Bacterial Molecular Properties and Genomic Position

Spatial Patterns of Molecular Traits in Bacterial Genomes

By Daniella F. Lato, BSc

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University (2021) Doctor of Philosophy
Hamilton, Ontario (Department of Biology)

TITLE: Spatial Patterns of Molecular Traits in Bacterial Genomes
AUTHOR: Daniella F. Lato BSc (McMaster University)
SUPERVISORS: G. Brian Golding
NUMBER OF PAGES: xvi, 181

Abstract

The placement of genetic information within bacterial genomes is intentionally organized, creating predictable gradients of molecular properties along the origin-terminus of replication axis. Previous studies have reported that genes located near the origin of replication generally have a higher expression level, increased dosage, and are more conserved than genes located near the terminus of replication. Additionally, substitution rates usually increases with increasing distance from the origin of replication. However, the constant reorganization of genetic information is often overlooked when considering spatial molecular trends.

Here, we explore the interplay of genomic reorganization along the origin and terminus of replication axis of gene expression and substitutions in *Escherichia coli*, *Bacillus subtilis*, *Streptomyces*, and *Sinorhizobium meliloti*. Using ancestral reconstruction to account for genome reorganization, we demonstrated that the correlation between the number of substitutions and distance from the origin of replication is significant but small and inconsistent in direction. In another study, we looked at the overall expression levels of all genes from the same bacteria, and confirmed that gene expression tends to decrease when moving away from the origin of replication. We looked specifically at how inversions - one type of genomic reorganization - impact gene expression between closely related strains of *E. coli*. Some inversions cause significant differences in gene expression compared to non-inverted regions, however, the variation in expression does not significantly differ between inverted and non-inverted regions. This change in gene expression may be due to the expression regulation properties of two nucleoid associated proteins, Histone-like Nucleoid-Structuring (H-NS) and Factor for inversion stimulation (Fis), who's binding sites had a significant positive correlation with inverted regions.

In conclusion, we highlight the impact that genomic rearrangements and location have on molecular trends in bacteria, illustrating the importance of considering spatial trends in molecular evolutionary analyses, and to ensure accurate generalization of previously determined trends. Assuming that molecular trends are exclusively in one direction can be problematic.

what is the greatest lesson a woman should learn

that since day one
she's already had everything she needs within herself
it's the world that convinced her she did not
- Rupi Kaur

Acknowledgements

To my supervisor Dr. Brian Golding, thank you for seeing potential in me as an undergraduate and taking the time to literally teach me how to code. I would not be able to do anything I am doing today without the basics your taught me on the command line. Thank you for consistently dropping whatever you were doing to help me work through a problem (especially when it was LaTeX related). Thank you for fostering my curiosity and love for genetics and the field of biology I did not know I needed, bioinformatics.

To my supervisory committee Dr. Marie Elliot and Dr. Ben Evans, thank you for your continuous support. Ben, you have heard me present my research from undergrad all the way until now and have constantly provided the enthusiasm in lab meeting I needed to keep talking about substitutions for 7 years. Marie, thank you for bringing new perspectives to my research and for so kindly always checking in on me to see how everything was progressing.

Thank you to my external examiner, Dr. Xuhua Xia, for your invaluable feedback which contributed to bettering this thesis.

Thank you to all the past and present members of the Golding Lab for being the crucial daily support system assisting in figuring out what the scripts was doing (or not doing). Being apart of the Golding Lab has been a blast over the years and I am going to miss everyone that I was fortunate enough to interact with. Thank you in particular to Caitlin, Yasser, Sharok, and Zach for the wonderful friendships and immense support inside and outside of the lab. Caitlin, I could not have made it through the last push of grad school without our long distance coffee breaks. Thank you to Shawn for showing me that you can be a graduate student and still have interests and goals other than your research, and for reminding me that I am doing amazing.

To my larger McMaster community, thank you for creating environments where I could foster friendships, explore my interests, excel in leadership roles and swim to my hearts content. It has been a fabulous decade (nearly) long roller coaster and I would do it all again in a heartbeat.

Finally, the biggest thank you goes out to my family. Thanks for the constant and infinite love and support. Mom and Dad, you are always there for me no matter what and have supported me in everything I do. Thank you for always calming my stressed phone calls, making the drive to Hamilton weekly at some points, and bragging about everything I do to anyone who will listen. You have been to every milestone event in my life, always in the stands or audience cheering and encouraging me to keep going. You are the best parents anyone could ask for. To my brother Thomas, I have never met anyone who works harder than you. Your dedication and strength during your Masters has always inspired me to dive deeper into my studies. You remind me not to compromise what I love and who I am for anything. The world is not ready for all the amazing things you are going to accomplish.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Daniella F. Lato, declare that this thesis titled, "Spatial Patterns of Molecular Traits in Bacterial Genomes" and the work presented in it are my own. I confirm that:

- Chapter 1 - I completed a literature search and wrote the manuscript that introduces the thesis. G.B. Golding contributed to the editing of the manuscript.
- Chapter 2 - I designed the experiment and completed all analyses described in the manuscript including compiling datasets, whole genome alignment and quality check, phylogenetic analyses, ancestral reconstruction, and statistical analyses. I wrote the manuscript and G.B. Golding supervised the analyses and revision of the manuscript. This chapter is published in Genome Biology and Evolution.
- Chapter 3 - I designed the experiment and completed all analyses described in the manuscript including compiling datasets, gene expression normalization, and statistical analyses. I wrote the manuscript and G.B. Golding supervised the analyses and revision of the manuscript. This chapter is published in the Journal of Molecular Evolution.
- Chapter 4 - I designed the experiment and completed all analyses described in the manuscript including compiling datasets, gene expression normalization, whole genome alignment and quality check, and statistical analyses. Q. Zeng completed data processing for raw RNA-seq count data under the supervision of myself. I wrote the manuscript and G.B. Golding supervised the analyses and revision of the manuscript. This chapter is formatted for submission to GENOME.
- Chapter 5 - I completed a literature search and wrote the manuscript that discuss the results of the thesis. G.B. Golding contributed to the editing of the manuscript.

Acronyms

AIMS Architecture IMparting Sequences

bp Base Pair

COG Clusters of Orthologous Groups of proteins

CPM Counts Per Million

d*N* Non-synonymous Substitution Rate

d*S* Synonymous Substitution Rate

DBRT Database of Bacterial Replication Terminus

DNAP Deoxyribonucleic Acid Polymerase

dNTP Deoxyribonucleotide triphosphate

Fis Factor for inversion stimulation

GEO Gene Expression Ombinbus

GI Genomic Island

HGT Horizontal Gene Transfer

H-NS Histone-like Nucleoid-Structuring

indel Insertion/Deletion

IS Insertion Sequence

Kbp Kilobase Pair

lac Lactose Operon

LCB Locally Colinear Block

LOO Leave One Out

Mbp Megabase Pair

NS Non-Significant

PI Pathogenicity Island

RNAP Ribonucleic Acid Polymerase

rRNA Ribosomal Ribonucleic Acid

SNP Single Nucleotide Polymorphism

TE Transposable Element

TF Transcription Factor

TMM Trimmed Mean of M values

tRNA Tranfer Ribonucleic Acid

Chapter 1

Introduction

Daniella F. Lato

1.1 Bacteria: evolutionarily efficient

Bacterial genomes have evolved over millions of generations to become some of the most efficient and compact genomes on the planet. The complicated control mechanisms for bacterial life such as replication and gene expression, need to be contained within (often) one tiny circular chromosome. This compaction has become so efficient that only approximately 12% of prokaryotic genomes are non-coding (Ahnert et al. 2008), compared to the human genome where about 99% is non-coding. One of the most notable ways that bacteria have become genomically efficient is through combined transcription and translation (Griswold 2008; Le and Laub 2014). Bacterial genomes have the ability to have transcription and translation co-occur when replicating their genomes (Byrne et al. 1964; Miller et al. 1970). This results in bacterial genomes being highly organized spatially (Le and Laub 2014), especially in the cases where bacterial genomes are spread across multiple replicons, or chromosome-like structures. The coupling of transcription and translation allows bacteria to process for example, environmental changes in real time and alter when replication or other molecular processes begin (Wang et al. 2013; Marczynski et al. 2015). This is continually regulated through a number of complex feedback loops and molecular machinery based on growth state, environmental conditions, and stress (Wang et al. 2013; Marczynski et al. 2015).

Through these molecular and physical mechanisms, some bacteria maintain robust fitness and gene expression, despite huge changes in genomic organization (Naseeb et al. 2016). This is seen across bacterial species and is thought to be an evolutionarily conserved mechanism (Hartman, JL et al. 2001). These efficient evolutionary advances would not be possible without the successful integration and restructuring of DNA.

1.2 Receiving and reorganizing genetic information

Bacteria have become remarkably efficient in many aspects of their lives such as antibiotic resistance and adaptation to new and changing environments and hosts. These modifications would not be possible without the impressive mechanisms for obtaining and reorganizing genetic information.

1.2.1 Horizontal gene transfer

For all life on earth, the acquisition of new genetic information is crucial for constant evolution and adaptation. Most organisms acquire new genetic information and re-organize current genetic data through sexual reproduction. Bacteria are asexual and therefore have to rely on alternative methods such as Horizontal Gene Transfer (HGT) (Daubin and Ochman 2004). HGT is the non-vertical transmission of DNA from one bacterium to another. This can happen between bacteria from the same strain or even between bacteria from different species. There have even been a few examples of HGT between eukaryotes and prokaryotes (Soucy et al. 2015), although DNA

transfer happens most often among bacteria. Tenaillon et al. (2010) estimated that in bacteria, a single base is 100 fold more likely to be involved in a transfer event than to be mutated. HGT is therefore considered one of the main mechanisms bacteria use to obtain genetic variation (Ochman et al. 2000; Daubin and Ochman 2004) and escape Muller's Ratchet (Koonin 2016).

The first example of DNA being transferred non-linearly, where mobile elements providing antibiotic resistance were gained (Ochman et al. 2000; Soucy et al. 2015). This phenomenon happens not just with respect to antibiotic resistance, but any beneficial gene. If one bacteria contains a gene that is well adapted to a certain environment, this genetic information can be shared between genomes without the need for other bacteria to have evolved the trait "on its own" (Daubin and Ochman 2004; Soucy et al. 2015). The benefits of acquiring genes periodically, eliminates the need to constantly maintain that gene within a plasmid or portion of the bacterial genome (Soucy et al. 2015). This allows for increased metabolic properties and promotes diversification of gene function through the re-assortment of existing capabilities (Ochman et al. 2000). For example, there have been instances where HGT has allowed for changes in gene expression and phenotype between bacterial strains (Rocha 2004a).

Mechanisms for HGT

There is variation in HGT among bacteria, meaning that some are frequently undergoing transfer events, while other only have a few (Ochman et al. 2000). HGT occurs through four main mechanisms: transformation, transduction, conjugation, and through gene transfer agents (Soucy et al. 2015). Transformation is rather rare, and involves the uptake and expression of DNA or RNA from the environment (Soucy et al. 2015). Transduction is the most common form of transferring DNA between bacterium through a virus such as phages (Soucy et al. 2015). Other mobile and selfish genetic elements also help promote HGT (Soucy et al. 2015). Bacterial conjugation involves the transfer of DNA involving a plasmid which is then take up from donor to recipient through cell-to-cell contact (Soucy et al. 2015). Conjugation typically happens between bacterial species although there have been cases where *Agrobacterium tumefaciens* has used conjugation to transfer information to plants (Soucy et al. 2015). Finally, gene transfer agents, which are virus-like elements encoded by the host, are able to transfer and uptake DNA (Soucy et al. 2015). These agents have been found in the order *Rhodobacterales* (Soucy et al. 2015). Introgression can additionally create a transfer-like scenario, occuring with the hybridization of two species (Soucy et al. 2015).

Insertion sequences

Insertion Sequences (ISs) provide another mechanism bacteria can utilize to reorganize and transfer genetic meterial. ISs have the ability to move on their own or through integration into phages and plasmids (Siguier et al. 2014). ISs can incorporate genes involved in a variety of functions, such as antibiotic resistance(Siguier et al. 2014). There are diverse families or classes of IS elements that range in properties such as accessory gene content (Siguier et al. 2014). IS elements

can undergo massive losses and expansions that can be influenced by the host lifestyle (Siguier et al. 2014). For example, symbionts typically contain high IS loads, presumably to assist with adaptation to changing host environments (Siguier et al. 2014). We see a higher number of IS elements on plasmids rather than on chromosome hosts (Siguier et al. 2014). This is most likely because IS elements have the ability to impact genome structure and replication, which are less restricted on a plasmid than on a primary chromosome (Siguier et al. 2014).

It appears as though there is no region of bacterial genomes that is particularly favourable for ISs (Lee et al. 2016). However, the presence of IS elements can impact neighbouring genomic content. IS elements can cause recombination at about the same rate as they are inserted (Lee et al. 2016). Rates of IS insertion and recombination are nearly constant across multiple strains (Lee et al. 2016). IS sites have additionally been found to create a transposition bias in regions proximal to the insertion element (Lee et al. 2016). This means that novel insertion sites are often near a pre-existing copy of the IS element, which is most often found in non-coding regions of the genome (Lee et al. 2016). Like most other homologous genes, recombination happens most frequently between two physically close copies of IS elements (Lee et al. 2016).

Successful genetic transfer

Although all bacterial genetic material has the ability to be transferred, not all transfers will result in the successful uptake of DNA in the recipient. The main steps for any successful transfer are firstly, the DNA needs to get from the donor to the recipient. This is done through transformation, transduction, conjugation, and other gene transfer agents. Secondly, the new DNA needs to then be incorporated into the recipient genome so that it can finally be expressed and maintained in the genome of the recipient, so long as it confers a sufficient benefit (Ochman et al. 2000). There are a number of factors that decrease the likelihood of a successful transfer. An increase in phylogenetic distance between two taxa will decrease the chance of a successful transfer (Soucy et al. 2015), and the most frequent donors of genetic material are lineages within the same phylogroup (Nowell et al. 2014). However, there are some cases where taxa that are closer geographically or share a similar type of environment, are more comparable than phylogenetically related taxa (Hanage 2016). In general, bacteria that occupy a similar niche or environment are more likely to have a successful transfer (Wiedenbeck and Cohan 2011; Soucy et al. 2015). Additionally, if two bacteria occupy the same niche, then genes that would be useful for one environment could also be beneficial for a similar environment (Hendrickson et al. 2018).

The genes themselves also play a role in the achievement of transfer. Shorter genes are often more easily transferred because of less restrictions to size, being able to better integrate into the genome (Wiedenbeck and Cohan 2011). Genes with similar functions and regulators are more likely to be integrated into the genomes successfully because the existing machinery is already present in the recipient bacteria (Rocha 2008). This includes the successful transfer of operons which needs to be completed with precision, meaning that the complete set of operons

and regulators need to be present for transferred genes to work properly in a new organism (Gogarten and Townsend 2005; Rocha 2008).

The physical location of genes on the chromosome also influences transfer. Genes with similar functions, especially important ones, are often located near each other on the chromosome and therefore can be easily transferred, replicated and regulated together (Rocha 2004a; Daubin and Ochman 2004; Wiedenbeck and Cohan 2011; Soucy et al. 2015). Genes in close proximity promote co-expression and the use of the same regulators (Rocha 2004a; Daubin and Ochman 2004). Genes that are involved in more complex functions like genome structure, transcription and translation are less likely to be transferred because altering these genes in any way could be detrimental to the organism (Wiedenbeck and Cohan 2011). Likewise, genes that have a high number of protein-protein interactions are also less likely to be involved in HGT because this could have harmful consequences on the health of the bacteria (Wiedenbeck and Cohan 2011). A comprehensive list of genes that are transferred most often in bacteria has been curated by Wiedenbeck and Cohan (2011).

The external environment is also important to consider for transfer events. Some environments are better at preserving DNA, making it easier for bacteria to integrate this external DNA into their genomes (Thomas and Nielsen 2005).

The above mentioned criteria for a successful transfer apply to other organisms, such as eukaryotes. However, in the case of transfer between prokaryotes and eukaryotes, the transfer needs to occur before the loss of the ancestral gene between the two taxa (Husnik and McCutcheon 2018).

Hotpots for HGT

As mentioned previously, there are certain molecular characteristics that create favourable scenarios for horizontal transfer, such as genomic location. Horizontally transferred genes are concentrated in only about 1% of chromosomal locations called hotspots (Oliveira et al. 2017). The number of these hotspots increases with genome size and transfer rate, however there may be exceptions to this due to differences in selective pressures (Oliveira et al. 2017). Most mobile genetic elements are hotspots, however most hotspots lack mobile genetic elements (Oliveira et al. 2017). The location of hotspots varies between bacterial genomes and greatly depends on the local positions of core genes and mobile genetic elements (Oliveira et al. 2017). In some cases, hotspots are preferentially located near regions of core genes that perform functions related to replication, recombination repair, and transcription (Oliveira et al. 2017). These core genes provide anchors for the HGT hotspots allowing for increased recombination between core gene regions, creating a genetically diverse hotspot (Oliveira et al. 2017). There is also evidence that shows the frequency of HGT hotspots increases linearly with distance from the origin of replication (Oliveira et al. 2017). The impact that distance from the origin of replication has on various molecular trends will be discussed in detail later in this dissertation.

Maintenance of genetic material

Once genetic material is introduced into a bacterial genome through HGT, it then needs to be beneficial or adaptive to the organism for it to be maintained within the genome and not lost (Ochman et al. 2000; Soucy et al. 2015; McInerney et al. 2017). If the recent horizontally transferred genes confer neutral or nearly neutral benefits and these remain neutral over time, then they will eventually be lost (Soucy et al. 2015). Newly acquired genes have higher rates of evolution and lower levels of gene expression (Soucy et al. 2015) than the rest of the genome, usually to assist with integration to the new host environment, such that regulatory networks are not disrupted (Wiedenbeck and Cohan 2011). Therefore, recently transferred genes will have genomic signatures similar to the donor rather than to the recipient (Ochman et al. 2000; Daubin et al. 2003). As evolutionary time passes, these genes will adopt genomic signatures that are more similar to the new host (Ochman et al. 2000; Daubin et al. 2003).

Most of the genetic material that is transferred is not initially adaptive however, this material possesses the potential to become adaptive under the right conditions (Wiedenbeck and Cohan 2011). This means that there are going to be massive gene loss events happening to remove any non-adaptive genetic information and to combat the huge amounts of genes gained from HGT (Wolf and Koonin 2013). As more time passes from the transfer event, selection and other evolutionary forces have the opportunity to act upon the transferred gene(s), altering their genomic features to be more similar to the new host (Daubin et al. 2003).

Interestingly, transfers that happened long ago seem to be usually located in gene rich regions, where as transfers that are more recent are found in typically non-conserved regions (Husnik and McCutcheon 2018). The genomic constraints for the insertion of new sequences could be more controlled in conserved regions and more relaxed in non-conserved regions, possibly allowing for the integration of new genetic material. Regions of increased HGT tend to be clustered near the terminus because they are often weakly expressed and found in regions with local hyper recombination to promote the insertion of new genetic material (Rocha 2004b).

Karcagi et al. (2016) have performed experiments where large numbers of horizontally transferred genes were deleted from *Escherichia coli* genomes, and they found that there was a decline in nutrient utilization and stress tolerance. This is consistent with the idea that horizontally transferred genes are especially important in nutrient limiting environments (Karcagi et al. 2016). They suggest that epistasis may also play a role in which horizontally transferred genes are maintained due to changing physiology being induced (Karcagi et al. 2016).

With respect to a phylogenetic context, just because bacteria share DNA via HGT, does not mean that the host and donor will eventually converge evolutionary (Wiedenbeck and Cohan 2011). One of the primary benefits of HGT is that it provides diversification of existing capabilities (Ochman et al. 2000). Bacteria are sampling the large gene pool, not necessarily constantly accumulating genetic information (Ochman et al. 2000). Typically taxa with shared genes will group together when considering genes that are transferred compared to taxa that

do not share any genes (Gogarten and Townsend 2005). When looking at a particular branch within a phylogenetic tree, we typically see that the number of acquired genes is higher than the number of lost genes (Daubin et al. 2003). This means that the construction of phylogenetic trees becomes tricky when considering HGT (Gogarten and Townsend 2005). However, higher taxonomic groups of bacteria show phylogenetic and ecological cohesion so there must be other factors constraining HGT (Hendrickson et al. 2018). Additionally, organisms that rely on symbiotic interactions can be exposed to new microbiotic environments through their host, which can promote HGT (Soucy et al. 2015).

Detecting HGT

It is easiest for us to detect recent HGT because it more closely resembles the genomic signatures from the donor bacteria rather than the recipient (Ochman et al. 2000; Daubin et al. 2003). The more time that has passed since the transfer event, the more evolutionary forces have the opportunity to act upon the transferred genes, therefore making it harder to detect HGT between distantly related taxa (Daubin et al. 2003; Wiedenbeck and Cohan 2011). The most popular method used to identify laterally transferred genes is through parametric methods, or looking for differences in genomic signatures when compared to the rest of the genome (Ravenhall et al. 2015). This includes looking for genes that are AT rich (Daubin et al. 2003), have different mutation biases (Daubin et al. 2003), or nucleotide, amino acid or codon bias (Ochman et al. 2000; Daubin et al. 2003; Gogarten and Townsend 2005; Soucy et al. 2015). These parametric methods do not rely on comparisons to other genomes which can be an asset (Ravenhall et al. 2015).

One can also use a variety of phylogenetic methods to detect HGT. For example looking at housekeeping genes is usually a good indication of genes that are vertically transmitted (Soucy et al. 2015), and any conserved genes usually have the same phylogenetic tree between species (Koonin 2016). Statistical differences in things like genomic signatures, between these housekeeping genes and any other genes gives a good indication of potential horizontally transferred genes (Soucy et al. 2015).

Unfortunately, there are a number of factors that produce phylogenetically similar scenarios to HGT, such as incomplete lineage sorting (Than et al. 2007) and gene duplication and loss (Gogarten and Townsend 2005). These can make determining the boundaries between species difficult (Daubin and Ochman 2004; Syvanen 2012) and causing discordance between the species and gene trees (Than et al. 2007; Koonin 2016).

It is best to use a variety of different methods in an attempt to identify HGT. If all methods are in agreement with certain genes, we can be more confident that they truly are horizontally transferred (Soucy et al. 2015).

1.2.2 Recombination

Similar to HGT, recombination allows for the reorganization of genetic material within a genome. These rearrangements can help with adaptation strategies (Rocha 2004a; Hanage 2016), gene conversion (Hanage 2016), and diversification by re-assortment of existing capabilities (Ochman et al. 2000) by shuffling genes around the genome (Hanage 2016).

Although recombination provides mostly beneficial services to bacteria by reorganizing genetic information, there are cases where recombination can cause a decrease in diversity. Recombination usually replaces a locus with another common locus, which can decrease diversification (Hanage 2016). Maddamsetti and Lenski (2018), showed that beneficial alleles were removed by increased recombination, although the bacteria did not decrease in fitness so there may be other introgressed genes from the donor that were good enough to offset the removal of previously beneficial genes.

Requirements for recombination

Recombination rate varies between bacteria where some recombine often and others, very little (Hanage et al. 2009). The rate of recombination is dependent on a number of factors such as differences in ecology (Rocha 2004a). Some speculate that bad environments or limited nutrients are necessary to promote rearrangements and recombination, however, in a long term evolution experiment, Raeside et al. (2014) determined that recombination happens frequently without adverse environmental conditions. Most recombination events are species specific (Hanage 2016), meaning that the greater the sequence divergence, the less homologous recombination (Daubin and Ochman 2004). We also see a decrease in recombination with decreased genetic variance (Casillas and Barbadilla 2017).

Additionally, sequence similarity, not necessarily homology, can promote homologous recombination (Hanage et al. 2009). Typically recombination works best when there are flanking regions of sequence similarity, and the sequence in between these similar regions could be anything (Hanage 2016). In bacteria, genes with a high GC content have evidence of increased recombination (Lassalle et al. 2015), and this has also been observed in humans (Lassalle et al. 2015) and does not appear to be linked to codon bias (Lassalle et al. 2015).

Detecting recombination

When recombination replaces like sequences with each other, it becomes more difficult to detect (Hanage 2016). It is therefore, often easier to detect recombination in bacteria with lots of genetic variation (Hanage 2016). One of the most confounding variables for detecting recombination is mutations. Recurrent mutation rates can produce the same signals as recombination (McVean et al. 2002). Likewise, recombination makes it hard to detect other phenomenon due to the change in sequence compositions and genomic signatures (McVean et al. 2002; Hanage 2016). One way to distinguish recombination and mutations is to look at allelic variation at multiple

loci (Hanage 2016). Recombination should show that the alleles share a locus with some other taxa in a population (Hanage 2016). If this does not happen, then we are most likely seeing the result of a mutation, not recombination (Hanage 2016). It is therefore important to check the frequency of a particular locus elsewhere in the genome and consider sequence homology across sections of a gene locus, because not all parts of a gene have the same homology (Hanage 2016).

RecA: recombination assistant

Recombination has a tight linkage with replication and particularly RecA, which is involved in the repair process of recombination in bacteria. Recombination is particularly important in repairing DNA, and is involved in translation and mismatch repair (Lusetti and Cox 2002; Kowalczykowski 2015). Replication pauses when there is DNA damage, allowing RecA to use recombination to fix the DNA and/or replication fork (Lusetti and Cox 2002). Recombination is required if there is damaged DNA that is bypassed because of replication fork stalling (Cox 2007). RecA is what assists in fixing this damaged DNA and ensuring that replication proceeds as planned (Cox 2007). It does this by binding homologous pieces of DNA on each strand and bringing them together to promote the exchange of DNA between strands (Cox 2007; Kowalczykowski 2015). RecA and it's complexes can be affected by regulatory proteins and single stranded DNA-binding proteins in a positive manner, by stimulating DNA strand exchange, and in a negative manner by inhibiting RecA complex formation (Kowalczykowski 2015). The multilevel and precise regulation of RecA prevents the accidental deletion or recombination of an important homolog (Cox 2007).

Recombination and genomic location

When considering distance from the origin of replication, it appears as though recombination repair may be more prevalent near the origin because genes at this location are at a higher copy number (Sharp et al. 1989; Schmid and Roth 1987). Additionally, low levels of recombination have been found within the left and right physically folded structures of the chromosome (macrodomains) near the origin of replication (Wang et al. 2013). Any rearrangements or recombination that significantly impact the folding of macrodomains are more deleterious than inversions within the domains (Rocha 2008). Interestingly, there have been cases in *E. coli* where recombination near the origin of replication is symmetrical in order to conserve distance between the origin and other replication initiators such as DnaA (Frimodt-Moller et al. 2015).

Selective pressures and recombination

Recombination does not operate without selective constraints. Rearrangements clash with the overall organization of genes on bacterial chromosomes because it has the potential to mess up this conserved organization (Rocha 2004a). So, there is a trade-off between positive selection and chromosome organization (Rocha 2004a).

The effect of recombination on gene composition is stronger at synonymous positions most likely due to purifying selection on protein coding sequences (Lassalle et al. 2015). Selection is

often acting to prevent loss, where as recombination helps maintain variation between closely related taxa (Ochman et al. 2000). Recombination is known to enhance the efficency of selection by breaking linkage among sites (Lassalle et al. 2015). For example, when recombination is low and selection is strong, beneficial genes can cause whole genomic regions to become fixed (Maddamsetti and Lenski 2018). Although this might not always be related to selection, genes that were physically linked to genes causing recombination had a strong transmission advantage, regardless of selective advantage (Maddamsetti and Lenski 2018). Additionally, recombination between chromosomes can break linkage between beneficial or deleterious mutations and the rest of the genome, which can cause individual genes to be fixed instead of whole genomes (Maddamsetti and Lenski 2018). This happens when there are particularaly high levels of recombination (Maddamsetti and Lenski 2018).

1.2.3 Inversions

Inversions are one particular type of genomic reorganization. They can help provide genetic diversity (Hughes et al. 2000; Belda et al. 2005) and assist in speciation and adaptation (Kresse et al. 2003). Inversions additionally promote spontaneous genome rearrangements (Sun et al. 2012). There has been evidence that inversions are non-random and provide specific functions in bacterial genome evolution (Kresse et al. 2003). In some cases inversions are the only source of rearrangement in bacteria (Romling et al. 1997).

Mechanisms for inducing inversions

Although we are not completely sure of the control processes for inversions, it has been speculated that a homologous recombination pathway might be involved (Cui et al. 2012). Large inversions tend to be caused by homologous recombination (Sekulovic et al. 2018). There have been many cases where inversions in bacteria happen because of changes in the external environment (Zieg et al. 1978; Hill and Gray 1988; Gally et al. 1993; Serkin and Seifert 2000; Rentschler et al. 2013; Blomfield 2015). Changes in the environment can also cause these inversions to revert back to their original state (Zieg et al. 1978). Inversions can also be induced depending on the growth phase of the bacteria (Zieg et al. 1978).

The sequence composition and genomic location also play a role in the likelihood of inversions. There are some "hotspots" for bacterial inversions where inversions repeatedly occur (Zieg et al. 1978; Schmid and Roth 1983; Mahan and Roth 1988; Segall et al. 1988; Segall and Roth 1989; Mahan and Roth 1991; Alm et al. 1999; Glaser et al. 2002; Sibley and Raleigh 2004; Raeside et al. 2014; Sekulovic et al. 2018). For example, some sections of the *Salmonella* genome invert frequently and are quite permissive to inversions (Segall et al. 1988). These permissive and non-permissive sections of the genome appear to be universal for *Salmonella* genomes (Segall et al. 1988). This is thought to be related to chromosome position, and not near by sequence composition (Segall et al. 1988). Additionally, the over expression of RecA appears to increase the reversion of some inversions (Cui et al. 2012). Repetitive regions (Naseeb et al.

2016), pathogenicity islands, mobile elements, or duplicated regions can be home to inversions (Furuta et al. 2010). Repeated sequences, such as duplications, have the ability to increase the frequency of inversions (Cui et al. 2012). Inversions happen between substantial stretches of sequence homology scattered within the genome such as Ribosomal Ribonucleic Acid (rRNA) and Transposable Elements (TEs) (Le Bourgeois et al. 1995; Gray 2000; Parkhill et al. 2003).

As with HGT, mobile genetic elements such as IS elements are involved with inversions or border inverted regions (Schneider and Lenski 2004). Interestingly, when homologous recombination occurs between IS elements, it can cause large inversions or deletions to occur (Reif and Saedler 1975; Louarn et al. 1985; Schneider et al. 2000).

Detecting inversions

Small genomic inversions can create specific and easily identifiable signatures in sequencing data sets (Sekulovic et al. 2018). This can be obtained through 454 sequencing technologies (Sun et al. 2012), microscopy (Zieg et al. 1978), restriction sites (Zieg et al. 1978), southern blot (Hill and Gray 1988), comparing sequencing data with a closely related organism (Cui et al. 2012), and/or through microarray gene expression (Cui et al. 2012). There have been cases where inversions tend to have a GC skew with more C's than G's, which can be used to identify inverted regions (Merrikh and Merrikh 2018).

Genomic location and inversions

As with many other molecular traits, there seems to be chromosomal organization of inversions that changes with distance from the origin of replication. Some researchers have found deleterious inversions in the terminus regions of bacterial genomes (Francois et al. 1990), however other studies have found an inversion tolerant section near the terminus of replication (Guijo et al. 2001). In a contrasting study in *E. coli*, inversions were six times more likely to happen near the origin of replication than the terminus (Hendrickson et al. 2018). Hendrickson et al. (2018) believe that this is linked to the disruption in Architecture IMparting Sequences (AIMS) distribution along the genome. AIMS are strand-biased repetitive elements which act during DNA segregation and are positively correlated with proximity to the replication terminus (Hendrickson et al. 2018). This has lead to the proposal that the terminus region of bacterial genomes is made up of both permissive and non-permissive zones interspersed (Guijo et al. 2001). If inversions include the terminus (Alokam et al. 2002), they can potentially shift the location of the terminus (Kresse et al. 2003), which could disrupt genes in good, bad, or neutral ways (Hill and Gray 1988; Segall et al. 1988; Kresse et al. 2003; Naseeb et al. 2016). Successive inversions can impact the replication of each half of the replicon by changing the replicon half lengths (Raeside et al. 2014).

Since genomic location impacts a number of molecular trends in bacterial genomes, there are typically strong selective forces acting to conserve organization along the origin-terminus of replication axis. Inversions can impact the genomic location of a gene and in particular, the

gene order. There have been a number of studies with examples of symmetric inversions around the origin of replication (Segall et al. 1988; Roth et al. 1996; Eisen et al. 2000; Suyama et al. 2000; Tillier and Collins 2000; Moran and Mira 2001; Suyama and Bork 2001; Nakagawa et al. 2003; Canchaya et al. 2006; Cui et al. 2012; Khedkar and Seshasayee 2016; Repar and Warnecke 2017). These symmetric inversions have also been found at the level of whole proteomes (Eisen et al. 2000). If inversions are symmetrical around the origin of replication, then it is possible for gene order and relative distance from the origin of replication to be conserved (Eisen et al. 2000; Hendrickson and Lawrence 2006; Touzain et al. 2011; Repar and Warnecke 2017). Symmetrical inversions should also retain co-linearity of chromosome arms (Repar and Warnecke 2017) and help maintain even replichore halves (Tillier and Collins 2000; Mackiewicz et al. 2001; Repar and Warnecke 2017). It has also been postulated that symmetric inversions are related to essential and highly expressed genes being predominantly located near the origin of replication (Zipkas and Riley 1975). Symmetric inversions are not solely found in bacteria, there have been cases of symmetric inversions being prevalent around the terminus in archaea with multiple origins of replication (Repar and Warnecke 2017).

Ramifications of inversions

Inversions can have a number of effects on a variety of molecular properties. For example, inversions can impact gene gain and loss (Furuta et al. 2010), gene orientation (Huynen et al. 2001), and intracellular signalling (Sekulovic et al. 2018). These can all impact how conserved genes are or how they are co-regulated depending on their orientation (Huynen et al. 2001). Since inversions can promote recombination (Segall et al. 1988), there is a potential for inversions to promote the evolution of genes with novel function (Korneev and O'Shea 2002). Inversions have been shown to affect various aspects of bacteria life such as antibiotic resistance, susceptibility to chemical compounds and interactions with a host (Cui et al. 2012). There is evidence that structural and transcription changes due to rearrangement can cause variability in growth (Krinos et al. 2001; Colson et al. 2004; Raeside et al. 2014; Naseeb et al. 2016). However, most identified functional inversions are often involved with altering the surface structure of the bacteria (Zieg et al. 1978; Abraham et al. 1985; Marrs et al. 1988; Bahrani and Mobley 1994; Krinos et al. 2001; Patrick et al. 2003; Honarvar et al. 2003; Kuwahara et al. 2004; Zimmerman et al. 2009; Somvanshi et al. 2012; Anjuwon-Foster and Tamayo 2017). There have also been examples of inversions disrupting the macro- and micro-domains of the physically folded chromosomes (Segall et al. 1988; Raeside et al. 2014; Naseeb et al. 2016), which could be extremely harmful for the growth and well being of the bacteria.

One interesting role that inversions occupy is the role of a control switch. Some inversions have the ability to revert or reverse (Hill and Gray 1988; Louarn et al. 1985; Cui et al. 2012). This switching appears to be random, but maintaining an inverted or reverted state is organized (Cui et al. 2012; Sekulovic et al. 2018). These inversions and rearrangements allow the bacteria to switch between various states (Borst and Greaves 1987) such as having a flagella or not in

Salmonella (Zieg et al. 1977; Johnson and Simon 1985; Li et al. 2019), switching the mating type of *Saccharomyces cerevisiae* (Hicks and Herskowitz 1976; Herskowitz and Oshima 1981), and changing the surface coat composition of *Trypanosoma brucei* to evade the host immune defence (Vickerman 1978; Lamont et al. 1986).

Inversions and gene expression

As mentioned previously, inversions can be a way for bacteria to alter their gene expression (Zieg et al. 1977; Zieg et al. 1978; Sekulovic et al. 2018; Li et al. 2019). In some cases inversions can bring a silent gene to the expression site, "turning on" expression for that gene (Cerdeño-Tárraga et al. 2005). Other times, the gene expression alteration is non-specific and inversions can cause genes in areas close to the inverted region to be differentially expressed (Cerdeño-Tárraga et al. 2005; Wong and Wolfe 2005; Naseeb et al. 2016; Sekulovic et al. 2018). This again depends on the organism and specific inversion, because there have been cases where some inversions do not alter expression of nearby genes (Meadows et al. 2010).

