal will define a single tree for many genes in a genome but there are a multitude of evolutionary histories present in a single genome.

Phylogenetic methods for HGT identification invoke parsimony, or 'simplest explanation', arguments to determine what genes, observed sporadically in trees, have been transferred in recently and which are the longer term residents. A simple example is given in Fig 5B. In this example, if the gene that is present in only one branch of the species tree depicted was present in the last common ancestor of all of these, then at least three separate loss events would need to be invoked to explain the irregular presence of the gene (inferred losses are denoted by the three x's). However the more parsimonious explanation is that a single gain of gene event occurred along the branch of the organism that has the gene currently. The complexity of this sort of argument increases with the number of genes being considered and the way that the trees themselves were built. The 16srRNA tree (see discussion below) is often used to establish relationships between organisms across the domains of life, however even this paragon of pedigree determination is susceptible to horizontal gene transfer (SCHOULS *et al.* 2003).

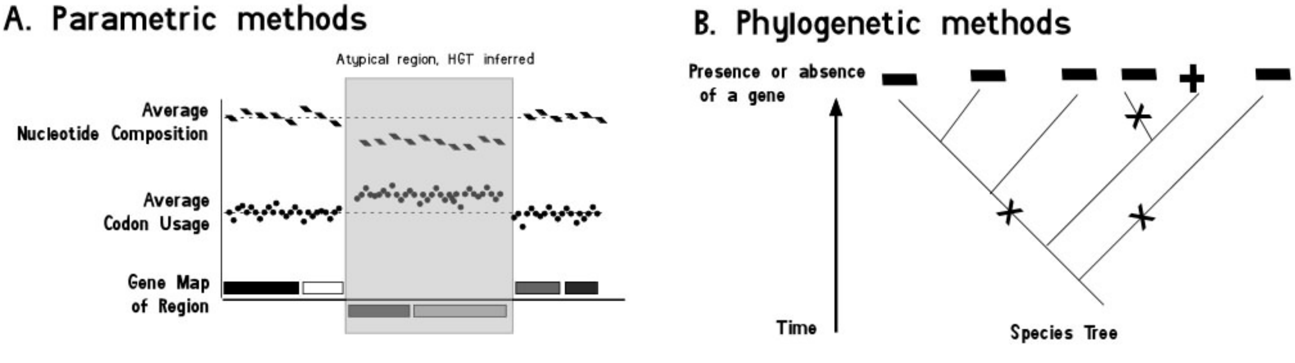

Figure 5 Parametric (A) and phylogenetic (B) methods for detection of HGT.

Naturally, we must recall that in nature there will be a simplest explanation for the patterns that we observe but that does not mean that the most parsimonious account is true. The two major methods for detection of HGT each have their own caveats and these actually lead to incompletely overlapping sets of predicted HGT for the same genomes (POPTSOVA 2007). Despite these conflicts it seems clear that HGT is, in fact taking place. By either method, newly acquired DNA is observed in nearly all genomes and as such, it is likely playing an interesting and important role in the evolution of the majority of life on the planet. Not only do nearly all bacterial genomes show evidence of HGT as an ongoing part of their evolution, but there do not appear to be genes that are immune to this process. The housekeeping genes apparently transfer less, probably because they are involved in complicated complexes or have dependencies on other genes that would not simultaneously transfer with ease and are therefore more difficult to replace. However, even these have been observed to transfer on occasion (JAIN *et al.* 1999). Although it is difficult to measure in real time, the completely sequenced genomes that we have tell us that HGT is widespread and frequent on an evolutionary time scale.

The observation of HGT requires that the novel DNA was available, integrated into the genome, was transcribed and translated, had no deleterious effect on replication and segregation and finally, advantageously affected the individuals carrying it to spread through the population. Many of these steps make up the commonly accepted model for HGT. Novel to this description is the absence of deleterious effect on the processes of replication and segregation. The notion of a conserved architecture of the chromosome which can absorb limited perturbations has not been brought to bear on the field of HGT. This has been examined in the case of inversions but once the Ter sites were identified it was thought that the search must be over. There is a more subtle, sequence based, architecture which must be maintained in the face of genome rearrangements in order for efficient cell division to occur. The architecture and its properties are described in Chapter 3. Interestingly, the consequences of the particular constraints that are imposed by this sequence architecture are salient to a previously troubling question raised by the universal tree of life.

1.3.3 The Conundrum of the Universal Tree of Life

Bacteria have always been trivial to assay biochemically and classify but these classification schemes have often proved empty of meaningful information about these organisms' phylogenetic relationships (WOESE 1994). The era of molecular evolutionary study helped to correct many of the errors in classification that had arisen. Even with these clarifications and tools there are complications to delineating a species concept in bacteria. Some of these include our lack of complete sampling and the fact that we know that HGT is taking place and interfering with taxonomic signals (KONSTANTINIDIS *et al.* 2006; MORENO 1997; STALEY 2006). When protein trees are analyzed for support of the three major domains of life only 20-50% of trees support monophyly.

Much of this thesis is dedicated to the elucidation of a previously unrecognized set of rules which govern HGT that are imposed by the DNA molecule itself. Why propose that such rules or constraints exist? One reason is that by two separate methods whereby whole genome sequence information is used to construct phylogenetic trees, there is support found for a single universal tree of life. One method used is the presence or absence of common genes as a shared, derived character trait for building trees (FITZ-GIBBON and HOUSE 1999; SNEL *et al.* 1999). The second method relies on the combination of many conserved genes together as a single "mega sequence" for analysis. Both methods show support for the three major domains of life being distinct. By that account it would appear that HGT has not had a significant effect on the evolution of these organisms. However, despite this fact we know that HGT and even long distance, or cross domain transfer, does occur (GOGARTEN *et al.* 1996; JAIN *et al.* 1999; MAKAROVA *et al.* 1999; OLENDZENSKI *et al.* 1998). In fact, as stated previously, newly acquired DNA can account in some cases for up to 24% of a single bacterial genome.

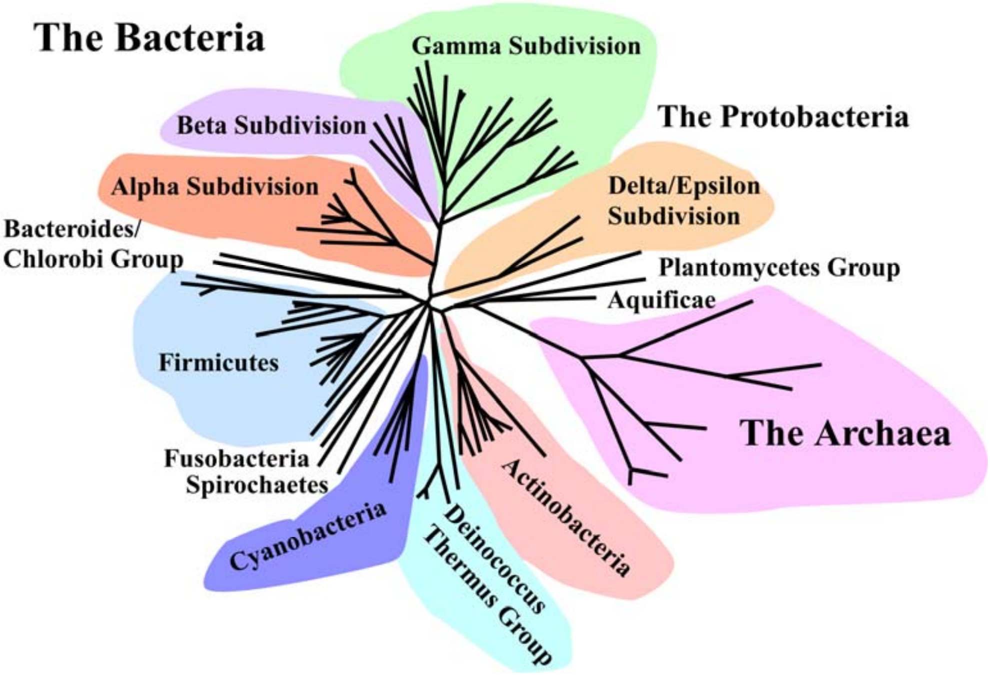

Figure 6 A 16s rRNA tree for Bacteria and Archaea.

If HGT is frequent and ubiquitous then why are the separations between the three domains of life so clear (Fig. 6)? What has lead to the retention of succinct families of bacteria that have traits as well as phylogenetic signals which speak to their cohesiveness as distinct groups? Why has HGT not rendered the microbial world to a web of inter-related and ill-defined blends of bacteria?

The answer to these questions likely lies in the frequency of transfer between groups. If HGT was a completely random process we might expect a web-like tree of life, since this is not what we observe, we can conjecture that there must be rules which govern this process which we have yet to elucidate. In this thesis I propose that the cohesion of bacterial families in the face of frequent HGT is due to frequent successful transfer among closely related groups, in conjunction with less frequent successful transfer between distantly related groups. The constraint to transfer that I propose is a simple, quantifiable and nearly universal cell biology limitation on what DNA that can be brought into chromosomes.

1.3.4 Quantifiable Constraints on the Process of Horizontal Gene Transfer

The particular constraint that I am describing falls out of a consideration of the events that must take place in order for HGT to be successful and observed (by any method):

1) Novel DNA enters the cytoplasm.
2) Recombination brings DNA into the chromosome.
3) Replication and segregation of the chromosome are intact.
4) Transcription and or translation are possible in this genomic context.
5) Positive selection or at least a lack of strong negative selection for new DNA.
6) New DNA can spread in the population.

It is clear that all of these steps must occur in order for HGT to be observed, consider the constraints on this process that might arise at each one;

1) Novel DNA enters the cytoplasm.

HGT begins with the entry of new DNA into the cytoplasm of the recipient cell. This takes place through conjugation, transduction or transformation. The relative frequencies at which these types of events occur in nature is not known. Frequencies of conjugation, transduction and transformation will depend on a number of unknown factors. In order for conjugation to take place two bacteria must come into close proximity and a conjugal tube must be maintained between the two of them for long enough to transfer at least some DNA. Little is known of the recipient range for conjugal plasmids. Similarly, natural transduction frequencies will depend first, on the occurrence of accidental incorporation during the phage lytic cycle of bacterial DNA and second, on the subsequent infection by that phage of a recipient in its range. The third event, transformation requires that the DNA from a donor be available in the recipient's environment and in most cases that the recipient turn on a transformation system that will bring the DNA into the cytoplasm directly. As I mentioned previously, transformation appears to be turned on in starvation situations which may be common in nature but transformation itself appears to be sporadically present in extant lineages and dependent on a variety of recognition sequences which may be selecting for closely related strains to be consumed before less related strains (REDFIELD 1988; REDFIELD 1993; REDFIELD 2001; REDFIELD *et al.* 1997).

2) Recombination brings DNA into the chromosome or other replicon.

There are a number of recombination systems and constraints that we are aware of at this time. The primary recombination protein in bacteria cells is RecA (CLARK 1991; EGGLESTON and WEST 1997; MAISNIER-PATIN *et al.* 2001; SMITH *et al.* 1995). RecA is not only the main homologous recombination protein but also acts as a regulator of the SOS response. In recombination RecA binds to the 3' end of single stranded DNA and scans local sister strands for homology for strand alignment and duplex formation. Once single stranded DNA is bound, RecA is in an activated state that is able to induce auto-cleavage of LexA, the SOS response regulator. The recombination machinery in bacterial chromosomes can be thought of as repair machinery, able to detect broken chromosomes and use homology to repair them (KUZMINOV 1999; SMITH *et al.* 1995). If enough breaks have taken place the SOS response can be turned on or amplified. The SOS response includes an array of ~20 repair proteins that act in cases of extreme DNA damage in the enteric bacteria and presumably others as well. DNA replication is halted and error prone polymerases and lesion repair proteins are turned on in a step wise manner. These genes are expressed in order to repair the offending DNA. As these emergency measures are enacted, accuracy is foregone and mistakes are made (MAGEE *et al.* 1992). It is not clear if SOS is induced during the process of HGT since it is so difficult for us to examine this process directly in nature. SOS induction is probably not strictly necessary for transduction, transformation and conjugation. However, high rates of these events are observed in RecA+ cells (MAJEWSKI *et al.* 2000; MIESEL and ROTH 1996; SMITH *et al.* 1995; ZAHRT and MALOT 1997).

For extremely close pairs of bacteria exchanging DNA, the mismatch repair system establishes a barrier which constrains transfer. This system involves recognition of mismatched DNA bases or insertion/deletion mismatches between sister DNA strands during DNA replication. Incongruities between two strands are repaired by establishing which one is the older strand and using that as the guide for repair. GATC sites along the DNA are methylated by a member of this repair system and so the methylated strand is marked as the older of the two strands (VULIC *et al.* 1999). The DNA mismatch repair system is highly conserved across the domains of life.

By way of example, *S. enterica* and *E. coli* cannot readily exchange DNA in laboratory experiments until the mismatch machinery, the Mut proteins, are deleted (RAYSSIGUIER *et al.* 1989; STAMBUK and RADMAN 1998; ZAHRT and MALOT 1997). However it appears that despite this constraint, homologous recombination of transferred DNA, leading to gene replacement does still occur (SCHOULS *et al.* 2003). There is a homology length limit to the mismatch correction system (MAJEWSKI and COHAN 1999; ZAWADZKI *et al.* 1995). Once this limit is reached HGT will occur by non-homologous recombination and be illegitimate in nature.

In the face of homologous recombination barriers and a lack of homology between very distant organisms it would seem that illegitimate recombination, that which by definition does not require sequence identity, is the rule for successful HGT (GOGARTEN *et al.* 2002).

In addition to these recombination mechanisms, the power of integration through site specific recombination systems, particularly phage integration to form prophages must not be overlooked. Recombination in this case happens through the action of integration sites like *attP* and *attB*. When these sites are compatible between chromosomes and plasmids or phage and with the help of a host integration factor, circular DNA can be brought into the chromosome. This type of recombination may be most frequent when there is a natural association between the players, as a pair of specific DNA sequences must be maintained in order for the integration to take place (GARCIA-RUSSELL *et al.* 2004; GHOSH *et al.* 2003; PIERSON and KAHN 1987).

It should be recalled that plasmid acquisition by transformation does not demand that the newly acquired DNA be incorporated into the genome. In such cases a mechanism for stable maintenance of the new DNA will depend not on recombination but on the plasmids' ability to replicate independently in the cell. There are two ways that this might occur: either replication initiation factors, present in the cytoplasm must recognize and act on the newly acquired replicon, or transcription or translation of the plasmids' own replication factors must be possible. Either of these possibilities would seem more likely if the donor and the recipient are closely related. Notably, plasmid acquisition is also complicated by incompatibility factors. The presence of certain plasmids will prevent additional plasmids, those that are too similar, from being acquired subsequently (ZAWADZKI *et al.* 1996).

Theses issues; recombination in the face of mismatch repair, integration specificity, plasmid maintenance and plasmid incompatibility are all issues which complicate the quantification of HGT in nature. However, it is also likely that they all favor transfer events which take place between closely related organisms. Though interesting, these are all currently difficult to quantify. The inability to calculate them makes their effects on a constraint on HGT difficult to elucidate. These are factors that are certainly affecting the frequency of HGT and they highlight the value of finding a constraint on transfer that is quantifiable, thereby empowering an exploration of the effect on evolution and an explanation for the cohesion of bacterial families in the face of HGT.

3) Replication and segregation of the chromosome is intact.

If replication of the chromosomes and subsequent segregation of the DNA into two individual daughter cells in an individual bacterium are interfered with by a newly acquired piece of DNA this HGT event will likely not increase in frequency in a population of superior competitors. This is a level of constraint on HGT that has not previously been discussed in the literature and which will, in part, be described in the later chapters of this work. The limitation that I will be formally describing involves the chromosomal architectures imposed by the AIMS sequences mentioned earlier and discussed in detail in this dissertation. AIMS-like sequence architecture is likely important for the processes of replication and segregation in the majority of bacterial genomes.

Other, less predictable events could interfere with chromosome replication and segregation immediately following the acquisition of new DNA. If, for example, there were structures in the incoming DNA which interfered with DNA polymerases such as hair-pins, such a transfer would be selected against. Additionally if particular sequences within the chromosome, such as the origin or terminus of replication were interrupted, these would likely be problematic HGT events. There are likely innumerable other accidents one could imagine. These share the quality of being random and rare and ultimately do not contribute to a framework for thinking about the likelihood of HGT events across a range of phylogenetic distances.

4) Transcription and or translation are possible in this genomic context.

Transcription and translation are not trivially regulated. It is probable that most foreign promoter regions will simply not be recognized in a new chromosome. A search for notable replacements of a particular promoter yielded 4 new promoter types out of 1000 novel DNA sequences tried (HORWITZ and LOEB 1986). These odds do not seem promising unless one considers the high standard (real antibiotic expression) called for in these experiments. It may only take a very low level of expression, conferred by most random sequences in order for newly acquired DNA to be exposed to selection.

In some ways our ability to sequence genomes is out-pacing our ability to understand the nuances that we now take for granted in our major model systems. Transcription and translation in a new genome are not impossible. It is possible that a gene inserts into a genome in a location that favors expression from some native promoter or that the gene carries a promoter-like sequence that is weakly recognized it its new cytoplasmic context. It has been suggested that operons might be formed in this way and that HGT may favor the formation of operons further by the subsequent deletion of unnecessary intermediate genes between those that function well in a new genomic context (LAWRENCE 1997; LAWRENCE 1999; LAWRENCE 2000; LAWRENCE and ROTH 1996; OMELCHENKO *et al.* 2003).

5) Positive selection or at least a lack of strong negative selection for new DNA.

Again, in cases where we have quantified the selection on newly acquired DNA in the ways that we can (namely experiments in laboratories where we may be poorly representing the conditions in which the genes were gained) we see at times, extremely small positive selection and sometimes none at all. The caveats to such experiments are clear. There is no telling what the environmental, population or historical circumstances of gene acquisition events were, let alone what transpired immediately after acquisition. These are necessarily going to continue to be ‘known unknowns’ for any given horizontally acquired DNA that we observe.

6) New DNA can spread in population.

In the event that genetic material has been 1) available, 2) incorporated stably into the genome, 3) has not interfered with replication or segregation, 4) is expressed and 5) has not had a negative effect on cell survival, there is the question of whether or not this becomes a general feature of the cell population. In the case of transformation, or plasmid acquisition in HGT, some plasmids will have the capability to conjugate and in this case can spread in a population quite easily.

In general, there are two primary ways that a newly acquired trait can spread in the population i) vertically and ii) horizontally. Vertical transmission will involve cell division and subsequent selection on this unique daughter cell population (those with the new DNA) amongst its parental type (those without the new DNA).

Vertical transmission of newly acquired DNA in a population can be thought of in terms of the new variant’s selection coefficient or the degree to which natural selection is acting to reduce the offspring of the new variant compared to the old variant. If the new DNA confers a fitness advantage, thereby increasing the number of offspring the unique population has relative to the parental type, then this new type will quickly take over the population. If however, the unique type has a selection coefficient of 0 compared to the parental type (if they are equally fit) then there is some probability that the new type may eventually still rise to fixation in the population but the dynamics of this process will function according to neutral drift dynamics in the population. In this case, the probability of fixation is equal to the initial frequency of the new type in the population. Therefore, if a population of 10^8 bacteria exists in some habitat and a single cell acquires some new DNA which confers no fitness advantage then the chances that this type will go to fixation in the population is roughly 1.0×10^{-8}.

Horizontal transmission in a population is the second way that newly acquired DNA can spread. Here, again we are dealing with homologous recombination among members of a population. Recombination between members of the same strain can spread an advantageous combination of alleles. The rate at which this sort of recombination takes place in nature has been analyzed in some groups and it has been observed that the rates at which populations of organisms recombine varies from species to species (FEIL *et al.* 2001).

Again, these constraints are not quantifiable or accessible to observation on the scale required to build notions or models of successful HGT in nature. Ideally, a set of constraints could be determined which would actually have predictive power to help us determine what sorts of HGT events might happen in the future. It is progress towards such a model that has been the goal of this dissertation and the constraint that is being elucidated has to do with the DNA molecule into which HGT is occurring. Chromosomes have a necessary structure.

DNA molecules carry information at a level above that of the genes. This information must be maintained in order to maintain the chromosome itself. The elucidation of this structure, as well as the constraint that it imposes on major chromosome rearrangements, including HGT, is the focus of this dissertation.

1.4 STRUCTURE OF THIS DISSERTATION

The work presented in this dissertation first examines the state of the field of HGT and the questions necessary to determine HGT's significance for bacterial evolution. The notion that HGT might be constrained by architectural sequences is couched along with other potential constraints (Chapter 2). Next, the structure of the chromosome is probed using bioinformatic tools to reveal that the major site of replication termination in a number of chromosomes is a particular site (the *dif* site), which is also used for chromosome concatemer resolution (Chapter 3). That work forms the groundwork for analyzing the sequence structure in the terminus region. A method to identify polarized architecture sequences is presented. Such sequences are found in most bacterial chromosomes examined and are demonstrated to be the result of selection and not simply stochastic mutational processes in those chromosomes (Chapter 4). In addition the impact of sequence architecture on the evolution of chromosomes is examined. This can be divided into evidence of architecture based constraints on inversions and HGT. Inversions destroy native chromosome architecture by reversing sequence polarity over large segments of the genome. Whereas HGT can both disrupt chromosome architecture by bringing in sequences in the incorrect orientation and can provide selectively beneficial genes that may out weigh the cost of this disruption (Chapter 5). A discussion of the work will be presented in the final chapter including implications for our notions of the tree of life and future work (Chapter 6).

2.0 LATERAL GENE TRANSFER: WHEN WILL ADOLESCENCE END?

"New gene pools are generated in every generation, and evolution takes place because the successful individuals produced by these gene pools give rise to the next generation."
~Ernst Mayr

2.1 SUMMARY

The scope and impact of horizontal gene transfer (HGT) in Bacteria and Archaea has grown from a topic largely ignored by the microbiological community to a hot-button issue gaining staunch supporters (on particular points of view) at a seemingly ever-increasing rate. Opinions range from HGT being a phenomenon with minor impact on overall microbial evolution and diversification, to HGT being so rampant as to obfuscate any opportunities for elucidating microbial evolution – especially organismal phylogeny – from sequence comparisons. This contentious issue has been fueled by the influx of complete genome sequences, which has allowed for a more detailed examination of this question than previously afforded. We propose that the lack of common ground upon which to formulate consensus viewpoints likely stems from the absence of answers to four critical questions. If addressed, they could clarify concepts, reject tenuous speculation and solidify a robust foundation for the integration of HGT into a framework for long-term microbial evolution, regardless of the intellectual camp in which you

reside. Herein we examine these issues, why their answers shape the outcome of this debate and the progress being made to address them.

2.2 COMING OF AGE

The first complete genome sequence of a free-living organism (*Haemophilus influenzae*) was released in 1995 and, as of this writing, more than 528 microbial genome sequences representing diverse lineages of Bacteria and Archaea have subsequently become available (FLEISCHMANN *et al.* 1995). The promise for new information held in complete genome sequences is vast and manifold, including (i) insight into previously unsuspected metabolic functions, (ii) elucidation of a microbe's underlying physiology even in the absence of a tractable genetic system or the ability to propagate the organism in pure culture, (iii) the identification of potential drug targets in pathogenic organisms, (iv) observations into the conservation of gene order, operon structure, variation in rates of evolution within and among genes, and so forth. The possibilities for mining novel answers to unasked questions also appear nearly endless, and the so-called "post-genomic era" has indeed brought about the publication of clever investigations that have called attention to hitherto unimagined aspects of microbiology. However, judging by surveys of the literature, it also seems that complete genome sequences have generated more debate, speculation, discussion and publication of works – both those of presenting objective analyses of new data and those proffering primarily interpretation and extrapolation of data according to one's point of view – regarding horizontal (lateral) gene transfer (HGT) than any other subject regarding the utilization of complete genome sequences. The availability of significant numbers of eukaryotic genome sequences has allowed the issue to be examined as a potent evolutionary force outside the prokaryotic domains.

This discussion of the scope and impact of HGT is not a young one, as the transmission of plasmid-borne antibiotic-resistance genes between organisms has been recognized for decades (DAVIES 1996). Yet at the time, this phenomenon was not thought to be widespread. Owing to the nature of bacterial reproduction, genes were viewed as being inherited primarily by vertical transfer, transmitted faithfully from mother cell to daughter cell during binary fission. HGT was an idea in its infancy – new and cute, but of no impact on the weightier matters of overall microbial evolution. More contemporary genome analyses often reach the same conclusions (SNEL *et al.* 1999) – that is, that vertical inheritance is the dominant mode of gene propagation – although the resolution becomes less staunch as more taxa are included for analysis and other evolutionary forces (gene loss and gene "genesis") are examined in more detail (SNEL *et al.* 2002). The trickle of DNA sequence data throughout the 1980's and early 1990's led to several compelling cases for HGT playing a role in the evolution of particular genes in some taxa [*e.g.*, the *gapA* gene in proteobacteria (DOOLITTLE *et al.* 1990)]. Even then there was no serious consideration of HGT as a major player in microbial evolution; vertical inheritance with periodic selection (LEVIN 1981) was still the dominant perspective of microbial evolution, even when DNA transfer between closely related strains of the same "species" was recognized (DYKHUIZEN and GREEN 1991).

As a conceptual brick on the edifice of biological thought, HGT made its mark via numerous analyses of complete genome sequences; two general approaches were employed. Phylogenetics could point out incongruent evolutionary histories of genes within the same genome (GOGARTEN 1995; GOGARTEN *et al.* 1992), while parametric analyses found genes displaying sequence patterns that could be interpreted as telltale signs of long-term evolution in another mutational (and therefore, genomic) context (MÉDIGUE *et al.* 1991). Yet despite the apparent surplus of data, HGT can still be considered to be an idea in its conceptual adolescence, so to speak. It has clearly shown promise in potentially changing directions of thought, allowing new insight into problems once thought tidily solved (GOGARTEN *et al.* 2002) and possibly offering new paradigms for interpreting microbial systematics, phylogeny and evolution. At its most dramatic interpretation by some readers, the apparently rampant and indiscriminate nature of HGT could dismantle the entire framework of bacterial phylogeny based on sequences of one or few genes (DOOLITTLE 1999); this would occur primarily because multiple phylogenies would better represent the mosaic nature of bacterial chromosomes (GOGARTEN *et al.* 2002).

For all its promise, HGT has not really established itself in any of these areas of opportunity. It has not reached scientific adulthood, where it would be accepted as a cornerstone of microbial evolution with well-defined roles, boundaries, causes and consequences (KURLAND 2000). The "genomic era" brought HGT to this point, and we propose four hurdles that must be passed for HGT to step out of the spotlight of debates between skeptics and champions – both often interpreting the same data from different viewpoints – and reach scientific maturity. We do not present here a comprehensive overview of the mechanism, elucidation, interpretation or impact of horizontal transfer [which have been reviewed extensively elsewhere, *e.g.* see (DOOLITTLE *et al.* 2003; GOGARTEN *et al.* 2002; KOONIN *et al.* 2001; OCHMAN *et al.* 2000), or provide an overarching framework for its role in microbial genome evolution. Rather we discuss these four questions, and the progress being made towards answering them. With these data in hand, perhaps microbiologists could proceed to outline with rigor and confidence the roles of gene transfer in microbial evolution.

2.3 HOW DOES HGT IMPACT THE EVOLUTIONARY HISTORY OF DIFFERENT GENES?

Perhaps nowhere has the HGT debate been more focused than on its relative influence on the evolutionary histories of different genes. No one denies that certain classes of genes (*e.g.,* those encoding antibiotic-resistance) are associated with mobile genetic elements and can experience high rates of transfer (HALL 1997). In contrast, the rRNA genes have long been considered relatively recalcitrant to transfer, allowing the foundations of bacterial phylogenetics (WOESE 1987). The phylogenies of other highly conserved genes, like tRNA synthetases (WOESE *et al.* 2000), primarily reflect that inferred from rRNA genes (LUDWIG *et al.* 1998), although some notable transfers are evident among these phylogenies (WOESE 2000). These data support the view that a "core" set of genes has been inherited by vertical descent and represent the "true" phylogeny of the bacteria that harbor them. Along these lines, it has been proposed that genes whose products interact with a large number of other proteins and RNAs would be those least likely to be transferred (JAIN *et al.* 1999). Newly-introduced orthologs would be unlikely to express a product that could out-perform one that had experienced long-term coevolution with its cognate partners.

Implicit (but unstated) in the idea that highly conserved genes would be subject to less transfer is the verity that there would be a smaller subset of strains that could benefit from receiving the new genes. Clearly most genomes would already contain a homologue of the transferred gene and an orthologous replacement would have to occur. Among less highly conserved genes, many lineages may be naïve to the gene's product, and a selective advantage could arise by the newly acquired gene(s) providing a novel function (LAWRENCE 1997; LAWRENCE and ROTH 1996; LAWRENCE and ROTH 1999). Among more highly conserved genes, orthologous replacement would occur at a rate of 50% at best, ignoring any detriments inherent to the retention of introgressed genes (which are discussed below) beyond their lack of coevolution with potentially interacting partners.

Yet there have been cases where genes involved in information transfer (replication, transcription, translation) have been subject to HGT (for some examples, see Table 1). Indeed, even rRNA genes have been shown to experience HGT (MYLVAGANAM and DENNIS 1992; YAP *et al.* 1999); their ability to be transferred lies in many of the same features originally cited as reasons that they would likely not be: they are ubiquitous in distribution, are highly conserved and perform the identical function in all cells. Yet these properties actually promote exchange of all or parts of the rRNA molecule, fueled by long regions of nucleotide identity (not encoding a protein, this gene lacks the variant bases that arise due to the degeneracy of the genetic code) and high degree of conservation of function (Fig. 7). Moreover, surveys of genomes for atypical genes show that many other genes have been acquired recently, up to 25% of the genome (GARCIA-VALLVE *et al.* 2000; HAYES and BORODOVSKY 1998; KARLIN *et al.* 1998; LAWRENCE and OCHMAN 1998; LAWRENCE and OCHMAN 2002; NELSON *et al.* 1999; OCHMAN and JONES 2000; OCHMAN *et al.* 2000; RAGAN 2001).

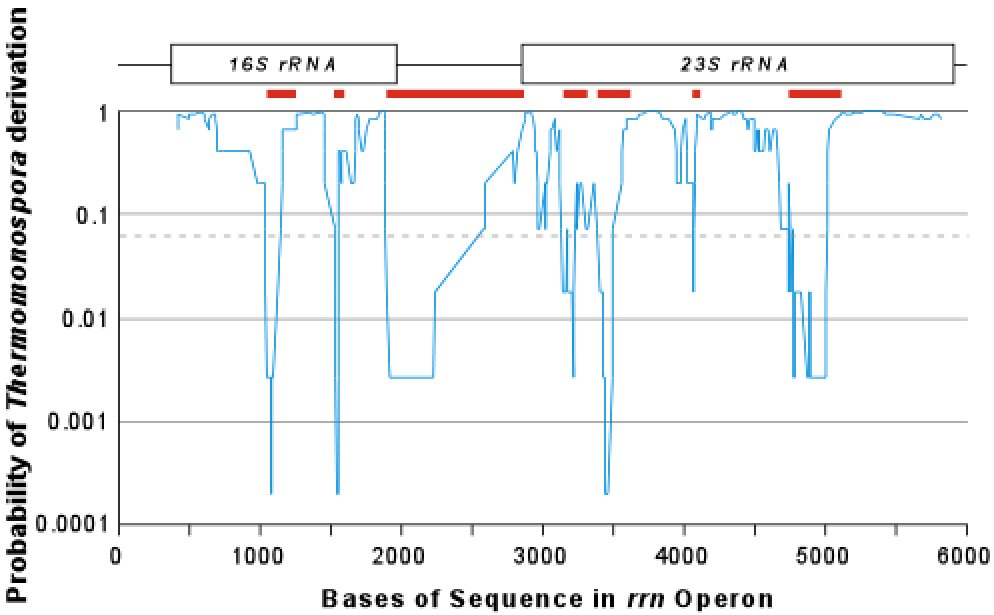

Figure 7 Evidence of HGT of portions of the 16s rRNA sequence.

2.3.1 Figure 7 legend

Mosaicism within the *Thermomonospora chromogena rrnB* operon, which bears regions of identity to the *rrn* operons of *Thermobispora bispora*. Informative sites were identified as positions where (a) three full-length *T. chromogena rrn* operons were identical, (b) two full-length *T. bispora rrn* operons were identical to each other, but differed from the *T. chromogena* sequences, and (c) the *T. chromogena rrnB* base matched one of the two. Of the 478 informative sites, 202 sites (42%) paired the *T. chromogena rrnB* operon with *T. bispora rrn* operons, while 276 showed identity across all four *T. chromogena rrn* loci examined. A window of 10 informative sites was used to calculate the probability of the *T. chromogena rrnB* operon matching the other 3 *T. chromogena rrn* loci (blue line); P=0.0002 indicates that all 10 sites within the window matched the *T. bispora rrn* loci. The red bars denote regions likely to be of *T. bispora* origin (P < 0.05). Figure adapted from (Gogarten *et al.*, 2002).

Here is where one can interpret data in different ways. One position may be "Look, no gene is immune to transfer, even if it is involved in a complex molecular machine with coevolving parts. Therefore, no consortium of coevolving genes defines the essence of a bacterial cell. As a result, one cannot simply deduce microbial evolution from molecular phylogenies as represented by a single, bifurcating tree; rather, this mosaicism is best represented by reticulation, where genomes contain genes with differing histories." Such an argument has been made convincingly for bacteriophage genome evolution (LAWRENCE *et al.* 2002); but in this case, the transferred fragments represent much larger fragments of the genome (up to 50%), and it is impossible to identify a common "core" of genes shared among all bacteriophage lineages. By analogy, is it valid to extend this argument to bacterial genomes as well?

Table 1 Notable recorded incidences of HGT.

Protein	Phylogenetic incongruities	Reference
Ribosomal RNA (*rrn)*	(i) *Thermomonospora contains rrn* operon donated from *Thermobispora* *(ii) Haloacula* contains *rrn* operon from a probable Halobacteial donor	(MYLVAGANAM and DENNIS 1992; YAP *et al.* 1999)
RNA polymerase	*Mycoplasma* branches at the bottom of the Bacterial domain	(KLENK *et al.* 1999)
Ribosomal protein L32 (RpmF)	*Lactococcus lactis* groups with the Proteobacteria	(MAKAROVA *et al.* 2001)
Ribosomal protein L33 (RpmG)	(i) *Deinococcus* groups with Aquifex instead of *Thermus* (ii) *Mycobacterium leprae* groups separately from *M. tuberculosis*	(MAKAROVA *et al.* 2001)
Ribosomal protein S14 (RpsN)	*(i) Mycoplasmas* are separate from other low-GC Gram-positive Bacteria (ii) *Deinococcus* is separated from *Thermus* and groups with some low-GC Gram-Positive Bacteria	(BROCHIER *et al.* 2000)
Ribosomal protein s18 (RpsR)	Three Mycoplasmatales species group with ε-protebacteria	(MAKAROVA *et al.* 2001)
Lysyl-tRNA synthase	*Borrelia* groups with Archaea	(IBBA *et al.* 1997)
Penylalanyl-tRNA synthase	Spirochaetes group with Archaea	(WOESE *et al.* 2000)
Prolyl-tRNA synthase	*(i) Deinococcus, Mycoplasma* and *Borrelia* groups with Archaea *(ii) Borrelia* does not group with the spirochaete *Treponema,* which remains within the Bacterial clade	(GOGARTEN and OLENDZENSKI 1999) (WOESE *et al.* 2000)
Seryl-tRNA synthase	The Acrhaeon *Haloarcula* groups with Bacteria	(DOOLITTLE and HANDY 1998) (WOESE *et al.* 2000)

An alternative and equally valid viewpoint is that the transfer of highly conserved genes (Table 1, Fig. 7) is relatively rare, and therefore does not affect the robustness of the underlying organismal phylogeny in an analogous fashion. Instead, much of HGT would be limited to genes that affect bacterial lifestyle, but do not have a large impact on the "core" set of genes involved in information transfer or central metabolism. Certainly no gene is immune to HGT, and one can always identify the occasional transfer event among any set of genes. Yet on the whole, these are exceptions to the rule of vertical inheritance of the as-yet-undesignated "core" set of genes that encode the consortium of essential gene products enabling cellular life. The impact of HGT on these genes is constrained by its rarity in this arena, thereby leaving organismal phylogeny – and all the biological inferences made from it – intact.

Recent analyses of orthologous sequences among diverse genomes supports their general congruence with the rRNA phylogeny, at least among the relatively closely-related genomes of some clades (DAUBIN *et al.* 2003; MAKAROVA *et al.* 1999; NESBO *et al.* 2001). These data support the idea that there may be core sets of genes recalcitrant to frequent HGT, although their numbers may be small, and the composition of these sets may vary among bacterial lineages. Indeed, the same data used to infer high rates of gene transfer among genomes (Ochman *et al.*, 2000) have been reanalyzed to infer that not all classes of genes – here, using a functional classification scheme (SERRES and RILEY 2000) – are found in proportional abundance among the newly acquired genes identified in numerous bacterial genomes (Lawrence, unpublished data). On the contrary, genes involved in "information transfer" are rarely, if ever, identified as "recently acquired," whereas genes encoding transporters and other more peripheral metabolic functions are highly represented in this group. These analyses support the idea that not all genes are transferred with equal likelihood among all lineages; this conclusion affects large-scale genomic analyses, like supergene trees (BROWN *et al.* 2001). In this example, the data set was reduced to 14 genes to generate a tree topology congruent with the rRNA phylogeny (BROWN *et al.* 2001).

Yet the questions as to how many genes remain primarily recalcitrant to transfer, how many experience frequent HGT to escape loss, and the nature of the continuum between these two extremes, remain unanswered in any quantitative fashion. More importantly, the issue as to how population structure and subdivision affect the likelihood of successful lateral transfer have only begun to be explored; a recent model shows that genes with low selective value are likely to be lost unless transferred into "patchy" populations allowing local fixation (BERG and KURLAND 2002), consistent with the predictions of the Neutral Model of molecular evolution (KIMURA 1983). Since a gene's selection coefficient is a function both of its identity and its genomic (and, hence, ecological) context, assigning genes along a spectrum of "readily transferred" to "rarely transferred" becomes even more difficult, as is discussed below.

Lastly, one can add another layer of complexity by asking the transferred genes to provide a selective value during transit. While this is not necessary if genes are introduced by transformation, many genes are introduced by transduction; indeed, the original conception of horizontally acquired "pathogenicity islands" was intimately associated with bacteriophages (BARINAGA 1996). Bacteriophages are highly mosaic and many contain genes typically thought to be "bacterial" in origin (PEDULLA *et al.* 2003). If a gene is in transit between bacterial genomes via a bacteriophage intermediate, it has a higher likelihood of completing the voyage successfully if it provides a useful function to the phage or to the prophage. Not all genes would satisfy this criterion. As a result, placing genes on an overall scale from "nearly immobile" to "highly transmissible" is a formidable task with a great number of variables to consider.

2.4 HOW DOES THE ROLE OF HGT DIFFER AMONG DIFFERENT LINEAGES?

As alluded to above, the rate of HGT of individual genes also must vary among bacterial lineages, owing to the different selective values they would impart in different genomic contexts. This constraint is obvious when examining the genomes of intracellular parasites and obligate pathogens, both of which are experiencing genome reduction (ANDERSSON and ANDERSSON 1999; ANDERSSON and ANDERSSON 1999; MORAN and WERNEGREEN 2000). Here, organisms that experience strong declines in effective population size and/or rate of gene exchange by homologous recombination cannot retain the genes they currently possess, since their thresholds for effectively neutral mutations have increased (LAWRENCE 2001). As a result, many of their ancestral genes cannot be retained as their benefits are insufficient to prevent their loss by mutation and genetic drift. In addition, their sheltered lifestyles limit access to the agents of HGT (bacteriophages, other bacteria with conjugative plasmids, *etc.*), also lowering the likelihood of gene acquisition. A comparison of insect endosymbionts shows remarkable genome stasis over 50 Myr (TAMAS *et al.* 2002), including the lack of genes acquired by HGT.

Genome reduction can also play the opposite, more counterintuitive, role in affecting a lineage's propensity for participating in HGT. While many of the lineages undergoing genome reduction will likely not give rise to descendents that undergo genome expansion, some will. For example, the *Mycoplasma pneumoniae* genome has significantly more DNA than its congener *M. genitalium*. Much of the "additional" DNA found in *M. pneumoniae* is atypical (OCHMAN *et al.* 2000), suggesting that a small genome has acquired new functions by HGT, and thus is experiencing genome expansion. Here, it is likely that the population size or recombination rate have increased so that the likelihood of retaining newly-introduced genes has increased. More importantly, the organism must be shifting into an ecological niche wherein the newly-acquired genes serve useful purposes. Therefore, genome reduction should not be viewed only as an indicator that a lineage will likely see reduced rates of HGT; some lineages may see enormous increases in their rates of HGT as they regain genes previously lost.

