DESZYCK, JOHN E.. M.S. Differential Evolution Rates Along Bacterial Chromosomes. (2024)
Directed by Dr. Louis Marie Bobay. 36 pp.

Substitution occurs at different rates at different positions around a circular bacterial chromosome. Studies have found various patterns in these differences, but all such studies to date have relied on a few genomes each from a few well-known species. This study examines every completely assembled genome in GenBank, 15,225 specimens from 329 species, and finds that there is no consistent linear correlation between substitution rates or strength of selection and position along the Ori-Ter axis, although outliers are more common near Ter.

DIFFERENTIAL EVOLUTION RATES ALONG BACTERIAL CHROMOSOMES

by

John E. Deszyck

A Thesis
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro

2024

Approved by

Dr. Louis Marie Bobay
Committee Chair

DEDICATION

To Mom, Dad and Annie. Turns out it worked. I'm back!

And to Victoria, who helped me see what's next.

APPROVAL PAGE

This thesis written by John E. Deszyck has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Louis Marie Bobay

Committee Members

Dr. Kasie Raymann

Dr. David Remington

July 18, 2024

Date of Acceptance by Committee

June 1, 2023

Date of Final Oral Examination

ACKNOWLEDGEMENTS

This work would not have been possible without the help of Dr. Kasie Raymann and Dr. Louis Marie Bobay. Kasie was my first teacher in microbiology and introduced me to the computational biology lab at UNCG. Louis Marie has been my advisor since I was still an undergrad. He helped me debug my code, write up my findings, navigate bureaucracy and has put up with more from me than he should have had to. They made me part of their lives, gave me important things to think about, and helped me find myself when I got lost. They are inspirational teachers and good people, and when I do science in the future, their way is the way I want to do it.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I: INTRODUCTION

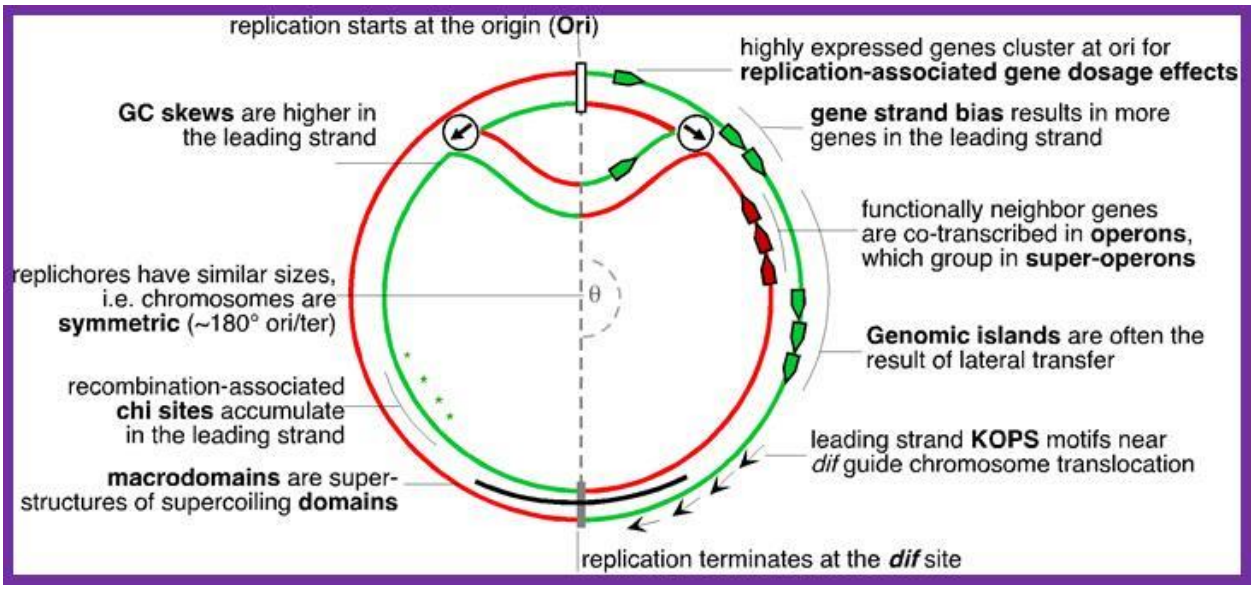
The concept of a gene is implicit but essential to the Central Dogma of molecular biology. It's also central to evolutionary theory: natural selection selects *genes* for their ability to reproduce themselves. All the genes of an organism compose its genome; and the selective pressures acting on genomes are the main focus of molecular evolution and population genomics studies. However, there are additional targets of natural selection besides genes and their regulatory sequences: the organization of bacterial genomes into chromosomes is an evolving trait that is also guided by selection and drift.

Bacterial genomes range in size from just over 100 kilobases (smaller than a large viral genome) (Moran and Bennet 2014) to 10,000 kb (larger than a small eukaryote genome) (Kuo *et al.* 2009). Although bacterial genomes are typically organized in a single circular chromosome—often supplemented by smaller additional rings of DNA called plasmids—more complex cases have been reported. *Vibrio cholerae* has two circular chromosomes, one large and one small, that replicate in synchrony (Dillon *et al.* 2018). *Streptomyces* and *Borrelia* have linear chromosomes, like eukaryotes' (Casjens 1998). Both have telomeres but their structures are evolutionary unrelated. *Thiobacillus* and *Klebsiella* have circular chromosomes and *linear* plasmids. *Agrobacterium tumefaciens'* genome is organized into a mix of circular and linear chromosomes and multiple plasmids. These examples highlight the diversity and complexity of chromosomal organizations that have evolved across bacterial lineages. However the evolutionary processes affecting the organization of bacterial genes *within* and *between* chromosomes remain poorly understood. The present study focuses on the evolutionary rate of genes relative to their organization on circular chromosomes.

Chromosome Organization

A bacterial circular chromosome has a characteristic geometry (Rocha 2008). There is a single origin of replication where the DNA is initially unwound. Two replication forks travel in opposite directions away from the origin (Fig. 1). Each fork can be pictured as a Y shape: the stem is two-stranded DNA that is pulled apart into single strands (the arms). DNA monomers are attached to the arms to form a complement for each single strand. The sugar backbone of DNA is asymmetrical, so strands have a direction to them. A double helix is a pair of oppositely oriented strands, and the replication machinery works in a single direction, so only one strand can be copied in one piece (Fig. 1). The *leading strand* is replicated continuously in the same direction as each replication fork. The other strand, the *lagging strand*, must be replicated in fragments and in the opposite direction of each replication fork (Fig. 1). The two forks, each with a leading and a lagging strand, travel in opposite directions around the chromosome until they meet, having covered approximately equal distances, at the replication terminus on the opposite side of the chromosome from the origin. The path traveled by one fork, half the chromosome, is called a replichore (Fig. 1) and each circular chromosome is therefore composed of two replichores.

Figure 1. Geometry of Chromosome Replication



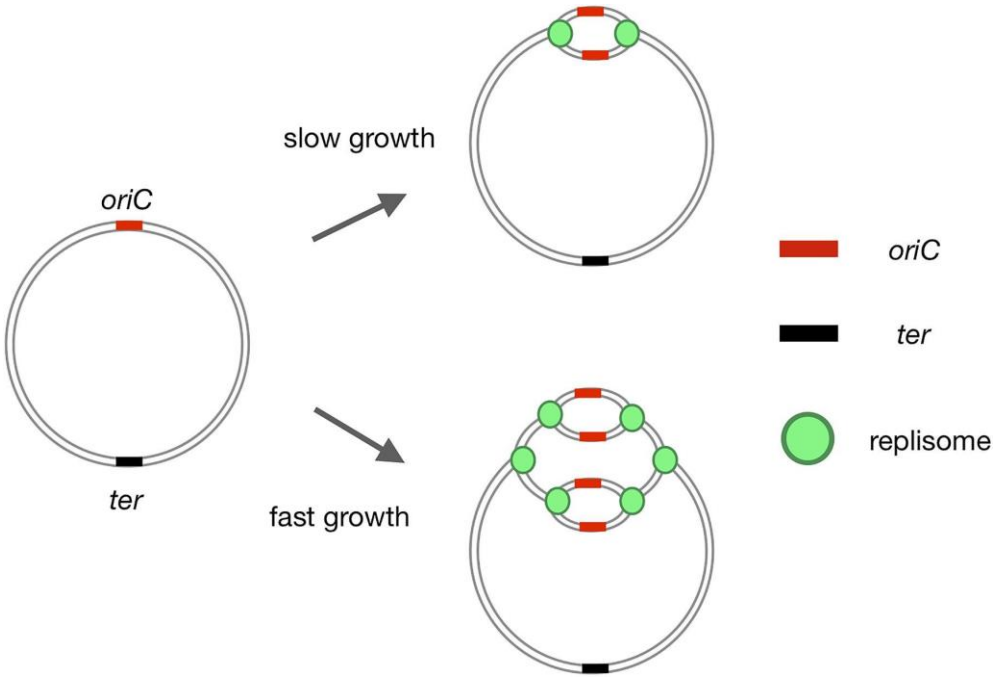
Black arrows are the replication forks. Leading strands are in green. Lagging strands are red. From Rocha 2008

A number of chromosome-structuring effects have been reported (reviewed in Rocha 2008). These effects are largely due to the mechanistic constraints affecting the chromosome and shaping its organization. A bacterial chromosome is constantly being replicated, transcribed, coiled and uncoiled, and subjected to many other processes. When chromosome structure changes through structural mutations (e.g., inversions or relocations), arrangements that favor the harmonious interplay between these processes are favored by selection (Rocha 2008).