This activation of gene expression mediated through inversions, is usually linked to moving genes closer to promoters or enhancers (Borst and Greaves 1987). Conversely, the removal of a promoter due to an inversion can cause inactivation of gene expression (Borst and Greaves 1987). There have been cases where short sequences flanking inversions are recognized by sequence specific recombinases which can allow the cell to determine which genes are available to the enzymes (Zieg et al. 1977; Borst and Greaves 1987). However with any rearrangements, errors occur which could negatively impact gene expression control (Borst and Greaves 1987).

Replication and inversions

The nature of bacterial replication partially dictates chromosome organization, therefore any genomic reorganization could impact this configuration. Small inversions tend to be tightly linked to replication (Gordon and Halliday 1995) and gene orientation. We know that genes are often found on the leading strand to avoid being in a head on collision with replication machinery (Rocha 2004b; Merrikh and Merrikh 2018). Small inversions can alter gene orientation and the location of a gene on the leading or lagging strand (Merrikh and Merrikh 2018). This could be used as a strategy for increasing the evolvability of a gene by inverting it to promote head on collisions with replication machinery, which are often under different selective pressures (Merrikh and Merrikh 2018). There have been some cases where transcription of a gene is altered via inversions. In yeast, there were large transcriptional changes in inverted genes that did not alter phenotype (Naseeb et al. 2016). Similarly, there have been cases where transcription of the whole genome is altered, not just the genes found within inverted segments (Naseeb et al. 2016).

Selective pressures on inversions

The rate of inversions varies between bacterial species (Furuta et al. 2010; Ely et al. 2019). Sometimes the same inversion can arise in two bacterial species through parallel evolution events

(Bochkareva et al. 2018). Other times inversions can be inherited vertically (Cui et al. 2012; Repar and Warnecke 2017) and fixed quickly (within 48 generations) in a population (Sun et al. 2012), and become stable over time (Segall et al. 1988; Badia et al. 1998; Kresse et al. 2003; Sun et al. 2012). In some cases inversions exist only for a short period of time and are eventually reorganized by other methods such as HGT or mutations (Parkhill et al. 2001; Sekulovic et al. 2018). Overall, inversion patters appear to be non-random and indicative of selection (Bochkareva et al. 2018).

1.3 Bacterial Genome Evolution

The way that bacteria receive and reorganize genetic information has shaped how bacterial genomes have evolved. Replication and other selective pressures have ultimately influenced the organization of the bacterial genome.

1.3.1 The role of replication

The nature of bacterial replication influences many aspects of the genome including genomic organization (Kopejtka et al. 2019). Replication creates more heterogeneity in bacterial genomes than what was previously through to be present, and assists in dictating gene distribution and sequence composition along the genome (Rocha 2004b). Many of the molecular biases we see in bacterial genomes are in fact due to the nature of replication (Rocha 2004b). Bacterial replication begins at the origin of replication and continues in both directions until the terminus of replication is reached. For bacteria with linear genomes there are two termini, one at each end of the chromosome arms. Bacteria also have the ability to concurrently have multiple rounds of replication happening on the same piece of DNA (Yoshikawa et al. 1964; Cooper and Helmstetter 1968). Replication is therefore thought to be one of the primary driving forces for creating gradients in genomic trends.

Physical chromosomal impacts of replication

When replication occurs in bacterial genomes there is a complicated physical condensing and unwinding of the DNA happening at the same time (Wang et al. 2013). This means that the DNA replication machinery and the physical chromosomal folding and movement are tightly coupled (Wang et al. 2013). Any time replication or the chromosomal movement is compromised, it could have detrimental effects on the bacteria (Wang et al. 2013).

Multi-repliconic bacteria and replication

Multi-repliconic bacteria have their genomes contained in multiple "chromosome like" structures. In these bacteria, coordination of replication between the primary chromosome and secondary replicons is very important. Each replicon has it's own factor that controls replication in that particular replicon, but it appears that in the case of *Vibrio cholerae* the replication control

factor for chromosome two can bind to a spot on chromosome one and control replication in chromosome two from the location on chromosome one (Baek and Chattoraj 2014).

Multi-repliconic bacteria have to control the initiation of replication in the chromosome and secondary replicons so that certain portions of each replicon are replicating at the correct time. In *Sinorhizobium meliloti* this is done through the delayed replication of the secondary replicons (pSymA and pSymB) (Flynn et al. 2010; Morrow and Cooper 2012). To maintain synchronization, due to the offset of different sequence lengths between secondary replicons and chromosome, the secondary replicons begin replication after the replication of the primary chromosome begins (Morrow and Cooper 2012). This reduces gene dosage throughout the secondary chromosome including near the origin of replication (Morrow and Cooper 2012; Cooper et al. 2010; Flynn et al. 2010). Species that grow faster require different timing in the replication of the primary chromosome and secondary replicons in order to allow replication to keep up with the growth of the organism (Morrow and Cooper 2012). This can impact various molecular trends such as causing increasing the gradient of substitution rates (Morrow and Cooper 2012). In contrast, organisms with a slow growth rate such as *Mycobacterium* and *Chlamydia*, have exhibited little or no variability in substitution rates (Mira and Ochman 2002).

Strand bias due to replication

One of the most prominent biases is the difference in molecular trends between the leading and lagging strand. Since replication continues in both directions starting at the origin of replication, and transcription and translation can occur at the same time, there are different selective pressures that impact each strand differently (Hyrien et al. 2013). It has been noted that most genes are located on the leading strand to avoid collisions between Deoxyribonucleic Acid Polymerase (DNAP) and Ribonucleic Acid Polymerase (RNAP) (Rocha 2004b). This makes replication more efficient and means that highly expressed genes are often found on the leading strand (Rocha 2004b). Genes that are in a "head on" orientation with DNAP and RNAP - genes that are in the opposite direction of DNAP and RNAP - frequently have a negative GC skew value (Merrikh and Merrikh 2018). Comparatively, most co-directional genes - genes that are in the same orientation as polymerase movement - have a positive GC skew (Merrikh and Merrikh 2018). Although, we do see some genes that are not in a co-directional orientation, which must confer some sort of benefit to having genes in the opposite orientation (Merrikh and Merrikh 2018). Merrikh and Merrikh (2018) explored this and found that head-on genes had a higher non-synonymous mutation rate compared to co-directional genes, and suggesting that this might be beneficial and impact evolvability of those genes. These head-on genes are retained over time and are enriched in common functions across species, which might mean that head-on genes are evolving at an accelerated rate, a potentially beneficial property (Merrikh and Merrikh 2018). This head-on orientation might also be a secondary effect for some other selective force at play (Merrikh and Merrikh 2018).

Genomic location and replication

Another major way that replication shapes genomic organization in bacteria is by creating differences between early and late replicating regions (Rocha 2004b). Typically polymerase is more accurate near the origin of replication and becomes less accurate as it moves towards the terminus (Niccum et al. 2019). This influences gene organization and sequence composition, creating a gradient between the origin of replication and the terminus (Rocha 2004b). Genes of similar function, particularly important ones, are often grouped together on the chromosome which can make transfer and replication easier (Rocha 2004a; Daubin and Ochman 2004; Wiedenbeck and Cohan 2011; Soucy et al. 2015).

1.3.2 Broad selective constraints

Previously, I have mentioned a few selective constraints to bacterial genomes as they pertain to the specific topics mentioned above. I will now go into more depth about broad selective constraints and phenomenon that occur in bacterial genomes. Any mutation that occurs in the genome, creates a basis for selection to act upon. These mutations produce a range of selective advantages and the effects of these mutations can range from strong to weak and anything in between (Ohta 1992). The stronger the constraint on the molecule, the slower the evolutionary rate and the lower the number of polymorphisms (Ohta 1992).

Estimating substitution rates

When talking about selection, substitution rates are often used as a metric to indicate what kind of selective forces are potentially acting on a gene or set of genes. There are two main types of substitutions: non-synonymous which cause a change in the amino acid sequence of a protein, and synonymous which do not result in a an amino acid replacement. The potential change in amino acid sequence means that selective pressures may act differently on each type of substitution. Synonymous substitutions are therefore more likely to accumulate within a gene or genome than non-synonymous substitutions.

The most common method to estimate the Non-synonymous Substitution Rate (dN) and the Synonymous Substitution Rate (dS), are described in detail in Li et al. (1985). This method uses the concept of degenerate sites, where 4-way degenerate site will code for the same amino acid regardless of a mutation at that codon site - like the third codon position of glycine - and a 0-way degenerate site results in a different amino acid for any mutation, like the second codon position of any amino acid. The number of each category of degenerate sites is counted up while keeping track of transitions, transversions and number of differences at each site. This is then used to calculate dN and dS. Since dS concerns the same amino acid being substituted, there is typically less variation in dS and there will be less changes between and within organisms. dN on the other hand concerns different amino acids being substituted, and therefore there is more variation in this rate between and within organisms.

When discussing selection, the ratio of dN and dS is often computed and visualized as $\omega = dN/dS$. The ω ratio allows us to predict if the genes will be maintained or deleted over time. If ω for a gene is larger than 1, the gene is likely under positive selection and therefore is beneficial to the organism and will likely be maintained in the genome over time. If ω is less than 1, the gene is likely under purifying or negative selection, and therefore is deleterious to the organism and will likely not be maintained in the genome over time. If ω is equal to (or close to) 1, the gene is under neutral selection, and is neither beneficial nor deleterious to the organism.

There are a number of different programs that estimate substitution rates and they all use slightly different theoretical concepts to do this. For any method, it is important to keep in mind the timescale being used to estimate substitution rates and if one is more interested in looking at a short term or long term rates (Ho et al. 2011). For example, one can look at spontaneous mutation rates measured over a small number of genomes or lower substitution rates over geological time frames (Ho et al. 2011). Typically, young estimates reflect non-lethal mutation rates and old estimates reflect substitution rates (Ho et al. 2011). Non-synonymous sites have a stronger time dependence than synonymous sites, and all rates can be over estimated if the time calibration is wrong, although this is most detrimental for short time scales (Ho et al. 2011). With increased sequence divergence, there is a decrease in the estimation bias rates (Ho et al. 2011). Substitutions are mutations that have been fixed in any of the diverging lineages being analyzed (Ho et al. 2011). Mutations appear all the time, but are usually lost within a few generations, therefore their contribution to evolution is relatively small (Ohta 1992). We will present data exploring mutations that have been subject to selection (substitutions), and the genomic patterns associated with them. Over time, natural selection will tend to remove deleterious alleles, which make up a large portion of spontaneous mutations, therefore, substitution rates are usually lower than mutation rates (Ho et al. 2011).

Gene characteristics impact selection

There are a number of factors related to the molecular make-up of a gene that can be used to predict what kind of selective pressures will be acting on that gene. If a protein or gene is useful to the organism, there is a selective advantage to maintain it within that organism (Penny 2015). Therefore, if an important protein is altered in function because of mutations, this could lead to lethality (Penny 2015). If a gene or protein is not important, then mutations build and it is eventually lost (Penny 2015). However, non-useful alleles can be important because they provide a base for beneficial mutations to occur, allowing the allele or gene to adapt (Penny 2015).

The length of a gene can also impact selection. Longer genes have stronger purifying selection and a negative correlation between the non-synonymous and synonymous mutation rate ratio (ω) and median protein coding length genes (Novichkov et al. 2009). However, this could be because current analytic methods are much better at detecting positive selection (Daubin and Ochman 2004; Yang and Dos Reis 2010).

1.4 Organizing bacterial genetic information

We have previously mentioned that bacteria are extremely efficient and have a strict organization of their genetic information within their genomes (Le and Laub 2014). There are a few broad categories of genomic information and structure that apply to almost all bacterial genomes.

1.4.1 Physical genome structure

Bacterial genomes come in three broad categories: circular, linear, and multi-repliconic. Most bacteria have their entire genome contained in a single circular chromosome, including *E. coli* and *Bacillus subtilis*. A few other bacteria have their genomes contained in a single linear chromosome such as *Streptomyces*. Additionally, there are other bacteria who's genomic content is split up into multiple chromosome-like structures called replicons. Multi-repliconic bacteria can have any number of replicons, including several megareplicons (Martin-Didonet et al. 2000). The numerous replicons within a multi-repliconic bacteria could be a combination of circular or linear repliconic structures. Some examples of multi-repliconic bacteria include *S. meliloti*, *V. cholerae*, and the *Azospirillum* species.

Within each bacterial replicon, there are common physical and structural constraints and phenomenon across bacterial species. The physical compaction and folding of genomes is another highly organized aspect of bacterial genomes. These compartmentalize bacterial genomes into Megabase Pair (Mbp) domains, which can vary in size and positioning in various bacteria (Le and Laub 2014; Badrinarayanan et al. 2015). These large domains are further broken down into macro and micro domains which are shaped by transcription (Le and Laub 2014; Badrinarayanan et al. 2015), and provide global organization of the replicon (Cagliero et al. 2013). Genes that are found within the same macro domain recombine more frequently and typically interact with loci from the same domain (Wang et al. 2013; Le and Laub 2014; Badrinarayanan et al. 2015). Interactions happen the most between functionally similar domains, so the most interaction happens within the same domain rather than between domains (Wang et al. 2013; Le and Laub 2014). This same logic extends to the chromosome arms or replichore halves, where genes on the same replicon half or chromosome arm interact more than genes on the opposites chromosome arm or replicon half (Wang et al. 2013; Le and Laub 2014). Interestingly, the domains near the origin of replication have a high frequency of interaction where as the domains near the terminus have a low frequency of interaction (Cagliero et al. 2013).

These domains can also impact rearrangements, physically bringing together areas of the chromosome to promote or inhibit recombination (Boccard et al. 2005; Esnault et al. 2007). The structural maintenance of the chromosome helps to promote colinearity of the chromosome arms (Wang et al. 2013; Le and Laub 2014), which results in similar packing density on both arms (Wang et al. 2013). Supercoils are particularly important in guiding recombination and they contain resolution sites, which are indicative of recombination rates (Badrinarayanan et al.

2015). These res sites can be found up to approximately 100 Kilobase Pairs (Kbps) apart (Rocha 2008; Le and Laub 2014). Supercoiling can additionally impact gene expression by bringing far genome segments closer together or vice versa (Rocha 2008).

There are a number of particular sequences or proteins that exhibit some level of control over the genomic architecture in bacteria. One such element is the AIMS. AIMS are repetitive elements preferentially found on the leading strand (Hendrickson et al. 2018), where movement to the opposite strand (lagging) has demonstrated detrimental effects in *E. coli* (**Ptac:06**). They act during DNA segregation and are positively correlated with proximity to the replication terminus (Hendrickson et al. 2018). Disruption in the distribution of AIMS can cause significant changes to the overall structure of the genome (Hendrickson et al. 2018).

Various nucleoid associated proteins are involved with not only genomic architecture, but gene expression regulation as well. The Histone-like Nucleoid-Structuring (H-NS) protein maintains and controls chromosome compaction and structure (Grainger et al. 2006), while also globally regulating transcription (Johansson et al. 2000; Kahramanoglou et al. 2011). H-NS has the ability to repress the transcription of non-essential genes (Browning et al. 2000; Hommais et al. 2001; Dorman 2004; Fang and Rimsky 2008; Dillon and Dorman 2010; Ali et al. 2012; Singh et al. 2016), playing an important role in silencing genes recently acquired via HGT (Dorman 2004; Oshima et al. 2006; Dorman 2007; Ali et al. 2014; Higashi et al. 2016). Similar to H-NS, the Factor for inversion stimulation (Fis) protein is a nucleoid associated protein that is involved in regulating expression (Kelly et al. 2004; Paul et al. 2004; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019) and a number of genomic architecture properties (Kahmann et al. 1985; Johnson et al. 1986; Thompson et al. 1987; Haffter and Bickle 1987; Ball and Johnson 1991; Messer et al. 1991; Filutowicz et al. 1992; Wold et al. 1996; Wu et al. 1996; Schneider et al. 1997; Schneider et al. 2001; Travers et al. 2001; Ryan et al. 2004; Dhar et al. 2009; Tsai et al. 2019; Dages et al. 2020). There is evidence that the Fis protein facilitates the activation of transcription through close interaction with promoters and RNAP or by altering local genome architecture (Kelly et al. 2004; Paul et al. 2004; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019), and any changes in the binding of Fis is impacted by genome-wide alterations in transcription (Grainger et al. 2006).

AIMS and nucleoid associated proteins, along with other genomic architecture elements, are responsible for maintaining and controlling the physical configuration of bacterial genomes, which is crucial to the proper functioning of the organism.

As mentioned previously, most chromosome organization is dictated by the nature of bacterial replication. Chromosome organization is defined as the distribution of genes relative to replication, segregation, or expression, which can be biased by expression levels, essentiality and function (Rocha 2004a). This chromosome organization is subject to genetic variability because of HGT, mutation rates, recombination, and other intra-genomic rearrangements (Rocha 2004a). Typically, closely related species have similar chromosome organization when compared to more

distantly related species (Tillier and Collins 2000). However, there are cases of chromosome divergence in closely related species of bacteria (Eisen et al. 2000; Hughes et al. 2000).

1.4.2 Gene classification

Although globally, bacterial genomes are organized into the above mentioned macro and micro domains, the specific genomic location of genetic elements depends on a number of factors such as the functional class of that element, mutations, recombination, and other intra-genomic rearrangements (Casillas and Barbadilla 2017). Bacterial genomes can be broadly split up into two categories: the core genome and the accessory genome (Galardini et al. 2013). The core genome consists of genes that are essential to the function of the organism and are generally conserved within strains (Galardini et al. 2013). The accessory genome encompasses genes that are used for non-essential functions such as local environmental adaptation. This accessory portion is generally used to distinguish phenotypically and genetically between different strains and strain specific behaviour (Galardini et al. 2013; Tettelin et al. 2008; Biondi et al. 2009). For example, in the case of *Streptomyces*, the core genome is broadly conserved across all *Streptomyces* species and located near the origin of replication (Redenbach et al. 1996; Choulet et al. 2006). Where as the accessory genome, located near the terminal ends of the chromosome, is highly variable (Redenbach et al. 1996; Choulet et al. 2006). The all encompasing bacterial genome is the pan-genome. A bacteria's pan-genome is simply all genes within the core and accessory genome (Medini et al. 2005).

The differences in gene content between the core and accessory regions of bacterial genomes dictate which selective pressures are primarily acting on each of those sections. With regards to horizontally transferred genes, non-core genes may actually be neutral or nearly neutral to the recipient (Daubin and Ochman 2004). The types of genes transferred to various regions differs. For example, conjugative elements were mainly inserted in the core regions of *E. coli* (Tidjani et al. 2019). However, Insertion/Deletion (indel) rates were about 5 times higher in the chromosome arms (accessory regions) compared to the core regions (Tidjani et al. 2019).

The placement of core and accessory genes within bacterial genomes is non-random and appears to be tightly linked to replication. The orientation of a gene is strongly correlated with the direction of DNA replication (Zeigler and Dean 1990; Kunst et al. 1997). The leading strand typically has more essential genes compared to the lagging strand (Rocha 2004b; Rocha 2008) and this appears to be conserved across bacteria (Rocha 2004b).

1.4.3 Genomic islands

The concept of Genomic Islands (GIs) is centred around the idea that genes of similar function or similar regulation methods, are grouped together on bacterial genomes (Rocha 2004a; Daubin and Ochman 2004; Rocha 2008; Wiedenbeck and Cohan 2011; Soucy et al. 2015). These GIs have a family of functions that can vary in size from about 10Kbps to 200Kbps, and have arisen multiple

times through convergent evolution (Juhas et al. 2009). GIs often differ between closely related strains and were once mobile (Juhas et al. 2009). They are often inserted at Tranfer Ribonucleic Acid (tRNA) genes flanked by repeats (Juhas et al. 2009). GIs are beneficial because they often carry insertion elements and transposons, which can offer a selective advantage by increasing genetic diversity (Juhas et al. 2009). These mobile elements allow some GIs to be self mobile, while others must rely on plasmids or HGT to be inserted into a new genome and replicate with the host genome (Juhas et al. 2009). Just as with normal HGT, the host background can help facilitate transfer if it is sufficiently similar to the donor genome (Juhas et al. 2009; Wiedenbeck and Cohan 2011). This transfer is thought to be linked to regulation and the environment and is not just random (Juhas et al. 2009). Similar to the typical HGT process, once the GIs are transferred to a new host cell, they need to be up-taken and integrated successfully in order for the information to become useful (Juhas et al. 2009). HGT is actually facilitated by GIs because the islands can transfer parts of the host genome with the GI when it is transferred (Juhas et al. 2009).

Pathogenicity Islands (PIs) are one subclass of GIs and as the name suggests, contain genes related to pathogenicity (Juhas et al. 2009). They have the ability to transform a pathogenic organism into a non-pathogenic organism, and vice versa (Ochman et al. 2000). In general, GIs carry novel genes compared to the rest of the genome with unique genomic signatures such as codon bias, GC content, and nucleotide frequencies (Juhas et al. 2009). GIs allow for adaptation and novel innovation like HGT (Juhas et al. 2009). Some of these novel gene classes include antibiotic resistance, virulence factors, and adaptation to new lifestyles and environments (Juhas et al. 2009). Again, the host plays a critical role in the expression of these new traits and genes found in GIs via regulators (Juhas et al. 2009).

1.4.4 Spatial molecular trends

As mentioned previously, the genomic location of various genes often depends on their function and links to replication. This non-random distribution of genes creates profound gradients of various molecular trends such as gene expression, mutation rates, and substitution rates across bacterial genomes. In this work, I will be exploring the patterns of these various molecular trends and provide data on their impact.

Spatial organization of the core and accessory genome

There have been broad findings with respect to the placement of core and accessory genetic information within bacterial genomes. It is generally accepted that core genes are typically located near the origin of replication, and accessory genes are located near the terminus of replication (Sharp et al. 2005; Couturier and Rocha 2006; Cooper et al. 2010; Morrow and Cooper 2012; Flynn et al. 2010; Kopejtka et al. 2019), although the exact location of the core and accessory regions varies between species. It is speculated that genes near the terminus are more prone to recombination, while genes near the origin have a higher prevalence of recombination

repair (Sharp et al. 1989; Flynn et al. 2010). Genes near the terminus therefore often have more variation and are less conserved compared to those near the origin of replication (Sharp et al. 1989; Flynn et al. 2010). Genes of similar function, especially essential ones, are often controlled by the same regulators or promoters, so having these genes located physically close together assists in this regulation (Rocha 2004a; Daubin and Ochman 2004; Rocha 2008; Wiedenbeck and Cohan 2011; Soucy et al. 2015).

However, there have been some bacteria where this does not appear to be the case. The accessory genome can move around the genome and become interspersed with core genome segments (Siguier et al. 2014). In a few species of *Rhodobacteraceae*, this creates a mosaic pattern of core and accessory genes dotted throughout the genome (Kopejtka et al. 2019). In still other species of this family, the complete opposite placement of core and accessory genes have been found. In these species, core genes were concentrated near the terminus of replication, not the origin (Kopejtka et al. 2019). It was speculated that this trend was due to uneven distribution of core genes between the replichores (Kopejtka et al. 2019).

Spatial organization of mutations and substitutions

Mutations are the basis for creating a variety of phenotypes. Selection can then act upon these differences to sustain or remove genotypes and associated phenotypes over time. Although mutation rates are relatively constant between bacteria (Hanage 2016), it is well known that substitutions have a non-random distribution around the genome which varies by gene and organism (Sharp et al. 1989; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012).

Substitution rates (dN and dS) and their ratio ω, typically increase with distance from the origin of replication (Cooper et al. 2010; Morrow and Cooper 2012). In some cases the mutation rate is up to two times higher near the terminus than near the origin of replication (Sharp et al. 1989). It is speculated that genes near the terminus are more prone to recombination, while genes near the origin have a higher prevalence of recombination repair (Sharp et al. 1989; Flynn et al. 2010). Genes near the terminus therefore often have more variation and are less conserved compared to those near the origin of replication (Sharp et al. 1989; Flynn et al. 2010). Additionally, genes found within the core genome are typically located near the origin of replication, while genes associated with the accessory genome are found near the terminus (Couturier and Rocha 2006; Flynn et al. 2010). The placement of these two gene categories may explain why near the origin, gene expression and essentiality are high and mutation rate is low (Sharp et al. 2005; Couturier and Rocha 2006; Flynn et al. 2010).

However, not all studies looking at mutation or substitution rate have seen the same positive correlation with distance from the origin of replication. In some cases it is said that only approximately 5% of variance in substitution rates along the chromosome are due to distance from the origin of replication (Rocha 2004b). Some studies found no correlation between distance from the origin of replication and the frequencies of mutations, but they did find mutation rate to

vary with position along the *E. coli* chromosome (Martina et al. 2012; Juurik et al. 2012). In some cases, substitution rates are more constant throughout the replichore except at the G+C poor terminus (Daubin and Perriere 2003). There have been a few other studies that found a significantly high peak of mutations near the terminus in *E. coli* (Senra et al. 2018) and in *Teredinibacter turnerae* (Yang et al. 2009). They found that mutations tend to cluster and are found close together (Senra et al. 2018), suggesting that spontaneous mutations may not be independent with respect to position (Amos 2010; Schrider et al. 2011; Sung et al. 2015). Other investigations found no positive correlation with mutation rates and distance from the origin of replication and instead found that intermediate positions had a higher non-synonymous mutation rate than positions farther from the origin in *E. coli* (Ochman 2003) and *Salmonella enterica* (Hudson et al. 2002; Ochman 2003). In a more recent study, Dillon et al. (2018) found that base substitution mutation rates vary in a wave like pattern over intervals of less than 1Kbp in *Burkholderia* and *Vibrio* (Dillon and Smith 2017; Wei et al. 2018). Concurrently replicated segments were found to have similar rates (Dillon et al. 2018). This wave like pattern was also seen in *E. coli* (Long et al. 2016; Niccum et al. 2019), *B. subtilis* (Niccum et al. 2019), *Pseudomonas fluorescens* (Long et al. 2014) and *Pseudomonas aeruginosa* (Dettman et al. 2016). Foster et al. (2013) observed the same wave like pattern in base pair substitutions in *E. coli*. The wave like pattern of these rates is thought to be related to cell cycle functions and not sequence composition (Dillon et al. 2018). Additionally, Deoxyribonucleotide triphosphate (dNTP) pools and the concentration of dNTPs can impact this wave like pattern (Niccum et al. 2019). There is some evidence that the types of mutations vary depending on the replication timing of certain genes (Dillon et al. 2015). It appears as though early replicating DNA has more AT mutations, and late replicating DNA has more GC mutations (Dillon et al. 2015). However, the authors note that this was determined using a small data set, and more data would be needed to extrapolate these findings (Dillon et al. 2015). All of these exceptions to the previously established molecular trends raises questions about just how universal these phenomenon are.

As with other molecular trends, the sequence composition of adjacent nucleotides influences substitution and mutation rates (Foster et al. 2018). In one extreme case, mutation rates were increased up to 75 fold in *B. subtilis* due to the composition of near by nucleotides, although there was no clear pattern about which mutations are impacted the most (Sung et al. 2015). This can have profound impacts on selection inferences, codon bias, and the appearance of multiple mutations at one site (Sung et al. 2015). It is thought that these context-dependent mutation patterns are symmetrical around the origin of replication and may arise from elevated mutation rates at certain motifs (Sung et al. 2015). Additionally, the chromosomal context around the mutation site is more important for modulating any frame shift mutations (Martina et al. 2012). In some cases the "near by" nucleotides are only a few base pairs away on either the $5'$ or $3'$ end (Dillon et al. 2018), but there have been instances where nucleotides (specifically a GATC sequence) are located up to 2Kbps away can have an impact on the mutated site (Martina et al. 2012). Chromosomal context can also affect frame shift (Martina et al. 2012) and mismatch

repair efficiency in both a positive and negative way (Foster et al. 2018). Mismatch deficient strains accumulate more (approximately 120 fold) base pair substitutions than mismatch repair proficient strains (Foster et al. 2018). However, this was dependent on the media the bacteria were grown in (Foster et al. 2018).

There has been a number of studies that found that Single Nucleotide Polymorphisms (SNPs) often reside near each other, forming clusters in the human genome (Amos 2010). These clusters can differ in size, frequency and mean SNPs (Amos 2010). The same trend is seen with substitutions, where simultaneous nucleotide substitutions happen within short stretches of DNA (Schrider et al. 2011).

The non-random distribution of synonymous and non-synonymous substitutions around the genome varies by gene and organism, creating codon bias within and between bacterial genomes (Sharp and Li 1986). Due to this bias, some segments of the genome are more mutable than others creating "hot spots" around the genome. Some of these hot spots are classified by mononucleotide runs which create an abnormal amount of base pair substitutions in these regions (Foster et al. 2018). In extreme cases, some sites can have over 500 frame shift mutations (Benzer 1961). Some hot spots have frequent indels (Streisinger et al. 1966; Farabaugh et al. 1978), which can be problematic and create sections of the genome where the DNA can become misaligned, causing indels in replication (Kunkel 2004).

Mutation rate has been found to be altered by the growth rate or stage of bacteria. When *E. coli* are growing slowly, mutation rate is higher compared to when *E. coli* is growing quickly (Maharjan and Ferenci 2018). Mutation rate in *E. coli* is also closely linked to oxygen availability and aerobic versus non-aerobic environments (Maharjan and Ferenci 2018). Varying growth media and temperatures can change the base pair substitution spectra (Foster et al. 2018), and this is most likely due to differences in growth rate and physiological state of the cells (Foster et al. 2018).

Spatial organization of gene expression

Gene expression plays a role in the distribution of genes as a function of distance from the origin of replication (Cooper and Helmstetter 1968; Chandler et al. 1975; Chandler and Pritchard 1975; Bremer and Churchward 1977; Schmid and Roth 1987; Sousa et al. 1997; Couturier and Rocha 2006; Bryant et al. 2014; Gerganova et al. 2015). This includes the chromosomal position of gene complexes consisting of genes and their promoters and regulators, which have the ability to turn genes on/off depending on the replicon location (Gerganova et al. 2015). Typically, genes with higher expression levels are located near the origin of replication (Couturier and Rocha 2006; Junier 2014; Gerganova et al. 2015; Lato and Golding 2020a) and this is thought to be a selective advantage for increased growth rate, copy number, and expression (Couturier and Rocha 2006; Junier 2014). Genes that are relocated to different repliconic positions can cause such a dramatic change in gene expression that there are changes in the cellular phenotype of

the bacteria, which can create differences in fitness (Gerganova et al. 2015). Although overall, there are a low number of highly expressed genes in bacteria (Rocha 2004b). Highly expressed genes are over represented on the leading strand to avoid collisions between DNAP and RNAP (Rocha 2004b; Mirkin and Mirkin 2005; Washburn and Gottesman 2011; Block et al. 2012). Interestingly, in *E. coli* the non-essential genes are often highly expressed, while essential genes are expressed at low levels (Rocha 2004b). Therefore, gene function could be the primary reason for impacting the expression timing of genes (Rocha 2004b).

Due to the bidirectional nature of bacterial replication and the ability for bacteria to undergo concurrent rounds of replication on the same piece of DNA, this creates an increased copy number for near the origin of replication (Sousa et al. 1997; Block et al. 2012). This establishes a gradient of higher gene dosage near the origin of replication and lower gene dosage near the terminus (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Gerganova et al. 2015; Sauer et al. 2016), which has been suggested as one of the ways that chromosome position alters gene expression (Block et al. 2012). This dosage effect can create expression differences of up to 5 fold higher near the origin of replication compared to the terminus (Schmid and Roth 1983; Block et al. 2012; Bryant et al. 2014; Sauer et al. 2016). However, some speculated that this dosage effect potentially impacts only constitutive expression (Block et al. 2012). Most of the studies that look at dosage effects move the gene and/or its promoters and operons to predetermined locations along the bacterial genome. This means that we are unsure of how dosage changes with naturally regulated operons (Garmendia et al. 2018). Garmendia et al. (2018), found that distance from the origin of replication had a small effect on growth rate in non-nutrient limiting envoronments. However, in growth limiting environments, there was a strong dosage effect when genes were located far from the origin of replication (Garmendia et al. 2018). It appears as though the bacteria are quickly adapting gene expression to this limited environment and can change protein concentrations to restore normal growth rate in nutrient limiting environments, overcoming gene dosage effects (Garmendia et al. 2018). Garmendia et al. (2018), therefore argues that gene dosage is not the primary force selecting for genes to be close to the origin of replication, it may be a co-regulation of many genes that require these elements to be located near the origin of replication (Garmendia et al. 2018). This may be why it is difficult to see gene dosage effects in highly expressed genes because they are often very well regulated (Yates et al. 1980; Jinks-Robertson and Nomura 1982; Aseev et al. 2008; Brandis et al. 2016), therefore masking gene dosage effects by correcting deficiencies in protein concentrations through expression regulation (Garmendia et al. 2018).

Having higher gene expression values near the origin of replication has been linked to physical constraints and processes of bacterial replication (Képes 2004; Peter et al. 2004; Jeong et al. 2004; Rocha 2004a; Block et al. 2012). For example, replication errors increase as replication moves farther from the origin of replication (Courcelle 2009). This impacts the placement of highly expressed and important genes where errors in replication could be detrimental to the gene product and the organism. Therefore, genes that are highly expressed and also essential to

the survival of the organism are often located near the origin of replication and on the leading strand to further avoid collisions between DNAP and RNAP (Rocha 2004b; Washburn and Gottesman 2011; Block et al. 2012). These core genes make up the majority of bacterial genomes, so intuitively we should have a higher concentration of genes near the origin of replication. The variation in gene expression with genomic location is predicted to be due to a number of complicated and intertwining factors such as transposon insertion events (Gerdes et al. 2003), gene order and conservation (Mackiewicz et al. 2001; Flynn et al. 2010), replication (Couturier and Rocha 2006), and nucleotide composition (Mackiewicz et al. 1999; Karlin 2001; Sharp et al. 2005). Additionally, it is important that transcription and translation are synchronized in the same replication state, which could be assisted via gene dosage (Garmendia et al. 2018).

Other features of chromosome organization have been shown to impact gene expression (Bryant et al. 2014). Genes that are controlled by the similar promoters and regulators are usually in the same spot on the replicon, creating conserved clusters throughout the genome (Rocha 2004a). These clusters make it easier to control expression of these genes all at once (Rocha 2004a). Operons located farther from the origin of replication show a smaller response to standard transcriptional activators (Rocha 2004a). Experiments done using a reporter gene cassette comprised the Lactose Operon (lac) promoter in *E. coli*, has shown that moving the reporter cassette to different locations causes changes in expression that were controlled by Transcription Factors (TFs) located outside of the mobile cassette region (Bryant et al. 2014). Therefore, genes can be controlled by not just copy number, but by TFs that are not necessarily located near the gene of interest (Bryant et al. 2014). However, there have been some experiments that moved genes controlled by only TFs located within the regions of DNA that was being re-located across the genome, and a decrease in expression was still found when moved to a location near the terminus of replication (Garmendia et al. 2018). There have been a few cases where neighbouring gene expression can influence promoters both up and downstream of the gene (Loconto et al. 2005; Bryant et al. 2014). It therefore appears as though any change in chromosomal location of a gene can impact the overall expression (Garmendia et al. 2018).

Gene expression being impacted by neighbouring molecular elements does not just stop at promoters. Local sequence composition can also alter the maximum expression levels at different chromosome positions (Block et al. 2012). Genes that are rich in GC content tend to be expressed at higher levels (Jansen and Gerstein 2000), which may reflect the inherent GC bias of heavily transcribed DNA (Jansen and Gerstein 2000; Marin et al. 2003).

IS elements can alter or switch gene expression in adjacent genes (Reynolds et al. 1981; Saedler et al. 1981; Ciampi et al. 1982). Just like in HGT, local sequence effects at some positions can impact expression from translocated genes (Block et al. 2012). Translocation has the ability to impact chromosome position, gene orientation, and distance between the target gene and its transcription factor, which all have the potential to alter gene expression (Block et al. 2012). For example, in *E. coli*, if a TF has a new location far from the target gene, it could potentially

have to diffuse before it can work on the target gene, which could take some time and alter the temporal expression pattern of that gene (Block et al. 2012). However, when tested it was found that expression but not TF activity was impacted by chromosome position in this particular case (Block et al. 2012).

The physical chromosome structure, such as supercoiling, can also impact gene expression (Miller and Simons 1993; Bryant et al. 2014; Gerganova et al. 2015). Although, only high supercoiling activity was affected by expression changes (Bryant et al. 2014), in some cases, supercoiling did not impact expression with changing genomic location of some *Pseudomonas* genes (Sousa et al. 1997).