There are certainly other biological limitations to the free exchange of DNA between all taxa. First, transmissible agents have restricted ranges; *e.g.,* bacteriophages have limited host ranges, as do many conjugative plasmids. Second, the apparati of transcription and translation become increasingly different with phylogenetic distance, imposing a barrier to facile gene exchange across large genetic distances. Only genes that provide large selective benefits would be retained following "long-distance" transfer since their initial expression levels would be poor. This is perhaps most dramatically illustrated by the difficulty in transfer of bacterial genes into eukaryotes, where operons cannot be expressed by a native promoter at the site of insertion. Rather, the eukaryotic transcription and translation machineries require independent expression of each gene, thereby imposing a barrier to gene transfer beyond the necessity for transit of the DNA to the nucleus of a germ-line cell, and its provision of a selectable function.

Data examining the effect of ecological niche on the propensity of gene exchange among cohabitants has had tantalizing beginnings [*e.g.,* among thermophilic Bacteria and Archaea (NELSON *et al.* 1999; WORNING *et al.* 2000), or between Bacteria and Fungi dwelling in the rumen (GARCIA-VALLVE *et al.* 2000)], but remain largely unexplored; in addition, caveats can always be raised in regards to methods employed to make these inferences (LOGSDON and FUGUY 1999). Since gene exchange – by either transformation or by transduction – does not require donors and recipients to cohabitate, it is not clear how dwelling in the same physical environment increases gene flow by HGT. Moreover, the breadth of ecologies explored by individual "species" is also a field of great interest but little data; preliminary work suggests that it may differ greatly among lineages (GORDON 2001; GORDON *et al.* 2002; OKADA and GORDON 2001; VOGEL *et al.* 2003), which is perhaps not unexpected. This variability makes taxon-to-ecology assignment difficult, if not infeasible, impractical and potentially misleading; *e.g.*, witness the strong and well-documented ecological differences among strains of *Salmonella enterica* as a pathogen (BAUMLER *et al.* 2000; RABSCH *et al.* 2002).

Lastly, bacterial chromosomes themselves may have higher-ordered structures that allow for proper replication termination and chromosome segregation. Such structures may be imparted by the asymmetric distribution of sequences arising naturally by strand-specific mutational biases (CAPIAUX *et al.* 2001; LOBRY 1996; LOBRY and LOUARN 2003; LOBRY and SUEOKA 2002). Unlike the factors discussed above, these sequence features can be examined quantitatively to test hypotheses in a rigorous fashion. We have examined such sequences in numerous taxa and have found that they are conserved only among phylogenetically related taxa (Hendrickson and Lawrence, unpublished results). The octomeric sequence shown in Fig. 8 displays a distribution indicative of participation in proper chromosome termination and segregation in its host *Mesorhizobium loti*, where it is counter-selected from appearing on the "improper" strand. This sequence, GGGCAGGG, has a similar distribution among closely-related α-proteobacteria, but is found in high abundance on both strands in distantly-related taxa, such as *Streptomyces coelicolor* (Hendrickson and Lawrence, unpublished results). If a DNA fragment were to be introduced into *Mesorhizobium* (or one of its relatives) from a donor taxon where this sequence was abundant on both strands, the presence of the DNA would incur a selective detriment that could potentially offset any benefits provided by the newly-acquired gene products. This barrier to gene exchange, unlike those discussed above, has only come to light with recent genomic analyses, which have both furthered our understanding of bacterial genome structure and shown us the depth of our ignorance.

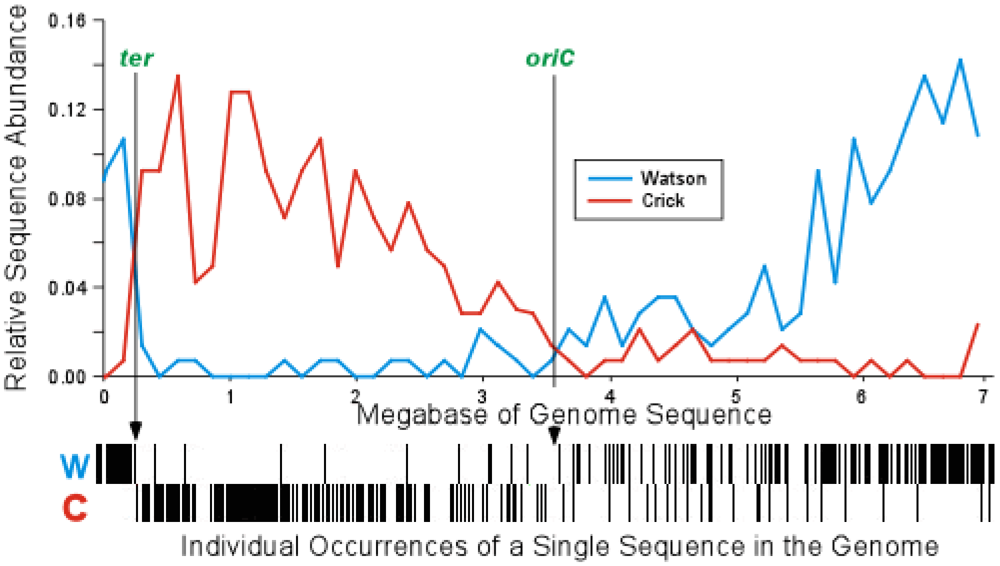

Figure 8 Occurences of a skewed, asymmetric sequence in *Mesorhizobium loti*.

2.4.1 Figure 8 legend

The lower panel depicts each sequence on either the Watson (top) or Crick (bottom) strand as a hash mark. The abundance of this sequence on each strand is tabulated in the graph above; the origin and terminus of replication can be inferred from analyses of GC skew as well as the distribution of the octomer depicted here (Hendrickson and Lawrence, unpublished results). The origin was distinguished from the terminus both by the position of the *dnaA* gene (which is typically origin-proximal) and by the orientation of *rrn* operons (typically transcribed away from the origin of replication). Comparable accumulation of octomeric sequences do not occur elsewhere in the genome, implying that this is not the result of chance. More importantly, analyses of di- and tri-nucleotide frequencies show they are uniform across the genome, rejecting the hypothesis that the accumulation of these sequences are the result of significant alteration in mutation bias during replication.

The distribution of such sequences, as well as the other factors detailed above (Fig.9A), would limit HGT frequency in a clade-specific fashion. That is, rates of HGT would be relatively high among closely-related taxa, but would decrease in efficiency with phylogenetic distance by the accumulation of these numerous problematic differences (*e.g.,* lack of proper ribosome binding sites, lack of proper promoter sequences, an excess of functionally-biased sequences on the "improper" DNA strand, *etc*). If this is true, then HGT would have a profoundly different impact on phylogenetic reconstruction than the genetic panmixia than had been previously envisioned by many. Here, bacterial clades would be self-reinforcing, since most of the HGT would be occurring among more closely-related taxa (Fig.9B). As a result, one would detect fewer long-range transfers of highly-conserved genes (Table 1), and many gene phylogenies would be congruent with that inferred from the rRNA sequences when examined at large phylogenetic scales. Hence, phylogenies based on gene content (FITZ-GIBBON and HOUSE 1999; SNEL *et al.* 1999; TEKAIA *et al.* 1999) may reflect the propensity for HGT among more closely related lineages as much as the retention of their ancestral genes (GOGARTEN *et al.* 2002).

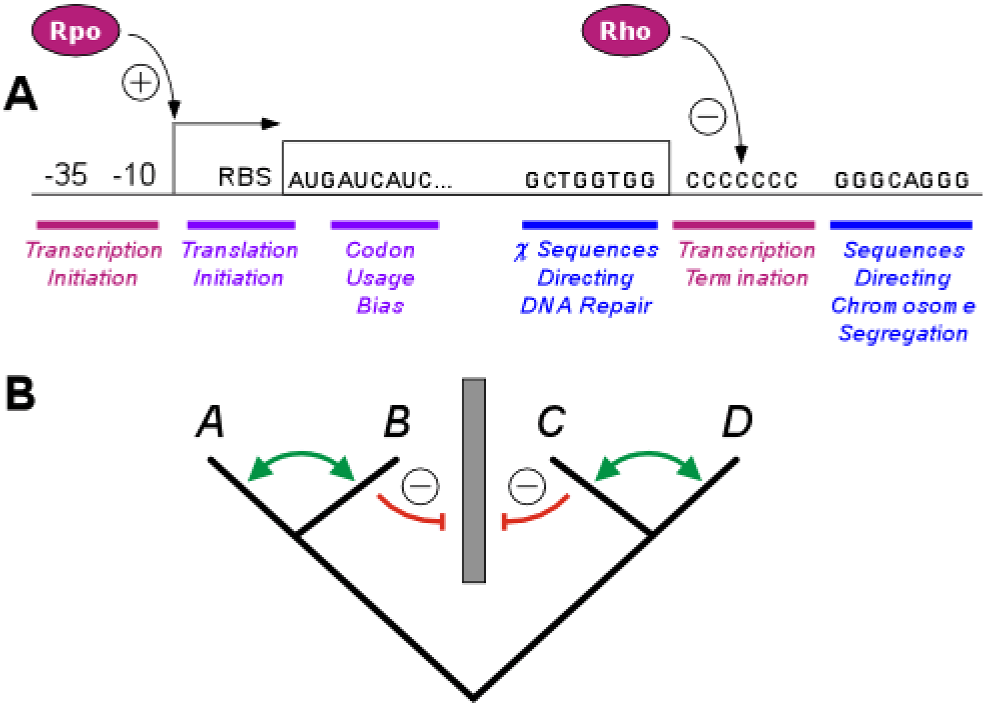

Figure 9 Limitations on HGT among taxa.

2.4.2 Figure 9 legend

A. Variable sequence features that can differ between taxa and decrease the likelihood of successful gene transfer, including those involved in transcription (magenta), translation (purple) and replication (blue). **B**. A model whereby the sequence features noted in panel A allow for more frequent transfer (green arrows) among more closely related organisms, but act as a barrier (albeit not an impervious one) to transfer between distantly related taxa (aborted red arrows).

2.5 HOW DOES ONE REACH ROBUST CONCLUSIONS ON THE PRESENCE OR ABSENCE OF HGT?

The availability of multiple complete genome sequences has created the opportunity for unprecedented sophistication in phylogenetic analyses, wherein dendrograms are no longer derived from selected, and fortuitously available, DNA sequences. Rather, the entire body of information contained in the genomes can be brought to bear. While this has solved some problems (like poor taxon sampling, or the necessity of employing single gene sequences), it has created problems of its own as new methodologies have been developed to analyzed genome sequences *en masse*. For example, does the creation of "supergene" trees (BROWN *et al.* 2001) amplify weak phylogenetic signals at the expense of masking the signals of gene transfer? Moreover, the dynamics of gene loss and growth of paralogous gene families can obfuscate the identification of horizontally acquired genes and the inference of genome evolution (JORDAN *et al.* 2001; KUNIN and OUZOUNIS 2003; MIRKIN *et al.* 2003; SNEL *et al.* 2002), and some inferences are open to misinterpretation regarding the role of HGT (see below). Yet these works clearly show that the balance between gene loss and gene acquisition – both by lateral gene transfer and by the expansion of preexisting gene families – will also vary among lineages, making an overall assessment of the impact of gene transfer alone in genome evolution only one part of a complex process we are only beginning to understand.

There lies an even more pressing issue beneath the questions regarding the impact of gene-specific, or taxon-specific variation of HGT on bacterial evolution: in many cases it is difficult to ascertain with any degree of certainty if HGT has or has not played a role in the evolutionary history of a gene. This lack of confidence stems from many sources, both those trivial to explain or to correct, and those that are more profoundly difficult to address [*e.g.,* see (KOONIN *et al.* 2001)]. For example, genes likely affected by HGT have been identified by numerous methods in bacterial genomes, but these lists of "alien" genes do not agree with each other (RAGAN 2001). In this case, many of the discrepancies can be attributed either to statistical artefacts in the methods employed, or to the different classes of genes that each method was designed to detect (LAWRENCE and OCHMAN 2002). In addition, parametric methods detecting atypical genes (presumably having evolved in a genome with different mutational biases) can lead to incorrect assignment of short ORFs as being atypical (and potentially newly acquired) due to lack of data, and may be unable to identify genes recently transferred from taxa with similar mutational biases. These methods will ultimately fail to detect genes that were introduced long ago, since the mutational proclivities of their current host will ameliorate any atypical sequence features over time (LAWRENCE and OCHMAN 1997; LAWRENCE and OCHMAN 1998).

Similarly, phylogenetic methods can be confounded by (i) the amplification of gene families in certain genomes, which interferes with the proper identification of orthologous genes, (ii) convergent evolution due to parallel phenotypic shifts (for example in the %GC content of the genome, or in thermal growth regime leading to predictable protein modifications), or (iii) phylogenetic artefacts such as variation in the rates of evolution between lineages or long branch-length attraction (RAGAN 2001; RAGAN 2001; SIMMONS *et al.* 2002; STILLER and HALL 1999). Ultimately, phylogenetic methods will also fail, in this case when evolutionary changes have become so numerous as to overwhelm a useful phylogenetic signal, making inferences regarding HGT a challenge in navigating the vagaries of phylogenic reconstruction methodology, which can always be called into question.

What we find even more disturbing is the failure of most investigators examining HGT to reach a consensus as to what null hypothesis should be tested. That is, regardless of approach, how one phrases a scientific question can bias the conclusions. In the “pre-genomic era”, it was assumed that genes were inherited vertically during cell division. Naturally, one tested the idea that a gene had been subject to HGT by stating vertical inheritance as the null hypothesis to be disproven by the weight of the data. If one could not disprove the null hypothesis, one then concluded that the gene was not subject to HGT. Yet one could just as easily begin with a null hypothesis whereby the genes being analyzed had been subject to HGT, and collect data to refute this hypothesis. Here one would conclude that the gene was not subject to HGT only if one refuted the null hypothesis, rather than having this conclusion be the default condition upon failure to disprove an alternate null hypothesis. In most phylogenetic analyses, the first scenario is the *de facto* approach; yet in many cases the data are of insufficient quality – for the reasons outlined above – to make robust conclusions regardless of which null hypothesis is taken. That is, if neither null hypothesis can be rejected, robust conclusions can not be made, and uncertainty must remain. This caveat is also applicable to the identification of putatively transferred genes by parametric approaches: the failure to identify a gene as atypical does not rule out the possibility that HGT has played a role in its evolution in this taxon. Is it fair to assume that genes have been inherited vertically and require evidence that HGT has played a role, rather than the converse?

2.6 HOW DOES ONE INTEGRATE HGT INTO THE CONTINUUM OF GENETIC EXCHANGE TO ARRIVE AT MEANINGFUL MICROBIOLOGICAL CONCEPTS?

Exchange of DNA among bacterial taxa can occur between very closely related strains, where it is often termed "recombination," and integration of DNA is mediated by homologous recombination (FEIL *et al.* 2001; GUTTMAN 1997). Therefore, transfer of DNA between closely-related taxa will be unlikely to result in a recombinant bearing two alleles of the same locus; rather, an orthologous replacement would occur. As sequence divergence increases, homologous recombination is precluded by the mismatch correction system (MAJEWSKI and COHAN 1999; ZAWADZKI *et al.* 1995), and only an illegitimate or site-specific recombination event can introduce the DNA into the genome. If the sequences are closely related, one copy will be retained and the other lost by deletion, as the genes would likely not encode proteins that conferred sufficiently distinct functions to allow selection for retention of both copies. The probability of gene retention likely increases as sequence divergence between donor and recipient lineages increases, since more time would elapse for functional differences to arise (Fig. 10). However, more distantly related taxa would experience the barriers to HGT discussed above (see also Fig. 9A), thereby reducing the probability of successful transfer. As a result, one can consider a "zone of paralogy" where it is most likely that sequences introduced by HGT could be retained. This "zone of paralogy" would also act to reinforce by HGT clade identities initially established by common ancestry.

The "zone of paralogy" also offers a cogent mechanism for the growth of gene families observed in many taxa (JORDAN *et al.* 2001, Snel, 2002 #3444; SNEL *et al.* 2002). The expansion of gene families by duplication and divergence of single genes within a single genome is an old idea, yet fraught with difficulty. Foremost among the difficulties is the problem of maintaining selection on both copies, thereby preventing loss of the duplicated gene, until each gene develops functionally distinct roles. While clever schemes have been devised to circumvent these problems [*e.g.,* see (LYNCH *et al.* 2001; STOLTZFUS 1999)], differential function may arise while genes reside in different cytoplasms and experience different selective constraints. HGT would then reunite previous orthologs in the same genome, where they would appear as paralogs; this process alleviates the need for a period of coexistence of multiple copies of the same gene without selection for differential function (GOGARTEN *et al.* 2002; LAWRENCE 2001). Therefore, one must consider carefully the mechanisms by which "gene genesis" (SNEL *et al.* 2002) occurs. Is HGT also playing a role here? Moreover, different rates of evolution among genes changes the taxonomic scope of organisms available for gene exchange by homologous recombination, and makes the "zone of paralogy" vary in a gene-specific manner.

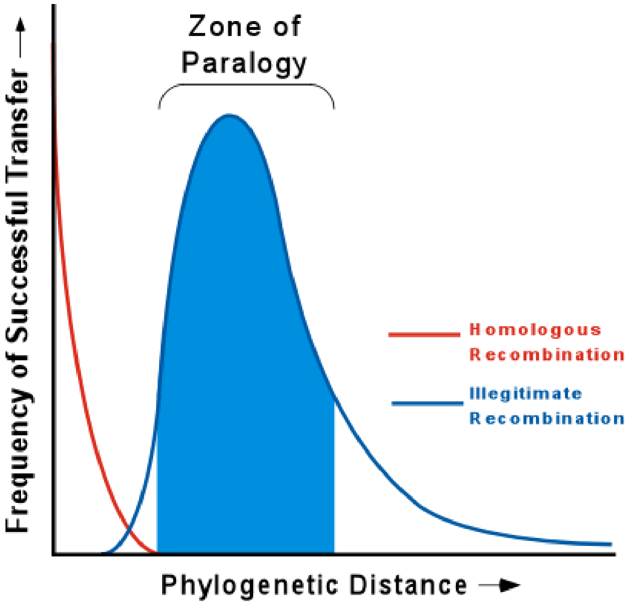

Figure 10 The interplay between HGT mediated by homologous & illegitimate recombination.

2.6.1 Figure 10 legend

The interplay between HGT mediated by homologous (red line) and illegitimate (blue line) recombination. Among closely related taxa, incoming DNA is likely integrated by homologous recombination, resulting in allelic replacement. More divergent sequences cannot recombine by this route, resulting in genomes with homologous genes; however, more distantly-related homologues are more likely to be retained as paralogues, since they are more likely to confer separate functions (axes are depicted using arbitrary units). However, the factors shown in Fig. 3 decrease the overall frequency of HGT as taxa become more distantly related. The interplay between these effects results in a "zone of paralogy," depicted in cyan, whereby sequences are most likely to be retained.

The rates of DNA exchange by homologous and illegitimate recombination are also intimately associated via the manner by which novel alleles are distributed in a population. If recombination among strains in a population is rare, then novel alleles arising by HGT are more likely to be lost by genetic drift than those able to be transmitted by homologous recombination. Therefore, increases in the rate of homologous recombination within populations serve not only to decrease the threshold of an effectively neutral mutation [increasing the likelihood of HGT (LAWRENCE 2001)] but also to disseminate newly-acquired genes and prevent their stochastic loss (BERG and KURLAND 2002). Yet the introduction of novel alleles by HGT will also allow for niche-specific adaptation, which will eventually lead to bacterial "speciation" (COHAN 2001; LAWRENCE 2002). Certain recombination events – those which disrupt such niche-specific loci – will produce less-fit offspring, leading to reproductive isolation at chromosomal loci surrounding genes introduced by HGT (LAWRENCE 2002).

One can view the interplay of gene exchange by these mechanisms as effectively blurring the lines between microbial taxa, making it difficult to delineate microbial "species" or groupings at higher taxonomic levels. It is difficult to apply the Biological Species Concept, as have Dykhuizen and Green (DYKHUIZEN and GREEN 1991), to groups of strains that are reproductively isolated at some loci and not others. Similarly, the variable domains of exchange among taxa at different levels of inclusiveness, as well as the variable rates of exchange among different genes, makes higher-ordered taxonomic classification difficult to quantify as well. As discussed previously (GOGARTEN *et al.* 2002), if higher-ordered taxonomy is dictated both by the presence of ancestral genes (as is the case in eukaryotes) as well as biased HGT within taxonomic groups, then bacterial taxonomy reflects both history (the patterns of speciation events) as well as ongoing processes (HGT). Hence, the conclusions of Zuckerkandl and Pauling (ZUCKERKANDL and PAULING 1965), that genes are documents of evolutionary history, becomes far more complex as we integrate patterns of gene exchange – and lineage specific gene loss – with histories of vertical inheritance.

2.7 CONCLUSIONS

Woese (WOESE *et al.* 2000) postulated that HGT was rampant early in microbial evolution, but plays a smaller role now, after passage through the "Darwinian Threshold". While the arguments that the role played by HGT differs now from the roles played in ancient lineages are compelling, it is still clear that HGT can be a potent process in microbial diversification. The questions remain as to how its impact can be quantified in lineages and genes of interest, and how these data can be integrated into a holistic understanding of how gene exchange mediates evolutionary change.

Answers are likely to come from multiple sources, including the accumulation of additional data that will allow for more conclusive identification of orthologs among distantly-related taxa, the development of more robust methods for phylogenetic inference that can be used on large data sets, integration of methods used to detect atypical genes and methods used to detect genes with aberrant phylogenetic histories, and the continued integration of the numerous evolutionary forces acting on genome evolution. More importantly, these advances must be accompanied by a holistic change in mindset among microbiologists. Critical, thoughtful evaluation and interpretation of all available data can assist in making inferences and conclusions that help clarify, rather than confound, these complex biological issues. Only in this way can horizontal gene transfer be discussed as a topic with a firm foundation in fact, rather than as a collection of anecdotes and seemingly arcane analyses.

3.0 MUTATIONAL BIAS SUGGESTS THAT REPLICATION TERMINATION OCCURS AT THE *DIF* SITE, NOT AT THE TER.

“The theory of evolution by cumulative natural selection is the only theory we know of that is in principle capable of explaining the existence of organized complexity.” ~Richard Dawkins

3.1 ABSTRACT

In bacteria, Ter sites bound to Tus/Rpt proteins halt replication forks moving only in one direction, providing a convenient mechanism to terminate them once the chromosome had been replicated. Considering the importance of replication termination and its position as a checkpoint in cell division, the accumulated knowledge on these systems has not dispelled fundamental questions regarding its role in cell biology: why are there so many copies of Ter, why are they distributed over such a large portion of the chromosome, why is the *tus* gene not conserved among bacteria, and why do *tus* mutants lack measurable phenotypes? Here we examine bacterial genomes using bioinformatics techniques to identify the region(s) where DNA polymerase III-mediated replication has historically been terminated. We find that in both *Escherichia coli* and *Bacillus subtilis*, changes in mutational bias patterns indicate that replication termination most likely occurs at or near the *dif* site. More importantly, there is no evidence from mutational bias signatures that replication forks originating at *oriC* have

terminated at Ter sites. We propose that Ter sites participate in halting replication forks originating from DNA repair events, and not those originating at the chromosomal origin of replication.

3.2 INTRODUCTION

The replication of chromosomal DNA is arguably the most important job a cell can perform. All other functions – including transcription, translation, protein targeting, energy generation, biosynthesis and metabolite transport – merely support the ultimate effort to reproduce the immense, information-bearing polymer that has been transmitted cell-to-cell for more than 3000 million years. Among bacteria, this has conservatively amounted to more than 1 million million million million million million million rounds of replication. Not surprisingly, bacteria have a single, well-regulated replication origin (*oriC*) that coordinates the synthesis of new DNA in an orderly fashion (KAGUNI 2006; KATO 2005; LEONARD and GRIMWADE 2005). Replication forks proceed bidirectionally from this position and, in circular chromosomes, terminate at some point ~180° away. When replication forks meet, the tremendous accumulation of positive supercoils in front of the colliding forks must be deftly dissipated to avoid rending the duplex DNA, chromosome dimers and catemers must be resolved, and the DNA must be apportioned faithfully to two daughter cells as the division septum creates them. One could consider replication termination and subsequent cell division to be the culmination of all metabolic efforts that took place in the previous cell cycle.

Yet given the importance of replication termination, its coordinated role in chromosome segregation and cell division (BARTOSIK and JAGURA-BURDZY 2005; HAYES and BARILLA 2006;

HAYES and BARILLA 2006; SHERRATT 2003; SHERRATT *et al.* 2001; SHERRATT *et al.* 2004; THANBICHLER *et al.* 2005), and the biophysical challenge of allowing replication forks to collide gracefully, it is somewhat surprising that the location of any replication terminus is ill-defined at best. A terminus zone was first described in the model organism *Escherichia coli* (BIRD *et al.* 1972; MASTERS and BRODA 1971), where replication forks appeared to terminate in a region corresponding to ~15 % of the chromosome (LOUARN *et al.* 1977; LOUARN *et al.* 1979), later refined to ~5 % (DE MASSY *et al.* 1987), located opposite of the replication origin. This activity was evident even if ectopic, terminus-proximal replication origins were fired (LOUARN *et al.* 1977), suggesting that termination had a molecular basis and was not merely the coincidental arrival of two replication forks traveling at similar rates. Investigation of this phenomenon led to the identification of Ter sites (HILL *et al.* 1987; HILL *et al.* 1988; PELLETIER *et al.* 1988), nonpalindromic sequences that arrest replication forks when DNA polymerase approaches them in the non-permissive orientation (Fig. 11A). Ter sites are located throughout the terminus-half of the *E. coli* chromosome (MULCAIR *et al.* 2006; NEYLON *et al.* 2005) and stall replication forks only when the Tus protein is bound there (HIDAKA *et al.* 1989; KOBAYASHI *et al.* 1989), where it acts as an antihelicase (HIDAKA *et al.* 1992; MULCAIR *et al.* 2006; MULUGU *et al.* 2001). A model (HILL 1992) was then proposed whereby the “inner-most” Ter sites act as a replication fork trap, wherein forks could enter but not leave (see first schema in Fig. 11B). This model was attractive in its elegance; termination would be, in essence, a passive process where forks were allowed to collide in a confined region of the chromosome, or at a Ter site in the non-permissive orientation if it were encountered first. Additional Ter sites were proposed to provide “back-ups” should a Ter-stalled fork regain processivity and bypass its initially-encountered Ter site (HILL 1992), and termination would not consistently occur at any other specific location.

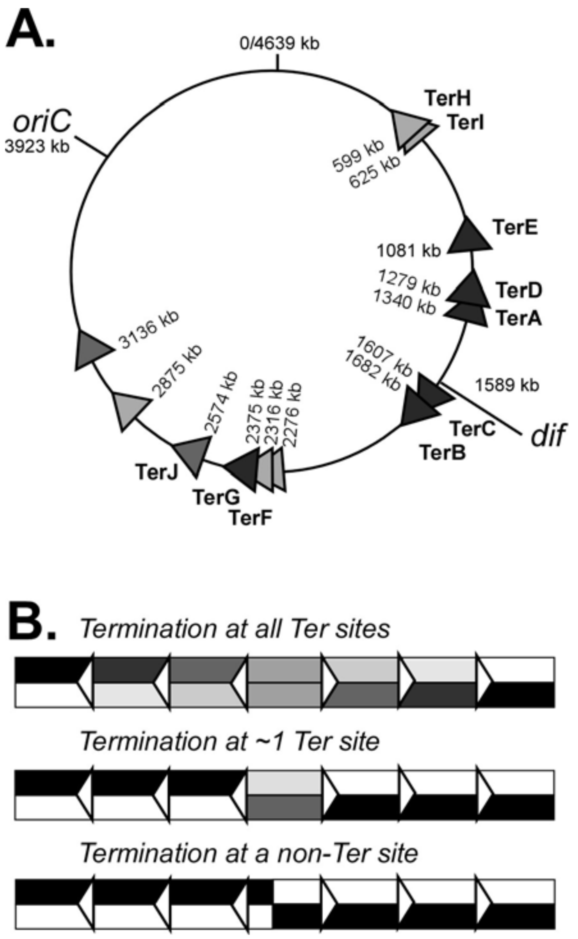

Figure 11 Models of replication termination in *E. coli*.

3.2.1.1 Figure 11 legend

A. Positions of Ter sites in *E. coli*; genome positions correspond to the *E. coli* K12 sequence. Ter sites are depicted as triangles; dark triangles are perfect matches to the consensus, medium and light grey triangles show one or two mismatches, respectively, at allowed variable positions. Ter sites are labeled according to those identified in Coskun-Ari and Hill (1997). **B**. Alternative models for replication termination. Triangles denote Ter sites; the color of the Watson and Crick strands denotes the strength of their leading-strand character.

While replication termination is a universal problem shared by all organisms with circular chromosomes, the Ter/Tus system is not. Unlike the broadly conserved *dnaA* gene which mediates replication initiation, plausible homologues of the *tus* gene are only found in close relatives of *E. coli* and on some plasmids (NEYLON *et al.* 2005). This distribution belies the central importance of replication termination and suggests that the Ter/Tus system is merely a recent addition to the enteric bacterial lineage. Replication forks are arrested by the analogous – but structurally non-homologous (BUSSIERE and BASTIA 1999; WAKE 1997) – Ter/*rtp* system in *Bacillus subtilis*, which is again restricted in its phylogenetic distribution. One might expect that proteins or other factors participating in such a central process would be broadly distributed, as are those involved in replication initiation and elongation, transcription initiation and elongation, and translation initiation, elongation and termination. Considering its central importance, more questions are perhaps raised by the Ter/*tus* and Ter/*rtp* systems than have been solved: (a) Why is the "replication trap" so large? The inner-most Ter sites are spaced ~270 kb apart in *E. coli*, or more than 5 % of the genome. In contrast, the structurally homologous Ter sites of plasmid R100 are separated by only 120bp, or 0.1 % of the genome (HIDAKA *et al.* 1988; HORIUCHI and HIDAKA 1988). (b) If the supposedly redundant Ter sites provide a "back-up" of the inner-most Ter sites, why are they found up to 1,500,000 bp away from those sequences in *E. coli*, some in closer proximity to the replication origin than to the supposed terminus (Fig. 11A)? (c) If the Ter/Tus interaction mediates the critical process of replication termination, especially in its role as a cell division check-point (PERALS *et al.* 2001; WANG *et al.* 2005), why can the *tus* gene be deleted with no obvious phenotype in otherwise wild-type cells (HILL 1992; ROECKLEIN *et al.* 1991; SKOKOTAS *et al.* 1994)? And why is this protein not conserved broadly among bacteria?

Although molecular biological assays demonstrate unequivocally that replication forks do pause at Ter sites in the presence of the Tus protein, it is not clear that (a) forks originating from the chromosomal origin of replication (*oriC*) – or other ectopic origins – have stalled at Ter site or at other nearby sites, or (b) if stalled forks detected at Ter sites originated from *oriC*. For example, synchronous DNA replication was achieved in an *oriC*^{TS} mutant using a unidirectional *ori*R1, and branched structures corresponding to stalled forks were detected at the TerA site (MAISNIER-PATIN *et al.* 2001). Yet it is not clear if these stalled forks originated from *ori*R1; indeed, their abundance was far less than expected if 100% of the cells had stalled replication there. Moreover, chromosome copy number was not measured at other loci to determine if replication termination occurred elsewhere. Ultimately, it is not clear if (a) Ter sites are retained because they halt replication forks originating from *oriC* as has been proposed, (b) Ter sites act primarily to halt replication forks that initiate upon the repair of DNA damage, or (c) stalled forks are a secondary effect of Tus binding, and the Ter/Tus interaction serves another primary purpose in the cell (just as LacI binding to *lac* operators results in transcription termination from upstream promoters while it also prevents activation of the *lacZYA* promoter by binding there as a repressor). While the Ter/*tus* model is tempting in its simplicity, similar concerns have been voiced almost since the model's inception (HILL 1992).

To assess the role of Ter sites in the termination of replication forks originating from *oriC*, we use a bioinformatics approach to locate the "historical" replication origin and terminus in bacterial chromosomes, provided these positions have been stable over evolutionary time (HENDRICKSON and LAWRENCE 2006). This is possible because mutational biases between leading and lagging strands make them compositionally distinct (LOBRY 1996; LOBRY and SUEOKA 2002); as a result, the replication origin and terminus are evident as locations where a continuous DNA strand switches from being replicated as a leading strand to being replicated as a lagging strand (CAPIAUX *et al.* 2001; GRIGORIEV 1998; LOBRY and LOUARN 2003; SALZBERG *et al.* 1998). Our purpose is not to locate the origin and terminus *per se*, but to use strand bias signatures to determine if the primary replication "terminus" maps to Ter site(s), or to some other non-Ter site. If replication termination occurs at Ter sites, we can quantify the fraction of termination events at each Ter site by quantitating changes in mutational bias. If the replication terminus is found elsewhere, we can identify this location as the position where strand identity changes from leading strand to lagging strand, and determine if this position is consistent across lineages.

3.3 METHODS

3.3.1 Genome sequences

The genome sequences for *Bacillus cereus* E33L, *Bacillus subtilis* 168, *B. licheniformis* ATCC14580, *Bacillus halodurans* C-125, *Corynebacterium glutamicum* ATCC13032, *Enterococcs faecalis* V583, *Erwinia carotovora* SCRI1043, *Escherichia coli* K12, *Frankia alni* ACN14a, *Haemophilus influenza*e Rd, *Listeria monocytogenes* 4b F2365, *Mycobacterium avium* K-10, *Nocardia farcinica* IFM 10152, *Pasteurella multocida* Pm70, *Propionibacterium acnes* KPA171202, *Pseudomonas aeruginosa* PAO1, *Pseudomonas syringae* DC3000, *Salmonella enterica* Typhimurium LT2, *Shewanella oneidensis* MR-1, *Staphylococcus aureus* MW2, *Thermobifida fusca* YX, *Vibrio cholerae* N16961, *Xanthomonas campestris* 8004 and *Yersinia pestis* CO92 downloaded from GenBank.

3.3.2 Detecting skewed octamers

Octamers were classified as matching IUB nondegenerate (GATC) and degenerate (RYMK) bases. Watson strands are defined as the DNA strand reported in GenBank files; Crick strands are complements of Watson strands. Leading strands are defined as Watson strands downstream, and Crick strands upstream, of the replication origin. Skewed octamers were detected as those sequences overrepresented on leading strands. AIMS (HENDRICKSON and LAWRENCE 2006) were detected as octamers with higher abundance near the replication terminus, as measured by χ^2 analysis, than predicted from the remainder of the genome. Skew is defined as the proportion of oligomers on the leading strand:

$$Skew = \frac{N_{Leading}}{N_{Leading} + N_{Lagging}}$$

3.3.3 A statistical test for change in skew

To detect a site where the degree of octamer skew changes, we quantified strand bias upstream ($\mathrm{Skew}_{\mathrm{Left}}$) and downstream ($\mathrm{Skew}_{\mathrm{Right}}$) of each octamer's position in the region analyzed. Skew differential (Differential) was defined as the absolute value of the difference between these values and the overall skew of the region ($\mathrm{Skew}_{\mathrm{Overall}}$), weighted by the number of octamers in each portion:

$$Differential = \frac{N_{Left}}{N}\left|Skew_{Left} - Skew_{Overall}\right| + \frac{N_{Right}}{N}\left|Skew_{Right} - Skew_{Overall}\right|$$

The position of change in octamer bias corresponds to position of maximum skew differential. To evaluate the significance of the skew differential, a randomization test was devised whereby strand identity – Watson or Crick – was randomly assigned to each octamer while preserving the overall strand bias. The significance was calculated as the fraction of randomized trials which yield maximum skew differentials at least as large as the original; a total of at least 10,000,000 randomization trials were performed to obtain a P-value.

3.3.4 Ter and *dif* sites

Ter sites in enteric bacteria were detected as those matching the 16 bp consensus sequence 5'-AGNATGTTGTAAYKAA, allowing substitutions at bases 1, 4 and 16 as described (COSKUN-ARI and HILL 1997). The *E. coli dif* site was defined as the sequence 5'-GGTGCGCATAATGTATATTATGTTAAAT (BLAKELY and SHERRATT 1994); the *dif* sites in the genomes of *S. enterica* and *E. carotovora* were found by virtue of both strong similarity to this sequence and similar location within the genome. A consensus sequence of 5'-RNTKCGCATAATGTATATTATGTTAAAT was used to locate putative *dif* sites in γ-proteobacterial genomes. Ter sites were detected in the *B. subtilis* genome as matching the consensus sequence 5'-KMACTAANWNNWCTATGTACYAAATNTTC as described (WAKE 1997). The *B. subtilis dif* site was defined as the sequence 5'-ACTTCCTAGAATATATATTATGTAAACT (SCIOCHETTI *et al.* 2001). A consensus sequence of 5’-ACTKYSTAKAATRTATATTATGTWAACT was used to locate putative *dif* sites in Firmicute genomes. A consensus sequence of 5’-TTSRCCGATAATVNACATTATGTCAAGT was used to locate putative *dif* sites in Actinobacterial genomes.

3.4 RESULTS

3.4.1 The *E. coli* genome has a single replication terminus

To characterize the nature of the replication terminus in *E. coli*, we measured genome-wide oligonucleotide skew, focusing on octamers. Briefly, compositional differences between leading and lagging strands result in differential abundance of nucleotides and oligonucleotides on these strands. We located the *E. coli* replication origin (*oriC*) at position 3923 kb as described (MEIJER *et al.* 1979), and we identified Ter sites as having strong matches to the published consensus sequence (COSKUN-ARI and HILL 1997), which detected all of the named Ter sites (Fig. 11A). We defined two replicores as the regions extending from *oriC* and continuing to the Ter sites located at positions 1081 kb (TerE) and 2315 kb (TerF), encompassing 73% of the genome. We excluded the TerE - TerF region to allow examination of octameric skew on both sides of the TerA, TerB, TerC and TerD sites.

To define a replication signature, we identified octamers that were overrepresented on leading strands in the TerF-*oriC*-TerE region. We found 136 non-degenerate octamers that were 70% skewed to the leading strand with at least 340 copies in the genome (Fig. 12A). We propose using this leading-strand signature to identify the replication terminus as the location where the leading strand moves from the Watson strand to the Crick strand. Alternatively, the lack of a specific replication terminus would result in a zone of low strand bias, where the Watson strand may be replicated as either a leading strand or a lagging strand (Fig. 11B); in this case, there would be more than one location of change in strand identity.

Complicating this analysis are Architecture Imparting Sequences (AIMS) (HENDRICKSON and LAWRENCE 2006), which are under selection for function and accumulate in abundance on leading strands towards replication termini (HENDRICKSON and LAWRENCE 2006; LAWRENCE and HENDRICKSON 2003; LAWRENCE and HENDRICKSON 2004); some have been proposed to direct the FtsK proteins towards the *dif* site (BIGOT *et al.* 2005). Therefore, AIMS do not provide an impartial indicator of mutation bias (HENDRICKSON and LAWRENCE 2006).

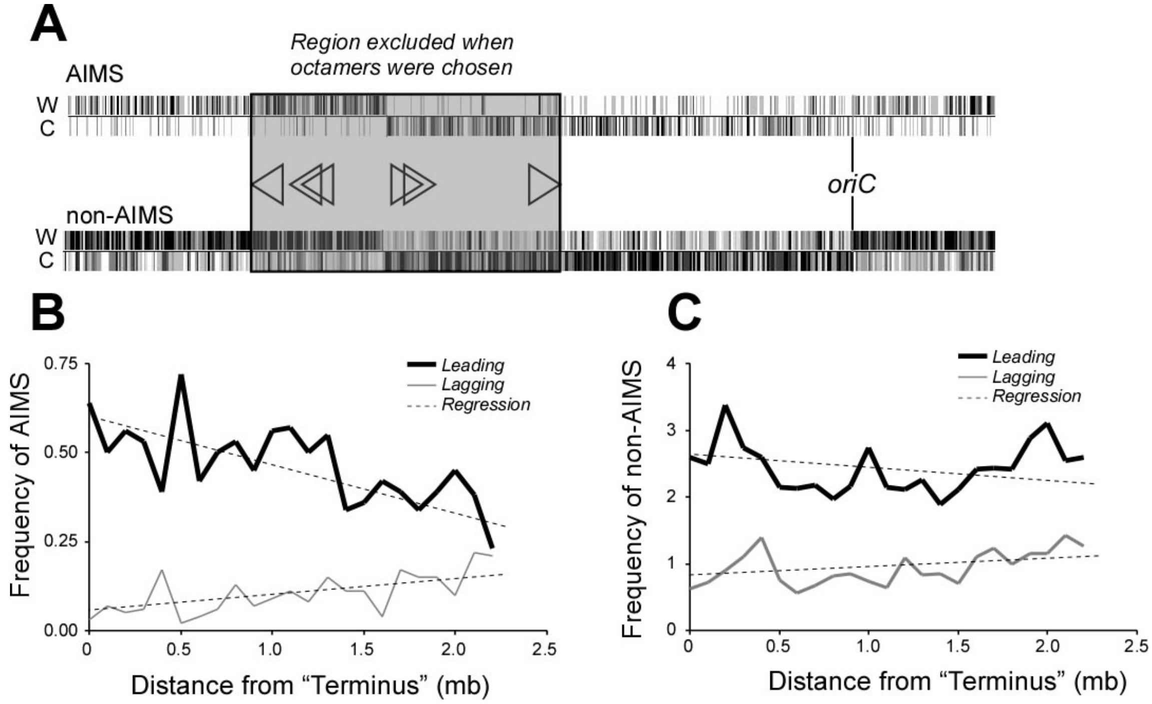

Figure 12 Use of replication strand bias to characterize the terminus region.