Under optimal conditions, an *E. coli* cell takes about 20 minutes to divide, but it takes this bacterium 40 minutes to replicate its chromosome (Rocha 2008). The answer to this riddle is that bacterial cells initiate replication of their chromosomes multiple times per cell division. This means that there might be two, four, or even eight copies of the origin and nearby genes, but only one copy of the terminus. Since each copy can be transcribed during replication, genes near the origin may be expressed two, four or eight times as much as those near the terminus. This is

called the *gene dosage effect* (Fig. 2). Selection is shaping bacterial chromosomes accordingly: highly expressed genes tend to cluster near the origin (Rocha 2008).

Figure 2. Multiple Replications



To speed up cell division, a single chromosome may be in the process of several waves of replication. In the image on the bottom right, two replications are underway and 2^2 =4 copies of the origin and nearby genes are present. From Trojanowski et al. 2018

The RNA Polymerase (which performs transcription) and the DNA polymerase (which replicates DNA) must both unwind a portion of the DNA to transcribe and replicate DNA, respectively. Since both processes can occur concurrently, mechanistic conflicts may arise. Transcription and replication are typically ongoing at the same time in a bacterial cell, and therefore collisions between the replication machinery and the transcription machinery frequently occur (Rocha 2008). These collisions are thought to impact the effectiveness of transcription—and possibly replication. As a result, gene orientation relative to the replichores has been shaped by selection (Fig. 1). When the replication and the transcription machineries are

traveling in the same direction, fewer collisions are expected to occur. When the collision is head-on, the DNA polymerase complex is more likely to fall off the DNA due to more frequent collisions and replication must be restarted. By definition, DNA polymerase always travels along the leading strand. RNA polymerase may travel along either strand, leading or lagging, whichever encodes the sequence it is transcribing. Genes transcribed on the lagging strand are thus prone to conflict, and since speed and fidelity of replication are major survival traits, they are often selected against. The preferential encoding of genes on the leading strand is called strand bias, and is observed to greater or lesser degree in all bacteria (Hendrickson *et al.* 2006). For instance, 75% of the genes of *Bacillus subtilis* are encoded on the leading strand (Tillier and Collins, 2000), whereas 55% of the genes of *E. coli* are found on the leading strand (Bobay *et al.* 2013). In contrast, some genes are thought to accumulate on the lagging strand because selection would favor their migration there (Merrikh and Merrikh 2018). When DNA and RNA polymerases collide head-on and DNA polymerase falls off the DNA, the unwound DNA strands no longer stabilize each other and both are much more prone to mutation. The collision itself also strains the DNA and can cause mutations as well. The most essential genes are highly conserved and most mutations are counter-selected in these genes; those are almost never found on the lagging strand (Merrikh and Merrikh 2018). For other genes, a higher rate of mutation might be neutral or advantageous, and the latter might appear preferentially on the lagging strand. They include transcription regulators, channel proteins, and especially virulence factors and drug resistance genes.

DNA takes up a substantial amount of space in a bacterial cell. One turn of helix, 10 base pairs, is 3.4 nm long (Albers 2015). A genome like *E. coli*'s, comprising some four million bases, is between one and two centimeters in length (Blattner *et al.* 1997). *E. coli* cell's size is

two microns, 10,000 times smaller (Gangan and Athale 2017). DNA must be packed very tightly to fit inside the cell. The packing must be loose enough that replication and transcription can proceed, but it imposes serious limits on the ability of different parts of the genome to interact with each other. Bacterial genomes are partitioned into zones called *macrodomains* (Fig. 1). The *E. coli* genome, for example, has four: Ori, containing the origin of replication, Ter, containing the terminus, and Left and Right macrodomains flanking Ter. There are also two unstructured regions flanking Ori (Valens et al 2004). DNA at one location in a macrodomain can interact with the rest of the DNA in the macrodomain, but interaction with locations in other macrodomains is much more limited. DNA in the unstructured regions can interact with DNA within the same region or DNA from the macrodomain on either side (which is Ori and either Right or Left). By limiting the scope of long-distance interactions, macrodomain structure could influence the evolution of genome structure, for example by favoring local interactions and discouraging global ones.

Hendrickson and Lawrence (2006) observed that most bacteria are enriched in various DNA octamers—different sequences for each species—that appear over-represented on the leading strand and under-represented on the lagging strand, and appear far more often than randomly expected. They called these sequences AIMS (Architecture IMparting Sequences) and they have a tendency to appear with greater frequency near the terminus of replication. Hendrickson and Lawrence observed that AIMS appeared in most bacteria, and hypothesized that they appear in all bacteria below their threshold of detection. They also speculated that some AIMS are used as navigation aids for proteins that locate the terminus such as the KOPS sites. Although their function is unknown, some types of AIMS are organized into a gradient along the genome and are enriched n proportion to their distance from the terminus of

replication. One well-characterized family of AIMS are Chi sites (Crossover Hotspot Instigator), which act as triggers for DNA repair by recruiting the RecBCD complex and thereby initiate homologous recombination (Buton and Bobay 2021).

Mutation accumulation experiments have shown that some areas of the chromosome might be more prone to mutation than others (Dillon *et al.* 2018). Mutation distribution is not constant along the replichores, and may occur in waves: low near the origin, then rising four times higher, then dropping down and finally rising again near the terminus. Interestingly, the previously-mentioned study observed that these patterns of mutations are symmetrical: the two replichores show similar fluctuations in mutation patterns along the Ori-Ter axis.

Figure 3. Patterns of Base Substitution

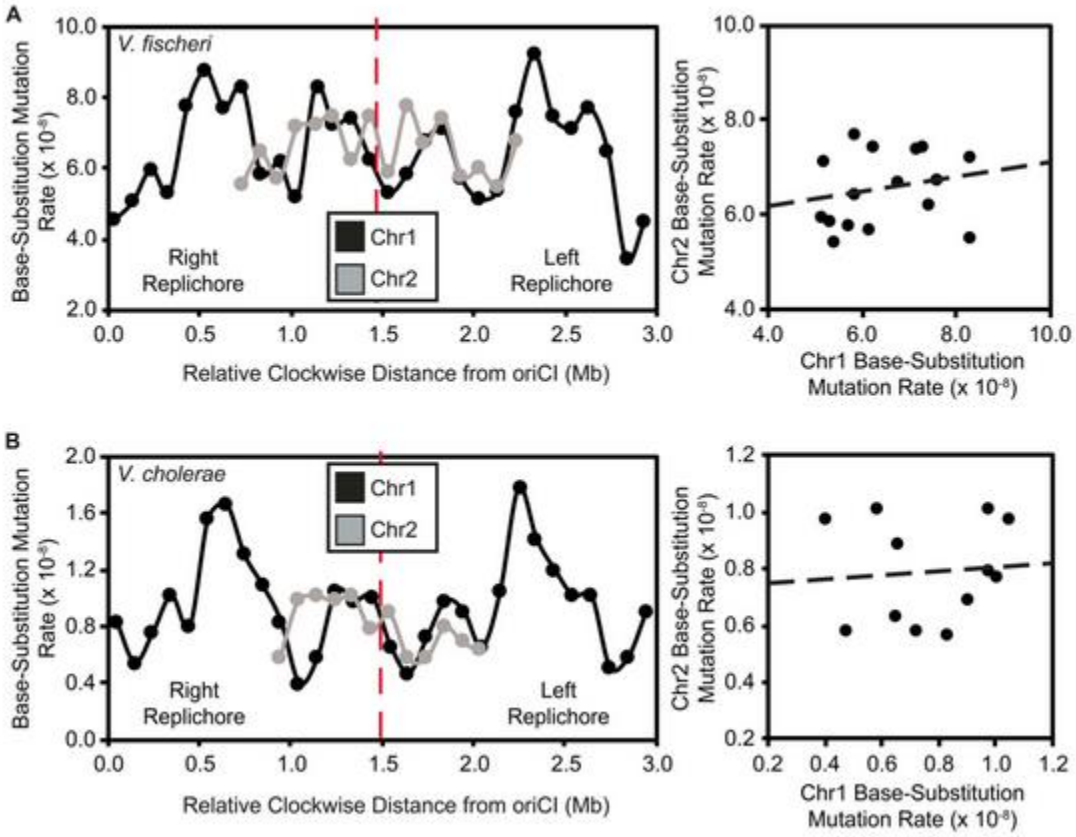


Figure 3: Base substitution patterns (bpsm) rates in 100kb intervals extending clockwise from the origin of replication of chromosome 1 and patterns of bpsm rates of concurrently replicated 100kb intervals on chromosome 2 for MMR-deficient Vibrio fischeri (A) and Vibrio cholerae (B). Patterns of bpsm rates on chr2 appear to map to those on concurrently replicated regions on chr1 in both species, but linear regressions between chromosomes are not significant in either species. From Dillon et al. 2018