The codon bias of a gene or region of the genome can also impact gene expression. Genes that are highly expressed have been found to have higher codon bias (Gouy and Gautier 1982; Cannarozzi et al. 2010) because they are usually full of codons that are recognized by the abundant tRNAs which can be translated faster (Wright et al. 2004; Quax et al. 2015). This creates selective pressures on codon bias mediated by tRNA abundance that may only be acting on highly expressed genes (Wright et al. 2002; Wright et al. 2004). The nucleotide composition around a codon is non-random (Gutman and Hatfield 1989; Buchan et al. 2006) and therefore can impact the translational efficiency of that gene (Berg and Kurland 1997; Gustafsson et al. 2004; Quax et al. 2015).

Although most previous studies have found that gene expression decreased when moving away from the origin of replication (Couturier and Rocha 2006; Junier 2014; Lato and Golding 2020a), there have been some experiments that did not find a clear negative correlation. Wright et al. (2007), looked at statistically correlated gene pairs in *E. coli* and found that they are often separated by 100Kbps and are often located in areas of high transcription. Other studies of *E. coli* observed that sections of the chromosome with increased transcription rates were periodically found throughout the genome over 700-800Kbps ranges (Jeong et al. 2004). It is speculated that this periodic phenomenon is due to a combination of physical constraints of the chromosome, such as supercoiling, and DNA composition (Jeong et al. 2004; Képes 2004; Peter et al. 2004; Allen et al. 2006; Block et al. 2012).

Other spatial molecular trends

There are a few other molecular trends that have varying patterns along bacterial replicons. A number of these molecular trends are related to bacterial replication as mentioned earlier. Genes associated with RNAP are often located closer to the origin of replication (Rocha 2004b; Rocha 2008). As mentioned previously, replication impacts the gene content and organization of genes on the leading and lagging strands of the genome. The leading strand tends to have more genes and is more genomically stable (Rocha 2004a). This also creates differences in the GC content and codon bias for each of the strands (Rocha 2004b), where the leading strand is typically G-rich (Perriere et al. 1996; Francino and Ochman 1997; Grigoriev 1998; McLean et al. 1998; Guo 2011).

The preferential location of certain genes on the leading or lagging strand could in part be due to this nucleotide skew between strands. Some nucleotides (GTP and CTP), are more energetically expensive than others (Rocha and Danchin 2002), and are found at lower concentrations within bacterial cells (Danchin et al. 1984). In particular, CTP has a much lower concentration in the cell than other nucleotides (Danchin et al. 1984). A combination of the low concentration and energetically expensive nature of CTP may offer a selective advantage in genes containing fewer C's (Marin and Xia 2008). This would be particularly important for highly expressed genes. Genes that are classified as essential in bacterial genomes are often highly expressed (Rocha and Danchin 2003). These essential and highly expressed genes likely need to be transcribed and translated efficiently, and are preferentially located on the G-rich leading strand (Rocha and Danchin 2003). Typically, the leading strand is less mutagenic than the lagging strand (Rocha 2004b), with more frequent base pair substitutions happening on the lagging strand (Lee et al. 2012; Dettman et al. 2016; Foster et al. 2018). Genes that are moved from one strand to the other adapt quickly to take on the new strand biases (Rocha 2004b).

Nucleotide proportions change with distance from the origin of replication. GC skew appears to oscillate a bit, but is generally positive for the right half of the replichore, and negative for the left half of the replichore (Blattner et al. 1997). GC skew additionally changes sign at the origin of replication in a number of bacteria (Lobry 1996; Kunst et al. 1997; Hyrien et al. 2013; Bhowmik et al. 2018). These changes in skew could be due to transcriptional effects (Freeman et al. 1998), the uptake of foreign DNA (Freeman et al. 1998), or inherant differences in GC content between the leading and lagging strands (Bhowmik et al. 2018).

Molecular trends in multi-repliconic bacteria

With respect to multi-repliconic bacteria, there have been uncommon findings, where the number of substitutions does not decrease with distance from the origin of replication. Dillon et al. (2015), found that substitution mutations are highest on the primary chromosomes and not the secondary replicons in *Burkholderia*. This appeared to have no relationship to the differences in nucleotide composition of these replicons, but rather due to some substitutions occurring at higher rates on particular replicons (Dillon et al. 2015). Purifying selection on the primary chromosome of *Burkholderia* must be substantially stronger to offset the effect of an elevated mutation rate (Dillon et al. 2015). In a more recent study by Dillon et al. (2018), base pair substitution rates were less variable on the smaller chromosomes of *Vibrio*. This provides yet another example where multi-repliconic substitution trends do not appear to be universal.

1.5 Thesis objectives

The state of knowledge about bacterial genome organization is extensive and covers a broad range of molecular topics, but lacking an in depth genomic analysis. There is evidence to support the organization of genomic content on bacterial genomes based on relative distance from the origin

of replication (Cooper and Helmstetter 1968; Chandler et al. 1975; Chandler and Pritchard 1975; Bremer and Churchward 1977; Schmid and Roth 1987; Sousa et al. 1997; Couturier and Rocha 2006; Bryant et al. 2014; Le and Laub 2014; Gerganova et al. 2015; Kopejtka et al. 2019; Lato and Golding 2020a). Prior research on spatial molecular trends when moving from the origin of replication to the terminus have determined that substitution rates (dN and dS), and their ratio (ω), increase with distance from the origin of replication (Cooper et al. 2010; Morrow and Cooper 2012). Nonetheless, these studies do not take into account genome reorganization such as rearrangements and inversions, which happen frequently in bacterial genomes and are an important source of genomic variation for bacteria (Ochman et al. 2000; Epstein et al. 2014). Nor do they consider the deep evolutionary history of these substitutions through methods such as ancestral reconstruction. Failing to account for these phenomenon could drastically alter the results. Additionally, gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012) and gene dosage (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Sauer et al. 2016) are increased near the origin, and genes become less conserved with increasing distance from the origin (Couturier and Rocha 2006). However, these studies do not take a genomic approach at looking at gene expression and often focus on one gene or a small subset of genes. The impact of inversions on gene expression in bacteria has previously been limited to creating inversions of a single gene and it's promoters, failing to consider "naturally occurring" inversions or all genes within the genome. The failure to incorporate genome rearrangements and lack of genome wide analysis on spatial molecular trends such as substitution rate and gene expression has lead to the following objectives:

1. Much is known about the influence genomic position has on substitutions in bacterial genomes. Most of these preceding studies failed to analyze secondary replicons of multi-repliconic genomes or take into account the effect of genomic rearrangements. I hypothesize that by using phylogenetic ancestral reconstruction and accounting for large scale genomic rearrangements, there will be a higher resolution of genomic substitution trends within a subset of bacterial genomes. I expect that the number of substitutions should increase when moving away from the origin of replication, based on these previous results.

2. Distance from the origin of replication has a profound impact on gene expression values. Previous studies have only looked at the impact genomic location has on a single gene or cluster of genes and promoters. I hypothesize that by looking at the expression patterns of all genes in a subset of bacterial genomes, we should see higher gene expression values found near the origin of replication and lower gene expression values found near the terminus of replication.

3. Multiple studies have shown that inversions can have various phenotypic effects in bacteria such as altering growth rate and gene expression. These studies primarily focus on creating novel inversions of one gene and it's promoters. I hypothesize that by using existing genomic gene expression and sequence data, "naturally occurring" inversions can be identified and

used to look at how gene expression differs between inverted and non-inverted segments of *E. coli* genomes.

1.5.1 Brief experimental objectives

This thesis aims to address the gaps in knowledge about spatial molecular trends in bacterial genomes, specifically number of substitutions, gene expression and inversions. This work will overall provide genomic wide in depth analysis for these three molecular traits that have previously not been studied. Firstly, this thesis will look at the impact distance from the origin of replication has on the number of substitutions in bacterial genomes by ancestrally reconstructing both position and substitutions in 24 bacterial genomes. Secondly, this work aims to assess difference in gene expression values between the origin and terminus of replication in four bacterial species including multi-repliconic bacterial taxa. Finally, this thesis will provide insight into the repercussions of "naturally occurring" inversions on gene expression within bacterial genomes. This thesis will seek to achieve the following three main objectives:

1. (a) Ancestrally reconstruct genomic substitutions and their genomic positions in 24 bacterial genomes while accounting for genomic reorganization, discussing the impact these substitutions have on genomic organization in bacteria.

   (b) Estimate the substitution rates and infer selective pressures acting upon the genomic protein coding substitutions mentioned above, and discuss the connection these selective pressures have on the organization of bacterial genomes.

2. Determine how gene expression is altered by the genomic location of the gene using existing genomic gene expression data from four bacterial species, examine the relationship this has with bacterial genome organization.

3. (a) Identify large scale and local "naturally occurring" inversions in four *E. coli* genomes.

   (b) Establish a link between gene expression and inverted or non-inverted segments of the *E. coli* genomes, discussing the possible implications inversions can have on gene expression and other functional bacterial traits.

   (c) Explore overlap between nucleoid associated associated proteins H-NS and Fis binding sites and identified inverted regions, and discuss the potential link to gene expression and genomic architecture.

Chapter 2

The Location of Substitutions and Bacterial Genome Arrangements

Daniella F. Lato and G. Brian Golding

As published in Genome Biology and Evolution, 2020
https://doi.org/10.1093/gbe/evaa260

2.1 Preface

Chapter 2 describes the identification and analysis of substitution, dN, dS, and ω patterns along various bacterial genomes. As described in Chapter 1, prior research on spatial molecular trends when moving from the origin of replication to the terminus have determined that substitution rates (Non-synonymous Substitution Rate (dN), Synonymous Substitution Rate (dS)), and their ratio (ω), increase with distance from the origin of replication. Nonetheless, these studies do not take into account genome reorganization such as rearrangements and inversions, which happen frequently in bacterial genomes, providing an important source of genomic variation for bacteria. Nor do previous studies consider the deep evolutionary history of these substitutions through methods such as ancestral reconstruction. In this work, we reconstruct ancestral genomic substitutions and their genomic positions in 24 bacterial genomes *Escherichia coli*, *Bacillus subtilis*, *Streptomyces*, and *Sinorhizobium meliloti*. This is done while accounting for genomic reorganization, discussing the impact these substitutions have on genetic organization in bacteria. We additionally estimated substitution rates (dN and dS) and infer selective pressures (ω) acting upon the genomic protein coding substitutions mentioned above, and discuss the connection these selective pressures have on the organization of bacterial genomes. This chapter is published in Journal of Molecular Evolution Genomics as: D. F. Lato and G. B. Golding (2020b). The Location of Substitutions and Bacterial Genome Arrangements. Genome Biol Evol. I made significant contributions to this study. I conceived the experiment jointly with G.B. Golding. I curated genomic datasets and evaluated the spatial genomic trends of substitutions, dN, dS, and ω in each species. I developed custom pipelines and scripts to complete this analysis including the modification of the `PAML` program (Yang 1997). I wrote the first version of this manuscript, which was edited and approved by G.B. Golding. G.B. Golding supervised the analyses and writing of the manuscript.

2.2 Abstract

Increasing evidence supports the notion that different regions of a genome have unique rates of molecular change. This variation is particularly evident in bacterial genomes where previous studies have reported gene expression and essentiality tend to decrease, while substitution rates usually increase with increasing distance from the origin of replication. Genomic reorganization such as rearrangements occur frequently in bacteria and allow for the introduction and restructuring of genetic content, creating gradients of molecular traits along genomes. Here, we explore the interplay of these phenomena by mapping substitutions to the genomes of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*, quantifying how many substitutions have occurred at each position in the genome. Preceding work indicates that substitution rate significantly increases with distance from the origin. Using a larger sample size and accounting for genome rearrangements through ancestral reconstruction, our analysis demonstrates that the correlation between the number of substitutions and distance from the origin of replication is significant but small and inconsistent in direction. Some replicons had a significantly decreasing trend (*E. coli* and the chromosome of *S. meliloti*), while others showed the opposite significant trend (*B. subtilis*, *Streptomyces*, pSymA and pSymB in *S. meliloti*). Non-synonymous Substitution Rate (dN), Synonymous Substitution Rate (dS) and ω were examined across all genes and there was no significant correlation between those values and distance from the origin. This study highlights the impact that genomic rearrangements and location have on molecular trends in some bacteria, illustrating the importance of considering spatial trends in molecular evolutionary analyses. Assuming that molecular trends are exclusively in one direction can be problematic.

[1] Department of Biology, McMaster University, Hamilton, ON, Canada

* Author for correspondence: G. Brian Golding, Department of Biology, Life Science Building, McMaster University, Hamilton, ON, Canada, L8S 4K1. Email: golding@mcmaster.ca.

Key Words: genome location, substitution, genomic structure, origin of replication, bacteria

Statement of Significance: Previous studies have demonstrated that genomic position in bacterial genomes impacts many molecular trends such as gene expression and substitution rate. However, these studies have failed to incorporate information about genomic reorganization, such as rearrangements, into their analyses and often used few taxa. Using ancestral reconstruction to account for genomic reorganization we have found that the number of substitutions significantly changes depending on bacterial genomic position. Utilizing information about genomic rearrangements, we demonstrate that although individual correlations between the number of substitutions and distance from the origin of replication are significant, the values

are small and inconsistent in direction. Consequently, varying substitution trends are detected when considering all bacterial species in this analysis.

2.3 Introduction

Bacterial genomes are subject to the introduction and reorganization of genetic information through processes such as Horizontal Gene Transfer (HGT), rearrangements, duplications, and inversions. These processes happen frequently and are important sources of genomic variation (Ochman et al. 2000; Epstein et al. 2014). Over a long term evolutionary experiment (25 years) it has been observed that there can be anywhere between 5 and 20 rearrangement events within a single lineage (identified from each population after 40,000 generations) (Raeside et al. 2014), and some of these spontaneous rearrangements (20% - 40%) persist in bacterial populations (Sun et al. 2012). DNA that is acquired through HGT or other genomic rearrangements can come from the same and/or different species of bacteria, allowing useful genes to be integrated into new genomes (Ochman et al. 2000). Genomic reorganization such as rearrangements, duplications, and inversions provide bacteria with the opportunity to fine tune existing gene expression, dosage, and replication. Bacteria can not escape genome reorganizations, and therefore incorporating past reorganizations is a crucial component of bacterial evolutionary analyses and can be done through multi-genome alignment programs such as `progressiveMauve` (Darling et al. 2010), which are rearrangement aware.

Changes in the genomic structure of a bacterial genome may provide new genomic landscapes capable of altering gene regulation. Here we will consider three main types of bacterial genomic structures: circular chromosomes, linear chromosomes, and multi-repliconic genomes. Secondary replicons of multi-repliconic bacteria are hypothesized to predominantly contain niche specific genes (Heidelberg et al. 2000; Egan et al. 2005). These replicons generally contain genes that have distinctive rates of evolution and selection acting upon them (Heidelberg et al. 2000). This allows the bacteria to thrive in rapidly changing environments, with varying molecular traits associated with each replicon (Heidelberg et al. 2000; Cooper et al. 2010; Morrow and Cooper 2012; Galardini et al. 2013; Jiao et al. 2018).

A previous multipartite genome investigation with four genomes of *Burkholderia* has shown that the primary chromosome is highly conserved and has higher gene expression compared to the secondary replicons which are less conserved (Morrow and Cooper 2012). A similar study using a minimum of four genomes from *Burkholderia*, *Vibrio*, *Xanthomonas*, and *Bordetella* also discovered that the primary chromosomes are conserved, with higher gene expression compared to the secondary replicons (Cooper et al. 2010). However, molecular differences between secondary replicons varies between bacterial species. In *S. meliloti*, there is evidence that pSymB is more transcriptionally integrated with the chromosome compared to pSymA and this could be a function of the difference in evolutionary time passed, with pSymB being older than pSymA, and the amount of gene flow between these secondary replicons (DiCenzo et al. 2018). Additionally,

primary chromosomes typically have lower substitution (Morrow and Cooper 2012) and evolutionary rates (Cooper et al. 2010) compared to the secondary replicons. Housekeeping genes usually reside on the primary chromosome, and the secondary replicons usually contain parts of the accessory genome, which could account for the substitution and evolutionary rate differences between primary and secondary replicons (Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012; Jiao et al. 2018). It has been suggested that the differences in gene content between replicons of multi-repliconic bacteria may be due to delays in replication (Flynn et al. 2010; Morrow and Cooper 2012). To maintain synchronization, due to the offset of different sequence lengths between primary and secondary replicons, the secondary replicons begin replication after the primary chromosome (Flynn et al. 2010; Morrow and Cooper 2012).

Prior research on molecular trends when moving from the origin of replication to the terminus have determined that gene expression is increased near the origin (Couturier and Rocha 2006; Kosmidis et al. 2020; Lato and Golding 2020a), and genes become less conserved with increasing distance from the origin (Couturier and Rocha 2006; Rocha and Danchin 2004). Analyses with a few bacterial species have replicated these results and found that gene expression decreases with increasing distance from the origin (*Burkholderia*; Morrow and Cooper 2012) and substitution rates (non-synonymous (dN), synonymous (dS), and dN/dS) increase with distance from the origin of replication (*Burkholderia*, *Vibrio*, *Bordetella*, *Xanthomonas*; Cooper et al. 2010: *Burkholderia*; Morrow and Cooper 2012). It is speculated that genes near the terminus are more prone to recombination, while genes near the origin have a higher prevalence of recombination repair (Sharp et al. 1989; Flynn et al. 2010). Genes near the terminus therefore often have more variation and are less conserved compared to those near the origin of replication (Sharp et al. 1989; Flynn et al. 2010). Additionally, genes found within the core genome are typically located near the origin of replication, while genes associated with the accessory genome are found near the terminus (Couturier and Rocha 2006; Flynn et al. 2010). The placement of these two gene categories may explain why near the origin, gene expression and essentiality are high (Couturier and Rocha 2006; Kosmidis et al. 2020; Lato and Golding 2020a) and substitution rate is low (Flynn et al. 2010).

It is well known that substitutions and mutations have a non-random distribution around the genome which varies by gene and organism (Sharp et al. 1989; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012; Dillon et al. 2015). But, not all studies have a clear positive correlation with distance from the origin of replication and mutation rate. Some studies found no correlation between distance from the origin of replication and the frequencies of mutations, but they did find mutation rate to vary with position along the *E. coli* chromosome (Martina et al. 2012; Juurik et al. 2012). Other investigations found no positive correlation with mutation rates and distance from the origin of replication and instead found that intermediate positions had a higher non-synonymous mutation rate than positions farther from the origin in *E. coli* (Ochman 2003) and *Salmonella enterica* (Hudson et al. 2002; Ochman 2003). With

respect to multi-repliconic bacteria, some studies have found a lack of positive correlation between mutation rate and distance from the origin of replication. Dillon et al. (2015) found that base-substitution mutation rates are highest on the primary chromosomes and not the secondary replicons in *Burkholderia*, opposing previous observed evolutionary rates in work by Cooper et al. (2010). This appeared to have no relationship to the differences in nucleotide composition of these replicons, but rather due to some types of substitutions occurring at higher rates on particular replicons (Dillon et al. 2015). In a more recent study, Dillon et al. (2018), found that base-substitution mutation rates vary in a wave like pattern in *Burkholderia* and *Vibrio*, where concurrently replicated segments have similar rates. This wave like pattern in mutations was also seen in *E. coli* (Long et al. 2016) and in mutation rates in *Pseudomonas aeruginosa* (Dettman et al. 2016). A similar wave like pattern in base pair substitutions has been observed in *E. coli* (Foster et al. 2013; Niccum et al. 2019). The wave like patterns are thought to be related to cell cycle functions and not sequence composition (Dillon et al. 2018). Interestingly there are noteworthy differences in the location of the core and accessory genomes in some bacterial species. In the *Rhodobacterales* family, some species have core genes concentrated near the terminus, not the origin of replication (Kopejtka et al. 2019). Other species of this family have a mosaic pattern of core genes dispersed throughout the genome (Kopejtka et al. 2019). It is speculated that other factors such as HGT, phage insertion, and replication may be responsible for the conflicting placement of core genes in various *Rhodobacterales* species (Kopejtka et al. 2019). All of these exceptions to the previously established molecular trends raise questions about how universal these trends are.

There are a number of additional factors that are dependent on distance from the origin such as transposon insertion events (Gerdes et al. 2003), gene order (Mackiewicz et al. 2001), number of replication forks (Couturier and Rocha 2006), and nucleotide composition (Mackiewicz et al. 1999; Karlin 2001). These phenomena are also important to consider when analyzing molecular trends with respect to distance from the origin of replication.

The majority of these studies used an average of three genomes per bacteria analyzed (Couturier and Rocha 2006; Flynn et al. 2010; Cooper et al. 2010; Morrow and Cooper 2012) and failed to analyze secondary replicons of multipartite genomes (Couturier and Rocha 2006; Flynn et al. 2010). In this study we examine the spatial substitution trends in *E. coli* (six genomes), *B. subtilis* (seven genomes), *Streptomyces* (five genomes), and *S. meliloti* (six genomes). These bacteria contain genomic structures that range from a single circular chromosomes (*E. coli* and *B. subtilis*), a linear chromosome (*Streptomyces*), and a multi-repliconic genome (*S. meliloti*). This selection of bacterial taxa provides a sample that covers broad lifestyles as well as representing a number of divergent phylogentic lineages, providing a diverse sample to determine if the number of substitutions increases with increasing distance from the origin of replication. This study aims to determine what spatial substitution trends appear in these bacterial genomes when including the effects of genomic reorganization. We use the ancestral states of substitutions and the ancestral genomic positions of the substitutions, leading to a more accurate estimation of

multiple substitutions and genomic position. Supplemental analyses on selection patterns was also performed to elucidate the potential influences on the substitution trends. We show here that the correlation between the number of substitutions and distance from the origin of replication is significantly inconsistent and small for the genomes we studied. For the majority of the replicons investigated, the number of substitutions increased when moving away from the origin of replication towards the terminus. But exceptions were the chromosomes of *E. coli* and *S. meliloti*, where the number of substitutions decreased with increasing distance from the origin. We did not find consistent significant correlations between dN, dS, and ω values and distance from the origin of replication. Possible causes and consequences of these patterns are discussed.

2.4 Materials and Methods

A complete list of version numbers and build dates for all the programs used in this analysis can be found in Supplementary Table S1.1 available on `GitHub` (`www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`).

2.4.1 Sequence Data

Whole genomes of different strains of *E. coli*, *B. subtilis*, and *S. meliloti*, as well as various species of *Streptomyces* were downloaded from NCBI. Access date and accession numbers are given in Supplementary Table S1.2. These bacteria inhabit a variety of different habitats and have contrasting genomic structures, providing a well rounded sample for this analysis. Although *E. coli*, *B. subtilis*, and *Streptomyces* contain small plasmids, they are not considered multi-repliconic bacteria and therefore their plasmids were not included in this analysis. *S. meliloti* is a multi-repliconic bacteria and its two large secondary replicons were included in the analyses (pSymA and pSymB). The replicons of *S. meliloti* are known to differ in genetic content, and therefore, all analyses were performed on each individual replicon of *S. meliloti*. The genomes used for each species consisted of as many reference genomes as were practically possible (Supplementary Table: S1.2).

2.4.2 Sequence Alignment

Alignments of each bacterial replicon were performed using `progressiveMauve` (default parameters) (Darling et al. 2010) to group the sequences of the replicons into Locally Colinear Blocks (LCBs). This method allows for rearrangements, duplications and inversions to be taken into account. A LCB is frequently found at different genomic positions in each of the taxa analyzed. `progressiveMauve` defines these segments of sequence as minimally being similar between at least two of the taxa, but not necessarily between all of them. To obtain accurate information for subsequent analyses, only the subset of LCBs that were present in all taxa were considered. Each locally co-linear block was then re-aligned with `MAFFT` (–auto) (Katoh et al. 2002) to obtain a more accurate local alignment. Although `progressiveMauve` is good at identifying large

scale rearrangements and inversions, it sometimes determined LCBs that were very small and contained questionably homologous or excessively gapped sequences (see Supplementary file for more information and examples). As a result, we used `trimAl` (Capella-Gutiérrez et al. 2009) to remove poorly aligned regions, which were defined as having poor homology and/or excessive gaps. We used the `-strictplus` setting in `trimAl` to automatically determine regions of unacceptable alignment.

A custom `Python` script was created to ensure that within each alignment LCB the correct coding frame was present. Codon position information was obtained for each base pair in the LCBs from the GenBank file for each taxa. Each column of the alignment was only kept if all taxa had the same codon position (1, 2, or 3). Alignment columns where the codon positions were not the same were removed from the analysis.

We found that using these alignment trimming criteria effectively removed portions of the alignment that had poor homology or were gaped. We imposed an additional minimum ungapped alignment length of 100Base Pair (bp) to each of the gene segments. We chose this number so that we could keep the maximum amount of information, while avoiding comparing potentially inaccurate and extremely short portions of a gene (less than 100bp). These trimmed alignments of genes and gene segments are used for the remainder of the analysis.

There is a delicate balance between capturing large amounts of recombination, while still ensuring a comparison of homologous sequences. The more distantly related taxa are, the less similar the genetic sequences are, which in the case of `progressiveMauve`, results in a large number of short LCBs. A high number of LCBs results in the potential comparison of non-homologous sequences, which would create incorrect results in any phylogenetic or evolutionary analysis. As a result, we had to limit the number sequences used in our analyses (see Supplementary Material for additional details, `www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`).

In addition, the number of sequences chosen for all bacteria was constrained by the computational time required to perform a `progressiveMauve` alignment. This computing time increases exponentially with additional genomes. For further information please see the Supplementary Material on `GitHub` at `www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangementss`.

Protein Coding Substitutions

To ensure that only homologous sequences were being compared, we are only considering the substitutions that reside in protein coding regions of the genome. Any site where a gap or ambiguous nucleotide was present, was removed from the analysis, and the remaining portions of the gene were separated and considered two distinct "genes". The remainder of the analyses was done on each of these gene segments separately.

2.4.3 Phylogenetic Trees

Rearrangements, duplications, and inversions happen frequently and must be considered when analyzing spatial genomic trends. Phylogenetic trees were created to trace the evolutionary history of large scale and local DNA rearrangements. These trees were used to determine the number of substitutions and record the genomic location of substitutions for each respective replicon. Whole genome alignments both including and excluding the outgroups were performed using `progressiveMauve` and split up into LCBs which were re-aligned with `MAFFT` (see the Sequence Alignment Methods section). Each of the LCBs specified by `progressiveMauve` were combined to create a single "super sequence". `RAxML` was used to estimate phylogenetic trees both including (`raxmlHPC-PTHREADS-SSE3 -T 20 -f a -x 12345 -o -N 100 -p 12345 -m GTRGAMMA`) and excluding (`raxmlHPC -f a -x 12345 -p 12345 -# 1000 -m GTRGAMMA`) the outgroup. The tree topology from the phylogenetic tree including the outgroup was used to optimize the branch lengths for the phylogenetic tree excluding the outgroup (`raxmlHPC -f T -t -p 12345 -m GTRGAMMA`). Bootstrap values for this tree was calculated using 1000 replicates (`raxmlHPC -f b -t -z -m GTRGAMMA`). Phylogenetic trees with bootstrap support values can be found in the Supplementary Material.

A SH-test (Shimodaira and Hasegawa 1999; Goldman et al. 2000) was performed to determine if there was a significant difference between the "super sequence" and the tree topology of each LCB individually. Any LCBs that had a topology that was significantly different (at the 5% significance level) from the "super sequence" topology, was removed from the remainder of the analyses. The SH-test was performed using `RAxML` (`raxmlHPC -f H -t -z -s -m GTRGAMMA`) (Stamatakis 2014).

2.4.4 Origin and Bidirectional Replication

For each bacteria the origin of replication was denoted as the beginning of the *oriC* region for the chromosomal replicons, and the beginning of the *repC* (Pinto et al. 2011) region for the secondary replicons of *S. meliloti* (Supplementary Table S1.7). This origin of replication position was calibrated to be the beginning of the genome, position 1, and remaining positions in the genome were all scaled around this origin of replication taking into account the bidirectional nature of bacterial replication (Figure 2.1).

The terminus of replication was determined using the Database of Bacterial Replication Terminus (Kono et al. 2011), which uses the prediction of *dif* sequences (normally found at the terminus), as a proxy for the location of the terminus (Clerget 1991; Blakely et al. 1993). For pSymA and pSymB of *S. meliloti* the terminus is not listed in the database, thus the terminus location was assigned to the midpoint between the origin of replication and the end of the replicon. Replication in the linear chromosome of *Streptomyces* begins at the origin of replication, located to the right of the middle of the replicon (Heidelberg et al. 2000), and terminates at each end of the chromosome arms (Heidelberg et al. 2000)(Supplementary Table S1.7).

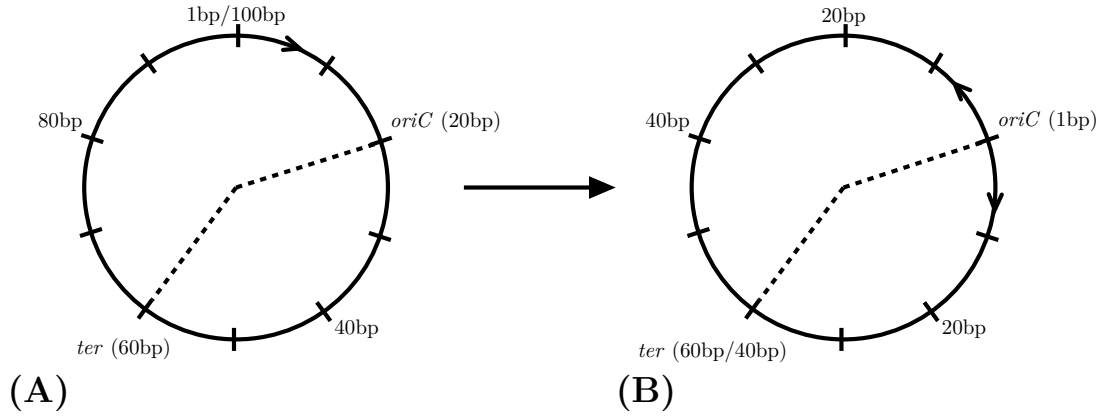


Figure 2.1: Schematic of the transformation used to scale the positions in the genome to the origin of replication and account for bidirectional replication. Circle (A) represents the original replicon genome without any transformation. Circle (B) represents the same replicon genome after the transformation. The origin of replication is denoted by "*oriC*" and the terminus of replication is denoted by "*ter*". The dashed line represents the two halves of the replicon. The replicon genome in this example is 100 base pairs in length. Every 10 base pairs is denoted by a tick on the genome. The origin in (A) is at position 20 in the genome and is transformed in (B) to become position 1. The terminus is at position 60 in (A) and position 60/40 in (B). The terminus has two positions in (B) depending on which replicon half is being accounted for. If the replication half to the right of the origin is considered, the terminus will be at position 40. If the replication half to the left of the origin is considered, the terminus will be at position 60. Position 40 in (A) becomes position 20 in (B). Position 80 in (A) becomes position 40 in (B), due to the bidirectional nature of bacterial replication. Figure from: D.F. Lato and G.B. Golding, Spatial Patterns of Gene Expression in Bacterial Genomes, Journal of Molecular Evolution, published June 2020, Springer Nature.

We have chosen a single base to represent the origin and terminus of replication. In reality, the origin of replication is often multiple base pairs long, and there has been no evidence for site-specific termination of replication, but rather a small genomic region where replication concludes based on various other factors (Duggin and Bell 2009). To determine the effect of the exact location of the origin and terminus, permutation tests shuffling the *oriC* position by 10,000bp increments in each direction from the original origin (Supplementary Table: S1.7) to a maximum of 100,000bp in each direction were performed. These results showed that moving the origin of replication does not affect the results of the analysis (Supplementary Table: S1.8). Based on this supplementary test, choosing a single base to represent the origin and terminus of replication has minimal impact on the analysis.

2.4.5 Ancestral Reconstruction

To track genome reorganization, nucleotide substitutions and genomic positions were reconstructed in extinct ancestors. We used the `PAML` (Yang 1997) package of programs, with slight

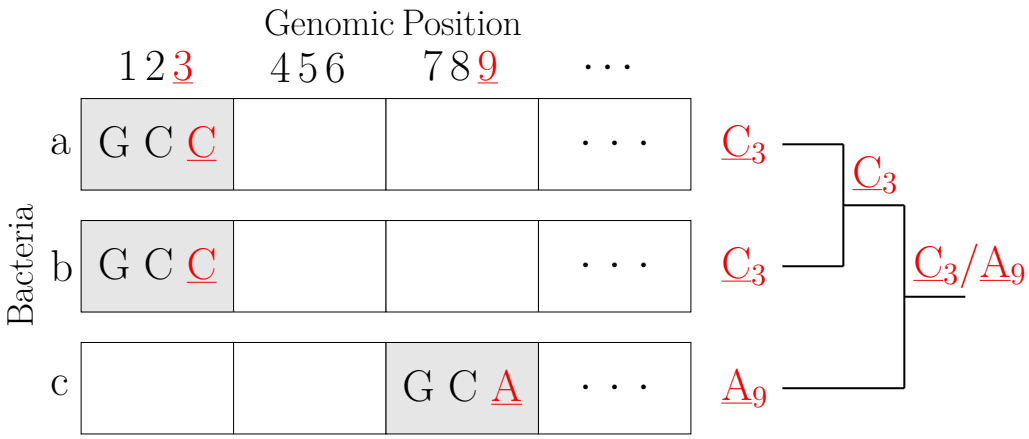


FIGURE 2.2: Schematic of the ancestral reconstruction of both the nucleotide and genomic position. Each horizontal row of rectangles represents three hypothetical bacterial genomes (a, b, c). The genomic position is indicated at the top of the diagram. The phylogenetic tree showing the relationship between all three bacteria is pictured on the right of the diagram. The light grey rectangle denotes the homologous genomic region. In bacteria (a) and (b), this segment is located at genomic positions 1-3. In bacteria (c), this segment is located at genomic positions 7-9. Within this genomic region of interest there is a substitution where the nucleotides changed from C → A, this is highlighted in red and underlined. This would mean that in bacteria (c) there was a substitution from C → A which is also associated with a genomic position of 9. This substitution is at position 3 in bacteria (a) and (b), and in position 9 in bacteria (c). This is depicted by the values ($\underline{C}_3$) and ($\underline{A}_9$). The ancestral reconstruction process in this analysis can be seen at the inner nodes of the phylogenetic tree by the values ($\underline{C}_3$). The most parsimonious reconstruction of the sequence and associated genomic position is having the value ($\underline{C}_3$) present at the ancestor of bacteria (a) and (b). The ancestral node of all three bacteria would have a reconstruction of the sequence and associated genomic position of ($\underline{C}_3$ / $\underline{A}_9$). In this situation where there is a "tie" for two most parsimonious options, the option with the highest likelihood estimate would be chosen using maximum-likelihood methods (see Yang 1997 for more details).

modification, to reconstruct genome location and substitutions in hypothetical ancestors (Figure 2.2).

Nucleotide Substitutions

The `baseml` program (model=0, Mgene=0, clock=1, fix_kappa=0, kappa=5, fix_alpha=1, alpha=0, Malpha=0, ncatG=5, nparK=0, nhomo=0, getSE=0, RateAncestor=2) in the `PAML` package (Yang 1997) was used to determine single nucleotide substitutions within each of the alignments. This program determined the ancestral state of each nucleotide in the alignment at each node in the phylogenetic tree (Figure 2.2). Multiple substitutions at one site were allowed and accounted for as separate substitutions. Any nucleotides, or columns, in the alignment that

had at least one gap present were not used in the analyses because the `baseml` program inaccurately classifies substitutions when a gap is involved. These gapped positions were categorized as missing data.

Genomic Position

Genomic reorganization was accounted for using the genome locations specified by `progressiveMauve` to determine the ancestral genome positions of each taxa (Figure 2.2). These locations were inferred for each nucleotide in the alignment. The `codeml` program (CodonFreq=F3X4, clock=0, aaDist=0, aaRatefile=dat/jones.dat, model=0, NSsites=0, Mgene=0, fix_kappa=0, kappa=2, fix_omega=0, omega=0.4, fix_alpha=1, alpha=0, Malpha=0, ncatG=8, getSE=0, RateAncestor=1) (Yang 1997) from the `PAML` package was modified to reconstruct the ancestral genome positions at each node within the phylogenetic tree (Supplementary Trees: S1.4 - S1.9) of each respective replicon for each position in the alignment (Figure 2.2).

A custom `Python` script (see `GitHub` `www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`) was used to associate each of the protein coding regions with their genomic positions and determine how many ancestral and extant substitutions were found in each region. Each branch in the tree possesses information on how each nucleotide in the alignment has moved throughout the genome to the current position in each of the taxa (Figure 2.2). Therefore, each segment of sequence has the opportunity to be present in one position in the genome of one taxa, and a completely different position in another taxa (Figure 2.2).