3.4.1.1 Figure 12 legend

The terminus zone is defined as the region between the three most origin-distal Ter sites on each replicores. **A**. Strand biased octamers were defined as those over-represented (70 %) on leading strands in the region outside the Ter zone. Positions on Watson and Crick strands (W, C) are shown as vertical lines. Positions of Ter sites are noted as triangles. **B.** Strand bias of AIMS octamers. AIMS octamers (HENDRICKSON and LAWRENCE 2006) were defined as those which increased 1.5-fold from origin to terminus. **C.** Strand bias of Non-AIMS octamers.

The presence of AIMS affects all measures of mutation bias, including widely-used GC-skew metrics and this potentially confounding influence must be removed. To arrive at an unbiased set of oligomers, we eliminated 30 AIMS which increased in abundance at least 1.5-fold toward the replication terminus (Fig. 12B). The remaining 310 octamers showed no significant increase in abundance towards the replication terminus (Fig. 12C) and therefore are taken to represent the signature of strand-specific mutational bias alone.

We examined the distribution of these non-AIMS octamers in the region between TerE and TerF (Fig. 13A). The Ter/Tus model predicts that there should be no change in strand identity between Ter sites. If termination has used both Ter sites with comparable frequency, then the Watson strand of the TerA-TerC region would be replicated sometimes as a leading strand and sometimes as a lagging strand (Fig. 11B) and strand bias would be less pronounced here than in the TerD-TerA or TerC-TerB regions. In addition, the TerA-TerC region should show no single point of unambiguous transition between leading and lagging strand signature. If termination favors either TerA or TerC (LOUARN *et al.* 1991), then the transition between leading and lagging strand identity should occur at one of these Ter sites (Fig. 11B). If replication forks have bypassed the TerA or TerC sites and halted when they encountered the TerD or TerB sites, then the strand bias of the TerE-TerD and TerB-TerF regions should exceed that of the "inner-Ter" region as well (Fig. 11B).

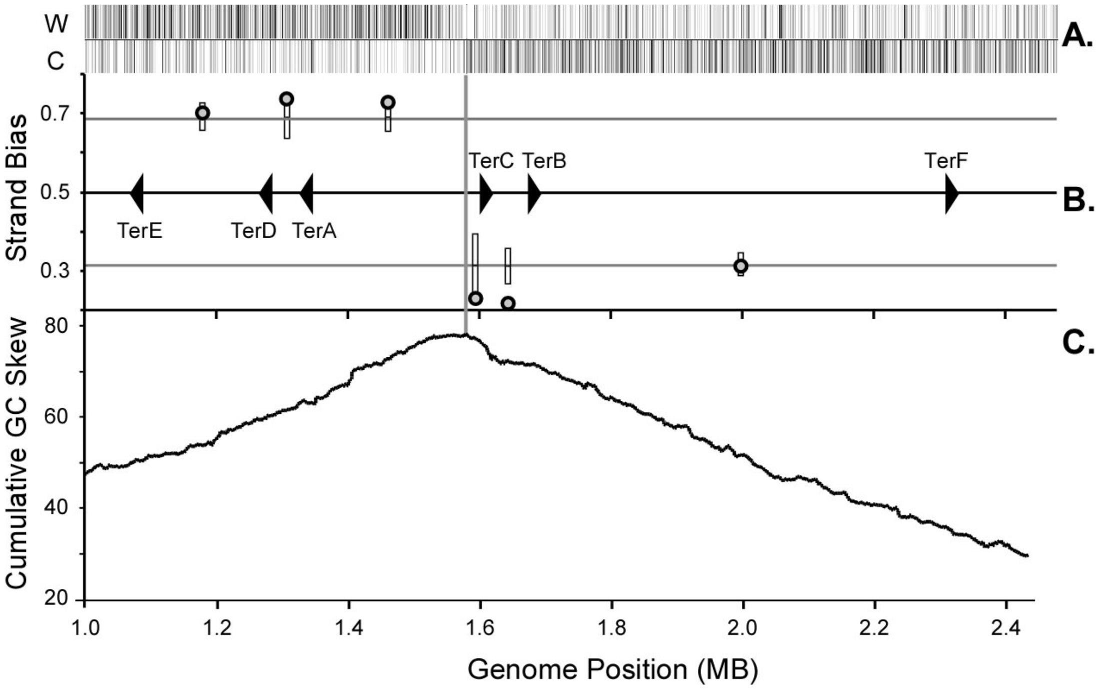

Figure 13 Replication strand bias used to quantify the replication termination site.

3.4.1.2 Figure 13 legend

A. Non-AIMS octamers (see Fig 2) were identified from the chromosome region outside the terminus zone; their positions are shown here within terminus zone. **B**. The strand bias of leading and lagging strand outside the terminus zone are shown as grey lines. The position of change in strand bias identity was determined by visual inspection and assigned to 1585 kb; this is noted with a vertical line. Triangles denote positions of Ter sites. The strand biases of regions between Ter sites – or between Ter sites and the site of change in strand identity – are shown as open circles. Bars depict intervals of 1 standard deviation above and below the mean bias of equally-sized intervals in the non-terminus zone. The vertical gray line indicates the apparent position of change in strand bias. **C**. The cumulative GC skew of the third codon positions of genes in this interval.

The distribution of non-AIMS octamers suggests that there is a single point where Watson-strand identity switches from the leading strand to the lagging strand (Fig. 13A), this breakpoint is not any of the previously-identified Ter sites; rather, it is between the two inner-most Ter sites. This site of change in strand bias is also seen in the plot of cumulative GC-skew (Fig. 13C); although this metric has not eliminated the potentially confounding influence of AIMS, it shows that octameric skew accurately reflects overall nucleotide skew. In addition, the DNA between any two Ter sites, or between Ter sites and the apparent point of change in strand bias, is no less strand-biased than other origin-proximal intervals (Fig. 13B). These data suggest both that replication termination has historically occurred primarily at a non-Ter location, and that no significant replication termination is apparent at any of the six most origin-distal Ter sites. To establish these points rigorously, we developed a statistical procedure for locating positions of change in strand bias and evaluating their significance.

3.4.2 Replication termination has historically occurred at a specific site between the two inner-most Ter sites

To determine if termination between the inner-most Ter sites is robust and significant, we enumerated strand-biased octamers in the origin-proximal 94.2% of the *E. coli* genome – outside the inner-most Ter sites, TerC and TerA. We eliminated the AIMS and used the remaining octamers as signatures for leading-strand identity, examining their distribution in the inner-Ter region. To quantify change in strand bias, we calculated the bias toward the Watson strand both upstream and downstream of each octamer's position; the skew differential was defined as the absolute value of the difference between these values and the skew of the overall region, weighted by the number of octamers in each region.

Figure 14A shows a plot of skew differential with genome position in the *E. coli* inner-Ter region. Upstream of 1580 kbp, strand-biased octamers are found 77.3% on the Watson strand, whereas downstream of this point these same octamers are found only 26.3 % on the Watson strand; weighting for the different lengths of these regions, this represents an average difference of about 10% from the overall bias of 71.1% on the Watson strand. These data suggest that replication termination has occurred at genomic position 1580 kb. To evaluate the significance of this skew differential, we used the randomization test described above; an example of one randomization trial is shown in grey in Fig. 14A. The distribution of maximum skew differentials for randomized octamer distributions is shown in the inset in Fig. 14A, where the mean differential is ~2%; it is clear that is it highly unlikely to have observed a skew differential of ~10% with randomized octamers ($P < 0.0000001$).

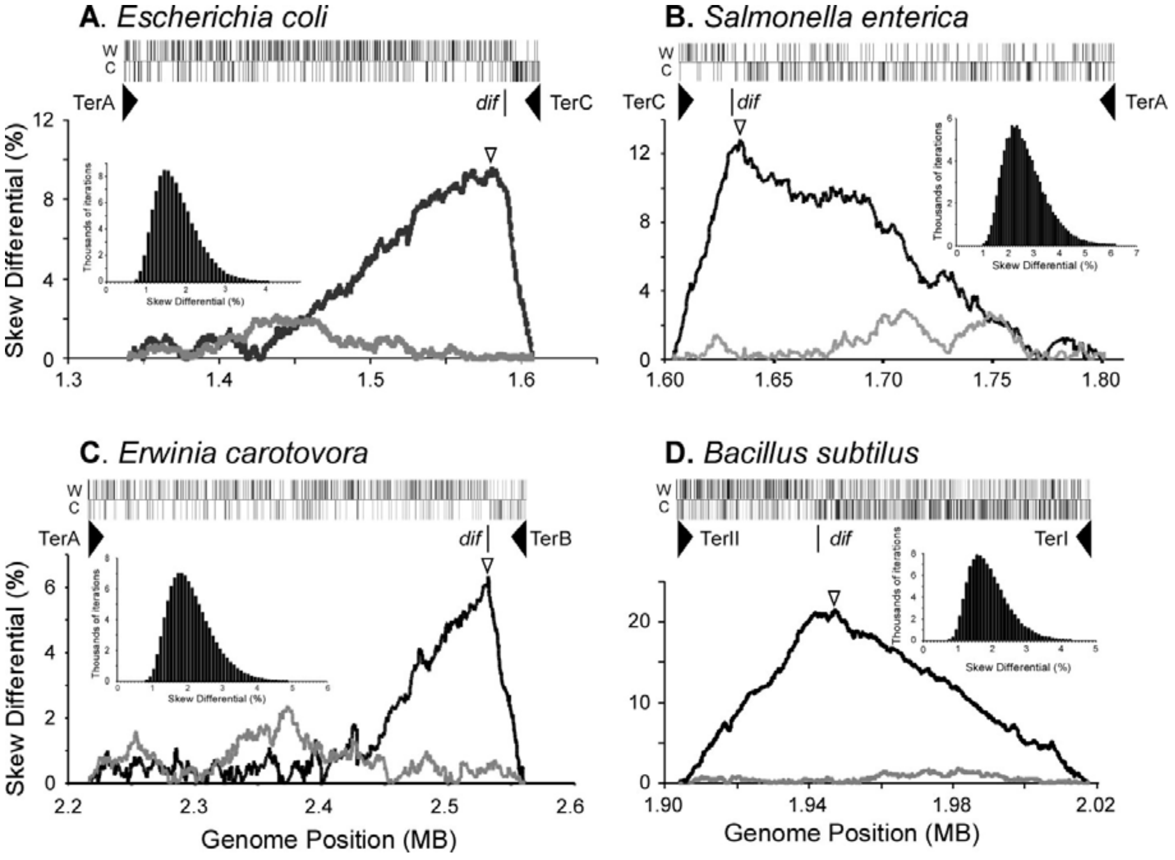

Figure 14 Detecting a shift in strand bias between the 'inner-most' Ter sites.

3.4.2.1 Figure 14 legend

Strand-biased octamers were enumerated in the region outside the two most origin-distal Ter sites; the positions of octamers within the inner-Ter regions were then determined. **A**. Sliding window analysis of change in strand bias in the *E. coli* K12 genome. Positions of strand-biased octamers on Watson and Crick strands (W, C) within the inner-Ter region are depicted above. Strand-bias is calculated as the percent of octamers on the Watson strand; strand bias differential is the absolute value of the difference in strand bias of the regions upstream and downstream of each point. The inset shows the distribution of values for maximum skew differential for when octamers' positions are randomized. The open triangle indicates the point of maximum skew differential. **B**. Sliding window analysis of change in strand bias in the *S. enterica* serovar Typhimurium genome. **C.** Sliding window analysis of change in strand bias in the *E. carotovora* genome. **D**. Sliding window analysis of change in strand bias in the *B. subtilis* genome.

To determine if these result are robust in the face of mutational change, and do not reflect a recent inversion in the region adjacent to a Ter site, we examine genomes of bacteria related to *E. coli*.. Genes in *S. enterica* serovar Typhimurium are ~85% identical to their *E. coli* homologues, so that the positions of individual octamers are typically not conserved; yet a statistically significant change in strand identity is again evident between the two inner-most Ter sites (P < 0.0000001, Fig. 14B). Similar results were seen for the genomes of *E. coli* O157 and *S. enterica* serovar Paratyphi (data not shown), as well as in the genome of the even most distantly-related enteric bacterium *E. caratovora* (P < 0.0000001, Fig. 14C).. The *dif* sequence, the site of action of the XerCD site-specific recombinase (BLAKELY and SHERRATT 1994), is located very close to the site of strand-bias change in the *E. coli, S. enterica* and *E. carotovora* genomes (Figs. 14ABC). These results suggest that the replication terminus maps close to the *dif* site, rather than to any Ter site, in enteric bacteria. The occurrence of a specific termination site between Ter sites is not excluded by any previous analysis[*e.g.,* (DE MASSY *et al.* 1987; KUEMPEL *et al.* 1977; MAISNIER-PATIN *et al.* 2001; PELLETIER *et al.* 1988)] which lack the resolution to discriminate between Ter sites and the *dif* site.

Strand bias was similarly examined in *Bacillus subtilis*, where replication termination has been associated with the analogous, but not homologous, Rtp protein acting at Ter sites (BUSSIERE and BASTIA 1999; WAKE 1997). As with the enteric bacteria, strand-biased octamers were enumerated in the region excluding all Ter sites, AIMS were ignored, and the positions of remaining octamers were determined in the region between the inner-most Ter sites (Fig. 14D). A change in strand bias was again observed between the inner-most Ter sites (P<0.0000001); the large skew differential – greater than 20% – reflects the stronger strand bias in Firmicutes

(Rocha 2004). As in the enteric bacteria, the *dif* site was located very near to this bioinformatically-determined site of change in strand bias.

To eliminate any confounding influence of transcription bias –lagging strands are more often template strands for transcription, especially in the Firmicutes (Rocha 2004) – we constructed derivatives of the *E. coli* and *B. subtilis* genomes with all genes encoding proteins, tRNAs, tmRNAs and rRNAs removed; as a result, these "genomes" contained only the non-coding spacers between genes. Due to the small size of these "genomes," strand bias was examined by calculating GC skew (the ratio of G-C to G+C) for 100 bp windows. The plot of cumulative GC skew with genome position shows a clear inflection point at the *B. subtilis dif* site, between the two inner-most Ter sites, again supporting the conclusion that Watson strands change from leading strands to lagging strands at this point (Fig. 15). Similar results are seen for *E. coli* and other enteric bacteria (data not shown), although the distance from the *dif* site to the nearest Ter site is far smaller in these geneless genomes (Fig. 14). While these data have not accounted for the potentially confounding influence of AIMS, the results of above analyses have shown that inclusion of AIMS does not change the conclusions drawn.

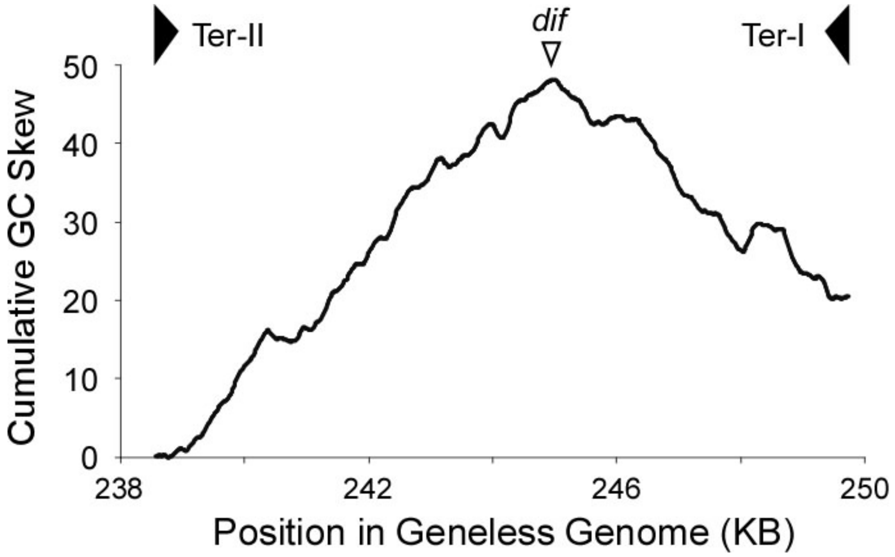

Figure 15 Cumulative skew of the geneless version of the *B. subtilis* genome.

3.4.2.2 Figure 15 legend

Cumulative GC skew is plotted for 100 nucleotide windows.

3.4.3 No detectable replication termination has historically occurred at the inner-most Ter sites

While the previous analysis demonstrates that replication termination has historically occurred at a position very near the *dif* site in both γ-Proteobacteria and Firmicutes, it is still possible that replication forks also arrest at Ter sites should they often fail to halt at the bioinformatically-defined terminus. If forks originating from *oriC* passed the *dif*-associated terminus and halted at the first Ter site they encountered, then the region between the Ter site and the *dif*-associated terminus would be less strand-biased than the region on the origin-side of at least one Ter site

(Fig. 11B). We examined the *E. coli* genome for strand biased octamers and assessed whether Watson strands were more biased on the origin-proximal sides of the two “inner-most” Ter sites (TerA and TerC) than the regions on the *dif*-proximal sides (Fig. 16A). If so, then we would expect to find a peak in skew differential at a Ter site, where the genome would be more biased on the origin side. Yet we found no change in strand bias associated with the Ter sites on either side of the *dif* site ($P > 0.05$). The sites of maximum skew differential in these regions were not located near Ter sites. More importantly, the change in skew at these sites showed that the *dif*-proximal region was actually somewhat more-strand biased, not less strand-biased (Fig. 16A). Therefore, these “peaks” do not correspond to cryptic Ter sites, but represent only the stochastic distribution of octamers. Similar results were observed for the *B. subtilis* genome (Fig. 16B), where there was no significant change in strand-bias across Ter sites ($P>0.05$) or any other location except the *dif* site. These data suggest that neither *E. coli* nor *B. subtilis* Ter sites participate significantly in stopping those replication forks that produce the mutational bias we are examining.

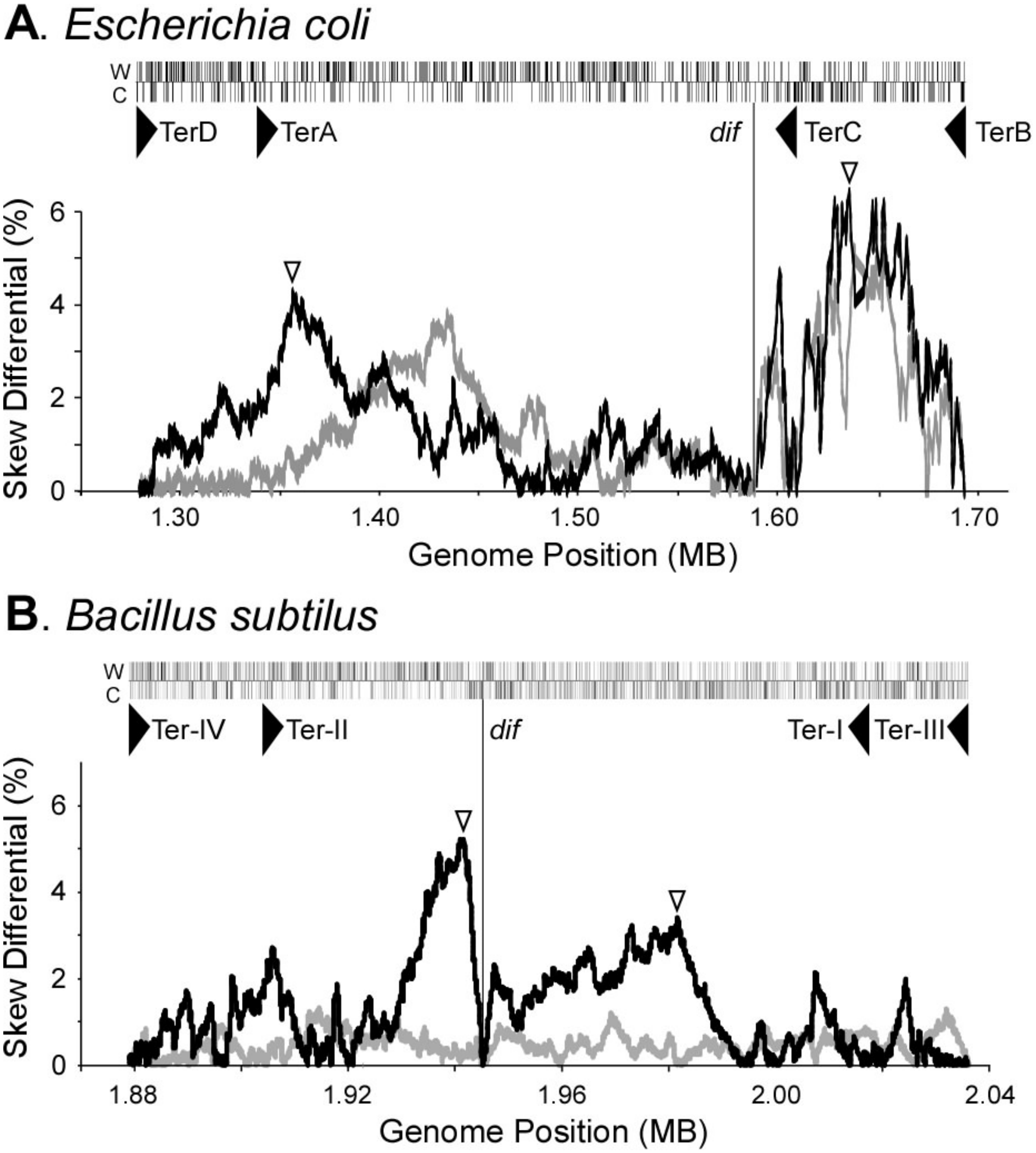

Figure 16 Lack of change in strand bias across Ter sites in the *E. coli* and *B. subtilis*.

3.4.3.1 Figure 16 legend

A. Lack of change in strand bias across Ter sites in the *E. coli* genome. Strand-biased octamers were enumerated in the region outside the two most origin-distal Ter sites; the positions of octamers on the Watson and Crick strands (W, C) within the TerB-TerD region were then determined. The regions from TerB to *dif*, and from *dif* to TerD were analyzed separately. Strand-bias is calculated as the percent of octamers on the Watson strand; strand bias differential is the absolute value of the difference in strand bias of the regions upstream and downstream of each point. Open triangles indicate the point of maximum skew differential for each analysis. **B**. Lack of change in strand bias across Ter sites in the *B. subtilis* genome; analysis was performed as in part A.

3.4.4 The mutational bias defining the *dif* site also defines *oriC*

We used octamers skewed on either side of the replication origin to locate the replication terminus (Figs. 12, 13, 14), postulating that the mutational bias defining the terminus was imparted by replication forks originating at *oriC*. If so, then octamers skewed on either side of the *dif* site should similarly identify the replication origin. To test this hypothesis, we analyzed the *E. coli* genome for octamers that were strand biased in particular 50%-genome intervals. In each case we defined a central "breakpoint" and identified octamers that were biased to the Watson strands in the 25% of the genome upstream – and to the Crick strands in the 25% of the genome downstream – of these points. We analyzed several hundred breakpoints throughout the *E. coli* genome. Not surprisingly, there were two locations where numerous octamers were over-abundant on different strands (Watson or Crick) upstream and downstream of these points (Fig. 7A); these positions correspond to the replication origin and replication terminus. The replication terminus has a stronger signal than does the replication origin; keeping in mind that only 50% of the genome is analyzed for any location, this signal may represent the overabundance of AIMS near the replication terminus, increasing the strand bias there (HENDRICKSON and LAWRENCE 2006; LAWRENCE and HENDRICKSON 2004).

We then analyzed the distributions of octamers which defined four particular breakpoints (Figs. 17BCDE). Not surprisingly, the few octamers over-represented on different strands on either side of positions located in the middle of replicores (genome positions 436 kb and 2756 kb) were completely unbiased in the portion of the genome not examined when these octamers were selected (Figs. 17BD). That is, the degree of strand-bias observed for these octamers in the regions analyzed was purely the result of stochastic processes, and outside these regions these octamers were equally abundant on both strands. In contrast, strand-biased octamers identified in the terminus region also showed a clear change in stand-bias at the replication origin (Fig. 17C), and vice-versa (Fig. 17E). These data establish that the mutational biases defining the replication terminus appear to have been imparted by forks originating from *oriC*.

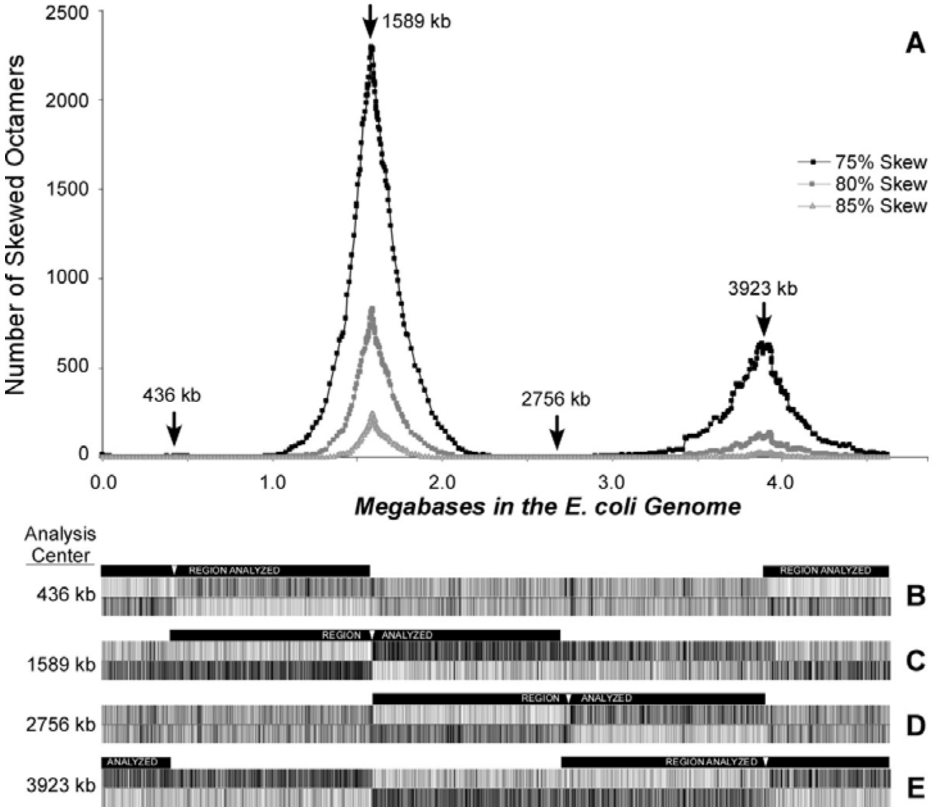

Figure 17 Analysis of strand bias in *E. coli* genome.

3.4.4.1 Figure 17 legend

A. Breakpoint permutation analysis. Strand-biased octamers are enumerated in regions corresponding to 25% of the length of the genome upstream and downstream of each genome position. A minimum of 50 octamers must be present in this region; curves are shown for sets of octamers that are 75%, 80%, 85% biased to the Watson strand downstream of each position and to the Crick strand upstream**.**

(Figure 17 cont.)

B-E. The positions of strand-biased octamers within the E. coli genome. The octamers used correspond to those detected in part A using the four genome positions indicated. Parameters were chosen to select ~500 octamers (allowing 2 bases of degeneracy) for each set. The regions used to detect octamers is shown above the octamers position map. **B**. Genome position 436 kb was selected as mid-way between the two peaks see in part A. N>21; bias > 72%. **C**. Genome position 1589 kb corresponded to the primary peak in part A. N>84; bias > 80%. **D**. Genome position 2756 kb was selected as mid-way between the two peaks see in part A. N>20; bias > 71%. **E**. Genome position 3923 kb corresponds to the secondary peak in part A. N>54; bias > 75%.

3.4.5 Replication termination occurs near the *dif* site in diverse γ-Proteobacteria and Firmicutes.

The proximity of the *dif* site to the bioinformatically-inferred replication terminus is observed in the genomes of other γ-proteobacteria and Firmicutes (Fig. 18). Here, we used the sequence of the *E. coli* and *B. subtilis dif* sites to search for similar sequences in genomes of representative members of the phyla γ-Proteobacteria and Firmicutes, respectively; we did not examine genomes where rearrangements have precluded the unambiguous identification of the replication origin. In most genomes, a single sequence with strong similarity to a molecularly-defined *dif* site was recognized. We inferred an approximate location for the replication origin and terminus using cumulative GC-skew of third codon positions and gene orientation bias as described (HENDRICKSON and LAWRENCE 2006). Skewed octamers were identified within the origin-proximal portion of each genome, eliminating potential AIMS from these data sets. We then refined the position of the replication terminus by determining the locations of skewed octamers within an 80-kb region flanking the approximate replication terminus. Our localization of replication termini closely matched those described in the Genome Atlas Database (HALLIN and USSERY 2004). Strikingly, the bioinformatically-defined replication termini – located at the peaks of the skew differential curves – were very close to the putative *dif* sites in the genomes of all γ-Proteobacteria and Firmicutes we analyzed (Fig. 18). While these data do not exclude the possibility that as-yet-unidentified Ter sites are acting at these locations, known Ter sites are more distantly situated, being tens of kilobases away from the replication termini we find (Figs.

14, 15). Therefore, we conclude that the replication terminus is generally associated with the *dif* site in γ-Proteobacteria and Firmicutes.

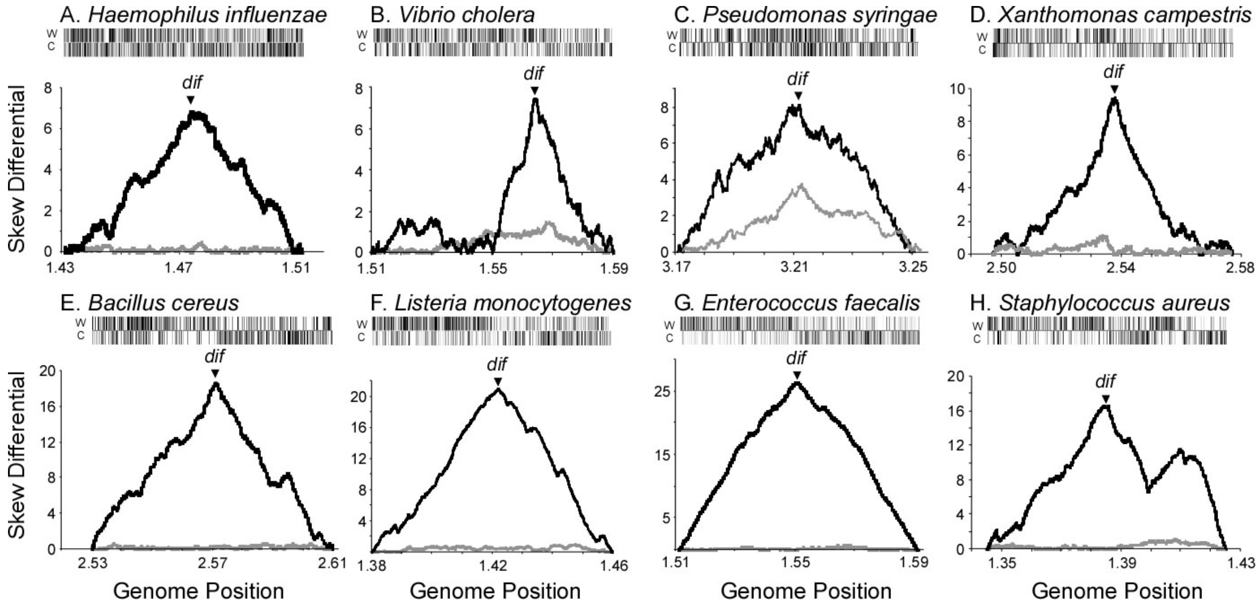

Figure 18 Bioinformatically defined replication termini and putative *dif* sites.

3.4.5.1 Figure 18 legend

Localization of bioinformatically-defined replication termini and putative *dif* sites in the genomes of γ-proteobacteria and Firmicutes. **A**. *H. influenzae* terminus, inferred as the site of octamer skew change, is genome position 1473765 bp; *dif* site is position 1472962 bp. **B**. *V. cholera* terminus, 1564066 bp; *dif* site, 1564104 bp. **C**. *P. syringae* terminus, 3209668; *dif* site, 3211773 bp. **D**. *X. campestris* terminus, 2537901 bp; *dif* site, 2537463 bp. **E**. *B. cereus* terminus, 2571079 bp; *dif* site 2570999 bp. **F**. *L. monocytogenes* terminus, 1421940 bp; *dif* site, 1421892 bp. **G**. *E. faecalis* terminus, 1550406 bp; *dif* site, 1550523 bp. **H**. *S. aureus* terminus, 1385620 bp; *dif* site, 1384864 bp.

3.5 DISCUSSION

Our data strongly suggest that replication termination is a far more active and controlled process than previously envisioned. Under the Ter/Tus model, replication forks are allowed to collide anywhere in the genome, but they will do so more often (a) at Ter sites, where one fork will be transiently stalled, and (b) in the region of the chromosome furthest from the replication origin. Yet our data suggest that replication forks originating from *oriC* only meet at the *dif*–associated terminus, preventing frequent collisions at any other location. If replication termination does not involve the action of Tus/Rtp at Ter sites, two questions are raised: 1) if *oriC*-born replication forks do not halt at Ter sites, what sequences do mediate termination? and (2) if they are not used for terminating *oriC*-born forks, what function do Ter sites serve?

3.5.1 The *dif* site is strongly associated with replication termination

The bioinformatically-defined replication terminus is found very close to the *dif* site in both γ-Proteobacteria and Firmicutes (Fig. 18). The XerCD recombinase acts at the *dif* site to resolve chromosome catemers following replication termination; it is activated and delivered there by the FtsK translocase (BIGOT *et al.* 2004; BIGOT *et al.* 2005; IP *et al.* 2003; MASSEY *et al.* 2004; YATES *et al.* 2006). FtsK, in turn, acts to apportion DNA among daughter cells, moving towards the *dif* site as directed by strand-biased sequences – termed AIMS (HENDRICKSON and

LAWRENCE 2006) or KOPS (BIGOT *et al.* 2005) – which originate from replication-induced strand bias acted upon by natural selection (HENDRICKSON and LAWRENCE 2006). While the proximity of the replication terminus to the *dif* site is likely not coincidental (Fig. 18), we do not believe that the *dif* site also acts as the replication terminus. The minimal 28 bp *dif* sequence alone is insufficient to act as a terminus because this sequence may be placed in additional, ectopic locations with no drastic phenotypic effects (CORNET *et al.* 1996; PÉRALS *et al.* 2000). In this regard, we infer that the *dif* site and the replication terminus are separate sites. However, if replication-imparted polarity is used to direct FtsK and other proteins to the *dif* site (CORRE and LOUARN 2002; HENDRICKSON and LAWRENCE 2006), then natural selection would favor close proximity of the *dif* site and the terminus. That is, the *dif* region represents the nexus of cell division, integrating the processes of chromosome mobilization, dimer resolution via XerCD recombination, and replication termination itself. As a caveat, we do note that transient cleavage of the *dif* site by the XerCD recombinase will prevent replication forks from proceeding, but it must be rejoined to allow completion of lagging strand synthesis. In addition, stalled forks – historically considered the hallmark of replication termination – would not be evident here due to strand cleavage. Alternatively, head-on collision with incoming FtsK could stall DNA polymerase in the vicinity of the *dif* site, without requiring a specific termination site.

While the resolution of our methods prevents us from defining the site of the replication terminus more precisely than within a kilobase, the proximity of the terminus to the *dif* site could be used to deduce its sequence and location in organisms lacking molecular characterization of this critical component of the cell division machinery. To explore this possibility, we determined the location of replication termination in members of the Actinobacteria. In the genome of *Frankia alni*, the terminus – defined as the site of strand bias change – is located at base-pair

4049160; this position lies within a sequence with strong similarity to the known *dif* sites in Firmicutes and Proteobacteria (Table 1). Using these sequences as a guide, a good consensus *dif* site for Actinobacteria is found near the site of strand bias change in the genomes of many Actinobacteria (Table 1). These results suggest that locating the position of strand bias change may be an effective way of selecting candidate *dif* sequences for molecular characterization.

Table 2 *dif* sites found and their consensus sequences.

Family	Species	Position	*dif* SiteSequence
γ-Proteobacteria			GGTTCGCATAA TGTATA TTATGTTAAAT
Enterobacteriacae	*E. coli*	1588773*	---G------- ------ -----------
Enterobacteriacae	*S. enterica*	1629676	---G------- ------ -----------
Enterobacteriacae	*E. carotovora*	2532120	----------- ------ -----------
Enterobacteriacae	*Y. pestis*	2562906	---G------- ------ -----------
Pasteurellaceae	*H. influenzae*	1473962	AT--------- -A--A- -----------
Pasteurellaceae	*P. multocida*	713837	AC--------- ------ -----------
Vibrionaceae	*V. cholera*	1564104	A--G--T--T- -----G -----------
Shewanellaceae	*S. oneidensis*	2476915	AC-G----C-- ------ -----------
Pseudomonadaceae	*P aeruginosa*	2443068*	-A--------- ------ -----------
Pseudomonadaceae	*P. syringae*	3211773	-T-A------- ------ -----------
Xanthomonadaceae	*X. campestris*	2537463	AT--------- ------ ------C-GGA
Firmicutes			ACTTCCTATAA TATATA TTATGTAAACT
Bacillaceae	*B. subtilis*	1941799	--------G-- ------ -----------
Bacillaceae	*B. cereus*	2570999*	---G------- ------ ------T----
Bacillaceae	*B. licheniformis*	2030751	------G-G-- ------ -----------
Bacillaceae	*B. halodurans*	2243235	GG--------- ------ -----------
Peptococcaceae	*D. hafniense*	1827925*	GGG-------- --G--- ---------G-
Enterococcaceae	*E. faeclis*	1550523	----TG----- –G---- ------T----
Listeriaceae	*L. monocytogenes*	1421892	----------- ------ -----------
Staphylococcacae	*S. aureus*	1384864*	----------- ------ -----------
Actinobacteria			TTCGCCGATAA TVNACA TTATGTCAAGT
Corynebacteriaceae	*C. glutamicum*	1551501*	--GT------- -GT--- --------TT-
Frankiaceae	*F. alni*	4049147	CA--------- -GC--- -----------
Mycobacteriaceae	*M. avium*	1888576*	-CTA------- GCG--- -----------
Nocardiaceae	*N. farcinica*	3131987	-A--------- -CT--- ------T----
Propionibacteriaceae	*P. acnes*	1340138	--GA------- GAG--- --------TT-
Nocardiopsaceae	*T. fusca*	1779148*	A---------- -AA-T- -----------
Bacterial Consensus			DBBBCSBATAA TRTAYA TTATGTHAANT

* Complement of the *dif* site begins at this position

3.5.2 Roles of Ter sites in recombination and repair

Although Ter sites stall replication forks, these forks need not originate from *oriC*. It is possible that Ter sites act primarily to impede retrograde replication forks originating during DNA break repair (KREUZER 2005; KUZMINOV 1999). While some models for dsDNA break repair do not invoke DNA synthesis (KOWALCZYKOWSKI *et al.* 1994), these models did not accommodate the roles of *dnaB* (BRESLER *et al.* 1973; BRESLER *et al.* 1968; STALLIONS and CURTISS 1971) or *priA* (KOGOMA *et al.* 1996; SANDLER *et al.* 1996) in recombination. Moreover, DNA damage repair via the RecBCD pathway has been shown to stimulate *oriC*-independent DNA synthesis (ASAI *et al.* 1993; KOGOMA 1997; MAGEE *et al.* 1992). In addition, it has been argued that DNA synthesis must follow strand invasion to avoid endless cycles of recombination initiated by dsDNA ends (SMITH 1991). Since replication forks initiated by DNA repair resemble those originating from *oriC* [*e.g.*, they depend on PriA and DnaT (LARK and LARK 1979; MASAI *et al.* 1994)], and the frequency of recombination in the terminus region is high (CORRE *et al.* 1997; LOUARN *et al.* 1994), it is reasonable to posit that Ter sites play a role in halting the retrograde motion of these forks. Alternatively, Ter sites could foil non-*oriC* replication origins, such as those found on integrated plasmids or prophages (HILL 1992).

This function for Ter sites is consistent with their dispersal over a large region of the chromosome (Fig. 11). Their abundance in the terminus-half of the chromosome may reflect either the increased abundance of retrograde forks arising there; dsDNA breaks may arise from the greater supercoiling stress near the terminus, thus causing more frequent recombination (LOUARN *et al.* 1994), where this excess is not entirely attributable to Ter-paused forks

(HORIUCHI *et al.* 1994). Alternatively, retrograde forks may be more problematic near the terminus, where extra chromosome segments or polymerase collisions befuddle the orderly segregation of DNA into daughter cells. The action of FtsK near the terminus would increase the problems associated with supernumerary chromosome regions, and the region of the genome with Ter sites also have an excess of FtsK-loading sites (HENDRICKSON and LAWRENCE 2006; SIVANATHAN *et al.* 2006).