In organisms with multiple chromosomes, areas of different chromosomes that are replicated at the same time also show similar mutation patterns, suggesting that this trend in mutation is related to the timing of replication. This mechanism is tightly regulated: replication is synchronized between the two replichores of the chromosome, and if multiple chromosomes are present, the chromosomes initiate replication at different times, so that all replichores finish replication simultaneously. The mutation patterns putatively arising from this timing mechanism

have implications for genome architecture. Since a gene's distance from the origin, and therefore its position in the wave-like pattern of mutation accumulation, changes very slowly, over time some chromosomal regions and the genes they encode will accumulate mutations faster than others. It is possible that genes with the same replication timing -- that is, at the same point on different replichores -- might be subjected to similar mutation rates and exhibit similar patterns of evolution.

In *E. coli*, recombination is reduced near the terminus (Touchon *et al.* 2009). GC content was 2% lower over the same area and both synonymous and non-synonymous substitutions were double the average over the rest of the chromosome. Touchon *et al.* rejected an elevated mutation rate as the explanation because inter-genome differences were low. Their explanation was background selection: When recombination rates are low, slightly deleterious mutations accumulate faster in the gene pool (non-synonymous mutations are usually deleterious), and these deleterious mutations will interfere with nearby mutations with neutral or beneficial fitness costs. The overall result of this reduction of recombination is a reduction in the efficiency of selection. To explain the lower GC content at the terminus Touchon *et al.* offered two possible explanations. For the first, they noted that most mutations are from GC to AT and that most such mutations are more frequently deleterious. Elsewhere on the chromosome, higher rates of recombination should increase the efficiency of selection, which would favor G/C mutations. So the lower GC content near the terminus would also be the consequence of the reduction of recombination in this region. As a result, the terminus region is relatively AT-enriched. The other explanation is that recombination itself favors mutation repairs from AT to GC (the biased gene conversion hypothesis). Under this hypothesis, GC content at the terminus is lower because recombination doesn't happen there as much.

Touchon *et al.* also propose explanations for why recombination is less frequent at the terminus. One is that fast-replicating cells contain many copies of origin-proximal DNA, so that there are more opportunities for recombination to happen there. Another possibility is that highly expressed genes near the origin are under stronger purifying selection. Observed mutation rates will be lower because most mutations will be strongly deleterious and will be weeded out. Observed rates of recombination may be higher if recombination preferably repairs critically damaged genes. Both trends would favor higher recombination rates. Finally, they suggest that reduced recombination near the terminus is due to the Ter macrodomain, which might act as a recombination insulator. But this last explanation is incomplete. There are three other macrodomains (Touchon *et al.* mention them by name), comprising more than half the chromosome, where recombination is *not* restricted. It is therefore unclear whyTer would prevent recombination. Regardless of its cause, the recombination-depleted, additionally-mutated, AT-enriched region around the terminus is a major feature of the chromosome of *E. coli*. Whether or not these patterns are general features across bacterial lineages remains to be determined.

In general, genes become less conserved with greater distance from the origin of replication (Rocha and Dauchin 2004; Couturier and Rocha 2006). Substitution rates, as measured by *dN*, *dS* and *dN/dS* all increase near the terminus. But this finding rests on observations of only a few species. Cooper et al. (2010) showed it to be true for *Burkholderia*, *Vibrio, Bordetella* and *Xanthomonas*; Morrow and Cooper provided more on *Xanthomonas* in 2012. Sharp et al. (1989) and Flynn et al. (2010) have suggested that genes near the terminus are more prone to recombination, but, as noted above, Touchon et al. (2009) have demonstrated that the opposite is true in *E. coli*.

Different species of bacteria have different nonrandom patterns of substitution around the chromosome (Sharp 1989; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012; Dillon et al. 2015). Some studies found no correlation between distance from the origin and substitution rates. Others found that substitutions were most frequent at intermediate points between the origin and terminus of replication (Ochman 2003). More elaborate patterns have also been detected. Dillon et al (2018), as noted above, found wavelike patterns of mutation rate that were synchronized across the multiple chromosomes of *Burkholderia* and *Vibrio*. Wave patterns have also been found in *E. coli* (Long et al. 2016) and *Pseudomonas aeruginosa* (Dellman et al. 2016).

The overarching trend in which substitution rates increase with increasing distance from the origin of replication, has been established in only a few studies noted above, mostly in *E. coli* and a few other bacterial species. The number of other patterns cited here raises questions as to how prevalent this trend actually is, and the extent to which the other patterns are trends in themselves or one-off deviations from it. Furthermore, many of the studies listed above used an average of only three genomes per species (Couturier and Rocha 2006; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012).

Recently, Lato and Golding (2020) used a broader sample: *E. coli, Bacillus subtilis, Streptomyces* and *Sinorhizobium melloti*. These species present distinct ways of structuring their genetic material. The first two species have single circular chromosomes. *Streptomyces* has a linear chromosome and *S. melloti* is multirepliconic. Lato and Golding were unable to establish a consistent relationship between distance to the origin of replication and substitution rate. Correlations between these quantities were frequently statistically insignificant; when significant they were always small and varied in sign. Lato and Golding concluded widely reported

molecular trends may not be as universal as previously thought; however their study remained restricted to a small set of species and it remains difficult to conclude to which extent these patterns are universal across prokaryotic lineages.

Approach and Innovation

Lato and Golding analyzed more data than previous work, but their sample is still limited to four species and 25 genomes. Even accounting for the fact that the data requirements for this analysis are stringent – only completely assembled genomes, a few percent of the total data available on GenBank will suffice – the available dataset is larger than their sample by several orders of magnitude. Small datasets can be informative when they are carefully chosen (Rocha 2006), or when a larger analysis is computationally intractable. But *ceteris paribus*, the natural approach to determining whether observed molecular trends are universal or particular is to use all the available data. To provide a more robust assessment of the impact of chromosome architecture on the rate of gene evolution, I analyzed 15,225 fully assembled genomes from 329 species, with each species having at least five representative genomes.

There are pros and cons to this approach. Some of these bacteria have linear chromosomes, or multiple chromosomes, or are outliers in some other way. It has been necessary at points to withhold this or that portion of the data. And computing with so many genomes is costly. But the reward is worth it: the chance to see whether what has been learned about the bacterial chromosome in the last 20 years is universal or parochial, broad or narrow.

CHAPTER II: MATERIALS AND METHODS

Data

Data were downloaded from NCBI's GenBank in 2020 and comprised every completely assembled genome available at that time: 15,225 genomes from 329 species. Species were included in my analysis when they were represented by at least five genomes; the most populous three are *E. coli* (902), *S. enterica* (751), and *B. pertussis* (546).

Computations were performed on the Longleaf supercomputing cluster at the University of North Carolina at Chapel Hill. Specialist programs and version numbers are as noted, but most work was conducted in custom-written scripts in the programming language Python. The version 3.9 python environment installed on Longleaf was used throughout this project.

Core Genome Extraction

Once downloaded, genomes were partitioned into datasets by species and the core genome of each species was built with *CoreCruncher* (Harris *et al.* 2020) using the stringent option and with a protein sequence identity threshold of 90%, a minimum gene length conservation of 80%. Orthologs were defined as core when present in at least 90% of the genomes. The core genome is the set of genes common to all—or nearly all—representatives. Orthologous proteins were collected and aligned with MUSCLE (Edgar 2004) with default settings, reverse transcribed into DNA and concatenated into a single merged alignment of the core genome of each species.

dN/dS Calculation

The species concatenates were processed with PAML (Phylogenetic Analysis by Maximum Likelihood) (Yang 1997) using the *yn00* algorithm, which determines the rate of synonymous substitutions (*dS*), non-synonymous substitutions (*dN*), and the ratio of the two quantities, *dN/dS*, for each pair of gene sequences of the core genome. Synonymous mutations occur when a base in a codon changes but the amino acid the codon represents does not. In a non-synonymous mutation, the amino acid does change, as may the fitness of the organism. The relative prevalence of these two types of substitutions can be used to show the efficiency with which selection acts on a gene. Indeed, synonymous mutations are expected to have near-neutral effects on fitness, whereas non-synonymous mutations predominantly decrease fitness and are therefore predominantly lost due to purifying selection. Thus, *dS* provides an estimate of the rate of evolution that is unaffected by direct selection. The ratio *dN/dS* provides an estimate of the strength of selection acting on a gene. Output files from PAML were parsed with custom Python scripts and stored for later use, when it would be combined with a second pipeline which constructed a representation of the geometry of the bacterial chromosomes under study.