For this portion of the analysis each genomic position was considered unique and distinct, including positions that were separated by one base pair. We performed a supplementary analysis to determine if clustering genomic positions based on how many base pairs separate substitutions, would significantly alter the overall spatial results (See Supplemental Material for more details). We determined that considering each genomic position to be unique and distinct or clustering the positions did not alter the results.

2.4.6 Logistic Regression

The binary nature of the data is ideal for a logistic regression to determine the statistical significance of substitution and position trends at protein coding regions of the genome in each bacterial replicon (Table 2.2). Any subset of points outside the inter quartile range were considered outliers and ignored.

A visualization of substitutions in relation to distance from the origin of replication can be found in Figures 2.3 and 2.4. The total number of substitutions in each 10Kilobase Pair (Kbp) region of the replicon was divided by the total number of protein coding sites within that 10Kbp region, to give the substitutions per 10Kbp (y-axis).

2.4.7 Selection

Within the protein coding regions of the genome, we wanted to observe how selection may be acting on each of the genes in the various bacterial replicons. Calculating the synonymous (dS) and non-synonymous (dN) substitution rates and the ratio of these two (ω) for each gene allows for an in depth analysis of the selective pressures throughout the genome while accounting for genomic reorganization between the bacterial taxa. We can then relate this information to the location of the genes in the genome and determine trends between selection and distance from the origin. It has been found previously that genes closest to the origin of replication are conserved (Couturier and Rocha 2006) and tend to be a part of the core genome (Couturier and Rocha 2006; Flynn et al. 2010). We therefore expect genes closer to the origin to have fewer substitutions and therefore lower values for dS and dN.

The datasets used for this portion of the analysis is the same as the one used in the substitutions analysis, with the exception that we ensured all genes and gene segments of the alignment start and end with complete codons for the selection analysis (this was done through a custom `Python` script). Gaps or ambiguous nucleotides were also removed from these genes (`Python`) and are subsequently missing in the graphical representation of the distribution (Figures 2.5 and 2.6).

Calculating *dN*, *dS* and ω

The `codeml` program (CodonFreq=2, clock=0, model=0, NSsites=0, icode=0, fix_omega=0, omega=0.4) in the `PAML` package (Yang 1997) was used to calculate the synonymous (dS) and non-synonymous (dN) substitution rates and to estimate a value for ω. dN, dS, and ω were calculated on each gene/gene segment separately. The varying nucleotide models have minimal impact on the dN and dS calculations because the overall number of synonymous and non-synonymous substitutions per site were small. There were some segments of the alignment that had poor homology (see Methods: Sequence Alignment for more information). As a result, some genes were split into multiple parts, removing those segments of poor alignment. Calculations and analyses were done separately for each of these gene "segments" for the remainder of the study.

Outliers for the selection data were determined using only the ω values. Any subset of ω points outside the inter quartile range were considered outliers and ignored. The associated dN and dS values for the same gene segment of each ω outlier were also considered outlier values. These points were subsequently removed from the analysis. We then used the dN, dS, and ω values of each gene or gene segment to calculate an arithmetic average of dN, dS, and ω for each replicon weighted by the length of each gene or gene segment. To prevent the use of undefined ω values, any genes where both dN and dS, or dS were equal to zero were removed from the weighted ω calculation. A summary of the average dN and dS results are found in Table 2.3.

Linear regressions were performed to determine if there was any correlation between dN, dS, and ω respectively and distance from the origin of replication while accounting for bidirectional replication. All linear regression results are summarized in Table 2.4.

2.5 Results

2.5.1 Average Number of Substitutions

Table 2.1 summarizes the average number of substitutions per base pair for each bacterial replicon. The *S. meliloti* chromosomes and species of *Streptomyces* chosen for this study are very similar, and therefore have highly conserved sequences. This strong sequence conservation is not seen for the other replicons (*E. coli*, *B. subtilis*, and the secondary replicons of *S. meliloti*). This low divergence between genomes is likely the cause for lower average number of substitutions per base pair in *Streptomyces* and the chromosome of *S. meliloti*. The smaller replicons of *S. meliloti* - pSymA and pSymB - have faster substitution rates compared to the larger chromosomal replicon of the same bacteria. This is likely due to the relative decreased divergence between strains used in the *S. meliloti* chromosome analysis. pSymB has a slightly faster substitution rate compared to pSymA. These results are consistent with the general knowledge of the gene content between the smaller replicons of *S. meliloti* and the chromosome. The smaller replicons are expected to evolve more quickly. It is curious that pSymB has a slightly higher average substitution rate compared to pSymA because pSymA has been shown to be more variable in gene content and function compared to pSymB (Galardini et al. 2013).

Bacteria and Replicon	Average Number of Substitutions per bp
E. coli Chromosome	6.48×10^{-3}
B. subtilis Chromosome	7.56×10^{-3}
Streptomyces Chromosome	4.23×10^{-4}
S. meliloti Chromosome	2.43×10^{-4}
S. meliloti pSymA	2.03×10^{-3}
S. meliloti pSymB	2.35×10^{-3}

Table 2.1: Average number of protein coding substitutions calculated per base across all bacterial replicons. Outliers and missing data are not included in the calculation.

2.5.2 Logistic Regression

The logistic regression and supporting statistical information for the substitution trends are found in Table 2.2. The number of substitutions decreased when moving away from the origin of replication for the protein coding regions of *E. coli* and the chromosome of *S. meliloti*. This

implies that the area near the terminus of replication in these replicon sections had fewer substitutions than the area near the origin of replication. pSymA and pSymB of *S. meliloti*, *B. subtilis*, and *Streptomyces* showed the opposite trend from the other bacterial replicons, with a decreased number of substitutions present near the origin of replication compared to the terminus. All of the correlation estimates between the number of substitutions and distance from the origin of replication are small and vary in their sign. From these inconsistent results, we conclude that there is no consistent, significant correlation between the number of substitutions and distance from the origin of replication.

Additional tests grouping the number of substitutions in varying windows of the genomes (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) were done to supplement the logistic regression results. The total number of substitutions per window size (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) was totalled and a linear regression was performed on those totals and distance from the origin of replication (Supplementary Tables S1.14 and S1.15). These results are inconsistent in sign when significant, mirroring the results from the logistic regression (Table 2.2). Based on these inconsistent supplemental results, we remain confident in saying that there is no consistent, significant correlation between the number of substitutions and distance from the origin of replication.

A non-linear analysis of the variation in the number of substitutions per 10Kbp with distance from the origin of replication was performed (Supplementary Figures S1.13 - S1.18). The results from this analysis complement the logistic regression results: the total number of substitutions varies with distance from the origin of replication, but the pattern and direction of this trend is inconsistent between bacterial replicons.

Additional analyses were done to ensure that the individual taxa chosen in this analysis were not influencing the overall conclusion about the distribution of substitutions along bacterial genomes. We systematically removed each taxa from the substitutions analysis (see Supplementary Material) to determine if any particular taxa were influencing the results. These results are summarized in Supplementary Table S1.18. From this supplemental analysis, we have come to the same conclusion that the number of substitutions significantly varies with distance from the origin of replication, but the direction of this trend is inconsistent in sign. In Supplementary Table S1.18, when most of the taxa in each species is removed, the correlation between the number of substitutions and distance from the origin of replication is significant and follows the same sign (positive or negative) within a replicon. However, occasionally the sign of this trend flips for particular strains/species that are removed. We determined this change was due to a new "outgroup" specified in the tree (via the removal of the previous "outgroup" in *Streptomyces* and pSymA of *S. meliloti*), or it is likely that the taxon removed was the ancestral genomic position for the substitutions and when it is removed, the ancestral genomic position changes (*B. subtilis* and pSymB of *S. meliloti*). A complete discussion of this can be found in the Supplementary Material. Future work exploring the ancestral states of nucleotides and genomic position using

different species/strains would be able to test for this.

Bacteria and Replicon	Protein Coding Sequences Coefficient Estimate
E. coli Chromosome	-2.66×10^{-8}***
B. subtilis Chromosome	2.76×10^{-8}***
Streptomyces Chromosome	6.97×10^{-8}***
S. meliloti Chromosome	-6.57×10^{-7}***
S. meliloti pSymA	2.74×10^{-7}***
S. meliloti pSymB	1.10×10^{-7}***

Table 2.2: Logistic regression analysis of the number of substitutions along all protein coding positions of the genome of the respective bacteria replicons. Grey coloured boxes indicate a negative logistic regression coefficient estimate. All results are statistically significant. Logistic regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. All results are marked with significance code as followed: $p < 0.001$ = '***'.

Areas of the bacterial genomes in this analysis with extremely high number of substitutions per 10Kbp region are regions that encode mostly small (65-150 amino acids long) hypothetical proteins (see Supplementary Table S1.13). These regions could have higher numbers of substitutions due to the small length of these genes and unclear characterization of the associated encoded proteins.

The density of ancestral and extant substitutions in protein coding regions across each bacterial replicon can be seen in Figures 2.3 and 2.4. These Figures supplement the logistic regression analysis and provide information on the frequency of substitutions in relation to the distance from the origin of replication while also taking into account the bidirectional replication (See Methods: Origin and Bidirectional Replication). Areas of these graphs that look sparse or appear to be "missing" data from some genomic regions have had data excluded in these regions because they did not meet the alignment quality and trimming requirements specified in this analysis (See Methods: Sequence Alignment).

2.5.3 Selection

The distribution of dN, dS, and ω values across each bacterial replicon can be seen in Figures 2.5 and 2.6. These Figures provide information on the values of dN, dS, and ω in relation to the distance from the origin of replication while taking into account bidirectional replication (See Methods: Origin and Bidirectional Replication). Areas of these graphs that look sparse or appear to be "missing" data from some genomic regions have had data excluded in these regions because they did not meet the alignment quality and trimming requirements specified in this analysis (See Methods: Sequence Alignment). High dS values in Figures 2.5 and 2.6 are reflective of divergent portions of a gene alignment. For a complete discussion of these values please see the Supplementary Material. dN and ω values of zero are produced by low numbers of substitutions, from in an overwhelming number of identical LCB (for a complete account of zero values please see the Supplementary Material).

	Genome Average		
Bacteria and Replicon	dS	dN	ω
E. coli Chromosome	0.2352	0.0101	0.0445
B. subtilis Chromosome	0.4134	0.0240	0.0712
Streptomyces Chromosome	0.0468	0.0011	0.0323
S. meliloti Chromosome	0.0122	0.0002	0.0042
S. meliloti pSymA	0.0839	0.0099	0.1760
S. meliloti pSymB	0.0956	0.0085	0.1148

Table 2.3: Weighted averages for dN, dS, and ω values calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean was calculated for the per gene averages for each bacterial replicon.

The genome average values of dN, dS, and ω for each replicon are found in Table 2.3. All bacterial replicons had average per genome dS values that were higher than the respective dN values. This is as expected, since most genes should be under purifying selection.

Linear regressions were performed to determine if there is any correlation between dN, dS, and ω respectively and distance from the origin of replication while accounting of bidirectional replication. All linear regression results are summarized in Table 2.4. All values for dN, dS, and ω, aside from any considered outliers (see Methods) were used in the regression analysis. We were unable to find significant linear regression coefficients for the majority of the bacterial replicons used in this analysis. The sporadic significant and non-significant positive and negative coefficient estimates do not provide a clear picture of how substitution rates and ω change with distance from the origin of replication, and we therefore can not conclude that there is one overarching spatial trend for dN, dS, or ω values.

Additional tests using the average dN, dS, or ω values in varying windows of the genomes

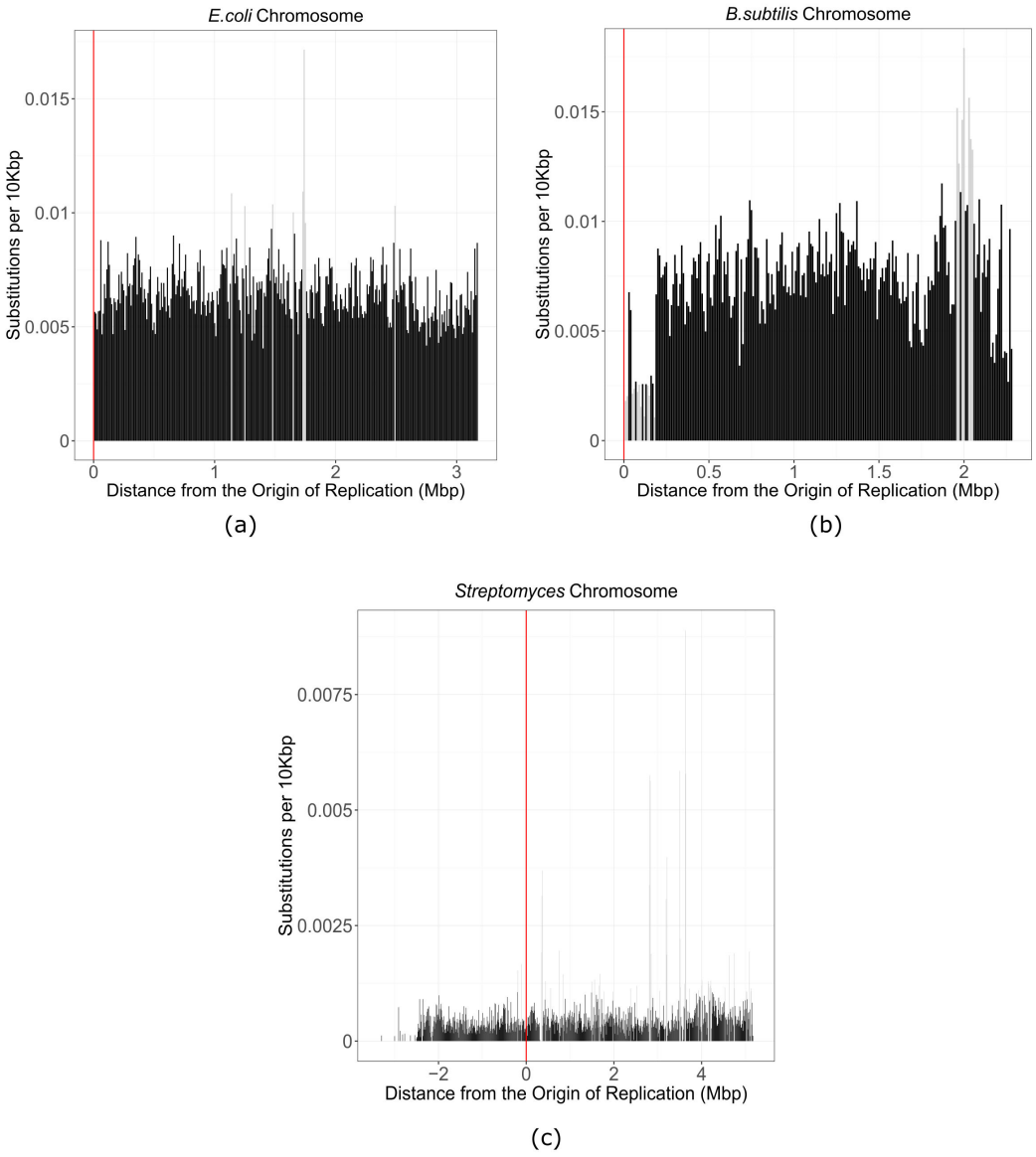


Figure 2.3: The bar graphs show the number of substitutions along the genomes of *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c). For *E. coli* and *B. subtilis*, the distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. This distance includes the distance from the origin in both replichores. For *Streptomyces* the origin of replication is denoted by position zero. The genome located on the shorter chromosome arm (to the left of the origin) has been given negative values, while the genome on the longer chromosome arm (to the right of the origin) has been given positive values. The origin of replication in the *Streptomyces* graph (c), has been highlighted at position zero by a red vertical line. The y-axis of the graphs indicate the number of substitutions per 10,000bp found at each position of the *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c) genomes. Each bar represents a section of the genome that spans 10Kbp. The total number of substitutions in each 10Kbp region of the replicon was divided by the total number of protein coding sites within that 10Kbp region, to give the substitutions per 10Kbp (y-axis). Outliers are represented in light grey bars.

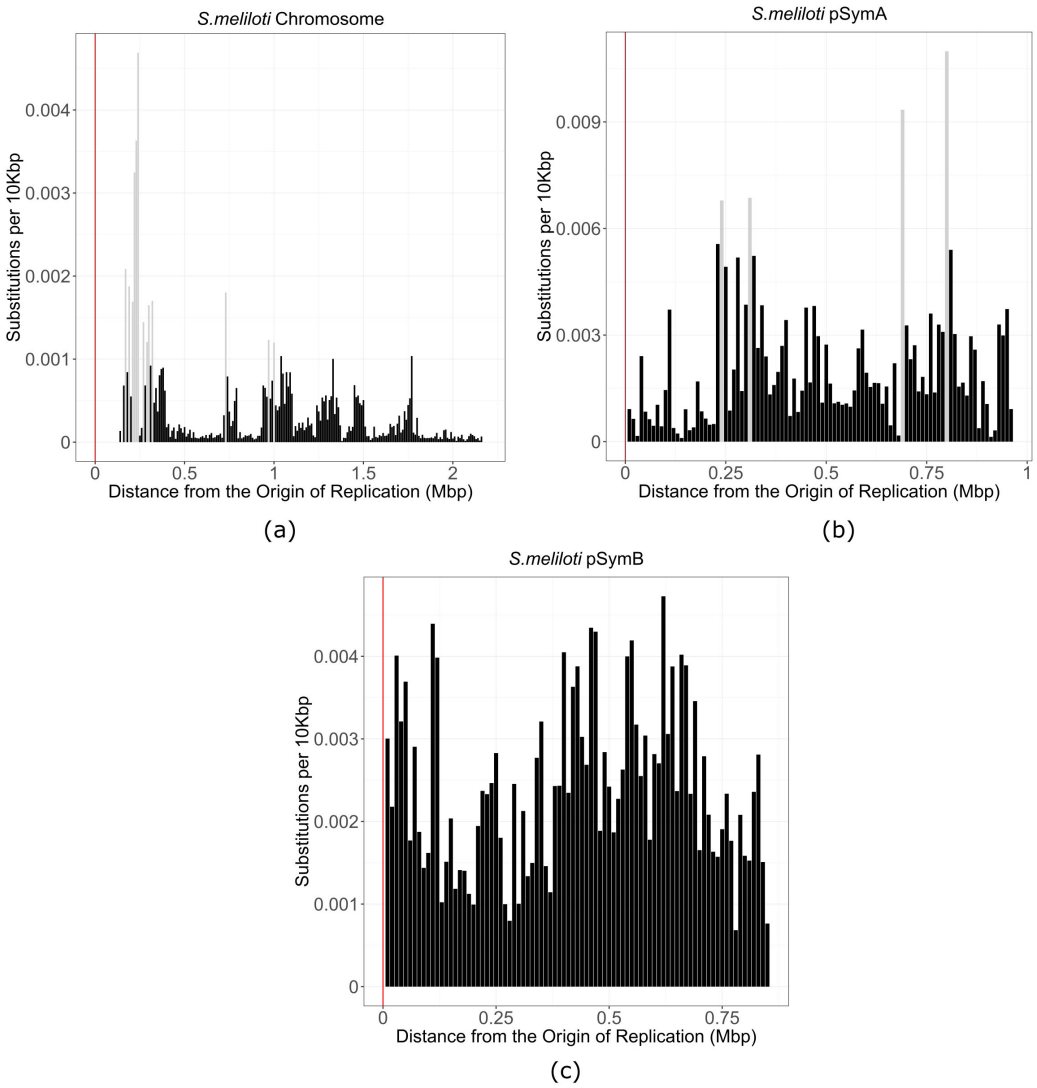


Figure 2.4: The bar graphs show the number of substitutions along the replicons of *S. meliloti*: chromosome (a), pSymA (b), and pSymB (c). Distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. This distance includes the distance from the origin in both replichores. The y-axis of the graph indicates the number of substitutions per 10,000bp of the replicons of *S. meliloti*: chromosome (a), pSymA (b), and pSymB (c). Each bar represents a section of the genome that spans 10Kbp. The total number of substitutions in each 10Kbp region of the replicon was divided by the total number of protein coding sites within that 10Kbp region, to give the substitutions per 10Kbp (y-axis). Outliers are represented by light grey bars.

(10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) were done to supplement the linear regression results done on all data points. The average dN, dS, or ω values per window size (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) was calculated and a linear regression was performed on those average values and distance from the origin of replication (Supplementary Table S1.21). These results are mostly not significant and ones that are significant are inconsistent in sign, mirroring the results from the linear regression on all data points (Table 2.4). Based on these inconsistent supplemental results, we are confident that there is no significant correlation between the value of dN, dS, or ω and distance from the origin of replication.

Bacteria and Replicon	*dN*	*dS*	ω
E. coli Chromosome	NS	NS	4.33×10^{-9}*** (0.007)
B. subtilis Chromosome	-6.03×10^{-9}*** (0.004)	NS	-6.80×10^{-9}*** (0.004)
Streptomyces Chromosome	1.40×10^{-10}* (0.002)	NS	NS
S. meliloti Chromosome	-1.67×10^{-10}* (0.003)	-8.67×10^{-9}*** (0.007)	-1.20×10^{-9}* (0.003)
S. meliloti pSymA	NS	NS	NS
S. meliloti pSymB	NS	NS	NS

Table 2.4: Linear regression to determine the correlations between dN, dS, and ω values and distance from the origin of replication. A regression was performed for each bacterial replicon with outliers removed. All results are marked with significance code as followed: p: < 0.001 = '***', > 0.05 = 'NS'. The R^2 values for each estimate are in brackets.

2.6 Discussion

To date there has been a large body of work looking at how molecular trends such as gene expression (Couturier and Rocha 2006; Cooper et al. 2010; Morrow and Cooper 2012; Kosmidis et al. 2020; Lato and Golding 2020a), substitution rates (Sharp et al. 1989; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012), and mutation rates (Hudson et al. 2002; Ochman 2003; Martina et al. 2012; Juurik et al. 2012; Dettman et al. 2016; Dillon et al. 2018) vary with genomic position. The general consensus is that substitution rate is highest near the terminus of replication and relatively low near the origin (Sharp et al. 1989; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012). Most of these studies used an average of 3 genomes per bacteria analyzed (Couturier and Rocha 2006; Flynn et al. 2010; Cooper et al. 2010; Morrow and Cooper 2012) and failed to analyze secondary replicons of multipartite genomes (Couturier and Rocha 2006; Flynn et al. 2010). However, there are also a number of studies that failed to observe this positive linear correlation in the absence of selection with mutations and mutation rates (Hudson et al. 2002; Ochman 2003; Martina et al. 2012; Juurik et al. 2012; Foster et al. 2013; Long et al. 2016; Dettman et al. 2016; Dillon et al. 2018). In this work we explored the spatial trends of substitutions and dN, dS, and ω values along bacterial genomes to add to the previous knowledge of spatial trends in bacteria. This study takes a unique approach to the analysis of how the number of substitutions changes with distance from the origin of replication by accounting for local and large scale genomic rearrangements by utilizing ancestral reconstruction techniques of both substitutions and genomic positions.

Although thousands of bacterial genomes have been sequenced for bacteria with different genomic structures, the majority of these genomes are incomplete and are composed of scaffolds or contigs. For this analysis, a complete genome, free of gaps or contigs, was necessary to accurately track substitutions and their genomic locations. Incomplete genomes would have gaps in genome positions, leaving missing information about substitutions for these segments of sequence. Therefore, we wished to consider only complete genomes. We would like to expand our analyses in the future to incorporate more genomes and taxa, but currently there are few that are suitable to our specific requirements.

We were unable to observe a consistent significant correlation between distance from the origin of replication and the number of substitutions per site as well as the values of $\mathrm{d}N$, $\mathrm{d}S$, and ω in the replicons that were analyzed. This necessitates further in-depth analysis of other molecular trends in bacterial genomes while accounting for genomic reorganization. Using tools such as ancestral reconstruction and the history of rearrangements, other spatial molecular trends in bacteria can be elucidated. This can be applied to gene expression and essentiality, to determine how these molecular components are impacted by rearrangements and what this tells us about the organization of genes along bacterial genomes.

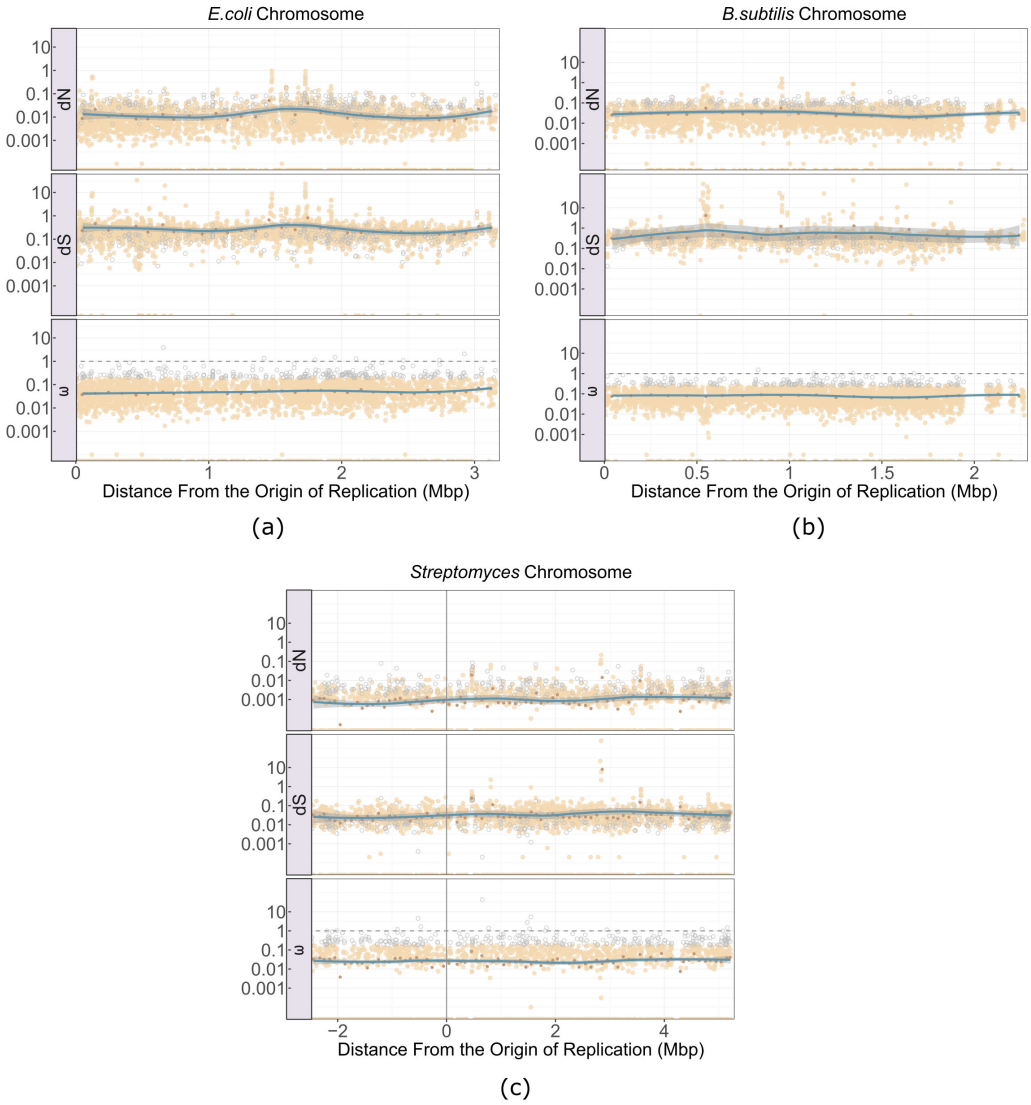


FIGURE 2.5: The graphs show the values of dN, dS, and ω along the genomes of *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c). For *E. coli* and *B. subtilis*, the distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. For *Streptomyces* the origin of replication is denoted by position zero. The genome located on the shorter chromosome arm (to the left of the origin) has been given negative values, while the genome on the longer chromosome arm (to the right of the origin) has been given positive values. The origin of replication in the *Streptomyces* graph (c), has been visualized at position zero by a grey vertical line. The y-axis of the graph indicates the value of dN, dS, and ω found at each gene segment position of the *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c) genomes. Outliers are represented by light grey open circles. The average dN, dS, and ω values for each 100,000bp region of the genome was calculated and represented by the dark brown points. A trend line represented in blue (using the `loess` method), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. For a complete list of outlier and zero value information, please see the Supplementary Material.

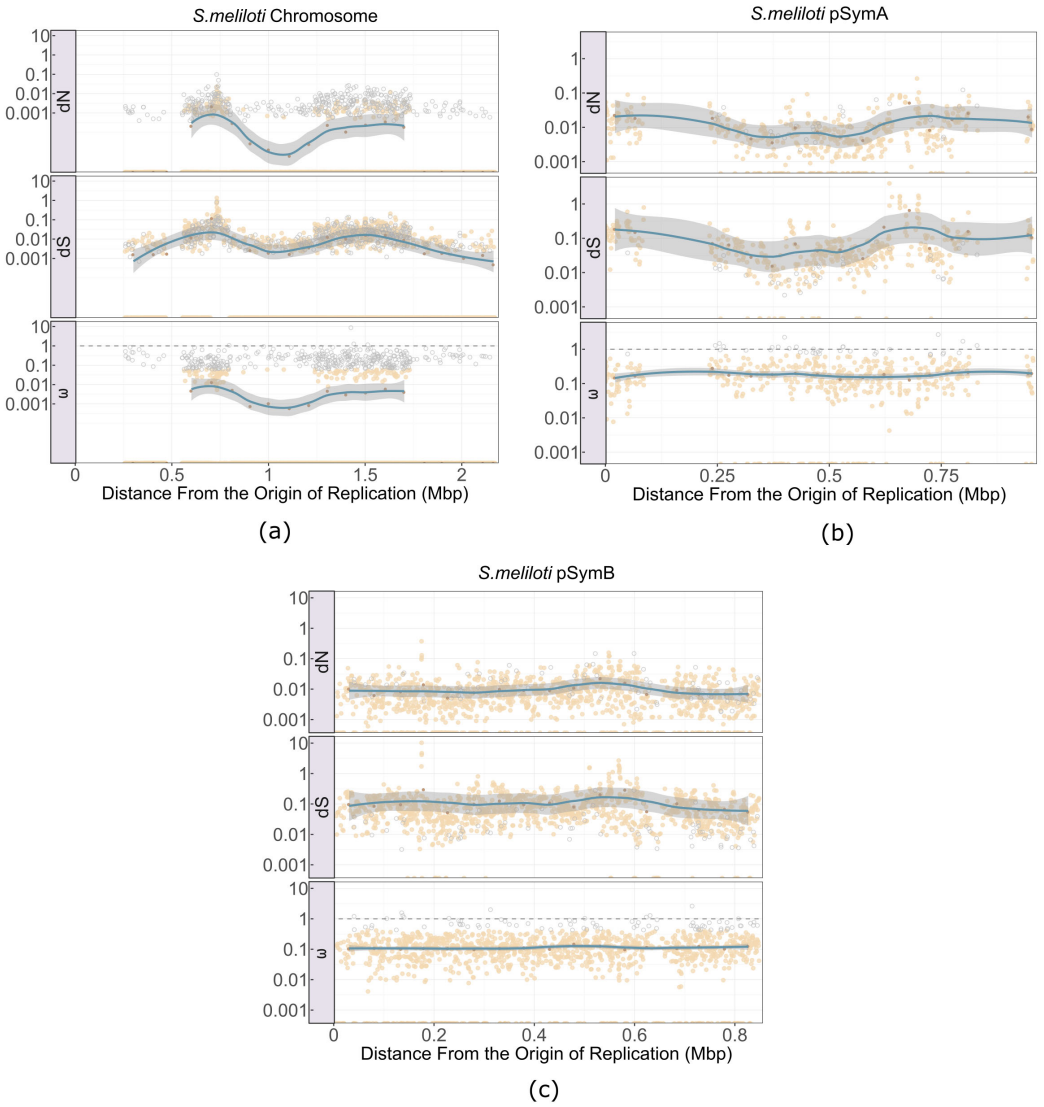


FIGURE 2.6: The graphs show the values of dN, dS, and ω along the replicons of *S. meliloti*, chromosome (a), pSymA (b), and pSymB (c). Distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. The y-axis of the graph indicates the value of dN, dS, and ω found at each gene segment position of the chromosome (a), pSymA (b), and pSymB (c) of *S. meliloti*. Outliers are represented by light grey open circles. The average *dN*, *dS*, and ω values for each 100,000bp region (for the chromosome) and 50,000bp region (for both pSymA and pSymB) of the replicons was calculated and represented by the dark brown points. A trend line represented in blue (using the `loess` method), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. For a complete list of outlier and zero value information, please see the Supplementary Material.

2.6.1 Spatial Substitution Trends

We have demonstrated here that any correlation between the number of substitutions and genomic position in our bacterial species is significant but small and inconsistent in sign (Table 2.2). In this analysis, we have looked at protein coding genes within the genomes of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*, including both core and accessory genes. Previous studies looking at substitution rates and genomic position typically looked at orthologous genes with similar genomic positions (Cooper et al. 2010; Morrow and Cooper 2012). The discrepancy between our results and previously published analyses may be due to our alignments having dissimilar genomic positions in some taxa and the inclusion of genomic reorganization. Some segments of the genomes have relatively high numbers of substitutions compared to the rest of the genome. For example, the high bars located near 2Megabase Pair (Mbp) from the origin in *B. subtilis* (Figure 2.3b) seem to have an increase in the number of substitutions in this genomic segment relative to the other 10Kbp regions. These high substitution regions are homologous genes or gene segments that happen to have amino acid changes which are driving the high number of substitutions in those bars. An illustrative example of one such gene segment can be found in the Supplementary Figures S11 and S12.

The multi-repliconic nature of *S. meliloti* appears to have a small effect on the overall spatial substitution trends of each replicon. For example, the opposing spatial substitution trends (Table 2.2, Figure 2.4) of different replicons in *S. meliloti* may be due to an over representation of highly expressed or essential genes located on the chromosome. We found an increased number of substitutions in the smaller replicons, pSymA and pSymB, compared to the chromosome. The smaller replicons are known to display less genomic conservation than the chromosome (Cooper et al. 2010; Morrow and Cooper 2012), and have genes used for local environmental adaptation (Medini et al. 2008; DiCenzo et al. 2019), which may explain the increased number of substitutions in pSymA and pSymB, compared to the chromosome.

A number of previous studies have complementary results regarding increasing substitution trends of bacterial replicons which was found in *B. subtilis*, *Streptomyces* and the small replicons of *S. meliloti* in this analysis. These previous studies observed gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012; Lato and Golding 2020a) decreases, while substitution rate was found to increase with increasing distance from the origin of replication (Prescott and Kuempel 1972; Morrow and Cooper 2012; Galardini et al. 2013). Genes that are less essential and often expressed less tend to evolve quickly compared to more conserved genes with higher expression levels (Sharp et al. 1989). pSymB of *S. meliloti* has been known to house essential genes (Cooper et al. 2010; Morrow and Cooper 2012), and *Streptomyces* has majority of it's essential genes concentrated near the origin of replication (Bentley et al. 2002; Kirby 2011). Additionally, pSymB has been shown to be more transcriptionally integrated with the chromosome compared to pSymA (DiCenzo et al. 2018), potentially contributing to the location of essential genes. Some of the proteins encoded on pSymB that are not necessarily deemed

essential, are still able to fulfill essential gene roles and functions (DiCenzo et al. 2018). These essential genes should have a decreased number of substitutions and therefore, coincide with the increasing substitution rate when moving away from the origin of replication in *Streptomyces* and pSymB of *S. meliloti*.

Molecular composition, gene content, and replication may all be factors contributing to the curious decreasing number of substitutions with increasing genomic distance found in *E. coli* and the chromosome of *S. meliloti* in this study. The integration of new genetic information through gene gain and loss sometimes occurs in particular regions along bacterial genomes termed "hotspots" (Farabaugh et al. 1978; Streisinger et al. 1966; Touchon et al. 2009; Oliveira et al. 2017). The frequency of these hotspots increases linearly with distance from the origin of replication (Oliveira et al. 2017), although different mobile elements such as integrative and conjugative elements and prophages, appear to have a different distribution (Oliveira et al. 2017). Variation in these preferential sites for gene gain and loss could be located near the origin of replication and may illuminate why we observed the number of substitutions to significantly decrease with distance from the origin of replication in the chromosomes of *E. coli* and *S. meliloti*. Some studies found inconsistencies, with the placement of core genes concentrated near the terminus or distributed evenly throughout the genome, rather than localized at the origin of replication (Kopejtka et al. 2019). Determining the distribution and placement of the core and accessory genes in *E. coli*, and *S. meliloti* could elucidate why these replicons appear to have a higher number of substitutions near the origin of replication. The distinct placement of genes across the genome is speculated to be in part due to the nature of replication. Translocations can happen at replication forks as they advance along the chromosome (Tillier and Collins 2000; Mackiewicz et al. 2001). If these replication forks were concentrated near the origin of replication, creating a hotspot for an increased number of translocations present in that area, providing an opportunity for new genomic signatures to arise, such as a minor increase in the number of substitutions near the origin of replication.