This model is supported by some otherwise paradoxical data regarding the frequency of usage of Ter sites in *E. coli*. Pelletier et al. (1988) created strains of *E. coli* with chromosomal inversions that moved the replication origin relative to the terminus. If replication from *oriC* were to terminate primarily at the initially-encountered Ter site, clearly the shorter replicore would finish first, and one “inner” Ter site would be used far more frequently than the other, since replication forks appear to move independently of one another (BREIER *et al.* 2005). Yet Ter sites in this inverted chromosome were used at the same frequency as in otherwise wild-type cells (PELLETIER *et al.* 1988). While recognized as inexplicable according to the conventional Ter model (HILL 1992), these data are entirely consistent with Ter usage primarily in halting repair-originating forks, since the creation and progress of these replication forks would be unaffected by the chromosomal inversions in those strains.

Similarly, the appearance of retrograde forks at artificial operator arrays near the replication origin has been attributed to their passage through Tus-bound Ter sites (POSSOZ *et al.* 2006). Yet forks do not arrive near the origin substantially more quickly in a *tus* mutant, demonstrating an additional impediment to retrograde forks. In addition, their arrival at the origin in *tus*$^{+}$ cells suggests that the *tetO* array is a more robust block to replication than eight or more Tus-bound Ter sites (Fig. 11A). Instead, we suggest that replication is blocked by the *dif* site, and that all forks arriving near the replication origin were spawned by DNA repair processes, explaining their arrival there at early time points even in *tus*$^+$ cells. Rather than removing blocks to retrograde forks, *tus* mutations increase the number of forks which can successful travel backwards to the *tetO* array.

Strand identity influences *in vivo* DNA metabolism. For example, ssDNA may be used for site-directed mutagenesis, but its efficacy is far higher when oligonucleotides are complementary to leading strands (used as lagging strand templates), likely because they are single-stranded when awaiting lagging strand synthesis at replication forks (COSTANTINO and COURT 2003; ELLIS *et al.* 2001). One could use such differences as reporters for strand identity at different chromosomal locations (PETERS and CRAIG 2001), potentially providing biochemical validation for bioinformatically-determined replication origins and termini. Yet the presence of replication forks having arisen from recombination and repair processes confounds the interpretation of these results, making unambiguous interpretations of strand identity difficult.

3.5.3 Could Tus act at a distance?

One interpretation of the changes in strand bias observed in Fig. 14 is that while Tus binds to Ter sites, it acts at a distance, halting replication forks near the *dif* site. We do not favor this interpretation for four reasons. First, the distance between the replication terminus and the closest Ter site is not constant (Fig. 14). Second, only a single site of change in strand bias was identified (Figs. 14, 16); similar sites were not observed adjacent to all Ter sites. Third, the position of strand bias change is located precisely at the Ter sites in plasmids R100 (data not shown), which are separated by only 120 bp. Since this plasmid carries no identifiable homologue of the *tus* gene, we posit that the enteric bacterial Tus protein mediates termination here. Lastly, the Tus protein has been demonstrated to halt termination < 100 bp from the Ter site (HILL and MARIANS 1990; MULCAIR *et al.* 2006); given the vagaries of DNA compaction, it is unlikely that a specific site of termination – as implicated by the sharp change in strand bias we observe – could be achieved kilobases away from the Tus binding site.

3.5.4 Could recombination at the *dif* site obscure termination occurring at Ter sites?

It has not escaped our attention that we are measuring strand bias as an historical archive of DNA replication, not the process of replication termination itself, and other processes may influence the patterns we observe. It is possible that both replication forks approach the *dif* site and pass it, each going on to terminate at their respective Ter sites. If so, then the region between the inner-most Ter sites would be replicated twice. Recombination at the *dif* site – mediated by the XerCD site-specific recombinase – could act to discard the "extra" DNA, and preserve the integrity of the strand-bias signature we observe. This model requires that replication forks must first collide and then pass each other on their way to their respective Ter sites. This behavior is not proposed for replication forks meeting at Ter sites or elsewhere (MULCAIR *et al.* 2006). While not impossible, replication forks passing one another is, at the molecular level, both non-trivial and nonsensical, since this action is precisely what a replication terminus is intended to prevent.

3.5.5 Could DNA polymerase move backwards?

One could postulate that a replication forks meet at a Ter site and then move in concert towards the *dif* site with one fork moving backwards – depolymerizing its nascent DNA strand as it moved away from the Ter site – until they reached the *dif* site. If so, then one would observe the mutational bias patterns we report. While this model does preserve the action of Ter sites, it still requires forks stop at the *dif* site. Therefore, this model reduces to the proposal that forks ultimately halt at the *dif* site. In addition, this model requires that DNA depolymerization occurs for a very large distance, especially if proceeding from origin-proximal Ter sites (*e.g.*, TerH or TerI).

3.6 CONCLUSIONS

The bioinformatically-defined replication terminus lies very near the molecularly defined *dif* site in members of the γ-Proteobacteria, the Firmicutes and likely the Actinobacteria. The existence of this clear, unique site for change in strand bias – and the lack of change in strand bias across Ter sites – was not predicted by previous models of replication termination invoking dispersed Ter sites engaged in polar replication arrest. We propose that the Ter sites act primarily to halt replication forks arising from DNA repair processes. In addition, our results suggest a more central role for the *dif* region in integrating chromosome mobilization, recombination and replication termination. Given the critical and intertwined roles of replication termination and DNA segregation in the prokaryotic life cycle, this scenario is not surprising. In bacteria, then, no success in life can compensate for failure at the *dif* site.

4.0 SELECTION FOR CHROMOSOME ARCHITECTURE IN BACTERIA

"Evolutionary speculation constitutes a kind of metascience, which has the same intellectual fascination for some biologists that metaphysical speculation possessed for some mediaeval scholastics. It can be considered a relatively harmless habit, like eating peanuts, unless it assumes the form of an obsession; then it becomes a vice." ~Roger Stanier (1970)

4.1 ABSTRACT

Bacterial chromosomes are immense polymers whose faithful replication and segregation is crucial to cell survival. The ability of proteins such as FtsK to move unidirectionally towards the replication terminus, and direct DNA translocation into the appropriate daughter cell during cell division, requires that bacterial genomes maintain an architecture for the orderly replication and segregation of chromosomes. We suggest that proteins that locate the replication terminus exploit strand-biased sequences that are overrepresented on one DNA strand, and that selection increases with decreased distance to the replication terminus. We report a generalized method for detecting these architecture imparting sequences (AIMS), and have identified AIMS in nearly all bacterial genomes. Their increased abundance on leading strands, and decreased abundance on lagging strands, towards replication termini are not the result of changes in mutational bias; rather, this reflects a gradient of long-term positive selection for AIMS. The maintenance of the

pattern of AIMS across the genomes of related bacteria independent of their positions within individual genes suggests a well-conserved role in genome biology. The stable gradient of AIMS abundance from replication origin to terminus suggests that the replicore acts as a target of selection, where selection for chromosome architecture results in the maintenance of gene order and in the lack of high-frequency DNA inversion within replicores.

4.2 INTRODUCTION

Bacterial chromosomes are not simply collections of genes; these polymers – up to 100,000-times longer than the cells that contain them – are organized into highly compacted nucleoids (HOLMES and COZZARELLI 2000; WU 2004) as super-coiled domains (DENG *et al.* 2004; HIGGINS *et al.* 1996; STEIN *et al.* 2005), are positioned at defined locations within the cytoplasm (GITAI *et al.* 2005; NIKI *et al.* 2000; TELEMAN *et al.* 1998; WU and ERRINGTON 1998) experience intricately-timed replication (CUNNINGHAM and BERGER 2005) and move through the cytoplasm in precise, choreographed ways (VIOLLIER and SHAPIRO 2004; VIOLLIER *et al.* 2004). Beyond encoding thousands of protein and RNA products, as well as signals for their production, DNA molecules must also carry information that controls the tempo and mode of their own replication and segregation into daughter cells. While numerous genetic, molecular biological and bioinformatic techniques serve to identify DNA sequences that are important because of the products they encode (that is, genes), finding sequences that are important for the maintenance of the DNA molecule itself has proven to be more difficult.

Global chromosome structure is suggested by the non-random distribution of genes within replicores. Single replication origins and termini typically apportion bacterial genes nearly symmetrically into two approximately equally-sized replicores. The locations of some genes relative to the replication origin is known to be important – *e.g.,* the proximity of the *Bacillus subtilus spoIIR* gene to the replication origin allows its transcription from the newly-formed forespore, whereas the origin-distal location of the *spoIIAB* gene prevents its encapsulation in the forespore, allowing for σ^F activation there (DWORKIN and LOSICK 2001).

Furthermore, *dnaA* genes are significantly associated with the replication origins. Outside of such special cases, little significance to the positions of other transcription units relative to the replication origin has been postulated beyond the potential for greater gene dosage of origin-proximal genes (LIU and SANDERSON 1995; LIU and SANDERSON 1996). Yet we can infer that gene order is constrained, since genetic maps retain order in the face of mechanisms that can rearrange them. More importantly, observed rearrangements are most often symmetrical with respect to replication origins and termini (EISEN *et al.* 2000; MACKIEWICZ *et al.* 2001; SANDERSON and LIU 1998; SUYAMA and BORK 2001; TILLIER and COLLINS 2000), suggesting that inversions that rearrange chromosome structure (*i.e.,* those that move genes from leading to lagging strands) are counter-selected.

Beyond the distribution of genes, the non-random distribution of certain oligomeric sequences is also consistent with global chromosome structure. One example is the χ recombination signal (EGGLESTON and WEST 1997; KOWALCZYKOWSKI *et al.* 1994; KUZMINOV 1995; MYERS and STAHL 1994); this octamer is highly abundant on leading strands (EL KAROUI *et al.* 1999; UNO *et al.* 2000), and serves to disable the RecD exonuclease, allowing the RecBC recombinase to repair double-stranded breaks efficiently via homologous recombination. The overabundance of χ sequences is consistent with their origin by mutational biases and maintenance by selection for function. That is, the signals that mediate global chromosome architecture could arise by mutational bias, where consistent replication from a single origin allows for differences to accumulate between leading and lagging strands (LOBRY 1996; LOBRY and LOUARN 2003; SALZBERG *et al.* 1998). Once placed under selection, differences between strands that arise by chance would be maintained, and disruption of these patterns would be detrimental (CAPIAUX *et al.* 2001; CORRE *et al.* 2000).That is, for the overabundance of χ sequences on leading strands to be identifiable, global chromosome structure must exist.

The processes described above – replication termination and DNA segregation – involve the action of proteins at the replication terminus. Therefore, sequences enabling proteins to locate the replication terminus are good candidates for those contributing to chromosome architecture. One may expect these sequences to accumulate near the replication terminus since it is there that selection for their function would be greatest. For example, the FtsK protein translocates along DNA towards the *dif* site at the replication terminus (PEASE *et al.* 2005), and may mediate segregation of chromosomes across the septum (LAU *et al.* 2003). FtsK delivers the XerCD recombinase to the *dif* site (BIGOT *et al.* 2004; IP *et al.* 2003; LI *et al.* 2003; MASSEY *et al.* 2004), where it acts to resolve entangled chromosomes during cell division (BLAKELY *et al.* 1991; CLERGET 1991). The FtsK protein must recognize strand-specific sequences to enable its directional movement towards the replication terminus. Since the frequency at which DNA translocases act is inversely proportional to the distance from the replication terminus, sequences would be under strongest selection – and therefore at highest abundance on their preferred strand – near the terminus. This increase in abundance towards the replication terminus – beyond what would be predicted by changes in mutational bias (DAUBIN and PERRIÈRE 2003) – can be taken as evidence for selection. Here, we describe methods for detecting such sequences and demonstrate that their distributions did not result from mutational biases or chance.

4.3 MATERIALS AND METHODS

4.3.1 Sequence analysis

Sequences were downloaded from GenBank and analyzed using DNA Master (cobamide2.bio.pitt.edu). Nucleotide skew was calculated as (G-C)/(G+C) or (A-T)/(A+T) at the third codon positions of protein-coding genes, corrected for the direction of transcription. Global pair-wise sequence alignments used the method of Needleman and Wunsch (1970); alignment scores were obtained using the PAM 250 matrix (ALTSCHUL 1991), normalized to the average length of the genes being compared. Octamers were classified as matching IUB nondegenerate (GATC) and degenerate (RYMK) bases. Watson strands are defined as the DNA strands – read 5' to 3' – reported in GenBank files; Crick strands are defined as the complements of Watson strands. Leading strands are defined as Watson strands downstream, and Crick strands upstream, of the replication origin. Skewed octamers were detected as those sequences overrepresented on leading strands; asymmetrically-distributed octamers were detected as sequences present in a particular region of a replicore at significantly higher abundance than predicted from their abundance in the remainder of the replicore as measured by χ^2 analysis.

4.3.2 Number of sequences defining the replication origin or terminus

Leading strands correspond to the Watson strands on one side of the replication origin or terminus, and to Crick strands on the other side. To locate these positions, a sliding-window analysis was performed, where windows were defined as encompassing 80% of a bacterial genome sequence, centered on a potential ‘break point.’ Strand-biased octamers defining a break point were enumerated as those that were overrepresented on the Watson strand upstream of the break point, but overrepresented on the Crick strand downstream of the break point. The numbers of biased octamers would be maximal when the break points lie close to either the replication origin or terminus, where Watson strands change from leading strands to lagging strands.

4.3.3 Detection of large inversions and insertions

Bacterial genomes were divided into segments (typically 10 – 100 kb in length) that were analyzed independently for strand-biased octamers. The libraries of strand-biased octamers generated for each genome segment were compared to each other. If similar sequences were biased on the Watson strands of both segments, these regions were viewed as being historically replicated in the same direction; if similar sequences were biased on the Watson strand of one segment and the Crick strand of another, these regions were viewed as being historically replicated in opposite directions. Pairwise similarity of octamer libraries was calculated as the Jaccard coefficient of similarity, S_J (JACCARD 1912). Most genomes could be described as having two large domains, where the Watson strands of segments in one domain were biased in a way similar to the Crick strands of segments in the other domain. Large inversions that did not include the origin or replication terminus were detected as regions where the strand bias of the Crick strand resembled the strand bias of the Watson strand of neighboring segments. Large insertions of foreign DNA – whose strand bias would be different from the remainder of the genome – were detected as regions where the libraries of strand-biased octamers resembled neither the Watson strand nor the Crick strand libraries of any chromosome segment. This pattern would also be reflected in old inversions that had begun to ameliorate their nucleotide composition (LAWRENCE and OCHMAN 1997).

4.3.4 Sequences accumulating near the replication terminus

Octamers that accumulated in abundance towards the replication terminus were initially detected as those that (a) exceeded 100 copies per genome, typically numbering at least one sequence per 10 kb of genomic sequence, (b) were over-represented on the leading strand, where typically >70% of the sequences were found on the leading strand, (c) showed abundance in a terminus-proximal window – typically defined as 10 to 25% of the genome length – that exceeded that predicted based on its abundance elsewhere, and (d) showed this pattern on both replicores. Consistent increase in abundance towards the replication terminus was verified by regression of local octamer abundance against distance from the terminus.

4.3.5 Correction for mutational bias

The abundances of nucleotides, dinucleotides, trinucleotides and tetranucleotides were calculated by sliding window analysis. The expected local abundance of octamers was calculated from the relative abundance of constituent nucleotides, dinucleotides, trinucleotides or tetranucleotides. For *n*-mers of length *j*, where j < 8, the expected frequency of an octamer E_j given the abundance of constituent *j*-mers is defined as

$$E_j = \frac{\prod_{i=1}^{9-j} O_i^j}{\prod_{i=2}^{9-j} O_i^{j-1}}$$

where O_i^j is the frequency of the sub-oligomer of length *j* at position *i* within the octamer. Therefore, the Expected frequency (*E*) of an octamer based on the constituent tetramers is calculated as $E_{ABCDEFGH}= (P_{ABCD}P_{BCDE}P_{CDEF}P_{DEFG}P_{EFGH})/(P_{BCD}P_{CDE}P_{DEF}P_{EFG})$.

4.3.6 Location of maximum octamer abundance

Sequences that accumulate towards the replication terminus were initially identified via their over-abundance in the region adjacent to the terminus. Linear regression was then used to determine if their abundance increased towards the replication terminus. To identify sequences which may accumulate to other, non-terminus locations within the chromosome (either by chance or by selection), we found those that were over-abundant in sequence windows away from the terminus. A quadratic regression of the local abundance of octamers against distance of the region from the terminus was then performed. Sequences that reached maximal abundance at a position away from the terminus would show a local maximum (the peak of the parabola) away from the terminus.

4.4 RESULTS

4.4.1 Identification of replication origins and termini

To identify sequences imparting chromosome architecture, replication origins and termini must be identified in a robust fashion that is consistent across genomes. Consistent replication initiation and termination at defined points results in strand biases due to the mutational differences between leading and lagging strands. Replication origins and termini can be detected as points of inflection in cumulative nucleotide skew plots (LOBRY 1996), where a single strand of DNA is synthesized as a lagging strand upstream, and as a leading strand downstream, of the replication origin. For example, replicore transitions can be identified in the *Rhodopseudomonas palustris* genome at ~ 500 kb and ~3100 kb as seen in plots of cumulative GC and AT skew (Fig. 11A). To increase the precision of our assignment, these positions were refined to within ~5 kb by the identification of highly skewed octamers that were tabulated based on crude localization of the origin and terminus (Fig. 11B). In the absence of a sufficiently strong single nucleotide bias to make an initial assignment, the change in octamer abundance alone was used to identify the replication origin and terminus by a sliding window analysis (Fig. 11D). Here, replication origins and termini were identified as those locations maximizing the numbers of octamers that were overrepresented on the Watson strands upstream – and on the Crick strands downstream – of a particular location.

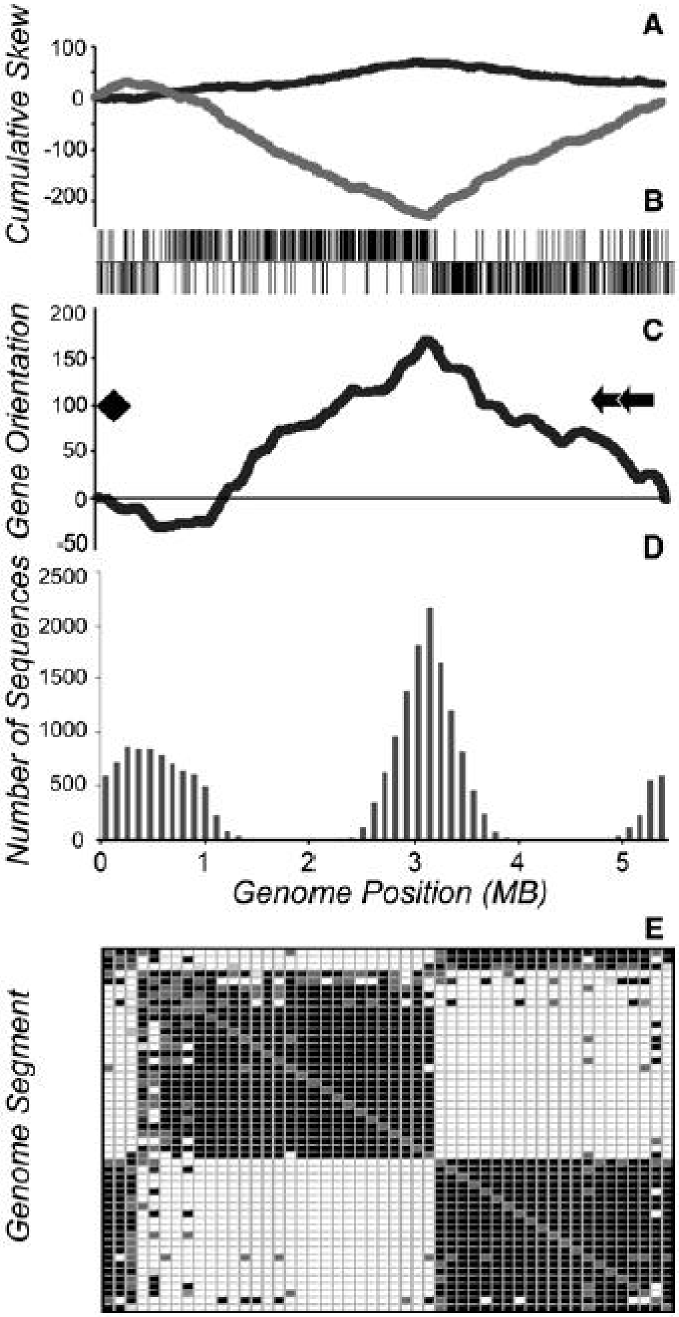

Figure 19 Finding the origin and teminus of replication *Rhodopseudomonas palustris.*

4.4.1.1 Figure 19 legend

Establishing the locations of the origin and terminus in completely sequenced bacteria: *Rhodopseudomonas palustris.* **A**. Cumulative third-codon-position nucleotide skew. **B**. Positions of five octamers (GAGGAGAG, GAGGAGGG, GAGGGGAG, GAGGGGGG and GGCGAGGG) are represented as vertical lines on either the Watson (W) or Crick (C) strand. **C**. Cumulative average gene orientation for a 100-gene sliding window, where values are calculated as the proportion of genes transcribed from the Watson strand. The diamond indicates the approximate location of the *dnaA* gene; arrows indicate the location and orientation of the rDNA cistrons. **D**. Break-point permutation analysis; the numbers of octamers overabundant on the Watson strand upstream of the break-point which are also overabundant on the Crick strand downstream of the break-point. **E**. Segmental analysis. Black squares denote regions where libraries of Watson strand-biased oligomers are congruent (see Methods), while white squares denote regions where libraries of Watson-strand-biased oligomers of one segment resemble Crick-strand-biased oligomer libraries of the other segment. Grey squares denote regions with equivocal data.

Examination of nucleotide skew alone does not identify which inflection point corresponds to the replication origin and which to the terminus. While the *dnaA* gene is often encoded near the replication origin and rRNA cistrons are often encoded on the leading strands (Fig. 19C), these are not rigorous criteria for localizing origins and termini. To augment these data, we examined gene orientation. Genes are preferentially encoded on the leading strand, perhaps to avoid polymerase collisions at genes under strong selection (ROCHA 2004; ROCHA and DANCHIN 2003; ROCHA and DANCHIN 2003). Although cumulative gene orientation bias is too crude to identify the replication origin precisely, it may be used to assign the origin and terminus to inflection points identified by mutational bias analysis (Fig. 19B). While more precise localization of the replication origin can be achieved by located *dnaA* boxes (MACKIEWICZ *et al.* 2004), our estimates were sufficiently accurate to enable the identification of strand-biased oligomers.

Single replication origins and termini were established in all large (>1000 kb) Bacterial genomes examined, indicating that mutational biases between leading and lagging strands are universal features of bacterial genomes. In most cases, the longest replicore represented between 50% and 55% of the chromosome length (Table 3), suggesting that selection operates to maintain replicores of approximately equal lengths. The positions of the *dnaA* genes were often, but not always, near the replication origin, and virtually all rRNA cistrons were replicated away from the origin.

Table 3 AIMS found in completely sequenced bacterial genomes.

Genome	Family	Size[1]	%GC	Origin	Terminus	Number Skewed[2]	Representative AIMS[3]
Mycobacterium tuberculosis	Actinobacteria	4412	65.60%	1	2232	1 (63)	CGGGGGAG, GGGGGAGC, TGGGGGAG
Nocardia farcinica	Actinobacteria	6021	70.80%	1	3137	78	CGGGGGAG, GAGGGGGA, GTGGGGGA, GCGGGGGA
Streptomyces coelicolor	Actinobacteria	8668	72.10%	4270	8667	45	TGGGGGAG
Symbiobacterium thermophilum	Actinobacteria	3566	68.70%	1	1957	517	GGGAGCTG, GGGGAGGA, TGGAGCGG, TGGTGGAG
Bacteroides thetaiotaomicron	Bacteroidetes/Chlorobi	6260	42.80%	4076	1212	32	**NF**
Chlorobium tepidum	Bacteroidetes/Chlorobi	2155	56.60%	3	1021	5 (380)	GGGGATGG, GGGGAGT, CAGGGGAK
Chlamydophila pneumoniae	Chlamydiae/Verrucomicrobia	1230	40.60%	842	213	568	GAGTTTTA, TAGGGGAA, TTAGGGGA
Parachlamydia sp.	Chlamydiae/Verrucomicrobia	2414	34.70%	1	1101	7	AAGGGGAG
Dehalococcoides ethenogenes	Chloroflexi	1470	48.90%	1	815	43	NF
Prochlorococcus marinus	Cyanobacteria	2411	50.70%	1	426	1356	TGGCTTTG
Deinococcus radiodurans	*Deinococcus-Thermus*	2649	67.00%	22	1362	7	AGGGGAGA
Bacillus subtilis	Firmicutes	4215	43.50%	1	1957	35	AAGAAGGG, GAAAAGGG, GAAGGGGA, GAGAAGGG
Clostridium acetobutylicum	Firmicutes	3941	30.90%	1	1982	39916	AAGAAGAT, GATGAGAT, ATAGATGA, GAAATGAA
Enterococcus faecalis	Firmicutes	3218	37.50%	1	1562	5685	TAGGGGATG, AGAGATGA
Lactococcus lactis	Firmicutes	2366	35.30%	1	1265	4082	AAGAAGAT,GAATTAGA, TGGAGAAA, TGGAGGAA
Oceanobacillus iheyensis	Firmicutes	3631	35.70%	1	1772	688	TAGAAGAG, AAAGGGAG, AAGGGAAA
Staphylococcus aureus	Firmicutes	2820	32.80%	1	1409	10536	AAGAACAA, AGAACAAG, GAAGATGA, ATGAAGAA
Fusobacterium nucleatum	Fusobacteria	2175	27.20%	642	1866	0	**NF**
Rhodopirellula baltica	Planctomycetes	7146	55.40%	5447	1859	0 (10)	**NF**
Agrobacterium tumefaciens cI	α-Proteobacteria	2841	59.40%	1	1479	99	AGGGCAGG, CGGGCAGG, GGGCAGGG,
Agrobacterium tumefaciens cII	α-Proteobacteria	2076	59.30%	1022	2075	33	GGGCAGGT, AGGGCAGG
Bradyrhizobium japonicum	α-Proteobacteria	9106	64.10%	617	4996	30	GGGCAGGG, GGGCAGGT, AGGGCAGG, GAGCAGGG
Brucella melitensis cI	α-Proteobacteria	2117	57.20%	1	956	128	AGGGCAGG, GGGCAGGG, GGGGCAGG
Brucella melitensis cII	α-Proteobacteria	1178	57.20%	94	758	69	GGCGAGGG, GGGCAGGG, GGTGAGGG
Mesorhizobium loti	α-Proteobacteria	7036	62.70%	3632	301	21	GGGCAGGG, GGCGAGGG, GGGAAGGG
Rhodopseudomas palustris	α-Proteobacteria	5459	65.00%	470	3156	74	AGGGCAGG, CGGGCAGG, GGGCAGGG, GAGCAGGG
Sinorhizobium meliloti	α-Proteobacteria	3654	62.70%	1	1726	31	GGGCAGGG, GAGCAGGG, AGGGCAGG

Sinorhizobium meliloti pSymA	α-Proteobacteria	1354	60.40%	1	654	0 (22)	**NF**
Sinorhizobium meliloti pSymB	α-Proteobacteria	1683	62.40%	57	1095	4	GGGCAGGG
Rickettsia conorii	α-Proteobacteria	1269	32.40%	1	697	1	AGAGCAGG, AGGGCAGG
Bordetella bronchiseptica	β-Proteobacteria	5339	68.10%	1	2957	431	GGGCAGGG, GGCAGGGC, GGCGGGGC
Bordetella parapertussis	β-Proteobacteria	4774	68.10%	1	2904	445	GGCAGGGC, GGCGGGGC
Escherichia coli	γ-Proteobacteria	4639	50.80%	3923	1589	36	AGAAGGGC, GGCAGGGC, GGGCAGGG
Haemophilus influenzae	γ-Proteobacteria	1830	38.20%	503	1471	5	**NF**
Pasteurella multocida	γ-Proteobacteria	2257	40.40%	1563	737	1	AGTATGTA
Salmonella typhimurium	γ-Proteobacteria	4857	52.20%	4084	1612	5	GGGAAGGG, GGGCAGGG, GGGGAAGG
Pseudomonas aeruginosa	γ-Proteobacteria	6264	66.60%	1	2445	85	AGGAGGGC, GGGCAGGG, GAGCAGGG, GAGGAGGG
Xanthomonas axonopodis	γ-Proteobacteria	5176	64.80%	1	2487	60	GGGCAGGG, GGGGCAGG, GGGTAGGG, GGGGCGGG
Geobacter sulfurreducens	δ-Proteobacteria	3814	60.90%	1	1892	2	GGGGAGGG, GGGTAGGG
Campylobacter jejuni	ε-Proteobacteria	1641	30.00%	1	777	180	TTAAGTGG, TTTGGGTG
Helicobacter pylori	ε-Proteobacteria	1644	39.20%	1643	685	12	AGTAGGGG
Borrelia burgdorferi	Spirochetes	911	28.60%	456	911	42076	TTTAGTTT
Leptospira interrogans	Spirochetes	4332	35.00%	1	2231	0	**NF**
Thermotoga maritima	Thermotogae	1861	46.20%	1086	156	0	**NF**

1) The genome size, replication origin and replication terminus are reported in kilobases or kilobases from the first base of the sequence, except that a value of '1' under 'Origin' denotes base 1 of the sequence.

2) Number of sequences with up to 2 degenerate bases, present at an abundance of 0.1/kb (0.05/kb), where 75% of the sequences were located on the leading strand.

3) Sequences were initially identified as those at least 1.4-fold more abundant in the terminus-proximal 10% of each replicore than expect from the origin-proximal 75% of each replicore. Increase in abundance towards the replication terminus were verified by linear regression of local abundance against genome position. Representative non-degenerate sequences are shown.

In some cases, more than two major inflection points in cumulative nucleotide skew plots were observed; for example, the *Pasteurella multocida* genome has six major inflection points (LAWRENCE and HENDRICKSON 2004). Such patterns could result from inversions, the insertion of foreign DNA with similar strand biases, or the presence of multiple replication origins and termini. In all such cases in Bacteria, we inferred that inversions within replicores had produced regions of the genome with nucleotide skew in the 'opposite' direction, because (a) the multiple regions did not reflect more than two symmetrical replicores as was the case in other bacterial genomes and Archaeal genomes with likely multiple replication origins (ZHANG and ZHANG 2003; ZHANG and ZHANG 2005), and (b) apparent large inversions were common in pathogens with reduced genome sizes where comparisons with the chromosomes of less-virulent relatives could delineate the extent of the inverted DNA (LIU and SANDERSON 1995; LIU and SANDERSON 1995; LIU and SANDERSON 1996; PARKHILL *et al.* 2003; READ *et al.* 2000; SUYAMA and BORK 2001).

4.4.2 Identification of large inversions and insertions

As noted above, the replication history of a DNA segment is reflected in its accumulation of strand bias. Therefore, large insertions and inversions that do not include the replication origin or terminus can be detected by their perturbation of nucleotide-skew and octamer-skew patterns. To identify these regions, a segmental analysis was performed, whereby the local strand biases of individual segments were assessed and compared (Fig. 19E). Here, regions of the chromosome that have historically been replicated in the same direction will have the same sets of octamers biased on their Watson strands. In contrast, large inversions could be identified as regions within well-defined replicores wherein octamers overrepresented on Watson strands were overrepresented on the Crick strands of neighboring segments (LAWRENCE and HENDRICKSON 2004). Older large inversions can be recognized as regions where the octamer strand bias is not congruent with either adjacent leading or lagging strands through the process of amelioration; large (> 25 kilobase) insertions will also give this appearance.

Recent large inversions within replicores are typically not evident in bacterial genomes. That is, most genomes showed two large replicores with consistent nucleotide and octamer skew. Exceptions fell into two classes. First, genomes of obligate endosymbionts and intracellular pathogens – typically less than 1000 kilobases in length – often showed signs of large-scale chromosome rearrangements; examples include the genomes of *Buchnera*, *Wolbachia, Mycoplasma pulmonis*, *M. genitalium* and *Ureaplasma urealyticum*. In most cases, chromosomes were sufficiently fragmented to preclude accurate identification of replication origins and termini. Second, pathogens with large genomes also showed rearrangements when compared to less virulent relatives with similarly-sized genomes. For example, *Salmonella enterica* serovar Typhi shows substantial rearrangement relative to less virulent salmonellae (LIU and SANDERSON 1995; LIU and SANDERSON 1996), and *Bordetella pertussis* is rearranged relative to *B. bronchiseptica* (PARKHILL *et al.* 2003). We also detected inversions in *E. coli* (~650-740 kb), *Fusobacterium nucleatum* (~530–650 kb)*, Helicobacter pylori* (many), *Pasteurella multocida* (~1480–1560 kb, and ~1880–1960 kb) and inversions shared between *Rickettsia prowezeckii* and *R. connori* (~360–400 kb, and ~1560–1600 kb).

4.4.3 Identification of sequences under selection

Chromosomes lacking large inversions were examined for octamers that increased in abundance towards the replication terminus only on leading strands. First, sequences that were overrepresented on leading strands on both replicores were identified; this bias in distribution could result solely from the mutational bias inherent in DNA replication and therefore does not in itself suggest that these sequences are under selection. Table 3 reports the number of abundant octamers (found at a frequency of more than 0.1 per kilobase) which showed strong strand bias (more than 75% – a 3:1 bias – were located on the leading strand). With these stringent criteria, between 0 and 42000 sequences were identified in 40 bacterial genomes examined. In genomes that lacked highly abundant oligomers that were skewed to this degree, we identified skewed sequences that were found at least once per 20 kilobases (Table 3).

Within these sets, we identified sequences under selection as those that increased in abundance on the leading strand towards the replication terminus. These sequences were initially identified as those that were overrepresented on leading strands in the terminus-proximal regions of each replicore. To eliminate sequences which were serendipitously overabundant in these regions – for example, if they were highly abundant in genomic islands integrated near the terminus region – the local abundance of each octamer was calculated for intervals spanning from the replication origin to the terminus. Sequences under selection were identified as those where the slope of the linear regression of abundance vs. position was significantly different from zero (Fig. 20). In most cases, the sequence also significantly decreased in frequency on the lagging strands, thus leading to greater strand bias near the terminus. In other cases, the abundance was extremely low on the lagging strand, precluding accurate assessment of changes in abundance on this strand.

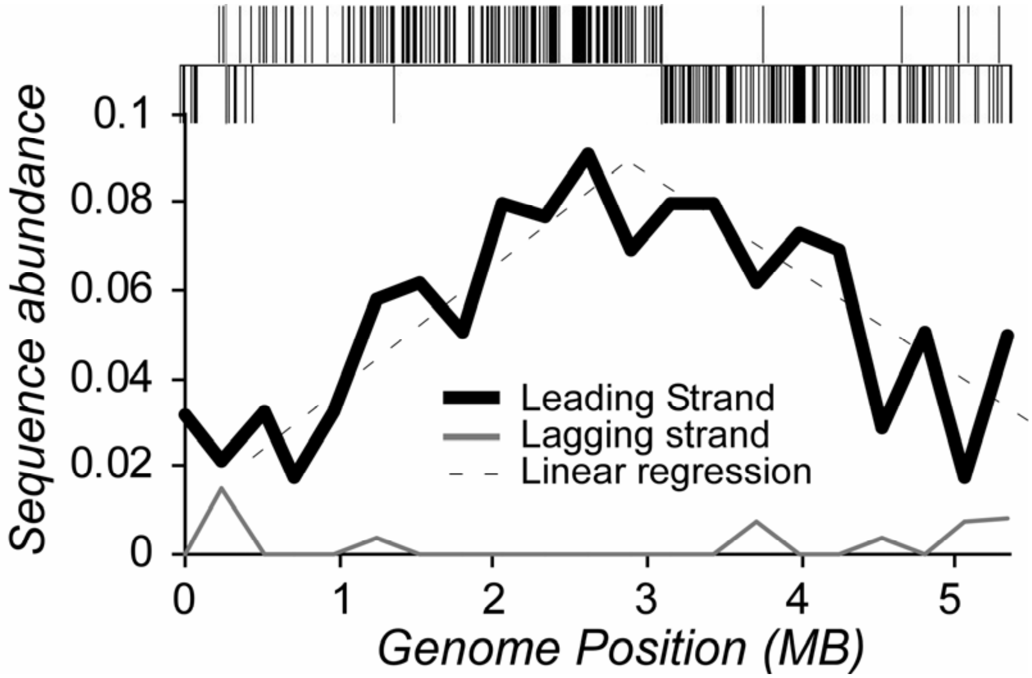

Figure 20 AIMS in *Rhodopseudomonas palustris.*

4.4.3.1 Figure 20 legend

This AIMS is present with 300 copies on the leading strand and 12 copies on the lagging strand. Sequence abundance is reported as number of AIMS per 50 kilobases within the ~290 kb window.

Table 3 shows examples of skewed sequences increasing towards the replication termini in bacterial genomes. For example, there are 312 copies of the GGGCAGGG octamer in the *R. palustris* genome; 96% of the occurrences are on the leading strand, and twice as many copies are found in the terminus-proximal region of both replicores than would be expected if sequences were distributed randomly (Fig 20). This sequence was found to increase towards the replication terminus in many genomes of proteobacteria (Table 3). We propose that skewed octamers increasing in abundance towards the replication terminus are under selection for maintenance of chromosome structure. Therefore, we term these octamers Architecture Imparting Sequences, or AIMS, to denote their potential involvement in one or more biological processes that use origin to terminus polarity. While the role of each AIMS in cell biology is unknown, it is clear that the distribution of AIMS represents selection operating above the level of the gene and that this selection structures – *i.e.*, provides an architecture to – bacterial chromosomes.

Table 3 presents only a sample of potential AIMS, not a definitive list of all sequences under selection for function. There are many sequences which are less numerous, less strand-biased, or which show a more modest increase in abundance towards the replication terminus which were excluded from this analysis. That is, we chose threshold values so that sequences that met our criteria could not have arisen by chance alone (see below). In genomes where no sequences were found to pass these criteria, we could identify sequences that increased in abundance towards the replication terminus that were less abundant, less strand-biased, or increased in abundance towards the terminus to a lesser degree. However, this set of sequences includes those whose distributions resulted by chance, thus potentially confounding conclusions drawn from their distributions.

AIMS were found in genomes of bacteria representing every major division, including multiple representatives of Actinobacteria, Chlamydiae, Cyanobacteria, Firmicutes, Proteobacteria and Spirochetes (Table 3). AIMS were easily identified in genomes of small size (*e.g.,* the TTTAGTTT octamer in the *Borrelia burgdorferi* genome, 911 kb) and large size (*e.g.,* the AGGAGGGC octamer in the *Pseudomonas aeruginosa* genome, 6264 kb). We could identify AIMS in genomes with high GC content (*e.g.,*TGGGGGAG in *Streptomyces coelicolor,* 72.1 % GC), high AT content (e.g., AAGAAGAT in *Clostridium acetobutylicum,* 30.9 % GC) or neutral composition (*e.g.,* TGGCTTTG in *Prochlorococcus marinus,* 50.7 % GC). AIMS were often GC-rich, even in genomes with high AT-content (*e.g.,* TAGGGGATG in *Enterococcus faecalis*, 37.5% GC). AIMS were also found in organisms with linear replicons (*e.g., Streptomyces, Borrelia, and Agrobacterium*), suggesting that functions utilizing at least some of the AIMS are required for replication and segregation of linear chromosomes. For example, such functions may include DNA translocation across the division septum.

In three instances, multiple, large replicons are found in the same organism: *Brucella melitensis, Agrobacterium tumefaciens* and *Sinorhizobium meliloti*. In *Brucella* and *Agrobacterium*, the AIMS identified from one large replicon also appeared to be skewed and increasing in abundance in the other replicon (Table 3); some sequences were less abundant on one replicon, and therefore were not reported in Table 3. This suggests that they are under selection in both replicons in each organism. *Sinorhizobium* has three replicons, including the large plasmids pSymA (1354 kb) and pSymB (1683 kb). AIMS found in the *Sinorhizobium* chromosome (*i.e.,* the largest replicons) were also AIMS in the pSymB sequence. While these sequences are not AIMS in the pSymA plasmid, they are skewed to leading strands. Since both plasmids harbor *repABC* partitioning operons near their respective replication origins, AIMS may not play as large a role in their maintenance.

As a rule, we did not identify AIMS in genomes less than 1000 kb in size; most genomes of this size class are found in obligate pathogens and intracellular parasites (*e.g., Buchnera, Mycoplasma*). We do not interpret this result as a lack of selection for polarity elements in these taxa. Rather, extensive chromosome rearrangements experienced by genomes of pathogens (MIRA *et al.* 2001), coupled with the very small size of these genomes limits the ability to find distributions that are statistically significant. Also, as noted above, replication origins and termini could not be located with confidence in these genomes. As a result, we were not confident of the sequence distributions we could infer.