Locating the Origin and Terminus of Replication

There is relatively little standardization among the 15,225 genomes in the dataset. They were collected over a number of years by many researchers conforming to different standards, using different conventions and techniques. This dataset records strings of bases in different orientations and from different starting points, some specially chosen and some not. This project focuses on the position of each gene within its replichore, but many of the researchers who compiled this data assigned the replichore boundaries, the origin and terminus, no special role. Thus, it was necessary to construct a representation of each circular chromosome in terms of its

replicating halves. Since circularity is implied in this and subsequent calculations, organisms with linear chromosomes were discarded from the analysis at this point.

There are several methods for locating the origin and terminus of replication (Luo and Gao 2019); this project uses a simple computational approach based on GC skew. Because the lagging strand is replicated discontinuously, this strand is found single-stranded for longer periods of time relative to the leading strand during replication. This is thought to lead to a mutational bias where the single-stranded DNA is more prone to cytosine deamination, which ultimately leads to an enrichment of thymine in the lagging strand. Gs and Cs are always paired, but because of this bias their frequencies are skewed: Gs appear more frequently on the leading strand, and Cs appear more frequently on the lagging strand (Lobry 1996). The extent of this bias varies along the chromosome, and the following GC skew calculation measures it.

First, I divided the chromosome into one kilobase wide windows, then I counted the number of Gs and Cs in each and calculate $\xi = (G-C)/(G+C)$. For each window I recorded the sum of the ξ-values up to that window. This is called the cumulative GC skew (CGC skew). Peak cumulative GC-skew were used to identify the origin and terminus of replication of each genome: the peak corresponds to the terminus; the trough corresponds to the origin

The locations of the origin and terminus in each genome were used to construct a new coordinate system for locating genes. The origin was assigned coordinate zero. Genes were assigned coordinates based on their distance in bases from the origin. Since the data files used were linear, in every case it was necessary the chromosomal coordinates of the genes were re-ordered to start at Ori. For each organism, positions of each gene were extracted from the *CoreCruncher* output and recorded, then merged with the PAML output, yielding the

relationship between the substitution rate and efficacy of selection on each gene and the position of that gene on its replichore.

Spearman's rank correlation coefficients were calculated between each gene's position and its value for each parameter of interest (*dN/dS*, *dN*, and *dS*). The coefficients for each species were collected and a distribution of ρ-values was tabulated for each parameter, giving a sense of the overall behavior of each parameter with respect to location for the entire dataset.

Window Analysis and Hotspot Detection

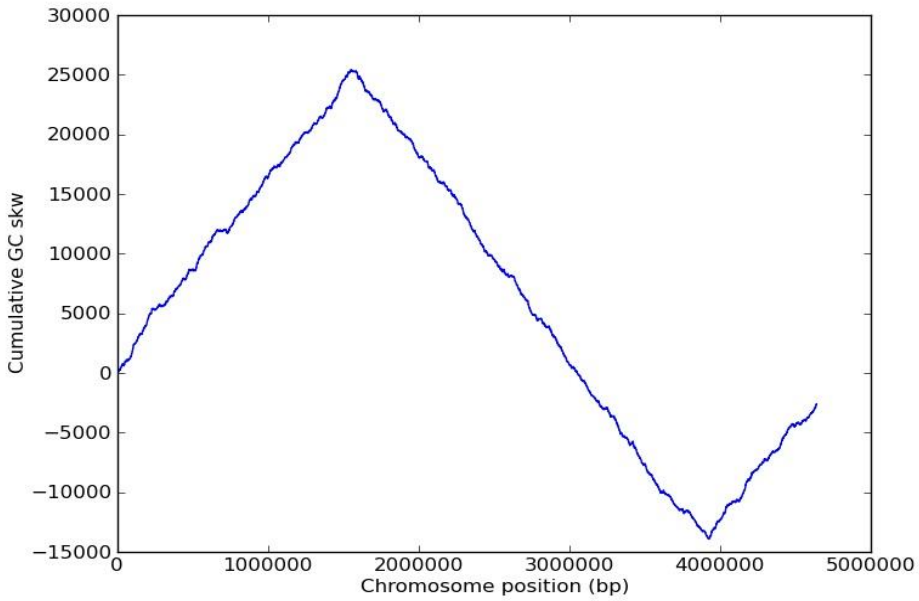
Starting at the origin of replication for each organism, the individual genomes were partitioned into windows of 50 kilobases each to facilitate further analysis. Average values for *dN/dS*, *dN*, and *dS* were calculated for genes falling within each window. Z-scores were calculated for each parameter average in each window, and the five highest and lowest Z-scores for each parameter were selected for each species. These collections of hotspots and coldspots were then plotted and visually inspected to see if they clustered near the origin of replication, near the terminus, somewhere in between or nowhere in particular. The distribution of hotspots and coldspots across species was also analyzed statistically.

CHAPTER III: RESULTS

GC Skew

I compiled a dataset of 329 bacterial species, which represents a total 15,225 completely assembled genomes. First, I identified the origin (Ori) and the terminus (Ter) of replication for each species. GC skew calculations were performed and graphed for one reference genome of each species (one genome was randomly used as a reference). The graphs were all individually inspected. They showed at most minor deviations from the sample graph, for *E. coli* (Fig. 4). The maximum and minimum values were used to identify the terminus and origin of replication, respectively: the peak corresponds to the terminus; the trough is the origin.

Figure 4. GC Skew for E. coli

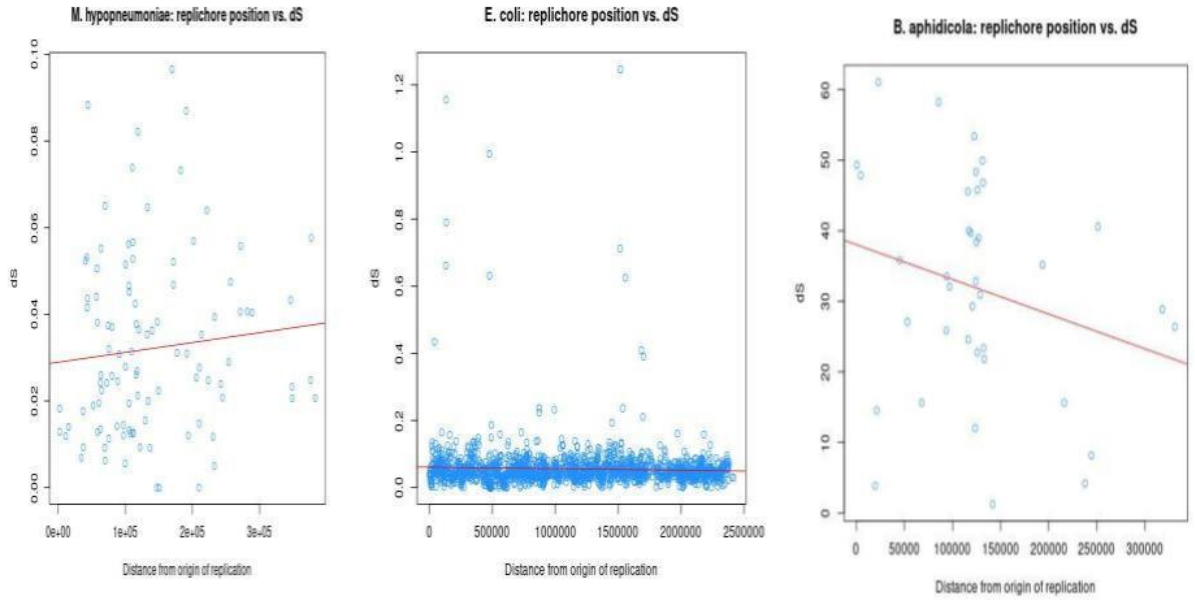


The peak is the terminus. The trough is the origin.

Synonymous Substitution

For each species, the rate of synonymous substitutions, *dS,* was tabulated for each gene and compared to each gene's position in base pairs away from the origin of replication. Spearman's rank correlation coefficients were then calculated between location and *dS.* Three representative species are plotted below (Fig. 5).