Additionally, potential genomic and pathogenicity islands have been found near the origin of replication in *Mycobacterium tuberculosis* and *Haloquadratum walsbyi* (Karlin 2001; Mira et al. 2010). These islands were found to have genomic signatures such as codon bias, that deviated from the rest of the genome (Karlin 2001). Deviations in these genomic signatures may extend to substitution rates and provide another potential explanation as to why some of the replicons in this study had a slight increase in the number of substitutions near the origin of replication. Other genomic signatures such as GC content or nucleotide composition have been found to significantly change around the origin of replication and terminus (Mackiewicz et al. 1999; Ikeda et al. 2003), and may be a contributing factor in explaining a higher number of substitutions near the origin of replication in *E. coli* and the chromosome of *S. meliloti*, and warrants further investigation.

Rearrangements, inversions, duplications, and HGT all play a major role in shaping gene order, gene expression, gene content, and substitutions in bacterial replicons. One study found that the density of transposon insertion events peaks at the origin of replication and is at a minimum at the terminus in *E. coli* (Gerdes et al. 2003). Once again, the differences in various genomic signatures caused by genome reorganization, in this case transposon insertion events, may be a justification for the high number of substitutions seen near the origin in some chromosomes in this analysis. The lack of a clear spatial genomic substitution trend in the genomes used, highlights the importance of accounting for genomic reorganization, such as rearrangements, in molecular analyses.

2.6.2 Spatial Selection Trends

Looking at the correlation between dN, dS, and ω values and distance from the origin of replication, we were unable to confirm a consistent linear correlation in the genomes analyzed (Table 2.4 and Figures 2.5 and 2.6). There are a few sparse areas in the distribution of dN, dS, and ω values across the genomes. These are areas where alignment data were removed due to poor homology, excessive gaps, or not being present in all taxa. We manually looked into genes with unusually high values of dN and dS, and we have determined that these values indeed represent genes with a high number of substitutions. The substitutions in these genes often have many (or only) substitutions of one type (i.e. synonymous or non-synonymous), skewing the dN or dS calculation, causing the unusually high values. These genes can be assumed to have a high degree of divergence between the taxa, and often encode for unconfirmed proteins such as hypothetical proteins (see Supplementary Material). Conversely, all *S. meliloti* chromosomes used in this analysis are extremely similar and therefore resulting in an overall low number of substitutions. The majority (61%) of the genes and gene segments in the chromosome of *S. meliloti* had dN values of 0, and therefore ω values of 0 (Supplementary Material). These zero values were not removed from the analysis or outlier calculations because they were too numerous to be outliers and they provide important information about the similarities between these strains of *S. meliloti*. The low number of substitutions, and consequently high numbers of zero dN, dS, and ω values in this bacteria are reflected in Figure 2.6.

As mentioned previously, the number of bacterial genomes used for each analysis was limited partially due to computational constraints completing the `progressiveMauve` whole genome alignment. Specialized alignment programs such as `Parsnp` (Treangen et al. 2014), identify and align only core regions of the genomes relatively quickly. Dealing with only core regions would reduce the potential for including alignments of poor sequence homology. This could allow the current analysis to be expanded to include more genomes of each bacterial species, and potentially add more phylogenetic diversity in the species chosen. However, using only the core genome removes valuable data from the analysis such as accessory genes, where most variations in mutation rate would be seen (Couturier and Rocha 2006; Flynn et al. 2010).

This work is not the first to observe diverging results from the general consensus of bacterial molecular trends. These notable exceptions to what are thought to be generally applicable rules of bacterial molecular trends, question the broad universal assumption of these phenomenon. With respect to mutations, there were a number of studies that were unable to confirm a positive linear correlation between distance from the origin of replication and mutation rates (Hudson et al. 2002; Ochman 2003; Martina et al. 2012; Juurik et al. 2012; Dettman et al. 2016; Dillon et al. 2018). Some of these patterns are thought to be a regional effect of sequence composition (Hudson et al. 2002), while others are more related to cell cycle function (Dillon et al. 2018). There are a number of other intertwining factors that impact the mutation spectra of bacteria such as transcription, replication, and growth state (Hudson et al. 2002; Ochman 2003; Juurik et al. 2012). When looking at differences in mutations between replicons of the multi-repliconic bacteria *Burkholderia*, substitutions are highest on the primary chromosomes compared to the secondary replicons (Dillon et al. 2015). This finding was unrelated to nucleotide composition and due to some substitutions occurring at higher rates on particular replicons (Dillon et al. 2015).

2.7 Conclusions

The integration of genomic reorganization, such as rearrangements and inversions, can have impacts on spatial molecular trends such as substitution rate. The general molecular trends previously found in bacteria when moving away from the origin of replication may not be as commonplace as expected, particularly when genome reorganization occurs. By utilizing ancestral reconstruction, we have demonstrated how information on genomic reorganization can be used to elucidate the spatial pattern of substitutions along bacterial genomes. We have illustrated that overarching spatial molecular trends may not be as universal as previously thought. We have found significant but small and inconsistent correlations between the number of substitutions and distance from the origin of replication in the genomes analyzed. We did not observe a consistent significant correlation between $\mathrm{d}N$, $\mathrm{d}S$, and *omega* values and distance from the origin of replication in the genomes analyzed. Combining genomic reorganization and current molecular pipelines through processes such as ancestral reconstruction, can add vital information to bacterial genome analyses. We believe that genomic location and genome reorganization are important to consider in future molecular evolutionary analyses in all areas such as gene expression, essential gene locations and functional classification of those genes. Observing other molecular trends through the lens of genomic reorganization will assist in answering questions about the evolution of bacteria.

2.8 Supplementary Material

Supplementary Figures S1.1 - S1.19 and Tables S1.1 - S1.22 are available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/). Further supplemental code, data, and information for this article are available on `GitHub` at `www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`.

2.9 Data Availability

The data underlying this article are available on `GitHub` at `www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`.

2.10 Acknowledgements

We thank Caitlin Simopoulos for comments on the manuscript. This work was supported by the Natural Sciences and Engineering Research Counsel (grant number RGPIN-2015-04477 to GBG).

Chapter 3

Spatial Patterns of Gene Expression in Bacterial Genomes

Daniella F. Lato and G. Brian Golding

As published in the Journal of Molecular Evolution, 2020, 88, pages 510–520
https://doi.org/10.1186/s12864-018-4665-2

3.1 Preface

Chapter 3 describes the identification and analysis of gene expression patterns along various bacterial genomes. As described in Chapter 1, gene expression and gene dosage are increased near the origin, and genes are typically less conserved with increasing distance from the origin of replication. However, these studies do not take a genomic approach to looking at gene expression and often focus on one gene or a small subset of genes. In this work, we combine RNA-seq data from eleven previously published experiments to evaluate genomic expression levels in *Escherichia coli*, *Bacillus subtilis*, *Streptomyces*, and *Sinorhizobium meliloti*. We have determined the expression landscape along the origin and terminus of replication axis in these bacterial replicons. This chapter is published in Journal of Molecular Evolution as: D. F. Lato and G. B. Golding (2020a). Spatial patterns of gene expression in bacterial genomes. J Mol Evol 88, 510–520. I made significant contributions to this study. I conceived the experiment jointly with G.B. Golding. I curated RNA-seq datasets and evaluated the spatial genomic trends of gene expression in each species. I developed custom pipelines and scripts to complete this analysis. I wrote the first version of this manuscript, which was edited and approved by G.B. Golding. G.B. Golding supervised the analyses and writing of the manuscript.

3.2 Abstract

Gene expression in bacteria is a remarkably controlled and intricate process impacted by many factors. One such factor is the genomic position of a gene within a bacterial genome. Genes located near the origin of replication generally have a higher expression level, increased dosage, and are often more conserved than genes located farther from the origin of replication. The majority of the studies involved with these findings have only noted this phenomenon in a single gene or cluster of genes that was re-located to pre-determined positions within a bacterial genome. In this work, we look at the overall expression levels from eleven bacterial data sets from *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. We have confirmed that gene expression tends to decrease when moving away from the origin of replication in the majority of the replicons analyzed in this study. This study sheds light on the impact of genomic location on molecular trends such as gene expression, and highlights the importance of accounting for spatial trends in bacterial molecular analysis.

3.3 Introduction

Gene expression in bacteria is complex and highly controlled. The regulation of bacterial gene expression is a crucial component of bacterial survival in order for these organisms to modulate gene expression and alter phenotypic properties such as growth rate (Garmendia et al. 2018) and motility (Ravichandar et al. 2017). Gene expression can be controlled through a variety of promoters, physical chromosome structure, and the DNA replication machinery. Therefore, different genes can be under distinct methods of regulation and be expressed at fluctuating levels depending on environmental conditions or growth stage. This variation in expression can be influenced by a myriad of effects such as differences in codon bias (Gutman and Hatfield 1989; Sharp et al. 1989; Buchan et al. 2006; Cannarozzi et al. 2010; Quax et al. 2015), gene orientation (Zeigler and Dean 1990; Kunst et al. 1997; Price et al. 2005), replication (Rocha 2004b; Washburn and Gottesman 2011; Block et al. 2012; Garmendia et al. 2018), and chromosomal location (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012). These phenomena can create predictable patterns that can be observed in many molecular traits across many bacterial species.

One set of patterns is related to the physical location of genes on the chromosome. Some studies have found certain genes and groups of genes to be expressed periodically around the chromosome. Wright et al. (2007), looked at statistically correlated gene pairs in *E. coli* and found that they are often separated by 100Kilobase Pairs (Kbps) and are often located in areas of high transcription. Other studies of *E. coli* observed that sections of the chromosome with increased transcription rates were periodically found throughout the genome over 700-800Kbps ranges (Jeong et al. 2004). It is speculated that this periodic phenomenon is due to a combination of physical constraints of the chromosome, such as supercoiling and DNA composition (Jeong et al. 2004; Képes 2004; Peter et al. 2004; Allen et al. 2006; Block et al. 2012). Prior research

on spatial molecular trends when moving from the origin of replication to the terminus have determined that gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012) and gene dosage (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Sauer et al. 2016) are increased near the origin, and genes become less conserved with increasing distance from the origin (Couturier and Rocha 2006). Additionally, substitution rates (non-synonymous (dN), synonymous (dS)), and the dN/dS ratio, increase with distance from the origin of replication (Cooper et al. 2010; Morrow and Cooper 2012). The variation in molecular trends with genomic location has been suspected to be due to a number of complicated and intertwining factors such as transposon insertion events (Gerdes et al. 2003), gene order and conservation (Mackiewicz et al. 2001; Flynn et al. 2010), replication (Couturier and Rocha 2006), and nucleotide composition (Mackiewicz et al. 1999; Karlin 2001; Sharp et al. 2005).

Gene expression in particular consistently varies with distance from the origin of replication. A number of previous studies have analyzed this spatial trend in a variety of bacteria such as *E. coli*, *Brucella*, and *Vibrio*. Both large- (Sharp et al. 2005; Couturier and Rocha 2006) and small-scale studies (Schmid and Roth 1987; Morrow and Cooper 2012; Block et al. 2012; Bryant et al. 2014; Garmendia et al. 2018), have detected decreasing gene expression values as genomic distance increases away from the origin of replication. However, the majority of these studies often only look at a single gene or cluster of genes and promoters (Schmid and Roth 1987; Block et al. 2012; Bryant et al. 2014; Garmendia et al. 2018). In these studies, genes or gene clusters are experimentally moved to per-determined locations around the replicon. This type of experiment can lead to biases stemming from the original location of the genes and the relative distance from the origin of replication. Additionally, the genes chosen are often selected because of their ability to be easily moved to various genomic locations. Choosing specific genes to manipulate and move around bacterial genomes is fundamental to understanding how the location of a gene on a chromosome impacts it's expression. However, observing one gene does not provide us with a complete picture of what is happening with gene expression from a genomic viewpoint.

Although many studies have found that gene expression decreases with increasing distance from the origin of replication, it is unclear if this phenomenon is persistent across diverse genomes and bacterial species. In this work we aim to answer this question by looking at the overall expression levels of all genes within eleven gene expression data sets from bacterial genomes of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. These bacteria inhabit a variety of different environments and cover a range of genomic structures and replication strategies. Some of the bacteria in this study have a single circular (*E. coli* and *B. subtilis*) or linear chromosome (*Streptomyces*) containing its genome, while others have the genome split up into multiple replicons (*S. meliloti*). Each of these genomic structures requires precise coordination between transcription and translation in order to replicate efficiently. This selection of bacterial taxa provides a sample that covers broad lifestyles as well as representing a number of divergent phylogentic lineages, providing a diverse sample for answering if gene expression decreased with increasing

distance from the origin of replication in across diverse bacterial genomes and species. Using whole genome expression data obtained from the Gene Expression Ombinbus (GEO) database (Barrett et al. 2012), we are able to observe genomic expression patterns in natural populations devoid of stress, while accounting for bidirectional replication. We have confirmed that gene expression indeed tends to be higher near the origin of replication and decreases with increasing distance from the origin. Understanding how the distance of a gene from the origin of replication can impact the expression level assists in explaining other spatial distance trends such as gene essentiality, gene conservation, and mutation rates.

3.4 Materials and Methods

3.4.1 Expression Data

The bacteria chosen for this analysis were *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. These bacteria inhabit a variety of different living environments and have contrasting genomic structures (i.e. circular, linear, multi-repliconic), providing a well rounded sample for this analysis. Although *E. coli*, *B. subtilis*, and *Streptomyces* contain small plasmids, they are not considered multi-repliconic bacteria and therefore their plasmids were not included in this analysis. *S. meliloti* is a multi-repliconic bacteria and its two large secondary replicons were included in the analysis (pSymA and pSymB). The replicons of *S. meliloti* are known to differ in genetic content, and therefore, all analyses were performed on each individual replicon of *S. meliloti*.

Gene expression data for *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti* was downloaded from the GEO (Barrett et al. 2012). The expression data sets for this analysis were only RNA-seq data sets for control data, where this was defined as the bacteria being grown in optimal growth conditions. Using strictly raw RNA-seq expression data allows the normalization to be standardized across all datasets, making the datasets directly comparable. The additional condition of using expression data where the bacteria were grown in control or stress free environments again allows for direct comparisons to be made between spatial gene expression trends between these bacterial species. Due to these constraints on our data, we were only able to retrieve a total of 11 gene expression datasets from GEO for this analysis.

Pseudogenes were excluded from this analysis. A complete list of expression data used is found in Supplementary Table S2.1. Correlation of gene expression across data sets was assessed for each bacteria with multiple data sets. For a detailed protocol, see Supplementary files on `GitHub` at `https://github.com/dlato/Spatial_Patterns_of_Gene_Expression.git`.

3.4.2 Normalization

The raw counts from control populations for each data set were used and normalized using the Trimmed Mean of M values (TMM) method (Robinson and Oshlack 2010). Raw counts were normalized to Counts Per Million (CPM) in `R` using the `edgeR` package (Robinson et al. 2010).

After normalization, any data sets that had multiple replicates were combined by finding the median CPM between replicates for each annotated gene. Only genes that had expression values in all data sets were used for this analysis.

3.4.3 Genomic Position

To relate the median CPM gene expression values to position in the genome a custom `Python` script was written to determine the midpoint position of each annotated gene in the bacterial genome. This allowed a single position location for each gene which simplifies the following regression calculations.

3.4.4 Origin and Bidirectional Replication

For each bacterium in this analysis, the beginning of the origin of replication was denoted as the beginning of the *oriC* region for the chromosomal replicons, and the beginning of the *repC* (Pinto et al. 2011) region for the secondary replicons of *S. meliloti* (Supplementary Table S2.2). This origin of replication position was calibrated to be the beginning of the genome, or position 1, and remaining positions in the genome were all scaled around this origin of replication (Figure 3.1).

To determine if specifying a single nucleotide as the origin of replication would alter the results, we performed permutation tests. These tests shuffled the *oriC* position by 10,000Base Pairs (bps) increments in each direction from the original origin (data not shown) to a maximum of 100,000bps in each direction. These results showed that moving the origin of replication does not affect the results of the analysis (data not shown).

The terminus of replication was determined using the Database of Bacterial Replication Terminus (DBRT) (Kono et al. 2011). DBRT uses the prediction of *dif* sequences as a proxy for the terminus location because the *dif* sequences are located in the replication termination region of the chromosome (Clerget 1991; Blakely et al. 1993). For pSymA and pSymB of *S. meliloti* the terminus is not listed in the database, thus the terminus location was assigned to the midpoint between the origin of replication and the end of the replicon. Replication in the linear chromosome of *Streptomyces* begins at the origin of replication, located to the right of the middle of the replicon (Heidelberg et al. 2000), and terminates at each end of the chromosome arms (Heidelberg et al. 2000) (Supplementary Table S2.2).

The origin scaling and bidirectional replication transformations were done in `R` (R Development Core Team 2014) and allow inferences to be made about gene expression while recording their distance from the origin of replication. A diagram of this transformation is outlined in Figure 3.1.

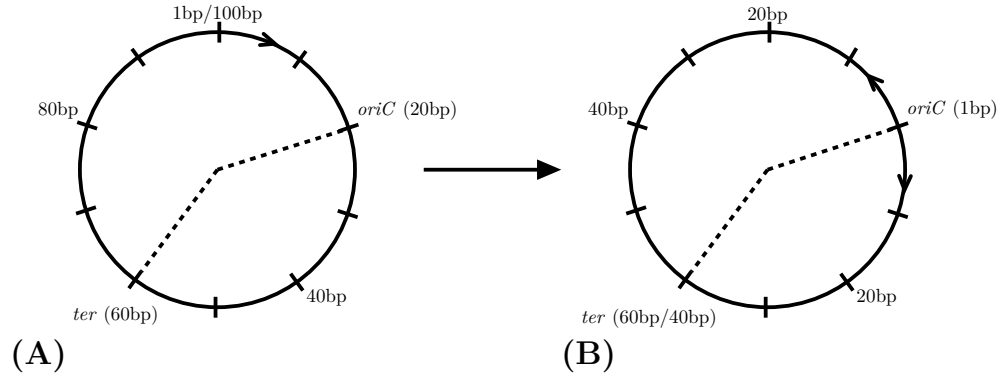


FIGURE 3.1: Schematic of the transformation used to scale the positions in the genome to the origin of replication and account for bidirectional replication. Circle (A) represents the original replicon genome without any transformation. Circle (B) represents the same replicon genome after the transformation. The origin of replication is denoted by "*oriC*" and the terminus of replication is denoted by "*ter*". The dashed line represents the two halves of the replicon separate by replication. The replicon genome in this example is 100 base pairs in length. Every 10 base pairs is denoted by a tick on the genome. The origin in (A) is at position 20 in the genome and is transformed in (B) to become position 1. The terminus is at position 60 in (A) and position 60 and 40 in (B). The terminus has two positions in (B) depending on which replicon half is being accounted for. If the replication half to the right of the origin is considered, the terminus will be at position 40. If the replication half to the left of the origin is considered, the terminus will be at position 60. Position 40 in (A) becomes position 20 in (B). Position 80 in (A) becomes position 40 in (B), because of the bidirectional nature of bacterial replication. "bp" denotes base pairs.

E. coli, *B. subtilis*, and all replicons of *S. meliloti* have a terminus of replication which is located roughly equidistant from the origin of replication (Supplementary Table S2.2). These bacteria therefore have approximately symmetrical chromosomal arms and as a result have genomic position labelling in Figures 3.2 and 3.3, accounting for bidirectional replication. *Streptomyces* on the other hand, is an acrocentric linear chromosome with one chromosomal arm being much shorter than the other (see Figure 3.2). The genomic position labelling of *Streptomyces* in Figure 3.2 has negative numbers to indicate the shorter chromosome arm, and positive numbers indicating the longer chromosome arm.

3.4.5 Average Gene Expression

The average gene expression per genome was calculated for each bacterial replicon. This was computed by taking the arithmetic mean of all normalized CPM gene expression values for the entire replicon.

A single median CPM per 10Kbps section of each bacterial genome was calculated. The gene expression information was summarized in bar graphs in `R` using `ggplot2` (Wickham 2009)

(Figures 3.2 and 3.3). Supplementary interactive figures can be found on GitHub (https://github.com/dlato/Spatial_Patterns_of_Gene_Expression.git).

3.4.6 Linear Regression

To assess the statistical significance of changes in expression with genomic position a simple linear regression was performed in R (R Development Core Team 2014). An average CPM expression value was calculated for each 10Kbps region of the genome. This was calculated by taking the sum of all CPM expression values over a 10Kbps region of the genome, and dividing this by the total number of genes present in that 10Kbps segment. A linear regression was performed on these 10Kbps average expression values to determine if there was a significant correlation between gene expression and distance from the origin of replication. Statistical outliers in this data set were removed from the linear regression. Outliers were defined as being outside the first quartile minus 1.5 times the interquartile range, and the third quartile plus 1.5 times the interquartile range. Additional linear regressions on a per gene basis, non-average expression values, and total additive expression values were also calculated. These results and methods can be found in the Supplementary Material (Supplementary Tables S2.3 - S2.5).

The total number of protein coding genes was determined for each 10Kbps region of the genome. To assess the statistical significance of the total number of genes in each 10Kbps region of the genome and position in the genome a simple linear regression was performed in R (R Development Core Team 2014).

A supplementary test to determine if gene expression differs between the leading and lagging strands of each bacterial replicon was performed. A two-sample Wilcox test was computed in R (R Development Core Team 2014) to compare expression of genes on the leading strand and the lagging strand. We found that there was no significant difference between gene expression on the leading and lagging strand in most of the bacterial replicons. The exceptions to this were *Streptomyces* and the chromosome of *S. meliloti*, which had a significant difference between gene expression on the leading and lagging strand, with higher gene expression on the leading strand. Full results can be found in the Supplementary Material. The percent of genes that reside on the leading strand of the various bacterial replicons was between approximately 54% and 74% (see Supplementary Material).

3.5 Results and Discussion

3.5.1 Origin and Bidirectional Replication

Bacterial chromosome replication begins at the origin of replication and proceeds away from the origin in both directions (Prescott and Kuempel 1972). Bidirectional replication affects the genomic location of the farthest point from the origin. Replication concludes at the terminus (Prescott and Kuempel 1972), which in circular replicons is usually located opposite from the

origin (Kono et al. 2011). However, in some bacteria the terminus is not exactly opposite from the origin. In a case like this, some of the distance measurements will only account for one of the replication halves (Figure 3.1). However, due to the nearly symmetrical location of the terminus to the origin, this effect is small.

In this analysis, a single base was chosen to represent the origin of replication. In reality, the origin of replication is often a number of base pairs long and choosing the first nucleotide position of this *oriC* region or the last nucleotide of this region may alter the subsequent bidirectional replication transformations and results. We performed permutation tests (data not shown) to determine the impact of altering the location of the origin of replication position. These results from our origin of replication permutation tests determined that moving the origin of replication does not affect the overall trends, providing a robust check for origin of replication location.

Bacteria and Replicon	Average Expression Value (CPM)
E. coli Chromosome	176.009
B. subtilis Chromosome	186.533
Streptomyces Chromosome	6.453
S. meliloti Chromosome	286.723
S. meliloti pSymA	764.793
S. meliloti pSymB	628.318

Table 3.1: Arithmetic mean gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million.

Bacteria and Replicon	Regression Slope of the Change in Gene Expression with Distance from the Origin of Replication
E. coli Chromosome	-3.65×10^{-5}***
B. subtilis Chromosome	-2.48×10^{-5}**
Streptomyces Chromosome	-1.41×10^{-7}**
S. meliloti Chromosome	NS
S. meliloti pSymA	NS
S. meliloti pSymB	NS

Table 3.2: Linear regression results of average expression and distance from the origin of replication. The average expression values were calculated by dividing the total counts per million expression value per 10kb section of the genome by the total number of genes in the respective 10kb section. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. Statistical outliers were removed from this linear regression calculation. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', > 0.05 = 'NS'. A grey row indicates a significant negative trend.

Bacteria and Replicon	Regression Slope of the Change in Number of Genes with Distance from the Origin of Replication
E. coli Chromosome	NS
B. subtilis Chromosome	-3.00×10^{-6}***
Streptomyces Chromosome	NS
S. meliloti Chromosome	-1.99×10^{-6}***
S. meliloti pSymA	NS
S. meliloti pSymB	-4.11×10^{-6}*

Table 3.3: Linear regression analysis of the total number of protein coding genes per 10Kbps along the genome of the respective bacteria replicons. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. All results are marked with significance codes as followed: < 0.001 = '***', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

3.5.2 Average Gene Expression

A summary of the average gene expression values per bacterial replicon can be found in Table 3.1. Most of the bacterial replicons have an average normalized expression value between 175 CPM - 765 CPM (Table 3.1). *Streptomyces* has an average gene expression value that is about two orders of magnitude lower than the other bacterial replicons (Table 3.1). This could be because there was only one data set available for this analysis (see Supplementary Table S2.1), and the mapped reads were assigned using the `Galaxy` streCoel (*Streptomyces coelicolor* 07/01/1996) Assembly (Afgan et al. 2018). This particular assembly and workflow may be why the *Streptomyces* gene expression data has consistently lower normalized CPM values across the genome compared to the other bacterial replicons which use a different suite of software including the Tuxedo Protocol (Trapnell et al. 2012).

3.5.3 Linear Regression

The average CPM gene expression values were calculated over 10Kbps regions. A linear regression was performed on those values to determine if there was a significant trend correlating gene expression and distance from the origin of replication. Gene expression decreases when moving away from the origin of replication for the chromosomes of *E. coli*, *B. subtilis*, and *Streptomyces* (Table 3.2). We were unable to detect a significant linear regression coefficient estimate for all replicons of *S. meliloti*. Previous work in similar bacterial species looking at the distribution of highly expressed (Couturier and Rocha 2006) and orthologous genes (Morrow and Cooper 2012), also found genes with higher expression values to be concentrated near the origin of replication. Our results are consistent with these studies as we see a decrease in gene expression with increasing distance from the origin of replication. All linear regression and supporting statistical information for the gene expression trends are found in Table 3.2. We performed additional

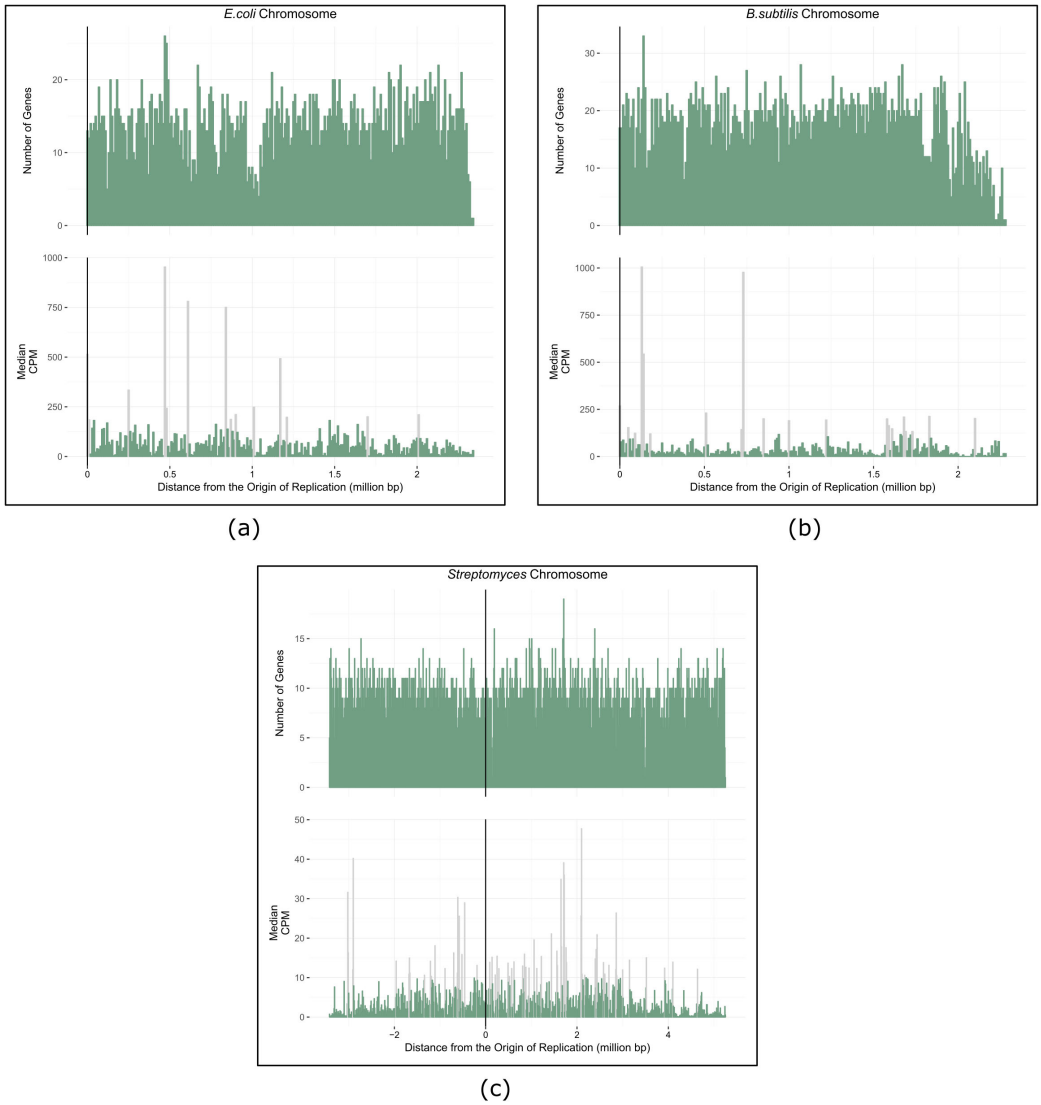


FIGURE 3.2: The top bar graphs show a count of the total number of genes (y-axis) at each position (x-axis) the genome of *E. coli* (a), *B. subtilis* (b) and *Streptomyces* (c). The bottom bar graphs show the median expression data along the genomes of *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c). The origin of replication is indicated by a black vertical line. For *E. coli* and *B. subtilis*, the distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. For *Streptomyces* the origin of replication is denoted by position zero. The genome located on the shorter chromosome arm (to the left of the origin) has been given negative values, while the genome on the longer chromosome arm (to the right of the origin) has been given positive values. The y-axis of the bottom graph indicates the total median CPM expression values found at each position of the *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c) genomes. Each bar represents a section of the genome that spans 10,000 base pairs. Light coloured bars represent statistical outliers

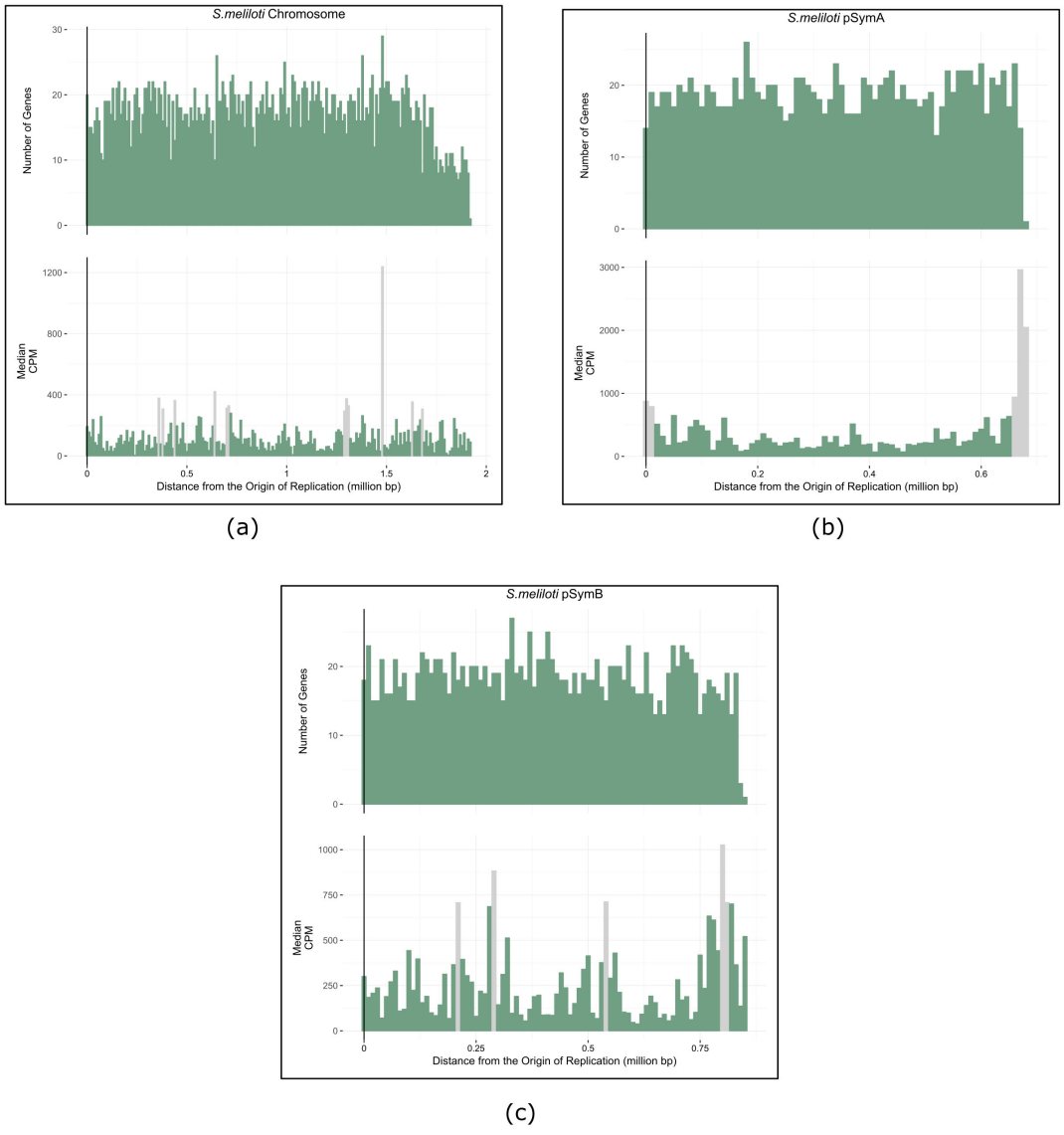


FIGURE 3.3: The top bar graphs show a count of the total number of genes (y-axis) at each position (x-axis) the replicons of *S. meliloti*: chromosome (a), pSymA (b) and pSymB (c). The bottom bar graphs show the median expression data along the *S. meliloti* replicons: chromosome (a), pSymA (b), and pSymB (c). The origin of replication is indicated by a black vertical line. The distance from the origin of replication is on the x-axis beginning with the origin of replication denoted by position zero on the left, and the terminus indicated on the far right. The y-axis of the bottom graph indicates the total median CPM expression values found at each position of the *S. meliloti* replicons: chromosome (a), pSymA (b), and pSymB (c). Each bar represents a section of the genome that spans 10,000 base pairs. Light coloured bars represent statistical outliers

statistical tests to look at how using different averaging methods for the gene expression values potentially altered the regression results. Some of these averaging methods included average gene expression over 10Kbps regions of the genome, and the total added expression over 10Kbps genomic regions. A full list of supplementary tests can be found in the Supplemental Material. We looked at the relationship between these averaged values and distance from the origin of replication and showed that there was no difference in averaging methods, and we still see gene expression decrease with increasing distance from the origin of replication. See Supplementary material for detailed methods of the additional regression tests.

Having higher gene expression values near the origin of replication has been linked to physical constraints and processes of the bacterial replicon (Képes 2004; Peter et al. 2004; Jeong et al. 2004; Allen et al. 2006; Block et al. 2012). For example, replication errors are thought to increase as replication moves farther from the origin of replication (Courcelle 2009). This impacts the placement of highly expressed and important genes where errors in replication could be detrimental to the gene product and the organism. Therefore, genes that are highly expressed and also essential to the survival of the organism might often be located near the origin of replication and on the leading strand to further avoid collisions between DNA and RNA polymerase (Rocha 2004b; Washburn and Gottesman 2011; Block et al. 2012). Genes that are part of the core genome of bacteria are typically located near the origin of replication (Sharp et al. 2005; Couturier and Rocha 2006; Flynn et al. 2010). These core genes make up the majority of bacterial genomes, so intuitively we should have a higher concentration of genes near the origin of replication. We determined that the total number of protein coding genes per 10Kbps decreases with distance from the origin of replication (Table S1.17). A higher concentration of genes is near the beginning of the genome, where we see increased expression, and a lower concentration of genes is near the terminus, where we observed decreased expression.