4.4.4 AIMS do not arise from changes in mutational bias

The underlying mutational biases vary along the chromosome (DAUBIN and PERRIÈRE 2003); that is, GC-skew at third codon positions differs between genes that are origin-proximal relative to those that are terminus-proximal. Therefore, one could infer that some octamers may increase in abundance towards the terminus strictly due to changes in mutational bias alone. To correct for changes in mutational bias in the terminus-proximal region, we quantitated the changes in the nucleotide, dinucleotide, trinucleotide and tetranucleotide frequencies from the origin to the terminus. For octamers under selection, their accumulation near the replication terminus cannot be explained by underlying changes in the distribution of nucleotides, dinucleotides, trinucleotides or tetranucleotides. For example, Figs. 21A and B show the abundance of the GGGCAGGG octamer in the *Rhodopseudomonas palustris* genome; towards the replication terminus, it clearly increases in abundance on the leading strand and decreases in abundance on the lagging strand. Yet the predicted abundance of this octamer – as inferred from the abundance of its constituent dinucleotides, trinucleotides and tetranucleotides – does not change appreciably. If any, predicted abundances decrease towards the terminus on the leading strand and increase on the lagging strand. These data suggest that a simple change in mutational bias from the replication origin to terminus is not responsible for the distribution of the GGGCAGGG octamer in the *R. palustris* genome. Similar results are seen for the GAAGGGGA octamer in the *Bacillus subtilus* genome (Fig. 21CD). We examined the distribution of all potential AIMS listed in Table 3 and conclude that changes in mutational biases alone can not explain the distribution of any octamer increasing in abundance near a replication terminus.

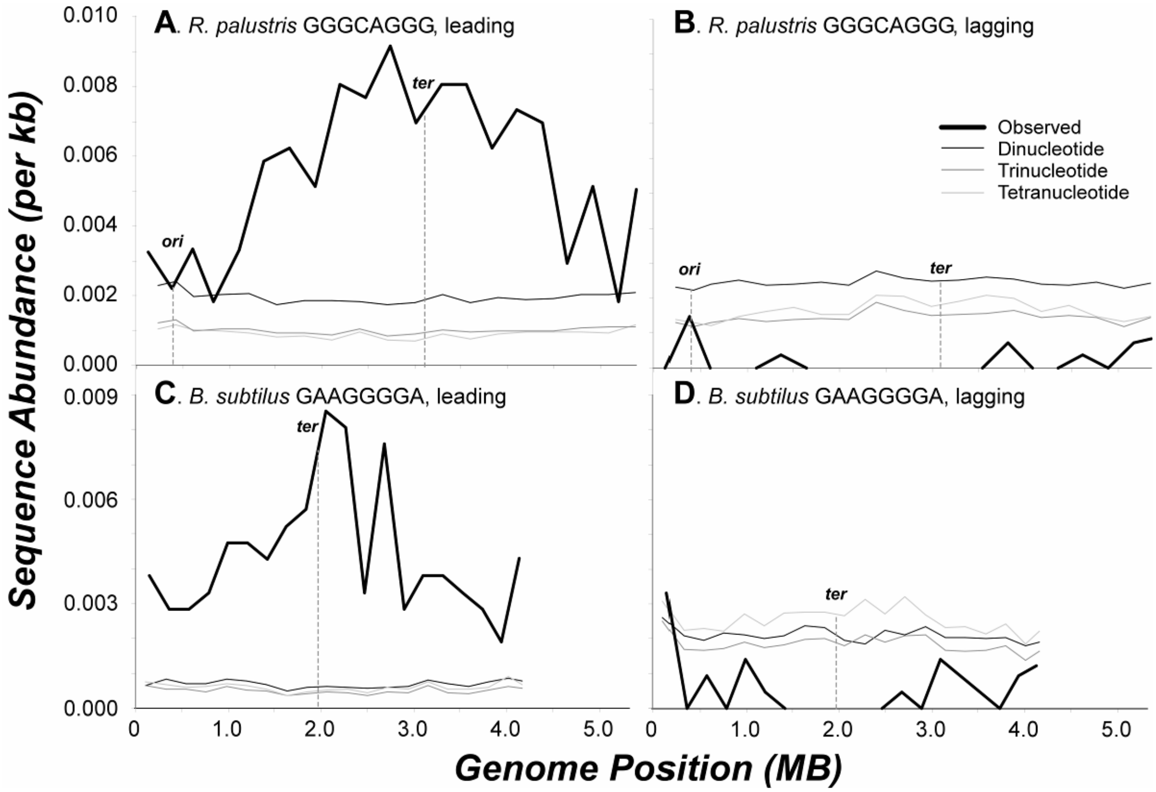

Figure 21 Actual and expected distribution of AIMS in *R. palustris* and *B. subtilis.*

4.4.4.1 Figure 21 legend

Distributions of AIMS are not explained by mutational changes from origin to terminus in chromosomes. The accumulation of the GGGCAGGG octamer on the (**A**) leading strand and (B) lagging strand in the terminus-proximal region of the *R. palustris* genome, and the accumulation of the GAAGGGGA octamer on the (C) leading strand and (D) lagging strand within the *Bacillus subtilus* genome. The observed local abundance of these sequences are shown along with the expected abundance predicted from the distributions of the 7 constituent

Figure 21 legend cont.

dinucleotides, 6 trinucleotides or 5 tetranucleotides as described in the Methods section. Sequence abundance is reported as number of AIMS (either observed or predicted) per kilobase.

4.4.5 Sequences only accumulate in abundance near the replication terminus

In identifying potential AIMS in bacterial genomes, we required both moderately high overall abundance as well as a strong increase in abundance towards the replication terminus. These criteria were established so that changes in abundance could not be attributed to chance. That is, given 16 million degenerate octamers that are examined, one would expect some to increase in abundance towards the replication terminus strictly by chance; asking for similar increases in both replicores reduces the number of false positives, but does not eliminate them.To ascertain how many sequences arise by chance that increase in abundance towards a particular location, we examined genomes for sequences which accumulated at other locations in the genome to the degree shown by AIMS. If the numbers of AIMS merely reflects chance, similar numbers of sequences should be identified that accumulate towards other locations in the genome.

As shown in Fig. 22, more sequences accumulate at the replication terminus than any other location in the genome. Moreover, those sequences appearing to reach maximum abundance outside the terminus region were found in lower copy numbers than AIMS, so their ‘accumulation’ at other chromosomal locations was interpreted as resulting from chance. That is, the numbers of sequences accumulating at a non-terminus location represented the ‘noise’ produced by examining 16 million octamers. Therefore, we interpret AIMS – that is, high-copy-number sequences that accumulate only at the replication terminus in a way unexplained by underlying mutational bias – as sequences under selection for function.

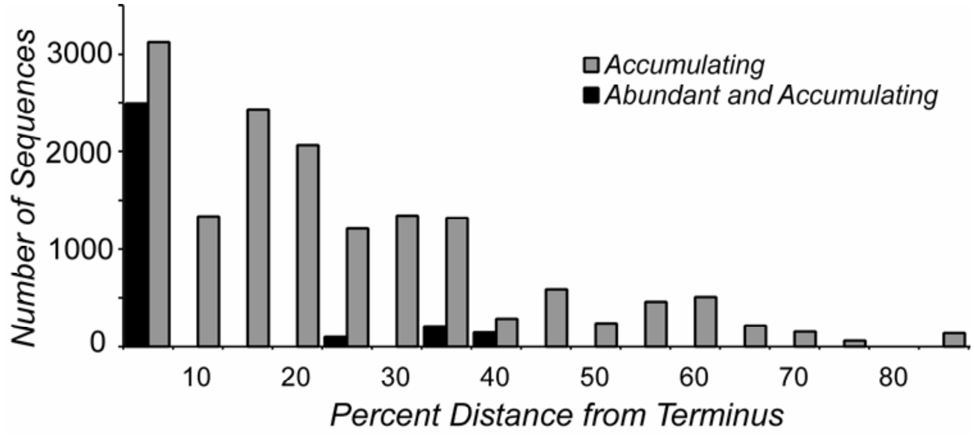

Figure 22 Sequences are abundant and accumulate only at the terminus.

4.4.5.1 Figure 22 legend

Sequences that accumulate towards a defined region with each replicore were identified; the total count of individual sequences is plotted (that is, the number of different sequences multiplied by their abundances). There are more sequences that accumulate gradually and have their highest point of abundance at the terminus than other regions of the genome. The grey bars show the numbers of sequences that are over-represented within the region specified. The black bars show the numbers of sequences that have their maximal abundance within the region specified.

4.4.6 The sequence distribution, not the individual sequences, are under selection

The accumulation of AIMS towards the replication terminus could result from the non-random distribution of genes within genomes. For example, if membrane proteins were located in the terminus proximal region, sequences encoding membrane-spanning domains may be overly abundant in this region. If this were the case, then one would expect that individual occurrences of AIMS themselves – not merely their distribution within the replicore – to be conserved among genomes of closely-related bacteria. As seen in Table 3, we have identified AIMS in several sets of closely related genomes, including the GGGCAGGG octamer in numerous α-proteobacteria.

To determine if AIMS were under selection for function in their resident proteins, we examined their locations within orthologous genes among closely-related taxa. We found that the locations of AIMS within orthologous genes were not conserved; rather, only their distribution – and increase in abundance towards the replication terminus – was shared among these genomes. For example, the distribution of the GGGCAGGG octamer increases in abundance among all α-proteobacteria examined; these genomes range in size from 3.7 to 9.1 MB (Fig. 23B). However, the precise locations of these individual sequences were not conserved among orthologous genes (Fig 23A), and the octamer was found in several reading frames, in both the template and non-template strands, and in intergenic regions. These data support the hypothesis that the distribution itself is under selection, suggesting a unit of selection at the level of the replicore, above the level of the individual gene.

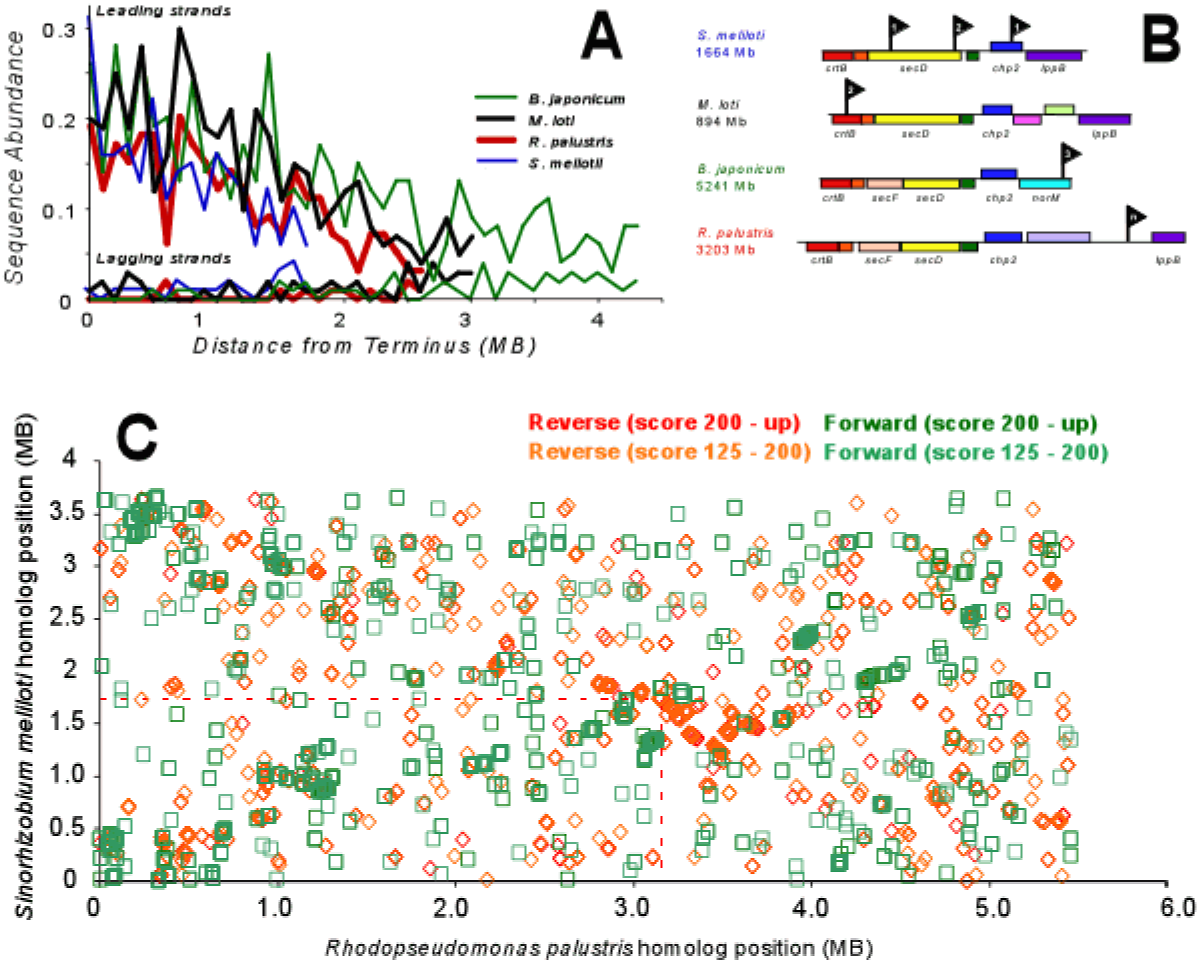

Figure 23 A) conservation of AIMS B) continued mutation & C) rearrangement.

4.4.6.1 Figure 23 legend

Among closely related bacteria it is the sequence distribution that is conserved, not the absolute positions of the sequences. **A**. The frequency of each octamer – not a cumulative frequency – within genomic regions is plotted as a function of the distance from the terminus of replication. Abundance on the two replicores is averaged. The GGGCAGGG octamer shows comparable distributions in genomes of four species of α-proteobacteria, increasing in abundance on leading strands towards the replication terminus.

(Figure 23 legend cont.)

B. AIMS within orthologous genes in the α-proteobacteria do not occur in the same positions. Arrows denote the positions of AIMS; the direction of the arrow denotes orientation. **C**. Orthologues shared between the *R. palustris* and *S. meliloti* genomes. A total of 1666 genes (50% of the *S. meliloti* gene complement) were reciprocal best matches with adjusted alignment scores of 125 or above, providing a conservative assignment of orthologues. Genes in the same orientation are shown as squares, and those in the opposite orientation as crosses.

More importantly, the cellular functions that require AIMS – candidates include the FtsK protein, which translocates to the dif site in the terminus region during cell division – appear to conserve their choice of AIMS. Table 3 shows several cases of related organisms which share AIMS, even though they share less than 90% sequence identity. For example, many α-proteobacteria share the GGGCAGGG octamer as an AIMS. As seen in Fig. 23C, AIMS may be retained even in the face of extensive chromosomal rearrangements, consistent with strong selection for AIMS.

4.5 DISCUSSION

4.5.1 AIMS are widespread among bacterial genomes

We have provided evidence that bacterial chromosomes contain sequences whose distributions suggest that they are under selection for a function unrelated to the genes in which they are found. The distributions of these sequences are consistent with their role in specifying strand identity. That is, differential abundance of sequences on leading and lagging strands can be used to locate the terminus; selection for this asymmetry will lead to increased abundance on leading strands, and decreased abundance on lagging strands, that is inversely correlated to distance from the replication terminus. While it is not clear precisely what these functions may be, their distributions are consistent with a role during DNA replication and segregation. We have termed these elements Architecture Imparting Sequences, or AIMS. It does not appear that the specific locations of AIMS with respect to genes or transcripts are under selection as is the case with transcription promoters, rho-independent transcription terminators, binding sites for regulatory proteins, translation start sites or translation stop sites. Rather, the distribution of AIMS across the replicore reflects a gradient in selection, where the entire replicore acts as a target of selection, functioning above the level of the individual gene or operon.

We have identified AIMS in nearly every bacterial genome we examined for which the identification of the replication origin and terminus was unambiguous (Table 3); the failure to identify AIMS in some genomes likely reflects the stringency of our search criteria rather than their absence from that genome. This suggests that AIMS are not under selection for a function that is found only in certain organisms, although the proteins that mediate this function may differ among organisms, leading to different AIMS being found in different genomes. For example, *ter* sites within the *E. coli* genome – bound by the Tus protein to halt retrograde replication forks – are found in the terminus-proximal region; but the *tus* gene is not found outside the proteobacteria (ANDERSEN *et al.* 2000). In other bacteria, sequences like AIMS may contribute to these functions. That is, the function is likely important to all bacteria, but particular sequences (like *ter*) will not be ubiquitous.

Inspection of Table 3 shows that genomes of closely related organisms often show similar AIMS. For example, the GGGCAGGG octamer – or some closely-allied sequence – is not only skewed in proteobacterial genomes, but is increasing in abundance on the leading strand towards the replication terminus – that is, it is an AIMS. Similarly, the AAGAAGAT octamer appears as an AIMS in the genomes of several Firmicutes, and Actinobacteria shared permutations of the YGGGGGAG octamer. As seen in Fig. 23, the common occurrence of AIMS in related genomes is not a result of the sequence being conserved within individual genes;

rather, the pattern of increasing abundance towards replication termini is shared. Moreover, AIMS are conserved in the face of extensive rearrangement of these chromosomes (Fig. 23C). The common sets of AIMS among related bacteria are consistent with shared, conserved mechanisms that maintain chromosome architecture in these organisms.

4.5.2 AIMS may represent longer, more degenerate sequences under selection for function

AIMS do not necessarily indicate the precise sequence acted upon by a molecular mechanism; rather, they are only sequences whose distributions must have arisen from selection for their overabundance near the replication terminus. The precise sequences acting as target of selection could be deduced from the library of AIMS within a genome. First, in many genomes sets of AIMS appear to represent a more degenerate sequence. For example, both permutations of the GGGMAGGG octamer are AIMS in *Mesorhizobium loti*, as well as both permutation of the GRGCAGGG octamer in *Pseudomonas aeruginosa* (Table 3). Therefore, the distributions of the non-degenerate sequences may reflect a more degenerate target of selection. Second, AIMS are detected as octamers, while either shorter or longer sequences may actually be under selection. In many genomes, there are AIMS which have overlapping sequences, such as the two octamer permutations of the AGGGCAGGG nonomer in *Sinorhizobium meliloti* and *Brucella melitensis*, or the three octamer permutations of the YGGGGGAGC nonomer in *Mycobacterium tuberculosis* (Table 3). Further inspection of these nonomers has not yielded evidence that these longer sequences might be the actual targets of selection; longer sequences do not accumulate towards the terminus to a larger degree than their constituent octamers. More thorough analyses may uncover some examples, but the low abundance of sequences longer than octamers precludes rigorous testing. More importantly, longer sequences may be insufficiently abundant to serve as polarizing elements. For example, the FtsK protein appears to recognize and reorient in response to sequence elements as frequently as once per 2 kilobases (6 times in 12 Kb) in the terminus region of the *E. coli* chromosome (PEASE *et al.* 2005), which may be accommodated by degenerate octamers, but likely not by longer sequences.

4.5.3 Selection for function does not always lead to accumulation near the replication terminus

AIMS represent one class of sequences that operates to maintain chromosome architecture; AIMS reflect selection for functions required at or near the replication terminus. Other sequences whose importance is not restricted to this region may show evidence for selection by virtue of their overrepresentation on leading strands throughout the genome. For example, the eight base pair χ sequence (GCTGGTGG) is recognized by *E. coli* RecBC helicase/exonuclease/ recombinase complex, halting the retro-translocation of Holliday Junctions at these sites and instigating resolution of recombination substrates (MYERS and STAHL 1994). It has been noted previously that χ sites are more abundant than would be expected (EL KAROUI *et al.* 1999). In the *E. coli* genome, χ sites are approximately 3.5 times more abundant across the length of the replicore than would be expected given its component tetramers (Fig. 24A). As discussed elsewhere (EL KAROUI *et al.* 1999; UNO *et al.* 2000), this increased abundance is taken as an indication of replicore-wide positive selection for function of this sequence. Here, the χ sequence prevents RecD-mediated degradation of DNA strands, allowing for rapid reestablishment of stalled replication forks. Since selection for the function of χ sequences is independent of genome position, we would not expect the abundance of χ to increase towards the terminus. Unlike AIMS, the abundance of the χ octamer does not increase towards the replication terminus. Also unlike AIMS, selection has not favored the increased abundance of the χ octamer on leading strands and decreased abundance on lagging strands, which would heighten strand bias; χ is actually somewhat more abundant on lagging strands than expected (Fig. 24A).

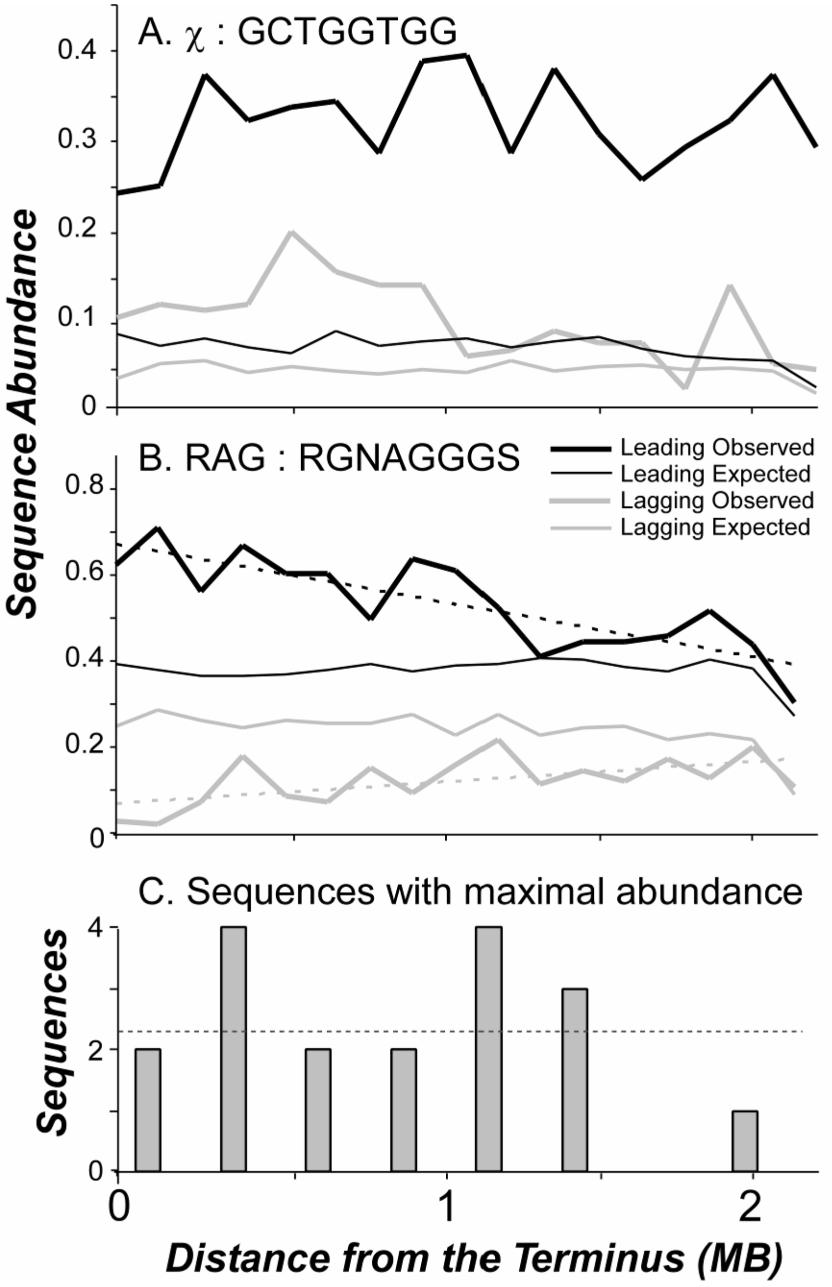

Figure 24 Distribution of octamers in the *E. coli* genome.

4.5.3.1 Figure 24 Legend (cont. from previous page)

The frequency of each octamer – not a cumulative frequency – within genomic regions is plotted as a function of the distance from the terminus if replication. Abundance on the two replicores is averaged. **A**. Abundance of the χ octamer. **B**. Abundance of the RAGS octamer. **C.** The distributions of octamers that are strand-biased to the same degree as the RAGS oligomer in the *E. coli* genome were analyzed for positions of maximal abundance; the numbers of oligomers whose distributions were maximal at 8 separate intervals (as determined by quadratic regression) are shown. The dashed line denotes the mean of 2.25 oligomers.

4.5.4 All sequences accumulating towards the terminus are not necessarily under selection

The RAG octamer (RGNAGGGS) was identified as a putative polarizing element in the *E. coli* chromosome, possibly aiding in positioning the *dif* site at the septum during cell division at the end of bacterial chromosome replication (BIGOT *et al.* 2004; CAPIAUX *et al.* 2001; CORRE and LOUARN 2002). Although the RAG octamer was postulated to act within the terminus-centered 10 kb *dif*-activity zone (CORNET *et al.* 1996; PÉRALS *et al.* 2000) or the 250 kb FtsK-zone (CORRE and LOUARN 2002; CORRE and LOUARN 2005), these boundaries reflect the resolution of the bacteriophage-excision assays used to assess the negative impact of placing AIMS in their non-permissive orientation.

In our analysis, the RAG sequence was not reported as an AIMS in the *E. coli* genome (Table 3), indicating that its distribution did not satisfy our threshold criteria. The degenerate RAG has 6 bases of information, making it a sufficiently abundant octamer to analyze, and does accumulate somewhat in abundance towards the replication terminus (Fig. 24B). What is not clear is if this increase in abundance is significant. The RAG sequence increases in abundance on leading strands towards the replication terminus 1.5-fold more than would be expected based on the distribution of underlying tetramers (Fig. 24B). Yet AIMS we identified increased to a much larger degree; for example, the degenerate AGGGCRGR octamer increased 3.2-fold in abundance. It is possible that the modest increase in abundance of the RAG octamer – and similar avoidance on the lagging strand – indicate that the RAG octamer is under selection as an AIMS and merely fails to exceed our threshold.

To determine if the degree to which the RAGS sequence accumulates towards the replication terminus is significant, we investigated whether sequences accumulated to this degree at other locations in the *E. coli* genome (Fig. 24C). We found that sequences that accumulate in abundance to the same degree as the RAG octamer were as likely to be found accumulating towards non-terminus locations (Fig. 24C). Since the 'accumulation' of octamers at non-terminus locations reflects baseline noise, one cannot conclude that the apparent increase in abundance of the RAG octamer towards the replication terminus reflects selection for function.

Importantly, a previously identified (LAWRENCE and HENDRICKSON 2003), widely-distributed (Table 3) AIMS among proteobacteria, GGGCAGGG, has now been implicated as a potential binding site for the FtsK protein in *Escherichia coli* (BIGOT *et al.* 2005; LEVY *et al.* 2005). This is gratifying, as the FtsK translocase is precisely the sort of protein that would interact with AIMS. Although Levy *et al.* (LEVY *et al.* 2005) point out that the GNGNAGGG octomer is biased in the genomes of several bacteria, strand-bias alone does not provide evidence for selection for function. Indeed, strand biased oligomers may arise by simple differences in mutational proclivities of the DNA polymerases replicating leading and lagging strands (LOBRY 1996), and Table 3 shows that genomes may have hundreds or even thousands of octameric sequences that are strand biased. Further analyses, such as those described herein, are required to demonstrate the footprint of natural selection.

4.5.5 Interplay of mutation and selection

Although the RAG octamer did not increase in abundance more than one would expect at random (Fig. 24BC), it may still be under selection for function. That is, the strand asymmetry we observe may be sufficient for chromosome polarity to be established. The increase in abundance towards the replication terminus accentuates strand asymmetry, which is also a feature we believe is under selection; if natural mutational biases yield both sufficient sequence abundance and sufficient strand asymmetry, then selection acting on these sequences will not change their distribution in any detectable fashion. The distribution of AIMS within a chromosome reflects a balance of mutation and selection, where a gradient of selection from the replication origin to terminus may increase the abundance of AIMS on leading strands if mutation acts to defeat the required asymmetry. When mutation does not defeat asymmetry, selection is less evident.

In some genomes, strand asymmetry – that is, nucleotide skew reflecting mutational biases – is more evident than in others. For example, Firmicutes show a much larger number of strand-biased oligomers than other taxa (Table 3). The pattern may reflect differences in DNA replication in these taxa; Firmicutes utilize DNA polymerase harboring different subunits to replicate their leading and lagging strands, potentially leading to stronger strand asymmetry (ROCHA 2004). In addition, the strong bias of genes to be encoded on leading strands (~80% in Firmicutes) may lead to stronger strand differences. Similarly, the prevalence of genes being encoded on leading strands will result in transcription-coupled repair processes acting differentially between the strands. As a result, AIMS may be less evident in such highly-skewed genomes since mutation does not defeat selected abundance distributions. That is, while the distribution of AIMS reflects selection, the absence of AIMS can not be regarded as an absence of selection.

4.5.6 Impact of AIMS on genome evolution

Most genomes show two large replicores with consistent strand asymmetry (Fig. 19E). In using this asymmetry as an indicator of chromosome rearrangement, we could detect inversions without genome comparison and without ambiguity regarding the polarity of the inversion (DARLING *et al.* 2004). Inversions have been described in many genomes that include the replication origin or terminus (EISEN *et al.* 2000; MACKIEWICZ *et al.* 2001); these rearrangements do not disrupt strand asymmetry and are not detected in our analysis. Our findings suggest that most genomes are recalcitrant to inversion within replicores; we found that only the genomes of obligate pathogens or symbionts contained significant numbers of large inversions within replicores. This finding is consistent with published findings for *Salmonella typhi* (LIU and SANDERSON 1995; LIU and SANDERSON 1996), *Bordetella pertussis* (PARKHILL *et al.* 2003), and *Wolbachia* (FOSTER *et al.* 2005). Therefore, one may ask why large inversions within replicores – that is, those not including the replication origin or terminus – are not found in genomes of free-living, non-pathogenic bacteria.

Selection against some inversions has been demonstrated in the *Salmonella enterica* genome (MAHAN and ROTH 1991; SEGALL *et al.* 1988) The lack of these 'forbidden' inversions does not reflect the inability to form them (MAHAN and ROTH 1991; SEGALL *et al.* 1988). We propose that disruption of the distribution of AIMS – rather than simply placing a gene on the lagging strand, or moving its position relative to the replication origin – counterselects organisms which contain large inversions within replicores. Such inversions would place large numbers of AIMS in their nonpermissive orientation and thus confer a fitness defect. For example, if the FtsK protein relies upon AIMS to translocate towards the replication terminus, the protein would receive incorrect orientation information within large inversions. It has not escaped our attention that selection would also act to limit the acquisition of genomic islands wherein AIMS were present in large numbers in the non-permissive orientation.

Just as genomes of pathogens show a large amount of gene loss (ANDERSSON and ANDERSSON 1999; ANDERSSON and ANDERSSON 1999; COLE *et al.* 2001) – reflecting an inability to select for gene retention (LAWRENCE 2001; LAWRENCE *et al.* 2001; LAWRENCE and ROTH 1999) – inversions also accumulate in these genomes. Such inversions would be insufficiently detrimental to prevent the persistence of strains bearing them. Pathogens often have reduced population sizes and reduced rates of recombination, thereby accelerating the fixation of deleterious changes. Yet mis-polarized AIMS would still be problematic, and the removal of this DNA may be beneficial. The deletion of inverted DNA would likely not be a strategy employed by most organisms, but it is a likely outcome for organisms experiencing genome reduction (ANDERSSON and ANDERSSON 1999; ANDERSSON and ANDERSSON 1999; COLE *et al.* 2001). The occurrence of large inversions in the genomes of some symbionts (MIRA *et al.* 2001) is consistent with this hypothesis. We speculate that the removal of inverted DNA may provide a selective advantage to DNA loss in organisms experiencing genome reduction. That is, deletion of DNA may not always be neutral or detrimental.

4.6 CONCLUSIONS

In bacterial genomes, where space is minimal and the DNA is information rich, AIMS represent an elegant solution to the problem of specifying the direction in which landmarks like the replication origin and terminus can be found. The large numbers of AIMS ensure that, even as the tide of random mutation disrupts individual sequences, the over-all distribution of these important signaling sequences are maintained. We believe that AIMS are a common feature among bacterial chromosomes and this previously unrecognized structure plays a role in influencing the evolution of these genomes. Though the mechanism by which most AIMS act has not been determined, it is possible that perturbations of these sequence patterns are sufficiently disruptive to chromosome maintenance that they are having, and have had, a major role to play in the shape and content of bacterial chromosomes as we see them today.

5.0 CHROMOSOME ARCHITECTURE SHAPES BACTERIAL EVOLUTION

5.1 INTRODUCTION

> *“I think that the most significant aspect of DNA is the support it gives to evolution by natural selection” ~Francis Crick 1989*

The chromosome is more than its list of protein- and RNA-encoding genes and their regulatory regions. The chromosome is also a massive polymer, capable of directing its own defense, repair, replication and segregation. To adopt Richard Dawkin’s phraseology; the organism is only the chromosome’s way of making more chromosomes. To accommodate all of these functions, and in particular to facilitate its replication and segregation into new organisms, the DNA molecule contains sequences which affect evolution by constraining the structure of the molecule itself. These sequences, termed Architecture Imparting Sequences (AIMS), are present in the majority of bacterial genomes. To proteins able to regognize them, AIMS indicate the relative position of the origin and terminus of replication via their overabundance on leading strands. The function of one family of AIMS, which has direct support in *Escherichia coli,* is to orient the action of the FtsK translocase, a protein which acts at the nexus of chromosome segregation, recombination and division. FtsK directs the movement of chromosomes into the appropriate daughter cells during cell division and this gene is present in the majority of bacterial genomes sequenced to date (BARTOSIK and JAGURA-BURDZY 2005).

Many sequences are simply ‘skewed’, being more overabundant on leading strands and underabundant on lagging strands as a result long term mutational tendencies of the replication machinery. Skewed sequences were first elucidated in 1996 when Lobry *et al*. used their strand asymmetry to identify orgins and termini of replication. In contrast, AIMS are identified as repeated eight base pair sequences that are overabundant on leading strands, underabundant on the lagging strand and, most importantly, increase in abundance near the terminus of replication (Figs. 25AB). The observed increase in number of AIMS near the terminus reflects an increased need for proteins to properly orient themselves in this region (Fig 25C). As the septum closes, partitioning both the cytoplasm and DNA into daughter cells, the late replicating terminus region, of one or both daughter cells, is most likely to be trapped in the division septum and require mobilization by proteins like FtsK. This function, terminus translocation, results in a gradient of selection for increased copies of these sequences nearest to the terminus. DNA that is more distant from the terminus experiences less selection for sequences to direct translocation, resulting in a gradual decrease in abundance and decreasing polarization of AIMS in origin proximal locations (Fig 25 ABC).

Bacterial chromosomes have other strand specific features in addition to sequence biases on leading and lagging strands. For example, genes tend to be oriented such that their transcription corresponds to the direction of DNA replication, that is, the majority of genes are transcribed using lagging strands as template strands. Rocha has proposed that this directionality minimizes the number of detrimental collisions between RNA- and DNA-polymerases (ROCHA 2004; ROCHA and DANCHIN 2003). Genes closer to the replication origin also tend to be expressed at a higher level and are more likely to be essential than those near the replication terminus (ROCHA 2004; ROCHA and DANCHIN 2003). This bias may represent an adaptation to the transient higher copy number of the origin proximal region during rapid growth.

These examples of informational structuring along bacterial chromosome are evidence that even small perturbations in the content of bacterial chromosomes can be potent forces in organismal evolution, which has often been reduced to considering only the nature and extent of an organism's gene inventory. This sensitivity can be attributed to the large population sizes of bacteria and the resulting competition experienced by individual bacterial cells; changes with very small selection coefficients, such as the non-random placement or orientation of genes are not effectively neutral. The very small detriment they incur is sufficient (in bacterial populations which may number more than 10^{20}) to lead to their removal by purifying selection. In these large competitive populations, every little bit helps.

Large scale chromosomal constraints were first revealed experimentally when Roth *et al.* investigated the nature of inversions in *Salmonella enterica*. They observed that some chromosomal inversions do not appear to form, even though they can be constructed by other means, suggesting that it is not the inverted DNA itself which is lethal. (MAHAN and ROTH 1991; SEGALL *et al.* 1988; SEGALL and ROTH 1989). The inversions which were not observered were termed "forbidden". The polar replication terminators (Ter sites) could prevent successful chromosomeal replication and most –but not all- of the "forbidden" inversions became permissive upon elimination of the Ter-binding Tus protein. The bacterial chromosome was recalcitrant to what seemed like a relatively innocuous evolutionary event (in that it did not change gene content) and this restraint involved the DNA in the rearranged segment, not the end points. To date, though the mechanism of replication arrest at Ter sites is well understood

(MULCAIR *et al.* 2006; VALJAVEC-GRATIAN *et al.* 2005), the actual function of these sites in the chromosome is not as well understood (HENDRICKSON and LAWRENCE 2007).

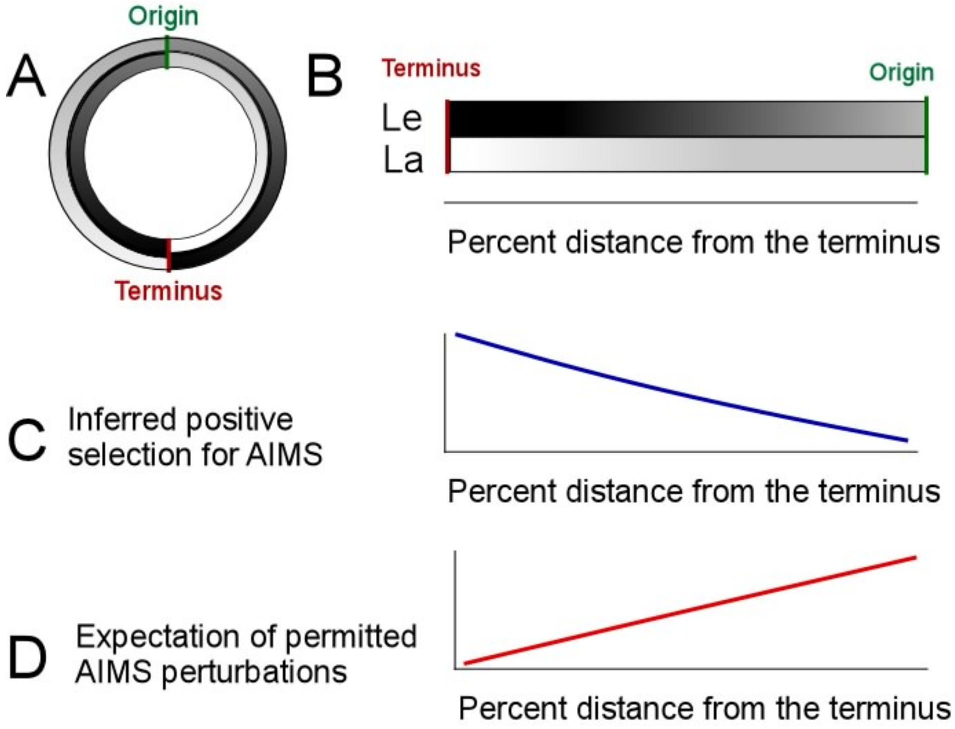
A
Origin
Terminus
B
Terminus
Origin
Le
La
Percent distance from the terminus
C
Inferred positive selection for AIMS
Percent distance from the terminus
D
Expectation of permitted AIMS perturbations
Percent distance from the terminus

Figure 25 Gradients of AIMS I bacterial chromosomes.

5.1.1.1 Figure 25 legend

Bacterial chromosome experiences gradients of selection due to Architecture Imparting Sequences (AIMS). A) A diagrammatic representation of a typical circular bacterial chromosome. Two strands of DNA are represented along with their approximate AIMS concentration (indicated by darker graying). The concentration of AIMS increases on the leading strand and decreases on the lagging strand with decreasing distance to the terminus. B) Same as in A) but now the two chromosome arms from origin to terminus are shown averaged and linearized. C) Positive selection for AIMS (plotted by distance from the terminus as in B)

Figure 25 legend cont.

increases with decreasing distance to the terminus. D) Permitted perturbations of AIMS in chromosomes (plotted as B and C) is expected to increase with increasing distance to the terminus.

Additional restrictions on chromosome evolution were revealed in a paper by Eisen *et al.* in 2000 (EISEN *et al.* 2000). The homologous genes between pairs of closely related bacteria were plotted with each axis corresponding to the position of the orthologue on each chromosome. If gene order were conserved one would expect a diagonal line, where orthologues lie at similar positions in the two chromosomes. Instead, X-shaped plots of homologous genes were observed in the majority of comparisons across many phyla of bacteria. The major diaganol represented genes at similar positions in the two chromosomes, but the anti-diaganol was unexpected. This distribution could result from the repeated inversion of large chromosomal regions that included either the replication origin or terminus of replication at the center. These observations independenly suggested that chromosomal rearrangements were not random with respect to the replication origin and terminus and, more importantly, the rearrangements observed would not affect the action of sequences like Ter (EISEN *et al.* 2000).

The strand bias of genes, clustering of essential genes near the origin, forbidden inversions in the laboratory and symmetrical inversions in nature are all examples of the non-random composition of bacterial chromosomes. Each reflects selection, and is the product of a mechanism which drives the pattern observed. In contrast, AIMS were identified because they were necessary to provide a specific function to the cell; they indicate the relative position of the origin and terminus to proteins like FtsK. One may ask, then, what selective constraints do AIMS impose on bacterial chromosomes?