Figure 5. Correlations between dS and Gene Location



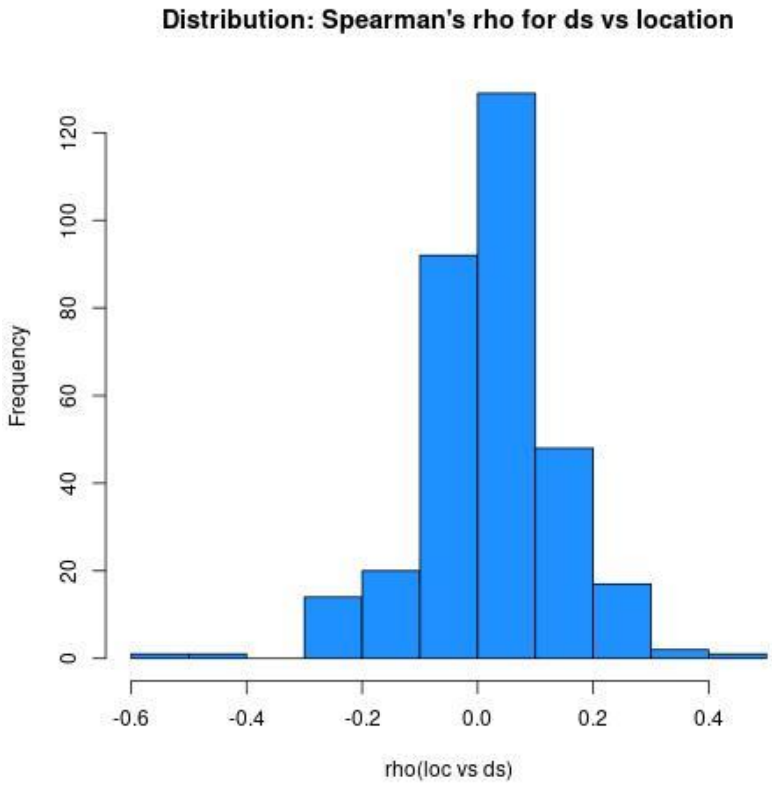
Mycoplasma hypopneumoniae (left, positive correlation), Escherichia coli (center, neutral correlation) and Buchnera aphidicola (right, negative correlation)

Of 329 total species, 146 had statistically significant correlations between average *dS* value and gene location (44%). Of the species that had significant correlations, 107 (73%) were positively correlated to chromosomal position along the Ori-Ter axis and 39 (27%) were negatively correlated to the position on the Ori-Ter axis. The p-values used to evaluate statistical significance were corrected for multiple testing with the Benjamini Hochberg method.

A histogram of the Spearman's coefficients is plotted for all 329 species on the next page (Fig. 6). The mean correlation coefficient for this sample is 0.03 and the standard deviation is 0.13. In other words, for all but a few outlier species, the correlations were neutral. Overall, I mostly observed no correlation or weak correlations between chromosomal position and the rate of substitution at synonymous positions, and those correlations that were significant were not systematically positive or negative.

Interestingly, these correlations were not evenly distributed across taxa: 45% of *Gammaproteobacteria* (n=99) presented a significant positive relationship between substitution rates and distance to Ori (Chi-square test, $P < 0.001$), whereas only 8% presented a significant negative relationship. In contrast, 46% of *Actinobacteria* (n=26) displayed a significant negative relationship (12% presented a significant positive relationship).

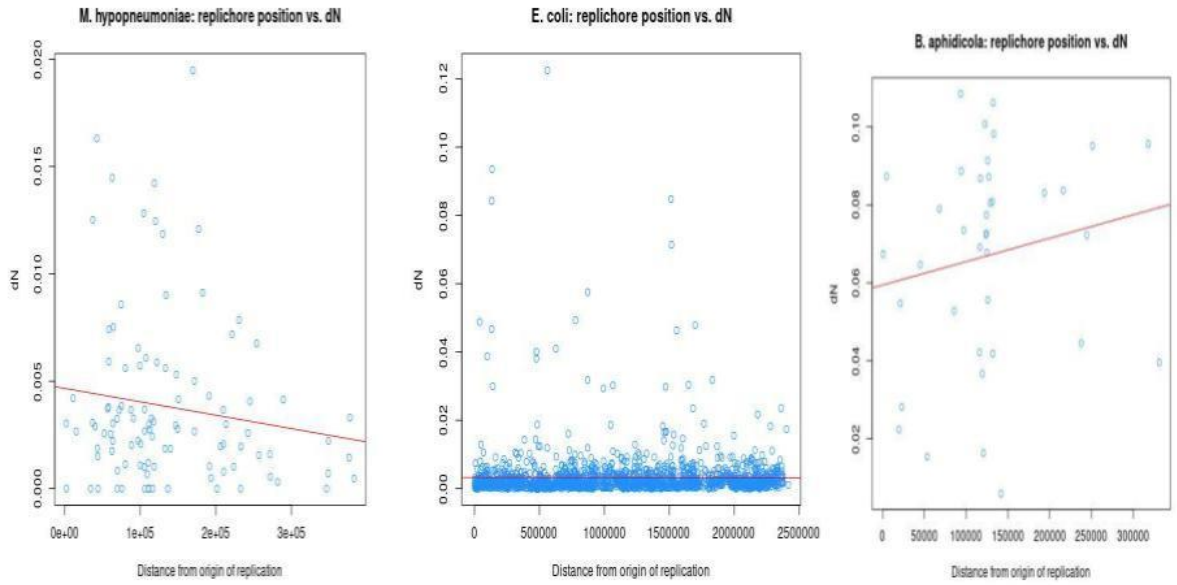
Figure 6. Histogram of Spearman coefficients for dS vs. Location



Non-Synonymous Mutation

The same analysis was repeated for *dN*, the rate of non-synonymous mutations. As described above, an average *dN* value was inferred with PAML for each gene in each species. Spearman correlations were calculated between the average *dN* values and the locations of each gene along the position on the Ori-Ter axis. The *dN*-location correlations from representative organisms are plotted in Fig. 7.

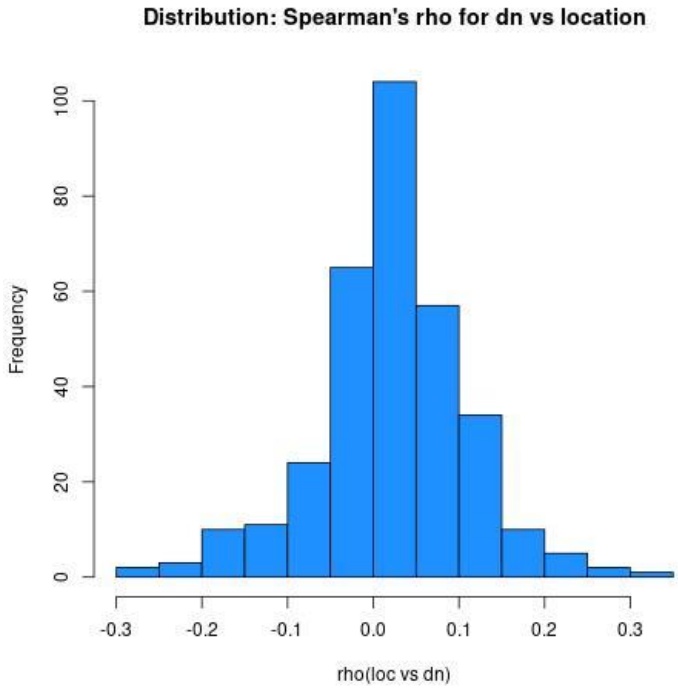
Figure 7. Correlations Between dN and Gene Location



Three species are pictured: Mycoplasma hypopneumoniae (left, negative correlation), Escherichia coli (center, neutral correlation) and Buchnera aphidicola (right, positive correlation)

Of 329 total species, 113 (34%) had significant correlations between *dN* and chromosomal position on the Ori-Ter axis. P-values used to assess significance were corrected for multiple testing using the BH method. Of those 113, 80 had positive correlations (71%) and 33 (29%) had negative correlations. The histogram of Spearman coefficients for all 329 species is plotted in Fig. 7. The mean correlation coefficient was 0.03 with standard deviation 0.10. Again, the vast majority of correlations were neutral, with only a few outliers positive or negative, indicating that most bacteria display no universal correlation between substitution rate at non-synonymous sites and location on the Ori-Ter axis. In addition, the species with significant correlations presented weak coefficients of correlation.