A number of studies suggest that it is the essentiality or function of the gene that impacts gene expression and organization of genes on the chromosome (Rocha and Danchin 2003; Rocha 2008). In particular, Couturier and Rocha (2006) found that only genes associated with transcription/translation were located close to the origin of replication, while other highly expressed genes are distributed randomly with respect to genomic location. To address this finding, we utilized the functional data available on the Clusters of Orthologous Groups of proteins (COG) database to assess how the functionality of genes change with distance from the origin of replication. A full account of the methods is found in the Supplementary Material. We found no clear pattern of genes with any functional COG category consistently being located near the origin of replication. This included genes that are associated with transcription and translation, which did not have a consistent correlation with distance from the origin of replication across all bacteria in this analysis. A full list of significant linear regression coefficients for all 24 COG functional categories can be found in the Supplementary Material. The lack of clear trends in functional categories changing with distance from the origin of replication leads us to believe that there may be mechanisms other than gene function dictating genomic gene expression trends in

bacterial genomes.

Gene dosage appears to play an important role in the location of genes along bacterial replicons (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Couturier and Rocha 2006; Block et al. 2012; Sauer et al. 2016). When gene expression is saturated, gene dosage (and the position of a gene near the origin of replication) can be used to increase transcription by increasing the number of transcripts of a gene (Couturier and Rocha 2006). This has implications for rapid growth periods in bacteria, allowing tighter control of growth in varying environmental conditions (Couturier and Rocha 2006). Faster growing species require overlapping replication cycles to allow replication to keep up with growth (Helmstetter 1996). This should therefore correlate with the strength in gradients of expression with distance from the origin of replication (Morrow and Cooper 2012). This allows for increased expression for genes replicated earlier, and decreased expression for genes replicated later (Sharp et al. 1989; Mira and Ochman 2002; Couturier and Rocha 2006; Dryselius et al. 2008) Both gene dosage and the growth rate of a bacteria could provide a mechanism by which selection could act to influence the locations of genes along bacterial replicons. The high concentration of highly expressed genes located near the origin of replication could be influenced by additional selective forces such as codon usage bias, can influence translational efficiency (Ikemura 1985; Kanaya et al. 1999; Sharp et al. 2005; Morrow and Cooper 2012).

We did not detect a significant relationship between gene expression and distance from the origin of replication for the replicons of *S. meliloti* (chromosome, pSymA and pSymB). Gene expression in this bacterium is not as well studied as the other bacteria used in this analysis (Martens et al. 2008). In our search for expression data, we identified fewer appropriate studies for *S. meliloti* to include in our data analysis. A smaller amount of gene expression data may be biasing the non-significant correlation between gene expression and distance from the origin of replication in this *S. meliloti*.

It has been suggested that the leading strand is favoured for the location of highly expressed genes to allow faster DNA replication and lower transcriptional losses (Brewer 1988). We found no statistical evidence for the leading strand to have higher expression levels compared to the lagging strand in most of the bacterial replicons, and have concluded that this is likely not driving the results of decreased gene expression with increased distance from the origin of replication. Previous studies have determined that the main factor that influences if a gene is on the leading or lagging strand is the essentiality of that particular gene, not expression (Rocha and Danchin 2003; Zheng et al. 2015). The number of bacterial genes on the leading strand varies between approximately 45% to 90% (Rocha 2002; Zivanovic et al. 2002; Koonin 2009; Mao et al. 2012). The bacterial replicons used in this analysis fall with this range, and therefore the leading and lagging strands are not influencing the results (see Supplementary Material).

Areas of the bacterial genomes with extremely high gene expression (Supplementary Table

S2.10), are regions that encode proteins involved in processes such as DNA repair and replication, RNA synthesis, metabolism, and ribosomal proteins. We expect these regions to have much higher expression levels compared to the rest of the genome because they encode proteins that are crucial to translation and replication processes. Shockingly, when accounting for bidirectional replication we see that some riboproteins in *E. coli*, *B. subtilis*, and *S. meliloti*, are not always located close to the origin of replication, and can be located up to 1.49Megabase Pairs (Mbps) away from the origin of replication (in the case of the chromosome of *S. meliloti*, see Supplementary Table S2.10 for more details).

3.6 Conclusions

The genomic location of a bacterial gene has a profound impact on the expression levels of that gene. Previous studies have focused on a small subset of genes (Schmid and Roth 1987; Block et al. 2012; Bryant et al. 2014; Garmendia et al. 2018), or expression trends in one bacterial species (Schmid and Roth 1987; Block et al. 2012; Morrow and Cooper 2012; Bryant et al. 2014; Garmendia et al. 2018). Here we assess gene expression levels across all protein coding genes within the bacterial genomes of *E. coli*, *B. subtilis*, and *Streptomyces*, and show there is a relationship with distance from the origin of replication. Most replicons in this study show that genes that are closer to the origin of replication have a higher expression level when compared to genes that are located farther from the origin of replication. This spatial variation is not unique to gene expression; other molecular trends such as gene conservation (Couturier and Rocha 2006) and substitution rate (Cooper et al. 2010; Morrow and Cooper 2012) also vary with distance from the origin. It is important to realize that the location of a gene within the genome will impact various molecular trends of that segment of DNA and may assist in explaining other phenomenon related to that gene. Further analyses on the spatial trends of other molecular traits such as substitution rate and gene essentiality will create a base of information on what molecular trends genomic location can alter.

3.7 Supplementary Material

Supplementary Figures S2.1 - S2.8 and Tables S2.1 - S2.10 are available on `GitHub` at `https://github.com/dlato/Spatial_Patterns_of_Gene_Expression.git`.

3.8 Acknowledgements

We thank Caitlin Simopoulos for comments on the manuscript. We thank the National Sciences and Engineering Research Council for funding for this project (Grant # RGPIN-2015-04477 to GBG).

Conflict of interest

The authors declare that they have no conflict of interest.

Chapter 4

Genomic Inversions in *Escherichia coli* Alter Gene Expression

Daniella F. Lato, Qing Zeng and G. Brian Golding

Formatted for submission to GENOME

4.1 Preface

Chapter 4 describes the identification of inversions and analysis of their impact on gene expression and genomic location in *E. coli*. As described in Chapter 1, multiple studies have shown that inversions can have various phenotypic effects in bacteria such as altering growth rate and gene expression. These studies primarily focus on creating novel inversions of one gene and it's promoters. They do not take a genomic approach to identifying inversions and how this can impact gene expression on a long-range scale. In this work, we used genomic sequence data to identify "naturally occurring" genomic inversions in various *E. coli* strains. We combined this information with previously published RNA-seq data to determine how gene expression differs between inverted and non-inverted segments of *E. coli* genomes. We have discussed the relationship between the genomic location of the identified inversions and the gene expression landscape along the origin and terminus of replication axis in these bacterial replicons. This chapter is formatted for submission in GENOME. I made significant contributions to this study. I conceived the experiment jointly with G.B. Golding. I curated genomic datasets and identified all inversions between various *E. coli* strains validating alignments with `DIAMOND`. I curated RNA-seq datasets. Q. Zeng assisted with data processing of the RNA-seq datasets and `DIAMOND` results under the supervision of D.F. Lato. I analyzed changes in gene expression induced by the identified inversions. I performed all statistical analysis. I developed custom pipelines and scripts to complete this analysis. I wrote the first version of this manuscript, which was edited and approved by G.B. Golding. G.B. Golding supervised the analyses and writing of the manuscript.

4.2 Abstract

Genomic reorganization, such as rearrangements, inversions, and duplications, influence how genetic information is organized along bacterial genomes. Inversions in particular, show evidence of being non-random and can facilitate bacterial genome evolution through gene gain and loss. Inverting a segment of sequence can impact location on the leading or lagging strand and may facilitate the gain and loss of genes, potentially promoting novel functions. Specific inversions have been know to alter gene expression, acting as control switches. Previous studies investigating the impact inversions have on gene expression typically induce inversions targeting specific genes or look at inversions between distantly related species. This fails to encompass a genome wide perspective on inversions and gene expression. Here we use whole genome alignment and BLAST techniques to identify genomic inversions and rearrangements between closely related strains of *E. coli*. We investigate the short- and long-range impact these inversions have on genomic gene expression using multiple RNA-seq datasets. We observed significant differences in gene expression between inverted and non-inverted regions of the *E. coli* genomes, and found that inverted genes had 1.27-85.58 fold higher gene expression in 75% of significant inverted regions. The nucleoid associated proteins H-NS and Fis have been associated with genome wide gene expression repression and activation respectively. We observed a significant positive correlation between the identified inversions and H-NS binding sites, and a significant positive correlation between identified inversions with a significant difference in gene expression within the alignment region and Fis binding sites. Inversions impact gene expression even between closely related strains of *E. coli*, and could provide a mechanism for the diversification of genetic content through controlled expression changes.

Key Words: Inversions, Gene Expression, RNA-seq, H-NS protein, Fis protein, *E. coli*, Genomics

4.3 Introduction

Genomic reorganization in bacteria, such as rearrangements and inversions, provides one way genetic diversity is created (Hughes et al. 2000; Belda et al. 2005), which can assist in adaptation (Kresse et al. 2003; Rocha 2004a; Hanage 2016), and gene conversion (Hanage 2016). There has been evidence that inversions are non-random and can promote bacterial genome evolution through the gain and loss of genes (Kresse et al. 2003). In some cases, inversions are the only source of rearrangement in bacteria (Romling et al. 1997).

Although the control processes for inversions are not completely certain, a homologous recombination pathway is speculated to be involved in causing inversions (Cui et al. 2012; Sekulovic et al. 2018). Other factors such as changes in the environment (Zieg et al. 1978; Hill and Gray

1988; Gally et al. 1993; Serkin and Seifert 2000; Rentschler et al. 2013; Blomfield 2015) or the growth phase of bacteria (Zieg et al. 1978), can additionally induce inversions and cause them to revert to a non-inverted state (Zieg et al. 1978).

The sequence composition and location of inversions play a role in their likelihood. Bacterial genomes have notable "hotspots" where inversions repeatedly occur (Zieg et al. 1978; Schmid and Roth 1983; Mahan and Roth 1988; Segall et al. 1988; Segall and Roth 1989; Mahan and Roth 1991; Alm et al. 1999; Glaser et al. 2002; Sibley and Raleigh 2004; Raeside et al. 2014; Sekulovic et al. 2018). For example, some locations of the *Salmonella* genome are universally permissive within the species, and invert frequently (Segall et al. 1988). The inversion permissive and non-permissive intervals in *Salmonella* are not randomly distributed, but show a regional distribution influenced by chromosomal position, rather than adjacent sequence composition (Segall et al. 1988).

Other studies however, did find some association with nearby sequence composition and the frequency of inversions. Areas of bacterial genomes with repetitive regions (Naseeb et al. 2016), pathogenicity islands (Furuta et al. 2011), mobile elements (Furuta et al. 2011), Insertion Sequences (ISs) (Schneider and Lenski 2004; Furuta et al. 2011), or duplicated regions (Furuta et al. 2011; Cui et al. 2012) can increase the frequency of inversions. For example, homologous recombination between IS elements can cause large inversions or deletions to occur (Reif and Saedler 1975; Louarn et al. 1985; Schneider et al. 2000).

Inversions can have a number of effects on a variety of molecular traits. Inversions can impact gene gain and loss (Furuta et al. 2011), gene orientation (Huynen et al. 2001), and consequently gene expression (Zieg et al. 1977; Zieg et al. 1978; Sekulovic et al. 2018; Li et al. 2019). Inversions can impact the conservation of a gene or how genes are co-regulated depending on their orientation (Huynen et al. 2001). Borst and Greaves (1987), give an overview of some predominant examples of how inversions and rearrangements can influence gene expression in a number of organisms ranging from bacteria to chicken.

One interesting role that inversions can play is acting as a "control switch" turning on/off different characteristics or states (Borst and Greaves 1987). Some examples include: antibiotic resistance (Cui et al. 2012), presence/absenct of a flagella in *Salmonella* (Zieg et al. 1977; Johnson and Simon 1985; Li et al. 2019), switching the mating type of *Saccharomyces cerevisiae* (Hicks and Herskowitz 1976; Herskowitz and Oshima 1981), and changing the surface coat composition of *Trypanosoma brucei* to evade the host immune defence (Vickerman 1978; Lamont et al. 1986). Some inversions have the ability to revert or reverse (Hill and Gray 1988; Louarn et al. 1985; Cui et al. 2012), where maintaining an inverted or reverted state is organized and controlled (Cui et al. 2012; Sekulovic et al. 2018).

Another example of inversion "control switches" is found in expression, where inversions provide a way for bacteria to alter their gene expression (Zieg et al. 1977; Zieg et al. 1978;

Sekulovic et al. 2018; Li et al. 2019). The activation of gene expression is usually linked to inversions moving genes closer to promoters or enhancers (Borst and Greaves 1987; Cerdeño-Tárraga et al. 2005). Conversely, the removal of a promoter due to an inversion can cause inactivation of gene expression (Borst and Greaves 1987). This can influence the expression of a specific gene (Cerdeño-Tárraga et al. 2005), or act in a non-specific manner, causing genes in areas close to the inversion to become differentially expressed (Wong and Wolfe 2005; Cerdeño-Tárraga et al. 2005; Naseeb et al. 2016; Sekulovic et al. 2018). However, this process depends on the organism and inversion, as there have been instances where inversions do not alter expression of nearby genes (Meadows et al. 2010).

Previous studies investigating the impacts inversions have on bacterial gene expression have largely focused on a single inversion (Zieg et al. 1978; Sekulovic et al. 2018), or the impact inversions have on the expression of a small number of genes (Zieg et al. 1977; Li et al. 2019). The few studies that have taken a whole genome approach to analyzing inversions and their impact on gene expression, have concluded that there is differential gene expression between inverted and non-inverted regions (Wong and Wolfe 2005; Alokam et al. 2002; Naseeb et al. 2016). Although, there appears to be no pattern of genes being consistently up- or down- regulated in inverted or non-inverted segments (Wong and Wolfe 2005; Alokam et al. 2002; Naseeb et al. 2016). Most of these whole genome analyses are focused on yeast (Wong and Wolfe 2005; Naseeb et al. 2016), with only one looking at genome wide inversions in bacteria (Alokam et al. 2002). Alokam et al. (2002) focus on comparing inversions between species (*Salmonella* and *E. coli*).

Here we aim to explore differences in gene expression due to inversions between closely related strains of *E. coli*. Although there has been work done to identify inversions between closely related strains of *Salmonella enterica* (Sun et al. 2012), to our knowledge, an in-depth analysis of gene expression and inversions between closely related strains of *E. coli* has not been investigated. Most inversions have an unknown affect on the genome (Raeside et al. 2014), including how they impact gene expression. In this analysis we identify inversions between four closely related strains of *E. coli* (K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922) and combine RNA-seq datasets from multiple previously published studies to examine the short- and long-range impact inversions have on gene expression. Within a few (8%) of the inversions identified, there is a significant difference in gene expression between the inverted and non-inverted sequences in the *E. coli* genome, where 75% of inverted genes had a 1.27-85.58 fold higher gene expression and 25% of genes had 1.3-100 fold lower gene expression. Inversions impact gene expression even between closely related strains of *E. coli* and could provide a mechanism for strains to diversify their genetic content through expression changes.

4.4 Materials and Methods

All custom scripts and commands used in this analysis can be found on `GitHub` at `https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/`.

4.4.1 Expression Data

Due to the limited number of appropriate control RNA-seq datasets, the following closely related strains of *E. coli* were chosen for this analysis: K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922. RNA-seq data for *E. coli* was downloaded from the Gene Expression Omnibus (GEO) (Barrett et al. 2012). The expression datasets for this analysis were only RNA-seq datasets for control data, where this was defined as the bacteria being grown in optimal growth conditions. Using strictly raw RNA-seq count expression data allows the normalization to be standardized across all datasets, making the datasets directly comparable. Due to these constraints on our data, we were only able to retrieve a total of nine gene expression datasets from GEO for this analysis. A complete list of gene expression datasets and accession numbers can be found in Supplementary Table S3.1.

Only genes that were present in each dataset were considered for this analysis. Pseudogenes and phage genes were excluded. Correlation of gene expression across datasets was assessed for each strain with multiple datasets. For a detailed protocol, see Supplementary files on `GitHub` at `https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/`.

4.4.2 Normalization

The raw counts from control populations for each dataset were used and normalized using the TMM method (Robinson and Oshlack 2010). Raw counts were normalized to Counts Per Million (CPM) in `R` using the `edgeR` package (Robinson et al. 2010). After normalization, any datasets that had multiple replicates were combined by finding the median CPM between replicates for each annotated gene. An average normalized gene expression value for each gene was calculated by averaging the normalized gene expression values from all datasets. The average normalized gene expression value for each gene per strain is what was used for the remainder of the analysis.

4.4.3 Sequence Data

Whole genomes of the four different strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113, and ATCC 25922, were downloaded from NCBI. These genomes correspond to the strains found in the gene expression data. Access date and accession numbers are given in Supplementary Table S3.2. Although *E. coli* contains small plasmids, the plasmids were not included in this analysis.

4.4.4 Identifying Inversions

Whole genome alignments of all *E. coli* strains were performed using the default `Parsnp` parameters (Treangen et al. 2014) which aligned the core regions of the *E. coli* genomes. The close phylogenetic relationships of these strains mitigates the amount of data that is considered not part of the core genome and therefore would be unused for this analysis. `Parsnp` is able to

identify local and large scale inversions between all taxa in this analysis. The alignment blocks specified by `Parsnp` are additionally genomic rearrangement aware, meaning that an alignment block can be present in varying genomic locations across each taxa (rearrangements). This allows for a particular alignment block to be rearranged or at a different genomic location than the same homologous sequence in another taxa. `Parsnp` defines these alignment blocks as minimally being similar in sequence between at least two of the taxa, but not necessarily between all of them. To validate the homologous regions identified by `Parsnp`, a reciprocal best hit analysis was performed using `DIAMOND` and a custom `Python` script.

All proteomes were downloaded from `UniProt` (May 4, 2020) and correspond to the reference strains used for the gene expression and sequence data (see Supplementary Tables S3.1 and S3.2). In some cases, the proteomes for the particular strains of *E. coli* used in this analysis were labelled as redundant to proteomes from other strains on `UniProt`. These redundant proteomes were treated as the same for both strains and a custom `Python` script was used to determine corresponding homologous gene names between the genomes of these redundant proteomes. A complete list of proteomes used for each strain can be found in Supplementary Table S3.3. A pair-wise alignment of each proteome was performed using `DIAMOND` (`-more-sensitive`). In the case of identical e-value scores for multiple reciprocal hits, the hit with genes that were closest in synteny were chosen as the reciprocal best hit. Various `BLAST` and `DIAMOND` parameters were tested to determine the optimal parameter selection for this analysis. A complete list of parameters tested can be found in the Supplemental Table S3.4. We found that all parameter combinations tested performed relatively equally so we decided to use the parameters that resulted in the maximum number of reciprocal best hits (`DIAMOND -more-sensitive`). Each homologous set of genes identified by `Parsnp` was compared to the `DIAMOND` reciprocal best hits. In instances where any reciprocal best hit between two taxa did not match the homologous genes specified by the `Parsnp` alignment, this gene (and all it's homologs in the other taxa) was removed from subsequent analyses.

4.4.5 Inversions and Gene Expression Correlation

To determine if there is a correlation between inverted regions of the *E. coli* genomes and differences in gene expression, various Wilcoxon signed-rank tests were performed in `R` (R Development Core Team 2014). This test was done to determine if the mean gene expression differs between all inverted alignment blocks and non-inverted alignment blocks. A Wilcoxon signed-rank test was done within each inverted block to determine if the mean gene expression differs between inverted and non-inverted sequences within an alignment block.

To explore differences in gene expression variation between various groups of inverted and non-inverted alignment blocks, we used the `R` package `cvequality` (Marwick and Krishnamoorthy 2019). This tests for significant differences in coefficient of variation in gene expression between various groups of inverted and non-inverted alignment blocks (Feltz and Miller 1996;

Krishnamoorthy and Lee 2014). The Feltz and Miller (1996) asymptotic test and Krishnamoorthy and Lee (2014) M-SLRT were used to test for equality of coefficient of variation in gene expression between the following groups: all inverted and non-inverted alignment blocks, all inverted and non-inverted sequences within the *E. coli* ATCC 25922 genome, and significant inverted alignment blocks (with a significant difference in gene expression between the inverted and non-inverted sequences) and non-significant inverted alignment blocks (Supplementary Table S3.5).

4.4.6 Inversions and Distance From the Origin of Replication

To determine if there is a spatial distribution of inversions and distance from the origin of replication, multiple logistic regressions were performed. The genomic position was determined to be the midpoint genomic location for each alignment block while accounting for distance from the origin of replication and bidirectional replication (using the same methods in Lato and Golding 2020a; Lato and Golding 2020b). A logistic regression was performed to look at the presence or absence of an inversion and the genomic position of that inversion in each *E. coli* genome. Similarly, the placement of inverted alignment blocks with a significant difference in gene expression (via a Wilcoxon signed-rank test) between inverted and non-inverted sequences within that alignment block, was investigated using a logistic regression. Additional logistic regressions were performed on each strain of *E. coli* to determine if there is a correlation between individual inverted sequences in each strain and distance from the origin of replication.

4.4.7 Nucleoid Associated Protein Binding

Some nucleoid associated proteins have been associated with regulating transcription within the *E. coli* genome (Johansson et al. 2000; Kelly et al. 2004; Paul et al. 2004; Kahramanoglou et al. 2011; Scholz et al. 2019) The Histone-like Nucleoid-Structuring (H-NS) protein maintains and controls chromosome compaction and structure (Grainger et al. 2006), while also globally regulating transcription (Johansson et al. 2000; Kahramanoglou et al. 2011). H-NS has the ability to repress the transcription of non-essential genes (Browning et al. 2000; Hommais et al. 2001; Dorman 2004; Fang and Rimsky 2008; Dillon and Dorman 2010; Ali et al. 2012; Singh et al. 2016), playing an important role in silencing genes recently acquired via Horizontal Gene Transfer (HGT) (Dorman 2004; Oshima et al. 2006; Dorman 2007; Ali et al. 2014; Higashi et al. 2016). H-NS binds to hundreds of targets across the *E. coli* genome (Grainger et al. 2006; Oshima et al. 2006), so we were interested to see if there was any correlation between the inversions we identified and H-NS binding sites.

Datasets containing H-NS binding information were downloaded from the following papers: Grainger et al. (2006), Oshima et al. (2006), Lang et al. (2007), Ueda et al. (2013), and Higashi et al. (2016). Information about the predicted genomic regions and/or genes bound by H-NS was extracted using custom `R` and `Python` scripts. A Pearson correlation was used to determine if there was a correlation between the binding sites of the H-NS protein and inverted regions. A

match between an H-NS binding site and alignment block that contained at least one inverted sequence was determined by any partial or complete overlap of the entire H-NS binding site and the inverted alignment block. This H-NS analysis was performed on each dataset separately (Grainger et al. 2006; Oshima et al. 2006; Lang et al. 2007; Ueda et al. 2013; Higashi et al. 2016). The Higashi et al. (2016) dataset had multiple criteria for what was considered an H-NS binding site for both the coding and non-coding regions of the *E. coli* K-12 W3110 genome. The H-NS correlation analysis described here was performed on each criteria separately.

Similar to H-NS, the Factor for inversion stimulation (Fis) protein is a nucleoid associated protein that is involved in regulating expression (Kelly et al. 2004; Paul et al. 2004; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019) and a number of genomic architecture properties (Kahmann et al. 1985; Johnson et al. 1986; Thompson et al. 1987; Haffter and Bickle 1987; Ball and Johnson 1991; Messer et al. 1991; Filutowicz et al. 1992; Wold et al. 1996; Wu et al. 1996; Schneider et al. 1997; Schneider et al. 2001; Travers et al. 2001; Ryan et al. 2004; Dhar et al. 2009; Tsai et al. 2019; Dages et al. 2020). There is evidence that the Fis protein facilitates the activation of transcription through close interaction with promoters and Ribonucleic Acid Polymerase (RNAP) or by altering local genome architecture (Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019), and changes in genome-wide transcription via external environment changes can trigger changes in the binding of Fis at particular operons and genes (Grainger et al. 2006). Datasets containing Fis binding information were downloaded from Grainger et al. (2006). Information about the binding regions of Fis was extracted using custom `R` and `Python` scripts. A Pearson correlation was used to determine if there was a correlation between the binding sites of the Fis protein and inverted regions. A match between an Fis binding site and alignment block that contained at least one inverted sequence was determined by any partial or complete overlap of the entire Fis binding site and the inverted alignment block.

4.5 Results

4.5.1 Identifying Inversions

Using a combination of the `Parsnp` alignment and our own more stringent restrictions on homologous regions (see Methods), we identified a total of 555 alignment blocks. Of these blocks, 68.29% had at least one sequence that was inverted. The inverted alignment blocks ranged from 96bp - 38570bp with an average of 11270bp. We refer to alignment blocks with at least one inverted sequence as inverted alignment blocks. As mentioned previously, `Parsnp` allows alignment blocks to be present at varying genomic locations in each taxa. Therefore, some of the inversions we have identified are in different genomic locations in each taxa. A summary of the various genomic locations of these inverted alignment blocks (which can include rearrangements) is visualized in Figure 4.1. The majority of the inversions identified (99.5%) involved the *E. coli*

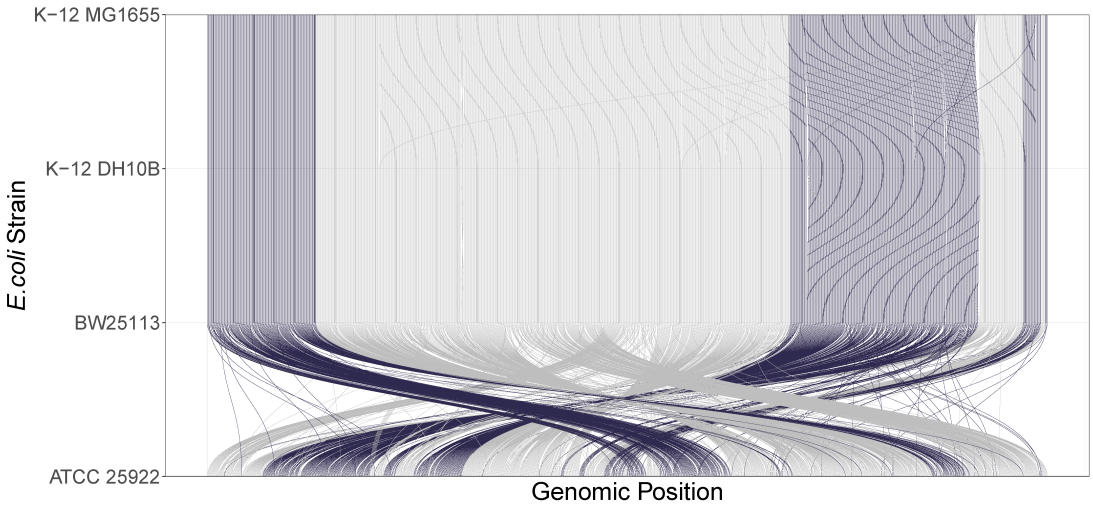


Figure 4.1: Visualization of rearrangements and inversions between all *E. coli* strains. The genome of each strain is represented as horizontal lines. Homologous alignment blocks in each taxa are connected with a winding vertical line. Alignment blocks can be found at varying genomic positions in each taxa. Inverted alignment blocks are coloured in dark purple and non-inverted alignment blocks (which may include rearrangements) are coloured in light grey.

ATCC 25922 strain being inverted relative to the other *E. coli* strains. The remaining few inverted alignment blocks (0.5%) had a pattern where both *E. coli* ATCC 25922 and K-12 DH10B strains were inverted relative to the *E. coli* K-12 MG1655 and BW25113 strains.

4.5.2 Inversions and Gene Expression Correlation

Wilcoxon signed-rank tests were performed to determine if there was a correlation between gene expression in all genes and inverted alignment blocks and individual sequences that were inverted. The results from these tests are summarized in Table 4.1. It should be noted that the sample size (number of genes) is large (~ 4500), which results in large W-statistics (coefficient estimates) for all Wilcoxon signed-rank tests in this analysis. We found that both correlations (W-statistics) are significant (Table 4.1). This indicates that there is a significant difference in gene expression between inverted alignment blocks (mean = 211 CPM, median = 50 CPM), and non-inverted alignment blocks (mean = 207 CPM, median = 44 CPM), and a significant difference in gene expression between individual inverted sequences and non-inverted sequences.

Taking a closer look at how gene expression varied within each inverted alignment block, 91.78% of inverted alignment had no significant difference in gene expression between inverted and non-inverted sequences. The remaining 8.22% of inverted alignment blocks had a significant 1.27-100 fold change in gene expression between the individual inverted and non-inverted sequences within each block. These inverted alignment blocks with significant differences in gene expression between the individual inverted and non-inverted sequences within each block will be

Datasets:	W-statistic
Inverted Blocks	15682115**
Inverted Sequences	11782352***

Table 4.1: Correlation coefficients for Wilcoxon signed-rank test on various datasets to determine the correlation between an inversion and difference in normalized gene expression. The "Inverted Blocks" dataset represents alignment blocks that have at least one taxa with an inverted sequence. The "Inverted Sequences" dataset represents all individual sequences from all alignment blocks that were inverted. The correlation between both datasets was computed using a Wilcoxon signed-rank test. All results are marked with significance codes as followed: p: < 0.001 = '***' and $0.001 < 0.01$ = '**'.

referred to as significant inverted alignment blocks. 75% of significant inverted alignment blocks had increased gene expression (1.27-85.58 fold change) within the individual inverted sequences in each block, and 25% had decreased expression (1.3-100 fold change). A visualization of average gene expression values for inverted and non-inverted sequences within each tested inverted alignment block can be found in Figure 4.2.

To examine if the coefficient of variation in gene expression differs between different groups of inverted and non-inverted alignment blocks, we performed Feltz and Miller (1996) asymptotic, and Krishnamoorthy and Lee (2014) Modified Signed Likelihood Ratio Tests (M-SLRT) tests on each group. The groups are as follows: all inverted and non-inverted alignment blocks, all inverted and non-inverted sequences within the *E. coli* ATCC 25922 genome, and significant inverted alignment blocks (which had a significant difference in gene expression between the inverted and non-inverted sequences) and non-significant inverted alignment blocks. There is a significant difference in the coefficient of variation in gene expression between significant inverted alignment blocks and non-significant inverted alignment blocks (Supplementary Table S3.5). We did not detect significant difference in the coefficient of variation in gene expression between inverted alignment blocks and non-inverted alignment blocks both overall and within the *E. coli* ATCC 25922 strain (Supplementary Table S3.5).

We detected a significant difference in gene expression between genes in inverted and non-inverted regions of the *E. coli* ATCC 25922 genome (Wilcoxon signed-rank test: $W = 1037272$, p-value = 0.015). Gene expression in the inverted alignment regions (mean = 246 CPM, median = 62 CPM) was higher than the non-inverted alignment regions (mean = 213 CPM, median = 55 CPM). We did not detect a significant difference in gene expression between genes found in inverted and non-inverted regions of the genomes of *E. coli* K-12 MG1655, K-12 DH10B, and BW25113.

4.5.3 Inversions and Distance From the Origin of Replication

We did not find a significant correlation between distance from the origin of replication and significant inverted alignment blocks. Figure 4.2 summarizes the distribution of differences in average gene expression between individual inverted and non-inverted sequences within inverted blocks along the origin-terminus replication axis. To simplify this visualization, the genomic position along the origin-terminus replication axis of the inverted alignment blocks was determined by the midpoint of the *E. coli* K-12 MG1655 strain within each alignment block.

A logistic regression combining the genomic location of inversions on the origin-terminus replication axis for all *E. coli* strains, was estimated to be significantly positive (Coefficient Estimate = 2.18e-7, p-value < 0.001). This indicates that inversions are preferentially located near the origin of replication when combining the genomic location from all *E. coli* strains. Since alignment blocks have the potential to be both inverted and rearranged, we explored the location of inversions on a per strain basis using logistic regressions. These results are summarized in Table 4.2. We did not detect a significant coefficient estimate between the genomic placement of inverted *E. coli* K-12 DH10B sequences along the origin-terminus replication axis (Table 4.2). This suggests that inverted sequences within *E. coli* K-12 DH10B do not have a preferential location along the origin-terminus replication axis. The coefficient estimate looking at the correlation between distance from the origin of replication and inverted sequences in *E. coli* ATCC 25922 was significantly negative (Table 4.2). This indicates that inverted sequences with the *E. coli* ATCC 25922 genome are concentrated near the origin of replication.

Strain	Coefficient Estimate
E. coli K-12 DH10B	NS
E. coli ATCC 25922	-1.90×10^{-7}***

Table 4.2: Logistic regression between inverted sequences within each strain and distance from the origin of replication for each strain. The *E. coli* strains K-12 MG1655 and BW25113 did not have any inversions identified within their sequences and therefore were not considered. All results are marked with significance codes $p < 0.001$ = '***' and > 0.05 ='NS'.

4.5.4 Nucleoid Associated Protein Binding

A visualization of the genomic distribution of the significant and non-significant inverted alignment blocks identified in this analysis and overlapping predicted H-NS binding sites is found in Figure 4.3. We therefore performed a Pearson Correlation test on each dataset separately (Table 4.3). We found a significant positive correlation between inverted alignment blocks and predicted H-NS binding sites for all binding criteria in the datasets from Oshima et al. (2006), Lang et al. (2007), and Higashi et al. (2016) (Table 4.3). We did not detect a significant correlation between significant inverted alignment blocks and predicted H-NS binding sites for any dataset (Table 4.3).

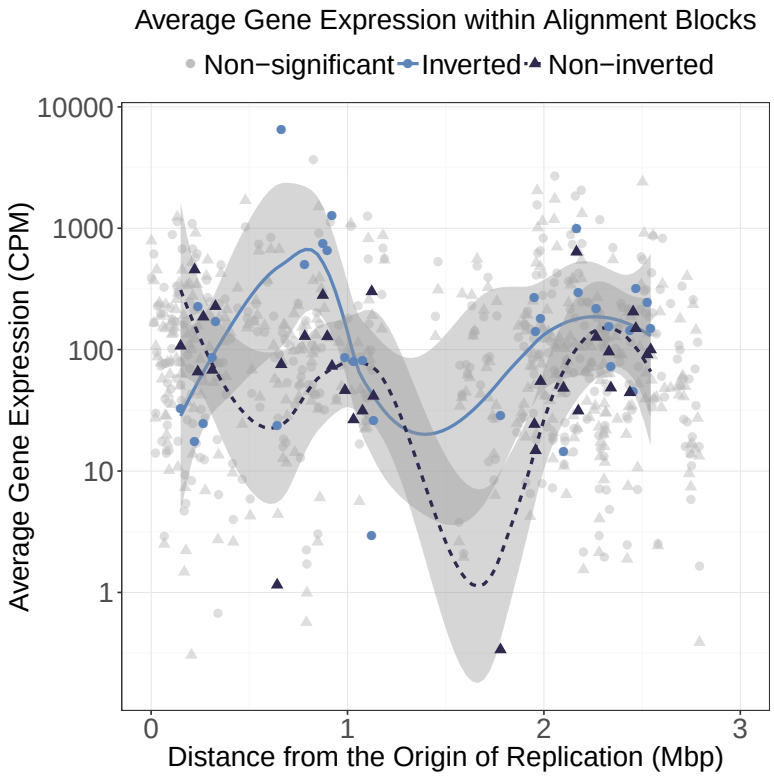


FIGURE 4.2: Visualization of the difference in gene expression between inverted and non-inverted sequences within alignment blocks. Each alignment block represents homologous sequences between the *E. coli* strains (Supplementary Table S3.2). *E. coli* K-12 MG1655 was used as the reference genome for the midpoint genomic position for each alignment block. Each alignment block has one point on the graph to represent the average expression value in **C**ounts **P**er **M**illion (**CPM**) for all inverted (circles) and non-inverted (triangles) sequences within the block. Blocks that had a significant difference in gene expression (using a Wilcoxon sign-ranked test, see Materials and Methods) have the inverted and non-inverted gene expression averages highlighted in blue circles and purple triangles respectively. All blocks that did not have a significant difference, have expression values coloured in light grey. A smoothing line (`span = 0.5, method = 'loess'`) was added to link the average gene expression values for the inverted (blue solid) and non-inverted (purple dashed) sequences within block that had a significant difference in gene expression (using a Wilcoxon sign-ranked test, see Materials and Methods).