Levy *et al.* observed that FtsK proteins are responsive to the orientation of a particular AIMS sequence in *E. coli* termed KOPS (LEVY *et al.* 2005; PEASE *et al.* 2005). Specifically, observations of FtsK moving on single DNA molecules showed that FtsK changed directions in response to encountering a particular AIMS sequence. These results imply that disrupting the distribution of AIMS in bacterial chromosomes will slow segregation by interfering with the processivity of FtsK. In general, because the distribution of AIMS reflects selection, perturbations in AIMS distributions will be detrimental to individual cells. Specifically, we predict that (a) reversal of polarized sequences, placing them in primarily the non-permissive orientation, or (b) the introduction of DNA carrying an abundance of mis-oriented sequences will be counter-selected. Moreover, this should be most evident at the replication terminus, where these sequences are at highest selection for function. As a result, selection for the conservation of AIMS will constrain chromosome evolution in bacteria.

We directly investigate this hypothesis by using genomic comparisons in two different ways. First we will evaluate the occurrence of inversions (intragenomic rearrangements) that do not include the replication origin or terminus. We predict that such inversions will persist only where the distribution of AIMS is least disrupted by them: near the replication origin. Second, we will evaluate the propensity for recently acquired DNA (intergenomic rearrangements) to integrate such that incoming AIMS are compatible with local AIMS skew. Lastly, because the likelihood of a bacterial donor genome to provide DNA with compatible AIMS is a function of its relatedness to the recipient genome, we discuss the effect of AIMS on the mode and tempo of gene exchange and evolution in bacteria.

5.2 METHODS

5.2.1 Genome comparisons

The genome sequences for *Aeropyrum pernix, Agrobacterium tumefaciens* str. C58*, Bacillus anthracis* str Ames*, Bacillus cereus* ATCC 14579, *Bacillus cereus* E33L, *Bacillus halodurans* C-125*, Bacillus licheniformis* ATCC14580, *Bacillus subtilis* subsp. subtilis str. 168, *Bacillus thuringiensis* str Al, *Bordetella bronchiseptica* RB50*, Bordetella pertussis* Tohama I*, Borrelia burgdorferi* B31*, Borrelia garinii* PBi*, Bradyrhizobium japonicum* USDA 110*, Bradyrhizobium* sp BTAi1*, Brucella abortus* biovar 1 str, *Brucella ovis* ATCC 25840 chrom*, Burkholderia mallei* ATCC 23344*, Burkholderia pseudomallei* 1106*, Campylobacter fetus* subsp fetus, *Campylobacter jejuni* subsp jejuni, *Candidatus Blochmannia florida, Clostridium perfringens* str 13*, Corynebacterium glutamicum* ATCC13032*, Deinococcus geothermalis* DSM 1*, Deinococcus radiodurans* R1*, Enterococcs faecalis* V583*, Erwinia carotovora* SCRI1043*, Escherichia coli* K12*, Frankia alni* ACN14a*, Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586*, Haemophilus influenzae* Rd*, Helicobacter hepaticus* ATCC 51*, Helicobacter pylori* 26695*, Lactobacillus acidophilus* NCFM*, Lactobacillus casei* ATCC 334*, Lactococcus lactis* subsp cremo*, Lactococcus lactis* subsp lacti, *Lactococcus lactis* subsp. lactis Il1403, *Listeria monocytogenes* 4b F2365*, Mycobacterium avium* 104*, Mycobacterium avium* K-10*, Mycobacterium bovis* AF2122/97, *Mycobacterium leprae* TN, *Mycobacterium smegmatis* str. MC2 155, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium ulcerans* Agy99*, Myxococcus xanthus* DK 1622, *Neisseria gonorrhoeae* FA 1090*, Nitrobacter hamburgensis* X14*, Nitrobacter winogradskyi* Nb-25*, Nocardia farcinica* IFM 10152*, Pasteurella multocida* Pm70*, Pasteurella multocida* subsp multocida*, Prochlorococcus marinus* str AS, *Prochlorococcus marinus* str MI, *Propionibacterium acnes* KPA171202, *Pseudomonas aeruginosa* PAO1, *Pseudomonas fluorescens* PfO-1, *Pseudomonas stutzeri* A1501, *Pseudomonas syringae* DC3000, *Rhodopseudomonas palustris* BisA53, *Rhodopseudomonas palustris* HaA, *Salmonella enterica* Typhimurium LT2, *Shewanella oneidensis* MR-1, *Sinorhizobium medicae* WSM419, *Sinorhizobium meliloti* 1021, *Staphylococcus aureus* MW2, *Staphylococcus aureus* subsp. aureus COL, *Staphylococcus epidermidis* ATC, *Staphylococcus haemolyticus* JC, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* D39, *Streptomyces avermitilits, Streptomyces coelicolor* A3(2), *Symbiobacterium thermophilum* IAM 14863, *Synechococcus* sp JA-3-3Ab, *Thermobifida fusca* YX, *Thermotoga maritima* MSB8, *Thermus thermophilus* HB27, *Thermus thermophilus* HB8, *Tropheryma whipplei* str Twist, *Tropheryma whipplei* TW08/27, *Ureaplasma parvum* serovar 3 st, *Vibrio cholerae* N16961, *Vibrio cholerae* O1 biovar eltor, *Vibrio parahaemolyticus* RIMD 2, *Xanthobacter autotrophicus* Py2, *Xanthomonas axonopodis* pv. citri str. 306, *Xanthomonas campestris* 8004, *Xanthomonas campestris* pv camp, *Xylella fastidiosa* 9a5c, *Yersinia enterocolitica* subsp, *Yersinia pestis* Antiqua, *Yersinia pestis* CO92, *Zymomonas mobilis subsp. mobilis* ZM4 were downloaded from GenBank (NIH 2007). Origins and termini of replication were found as described previously (HENDRICKSON and LAWRENCE 2006).

5.2.2 Phylogeny construction

Phylogeny construction was performed using 16s rRNA sequences from genomes and alignment was performed using the on-line clustalW program (THOMPSON *et al.* 1994). Dendrograms were generated from alignment files using TreeView (PAGE 1996).

5.2.3 Ortholog identification

Orthologs were identified in genome comparisons of two or more genomes by running a pairwise BLAST comparison between all genes; orthologues were taken as reciprocal best matches, discarding ties.

5.2.4 Inversion detection

Best orthologs were taken as described above and plotted by position in their respective genomes with a genome on each axis. For every pair of orthologs, if the direction of transcription of the gene in Genome A (X axis) matched the direction of transcription of the best match in Genome B (Y axis) these orthologs were labeled as being co-oriented in paired genome plots, else the labeling was anti-oriented. Inversions which include either the origin or the terminus will not be seen in these plots as they do not change the relative direction of transcription, neither do they change leading strands to lagging strands. From these plots inversions were identified as cases where the direction of transcription for a series of genes (n >= 4) in a row had been reversed but had not left the main transect of synteny *(i.e.*, regions that appeared to have experienced transposition were not included in the analysis). This cut off was chosen as one where simple inversions could be identified as plainly not being transpositions and were probably a meaningful size in terms of AIMS inverted. Smaller inversions are not likely to invert enough AIMS to be meaningful in this analysis. Inversions that change formerly leading strands to lagging strands and vice-versa would also reverse the direction of AIMS from high abundance on the leading strand to high abundance on the lagging strand.

Inversions were plotted as a function of base pairs of DNA included in the inversion versus distance from the terminus in the genome on the x axis of the graph. If origins or termini of replication appeared to have shifted over the evolutionary time between two genomes having shared a last common ancestor then these comparisons were not analyzed as the position of the origin or terminus is critical to deciding if strand parity will be maintained after an inversion occurs.

5.2.5 Insertion detection

Ortholog lists were generated between sets of genomes (as above). Comparisons were made of presence and absence of genes amongst closely related sets of bacteria in the same genus. Genes that were found in only one of these were categorized as ‘unique genes’. The list of unique genes to a particular genome was taken to be an approximation of recent insertions or HGT into that strain. Newly acquired genes were deleted from the genomes and AIMS were found in the remaining chromosomes as previously described (HENDRICKSON and LAWRENCE 2006). The mean of AIMS in the newly acquired AIMS was plotted as a function of distance from the terminus. The average substitution rate for 16s rRNA in eubacteria is about 1%/50 Myr (OCHMAN and WILSON 1988).

5.2.6 Estimation of genomic HGT compatibility between species

Three genomes, *B. subtilis*, *S. meliloti*, and *M. tuberculosis* were selected as target genomes for estimating the genomic compatibility of a selection of potential donor genomes from 8 phyla. The compatibility index *(C)* was calculated as:

$$C = [\,{}_{d}S^{W} - 0.50] \times \left(\frac{{}_{r}S^{W}}{{}_{r}L}\right)$$

Where S^{w} is the weighted skew for r (the recipient) or d (the donor) and L is the length of the chromosome.

AIMS were selected for each other recipient genomes according to the following criteria;

1) *B. subtilis*, N = 300, >100 per arm, 75% skew, 1.4 fold increase in last bin, 5% increase in skew in last bin, both arms. 24 sequences of degeneracy <= 2. 8639 copies in the genome.

2) *S. meliloti*, N = 200, >75 per arm, 70% skew, 1.2 fold increase in last bin, 5% increase in skew in last bin, both arms. 28 sequences found with degeneracy <= 2. 11,740 copies in the genome.

3) *M. tuberculosis*, N = 250, >100 per arm, 70% skew, 1.2 fold increase in last bin, 5% increase in skew in last bin, both arms. 26 sequences found with degeneracy of <=2. 8176 copies in genome. For more details on this approach to reliably finding robust AIMS see (Hendrickson and Lawrence 2006).

5.3 RESULTS

5.3.1 Disruptive inversions occur away from the terminus

Inversions that do not include either the origin or the terminus of replication will reverse the polarity of leading and lagging strands and place AIMS in a predominantly nonpermissive orientation. If perturbing AIMS distributions in bacterial chromosomes is detrimental, counterselection of these inversions will be strongest where the selection for AIMS is strongest: the terminus region (Fig 25CD). Groups of syntenic genes were identified by plotting the positions of orthologous genes shared in two species. Inversions were identified as genes that had different directions of transcription relative to the replication origin in the two genomes. The genome in which the inversion has actually occurred is not known in this analysis.

We analyzed 8 sets of genome pairs, representing taxa from 3 different bacterial divisions (Fig. 26). Taxa were chosen for which the map order of orthologous genes was largely preserved, so that groups of syntenic genes were unambiguously identified. The chromosomal position of an inversion was assigned as the distance of the midpoint of the inversion from the replication terminus, averaged for the 2 genomes. If the positions of inversions were equally distributed throughout the chromosome arms, then one might expect a mean of their positions of approximately one-half the distance of the origin to the terminus, or 25% of the genome length. For the collection of all data (Fig. 26), as well as data from individual genomes, this null hypothesis was rejected with high significance ($\underline{P}$ <0.0001, one-tailed *t*-test). In all cases, there was a strong tendency to observe greater numbers of inversion within chromosome arms closer to the replication origin. Here, selection for AIMS is very weak. Very few inversions were observed in the terminus region, where we predict that inverting the orientation of AIMS would confound the process of chromosome segregation into daughter cells.

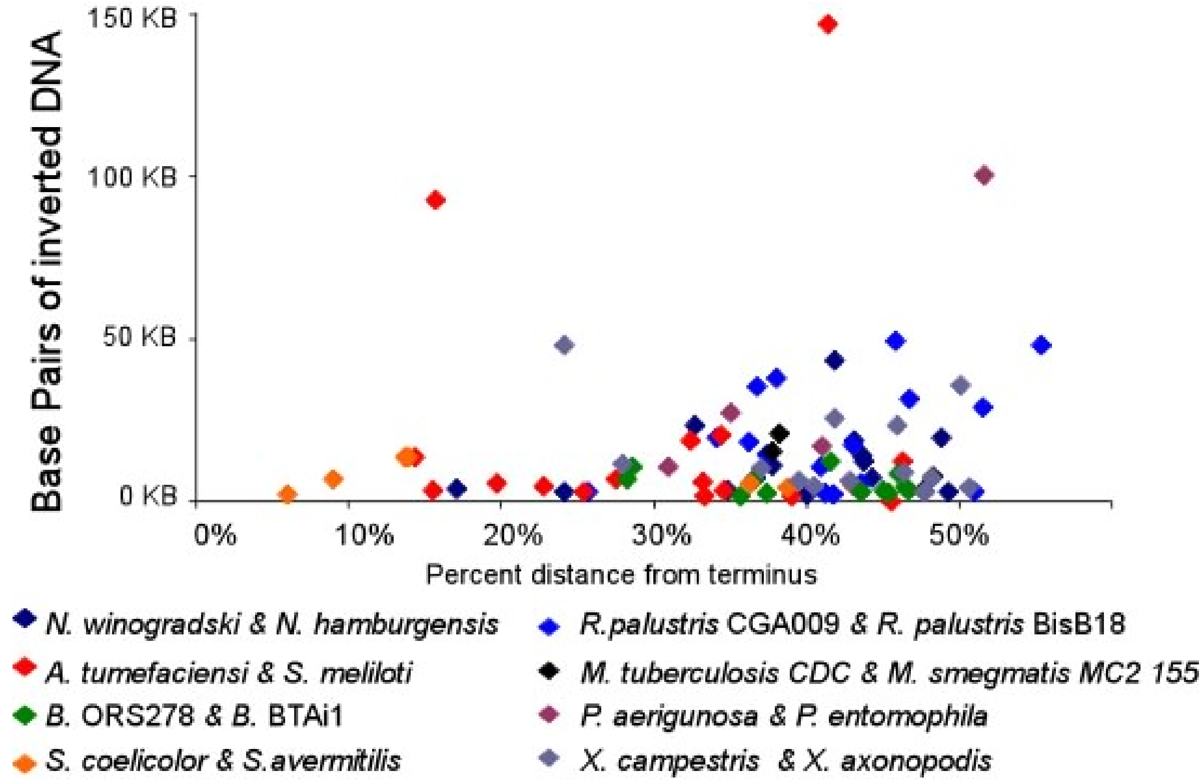

Figure 26 Observed positions of inversions between eight pairs of bacteria.

5.3.1.1 Figure 26 legend

Inversions that would disrupt AIMS in eight pair-wise genome comparisons. Inversions are identified as four or more genes that have reversed their orieintation of transcription relative to the terminus in one genome and not the other. This is the class of inversions that would convert leading strands into lagging strands, thereby reversing the orientation of AIMS.

The total DNA inverted as a function of distance from the terminus is summarized in Fig. 27. Across the taxa analyzed here, there is a strong tendency for inversions to occur more frequently, and to be larger, as a function of distance from the replication terminus.

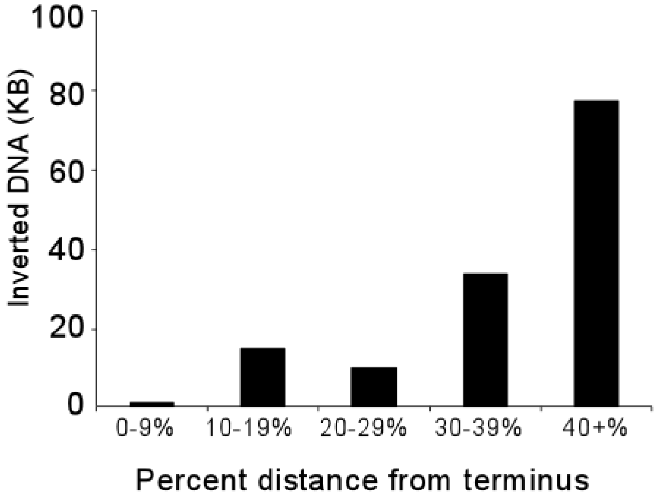

Figure 27 Summary of amount of observed DNA inversion by distance from the terminus.

5.3.1.2 Figure 27 legend

Summed inversions plotted by distance from the terminus. The inversions depicted are those described in Figure 26.

5.3.2 Recent insertions tend to introduce AIMS that are locally compatible

The above data demonstrate that AIMS constrain intragenomic rearrangements. Here, placing AIMS in primarily non-permissive orientations was detrimental, especially near the replication terminus. Horizontal gene transfer introduces foreign DNA into bacterial genomes; these events have been implicated in changing organisms niches and in such cases can be subject to strong positive selection or selective sweeps (DAVISON 1999; GARCIA-VALLVE *et al.* 2000; TOTH *et al.* 2006). Because the incoming DNA has not necessarily experienced selection for the AIMS present in the recipient genome, such intergenomic rearrangements may also introduce large numbers of AIMS in nonpermissive orientations. If the successful acquisition of new DNA is affected by AIMS, there will be selection for donor DNA to recombine into the recipient to maximize the numbers of AIMS in the permissive orientation, and this pressure should be strongest near the replication terminus (Fig. 25 CD).

To determine if AIMS content affects the probability of successful horizontal transfer, we identified recent gene acquisitions and measured the distribution of AIMS within them. Ancient HGT events, such as those available in public databases or found by parametric means, must be excluded from such an analysis since long term amelioration in a novel chromosomal context (LAWRENCE and OCHMAN 1997) will tend to deposit AIMS in acquired DNA where they did not previously exist. To begin, we identified recently acquired (<10 Myr) genes in 11 genomes as open reading frames present in one genome, and absent from at least 2 other closely related organisms, often including other strains of the same species. In this way, we identified recent gene gains and excluded ancient gene gains. Second, we identified AIMS in these genomes, excluding these unique genes from the genomes before analysis of AIMS, thereby allowing an estimation of the AIMS present at the time of acquisition. Using these data, we evaluated the fragments of recently-acquired DNA for the number and skew of the AIMS identified in their respective recipient chromosomes.

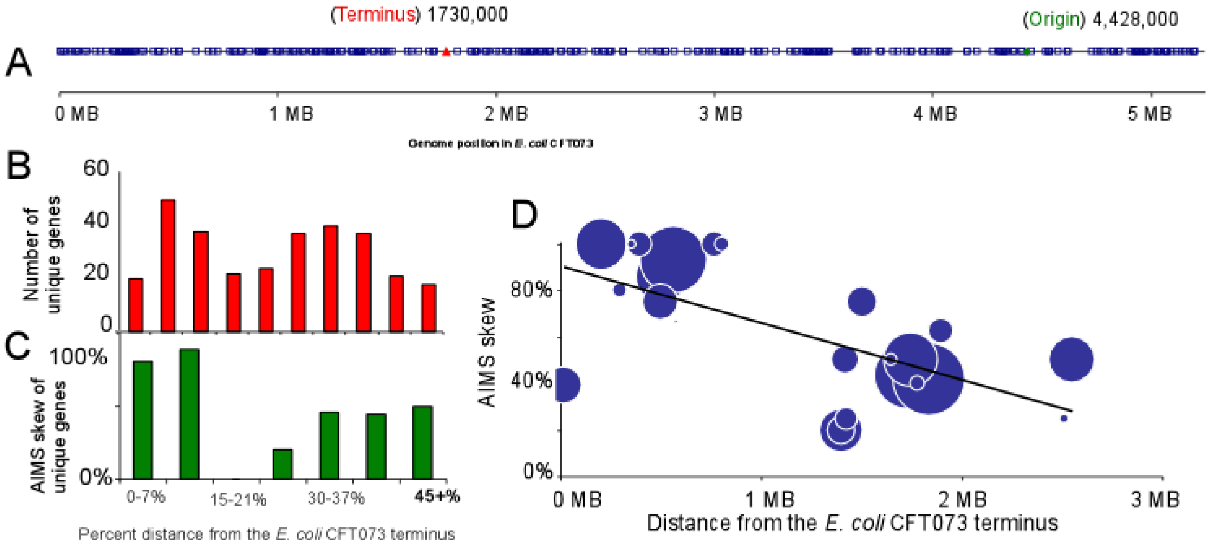

Figure 28 Analyzing AIMS distributions in recently acquired DNA in *E. coli* CFT073.

5.3.2.1 Figure 28 legend

Newly acquired DNA was identified in *E. coli* CFT073 and analyzed for AIMS contribution by area inserted. A) Map of *E. coli* CFT073 showing the positions of the genes identified as recent acquisitions (Blue squares) based on phylogenetic uniqueness in this organism among its close relatives. The approximate origin and terminus of replication are shown for scale. B) The number of unique genes identified as a function of percent distance from the terminus of replication. There does not appear to be an accumulation of recently acquired DNA in this genome in any particular location. C) Skew of recently acquired DNA as a function of percent distance from the terminus of replication. D) A plot of recently acquired DNA that contained at least 4 AIMS. The size of the data points is correlated to the size of the DNA acquired. There is a strong negative correlation between the AIMS skew (AIMS Leading /AIMS Leading + AIMS Lagging) and distance from the terminus in this genome and no evidence of terminus avoidance during acquisition.

Table 4 Number of unique genes (insertions) found in each of 11 taxa.

Genome name	Neighbor 16s rRNA identity	Unique genes found
E. coli CFT073	99	306
E. coli K12	99	96
E. coli UTI	99	100
P. fluoresens Pf01	98	557
P. syringiae 1448a	98	244
P. syringiae B782	99	283
P. syringiae DC3000	98	651
S. enterica serovar typhi	99	16
V. cholerae biovar eltor	95	487
V. parahaemolytica	95	596
V. fisheri ES114	97	229

Recent insertions in the E. coli CFT073 genome are presented in Fig 28. Phylogenetically unique insertions (recently acquired DNA) are represented as blue squares on the genetic map of *E. coli* CFT073 (Fig 28A). There is not, in this organism, a preference for insertions to avoid the terminus region all together, or for insertions to be smaller in the vicinity of the terminus (Fig. 28B). There is, however, a strong tendency for the insertions observed to have high skew if they are retained near the terminus of replication, (Fig. 28 CD). That is, insertions near the replication terminus have AIMS in the appropriate orientations (primarily on leading strands, thus providing strong skew) whereas insertions away from the terminus region have AIMS on both strands. To allow a robust assessment of skew, insertions with fewer then 4 AIMS were excluded, although their inclusion did not eliminate this trend. The relative sizes of the circles in Fig 28D represent the relative sizes of the inserted regions. If all insertions had come into the genome irrespective of the skew of the recipient's AIMS that they contained, the skew of the inserted DNA would not be a function of chromosomal position. Yet these data show significantly higher skew in terminus-proximal insertion (*P*= 0.0083; R^2 =0.28, Spearman rank correlation test).

The summary of recent insertions into 11 independent genomes is presented in Fig. 29. Insertions that occur close to replication terminus carry AIMS in a locally compatible distribution. The number of insertions found per genome is described in Table 4. These data suggest that when the recipient genome's AIMS are not themselves skewed in the genome of a potential donor, successful insertion will occur only near the origin of replication, reducing the potential target size for successful insertions from distant donors. In contrast, genomes with compatible AIMS – where these sequences are skewed in the donor genomes – can contribute genes both near the replication origin and near the replication terminus. We therefore conclude that AIMS are a discriminating force during the acquisition of novel genetic material in bacteria, and has the potential of biasing the pool of potential gene donors towards those with compatible genome architecture.

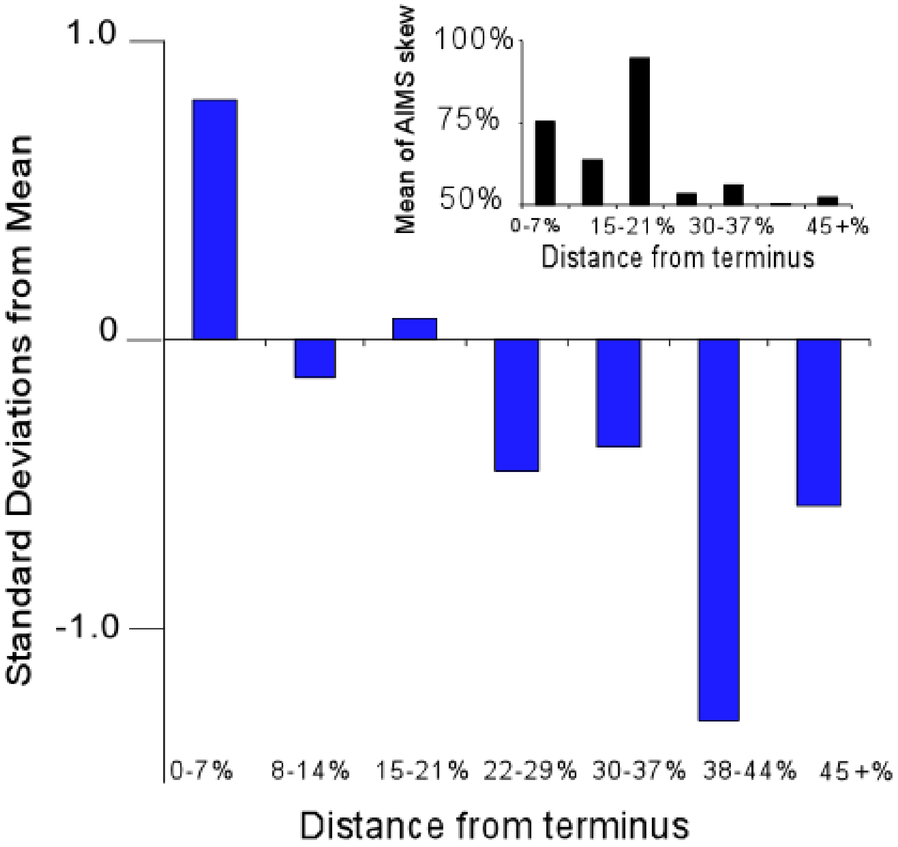

Figure 29 AIMS distributions in recently acquired DNA for 11 bacterial genomes.

5.3.2.2 Figure 29 legend

AIMS in recently acquired DNA insert such that the AIMS that they introduce are skewed compatibility with local AIMS. The standard deviation of each bin away from the average skew across all bins in each chromosome was averaged by bin to produce the standard deviation from the mean plot. The untransformed data for mean across bins are shown in the inset.

5.4 DISCUSSION

5.4.1 Inversions are counter-selected near the replication terminus

We observed that inversions were more frequent near the replication origin than near the terminus. This is satisfying in that disruption of AIMS distributions would be maximal in the terminus region, but there are other elements of chromosomal organization in the origin region which may be disrupted by inversions. For example, genes tend to be oriented to transcribe in the same direction as replication. The strength of this bias has been observed anywhere between 52% and 83% and is highest in some genomes at the origin (BENTLEY and PARKHILL 2004). It has been suggested that this tendency towards co-directionality avoids disruptive collisions which might otherwise occur between DNA and RNA polymerases (ROCHA 2004). In addition, essential genes tend to be near the origin and expressed from leading strands and these inversions would be disrupting those strand biases as well. Thus, without considering AIMS, one might expect inversion to be avoided near the replication origin; our data suggest exactly the opposite, suggesting that selection for retaining AIMS is more potent than selection retaining gene-strand bias.

One might ask if inversions occur near the terminus at all; perhaps their under-representation in that region reflects a local death of the recombination events which form them. But such rearrangements have been observed to occur in the laboratory when appropriate recombinant DNA substrates were provided (SEGALL *et al.* 1988), and the terminus region experiences a higher level of recombination than elsewhere (LOUARN *et al.* 1994). In addition, Rocha examined the distribution of small repeats – the substrates which provide endpoints for inversion formation – and found they did not vary in abundance from origin to terminus. One might suggest that Ter sites would limit inversion near the terminus. But in at least 4 of these organisms the locations of the Ter sites are known and the inversions that we were looking for could be quite small; eliminating inversions that would have included Ter sites does not explain the lack of inversions in the entire terminus region. We conclude that inversions occur near the terminus of replication but are counter-selected due to the disruption of AIMS distributions.

5.4.2 HGT in bacterial chromosomes match local AIMS distributions

DNA recently inserted into bacterial genomes will not have experienced selection for the AIMS present in the recipient genome. Rather, these sequences will be distributed according to the mutational proclivities of their donor genome. Here, they may or may not be skewed. Our data showed that DNA inserted near the replication terminus contains more AIMS in the permissive orientation than do insertions near the replication origin (Figs. 28 & 29). That is, insertions near the replication origin are unbiased, whereas insertions near the replication terminus are biased to contain AIMS in the proper orientation. The insertions analyzed are chosen as unique genes because they have been gained within the past 10 million years. Therefore we believe that the terminus-proximal insertions arrived with the observed level of skew since the time since insertion is insufficient for amelioration to have erased and re-written significant number of sequences (LAWRENCE and OCHMAN 1997; LAWRENCE and OCHMAN 1998; OCHMAN and LAWRENCE 1996).

We suggest that the while insertions with nonpermissive AIMS can insert near the replication terminus, they are counter-selected. The variance in these data indicates that there are insertions with nonpermissive AIMS near the terminus. We argue that these represent two classes: a) those that are so recent that negative selection has not had time to act against the detrimental AIMS and b) those that are experiencing strong selection on the functions encoded by the inserted DNA, so that the benefits outweigh the detriments incurred by improper distributions of AIMS. We propose that a compatible distribution of AIMS will arise by mutation and amelioration.

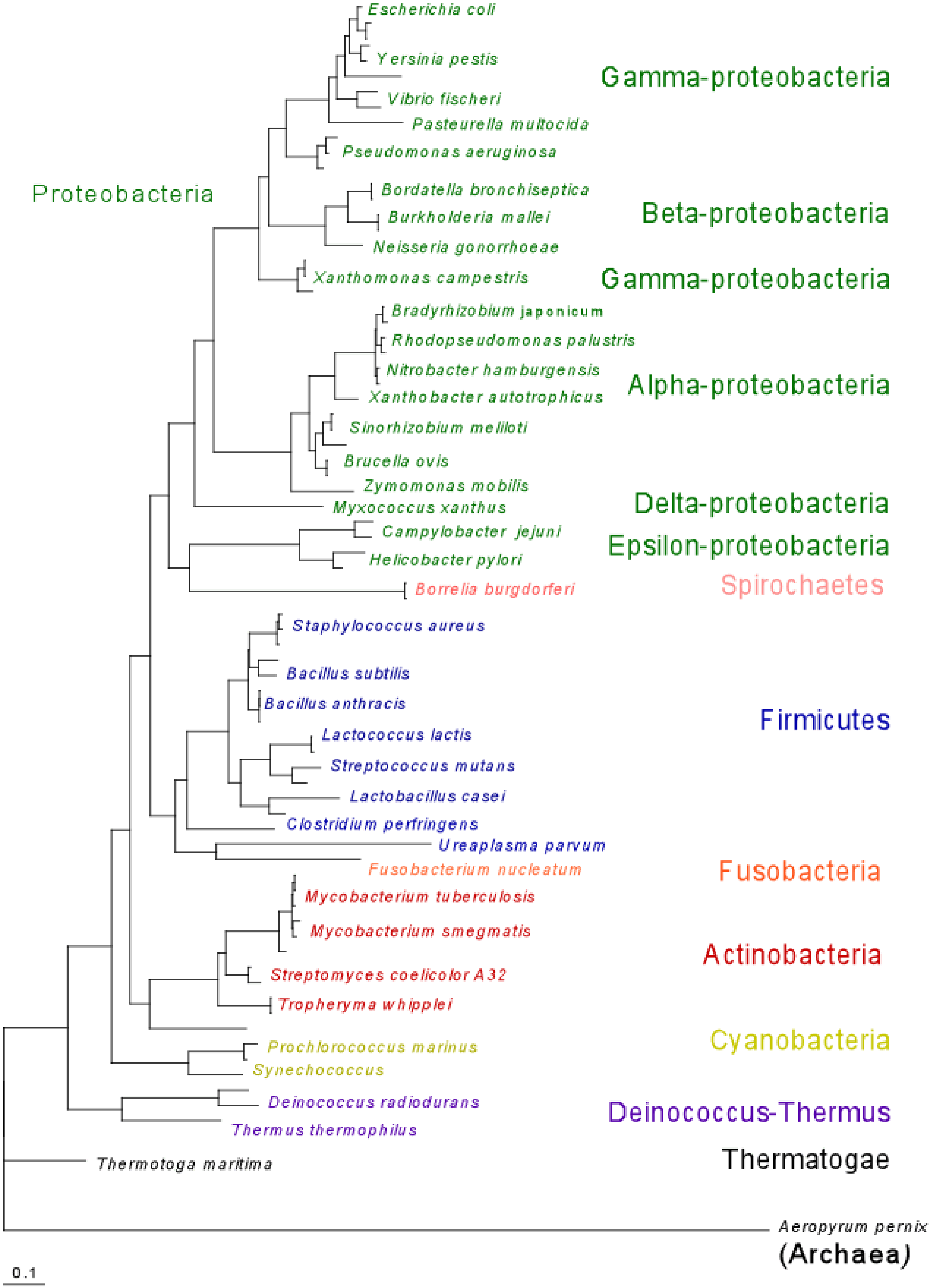

Figure 30 Eight phyla included in the AIMS compatibility test and out-group.

Figure 30 legend (cont. from previous page)

Relationships among 75 bacterial species representing 8 phyla that were used in HGT compatibility modeling as inferred from the sequences of 16S rDNA; a member of the Archaea was used as an outgroup. The dendrogram was constructed in PhyML (GUINDON and GASCUEL 2003) using maximum likelihood methods using the HKY93 substitution model with up to 8 evolutionary rate groups; the model was optimized for transitions/transversion ratio, the number of invariant sites and the gamma parameter. Though some of the very deep branches are suspect, the major divisions between the phyla are represented in accordance with generally accepted notions of relatedness.

5.4.3 HGT is shaped by AIMS

The comparative genomic analyses presented here provide evidence that both intra-genomic and inter-genomic rearrangements which perturb local AIMS structure are counter-selected and subsequently lost. Critically, inserted DNA arriving from genomes wherein the AIMS of the recipient genome are already skewed have a much higher likelihood of being retained. In these cases, the detrimental effects of perturbing AIMS distributions are minimized. Among insertions where AIMS are not skewed in the incoming DNA, the detriment incurred may outweigh the benefits provided by the encoded functions, thereby preventing retention of the newly acquired DNA.

We have previously described the conservation of AIMS amongst phylogenetic neighbors (HENDRICKSON and LAWRENCE 2006); because AIMS are under selection for function, it is not surprising that they are shared among phylogenetically-related taxa. Such conservation has enormous consequences on HGT frequencies between families of bacteria. Transfer between more closely related organisms – for example, bacteria belonging the same family or Division – will be more likely to introduce DNA with the recipient's AIMS in a properly skewed distribution. DNA from more distantly-related organisms will introduce AIMS in non-permissive orientations, and these transfer events will be counter-selected, especially near the replication terminus. This difference in the relative frequencies of successful HGT donors will lead to organisms exchanging DNA most frequently with members of the same taxonomic group. As a result, the taxonomic group will gain cohesion manifested as a shared gene pool. In this way, higher taxonomic units of bacteria may be delineated by virtue of high rates of within-taxa lateral gene transfer, much in the same way that eukaryotic species are delineated as groups which share high rates of homologous recombination.

To measure the compatibility of DNA between donors and recipients, we identified AIMS within recipient genomes and measure their abundance and skew within donor genomes. The phylogenetic relationships between the 75 taxa (78 chromosomes) chosen for this analysis are shown in Fig. 30. The 3 recipients chosen – *Bacillus subtilus*, *Sinorhizobium meliloti*, and *Mycobacterium tuberculosis* – have been shown to harbor distinct classes of AIMS (HENDRICKSON and LAWRENCE 2006). Figure 30 shows the relationship between genome compatibility and phylogenetic distance, here measured as similarity of the 16S rDNA locus. For each recipient analyzed, the most compatible donors include the recipients themselves as well as other members of its Division. The least compatible donors are most often members of other bacterial Divisions. A Spearman's rank correlation of the relationship between 16s rRNA identity and HGT compatibility was performed for the 78 potential donors for each of the three recipients. *S. meliloti* showed the highest correlation between these factors with an $R = 0.70$ and an R^2 of 0.49 (2-tailed $P<0.0001$); a summary for all three recipients is presented in Table 5.

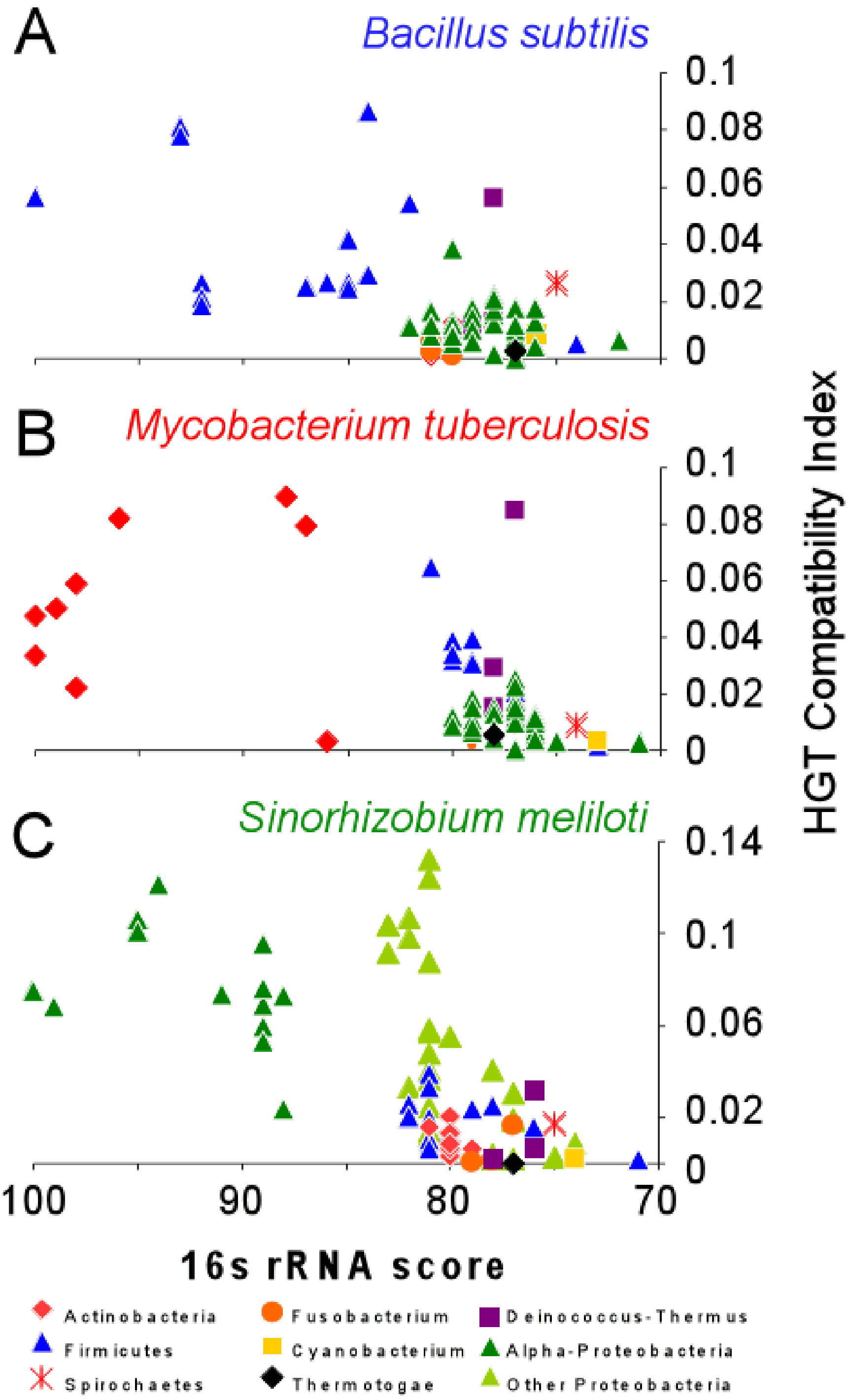

Figure 31 Compatibility for three recipient bacteria from 8 donor Phyla (25 total).

5.4.3.1 Figure 31 legend (cont. from previous page)

HGT compatibility for 78 donor chromosomes into three recipients. A) *Bacillus subtilis.* B) *Mycobacterium tuberculosis* CDC and C) *Sinorhizobium meliloti*. HGT compatibility was calculated as described in the methods. All comparisons are unidirectional and represent only how each of the 78 chromosomes rates as a donor to each of the three recipients, not vice-versa. Colors correspond to those for major divisions as in Figure 30 with the exception of the *S. meliloti* plot where the Alpha-proteobacteria and all other Proteobacteria are shown in dark and light green respectively. In other cases all Proteobacteria are shown in dark green.

The relationship between phylogenetic distance and genome compatibility suggests that the most compatible donors would lie within a species' Division, and DNA inserted from donors outside that Division would be counter-selected more readily. An ANOVA was performed on the compatibility indexes of within-Division vs. out–of-Division species to determine if the compatibility indexes were significantly different between these groups; the means and standard deviations are reported in Table 4. The analysis of variance revealed a significant difference in group HGT compatibility between the Firmicutes and all other phyla for *B. subtilis, F(17,61)* = 67.82 (*p* <0.0001). *M. tuberculosis* and the other Actinobacteria were significantly different in their compatibility from genomes outside of that phylum with an *F (10,68)* = 34.30 (*p* <0.0001). *S. meliloti* and other Proteobacteria were significantly different in their compatibility from the genomes outside of that phylum with an *F(41,37)* = 46.81 (*p* <0.0001). We therefore conclude that there is a statistically significant difference in HGT compatibility between donors in the same Division as the recipient and donors that are outside of the recipient's phylum. In all cases the mean of the compatibility within groups was at least 3 fold higher than the mean of the compatibility outside of the group.