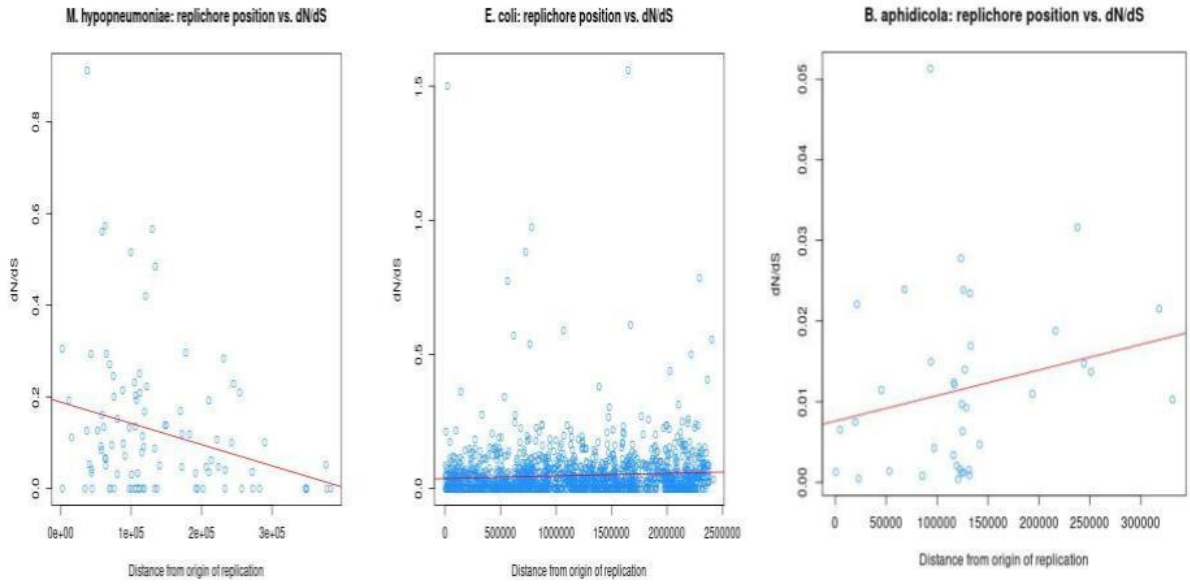
Figure 8. Spearman's coefficients for dN-location Correlation



Strength of Selection

The preceding analysis was carried out a third time for *dN/dS*, which is frequently used to evaluate the strength of selection acting on protein coding genes. Correlations between *dN/dS* and location are plotted for sample organisms (Fig. 9).

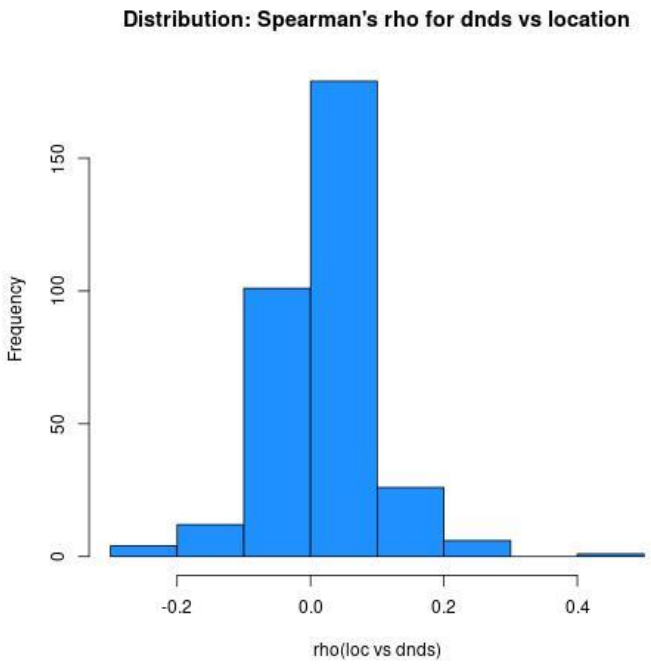
Figure 9. Correlations between dN/dS and Gene Location



Mycoplasma hypopneumoniae (left, negative correlation), Escherichia coli (center, neutral correlation) and Buchnera aphidicola (right, positive correlation)

Only 56 of 329 species had significant correlations (17%) (again, BH correction was used to correct significance estimates for multiple testing). Forty of those correlations (71%) were positive. The remaining 16 (29%) were negative. The histogram of *dN/dS* correlation coefficients is below (Figure 10). The mean correlation coefficient was 0.02 with standard deviation 0.08. These results indicate that the strength of selection is not correlated to the chromosomal position along the Ori-Ter axis for most species. Significant but weak correlations were observed for a minority of species.

Figure 10. Spearman's Coefficients for dN/dS-Location Correlation



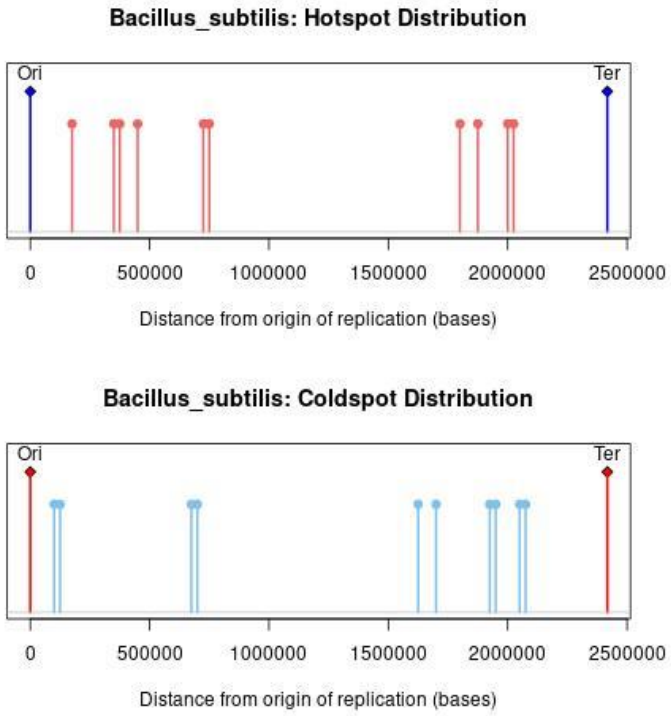
Window Analysis and Hotspots

Spearman coefficients capture whether there is an *overall* trend of increasing selection as one moves towards the terminus. It does not test the possibility that a few regions with highly biased substitution rates may be clustered in some chromosomal areas. I checked for this possibility by breaking each genome into windows of 50 kilobases each, calculating average values of *dN/dS*, *dN* and *dS* for each gene within each window, calculating z-scores, and taking the ten windows with the highest and lowest scores (i.e., the most significantly biased). I then plotted these windows for each species and visually inspected them for clustering.

Overall, the inspection of the positions of *dN/dS* hotspots and coldspots for individual species did not reveal any obvious trend. The graph for extreme windows in *B. subtilis* is typical (Fig. 11): there is an empty region in the middle and some slight bunching near the

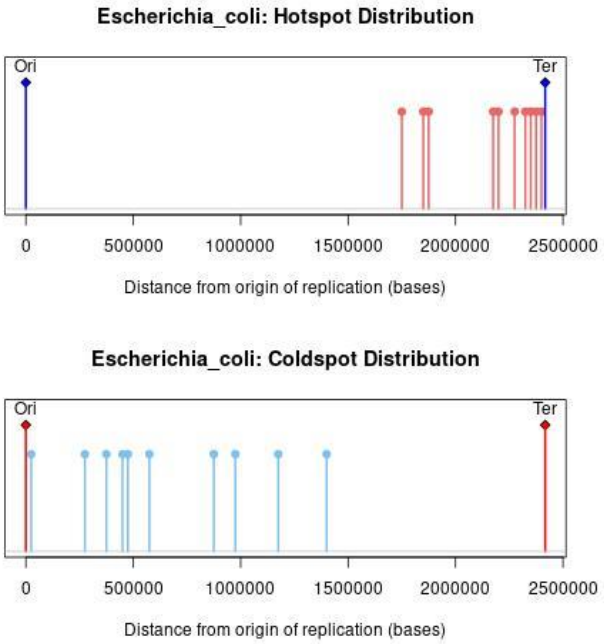
terminus, but the distribution is fairly uniform. The ten least strongly selected windows (coldspots) appear uniformly distributed as well.

Figure 11. Most Biased Window Locations in B. subtilis



Some species had more pronounced patterns, however. *E. coli*'s 10 hotspots appeared only in the most distant 30% of the chromosome, and seven were in the last 10% (Fig. 12). *E. coli*'s coldspots, in contrast, were distributed uniformly along the first 60% of the chromosome.

Figure 12. Most Biased Window Locations in E. coli



I also combined the data for all *dN*, *dS* and *dN/dS* hotspots and coldspots into six datasets and plotted histograms of their locations. Hot and coldspots are shown for *dS* in Fig. 13 and 14, for *dN* in Fig. 15 and 16, and for *dN/dS* in Fig. 17 and 18.

The distribution of *dS* hotspots is uniform for the first 60% of the Ori-Ter axis, and then curves upward toward a peak at Ter. The distribution of coldspots is flat everywhere except for a small spike at Ori and a larger spike at Ter. Interestingly, the bias of core genes with higher *dS* values toward Ter was much more common for *Gammaproteobacteria* as these represented 61% of species with a significant bias toward Ter ($P < 0.004$, Chi-square test), although *Gammaproteobacteria* represented only 36% of the dataset.