H-NS Binding Study	All Inversions and H-NS Binding	Significant Inversions and H-NS Binding	Total Number of H-NS Binding Sites Within All Alignment Blocks
Grainger et al. (2006)	NS	NS	53
Ueda et al. (2013)	NS	NS	275
Higashi et al. (2016)			
criteria A	0.0467*	NS	371
criteria B	0.0540**	NS	343
criteria C	0.0540**	NS	343
criteria D	0.0540**	NS	343
criteria E	0.0544**	NS	340
criteria F	0.0544**	NS	340
Lang et al. (2007)	0.0574**	NS	115
Oshima et al. (2006)	0.0390*	NS	664

Table 4.3: Pearson correlation between H-NS binding sites and inverted regions of the *E. coli* K-12 MG1655 genome. A genomic region was considered inverted if this sequence was inverted in any of the following four strains of *E. coli*: K-12 MG1655,K-12 DH10B, BW25113, and ATCC. The genomic positions of these inversions in *E. coli* K-12 MG1655 was used for reference. The binding sites for the H-NS protein are in the genomic coordinates of *E. coli* K-12 MG1655, chosen as a reference. The second column "All Inversions and H-NS Binding" represents the correlation coefficient between inverted regions and H-NS binding sites. The third column "Significant Inversions and H-NS Binding" represents the correlation coefficient between inverted regions with significant differences in normalized gene expression between inverted and non-inverted taxa (via a Wilcoxon signed-rank test) and H-NS binding sites. More information on the Higashi et al. (2016) binding criteria can be found in the Supplementary Material. All results are marked with significance codes as followed: p: $0.001 < 0.01 =$ '**', $0.01 < 0.05 =$ '*', $> 0.05 =$ 'NS'. The sample size for the second column correlation tests ("All Inversions and H-NS Binding") was 2908. The sample size for the third column correlation tests ("Significant Inversions and H-NS Binding") was 2023.

We observed a significant difference in expression between H-NS bound and H-NS un-bound genes (Wilcoxon signed-rank test: W= 13157398, p-value < 0.001), where predicted H-NS bound genes had higher average expression than predicted H-NS un-bound genes. Additionally, a significant difference in gene expression between inversions with predicted H-NS binding and non-inversions with predicted H-NS binding was detected (Wilcoxon signed-rank test: W= 6890889.5, p-value < 0.001), where inversions with predicted H-NS binding sites had higher average expression than non-inversions with predicted H-NS binding sites. We did not detect a significant difference in gene expression between significant inverted alignment blocks with predicted H-NS binding sites and non-significant inverted alignment blocks with predicted H-NS binding.

A visualization of the genomic distribution of the significant and non-significant inverted alignment blocks identified in this analysis and overlapping predicted Fis binding sites is found in Figure 4.4. We did not detect a significant correlation between inverted alignment blocks and predicted Fis binding (Table 4.4). We found a significant positive correlation between significant

Fis Binding Study	All Inversions and Fis Binding	Significant Inversions and Fis Binding	Total Number of Fis Binding Sites Within All Alignment Blocks
Grainger et al. (2006)	NS	0.068**	205

Table 4.4: Pearson correlation between Fis binding sites and inverted regions of the *E. coli* K-12 MG1655 genome. A genomic region was considered inverted if this sequence was inverted in any of the following four strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113, and ATCC. The genomic positions of these inversions in *E. coli* K-12 MG1655 was used for reference. The binding sites for the Fis protein are in the genomic coordinates of *E. coli* K-12 MG1655, chosen as a reference. The second column "All Inversions and Fis Binding" represents the correlation coefficient between inverted regions and Fis binding sites. The third column "Significant Inversions and Fis Binding" represents the correlation coefficient between inverted regions with significant differences in normalized gene expression between inverted and non-inverted taxa (via a Wilcoxon signed-rank test) and Fis binding sites. All results are marked with the following significance code: p: $0.001 < 0.01 =$ '**'. The sample size for the second column correlation tests ("All Inversions and Fis Binding") was 2908. The sample size for the third column correlation tests ("Significant Inversions and Fis Binding") was 2023.

inverted alignment blocks and predicted Fis binding sites (Table 4.4).

We explored the potential impact that Fis binding had on differences in expression. We observed a significant difference in expression between Fis bound and Fis un-bound genes (Wilcoxon signed-rank test: W= 15945296, p-value < 0.001), where predicted Fis bound genes had higher average expression than predicted Fis un-bound genes. Additionally, a significant difference in gene expression between inversions with predicted Fis binding and non-inversions with predicted Fis binding was detected (Wilcoxon signed-rank test: W= 6890889.5, p-value < 0.001), where inversions with predicted Fis binding sites had higher average expression than non-inversions with predicted Fis binding sites. We observed a significant difference in gene expression between significant inverted alignment blocks with predicted Fis binding sites and non-significant inverted alignment blocks with predicted Fis binding (Wilcoxon signed-rank test: W= 296280, p-value $<$ 0.001).

4.6 Discussion

We identified 379 inversions between four strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922, and combined information from multiple previously published RNA-seq datasets to examine the short- and long-range genomic impacts these inversions have on gene expression. Within 92% of the inverted regions identified, there was no significant difference in gene expression between inverted and non-inverted sequences in that region. However, inverted sequences had a 1.27-85.58 fold higher gene expression in 75% of the significant inverted alignment blocks (8% of the total inverted alignment blocks). Most (99.5%) identified inversions

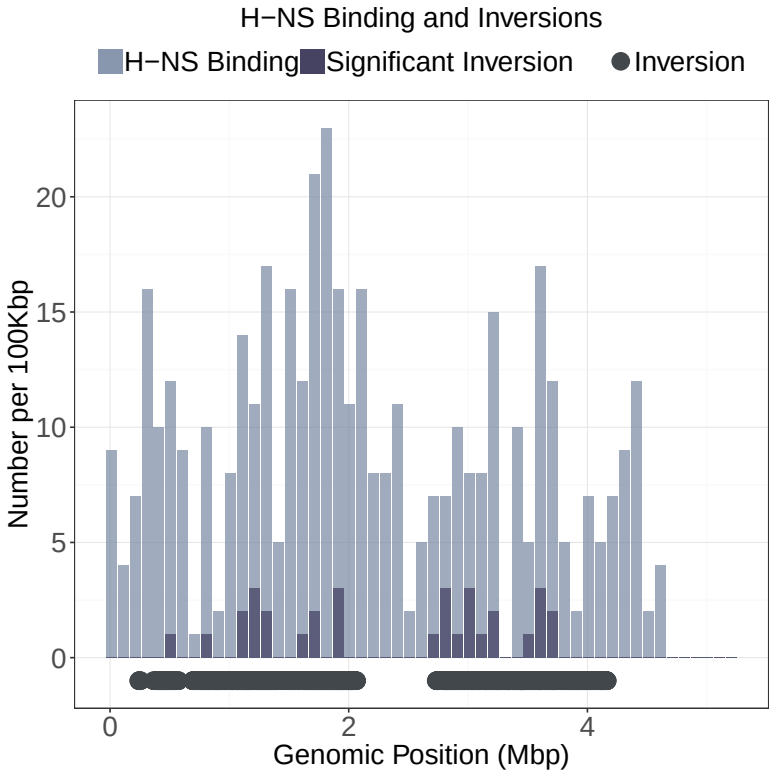


Figure 4.3: Histogram of the distribution of **H**istone-like **N**ucleoid-**S**tructuring (H-NS) binding sites and significant inversions. Visualization of the genomic locations of all tested inversion alignment blocks (dark grey filled circles on the x-axis, see Methods for details on test) identified between four strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113, and ATCC 25922. The data is plotted on the genome of *E. coli* K-12 MG1655 which is used as a reference. Each inversion alignment block has a single genomic location chosen to be the midpoint of the tested inverted region calculated to be the genomic distance from the *E. coli* K-12 MG1655 origin of replication. The total number of H-NS protein binding sites per 100Kbp in the *E. coli* K-12 MG1655 (light blue histogram bars). Data for the H-NS binding information is from Higashi et al. (2016) datasets and all H-NS binding sites identified in this dataset is shown. The total number of significant inversion alignment blocks (which had a significant difference in gene expression between the inverted and non-inverted sequences within the block using a Wilcoxon sign-ranked test, see Materials and Methods), are indicated by the dark purple histogram bars.

occurred in the *E. coli* ATCC 25922 strain relative to the other strains. Within the *E. coli* ATCC 25922 genome, the identified inversions significantly decreased with increasing distance from the origin of replication. A significant positive correlation between the identified inversions and H-NS binding sites and a significant positive correlation between the identified significant inversions and Fis binding sites was observed. We have provided an overview on the genomic impacts inversions can have on gene expression between closely related strains of *E. coli*.

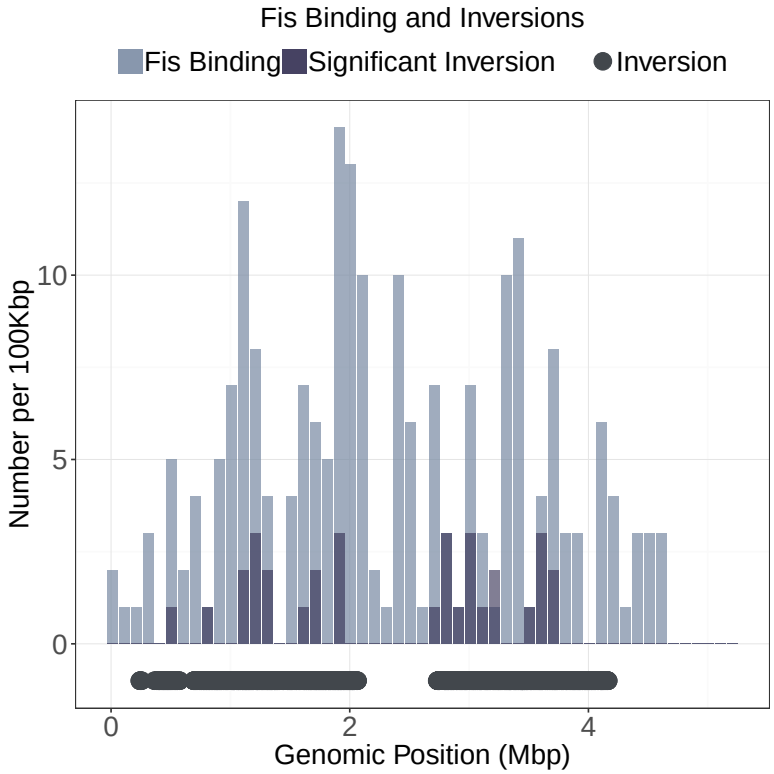


Figure 4.4: Histogram of the distribution of Fis binding sites and significant inversions. Visualization of the genomic locations of all tested inversion alignment blocks (dark grey filled circles on the x-axis, see Methods for details on test) identified between four strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113, and ATCC 25922. The data is plotted on the genome of *E. coli* K-12 MG1655 which is used as a reference. Each inversion alignment block has a single genomic location chosen to be the midpoint of the tested inverted region calculated to be the genomic distance from the *E. coli* K-12 MG1655 origin of replication. The total number of Fis protein binding sites per 100Kbp in the *E. coli* K-12 MG1655 (light blue histogram bars). Data for the Fis binding information is from Grainger et al. (2006) datasets and all Fis binding sites identified in this dataset is shown. The total number of significant inversion alignment blocks (which had a significant difference in gene expression between the inverted and non-inverted sequences within the block using a Wilcoxon sign-ranked test, see Materials and Methods), are indicated by the dark purple histogram bars.

Due to the limited amount of appropriate gene expression data available, we were unable to include more than four strains of *E. coli* in this analysis. The small sample size should be taken into consideration when interpreting the results of this analysis. Future studies increasing the amount of closely related *E. coli* genomes used, would provide insight into how widely observed the results presented in this work are.

4.6.1 Inversions and Gene Expression

The coefficient of variation in gene expression does not significantly differ between inverted and non-inverted alignment blocks (Supplementary Table S3.5). We did detect a significant difference in coefficient of variation in gene expression between significant inverted alignment blocks and non-significant inverted alignment blocks (Supplementary Table S3.5). However, this may be due to the initially present difference in gene expression between the significant and non-significant inverted alignment blocks. Therefore, the identified inversions overall do not disrupt variation in gene expression. We believe the inversion events observed had the greatest impact on gene expression in the *E. coli* ATCC 25922 strain, and may facilitate adaption strategies of this genome.

As mentioned previously, inversions offer a way for bacteria to alter their gene expression (Zieg et al. 1977; Zieg et al. 1978; Sekulovic et al. 2018; Li et al. 2019), through a "control switch" impacting specific genes (Cerdeño-Tárraga et al. 2005), or on a long-range scale altering expression in nearby genes (Wong and Wolfe 2005; Cerdeño-Tárraga et al. 2005; Naseeb et al. 2016; Sekulovic et al. 2018). It is speculated that inversions can help provide genetic diversity (Hughes et al. 2000; Belda et al. 2005) by promoting recombination (Segall et al. 1988), leading to the introduction of genes with novel functions (Korneev and O'Shea 2002) and assisting in speciation and adaptation (Kresse et al. 2003). Perhaps the significant differences in gene expression in the *E. coli* ATCC 25922 strain have allowed for this strain to adapt to different environmental niches compared to the *E. coli* K-12 MG1655, K-12 DH10B, and BW25113 strains. The ability for the inversions identified in this study to provide sufficient genetic variation to allow the *E. coli* ATCC 25922 strain to occupy unique environments compared to the other strains warrants further investigation.

4.6.2 Inversions and Distance From the Origin of Replication

We observed a significant positive trend between distance from the origin of replication and inverted alignment blocks. In the inverted alignment blocks, the *E. coli* strains K-12 MG1655, K-12 DB10B and BW25113 all have similar genomic positions, where as the ATCC 25922 strain is rearranged. We believe that the genomic positions of the inversions in the *E. coli* strains K-12 MG1655, K-12 DB10B and BW25113 are driving the positive trend observed between distance from the origin of replication and inverted alignment blocks.

The distribution of inversions along the *E. coli* ATCC 25922 genome (concentrated near the origin of replication; Table 4.2) suggest that the broad organization of genetic material along the *E. coli* genome influences inversions, among other molecular traits. For example, bacterial genomes are made up of interspersed permissive and non-permissive inversion zones (Segall et al. 1988; Guijo et al. 2001). These are constrained by selective forces acting to preserve the macro- and micro-domains of the folded chromosomes (Segall et al. 1988; Raeside et al. 2014; Naseeb et al. 2016). In a study by Hendrickson et al. (2018), six times as much inverted

DNA was found near the origin of replication rather than the terminus. This may be due to constraints on the distribution of Architecture IMparting Sequences (AIMS) (Hendrickson et al. 2018). AIMS act during DNA segregation and their improper distribution can disrupt AIMS-based genome architecture which could have detrimental consequences to the well being of the bacteria (Hendrickson et al. 2018). Hendrickson et al. (2018) conclude that the intragenomic rearrangements (likely causing inverted DNA), are counter-selected because they disrupt AIMS distribution. In *E. coli* ATCC 25922 most of the inversions identified here were found closer to the origin of replication rather than the terminus, which could be to avoid disturbing vital AIMS distribution. Future studies assessing the location of AIMS and other genomic architecture sequences along the *E. coli* ATCC 25922 genome, will assist in exploring this theory.

4.6.3 Nucleoid Associated Protein Binding

H-NS acts as a global transcription regulator (Johansson et al. 2000; Kahramanoglou et al. 2011) repressing the transcription of non-essential genes (Browning et al. 2000; Hommais et al. 2001; Dorman 2004; Fang and Rimsky 2008; Dillon and Dorman 2010; Ali et al. 2012; Singh et al. 2016) and plays an important role in silencing genes recently acquired via HGT (Dorman 2004; Oshima et al. 2006; Dorman 2007; Ali et al. 2014; Higashi et al. 2016). We determined that there is a significant positive correlation between all identified inverted alignment blocks and predicted H-NS binding sites, suggesting that H-NS plays a role in regulating expression in inverted regions. However, we did not detect a significant correlation between recent HGT genes in *E. coli* K-12 MG1655 and inverted alignment blocks, suggesting that horizontally transferred genes may not be preferentially located in inverted regions. Most of the identified inversions, and corresponding H-NS binding sites, occured in *E. coli* ATCC 25922 strain, suggesting that H-NS may be regulating expression in the inverted regions of this strain. There is little information on the exact H-NS binding sites within the *E. coli* ATCC 25922 strain, and determining these genomic locations would clarify if the strain specific inversions identified are frequently bound by H-NS. The repression of genes via H-NS binding is a complicated process (Wade and Grainger 2014; Singh et al. 2014; Chintakayala et al. 2013; Singh and Grainger 2013; Grainger 2016), and there may be other gene expression regulatory forces (Martinez-Antonio et al. 2009; Meyer et al. 2018), such as maintaining genomic architecture, acting on the identified inverted alignments in this work.

Opposing H-NS, there is evidence that the Fis protein activates transcription (Scholz et al. 2019). We observed a significant positive correlation between Fis binding sites and significant inverted alignment blocks. Interestingly, these significant inverted alignment blocks had higher average expression than non-inverted alignment blocks and higher average expression in inverted sequences within these blocks. Fis may be playing a role in activating expression within these significant inverted regions of the *E. coli* genome, potentially causing this increase in expression in these regions relative to other regions of the genome. However, Fis was found by Warnecke et al. (2012) to only be found at detectable levels during mid- and late exponential phase.

Exploring differences in expression between the identified inverted and non-inverted regions at various growth phases in *E. coli* will assist in understanding how nucleoid associated proteins such as Fis can impact expression in a temporal manner. Future studies exploring the roles that other nucleoid associated proteins play in altering expression within inverted regions can provide prospective on the interplay between genomic architecture and genomic reorganization through inversions.

Here, we observed that inversions can alter gene expression, demonstrating that this type of genomic reorganization has wide ranging impacts on bacteria, from molecular (this work) to physical characteristics (Hicks and Herskowitz 1976; Zieg et al. 1977; Vickerman 1978; Herskowitz and Oshima 1981; Johnson and Simon 1985; Lamont et al. 1986; Borst and Greaves 1987; Li et al. 2019).

4.7 Conclusions

In this work we identified hundreds of naturally occurring inversions between four strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922. We combine information from multiple previously published RNA-seq datasets to examine the short- and long-range genomic impacts these inversions have on gene expression. Within most of the inverted regions identified, there was no significant difference in gene expression between inverted and non-inverted sequences. Within the inverted alignment blocks that did have a significant difference in gene expression between inverted and non-inverted sequences (8%), 75% had higher gene expression in inverted sequences. We did not detect a general disruption of gene expression (as measured by a significant difference in the coefficient of variation in expression) between inverted and non-inverted alignment blocks. In the *E. coli* ATCC 25922 strain, where most of the identified inversions were located, the number of inversions significantly decreases with increasing distance from the origin of replication. This could be to avoid disruption in the distribution of AIMS in the *E. coli* ATCC 25922 genome, and warrants further investigation. There was a significant difference in expression between inverted and non-inverted regions of the *E. coli* ATCC 25922 genome, where inverted regions had a higher average expression. These observed difference in gene expression could have created space for increased adaptation strategies in the *E. coli* ATCC 25922 strain.

The H-NS and Fis proteins are well known transcription regulators (Johansson et al. 2000; Kelly et al. 2004; Paul et al. 2004; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019), and the predicted binding sites of both are significantly positively correlated with identified inversions in this analysis. This suggests that nucleoid associated proteins may play a role in the regulation of gene expression in inverted regions.

We have provided an overview of the short- and long-range impacts inversions can have on gene expression between closely related strains of *E. coli*. Inversions can alter genomic content,

which can lead to changes in gene expression. Further classification and investigation of bacterial inversions will provide more information as to how inversions can facilitate bacterial evolution.

4.8 Supplementary Material

Supplementary Figures S3.1 - S3.3 and Tables S3.2 - S3.5 are available at GENOME online (https://cdnsciencepub.com/journal/gen). Further supplemental code and the most recent information are available on GitHub at https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/.

4.9 Acknowledgments

We thank Natural Sciences and Engineering Research Counsel for funding for this project (grant RGPIN-202-05733 to GBG).

Chapter 5

Conclusion

Daniella F. Lato

5.1 Thesis summary

The position of a gene within a bacterial genome has profound impacts on the molecular properties of that gene. Previous research implies that there are predictable spatial patterns when moving from the origin of replication to the terminus. Specifically, as distance from the origin of replication becomes larger, gene expression typically decreases (Schmid and Roth 1987; Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012; Block et al. 2012; Bryant et al. 2014; Garmendia et al. 2018; Lato and Golding 2020a) and substitution rate customarily increases (Cooper et al. 2010; Morrow and Cooper 2012). In addition to these spatial trends, bacterial genomes routinely go through genome reorganization via processes such as Horizontal Gene Transfer (HGT), rearrangements, and inversions (Ochman et al. 2000; Epstein et al. 2014). Little has been done to merge the knowledge of bacterial genome reorganization and spatial molecular trends. We provide here an analysis of the impact genomic reorganization has on spatial molecular trends in various bacterial genomes.

We first identified genomic rearrangements and inversions in *Escherichia coli*, *Bacillus subtilis*, *Streptomyces* and *Sinorhizobium meliloti*. We were able to broadly group the genome regions into homologous sections that had unique genomic positions in each taxa. We mapped each of these genomic positions and respective sequences onto the phylogeny of each bacteria to obtain the ancestral history of reorganization movement. This allowed us to gain a complete picture of the rearrangements. The total number of extant and ancestral substitutions were estimated for each bacteria and the results were surprising. We determined that the number of substitutions significantly changes with increasing distance from the origin of replication, however the correlation is small and inconsistent in sign. Some replicons had a significantly decreasing substitution trend (the chromosomes of *E. coli* and *S. meliloti*), while others showed the opposite significant trend (*B. subtilis*, *Streptomyces* and the secondary replicons of *S. meliloti*: pSymA and pSymB). We did not observe a predictable increase in substitutions with increasing distance from the origin of replication, and it appears as though accounting for rearrangements may have partly been the culprit. We decided to look into evolutionary rates (Non-synonymous Substitution Rate (dN) and Synonymous Substitution Rate (dS)) and their ratio ω, which can give an indication of the selective pressures acting on the genome. We examined dN, dS and ω across all genes in each bacteria and did not detect a significant correlation between those values and distance from the origin. Genes having higher or lower values of dN, dS and ω do not appear to be clustered at any one particular genomic position. Our study is not the first to find molecular traits deviate from the customary trend directions. There are a number of studies that were unable to confirm a positive linear correlation between distance from the origin of replication and mutation rates (Hudson et al. 2002; Ochman 2003; Martina et al. 2012; Juurik et al. 2012; Dettman et al. 2016; Dillon et al. 2018). There are a number of other intertwining factors that impact the mutation spectra of bacteria such as transcription, replication, and growth state (Hudson et al. 2002; Ochman 2003; Juurik et al. 2012). Our analysis illustrates the importance of connecting various

bacterial processes, such as genome reorganization, to molecular evolution studies. Neglecting to account for rearrangements and inversions in bacterial genomic studies can obscure the observed spatial correlations. The number of substitutions certainly varies with genomic position, but the directional commonality of this correlation remains unclear.

Our second analysis focused on the spatial trends of gene expression in the same bacteria (*E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti*). To date, the majority of studies that found gene expression to generally be higher near the origin of replication than the terminus, moved a small number of easily manipulable genes to varying predetermined genomic locations (Schmid and Roth 1987; Block et al. 2012; Bryant et al. 2014; Garmendia et al. 2018). These studies are crucial for examining individual changes in gene expression among genes as a result of a (synthetic) reorganization. However, a broader context of the overall genomic pattern of gene expression along the origin-terminus of replication axis is lacking. Specifically, there is no current analysis that examines the genomic patterns of gene expression combining data from existing bacterial RNA-seq datasets. Utilizing preexisting control gene expression data, we were able to analyze the expression of all genes across the genomes of *E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti*. This was done in the absence of genomic reorganization to obtain a general understanding of what gene expression patterns are present using multiple existing bacterial RNA-seq datasets. We observed that normalized gene expression on a genomic scale indeed decreases with increasing distance from the origin of replication in *E. coli*, *B. subtilis* and *Streptomyces*. We were unable to observe a significant correlation between distance from the origin of replication and gene expression in *S. meliloti*. Genes with higher expression are typically located near the origin while genes with lower expression are found near the terminus. We explored the functional category of the genes in this analysis using the Clusters of Orthologous Groups of proteins (COG) functional categories, and found that there was no category that was significantly concentrated near the origin or terminus of replication. It has been suggested that the leading strand is favoured for the location of highly expressed genes to allow faster DNA replication and lower transcriptional losses (Brewer 1988). We found no statistical evidence for the leading strand to have higher expression levels compared to the lagging strand in most of the bacterial replicons, and have concluded that this is likely not driving the results of decreased gene expression with increased distance from the origin of replication. Previous studies have determined that the main factor influencing the location of a gene on the leading or lagging strand is the essentiality of that particular gene, not expression (Rocha and Danchin 2003; Zheng et al. 2015). Once again, the genomic position of a gene has a profound impact on the properties of that gene, in this case expression.

Our third analysis focused on integrating genomic reorganization and bacterial gene expression. Bacteria utilize genomic reorganization, such as rearrangements and inversions, as a tool to create genetic diversity (Hughes et al. 2000; Belda et al. 2005) and assist in speciation, adaptation (Kresse et al. 2003; Rocha 2004a; Hanage 2016), and gene conversion (Hanage 2016). Inversions

are one particular type of genomic reorganization that can promote spontaneous genome rearrangements (Sun et al. 2012). There has been evidence that inversions are non-random and provide specific functions in bacterial genome evolution (Kresse et al. 2003). In some cases, inversions are the only source of rearrangement in bacteria (Romling et al. 1997). Inversions indirectly impact the conservation of a gene or how genes are co-regulated depending on their orientation and location (Huynen et al. 2001). Since inversions can promote recombination (Segall et al. 1988), there is a potential for inversions to promote the introduction of genes, potentially leading to the evolution of novel functions (Korneev and O'Shea 2002). These processes allow inversions to provide a way for bacteria to alter their gene expression (Zieg et al. 1977; Zieg et al. 1978; Sekulovic et al. 2018; Li et al. 2019) by "turning on/off" expression of particular genes (Cerdeño-Tárraga et al. 2005) or altering expression in a non-specific way by causing genes in areas close to the inversion to be differentially expressed (Wong and Wolfe 2005; Cerdeño-Tárraga et al. 2005; Naseeb et al. 2016; Sekulovic et al. 2018). Previous studies investigating the impacts inversions have on bacterial gene expression have largely focused on a single inversion (Zieg et al. 1978; Sekulovic et al. 2018), in a small number of genes (Zieg et al. 1977; Li et al. 2019), or focus on distantly related organisms (Alokam et al. 2002; Wong and Wolfe 2005; Naseeb et al. 2016). In Chapter 4, we explored differences in gene expression due to inversions between closely related strains of *E. coli*. In this analysis we identify inversions between four closely related strains of *E. coli* (K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922) and combine RNA-seq datasets from multiple previously published studies to examine the short- and long-range impact inversions have on gene expression. We determined that there is a significant difference in gene expression between inverted and non-inverted regions of the genome, however, the variation in expression does not significantly differ between inverted and non-inverted regions. Within a few of the inversions identified (8%), there is a significant difference in gene expression between the inverted and non-inverted sequences, 75% of inverted genes had increased gene expression (1.27-85.58 fold change), and 25% of inverted genes had decreased expression (1.3-100 fold change). The Histone-like Nucleoid-Structuring (H-NS) and Factor for inversion stimulation (Fis) proteins have been associated with genome wide gene expression regulation (Johansson et al. 2000; Kelly et al. 2004; Paul et al. 2004; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019). We observed a significant positive correlation between the identified inversions and H-NS binding sites. Additionally, we detected a significant positive correlation between significant inverted alignment blocks (where there was a significant difference in expression between the inverted and non-inverted sequences within the block) and Fis binding. These correlations suggest that genomic architecture may be playing a role in regulating expression within these inverted regions. Inversions impact gene expression even between closely related strains of *E. coli* and could provide a mechanism for strains to diversify their genetic content through controlled expression changes.

5.2 The Impact of Genomic Reorganization on Substitution in Bacterial Genomes

To date there has been a large body of work looking at how molecular trends such as gene expression (Couturier and Rocha 2006; Cooper et al. 2010; Morrow and Cooper 2012; Lato and Golding 2020a) and substitution rates (Sharp et al. 1989; Sharp et al. 2005; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012; Lato and Golding 2020b) vary with genomic position. The general consensus is that substitution rates are highest near the terminus of replication and relatively low near the origin (Sharp et al. 1989; Cooper et al. 2010; Flynn et al. 2010; Morrow and Cooper 2012). The majority of these studies used an average of 3 genomes per bacteria analyzed (Couturier and Rocha 2006; Flynn et al. 2010; Cooper et al. 2010; Morrow and Cooper 2012) and failed to analyze secondary replicons of multipartite genomes (Sharp et al. 2005; Couturier and Rocha 2006; Flynn et al. 2010). Additionally, frequent processes that cause genomic reorganization in bacteria, such as rearrangements and inversions, have not been incorporated into the analysis of spatial molecular trends such as gene expression and substitution rates. A complete picture involving common aspects of bacterial genome evolution, such as genomic reorganization, will provide a more accurate and in-depth representation of substitution rates along the origin-terminus of replication axis. In Chapter 2, we integrated historical rearrangement information into a spatial analysis of substitutions. We present a unique approach to the investigation of the location of substitutions along bacterial genomes, by accounting for local and large scale genomic rearrangements. This was done by utilizing ancestral reconstruction techniques of both sequences and genomic positions. This phylogenetic reconstruction analysis provides a thorough account of how genomic reorganization processes can impact the trends seen in substitutions with increasing distance from the origin of replication.

5.2.1 Data Quality and Genome Reorganization Challenges

In order to provide accurate ancestral reconstruction of both the genomic sequence and genomic position, it was crucial to only compare homologous sequences between complete bacterial genomes. We discuss the constraints on the number of sequences chosen in Chapter 2 in the Appendix A0.2. Using the genome alignment program presented in Chapter 2, (`progressiveMauve` (Darling et al. 2010)) it was difficult to align homologous sequences. The program `progressiveMauve` identifies conserved segments of sequence that seem to be internally vacant from any genome rearrangements, known as Locally Colinear Blocks (LCBs) (Darling et al. 2010). In our experience, when more divergent taxa are provided to `progressiveMauve`, the LCBs become increasingly unreliable. We have illustrated this circumstance in Figure S1.1 with the alignment of six divergent *Streptomyces* genomes (see Figure S1.1 caption for sequence information). Although these taxa are relatively closely related, `progressiveMauve` is unable to resolve clear LCBs. `progressiveMauve` identified a total of 521 rearrangements (including inversions), ranging from 107bp - 0.6Mbp in length, with an average length of 8598bp. Unfortunately,

the LCBs established in this particular alignment (Figure S1.1) did not compare homologous sequences, resulting in inaccurate substitution estimates.

We performed a number of additional whole genome alignments using our sequence alignment pipeline (Chapter 2) to determine how many sequences we could use in this analysis, while still obtaining accurate alignment information. Using the alignment pipeline outlined in Chapter 2, we increased the total number of genomes to 26 for *B. subtilis* and *E. coli* as a starting point to determine if it would be possible to increase the data set size. We chose 26 distantly related complete reference genomes for *B. subtilis* and *E. coli* (Tables S1.3 and S1.4). These were aligned with `progressiveMauve` and subsequently `MAFFT`, and trimmed using `trimAl` and our custom codon aware `Python` script (using the same methods as in Chapter 2). After our conservative alignment trimming methods (Chapter 2), we found that only 0.13% - 0.4% of the whole genome alignments in *E. coli* and *B. subtilis* were retained. In our experience with `progressiveMauve`, the more divergent the genomes are - even slightly more divergent in this case - the more inaccurate the specification of LCBs by `progressiveMauve`. These LCBs contain very poor alignments that do not align homologous genes. As a result, through our trimming methods, the majority of these alignments are classified as poor and discarded. This would become increasingly problematic when more genomes are added. Therefore, given our current methods and pipelines, we do not believe it is possible to increase the number of genomes substantially more than what we have presented. We believe that the benefits of the in-depth analysis and quality of data we have provided outweigh the inaccurate data that would be produced with more genomes.

The issue of aligning divergent genomes does not stop with the LCBs classification. `progressiveMauve`, like any other alignment program, requires increased computational time to correctly align divergent sequences. An increase in divergence, increases computational time. We observed this directly when testing how many sequences we were able to use in our analysis for Chapter 2. A randomized selection of 26 divergent *Streptomyces* genomes (Table S1.5) with unknown or unspecified strain identification were chosen for whole genome alignment with `progressiveMauve`. It took nearly twice as long to align 26 divergent *Streptomyces* genomes ($\sim$ 30 days), compared to the same number of more closely related *E. coli* genomes ($\sim$ 14 days). This imparted yet another disappointing constraint on the number of sequences we were able to utilize.

When accounting for genomic reorganization it is important to ensure the accurate identification of rearrangements and inversions, while also maintaining comparisons between homologous regions. We chose to analyze a total of 24 bacterial genomes from *E. coli* (6 genomes), *B. subtilis* (7 genomes), *Streptomyces* (5 genomes) and *S. meliloti* (6 genomes). Although our number of genomes per bacteria was relatively small (5-7), we maintained a high level of quality to our data, allowing for a deep analysis of the evolutionary history of rearrangements and substitutions along bacterial genomes.

5.2.2 Genomic Reorganization and Spatial Patterns

The dilemma of having a large enough data set while still maintaining accurate orthologous gene comparisons was not the only complication of this analysis. Accounting for the abundant genomic reorganization that occurs in bacteria, is a complicated matter. Rearrangements and inversions can cause orthologous genes to be present in different genomic locations in distinct taxa. Determining the present location of these genes is relatively trivial when comparing sequence alignments through programs such as `BLAST` (Altschul et al. 1990). However, to gain an accurate representation of the history of these rearrangements, an ancestral reconstruction must be performed. In Chapter 2, we modified the nucleotide and protein ancestral sequence reconstruction software `PAML` (Yang 1997), to determine the ancestral nucleotide and genomic position for each protein coding base pair in our data sets. This allowed us to form a complete picture of the history of genomic reorganization through the associated nucleotides (and substitutions) and genomic positions. To date, there are many studies that look at the spatial organization of substitutions along bacterial genomes (Couturier and Rocha 2006; Flynn et al. 2010; Cooper et al. 2010; Morrow and Cooper 2012), which found that substitution rate typically increases with increasing distance from the origin of replication. Yet, none of these studies incorporate genomic reorganization such as rearrangements and inversions, which are known to happen frequently in bacterial genomes. In Chapter 2 we explored the spatial trends of substitutions and $\mathrm{d}N$, $\mathrm{d}S$, and ω values along bacterial genomes to add to the previous knowledge of spatial trends in bacteria.

Using ancestral reconstruction in a novel way to account for genomic reorganization, we observed a significant but inconsistent correlation between distance from the origin of replication and the number of substitutions. We were unable to detect a significant consistent relationship between values of $\mathrm{d}N$, $\mathrm{d}S$, and ω and distance from the origin of replication. This necessitates further in-depth analysis of other molecular trends in bacterial genomes while accounting for genomic reorganization. Using tools such as ancestral reconstruction to determine the history of rearrangements, other spatial molecular trends in bacteria can be accurately elucidated. This can be applied to gene expression and essentiality, to determine how these molecular components are impacted by rearrangements and what this tells us about the organization of genes along bacterial genomes. Determining how the number of substitutions are distributed spatially throughout bacterial genomes while considering rearrangements, broadens our knowledge of bacterial adaptability and evolution. Chapter 2 provides evidence that all aspects of genome (re-)organization need to be incorporated into spatial genomic analysis in bacteria, as it can have a profound impact on what are thought to be "universal" spatial molecular trends.