Table 5 HGT compatibility, Spearman's correlation of 16s r RNA and ANOVA for within phyla compatibility

Recipient	R_S	R_S^2	*P*		Category	N	Mean	FOLD	SD
B. subtilis	0.38	0.14	0.007		In phylum	17	0.0451	3.9	0.0272
					Out	61	0.0115		0.0090
M. tuberculosis	0.46	0.21	>0.0001		In phylum	10	0.0486	3.3	0.0293
					Out	68	0.0148		0.0146
S. meliloti	0.70	0.49	>0.0001		In phylum	41	0.0577	4.3	0.0097
					Out	37	0.0133		0.0384

There were three cases where potential donor genomes comprised two large chromosomes; these were *Burkholderia pseudomallei*, *Vibrio cholerae* and *Vibrio parahaemolyticus*. In each case the 2 major replicons were analyzed as separate donors. The 2 chromosomes often varied in genome compatibility, with the ratio of compatibility index ranging from 0.04 fold to 468 fold. Therefore, primary and secondary chromosomes may vary substantially in their propensity to donate genes to different recipients (for more discussion, see Chapter 6).

5.4.4 The effect of AIMS constraints on concepts of bacterial relatedness

The frequency of HGT can affect the appearance of relatedness between bacterial chromosomes (GOGARTEN *et al.* 2002). Those authors suggested that HGT could make groups of organisms appear to be closely related not because they shared a common ancestor, but because they had exchanged large quantities of DNA frequently. Here we suggest that HGT is likely to be most frequent between bacteria within the same Division, owing to the compatibility of their AIMS (Fig. 31). The lack of compatibility in AIMS results in a barrier to gene exchange between species that reside in different bacterial divisions (Fig. 32). The constraint on HGT we propose here is clearly dependent on the size of the fragment of DNA that is initially recombined into the genome. Very small fragments of DNA (100 bp for example) from nearly any donor are unlikely to bring AIMS into a recipient chromosome in any orientation and therefore will not be susceptible to chromosome compatibility constraints. In addition, any gene experiencing strong selection for function – *e.g.*, an antibiotic resistance gene – could provide a function so beneficial that the detriment incurred by disrupting AIMS distributions can be offset.

Here we have presented the presumed effect on HGT based on the AIMS in three phyla: the Actinobacteria, the Firmicutes and the Proteobacteria. There are also gradations of AIMS within these phyla, as well as other phyla with their own ‘specific AIMS’. For example, AIMS in *T. thermophilus* do not appear to be shared with *D. radiodurans,* a species within its Division. It will be necessary to expand this analysis to gain a complete view of the complete set of rules set by AIMS that govern the process of HGT. This will be feasible when much greater numbers of bacterial genome sequences become available, especially for those Divisions not well represented in the currently-available data set.

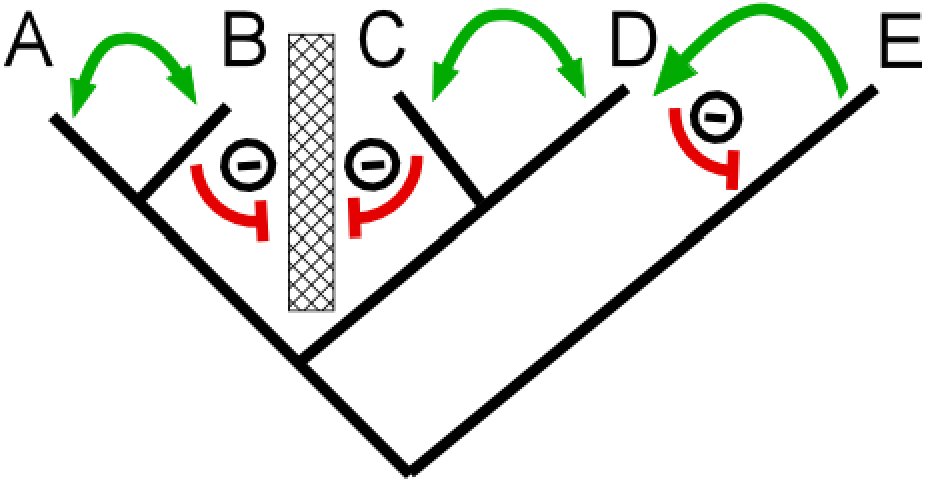

Figure 32 The effect of AIMS compatibility on HGT between organisms.

5.4.4.1 Figure 32 legend

A new model whereby the AIMS allow for more frequent transfer (green arrows) among more closely related organisms, but act as a barrier (albeit not an impervious one) to transfer between distantly related taxa (aborted red arrows). The transfer between D and E is representative of a non-reciprocal transfer constraint which is unique to this model of HGT constraint.

5.4.4.2 HGT permissivity does not have to be reciprocal

If constraints on horizontal gene transfer had environmental or mechanical origins, such as increased transfer among members of the same microenvironments, or among bacteria which share promoter sequences, then permissive HGT in one direction implies permissive HGT in the other direction. Yet in *T. thermophilus* HB27, compatibility in one direction did not necessitate compatibility in the other. *T. thermophilus* is compatible as a donor to differing degrees with both *B. subtilis* (compatibility similar to *B. subtilis* self-compatibility) and *M. tuberculosis* (compatibility 2.5 fold higher than *M. tuberculosis* self-compatibility). This was surprising considering the phylogenetic distance between this member of the Deinococus-Thermus Division and the recipients in question, a Firmicute and an Actinobacterium (Fig 30).

While two *T. thermophilus* representatives were compatible both with each other and with *B. subtilis* and *M. tuberculosis*, both showed limited compatibility as recipients of HGT with all other genomes tested (Fig 33). Division member *D. radiodurans* was also incompatible; given its low similarity to *T. thermophilus*, this is not surprising. *T. thermophilus* is a reasonable gene donor to species outside of its Division but is not an equally reasonable recipient of DNA from those same genomes. This is a new insight into the process of HGT.

AIMS-mediated HGT compatibility can be non-reciprocal (Fig 32). This represents a dramatic departure from other models that have been proposed; reciprocality must be tested in both directions (Figs. 31 and 33). As a result, HGT is less analogous to a highway (BEIKO *et al.* 2005) and more like city streets, where some are two-way and some are one-way. This brings to mind a phrase intoned in Pittsburgh but credited as originated in the state of Maine; 'you can't get there from here'. We predict that more of these non-reciprocal paths will emerge as this constraint on HGT is analyzed with more taxonomic breadth.

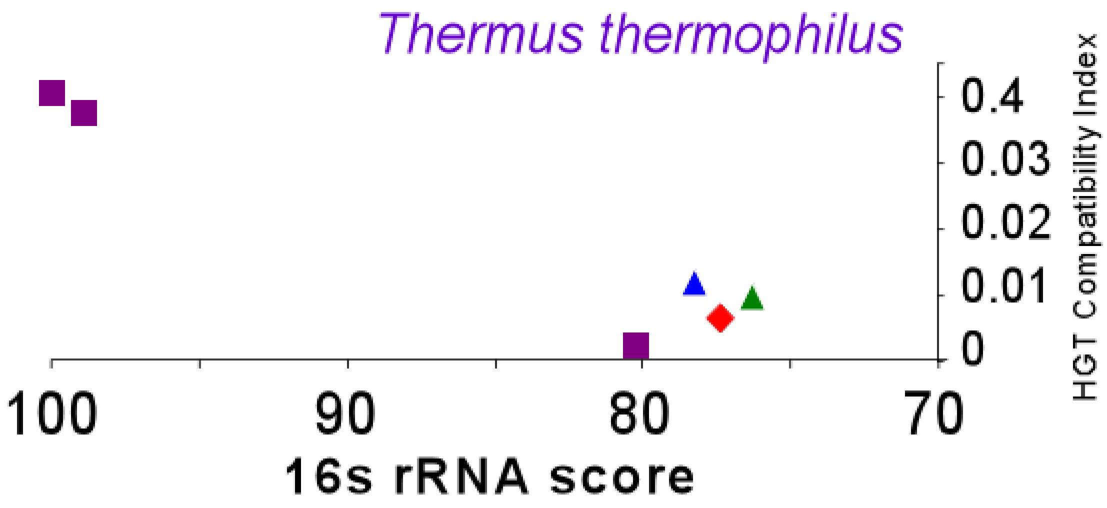

Figure 33 *Thermus thermophilus* HB27 is not a compatible recipient of DNA from 4 taxa.

5.4.4.3 Figure 33 legend

Thermus thermophilus HB27 is not compatible with any of the taxa tested here with the exception of *Thermus thermophilus* HB8. HGT compatibility was calculated as described in the methods. All comparisons are unidirectional and represent only how each of the 6 chromosomes rates as a donor to this recipient, not vice-versa. Colors correspond to those for major divisions as in Figures 30 and 32. This distantly related Deinococcus-Thermus member at a 16s rRNA score of 80 is *D. radiodurans.*

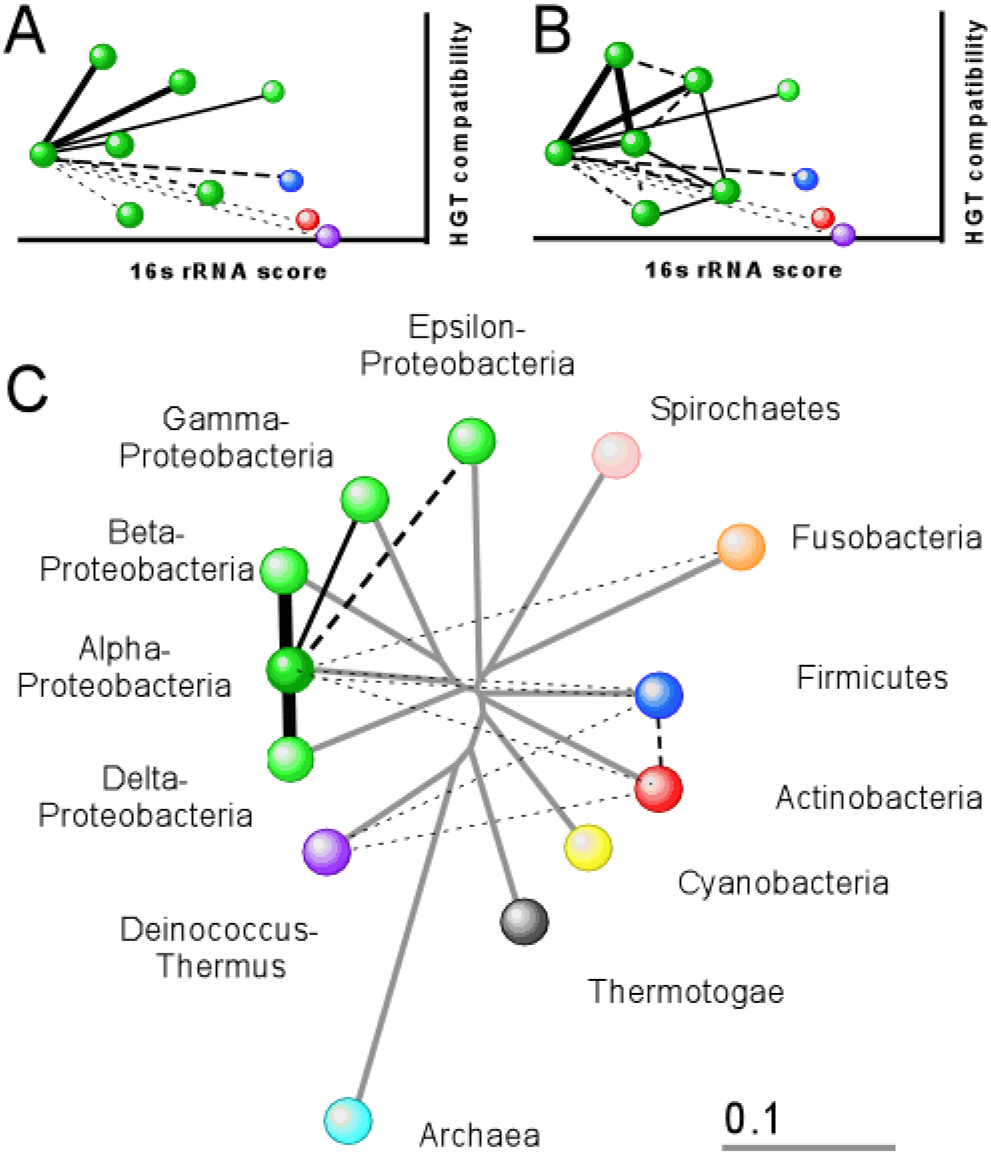

Figure 34 A network diagram of major HGT compatibility observed in this paper.

5.4.4.4 Figure 34 legend

An approximation of the major compatibility observed in this paper. A) A caricature of the data presented in Fig. 31 A of this chapter. Each ball representing a single replicon for which compatibility was tested. We have information about the uni-directional compatibility for many species into *S. meliloti.* Compatibility can be thought of as a proxy for predicted frequencies of transfer permitted according to this constraint on HGT. These expected frequencies are superimposed as weighted and dashed lines to represent the amount of transfer. B) The evidence that we have suggests strongly that with-in group transfer is frequent because of intra-division compatibility. Therefore, though we do not know the weights we can assume that the Alpha-proteobacteria form a tightly knit group of co-compatible DNA structure. These exchange DNA at a lower frequency with taxa outside of their division. C) A diagram representing the exchange frequencies observed in this chapter. The 16s rRNA dendrogram (grey) was made using representative 16s rDNA sequences from each of the phyla depicted as described previously. Each of the balls can now be thought of as the cluster of frequently exchanging groups in part B. The lines between the balls are drawn according to the frequencies implied by the division representatives analyzed (see Fig 31). This figure is speculative but using social network software we will be able to produce rigorous diagrams of this sort (see chapter 6).

5.5 SUMMARY

The work presented here describes a constraint on both intra- and inter-chromosomal rearrangement. The compatibility of incoming DNA with the chromosome structure of a potential recipient will change the likelihood of a successful HGT event. We have previously presented data which indicate that AIMS are a conserved feature of closely related bacteria (HENDRICKSON and LAWRENCE 2006) This pattern of conserved AIMS constraining distant transfer will shape the flow of incoming DNA such that the majority of transfers will occur between related groups, thereby increasing the genomic cohesion of those groups. AIMS are playing a large role in shaping the networks of genetic transfer between organisms. This provides a mechanism to explain the conundrum of the universal tree of life, whereby seemingly robust taxonomic groups are maintained in the face of high rates of lateral gene transfer. The tree of life is sustained in the face of frequent natural HGT because that HGT is constrained to close relatives and therefore strengthens the similarity between them. In other words, a Proteobacterium looks like a Proteobacterium not just because it shares a last common ancestor with the rest of the Proteobacteria, but because it exchanges DNA with them most frequently as well.

6.0 ADDITIONAL PERSPECTIVES REGARDING BACTERIAL EVOLUTION

"Most species do their own evolving, making it up as they go along, which is the way Nature intended. And this is all very natural and organic and in tune with mysterious cycles of the cosmos, which believes that there's nothing like millions of years of really frustrating trial and error to give a species moral fiber and, in some cases, backbone." ~Terry Pratchett

During the course of looking for architecture in available completely sequenced genomes my mind has wandered into the exciting but dangerous realm of speculation. This chapter contains a handful of explorations which, seem to me fruitful new directions, worthy of serious further consideration, but not entirely developed as yet. In some cases these are issues I intend to pursue immediately and in some they are merely proposals to the scientific community in general.

6.1 HISTORICAL DNA TOPOGRAPHY: BETTER MOLECULAR BIOLOGY THROUGH BIOINFORMATICS

Bioinformatics can be applied to test not only how a bacterium is currently using its sequence architecture, but what has happened in the past. Examining repetitive sequences in genomes to look for signals that indicate how error prone, repetitive processes have been occurring over time is similar to looking along a creek bed at the strata of rock to understand how a particular piece of land was formed. I define Historical DNA Topography (HDT) as the study of DNA sequences to obtain information about the history of the processes that shape those sequences. Replication, repair, transcription and translation all processes which have weak mutational affects on the DNA and therefore are processes that can be detected through examination of the DNA sequence.

During this dissertation, bioinformatics was used to reveal the origin and terminus of replication using the repetitive sequences deposited by the mutational proclivities of the leading vs. lagging strand machinery (HENDRICKSON and LAWRENCE 2006). That pursuit required an understanding of the structure of mutational change in the terminus region in particular as that is where AIMS are most important. We discovered to our surprise that skewed sequences were indicating that replication termination was occurring most frequently at a single position in the terminus region and that this closely corresponded with the *dif* site. Our original expectations for mutational structure and skew in the terminus region had incorporated the notion that replication termination might halt at any one of a set of polar replication terminators, thereby making skew weak in this region. Upon characterization of the replication patterns however we found that a single location was generally utilized in the terminus region (HENDRICKSON and LAWRENCE 2007). This was a surprising observation to many in the field and is being investigated further by other laboratories. That is not however, the end of the surprises revealed by skewed sequences about replication.

There are similar inconsistencies, which have yet to be explored, regarding the origin of replication. Figures 17 (*E. coli*) and 19D (*R. palustris*) contain plots of the number of sequences in a proportion of the chromosomes (20%) that define a break point in skew at a locations along the axis of the chromosome. Both of these figures, and many more for many other genomes, suggest that the origin of replication is not the well defined single point of replication initiation that we have envisioned it to be after decades of molecular and *in vitro* characterization. DNA replication is described as originating at a specific location (the location at which DnaA binds), we expected leading and lagging strand character to be rigorously defined on either side of this point. Yet, unlike the replication terminus, no such clear transition between leading- and lagging-strand character is observable in the origin region.

These patterns need to be studied further and are the basis of a seed project for my post-doctoral work. If the patterns stand up and are not the result of extreme instability (for example, very frequent inversions) then the implication is that the precise location of replication start is variable while the precise location of replication termination is more reliable. If the location of replication start is variable this could imply 1) that there is variation (or physical shift in location) of this initiation site over time in all bacterial lineages, 2) that there is regulated variation in the exact location that replication forks start at in response to different cellular cues from the environment (this could change the effective ploidy of different sets of genes in the origin region), or 3) that there is stochastic variation in the absolute location of DNA strand melting, despite a consistent location of replication initiation at the DnaA boxes (the sites where the initiation protein, DnaA binds and begins the process of replication). HDT is an approach which can be used to learn about processes and mechanism that are taking place in chromosomes. These hypotheses can be explored using molecular techniques but it is the bioinformatics which indicates that such hypotheses can be formed.

A method such as HDT gives us information about the mechanisms that we can not directly observe which may vary either on a population level or on evolutionary time scales. The biochemistry and *in vitro* work that has forged the foundations of our understanding of the processes taking place in bacterial cells have left us with a false sense of certainty about processes which may vary at these levels. This additional complexity in bacterial life may have large consequences for us. For example, if a drug were to be designed to target the origin region and block bacterial replication forks from a certain direction, my data suggest that a random subset of bacteria in any population would escape such a selection if origins are more variable than we have historically believed.

6.2 AIMS AND PLASMID EVOLUTION

AIMS have influenced the evolution of bacteria for a very long time. These sequences aid in solving ancient and ubiquitous problems of DNA segregation in the Bacteria. AIMS do not however appear to be the only solution to segregating replicons. For example, many bacterial plasmids carry the Par system for partitioning. ParM is an ancient actin homolog that polymerizes at mid-cell and pushes the copies of the plasmid out to the new mid-cell positions of the dividing bacteria (GARNER *et al.* 2004; MOLLER-JENSEN and GERDES 2004). This avoids two problems: 1) needing to maintain the skew of a particular host in order to be compatible with the host's segregation system (increasing effective host range) and 2) being able to replicate and segregate without being integrated into the host chromosome, again increasing host range by avoiding the problems of integrating with incompatible host DNA.

It is possible that plasmids have evolved or at least continued to be strongly selected as independent replicons because they have maintained their autonomy from bacterial chromosomes in these two ways. This frees these entities to act selfishly when they would otherwise be tied more directly to host chromosome health and segregation. Perhaps plasmids are plasmids in order to avoid being constrained by a single taxon's AIMS.

This leads to another point and that is the issue of secondary large replicons. I discussed previously the issue of whether a large replicon should be called a secondary chromosome or whether it should be called a plasmid. The current nomenclature involves the somewhat arbitrary identification on the replicon of "necessary genes" as evaluated by homology and or lab media growth detriments in the face of replicon loss. Establishing whether or not the secondary replicon though large or small has the same AIMS as the larger replicon would be a biologically driven method for chromosome nomenclature.

In some cases a secondary chromosomes appears to be a plasmid with an independent segregation mechanism which has simply gained large amounts of DNA (often recently) so as to be the size of the large replicon (GERDES *et al.* 2000). The important difference here is the degree to which the secondary replicon is really 'part' of the genome of the host vs. being a guest or a transiently antagonistic independent entity.

There are likely two stages to developing AIMS in a large replicon. First, the polymerases that act on the DNA must act for a long enough period of time that the native accumulation of mutationally biased sequences must develop on the replicon. At this point the skewed sequences will be compatible between the major and minor replicons but the secondary replicon does not yet contain AIMS. Second, the segregation machinery that acts on the major replicon, be it FtsK or some other mechanism, must start to act on this secondary chromosome along with the first one. AIMS are therefore a signal that two separate essential sets of cellular machineries, replication and segregation are acting on both of the chromosomes in the cell.

Adaptation in the lineage to this extent would then make it a chromosome in that genome as opposed to a less permanent feature of the genome.

This latter idea also implies that secondary replicons may actually be easy targets for recently acquired DNA. As a small unstructured replicon that does not carry AIMS to inhibit compatibility, a plasmid will be a target for recombination with incoming DNA approximately equal to the ratio of plasmid to chromosomal DNA. This will start to expand the physical size of the plasmid .Once HGT increases the size to a degree the target is large, not only because of its actual size but because of the fact that this secondary replicon is a more permissive target for HGT because of its absence of AIMS structure along with the smaller selection coefficients of the newly acquired DNA. Once the process has begun and expansion of the plasmid is underway a rapid succession of genome contents (along a backbone responsible for replication and ParM like segregation) will occur. This may explain the extreme synteny disruptions in the secondary chromosomes in *Vibrio* and *Agrobacterium* as compared to the conservation of synteny amongst their primary chromosomes (Hendrickson, unpublished results).

6.3 AMELIORATION OF NOVEL DNA BEFORE TRANSFER WITHIN A PHYLUM

If positive selection on novel function is sufficient to retain a recent acquisition in the face of deleterious AIMS, the process of amelioration will eventually work to correct the sequence skew in the new DNA. Once a particular member of a phylum has adjusted the AIMS distribution (through random mutational changes over evolutionary time) to more closely match the rest of the phylum, the new DNA will also be susceptible to increased transfer with-in the phylum. In other words, HGT compatibility of a particular piece of disruptive but useful DNA will tend to increase with time in the Division. In this way a particular bacterial species, experiencing strong selection to shift into a new niche, can become a reservoir of novel adaptive DNA for the phylum.

6.4 BACTERIAL NETWORKS PARALELL SOCIAL NETWORKS: INFORMATION EXCHANGE AND DEFINING GROUPS

One of the new observations which come from this work is that there is a natural flow of genetic information between different groups of organisms. This idea bears a remarkable resemblance to ideas discussed in the field of social networks with regard to the flow of information between members of a network. An example of a social network is a group of friends who exchange e-mails and news. Applying social network theory to the study of genetic transfer has the potential introduce us to new ways of thinking about information flow. For example, Everret M. Rogers has studied the diffusion of innovations in social networks. His research suggests that there are different stages that an individual goes through (a single bacterium in our case) during the evaluation and adoption of new technology as well as stages that a population goes through as more individuals adopt a new technology (new ability conferred by HGT). Application of this field of study on the analysis of bacterial exchange networks predicts that innovations will spread within a population in an S-curve. Meaning that a trait (or new technology) will be acquired by a small number in the beginning but at some point it will ‘catch on’ and in a short time the majority of the population will have acquired it. Towards the end of the spread of the new trait there will be a plateau in the rate of adoption since 1) some individuals will just be recalcitrant to the change and 2) in this late stage the number of individuals who can take on the new trait has been reduced dramatically. Those who could or would already did. Social networking theory also states a number of exceptions to such a rule. Exceptions include disruptive technologies (niche changing innovations which isolate recombining populations) and the path dependence of certain changes such as innovations which are incompatible with other possible innovations. This is an interesting line of thought to consider because one might be able to reverse engineer the trait in a bacterial population by examining its spread through a population or populations of bacteria.

Another interesting potential for thinking of HGT in bacteria in terms of social network analysis is as an aid to concepts of species in bacteria. Algorithms are being developed to define what a true “group” is in network theory (Rosvall and Bergstrom 2007). It is tempting, in the face of the mixing effect of HGT to abandon the bacterial species concept. From a social networking paradigm, groups can be defined based on the perceived amount of exchange between individuals is a common and solvable problem. In fact, there is a very nice piece of formal theory in social network analysis which refers to the idea of weak vs. strong ties in social networks. A strong tie is defined as two individuals who exchange information frequently. For our purposes these might be two members of the same species. Small groups of individuals who all exchange information very frequently are said to be limited in their knowledge to colloquial news (a small limited gene pool). However, if an individual in this group has what is called a ‘weak tie’ to an individual from outside of this group there is an occasional influx of highly novel information which can then be shared with the other members of the small, tightly associated clique. Weak ties actually make larger groups more cohesive (Granovetter 1973). The value of such an idea for HGT theory is immediately obvious. Loosely associated groups of organisms, those with ‘weak ties’ to other Divisions, will occasionally see an influx of novel genetic material which may be useful. Organisms that do not have access to this sort of influx are necessarily left to their colloquial ways.

It is my hope that these two fields of study come together in the near future. There is a very useful ‘weak tie’ to be forged there. It has, in a way already begun. Network analysis was recently undertaken by a biologist in a novel way in the case of Carl Bergstrom (Bergstrom 2007). In that work a network analysis of journal citations was performed with an aim to rank journals according to the quality of citations they received.

Using the predictions for gene exchange frequency that I have discovered as well as the tools for modeling available in social networking theory should allow us to incorporate notions of lateral gene transfer into our species concept in the future (see Fig 34).

6.4.1 Summary

I have used this final chapter to briefly mention some of the consequences that I think fall out of the work I have presented that are not directly supported by data in this dissertation. historical DNA topography, plasmid or second chromosome evolution, amelioration of novel HGT and modeling bacterial groups as social networks are all potentially fertile ground for further research that have sprung from the idea that AIMS act as a constraint during bacterial evolution.

BIBLIOGRAPHY

ALTSCHUL, S. F., 1991 Amino acid substitutions matrices from an information theoretic perspective. J. Mol. Biol. **219:** 555-565.

ANDERSEN, P. A., A. A. GRIFFITHS, I. G. DUGGIN and R. G. WAKE, 2000 Functional specificity of the replication fork-arrest complexes of *Bacillus subtilis* and *Escherichia coli*: significant specificity for Tus-Ter functioning in *E. coli*. Mol. Microbiol. **36:** 1327-1335.

ANDERSSON, J. O., and S. G. ANDERSSON, 1999 Genome degradation is an ongoing process in *Rickettsia*. Mol. Biol. Evol. **16:** 1178-1191.

ANDERSSON, J. O., and S. G. ANDERSSON, 1999 Insights into the evolutionary process of genome degradation. Curr. Opin. Genet. Dev. **9:** 664-671.

ANDERSSON, S. G., and C. DEHIO, 2000 *Rickettsia prowazekii* and *Bartonella henselae*: differences in the intracellular life styles revisited. Int J Med Microbiol **290:** 135-141.

ARIGONI, F., F. TALABOT, M. PEITSCH, M. D. EDGERTON, E. MELDRUM *et al.*, 1998 A genome-based approach for the identification of essential bacterial genes. Nat. Biotechnol. **16:** 851-856.

ASAI, T., S. SOMMER, A. BAILONE and T. KOGOMA, 1993 Homologous recombination-dependent initiation of DNA replication from DNA damage-inducible origins in *Escherichia coli*. EMBO J **12:** 3287-3295.

AZAD, R. K., and M. BORODOVSKY, 2004 Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. Brief Bioinform. **5:** 118-130.

AZAD, R. K., and J. G. LAWRENCE, 2005 Use of artificial genomes in assessing methods for atypical gene detection. PLoS Computational Biology **1:** e56.

BALBINDER, E., 1993 Multiple pathways of deletion formation in Escherichia coli. Mutat Res **299:** 193-209.

BARINAGA, M., 1996 A shared strategy for virulence. Science **272:** 1261-1263.

BARTOSIK, A. A., and G. JAGURA-BURDZY, 2005 Bacterial chromosome segregation. Acta Biochim. Pol. **52:** 1-34.

BAUMLER, A. J., B. M. HARGIS and R. M. TSOLIS, 2000 Tracing the origins of *Salmonella* outbreaks. Science **287:** 50-52.

BEIKO, R. G., T. J. HARLOW and M. A. RAGAN, 2005 Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci., USA **102:** 14332-14337.

BENTLEY, S. D., and J. PARKHILL, 2004 Comparative Genomic Structure of Prokaryotes. Annual Review of Genetics **38:** 771-791.

BERG, O. G., and C. G. KURLAND, 2002 Evolution of microbial genomes: sequence acquisition and loss. Mol. Biol. Evol. **19:** 2265-2276.

BERGSTROM, C. T., 2007 Eigenfactor: Measuring the value and prestige of scholarly journals. C&RL News **68**.

BERGTHORSSON, U., and H. OCHMAN, 1995 Heterogeneity of genome sizes among natural isolates of Escherichia coli. Journal of Bacteriology **177:** 5784-5789.

BERGTHORSSON, U., and H. OCHMAN, 1998 Distribution of chromosome length variation in natural isolates of Escherichia coli. Molecular Biology & Evolution **15:** 6-16.

BIGOT, S., J. CORRE, J. M. LOUARN, F. CORNET and F. X. BARRE, 2004 FtsK activities in Xer recombination, DNA mobilization and cell division involve overlapping and separate domains of the protein. Mol. Microbiol. **54:** 876-886.

BIGOT, S., O. A. SALEH, C. LESTERLIN, C. PAGES, M. EL KAROUI *et al.*, 2005 KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. EMBO J **24:** 3770-3780.

BIRD, R. E., J. LOUARN, J. MARTUSCELLI and L. CARO, 1972 Origin and sequence of chromosome replication in *Escherichia coli*. J. Mol. Biol. **70:** 549-566.

BLACK, J. G., 1996 *Microbiology. Principles and Applications*. Prentice Hall., Upper Saddle River, New Jersey.

BLAKELY, G., S. COLLOMS, G. MAY, M. BURKE and D. SHERRATT, 1991 *Escherichia coli* XerC recombinase is required for chromosomal segregation at cell division. New Biol. **3:** 789-798.

BLAKELY, G., and D. SHERRATT, 1994 Determinants of selectivity in Xer site-specific recombination. Genes Dev. **10:** 762-773.

BOCCARD, F., E. ESNAULT and M. VALENS, 2005 Spatial arrangement and macrodomain organization of bacterial chromosomes. Mol. Microbiol. **57:** 9-16.

BREIER, A. M., H. U. WEIER and N. R. COZZARELLI, 2005 Independence of replisomes in *Escherichia coli* chromosomal replication. Proc. Natl. Acad. Sci., USA **102:** 3942-3947.

BRESLER, S. E., V. A. LANZOV and V. T. LIKHACHEV, 1973 On the mechanism of conjugation in *Escherichia coli* K12. 3. Synthesis of DNA in the course of bacterial conjugation. Mol. Gen. Genet. **120:** 125-131.

BRESLER, S. E., V. A. LANZOV and A. A. LUKJANIEC-BLINKOVA, 1968 On the mechanism of conjugation in *Escherichia coli* K 12. Mol. Gen. Genet. **102:** 269-274.

BROCHIER, C., H. PHILIPPE and D. MOREIRA, 2000 The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet. **16:** 529-533.

BROWN, J. R., C. J. DOUADY, M. J. ITALIA, W. E. MARSHALL and M. J. STANHOPE, 2001 Universal trees based on large combined protein sequence data sets. Nature Genet. **28:** 281-285.

BUSSIERE, D. E., and D. BASTIA, 1999 Termination of DNA replication of bacterial and plasmid chromosomes. Mol. Microbiol. **31:** 1611-1618.

CAPIAUX, H., F. CORNET, J. CORRE, M. GUIJO, K. PERALS *et al.*, 2001 Polarization of the *Escherichia coli* chromosome. A view from the terminus. Biochimie **83:** 161-170.

CAPIAUX, H., C. LESTERLIN, K. PERALS, J. M. LOUARN and F. CORNET, 2002 A dual role for the FtsK protein in *Escherichia coli* chromosome segregation. EMBO Rep. **3:** 532-536.

CARLSON, C. R., and A. B. KOLSTO, 1994 A small (2.4 Mb) Bacillus cereus chromosome corresponds to a conserved region of a larger (5.3 Mb) Bacillus cereus chromosome. Molecular Microbiology **13:** 161-169.

CASJENS, S., 1998 The diverse and dynamic structure of bacterial genomes. Annu. Rev. Genet. **32:** 339-377.

CLARK, A. J., 1991 *rec* genes and homologous recombination proteins in *Escherichia coli*. Biochimie **73:** 523-532.

CLERGET, M., 1991 Site-specific recombination promoted by a short DNA segment of plasmid R1 and by a homologous segment in the terminus region of the *Escherichia coli* chromosome. New Biol. **3:** 780-788.

COHAN, F. M., 2001 Bacterial species and speciation. Syst. Biol. **50:** 513-524.

COLE, S. T., K. EIGLMEIER, J. PARKHILL, K. D. JAMES, N. R. THOMSON *et al.*, 2001 Massive gene decay in the leprosy bacillus. Nature **409:** 1007-1011.

CORNET, F., J. LOUARN, J. PATTE and J. M. LOUARN, 1996 Restriction of the activity of the recombination site *dif* to a small zone of the *Escherichia coli* chromosome. Genes Dev. **10:** 1152-1161.

CORRE, J., F. CORNET, J. PATTE and J. M. LOUARN, 1997 Unraveling a region-specific hyper-recombination phenomenon: genetic control and modalities of terminal recombination in *Escherichia coli*. Genetics **147:** 979-989.

CORRE, J., and J. M. LOUARN, 2002 Evidence from terminal recombination gradients that FtsK uses replichore polarity to control chromosome terminus positioning at division in *Escherichia coli*. J. Bacteriol. **184:** 3801-3807.

CORRE, J., and J. M. LOUARN, 2005 Extent of the activity domain and possible roles of FtsK in the *Escherichia coli* chromosome terminus. Mol. Microbiol. **56:** 1539-1548.

CORRE, J., J. PATTE and J. M. LOUARN, 2000 Prophage lambda induces terminal recombination in *Escherichia coli* by inhibiting chromosome dimer resolution. An orientation-dependent *cis-* effect lending support to bipolarization of the terminus. Genetics **154:** 39-48.

COSKUN-ARI, F. F., and T. M. HILL, 1997 Sequence-specific interactions in the Tus-Ter complex and the effect of base pair substitutions on arrest of DNA replication in *Escherichia coli*. J Biol Chem **272:** 26448-26456.

COSTANTINO, N., and D. L. COURT, 2003 Enhanced levels of lambda Red-mediated recombinants in mismatch repair mutants. Proc. Natl. Acad. Sci., USA **100:** 15748-15753.

CRAWFORD, I. P., and R. MILKMAN, 1991 Orthologous and paralogous divergence, reticulate evolution, and lateral gene transfer in bacterial *trp* genes, pp. 77 - 95 in *Evolution at the Molecular Level*, edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM. Sinauer Associates, Sunderland, MA.

CUNNINGHAM, E. L., and J. M. BERGER, 2005 Unraveling the early steps of prokaryotic replication. Curr. Opin. Struct. Biol. **15:** 68-76.

DARLING, A. C., B. MAU, F. R. BLATTNER and N. T. PERNA, 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:** 1394-1403.

DARWIN, C., 1859 *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life*. John Murray, London.

DAUBIN, V., N. A. MORAN and H. OCHMAN, 2003 Phylogenetics and the cohesion of bacterial genomes. Science **301:** 829-832.

DAUBIN, V., and G. PERRIÈRE, 2003 G+C3 structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. **20:** 471-483.

DAVIES, J., 1996 Origins and evolution of antibiotic resistance. Microbiologia **12:** 9-16.

DAVISON, J., 1999 Genetic exchange between bacteria in the environment. Plasmid **42:** 73-91.

DE MASSY, B., S. BEJAR, J. LOUARN, J. M. LOUARN and J. P. BOUCHE, 1987 Inhibition of replication forks exiting the terminus region of the *Escherichia coli* chromosome occurs at two loci separated by 5 min. Proc. Natl. Acad. Sci., USA **84:** 1759-1763.

DENG, S., R. A. STEIN and N. P. HIGGINS, 2004 Transcription-induced barriers to supercoil diffusion in the *Salmonella typhimurium* chromosome. Proc. Natl. Acad. Sci., USA **101:** 3398-3403.

DOOLITTLE, R. F., D. F. FENG, K. L. ANDERSON and M. R. ALBERRO, 1990 A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J. Mol. Evol. **31:** 383-388.

DOOLITTLE, R. F., and J. HANDY, 1998 Evolutionary anomalies among the aminoacyl-tRNA synthetases. Curr. Opin. Genet. Dev. **8:** 630-636.

DOOLITTLE, W. F., 1999 Phylogenetic classification and the universal tree. Science **284:** 2124-2129.

DOOLITTLE, W. F., Y. BOUCHER, C. L. NESBO, C. J. DOUADY, J. O. ANDERSSON *et al.*, 2003 How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos. Trans. R. Soc. Lond. B Biol. Sci. **358:** 39-57.

DWORKIN, J., and R. LOSICK, 2001 Differential gene expression governed by chromosomal spatial asymmetry. Cell **107:** 339-346.

DWORKIN, J., and R. LOSICK, 2002 Does RNA polymerase help drive chromosome segregation in bacteria? Proceedings of the National Academy of Sciences of the United States of America **99:** 14089-14094.

DYKHUIZEN, D. E., and L. GREEN, 1991 Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. **173:** 7257-7268.

ED. APPLEMAN, P., 1979 *Darwin*. W.W. Norton & Company, New York.

EGGLESTON, A. K., and S. C. WEST, 1997 Recombination initiation: easy as A, B, C, D. chi? Curr. Biol. **7:** R745-749.

EISEN, J. A., J. F. HEIDELBERG, O. WHITE and S. L. SALZBERG, 2000 Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1:** 1-11.

EISEN, J. A., J. F. HEIDELBERG, O. WHITE and S. L. SALZBERG, 2000 Evidence for symmetric chromosomal inversions around the replication origin in bacteria.[see comment]. Genome Biology **1:** RESEARCH0011.

EL KAROUI, M., V. BIAUDET, S. SCHBATH and A. GRUSS, 1999 Characteristics of Chi distribution on different bacterial genomes. Res. Microbiol. **150:** 579-587.

ELLIS, H. M., D. YU, T. DITIZIO and D. L. COURT, 2001 High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. Proc. Natl. Acad. Sci., USA **98:** 6742-6746.

FEIL, E. J., E. C. HOLMES, D. E. BESSEN, M. S. CHAN, N. P. DAY *et al.*, 2001 Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci., USA **98:** 182-187.

FEKETE, R. A., and D. K. CHATTORAJ, 2005 A *cis*-acting sequence involved in chromosome segregation in *Escherichia coli*. Mol. Microbiol. **55:** 175-183.

FITZ-GIBBON, S. T., and C. H. HOUSE, 1999 Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. **27:** 4218-4222.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE, R. A. CLAYTON, E. F. KIRKNESS *et al.*, 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:** 496-512.

FOSTER, J., M. GANATRA, I. KAMAL, J. WARE, K. MAKAROVA *et al.*, 2005 The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. PLoS Biol. **3:** e121.

GALTIER, N., M. GOUY and C. GAUTIER, 1996 SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci **12:** 543-548.

GARCIA-RUSSELL, N., T. G. HARMON, T. Q. LE, N. H. AMALADAS, R. D. MATHEWSON *et al.*, 2004 Unequal access of chromosomal regions to each other in *Salmonella*: probing chromosome structure with phage lambda integrase-mediated long-range rearrangements. Mol. Microbiol. **52:** 329-344.

GARCIA-VALLVE, S., A. ROMEU and J. PALAU, 2000 Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. **10:** 1719-1725.

GARCIA-VALLVE, S., A. ROMEU and J. PALAU, 2000 Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. Mol. Biol. Evol. **17:** 352-361.

GARNER, E. C., C. S. CAMPBELL and R. D. MULLINS, 2004 Dynamic instability in a DNA-segregating prokaryotic acting homolo. Science **306:** 1021-1025.

GERDES, K., J. MOLLER-JENSEN and R. B. JENSEN, 2000 Plasmid and chromosome partitioning: surprises from phylogeny. Molecular Microbiology **37:** 455-466.

GEVERS, D., P. DAWYNDT, P. VANDAMME, A. WILLEMS, M. VANCANNEYT *et al.*, 2006 Stepping stones towards a new prokaryotic taxonomy. Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences **361:** 1911-1916.

GHOSH, P., A. I. KIM and G. F. HATFULL, 2003 The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of *attP* and *attB*. Mol. Cell **12:** 1101-1111.

GITAI, Z., M. THANBICHLER and L. SHAPIRO, 2005 The choreographed dynamics of bacterial chromosomes. Trends Microbiol. **13:** 221-228.