Figure 13. Distribution of dS Hotspot Locations for all Species

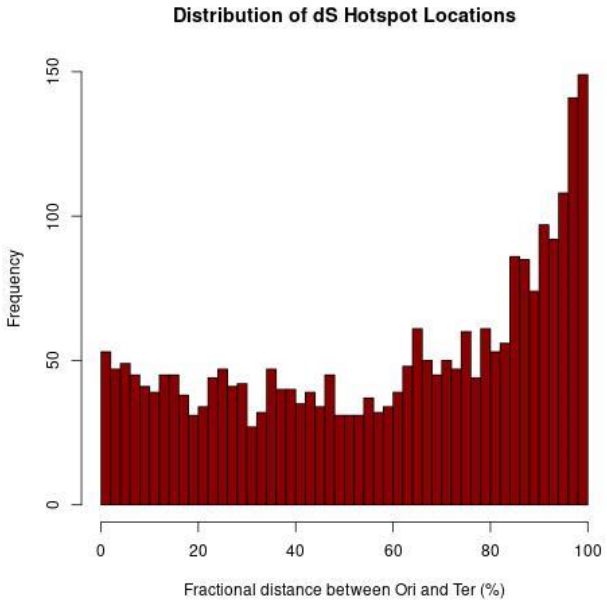


Figure 14. Distribution of dS Coldspot Locations for all Species

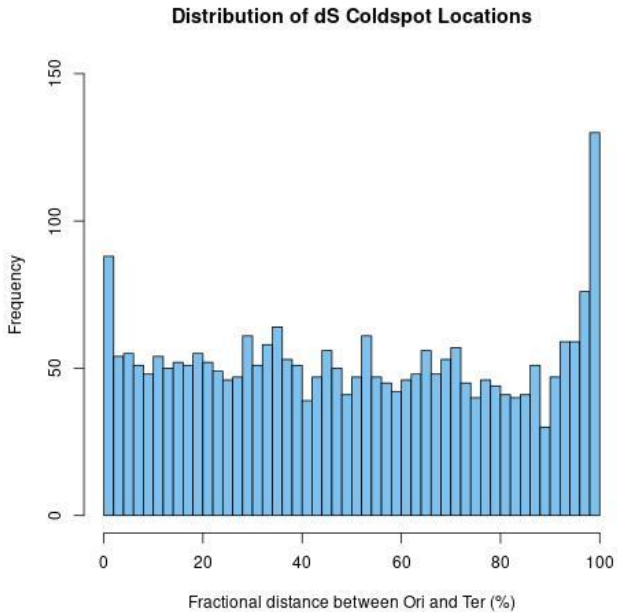


Figure 15. Distribution of dN Hotspot Locations for all Species

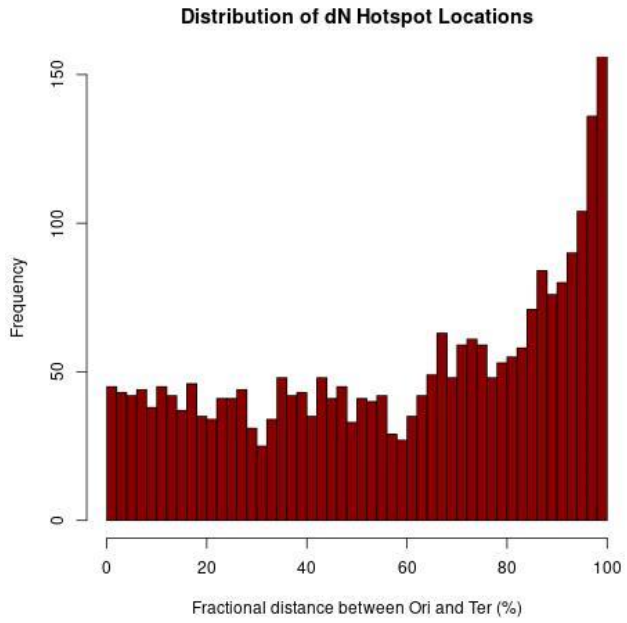
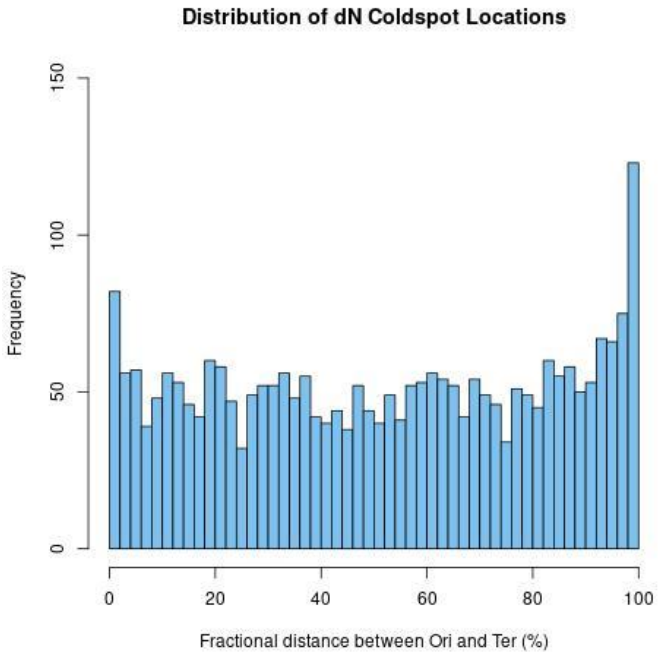


Figure 16. Distribution of dN Coldspot Locations for all Species



Distribution of *dN* hotspots and coldspots (see above) are similar to those for *dS.* For hotspots there is a flat distribution for the first 60% of the chromosome, followed by a curve upward toward a peak at Ter. The coldspot distribution is flat except for a modest spike very close to Ori and a larger spike very close to Ter.

Figure 17. Distribution of dN/dS Hotspot Locations for all Species

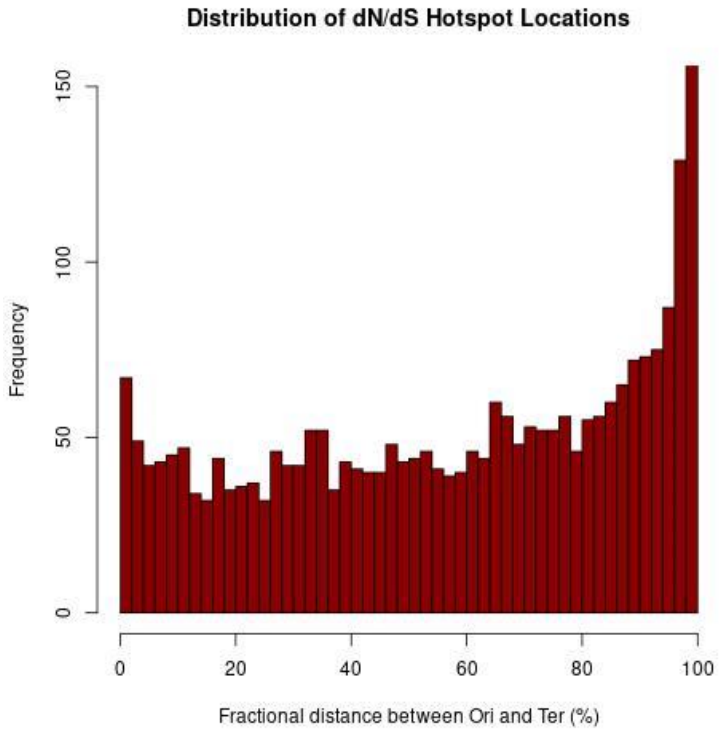
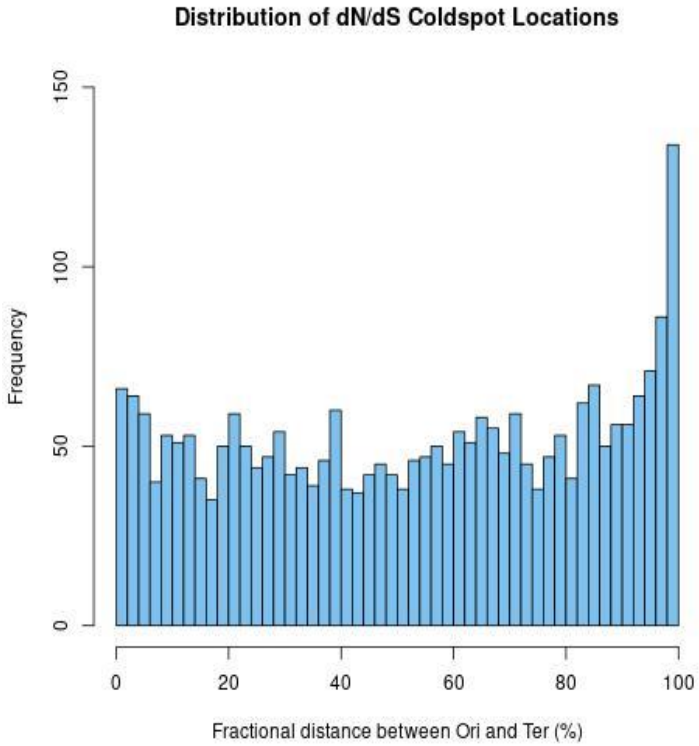


Figure 18. Distribution of dN/dS Coldspot Locations for all Species



Distribution of *dN/dS* hotspots was flat for the first 40% of the genome, sloped gently upward for the next 20%, and then curved sharply upward for the final 40%. There was a prominent peak at Ter, but no small peak at Ori. Coldspot distribution was uniform across the first 80% of the genome before a curve upward to a peak at Ter. Overall, these results indicate that regions of the chromosome presenting more extremely-biased rates of substitution tend to cluster near the Ter.