GOGARTEN, J. P., 1995 The early evolution of cellular life. Trends in Ecology and Evolution **10:** 147-151.

GOGARTEN, J. P., W. F. DOOLITTLE and J. G. LAWRENCE, 2002 Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. **19:** 2226-2238.

GOGARTEN, J. P., J. FICHMANN, Y. BRAUN, L. MORGAN, P. STYLES *et al.*, 1992 The use of antisense mRNA to inhibit the tonoplast H+ ATPase in carrot. Plant Cell **4:** 851-864.

GOGARTEN, J. P., E. HILARIO and L. OLENDZENSKI, 1996 Gene duplications and horizontal gene transfer during early evolution, pp. 267-292 in *Symposium of the Society for General Microbiology; Evolution of microbial life*, edited by D. M. ROBERTS, P. SHARP, G. ALDERSON and M. A. COLLINS. Cambridge University Press, Cambridge.

GOGARTEN, J. P., and L. OLENDZENSKI, 1999 Orthologs, paralogs and genome comparisons. Curr Opin Genet Dev **9:** 630-636.

GORDON, D. M., 2001 Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. Microbiology **147:** 1079-1085.

GORDON, D. M., S. BAUER and J. R. JOHNSON, 2002 The genetic structure of *Escherichia coli* populations in primary and secondary habitats. Microbiology **148:** 1513-1522.

GORDON, G. S., and A. WRIGHT, 2000 DNA segregation in bacteria. Annual Review of Microbiology **54:** 681-708.

GRANOVETTER, M. S., 1973 The strength of weak ties. American Journal of Sociology **78:** 1360-1380.

GRIGORIEV, A., 1998 Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. **26:** 2286-2290.

GUINDON, S., and O. GASCUEL, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:** 696-704.

GUTTMAN, D. S., 1997 Recombination and clonality in natural populations of *Escherichia coli*. Trends Ecol. Evol. **12:** 16-22.

HALL, R. M., 1997 Mobile gene cassettes and integrons: moving antibiotic resistance genes in gram-negative bacteria. Ciba Found. Symp. **207:** 192-202.

HALLIN, P. F., and D. W. USSERY, 2004 CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. Bioinformatics **20:** 3682-3686.

HAMADY, M., M. D. BETTERTON and R. KNIGHT, 2006 Using the nucleotide substitution rate matrix to detect horizontal gene transfer. BMC Bioinformatics **7:** 476.

HARRY, E. J., 2001 Bacterial cell division: regulating Z-ring formation. Molecular Microbiology **40:** 795-803.

HAYES, F., and D. BARILLA, 2006 Assembling the bacterial segrosome. Trends Biochem. Sci. **31:** 247-250.

HAYES, F., and D. BARILLA, 2006 The bacterial segrosome: a dynamic nucleoprotein machine for DNA trafficking and segregation. Nat. Rev. Microbiol. **4:** 133-143.

HAYES, W. S., and M. BORODOVSKY, 1998 How to interpret an anonymous bacterial genome: machine learning approach to gene identification. Genome Res. **8:** 1154-1171.

HENDRICKSON, H., and J. G. LAWRENCE, 2006 Selection for chromosome architecture in bacteria. J. Mol. Evol. **62:** 615-629.

HENDRICKSON, H., and J. G. LAWRENCE, 2007 Mutational bias suggests that replication termination occurs near the *dif* site, not at Ter sites. Molecular Microbiology **64:** 42-56.

HENDRICKSON, H., E. S. SLECHTA, U. BERGTHORSSON, D. I. ANDERSSON and J. R. ROTH, 2002 Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. PNAS **99:** 2164-2169.

HIDAKA, M., M. AKIYAMA and T. HORIUCHI, 1988 A consensus sequence of three DNA replication terminus sites on the *E. coli* chromosome is highly homologous to the *terR* sites of the R6K plasmid. Cell **55:** 467-475.

HIDAKA, M., T. KOBAYASHI, Y. ISHIMI, M. SEKI, T. ENOMOTO *et al.*, 1992 Termination complex in *Escherichia coli* inhibits SV40 DNA replication in vitro by impeding the action of T antigen helicase. J. Biol. Chem. **267:** 5361-5365.

HIDAKA, M., T. KOBAYASHI, S. TAKENAKA, H. TAKEYA and T. HORIUCHI, 1989 Purification of a DNA replication terminus (ter) site-binding protein in *Escherichia coli* and identification of the structural gene. J. Biol. Chem. **264:** 21031-21037.

HIGGINS, N. P., X. YANG, Q. FU and J. R. ROTH, 1996 Surveying a supercoil domain by using the gamma delta resolution system in *Salmonella typhimurium*. J Bacteriol **178:** 2825-2835.

HILL, T. M., 1992 Arrest of bacterial DNA replication. Annu. Rev. Microbiol. **46:** 603-633.

HILL, T. M., J. M. HENSON and P. L. KUEMPEL, 1987 The terminus region of the *Escherichia coli* chromosome contains two separate loci that exhibit polar inhibition of replication. Proc. Natl. Acad. Sci., USA **84:** 1754-1758.

HILL, T. M., and K. J. MARIANS, 1990 *Escherichia coli* Tus protein acts to arrest the progression of DNA replication forks *in vitro*. Proc. Natl. Acad. Sci., USA **87:** 2481-2485.

HILL, T. M., A. J. PELLETIER, M. L. TECKLENBURG and P. L. KUEMPEL, 1988 Identification of the DNA sequence from the *E. coli* terminus region that halts replication forks. Cell **55:** 459-466.

HOLMES, V. F., and N. R. COZZARELLI, 2000 Closing the ring: links between SMC proteins and chromosome partitioning, condensation, and supercoiling. Proc. Natl. Acad. Sci., USA **97:** 1322-1324.

HORIUCHI, T., Y. FUJIMURA, H. NISHITANI, T. KOBAYASHI and M. HIDAKA, 1994 The DNA replication fork blocked at the Ter site may be an entrance for the RecBCD enzyme into duplex DNA. J. Bacteriol. **176:** 4656-4663.

HORIUCHI, T., and M. HIDAKA, 1988 Core sequence of two separable terminus sites of the R6K plasmid that exhibit polar inhibition of replication is a 20 bp inverted repeat. Cell **54:** 515-523.

HORWITZ, M. S., and L. A. LOEB, 1986 Promoters selected from random DNA sequences. PNAS **83:** 7405-7409.

IBBA, M., J. L. BONO, P. A. ROSA and D. SOLL, 1997 Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. Proc. Natl. Acad. Sci., USA **94:** 14383-14388.

IP, S. C., M. BREGU, F. X. BARRE and D. J. SHERRATT, 2003 Decatenation of DNA circles by FtsK-dependent Xer site-specific recombination. EMBO J. **22:** 6399-6407.

JACCARD, P., 1912 The distribuition of flora in the alpine zone. The New Phytologist **11:** 37-50.

JAIN, R., M. C. RIVERA and J. A. LAKE, 1999 Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci., USA **96:** 3801-3806.

JORDAN, I. K., K. S. MAKAROVA, J. L. SPOUGE, Y. I. WOLF and E. V. KOONIN, 2001 Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. **11:** 555-565.

JUN, S., and B. MULDER, 2006 Entropy-driven spatioal organization of highly confined polymers: lessons for the bacterial chromosome. PNAS USA **103:** 12388-12393.

KAGUNI, J. M., 2006 DnaA: Controlling the Initiation of Bacterial DNA Replication and More. Annu. Rev. Microbiol. **60:** 351-375.

KARLIN, S., A. M. CAMPBELL and J. MRÁZEK, 1998 Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32:** 185-225.

KATO, J., 2005 Regulatory network of the initiation of chromosomal replication in *Escherichia coli*. Crit. Rev. Biochem. Mol. Biol. **40:** 331-342.

KELLY, C. D., RAHN, O., 1931 The Growth Rate of Individual Bacterial Cells. Journal of Bacteriology **23:** 147-153.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base subsitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111-120.

KIMURA, M., 1981 Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78:** 454-458.

KIMURA, M., 1983 *The Neutral Allele Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

KLECKNER, N., J. BENDER and S. GOTTESMAN, 1991 Uses of transposons with emphasis on Tn*10*. Methods Enzymol. **204:** 139-180.

KLENK, H. P., R. A. CLAYTON, J. F. TOMB, O. WHITE, K. E. NELSON *et al.*, 1997 The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature **390:** 364-370.

KLENK, H. P., T. D. MEIER, P. DUROVIC, V. SCHWASS, F. LOTTSPEICH *et al.*, 1999 RNA polymerase of *Aquifex pyrophilus*: implications for the evolution of the bacterial *rpoBC* operon and extremely thermophilic bacteria. J. Mol. Evol. **48:** 528-541.

KOBAYASHI, T., M. HIDAKA and T. HORIUCHI, 1989 Evidence of a ter specific binding protein essential for the termination reaction of DNA replication in *Escherichia coli*. EMBO J **8:** 2435-2441.

KOGOMA, T., 1997 Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. Microbiol. Mol. Biol. Rev. **61:** 212-238.

KOGOMA, T., G. W. CADWELL, K. G. BARNARD and T. ASAI, 1996 The DNA replication priming protein, PriA, is required for homologous recombination and double-strand break repair. J. Bacteriol. **178:** 1258-1264.

KONSTANTINIDIS, K. T., A. RAMETTE and J. M. TIEDJE, 2006 The bacterial species definition in the genomic era. Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences **361:** 1929-1940.

KOONIN, E. V., 2000 How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. **1:** 99-116.

KOONIN, E. V., K. S. MAKAROVA and L. ARAVIND, 2001 Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. **55:** 709-742.

KOSKI, L. B., R. A. MORTON and G. B. GOLDING, 2001 Codon bias and base composition are poor indicators of horizontally transferred genes. Mol. Biol. Evol. **18:** 404-412.

KOWALCZYKOWSKI, S. C., D. A. DIXON, A. K. EGGLESTON, S. D. LAUDER and W. M. REHRAUER, 1994 Biochemistry of homologous recombination in *Escherichia coli*. Microbiol. Rev. **58:** 401-465.

KRAWIEC, S., and M. RILEY, 1990 Organization of the bacterial chromosome. Microbiol. Rev. **54:** 502-539.

KREUZER, K. N., 2005 Interplay between DNA replication and recombination in prokaryotes. Annu. Rev. Microbiol. **59:** 43-67.

KUEMPEL, P. L., S. A. DUERR and N. R. SEELEY, 1977 Terminus region of the chromosome in *Escherichia coli* inhibits replication forks. Proc. Natl. Acad. Sci., USA **74:** 3927-3931.

KUNIN, V., and C. A. OUZOUNIS, 2003 The balance of driving forces during genome evolution in prokaryotes. Genome Res. **13:** 1589-1594.

KURLAND, C. G., 2000 Something for everyone. EMBO Rep. **1:** 92-95.

KUZMINOV, A., 1995 Collapse and repair of replication forks in *Escherichia coli*. Mol. Microbiol. **16:** 373-384.

KUZMINOV, A., 1999 Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. Microbiol. Mol. Biol. Rev. **63:** 751-813.

LARK, K. G., and C. A. LARK, 1979 *recA*-dependent DNA replication in the absence of protein synthesis: characteristics of a dominant lethal replication mutation, *dnaT*, and requirement for *recA*$^+$ function. Cold Spring Harb. Symp. Quant. Biol. **43 Pt 1:** 537-549.

LAU, I. F., S. R. FILIPE, B. SOBALLE, O. A. OKSTAD, F. X. BARRE *et al.*, 2003 Spatial and temporal organization of replicating *Escherichia coli* chromosomes. Mol. Microbiol. **49:** 731-743.

LAWRENCE, J. G., 1997 Selfish operons and speciation by gene transfer. Trends Microbiol. **5:** 355-359.

LAWRENCE, J. G., 1999 Gene transfer and minimal genome size., pp. 32-38 in *Size Limits of Very Small Organisms*, edited by A. KNOLL, M. J. OSBORN, J. BAROSS, H. BERG, N. R. PACE *et al.* National Research Council, Washington DC.

LAWRENCE, J. G., 1999 Selfish operons: the evolutionary impact of gene clustering in the prokaryotes and eukaryotes. Curr. Op. Genet. Dev. **9:** 642-648.

LAWRENCE, J. G., 2000 Clustering of antibiotic resistance genes: Beyond the selfish operon. ASM News **66:** 281-286.

LAWRENCE, J. G., 2001 Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom. Syst. Biol. **50:** 479-496.

LAWRENCE, J. G., 2002 Gene transfer in bacteria: Speciation without species? Theor. Pop. Biol. **61:** 449-460.

LAWRENCE, J. G., G. F. HATFULL and R. W. HENDRIX, 2002 Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J. Bacteriol. **184:** 4891-4905.

LAWRENCE, J. G., and H. HENDRICKSON, 2003 Lateral gene transfer: when will adolescence end? Mol. Microbiol. **50:** 739-749.

LAWRENCE, J. G., and H. HENDRICKSON, 2004 Chromosome structure and constraints on lateral gene transfer. Dev. Genet. **2004:** 319-336.

LAWRENCE, J. G., R. W. HENDRIX and S. CASJENS, 2001 Where are the pseudogenes in bacterial genomes? Trends Microbiol. **9:** 535-540.

LAWRENCE, J. G., and H. OCHMAN, 1997 Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:** 383-397.

LAWRENCE, J. G., and H. OCHMAN, 1998 Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci., USA **95:** 9413-9417.

LAWRENCE, J. G., and H. OCHMAN, 2002 Reconciling the many faces of gene transfer. Trends Microbiol. **10:** 1-4.

LAWRENCE, J. G., and J. R. ROTH, 1996 Selfish operons: Horizontal transfer may drive the evolution of gene clusters. Genetics **143:** 1843-1860.

LAWRENCE, J. G., and J. R. ROTH, 1999 Genomic flux: genome evolution by gene loss and acquisition, pp. 263-289 in *Organization of the Prokaryotic Genome*, edited by R. L. CHARLEBOIS. ASM Press, Washington, D.C.

LEONARD, A. C., and J. E. GRIMWADE, 2005 Building a bacterial orisome: emergence of new regulatory features for replication origin unwinding. Mol. Microbiol. **55:** 978-985.

LERAT, E., V. DAUBIN, H. OCHMAN and N. A. MORAN, 2005 Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. **3:** e130.

LEVIN, B., 1981 Periodic selection, infectious gene exchange, and the genetic structure of *E. coli* populations. Genetics **99:** 1-23.

LEVY, O., J. L. PTACIN, P. J. PEASE, J. GORE, M. B. EISEN *et al.*, 2005 Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase. Proc. Natl. Acad. Sci., USA.

LI, Y., B. YOUNGREN, K. SERGUEEV and S. AUSTIN, 2003 Segregation of the *Escherichia coli* chromosome terminus. Mol. Microbiol. **50:** 825-834.

LIU, S. L., and K. E. SANDERSON, 1995 The chromosome of *Salmonella paratyphi* A is inverted by recombination between *rrnH* and *rrnG*. J. Bacteriol. **177:** 6585-6592.

LIU, S. L., and K. E. SANDERSON, 1995 Rearrangements in the genome of the bacterium *Salmonella typhi*. Proc. Natl. Acad. Sci., USA **92:** 1018-1022.

LIU, S. L., and K. E. SANDERSON, 1996 Highly plastic chromosomal organization in *Salmonella typhi*. Proc. Natl. Acad. Sci., USA **93:** 10303-10308.

LOBRY, J. R., 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:** 660-665.

LOBRY, J. R., and J. M. LOUARN, 2003 Polarisation of prokaryotic chromosomes. Curr. Opin. Microbiol. **6:** 101-108.

LOBRY, J. R., and N. SUEOKA, 2002 Asymmetric directional mutation pressures in bacteria. Genome Biol. **3:** RESEARCH0058.0051-0058.0014.

LOGSDON, J. M., and D. M. FUGUY, 1999 *Thermotoga* heats up lateral gene transfer. Curr. Biol. **9:** R747-R751.

LOUARN, J., F. CORNET, V. FANCOIS, J. PATTE and J.-M. LOUARN, 1994 Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: possible relation to nucleoid organization. J. Bacteriol. **176:** 7524-7531.

LOUARN, J., J. PATTE and J. M. LOUARN, 1977 Evidence for a fixed termination site of chromosome replication in *Escherichia coli* K12. J. Mol. Biol. **115:** 295-314.

LOUARN, J., J. PATTE and J. M. LOUARN, 1979 Map position of the replication terminus on the *Escherichia coli* chromosome. Mol. Gen. Genet. **172:** 7-11.

LOUARN, J. M., J. LOUARN, V. FRANCOIS and J. PATTE, 1991 Analysis and possible role of hyperrecombination in the termination region of the *Escherichia coli* chromosome. J. Bacteriol. **173:** 5097-5104.

LUDWIG, W., O. STRUNK, S. KLUGBAUER, N. KLUGBAUER, M. WEIZENEGGER *et al.*, 1998 Bacterial phylogeny based on comparative sequence analysis. Electrophoresis **19:** 554-568.

LYNCH, M., M. O'HELY, B. WALSH and A. FORCE, 2001 The probability of preservation of a newly arisen gene duplicate. Genetics **159:** 1789-1804.

MACKIEWICZ, P., D. MACKIEWICZ, M. KOWALCZUK and S. CEBRAT, 2001 Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol **2:** INTERACTIONS1004.

MACKIEWICZ, P., J. ZAKRZEWSKA-CZERWINSKA, A. ZAWILAK, M. R. DUDEK and S. CEBRAT, 2004 Where does bacterial replication start? Rules for predicting the *oriC* region. Nucleic Acids Res. **32:** 3781-3791.

MAGEE, T. R., T. ASAI, D. MALKA and T. KOGOMA, 1992 DNA damage-inducible origins of DNA replication in *Escherichia coli*. EMBO J **11:** 4219-4225.

MAHAN, M. J., and J. R. ROTH, 1991 Ability of a bacterial chromosome segment to invert is dictated by included material rather than flanking sequence. Genetics **129:** 1021-1032.

MAISNIER-PATIN, S., K. NORDSTROM and S. DASGUPTA, 2001 RecA-mediated rescue of *Escherichia coli* strains with replication forks arrested at the terminus. J. Bacteriol. **183:** 6065-6073.

MAJEWSKI, J., and F. M. COHAN, 1999 DNA sequence similarity requirements for interspecific recombination in *Bacillus*. Genetics **153:** 1525-1533.

MAJEWSKI, J., P. ZAWADZKI, P. PICKERILL, F. M. COHAN and C. G. DOWSON, 2000 Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J. Bacteriol. **182:** 1016-1023.

MAKAROVA, K. S., L. ARAVIND, M. Y. GALPERIN, N. V. GRISHIN, R. L. TATUSOV *et al.*, 1999 Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res. **9:** 608-628.

MAKAROVA, K. S., V. A. PONOMAREV and E. V. KOONIN, 2001 Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome Biol. **2:** 1-14.

MASAI, H., T. ASAI, Y. KUBOTA, K. ARAI and T. KOGOMA, 1994 Escherichia coli PriA protein is essential for inducible and constitutive stable DNA replication. EMBO J **13:** 5338-5345.

MASSEY, T. H., L. AUSSEL, F. X. BARRE and D. J. SHERRATT, 2004 Asymmetric activation of Xer site-specific recombination by FtsK. EMBO Rep. **5:** 399-404.

MASSEY, T. H., C. P. MERCOGLIANO, J. YATES, D. J. SHERRATT and J. LOWE, 2006 Double-stranded DNA translocation: structure and mechanism of hexameric FtsK. Molecular Cell **23:** 457-469.

MASTERS, M., and P. BRODA, 1971 Evidence for the bidirectional replications of the *Escherichia coli* chromosome. Nat. New Biol. **232:** 137-140.

MÉDIGUE, C., T. ROUXEL, P. VIGIER, A. HÉNAUT and A. DANCHIN, 1991 Evidence of horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222:** 851-856.

MEIJER, M., E. BECK, F. G. HANSEN, H. E. BERGMANS, W. MESSER *et al.*, 1979 Nucleotide sequence of the origin of replication of the *Escherichia coli* K-12 chromosome. Proc. Natl. Acad. Sci., USA **76:** 580-584.

MIESEL, L., and J. R. ROTH, 1996 Evidence that SbcB and RecF pathway functions contribute to RecBCD- dependent transductional recombination. J Bacteriol **178:** 3146-3155.

MIRA, A., H. OCHMAN and N. A. MORAN, 2001 Deletional bias and the evolution of bacterial genomes. Trends Genet. **17:** 589-596.

MIRKIN, B. G., T. I. FENNER, M. Y. GALPERIN and E. V. KOONIN, 2003 Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. **3:** 2.

MOLLER-JENSEN, J., and K. GERDES, 2004 Dynamic instability of a bacterial engine. Science **306:** 987-988.

MORAN, N. A., and J. J. WERNEGREEN, 2000 Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol. Evol. **15:** 321-326.

MORENO, E., 1997 In search of a bacterial species definition. Rev. Biol. Trop. **45:** 753-771.

MULCAIR, M. D., P. M. SCHAEFFER, A. J. OAKLEY, H. F. CROSS, C. NEYLON *et al.*, 2006 A molecular mousetrap determines polarity of termination of DNA replication in *E. coli*. Cell **125:** 1309-1319.

MULLER-HILL, B., 1996 *The lac Operon: A Short History of a Genetic Paradigm*. Walter de Gruyter, Berlin.

MULUGU, S., A. POTNIS, SHAMSUZZAMAN, J. TAYLOR, K. ALEXANDER *et al.*, 2001 Mechanism of termination of DNA replication of *Escherichia coli* involves helicase-contrahelicase interaction. Proc. Natl. Acad. Sci., USA **98:** 9569-9574.

MYERS, R. S., and F. W. STAHL, 1994 Chi and the RecBC D enzyme of Escherichia coli. Annu Rev Genet **28:** 49-70.

MYLVAGANAM, S., and P. P. DENNIS, 1992 Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics **130:** 399-410.

NEEDLEMAN, S. B., and C. D. WUNSCH, 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:** 443-453.

NELSON, K. E., R. A. CLAYTON, S. R. GILL, M. L. GWINN, R. J. DODSON *et al.*, 1999 Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399:** 323-329.

NESBO, C. L., S. L'HARIDON, K. O. STETTER and W. F. DOOLITTLE, 2001 Phylogenetic analyses of two "Archaeal" genes in *Thermotoga maritima* reveal multiple transfers between Archaea and Bacteria. Mol. Biol. Evol. **18:** 362-375.

NEYLON, C., A. V. KRALICEK, T. M. HILL and N. E. DIXON, 2005 Replication termination in *Escherichia coli:* structure and antihelicase activity of the Tus-Ter complex. Microbiol. Mol. Biol. Rev. **69:** 501-526.

NIELSEN, H. J., LI, Y., YOUNGREN, B., HANSEN, F.G., AUSTIN, S., 2006 Progressive segregation of the *Escherichia coli* chromosome. Molecular Microbiology **61:** 383-393.

NIH, 2007 Entrez Genome, pp.

NIKI, H., Y. YAMAICHI and S. HIRAGA, 2000 Dynamic organization of chromosomal DNA in *Escherichia coli*. Genes Dev. **14:** 212-223.

NORDMAN, J., O. SKOVGAARD and A. WRIGHT, 2007 A novel class of mutations that affect DNA replication in *E. coli*. Molecular Microbiology **64:** 125-138.

OCHMAN, H., and I. B. JONES, 2000 Evolutionary dynamics of full genome content in *Escherichia coli*. EMBO J. **19:** 6637-6643.

OCHMAN, H., and J. G. LAWRENCE, 1996 Phylogenetics and the amelioration of bacterial genomes, pp. 2627-2637 in *Escherichia coli and Salmonella typhimurium: Cellular and molecular biology, 2nd edition*, edited by F. C. NEIDHARDT, R. CURTISS III, J. L. INGRAHAM, E. C. C. LIN, K. B. LOW *et al.* American Society for Microbiology, Washington, D.C.

OCHMAN, H., J. G. LAWRENCE and E. GROISMAN, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature **405:** 299-304.

OCHMAN, H., and A. C. WILSON, 1988 Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. **26:** 74-86.

OKADA, S., and D. M. GORDON, 2001 Host and geographical factors influence the thermal niche of enteric bacteria isolated from native Australian mammals. Mol. Ecol. **10:** 2499-2513.

OLENDZENSKI, L., E. HILARIO and J. P. GOGARTEN, 1998 Horizontal Gene Transfer and Fusing Lines of Descent: the Archaebacteria - a Chimera? pp. 349-362 in *Horizontal Gene Transfer*, edited by M. SYVANEN and C. KADO. Chapman and Hall, London.

OMELCHENKO, M. V., K. S. MAKAROVA, Y. I. WOLF, I. B. ROGOZIN and E. V. KOONIN, 2003 Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*. Genome Biol. **4:** R55.

PAGE, R. D., 1996 TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:** 357-358.

PARKHILL, J., M. SEBAIHIA, A. PRESTON, L. D. MURPHY, N. THOMSON *et al.*, 2003 Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. **35:** 32-40.

PEASE, P. J., O. LEVY, G. J. COST, J. GORE, J. L. PTACIN *et al.*, 2005 Sequence-directed DNA translocation by purified FtsK. Science **307:** 586-590.

PEDULLA, M. L., M. E. FORD, J. M. HOUTZ, T. KARTHIKEYAN, C. WADSWORTH *et al.*, 2003 Origins of highly mosaic mycobacteriophage genomes. Cell **113:** 171-182.

PELLETIER, A. J., T. M. HILL and P. L. KUEMPEL, 1988 Location of sites that inhibit progression of replication forks in the terminus region of *Escherichia coli*. J. Bacteriol. **170:** 4293-4298.

PERALS, K., H. CAPIAUX, J. B. VINCOURT, J. M. LOUARN, D. J. SHERRATT *et al.*, 2001 Interplay between recombination, cell division and chromosome structure during chromosome dimer resolution in *Escherichia coli*. Mol. Microbiol. **39:** 904-913.

PÉRALS, K., F. CORNET, Y. MERLET, I. DELON and J. M. LOUARN, 2000 Functional polarization of the *Escherichia coli* chromosome terminus: the *dif* site acts in chromosome dimer resolution only when located between long stretches of opposite polarity. Mol. Microbiol. **36:** 33-43.

PETERS, J. E., and N. L. CRAIG, 2001 Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. Genes Dev. **15:** 737-747.

PIERSON, L. S. D., and M. L. KAHN, 1987 Integration of satellite bacteriophage P4 in *Escherichia coli*. DNA sequences of the phage and host regions involved in site-specific recombination. J Mol Biol **196:** 487-496.

POPTSOVA, M., GOGARTEN, JP, 2007 The power of phylogenetic approaches to detect horizontally transferred genes. BMC Evolutionary Biology **7:** 17.

POSSOZ, C., S. R. FILIPE, I. GRAINGE and D. J. SHERRATT, 2006 Tracking of controlled *Escherichia coli* replication fork stalling and restart at repressor-bound DNA *in vivo*. EMBO J. **25:** 2596-2604.

RABSCH, W., H. ANDREWS, R. A. KINGSLEY, R. PRAGER, H. TSCHAPE *et al.*, 2002 *Salmonella enterica* serotype Typhimurium and its host-adapted variants. Infect. Immun. **70:** 2249-2255.

RAGAN, M. A., 2001 Detection of lateral gene transfer among microbial genomes. Curr. Opin. Genet. Dev. **11:** 620-626.

RAGAN, M. A., 2001 On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. **201:** 187-191.

RAPPLEYE, C. A., and J. R. ROTH, 1997 A Tn10 derivative (T-POP) for isolation of insertions with conditional (tetracycline-dependent) phenotypes. J Bacteriol **179:** 5827-5834.

RAYSSIGUIER, C., D. S. THALER and M. RADMAN, 1989 The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature **342:** 396-401.

READ, T. D., R. C. BRUNHAM, C. SHEN, S. R. GILL, J. F. HEIDELBERG *et al.*, 2000 Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. **28:** 1397-1406.

READ, T. D., S. L. SALZBERG, M. POP, M. SHUMWAY, L. UMAYAM *et al.*, 2002 Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis.[see comment]. Science **296:** 2028-2033.

REDFIELD, R. J., 1988 Evolution of bacterial transformation: is sex with dead cells ever better than no sex at all? Genetics **119:** 213-221.

REDFIELD, R. J., 1993 Evolution of natural transformation: testing the DNA repair hypothesis in Bacillus subtilis and Haemophilus influenzae. Genetics **133:** 755-761.

REDFIELD, R. J., 2001 Do bacteria have sex? Nature Reviews Genetics **2:** 634-639.

REDFIELD, R. J., M. R. SCHRAG and A. M. DEAN, 1997 The evolution of bacterial transformation: sex with poor relations. Genetics **146:** 27-38.

REEVES, P., 1993 Evolution of *Salmonella* O antigen variation by interspecific gene transfer on a large scale. Trends Genet **9:** 17-22.

ROCHA, E. P., 2004 The replication-related organization of bacterial genomes. Microbiology **150:** 1609-1627.

ROCHA, E. P., and A. DANCHIN, 2003 Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat. Genet. **34:** 377-378.

ROCHA, E. P., and A. DANCHIN, 2003 Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. **31:** 6570-6577.

ROECKLEIN, B., A. PELLETIER and P. KUEMPEL, 1991 The tus gene of *Escherichia coli*: autoregulation, analysis of flanking sequences and identification of a complementary system in *Salmonella typhimurium*. Res. Microbiol. **142:** 169-175.

ROSVALL, M., and C. T. BERGSTROM, 2007 An information-theorhetic framework for resolving structure in complex networks. PNAS **104:** 7327-7331.

ROTHFIELD, L., S. JUSTICE and J. GARCIA-LARA, 1999 Bacterial cell division. Annual Review of Genetics **33:** 423-448.

ROTHFIELD, L., S. JUSTICE and J. GARCIA-LARA, 1999 Bacterial Cell Division. Annual Review of Genetics **33:** 423-448.

SALZBERG, S. L., A. J. SALZBERG, A. R. KERLAVAGE and J. F. TOMB, 1998 Skewed oligomers and origins of replication. Gene **217:** 57-67.

SANDERSON, K. E., and S. L. LIU, 1998 Chromosomal rearrangements in enteric bacteria. Electrophoresis **19:** 569-572.

SANDLER, S. J., H. S. SAMRA and A. J. CLARK, 1996 Differential suppression of *priA*2:kan phenotypes in *Escherichia coli* K-12 by mutations in *priA, lexA,* and *dnaC*. Genetics **143:** 5-13.

SCHOULS, L. M., C. S. SCHOT and J. A. JACOBS, 2003 Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. Journal of Bacteriology **185:** 7241-7246.

SCIOCHETTI, S. A., P. J. PIGGOT and G. W. BLAKELY, 2001 Identification and characterization of the *dif* site from *Bacillus subtilis*. J. Bacteriol. **183:** 1058-1068.

SEGALL, A., M. J. MAHAN and J. R. ROTH, 1988 Rearrangement of the bacterial chromosome: forbidden inversions. Science **241:** 1314-1318.

SEGALL, A. M., and J. R. ROTH, 1989 Recombination between homologies in direct and inverse orientation in the chromosome of *Salmonella*: intervals which are nonpermissive for inversion formation. Genetics **122:** 737-747.

SERRES, M. H., and M. RILEY, 2000 MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb. Comp. Genomics **5:** 205-222.

SHARP, P. M., E. BAILES, R. J. GROCOCK, J. F. PEDEN and R. E. SOCKETT, 2005 Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Research **33:** 1141-1153.

SHERRATT, D. J., 2003 Bacterial chromosome dynamics. Science **301:** 780-785.

SHERRATT, D. J., I. F. LAU and F. X. BARRE, 2001 Chromosome segregation. Curr. Opin. Microbiol. **4:** 653-659.

SHERRATT, D. J., B. SOBALLE, F. X. BARRE, S. FILIPE, I. LAU *et al.*, 2004 Recombination and chromosome segregation. Philos. Trans. R Soc. Lond. B Biol. Sci. **359:** 61-69.

SIMMONS, M. P., C. P. RANDLE, J. V. FREUDENSTEIN and J. W. WENZEL, 2002 Limitations of relative apparent synapomorphy analysis (RASA) for measuring phylogenetic signal. Mol. Biol. Evol. **19:** 14-23.

SIVANATHAN, V., M. D. ALLEN, C. DE BEKKER, R. BAKER, L. K. ARCISZEWSKA *et al.*, 2006 The FtsK gamma domain directs oriented DNA translocation by interacting with KOPS. Nat. Struct. Mol. Biol. **13:** 965-972.

SKOKOTAS, A., M. WROBLESKI and T. M. HILL, 1994 Isolation and characterization of mutants of Tus, the replication arrest protein of *Escherichia coli*. J. Biol. Chem. **269:** 20446-20455.

SMITH, D. R., L. A. DOUCETTE-STAMM, C. DELOUGHERY, H. LEE, J. DUBOIS *et al.*, 1997 Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. J Bacteriol **179:** 7135-7155.

SMITH, G. R., 1991 Conjugational recombination in *E. coli*: myths and mechanisms. Cell **64:** 19-27.

SMITH, G. R., S. K. AMUNDSEN, P. DABERT and A. F. TAYLOR, 1995 The initiation and control of homologous recombination in *Escherichia coli*. Philos. Trans. R. Soc. Lond. B Biol. Sci. **347:** 13-20.

SMITH, M. G., T. A. GIANOULIS, S. PUKATZKI, J. J. MEKALANOS, L. N. ORNSTON *et al.*, 2007 New insights into Acinetobacter baumannii pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. Genes & Development **21:** 601-614.

SMITH, M. W., D.-W. FENG and R. F. DOOLITTLE, 1992 Evolution by acquisition: the case for horizontal gene transfers. Trends in Biochem. Sci. **17:** 489-493.

SNEL, B., P. BORK and M. HUYNEN, 1999 Genome phylogeny based on gene content. Nature Genet. **21:** 108-110.

SNEL, B., P. BORK and M. A. HUYNEN, 2002 Genomes in flux: the evolution of Archaeal and proteobacterial gene content. Genome Res. **12:** 17-25.

STALEY, J. T., 2006 The bacterial species dilemma and the genomic-phylogenetic species concept. Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences **361:** 1899-1909.

STALLIONS, D. R., and R. CURTISS, 3RD, 1971 Chromosome transfer and recombinant formation with deoxyribonucleic acid temperature-sensitive strains of *Escherichia coli*. J. Bacteriol. **105:** 886-895.

STAMBUK, S., and M. RADMAN, 1998 Mechanism and control of interspecies recombination in *Escherichia coli*. I. Mismatch repair, methylation, recombination and replication functions. Genetics **150:** 553-542.

STEIN, R. A., S. DENG and N. P. HIGGINS, 2005 Measuring chromosome dynamics on different time scales using resolvases with varying half-lives. Mol. Microbiol. **56:** 1049-1061.

STILLER, J. W., and B. D. HALL, 1999 Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol. Biol. Evol. **16:** 1270-1279.

STOLTZFUS, A., 1999 On the possibility of constructive neutral evolution. J. Mol. Evol. **49:** 169-181.

SUYAMA, M., and P. BORK, 2001 Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. **17:** 10-13.

TAMAS, I., L. KLASSON, K. NÄSLUND, A.-S. ERIKSSON, B. CANBÄCK *et al.*, 2002 Fifty million years of genomic stasis in endosymbiotic bacteria. Science **296:** 2376-2379.

TEKAIA, F., A. LAZCANO and B. DUJON, 1999 The genomic tree as revealed from whole proteome comparisons. Genome Res. **9:** 550-557.

TELEMAN, A. A., P. L. GRAUMANN, D. C. LIN, A. D. GROSSMAN and R. LOSICK, 1998 Chromosome arrangement within a bacterium. Curr. Biol. **8:** 1102-1109.

THANBICHLER, M., P. H. VIOLLIER and L. SHAPIRO, 2005 The structure and function of the bacterial chromosome. Curr. Opin. Genet. Dev. **15:** 153-162.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN and D. G. HIGGINS, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876-4882.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

TILLIER, E. R., and R. A. COLLINS, 2000 Genome rearrangement by replication-directed translocation. Nat. Genet. **26:** 195-197.

TOTH, I. K., L. PRITCHARD and P. R. BIRCH, 2006 Comparative genomics reveals what makes an enterobacterial plant pathogen. Annu. Rev. Phytopathol. **44:** 305-336.

UNO, R., Y. NAKAYAMA, K. ARAKAWA and M. TOMITA, 2000 The orientation bias of Chi sequences is a general tendency of G-rich oligomers. Gene **259:** 207-215.

VALJAVEC-GRATIAN, M., T. A. HENDERSON and T. M. HILL, 2005 Tus-mediated arrest of DNA replication in *Escherichia coli* is modulated by DNA supercoiling. Mol. Microbiol. **58:** 758-773.

VIOLLIER, P. H., and L. SHAPIRO, 2004 Spatial complexity of mechanisms controlling a bacterial cell cycle. Curr. Opin. Microbiol. **7:** 572-578.

VIOLLIER, P. H., M. THANBICHLER, P. T. MCGRATH, L. WEST, M. MEEWAN *et al.*, 2004 Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. Proc. Natl. Acad. Sci., USA **101:** 9257-9262.

VOGEL, J., P. NORMAND, J. THIOULOUSE, X. NESME and G. L. GRUNDMANN, 2003 Relationship between spatial and genetic distance in *Agrobacterium* spp. in 1 cubic centimeter of soil. Appl. Environ. Microbiol. **69:** 1482-1487.

VULIC, M., R. E. LENSKI and M. RADMAN, 1999 Mutation, recombination, and incipient speciation of bacteria in the laboratory. Proc. Natl. Acad. Sci., USA **96:** 7348-7351.

WAKE, R. G., 1997 Replication fork arrest and termination of chromosome replication in *Bacillus subtilis*. FEMS Microbiol. Lett. **153:** 247-254.

WANG, A., and J. R. ROTH, 1988 Activation of silent genes by transposons Tn*5* and Tn*10*. Genetics **120:** 875-885.

WANG, X., C. POSSOZ and D. J. SHERRATT, 2005 Dancing around the divisome: asymmetric chromosome segregation in *Escherichia coli*. Genes Dev. **19:** 2367-2377.

WEISS, D. S., 2004 Bacterial cell division and the septal ring. Molecular Microbiology **54:** 588-597.

WICK, L. M., W. QI, D. W. LACHER and T. S. WHITTAM, 2005 Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. J. Bacteriol. **187:** 1783-1791.

WOESE, C. R., 1987 Bacterial evolution. Microbiol. Rev **51:** 221-271.

WOESE, C. R., 1994 There must be a prokaryote somewhere: microbiology's search for itself. Microbiological Reviews **58:** 1-9.

WOESE, C. R., 2000 Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci., USA **97:** 8392-8396.

WOESE, C. R., and G. E. FOX, 1977 Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci., USA **74:** 5088-5090.

WOESE, C. R., G. J. OLSEN, M. IBBA and D. SOLL, 2000 Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. **64:** 202-236.

WORNING, P., L. J. JENSEN, K. E. NELSON, S. BRUNAK and D. W. USSERY, 2000 Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. Nucl. Acids Res. **28:** 706-709.

WU, L. J., 2004 Structure and segregation of the bacterial nucleoid. Curr. Opin. Genet. Dev. **14:** 126-132.

WU, L. J., and J. ERRINGTON, 1997 Septal localization of the SpoIIIE chromosome partitioning protein in *Bacillus subtilis*. EMBO **16:** 2161-2169.

WU, L. J., and J. ERRINGTON, 1998 Use of asymmetric cell division and *spoIIIE* mutants to probe chromosome orientation and organization in *Bacillus subtilis*. Mol. Microbiol. **27:** 777-786.

WU, M., Q. REN, A. S. DURKIN, S. C. DAUGHERTY, L. M. BRINKAC *et al.*, 2005 Life in hot carbon monoxide: the complete genome sequence of Carboxydothermus hydrogenoformans Z-2901.[erratum appears in PLoS Genet. 2006 Apr;2(4):e60 Note: Haft, Daniel H [added]]. PLoS Genetics **1:** e65.

YAP, W. H., Z. ZHANG and Y. WANG, 1999 Distinct types of rRNA operons exist in the genome of the Actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. **181:** 5201-5209.

YATES, J., I. ZHEKOV, R. BAKER, B. EKLUND, D. J. SHERRATT *et al.*, 2006 Dissection of a functional interaction between the DNA translocase, FtsK, and the XerD recombinase. Mol. Microbiol. **59:** 1754-1766.

ZAHRT, T. C., and S. MALOT, 1997 Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. Proc. Natl. Acad. Sci., USA **94:** 9786-9791.

ZAWADZKI, P., M. A. RILEY and F. M. COHAN, 1996 Homology among nearly all plasmids infecting three Bacillus species. J Bacteriol **178:** 191-198.

ZAWADZKI, P., M. S. ROBERTS and F. M. COHAN, 1995 The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. Genetics **140:** 917-932.

ZHANG, R., and C. T. ZHANG, 2003 Multiple replication origins of the archaeon *Halobacterium* species NRC-1. Biochem. Biophys. Res. Commun. **302:** 728-734.

ZHANG, R., and C. T. ZHANG, 2005 Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea **1:** 335-346.

ZUCKERKANDL, E., 1965 The evolution of hemoglobin. Sci. Amer. **212:** 110-118.

ZUCKERKANDL, E., and L. PAULING, 1965 Molecules as documents of evolutionary history. J. Theoret. Biol. **8:** 357-366.