CHAPTER IV: DISCUSSION

Many studies over the past 20 years have produced a consensus that the substitution rates of genes increase with increasing distance from the origin of replication (Rocha and Dauchin 2004; Courtier and Rocha 2006). There have also been several variations on this theory. Ochman (2003) found that substitutions were most common halfway between the origin and terminus, and Dillon *et al.* (2018) detected wavelike patterns of mutation. These results have all been obtained using few genomes from only a few species. Touchon *et al.* (2009) used only *E. coli* for their research. Dillon *et al.* used *V. fischeri* and *V. cholerae*; Cooper *et al* (2010) used *Burkholderia, Vibrio, Bordetella* and *Xanthomonas.*

Lato and Golding set out to test the consensus with more data in 2020, but their expanded sample was limited by their computationally intensive approach to five genomes each of *E. coli, B. subtilis, S. meliloti* and *Streptomyces*. Even with this limited dataset, they found the consensus open to question. They found that *dN*, *dS* and *dN/dS* were not always correlated with distance from the origin of replication in a statistically significant way, and that those correlations that were significant were weak and went in either direction.

This study takes up the task of establishing to what extent these patterns are universal in prokaryotes where Lato and Golding left off. It examines every completely assembled bacterial genome that was available when it was begun in 2020, 15,225 in all, spread across 329 species. After computing *dN, dS* and *dN/dS* values for every core gene in every species, and assigning origin-relative coordinates to every gene, I calculated correlation coefficients for every species between its genes' substitution rates and their locations.

The results are consistent with Lato and Golding's. The correlation between *dN* and the distance from origin is significant for only 34% of the analyzed species. Correlations with *dS* are

significant for only 42% of species. For *dN/dS*, only 17% of the species display a significant pattern. For all three parameters, mean and median correlation coefficients are near zero and variances are low. Organisms with the strongest positive or negative correlations, such as *Buchnera apidicola* and *Mycoplasma hypopneumoniae*, tend to have very small genomes. Some clades such as *Gammaproteobacteria* tend to present positive correlations, while others, such as *Actinobacteria*, predominantly present negative relationships. However, those are trends and there are no clear patterns that are lineage-specific.

Through the analysis of genomic hotspots and coldspots of substitution, I inspected the spatial distribution of the most- and least-substituted regions of species' chromosomes, and found that, while there is no consistent linear correlation of substitution rates or strength of selection along the Ori-Ter axis, outliers are more common near Ter. For *dN*, *dS* and *dN/dS*, and for both hotspots and coldspots, the distribution of substitutions is uniform for most of the aggregate genome but are much higher near Ter (twofold to threefold increase). This supports the consensus view, *in the aggregate*, that the chromosomal region near Ter presents more variable substitution rate but patterns are often unclear when analyzing individual species. Most of the species that did follow this pattern individually were *Gammaproteobacteria*. Coldspots also appeared more frequently near Ter, so these results should be interpreted as showing that variability, in general, is higher in the Ter region of the chromosome. The biological mechanisms shaping more heterogeneous rates of evolution of the Ter region are unknown. It is possible that the different levels of compaction and chromosome structure differ at the Ter macrodomain, limiting access to certain regions for the homologous recombination machinery (for DNA repair) and favoring its access to other regions. Based on this larger sample, we can conclude that the sharpness of previous results were due mostly to their small sample sizes. I can confirm, for

example, that *E. coli* does exhibit elevated *dN/dS* values near the terminus of replication, as Touchon *et al.* found in 2009. But the widely reported molecular trends based on these small datasets do not generalize to most species, as this broader analysis shows.

It may be that the trend observed in previous studies occurs in restricted sets of lineages, however. In the 39 species that have a positive correlation between average *dN/dS* and gene location, certain genera recur: species of *Bacillus* four times, two species of *Clostridium*, six species of *Enterobacter*, five species of *Klebsiella* , six species of *Pseudomonas*, and four species of *Serratia*. In species with significantly negative correlations, three species of *Lactobacillus* were identified, two species of *Phaeobacter*, and three species of *Staphylococcus*. Genera are not split between the sets of positively and negatively correlated bacteria.

REFERENCES

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. Molecular Biology of the Cell, 6th Edition. 2015 Garland Science, New York, NY

Blattner FR, Plunkett G, Bloch CA, Perna NT, *et al.* The Complete Genome Sequence of *Escherichia coli* K-12. *Science* 1997 **277**:5331 1453-62

Bobay LM, Rocha EPC, Touchon M. *MBE* 2013 **30**:4 737-51

Bobay LM. Personal communication, 2021

Brilli M, Lio P, Lacroix V, Sagot MF. Short and long-term genome stability analysis of prokaryotic genomes. *BMC Genomics*. 2013 **14**:309

Buton A, Bobay LM. Evolution of Chi motifs in Proteobacteria. *Genes|Genomes|Genetics* 2021 **11**:1

Casjens S.. The Diverse and Dynamic Structure of Bacterial Chromosomes. *Annu. Rev. Genet.* 1998 **32**:339-77

Clokie MRJ, Millard AD, Heaphy S. Phages in nature. *Bacteriophage* 2011 **1**:1 31-45

Darling, AE, Miklos I,.Ragan MA. Dynamics of Genome Rearrangement in Bacterial Populations. *PLoS Genet.* 2008 4(7): e1000128

Dillon MM, Sung W, Lynch M, Cooper VS. Periodic variation of mutation rates in bacterial genomes associated with replication timing. *mBio* 2018 9:e01371-18.

Gangan MS, Athale CA. Threshold effect of growth rate on population variability of *Escherichia coli* cell lengths. *Royal Society Open Science* **4**:2

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and throughput. *Nucleic Acids Res.* 2004 **32**(5): 1792-1797

Harris CD, Torrance EL, Raymann K, Bobay LM. Core Cruncher: Fast and Robust Construction of Core Genomes in Large Prokaryotic Datasets. *Molecular Biology and Evolution* (2020) **38**(2):727-734

Hendrickson H, Lawrence JG. Selection for Chromosome Architecture in Bacteria. *J Mol Evol* (2006) **62**:615-629

Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 2009 **19**:8 1450-4

Lato DF, Golding GB. The Location of Substitutions and Bacterial Genome Arrangements. *Genome Biol. Evol.* 2020 **13**(1)

Lobry JR. Asymmetric substitution patterns in the two DNA strands

Luo H, Gao F. DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. *Nucleic Acids Research* 2019 **47**:D1 D74-D77

Moran N, Bennett G. The Tiniest Tiny Genomes. *Annu Rev Microbiol.* 2014;68:195-215

Merrikh CN, Merrikh H. Gene Inversion Potentiates Bacterial Evolvability and Virulence. *Nature Communications* 2018 **9**:4662

Niccum BA, Lee H, MohammedIsmail W, Tang H, Foster PL. 2019. The symmetrical wave pattern of base-pair substitution rates across the Escherichia coli chromosome has multiple causes. *mBio* 10:e01226-19.

Repar J, Supek F, Klanjscek T, Warnecke T, Zahradka K, Zahradka D. Elevated Rate of Genome Rearrangements in Radiation-Resistant Bacteria. *Genetics* 2017 **205**:4 1677-89

Rocha, EPC. Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol* 2006 **23**:3 513-22

Rocha, EP.C. The Organization of the Bacterial Genome. *Annu. Rev. Genet.* 2008 **42**:211-33

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014 **30**:9 1312-1313

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, et al. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. PLoS Genet (2009) 5(1): e1000344.

Touchon M, Bobay LM, Rocha EPC. The Chromosomal accommodation and domestication of mobile genetic elements. *Current Opinion in Microbiology* (2014) **22**:22-29

Trojanowski D, Holowka J, Zakrzewska-Czerwinska J. Where and When Bacterial Chromosome Replication Starts: A Single Cell Perspective. *Front. Microbiol.* 26 November 2018

Valens M, Penaud S, Rossignol M, Cornet F, Boccard F. Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.* (2004) **23**:21 4330-41

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 1997 **13**:5 555-6