5.3 Gene Expression Along the Origin and Terminus of Replication Axis in Bacteria

In addition to the number of substitutions, previous studies have found gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012) and gene dosage (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Sauer et al. 2016) vary when a gene is moved to different genomic locations. This phenomenon is pervasive across bacteria and creates a predictable pattern about where highly expressed genes are found within bacterial genomes. Typically gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012) and gene dosage (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Sauer et al. 2016) are increased near the origin of replication, and decreased near the terminus of replication (Couturier and Rocha 2006). Although many studies have found this linear gene expression trend to be universal, it is unclear if this phenomenon is persistent across diverse genomes and bacterial species. In particular, there have been no studies that look at how gene expression varies with genomic position when combining gene expression data from multiple experiments. Chapter 3 addressed this gap in knowladge by looking at the overall expression levels of all genes within eleven gene expression data sets from bacterial genomes of *E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti*. We have combined information from multiple previously published RNA-seq experiments to determine the spatial pattern of gene expression using this amalgamation of data. Using whole genome RNA-seq expression data obtained from the Gene Expression Ombinbus (GEO) database (Barrett et al. 2012), we are able to observe genomic expression patterns in natural populations devoid of stress, while accounting for bidirectional replication. We have confirmed that gene expression indeed tends to be higher near the origin of replication and decreases with increasing distance from the origin when multiple datasets are used. Understanding how the distance of a gene from the origin of replication can impact the expression level assists in explaining other trends along the origin-terminus of replication axis such as gene essentiality, gene conservation, and mutation rates.

5.3.1 Establishing a Baseline Trend for Genomic Traits in Bacteria

The highly organized structure of bacterial genomes can aid in the prediction of various traits of a gene. In Chapter 2 we demonstrated the profound impact that genomic reorganization can have on what is typically thought of as "universal" patterns for substitutions in bacterial genomes. However, to properly assess the impact genomic reorganization can have on various molecular traits, it is important to first establish what the prevailing trend is in the absence of reorganization. To our knowladge, there is no information on how gene expression changes along the origin-terminus of replication axis when considering data from multiple bacterial RNA-seq datasets. In Chapter 3 we identified what the genomic landscape of bacterial gene expression

is when combining previously published RNA-seq experimental data. We found that gene expression decreases linearly with increasing distance from the origin of replication, corroborating results from previous studies that used only one RNA-seq dataset. Determining the pattern of variation in gene expression along the origin-terminus of replication axis gives us greater understanding of how bacteria can utilize the organization of genetic information to influence expression. This general organization can inform how we expect expression to change if the genomic location of a gene changes which can be useful when considering HGT. We can then use these overarching patterns of spatial organization to explore how more complex phenomenon - such as genomic reorganization - can alter the genomic landscape of a molecular trait.

5.4 Genomic Reorganization and Gene Expression in Bacterial Genomes

In Chapter 2, we illustrated the impact genomic reorganization can have on substitutions in bacterial genomes. Accounting for genomic reorganization can alter the "typical" spatial pattern of substitutions along bacterial genomes. After establishing a baseline trend for gene expression and how it changes with distance from the origin of replication using previously published RNA-seq data (Chapter 3), we wanted to examine how this trend changes when genome reorganization is incorporated into the analysis. Inversions are one particular type of genomic reorganization that can promote spontaneous genome rearrangements (Sun et al. 2012), and in some cases, are the only source of rearrangement in bacteria (Romling et al. 1997). Inversions impact a multitude of molecular traits in bacteria, but the most intriguing is providing a way for bacteria to alter their gene expression (Zieg et al. 1977; Zieg et al. 1978; Sekulovic et al. 2018; Li et al. 2019), having a short- (Cerdeño-Tárraga et al. 2005) and long-range (Wong and Wolfe 2005; Cerdeño-Tárraga et al. 2005; Naseeb et al. 2016; Sekulovic et al. 2018) impact on gene expression.

Previous studies investigating the influence inversions have on bacterial gene expression have largely focused on a single inversion (Zieg et al. 1978; Sekulovic et al. 2018), a small number of genes (Zieg et al. 1977; Li et al. 2019), or focus on comparing inversions between distantly related species (Alokam et al. 2002; Wong and Wolfe 2005; Naseeb et al. 2016). In Chapter 4, we explored differences in gene expression due to inversions between closely related strains of *E. coli*. We identified hundreds of inversions and combined RNA-seq datasets from multiple previously published studies to examine the short- and long-range impact inversions have on gene expression. We determined that there is a significant difference in gene expression between inverted and non-inverted regions of the genome. Within a few of the inversions identified, there is a significant difference in gene expression between the inverted and non-inverted sequences, with inverted sequences having higher gene expression 75% the time. The H-NS and Fis nucleoid associated proteins have been associated with gene expression regulation (Johansson et al. 2000; Kelly et al. 2004; Paul et al. 2004; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019; Kahramanoglou et al. 2011), and we observed a significant positive

correlation between some identified inversions and H-NS and Fis binding sites. These correlations suggest that genomic architecture may play a role in regulating expression in inverted regions. Inversions can impact gene expression even between closely related strains of *E. coli* and could provide a mechanism for strains to diversify their genetic content through controlled expression changes.

5.5 Bacterial Molecular Analysis and Genomic Reorganization

Throughout this work we have established that the number of substitutions (Chapter 2), gene expression (Chapter 3), and inversions (Chapter 4) vary with genomic position. These are similar conclusions reached by previous studies looking at how substitution rates (dN and dS) (Cooper et al. 2010; Morrow and Cooper 2012), gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012), gene dosage (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Sauer et al. 2016), and gene conservation (Couturier and Rocha 2006) change with increasing distance from the origin of replication. However, when genomic reorganization, such as rearrangements and inversions, are accounted for, the distribution of these molecular traits along the origin-terminus of replication axis can change. We demonstrated that genomic reorganization can have profound impact on the landscape of both substitutions (Chapter 2) and gene expression (Chapter 4). To obtain an accurate representation of how molecular traits change with distance from the origin of replication, genomic reorganization needs to be considered. This can assist in determining evolutionary forces working to diversify bacterial genomes, and elucidate how bacteria can utilize the organization of genetic information along the origin-terminus of replication axis to adapt on a molecular level.

5.6 Future Studies

5.6.1 Extensive and Detailed Control RNA-seq Data

In Chapter 3 we demonstrated that combining gene expression information from multiple previously published experiments deepens the understanding of the spatial genomic gene expression trends. Studying the changes in gene expression of a particular gene when it is re-located to predetermined positions within bacterial genomes is necessary to elucidate the individual impact genomic location has on genes. However, it does not provide a complete picture of the interplay between genomic location and gene interactions. A genomic approach to gene expression profiles of bacteria is crucial to capture the variety of expression levels and their interactions within a bacterial genome. In order to elucidate any genomic trends that may occur in gene expression along the origin-terminus of replication axis, a number of control datasets must be examined.

The definition of a "controlled" environment depends on the experiment in question. This could be environmental conditions absent of any stress, sufficient resources, or strains absent of particular mutations. The nature of the original experiment often dictates what is considered a "control" condition. The GEO (Barrett et al. 2012) is a common repository for RNA-seq experiments, providing a wealth of publicly available experimental data. There are a number of specifications and fields that authors can fill out to ensure that important information about the datasets is present. However, many of these helpful meta-data fields are not mandatory and left to author discretion for completion. This creates vast inconsistencies in the annotation and information provided for each experimental dataset. Sometimes there are long descriptions about the nature of the experiment, specific strains used, read mapping tools, and control environment information. In other cases, the material provided is sparse. When performing strictly computational analysis that attempts to combine all available RNA-seq datasets, this lack of detailed information becomes problematic for identifying complementary experiments where data can reasonably be combined. Without proper information on what constitutes a "control" environment, it is difficult to confidently state that two experiments performed in different laboratories at different times can be considered similar enough to combine that information. If two experiments are deemed to have the same "control" environment, the number of datasets available proves a constant concern for statistical power. In the "popular" model organisms such as *E. coli*, *Drosophila melanogaster*, and *Mus musculus*, the amount of data available is not often an issue. However, in less "popular" organisms such as *S. meliloti*, the amount of available data is sparse and often not containing the same "control" environments.

Having a more rigorous and structured format for submitting meta-data, and in particular RNA-seq meta-data, to publicly available databases will avoid some of these issues. Creating a specific submission category that requires researchers to identify the exact conditions their "control" bacteria were reared in will simplify the process of assessing if multiple datasets are comparable. Requiring extra steps to search through the published paper (if applicable) to determine key information about the experiments is unnecessary and time consuming. All the information needed about a particular experiment should be easily accessible in the GEO meta-data fields, making the automated process of gathering and scanning for appropriate datasets fast and simple.

Additionally, there needs to be a focus on gathering and storing data, particularly RNA-seq data, from non-model organisms. In Chapter 3, we were unable to detect a significant linear relationship between distance from the origin of replication and gene expression for all replicons of *S. meliloti* (chromosome, pSymA and pSymB). Gene expression in this bacteria is not as well studied as the other bacteria used in this analysis (Martens et al. 2008). In our search for expression data, we identified fewer appropriate studies for *S. meliloti* to include in our data analyses. A smaller amount of gene expression data may be biasing the non-significant correlation between gene expression and distance from the origin of replication in this analysis. In biology there are constant exceptions to the "rules" and these are typically only found when

considering non-model organisms. For example, when examining the typical organization of genes and molecular traits along the origin-termins of replication axis, there are a number of bacteria and strains that do not follow these general rules (Hudson et al. 2002; Ochman 2003; Martina et al. 2012; Juurik et al. 2012; Dettman et al. 2016; Dillon et al. 2018). Without sufficient amounts of non-model organism data, it is dangerous to label spatial molecular trends as "universal". Having a wealth of non-model organism data and specifying more meta-data, will eliminate some of the concerns about statistical power of an analysis and enable the amalgamation of all comparable available RNA-seq datasets.

5.6.2 Expanding Spatial Molecular Trends to Other Conditions and Strains

To date, there are many detailed accounts of how bacterial genomes are organized along the origin-terminus of replication axis (Cooper and Helmstetter 1968; Chandler et al. 1975; Chandler and Pritchard 1975; Bremer and Churchward 1977; Schmid and Roth 1987; Sousa et al. 1997; Couturier and Rocha 2006; Bryant et al. 2014; Le and Laub 2014; Gerganova et al. 2015; Kopejtka et al. 2019). Prior research on spatial molecular trends when moving from the origin of replication to the terminus have determined that substitution rates (dN and dS) increase with distance from the origin of replication (Cooper et al. 2010; Morrow and Cooper 2012). Additionally, gene expression (Sharp et al. 2005; Couturier and Rocha 2006; Morrow and Cooper 2012), gene dosage (Cooper and Helmstetter 1968; Schmid and Roth 1987; Rocha 2004a; Block et al. 2012; Sauer et al. 2016), and gene conservation (Couturier and Rocha 2006) are increased near the origin, and decrease near the terminus (Couturier and Rocha 2006). However, these studies do not take into account genome reorganization such as rearrangements and inversions, which happen frequently in bacterial genomes and are an important source of genomic variation for bacteria (Ochman et al. 2000; Epstein et al. 2014). Nor do they consider the deep evolutionary history of these substitutions through methods such as ancestral reconstruction. In Chapter 2, we addressed this gap in knowladge by exploring the spatial pattern of substitutions along the replicons of *E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti*. Using ancestral reconstruction, we were able to shed light on the evolutionary history of the substitutions and their respective genomic positions. We determined that the number of substitutions significantly varies with distance from the origin of replication in all replicons studied, however this correlation is inconsistent in sign. Some replicons had a significantly decreasing trend (*E. coli* and the chromosome of *S. meliloti*), while others showed the opposite significant trend (*B. subtilis*, *Streptomyces*, pSymA and pSymB in *S. meliloti*). dN, dS and ω were examined across all genes and there was no consistent significant correlation between those values and distance from the origin. This provides an in-depth analysis of how genomic rearrangements can impact the location of substitutions and substitution rates along bacterial genomes.

For the analyses described in Chapter 2, we used only complete bacterial genomes from a few different species of bacteria to obtain the most accurate information for the ancestral

reconstruction. This gives a good indication on how accounting for genomic reorganization, such as rearrangements and inversions, can impact the general spatial pattern of substitutions in bacterial replicons under ordinary circumstances. It is well known that bacteria are accomplished in adapting to ever changing environmental conditions. It would be valuable to determine how bacterial lineages reared in certain environments or possessing particular mutations or genetic abilities, differ in their spatial substitutions trends. Utilizing the general ancestral reconstruction methods presented in Chapter 2, genomic reorganization could be identified and accounted for in any number of genetically or environmentally altered bacteria.

For example, there are strains of *S. meliloti* available that can successfully propagate while missing the smaller replicons pSymA and pSymB (DiCenzo et al. 2014). These strains have had the essential and necessary genes from pSymB successfully integrated into the chromosome replicon, eliminating the need for the secondary replicons (DiCenzo et al. 2014). As mentioned previously, the secondary replicons of multi-repliconic bacteria provide further organization of the genome, acting as an evolutionary test bed for recently acquired genes (Cooper et al. 2010). Completely removing these secondary replicons would have a profound impact on the genetic landscape of the remaining primary chromosome. Investigating the pattern of substitutions with increasing distance from the origin of replication in these mutant *S. meliloti* strains could provide information on where rearrangements are located and how the interaction between rearrangements and the loss of secondary replicons impacts the placement of substitutions. This type of analysis could be extended to look at the evolutionary history of substitutions in any mutant strain, such as antibiotic resistant strains of *E. coli*. The ancestral analysis need not stop at substitutions. The evolutionary history of gene expression, dosage, and function and their distance from the origin of replication can all be analyzed using ancestral reconstruction. This can inform us on the reorganization patterns in "mutant" phenotypes and their impact on spatial molecular trends.

5.6.3 Identification of sequences and proteins involved in genomic architecture within the *E. coli* ATCC 25922 strain.

Genomic inversions provide a way for bacteria to alter the expression of genes (Zieg et al. 1977; Zieg et al. 1978; Sekulovic et al. 2018; Li et al. 2019). Previous studies investigating the impacts inversions have on bacterial gene expression have largely focused on a single inversion (Zieg et al. 1978; Sekulovic et al. 2018), or the impact inversions have on the expression of a small number of genes (Zieg et al. 1977; Li et al. 2019). Studies that took a genomic approach to analyzing inversions and their impact on gene expression are typically focused on yeast (Wong and Wolfe 2005; Naseeb et al. 2016), with few studies examining genome wide inversions in distantly related bacteria (Alokam et al. 2002). In Chapter 4 we identified hundreds of naturally occurring inversions between four closely related strains of *E. coli*: K-12 MG1655, K-12 DH10B, BW25113 and ATCC 25922. We combine information from multiple previously published RNA-seq datasets to examine the short- and long-range genomic impacts these inversions have on gene

expression. Within most of the inverted regions identified, there was no significant difference in gene expression between inverted and non-inverted sequences. However, in the inverted alignment blocks that did have a significant difference in gene expression between inverted and non-inverted sequences, the inverted sequences had higher gene expression 75% of the time. Most of the inversions that were identified occurred in the *E. coli* ATCC 25922 strain relative to the other strains. In the *E. coli* ATCC 25922 strain, the location of the inversions significantly decreased with increasing distance from the origin of replication. The location of these inversions could be constrained by sequences and proteins involved in maintaining correct genomic architecture, such as Architecture IMparting Sequences (AIMS), H-NS and Fis, and warrants further investigation. The H-NS and Fis proteins are additionally well known transcription regulators (Johansson et al. 2000; Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019), so we investigated correlations between H-NS and Fis binding sites and the identified inversions. H-NS and Fis are significantly positively correlated with various inversions identified in our analysis, suggesting that nucleoid associated proteins may play a role in the regulation of gene expression in these regions. We have provided an overview of the short- and long-range impacts inversions can have on gene expression between closely related strains of *E. coli*. Inversions can be used as a tool to alter genomic content, which can lead to changes in gene expression. Further classification and investigation of bacterial inversions can provide information on other molecular ways inversions can facilitate bacterial evolution. Inversions impact gene expression even between closely related strains of *E. coli* and could provide a mechanism for strains to diversify their genetic content through controlled expression changes.

The majority of the inversions identified in Chapter 4 occurred in the *E. coli* ATCC 25922 strain relative to the other strains (K-12 MG1655, K-12 DH10B, and BW25113). We therefore propose further exploration into various molecular aspects of the *E. coli* ATCC 25922 genome to investigate potential reasons for the large number of inversion and their impact. In Chapter 4, we detected a significant negative trend between inversions and distance from the origin of replication in the *E. coli* ATCC 25922 genome. This suggests that the majority of the inversions identified are concentrated closer to the origin of replication rather than the terminus. Some examples of inversions disrupting the macro- and micro-domains of the physically folded chromosomes can be extremely harmful for the growth and well being of the bacteria (Segall et al. 1988; Raeside et al. 2014; Naseeb et al. 2016). This concept has been related to the distribution of AIMS, which act during DNA segregation. A change in the distribution of AIMS can disrupt AIMS-based genome architecture (Hendrickson et al. 2018). The inversions identified in *E. coli* ATCC 25922 could be located closer to the origin of replication because in this position they do not disrupt AIMS distribution. Most of the information regarding the profile of AIMS along the *E. coli* genome is specific to the *E. coli* K-12 strain. The particular arrangement of AIMS in the context of the *E. coli* ATCC 25922 genome is not well known. A detailed analysis of the location of AIMS within the *E. coli* ATCC 25922 genome would give insight into how the inversions identified in Chapter 4 relate to the AIMS distribution.

In addition to maintaining and controlling genomic structure, H-NS acts as a global transcription regulator (Johansson et al. 2000; Kahramanoglou et al. 2011) repressing the transcription of non-essential genes (Browning et al. 2000; Hommais et al. 2001; Dorman 2004; Fang and Rimsky 2008; Dillon and Dorman 2010; Ali et al. 2012; Singh et al. 2016) and playing an important role in silencing genes recently acquired via HGT (Dorman 2004; Oshima et al. 2006; Dorman 2007; Ali et al. 2014; Higashi et al. 2016), assisting in their integration into the host bacterial genome (Dorman 2007). Fis on the other hand, activates transcription, increasing expression (Bradley et al. 2007; Cho et al. 2008; Kahramanoglou et al. 2011; Scholz et al. 2019). In Chapter 4, we detected significant overlap between H-NS binding and almost all of our identified inversions, and significant overlap between Fis binding and significant inverted alignment blocks. These associations indicate that H-NS and Fis could be playing a role in altering gene expression in the *E. coli* ATCC 25922 strain, where the majority of the identified inversions occurred. As with the information on the distribution of AIMS, most information on potential H-NS and Fis binding sites is in reference to the *E. coli* K-12 strain. There is little information on the exact binding sites of H-NS within the *E. coli* ATCC 25922 genome. Determining the genomic locations of H-NS and Fis binding sites would clarify if the strain specific inversions identified in *E. coli* ATCC 25922 are frequently bound and potentially under the influence of these nucleoid associated proteins. To date there are many studies that have identified sequences specific to the binding of the H-NS protein (Grainger et al. 2006; Oshima et al. 2006; Bouffartigues et al. 2007; Lang et al. 2007; Higashi et al. 2016; Rangarajan and Schnetz 2018). We propose that using similar techniques used in Grainger et al. (2006), Oshima et al. (2006), Bouffartigues et al. (2007), Lang et al. (2007), Higashi et al. (2016), and Rangarajan and Schnetz (2018), potential binding sites for the H-NS and Fis proteins can be identified in the *E. coli* ATCC 25922 genome. These can then be investigated to determine how various nucleoid associated proteins influence gene expression in bound areas and in nearby genes.

As mentioned previously, H-NS can silence genes recently acquired via HGT (Dorman 2004; Oshima et al. 2006; Dorman 2007; Ali et al. 2014; Higashi et al. 2016). Spurious transcription in recently horizontally transferred genes could decrease fitness (Singh et al. 2014; Lamberte et al. 2017). It has been proposed that the repression of recently transferred genes via H-NS may assist in their integration into the host bacterial genome (Dorman 2007) and may ultimately assist the bacteria in thriving under new environmental conditions (Hommais et al. 2001; Navarre et al. 2007; Higashi et al. 2016). In this work we did not detect a significant correlation between horizontally acquired genes and H-NS binding sites. Similar to the knowledge surrounding AIMS, H-NS, and Fis, there is insufficient information on genes acquired via HGT in the *E. coli* ATCC 25922 strain. Further investigation and identification of genes recently obtained by HGT in the *E. coli* ATCC 25922 strain should be conducted. This list of transferred genes in *E. coli* ATCC 25922 will provide additional insight into why certain inversions have significantly different levels of gene expression and potentially how this relates to most of the identified inversions being located near the origin of replication.

5.7 Conclusion

In this work we determined the spatial genomic trends of substitutions, substitution rates, and gene expression along the origin-terminus of replication axis in all replicons of *E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti*. These trends were determined while accounting for genomic reorganization, such as rearrangements and inversions, using a variety of bioinformatic techniques. We used the `PAML` (Yang 1997) program to reconstruct the ancestral nucleotide sequence and modified this program to allow for the reconstruction of the corresponding genomic positions for protein-coding regions within the genomes of *E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti*. The synonymous (dS) and non-synonymous (dN) substitution rates and their ratio (ω) were calculated for each of these protein-coding regions. This information was used to elucidate the distribution of substitutions, substitution rates, and ω along the origin-terminus of replication axis. We observed that the number of substitutions significantly changes with distance from the origin of replication, but the sign of this correlation is inconsistent. Some replicons had the number of substitutions significantly decrease with increasing distance from the origin of replication (the chromosomes of *E. coli* and *S. meliloti*), while others showed the opposite significant trend (*B. subtilis*, *Streptomyces* and the two smaller replicons of *S. meliloti*: pSymA and pSymB). Genes having higher or lower values of dN, dS and ω do not appear to be clustered at any one particular genomic position and there was no significant correlation between those values and distance from the origin. We did not observe a predictable increase in substitutions with increasing distance from the origin of replication, and suspect that accounting for genomic reorganization may have influenced the results.

We wanted to investigate if the spatial patterns of gene expression in the above mentioned bacteria were similarly impacted by accounting for rearrangements. To examine this phenomenon it was first necessary to determine what the genomic gene expression landscape is when combining publicly available gene expression data from multiple experiments. Utilizing previously published RNA-seq datasets, we established the distribution of gene expression along the origin and terminus of replication axis. We were able to detect a significant negative linear relationship, where gene expression decreases with increasing distance from the origin of replication. Finally, we used this information to examine the genomic impacts certain genomic reorganization, such as inversions, has on gene expression in several strains of *E. coli*. We identified hundreds of naturally occurring inversions between various strains of *E. coli*. Using existing RNA-seq data, we were able to study the short- and long-range genomic impacts inversions have on gene expression. Within most of the inverted regions identified, there was no significant difference in gene expression between inverted and non-inverted sequences, and no significant difference in the variation in expression between inverted and non-inverted regions. However, in the inverted alignment blocks that did have a significant difference in gene expression between inverted and non-inverted sequences, the inverted sequences had higher gene expression (1.27-85.58 fold) 75% of the time and lower gene expression (1.2-100 fold) 25% of the time. Most of the inversions that

were identified occurred in the *E. coli* ATCC 25922 strain and the location of these inversions significantly decreased with increasing distance from the origin of replication. H-NS and Fis are significantly positively correlated with some inversions identified in our analysis, suggesting that nucleoid associated proteins may play a role in the regulation of gene expression in these regions.

Our analyses takes a novel approach by accounting for genomic reorganization and applying ancestral reconstruction for both nucleotides and genomic positions. This provided an evolutionary history of substitutions and where they have been rearranged within bacterial genomes. We were able to successfully replicate the gene expression trends identified in single experiment studies, amalgamating RNA-seq data from multiple previously published studies. We conducted a whole genome review of the impacts inversions can have on gene expression in *E. coli*, and how this relates to the origin-terminus of replication axis. Finally, we demonstrated that accounting for various forms of genomic reorganization, such as rearrangements and inversions, is crucial to the analysis of molecular traits in bacteria. This is particularly important when considering how molecular traits vary with genomic position. We demonstrated that including information about genomic reorganization can produce patterns that do not align with the "most common" spatial trend. This challenges the notion that these trends are "universal", and can be significantly altered using genomic reorganization history.

Bibliography

Abraham, J. M., Freitag, C. S., Clements, J. R., and Eisenstein, B. I. (1985). An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. Proc Natl Acad Sci 82(17), 5724–5727.

Afgan, E, Baker, D, Batut, B, Van Den Beek, M, Bouvier, D, Čech, M, Chilton, J, Clements, D, Coraor, N, Grüning, B. A., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46(W1), W537–W544.

Ahnert, S. E., Fink, T. M. A., and Zinovyev, A (2008). How much non-coding DNA do eukaryotes require? J Theor Biol 252(4), 587–592.

Ali, S. S., Soo, J, Rao, C, Leung, A. S., Ngai, D. H., Ensminger, A. W., and Navarre, W. W. (2014). Silencing by H-NS potentiated the evolution of *Salmonella*. PLoS Pathog 10, e1004500.

Ali, S. S., Xia, B, Liu, J, and Navarre, W. W. (2012). Silencing of foreign DNA in bacteria. Curr Opin Microbiol 15, 175–181.

Allen, T. E., Price, N. D., Joyce, A. R., and Palsson, B. Ø. (2006). Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. PLoS Comp Biol 2(1), e2.

Alm, R. A., Ling, L.-S. L., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B, Guild, B. C., and Carmel, G (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397(6715), 176.

Alokam, S, Liu, S.-L., Said, K, and Sanderson, K. E. (2002). Inversions over the terminus region in *Salmonella* and *Escherichia coli*: IS200s as the sites of homologous recombination inverting the chromosome of *Salmonella enterica serovar typhi*. J Bacteriol 184(22), 6190–6197.

Altschul, S. F., Gish, W, Miller, W, Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Amos, W (2010). Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proc Royal Soc B: Biol Sci 277(1686), 1443–1449.

Anjuwon-Foster, B. R. and Tamayo, R (2017). A genetic switch controls the production of flagella and toxins in *Clostridium difficile*. PLoS Genet 13(3), e1006701.

Aseev, L. V., Levandovskaya, A. A., Tchufistova, L. S., Scaptsova, N. V., and Boni, I. V. (2008). A new regulatory circuit in ribosomal protein operons: S2-mediated control of the rpsB-tsf expression in vivo. RNA 14(9), 1882–1894.

Badia, J, Ibanez, E, Sabate, M, Baldoma, L, and Aguilar, J (1998). A rare 920-kilobase chromosomal inversion mediated by IS1 transposition causes constitutive expression of the yiaK-S operon for carbohydrate utilization in *Escherichia coli*. J Biol Chem 273, 8376–8381.

Badrinarayanan, A, Le, T. B. K., and Laub, M. T. (2015). Bacterial chromosome organization and segregation. Annual Rev Cell Devel Biol 31, 171–199.

Baek, J. H. and Chattoraj, D. K. (2014). Chromosome I controls chromosome II replication in *Vibrio cholerae*. PLoS Genet 10(2), e1004184.

Bahrani, F. K. and Mobley, H. L. (1994). Proteus mirabilis MR/P fimbrial operon: genetic organization, nucleotide sequence, and conditions for expression. J Bacteriol 176(11), 3412–3419.

Ball, C. A. and Johnson, R. C. (1991). Efficient excision of phage lambda from the *Escherichia coli* chromosome requires the Fis protein. J Bacteriol 173, 4027–4031.

Barrett, T, Wilhite, S. E., Ledoux, P, Evangelista, C, Kim, I. F., Tomashevsky, M, Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M, et al. (2012). NCBI GEO: archive for functional genomics data sets. Nucleic Acids Res 41(D1), D991–D995.

Belda, E., Moya, A., and Silva, F. J. (2005). Genome rearrangement distances and gene order phylogeny in gamma-proteobacteria. Mol Bio Evol 22(6), 1456–1467.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H, Harper, et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor A3(2)*. Nature 417(6885), 141–147.

Benzer, S (1961). On the topography of the genetic fine structure. Proc Natl Acad Sci U S A 47(3), 403.

Berg, O. G. and Kurland, C. G. (1997). Growth rate-optimised tRNA abundance and codon usage1.

Bhowmik, B. K., Clevenger, A. L., Zhao, H, and Rybenkov, V. V. (2018). Segregation but Not Replication of the *Pseudomonas aeruginosa* Chromosome Terminates at Dif. mBio 9(5), e01088–18.

Biondi, E. G., Tatti, E, Comparini, D, Giuntini, E, Mocali, S, Giovannetti, L, Bazzicalupo, M, Mengoni, A, and Viti, C (2009). Metabolic capacity of *Sinorhizobium* (*Ensifer*) *meliloti* strains as determined by phenotype MicroArray analysis. Appl Environ Microbiol 75, 5396–5404.

Blakely, G, May, G, McCulloch, R, Arciszewska, L. K., Burke, M, Lovett, S. T., and Sherratt, D. J. (1993). Two related recombinases are required for site-specific recombination at dif and cer in *E. coli K12*. Cell 75(2), 351–361.

Blattner, F. R., PlunkettG, r., Bloch, C. A., Perna, N. T., Burland, V, Riley, M, Collado-Vides, J, Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of *Escherichia coli K-12*. Science 277, 1453–1462.

Block, D. H. S., Hussein, R, Liang, L. W., and Lim, H. N. (2012). Regulatory consequences of gene translocation in bacteria. Nucleic Acids Res 40(18), 8979–8992.

Blomfield, I. C. (2015). Sialic acid and N-acetylglucosamine regulate type 1 fimbriae synthesis. In: Metabol Bacter Pathogen. American Society of Microbiology, 95–103.

Boccard, F, Esnault, E, and Valens, M (2005). Spatial arrangement and macrodomain organization of bacterial chromosomes. Mol Microbiol 57(1), 9–16.

Bochkareva, O. O., Moroz, E. V., Davydov, I. I., and Gelfand, M. S. (2018). Genome rearrangements and selection in multi-chromosome bacteria *Burkholderia spp.* BMC Genomics 19(1), 965.

Borst, P and Greaves, D. R. (1987). Programmed gene rearrangements altering gene expression. Science 235(4789), 658–667.

Bouffartigues, E, Buckle, M, Badaut, C, Travers, A, and Rimsky, S (2007). H-NS cooperative binding to high-affinity sites in a regulatory element results in transcriptional silencing. Nat Struct Mol Biol 14, 441–448.

Bradley, M. D., Beach, M. B., DeKoning, A. P. J., Pratt, T. S., and Osuna, R (2007). Effects of Fis on *Escherichia coli* gene expression during different growth stages. Microbiology (Reading) 153, 2922–2940.

Brandis, G, Bergman, J. M., and Hughes, D (2016). Autoregulation of the tufB operon in Salmonella. Mol Microbiol 100(6), 1004–1016.

Bremer, H and Churchward, G (1977). An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions. J Theor Biol 69(4), 645–654.

Brewer, B. J. (1988). When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell 53, 679–686.

Browning, D. F., Cole, J. A., and Busby, S. J. (2000). Suppression of FNR-dependent transcription activation at the *Escherichia coli* nir promoter by Fis, IHF and H-NS: modulation of transcription initiation by a complex nucleo-protein assembly. Mol Microbiol 37, 1258–1269.

Bryant, J. A., Sellars, L. E., Busby, S. J. W., and Lee, D. J. (2014). Chromosome position effects on gene expression in *Escherichia coli K-12*. Nucleic Acids Res 42(18), 11383–11392.

Buchan, J. R., Aucott, L. S., and Stansfield, I (2006). tRNA properties help shape codon pair preferences in open reading frames. Nucleic Acids Res 34(3), 1015–1027.

Byrne, R, Levin, J. G., Bladen, H. A., and Nirenberg, M. W. (1964). The in vitro formation of a DNA-ribosome complex. Proc Natl Acad Sci U S A 52, 140–148.

Cagliero, C, Grand, R. S., Jones, M. B., Jin, D. J., and O'sullivan, J. M. (2013). Genome conformation capture reveals that the *Escherichia coli* chromosome is organized by replication and transcription. Nucleic Acids Res 41(12), 6058–6071.

Canchaya, C., Claesson, M. J., Fitzgerald, G. F., Van Sinderen, D., and O'Toole, P. W. (2006). Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. Microbiol 152(11), 3185–3196.

Cannarozzi, G, Schraudolph, N. N., Faty, M, Rohr, P von, Friberg, M. T., Roth, A. C., Gonnet, P, Gonnet, G, and Barral, Y (2010). A role for codon order in translation dynamics. Cell 141(2), 355–367.

Capella-Gutiérrez, S, Silla-Martinez, J. M., and Gabaldón, T (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinfor 25(15), 1972–1973.

Casillas, S and Barbadilla, A (2017). Molecular population genetics. Genetics 205(3), 1003–1035.

Cerdeño-Tárraga, A. M., Patrick, S, Crossman, L. C., Blakely, G, Abratt, V, Lennard, N, Poxton, I, Duerden, B, Harris, B, and Quail, M. A. (2005). Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. Science 307(5714), 1463–1465.

Chandler, M, Bird, R. E., and Caro, L (1975). The replication time of the *Escherichia coli* K12 chromosome as a function of cell doubling time. J Mol Bio 94(1), 127–132.

Chandler, M. G. and Pritchard, R. H. (1975). The effect of gene concentration and relative gene dosage on gene output in Escherichia coli. Mol Gen Genet MGG 138(2), 127–141.

Chintakayala, K, Singh, S. S., Rossiter, A. E., Shahapure, R, Dame, R. T., and Grainger, D. C. (2013). *E. coli* Fis protein insulates the cbpA gene from uncontrolled transcription. PLoS Genet 9, e1003152.

Cho, B. K., Knight, E. M., Barrett, C. L., and Palsson, B. O. (2008). Genome-wide analysis of Fis binding in *Escherichia coli* indicates a causative role for A-/AT-tracts. Genome Res 18, 900–910.

Choulet, F, Aigle, B, Gallois, A, Mangenot, S, Gerbaud, C, Truong, C, Francou, F.-X., Fourrier, C, Guérineau, M, Decaris, B, et al. (2006). Evolution of the terminal regions of the *Streptomyces* linear chromosome. Mol Bio Evol 23(12), 2361–2369.

Ciampi, M. S., Schmid, M. B., and Roth, J. R. (1982). Transposon Tn10 provides a promoter for transcription of adjacent sequences. Proc Natl Acad Sci 79(16), 5016–5020.

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(8), 780–788.

Colson, I, Delneri, D, and Oliver, S. G. (2004). Effects of reciprocal chromosomal translocations on the fitness of *Saccharomyces cerevisiae*. EMBO Rep 5, 392–398.

Cooper, S and Helmstetter, C. E. (1968). Chromosome replication and the division cycle of *Escherichia coli B/r*. J Mol Bio 31(3), 519–540.

Cooper, V. S., Vhor, S. H., Wrocklage, S. C., and Hatcher, P. J. (2010). Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comp Biol 6(4), e1000732.

Courcelle, J (2009). Shifting replication between IInd, IIIrd, and IVth gears. Proc Natl Acad Sci 106(15), 6027–6028.

Couturier, E and Rocha, E. P. (2006). Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol 59(5), 1506–1518.

Cox, M. M. (2007). Motoring along with the bacterial RecA protein. Nat Rev Mol Cell Biol 8(2), 127.

Cui, L, Neoh, H, Iwamoto, A, and Hiramatsu, K (2012). Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. Proc Natl Acad Sci 109(25), E1647–E1656.

Dages, S, Zhi, X, and Leng, F (2020). Fis protein forms DNA topological barriers to confine transcription-coupled DNA supercoiling in *Escherichia coli*. FEBS Lett 594, 791–798.

Danchin, A, Dondon, L, and Daniel, J (1984). Metabolic alterations mediated by 2-ketobutyrate in Escherichia coli K12. Mol Gen Genet 193, 473–478.

Darling, A. E., Mau, B, and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS one 5(6), e11147.

Daubin, V, Lerat, E, and Perrière, G (2003). The source of laterally transferred genes in bacterial genomes. Genome Biol 4(9), R57.

Daubin, V and Ochman, H (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 14(6), 1036–1042.

Daubin, V and Perriere, G (2003). G+ C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol 20(4), 471–483.

Dettman, J. R., Sztepanacz, J. L., and Kassen, R. (2016). The properties of spontaneous mutations in the opportunistic pathogen *Pseudomonas aeruginosa*. Bmc Genomics 17(1), 27.

Dhar, G, Heiss, J. K., and Johnson, R. C. (2009). Mechanical constraints on Hin subunit rotation imposed by the Fis/enhancer system and DNA supercoiling during site-specific recombination. Mol Cell 34, 746–759.

DiCenzo, G. C., Benedict, A. B., Fondi, M, Walker, G. C., Finan, T. M., Mengoni, A, and Griffitts, J. S. (2018). Robustness encoded across essential and accessory replicons of the ecologically versatile bacterium *Sinorhizobium meliloti*. PLoS Genet 14, e1007357.

DiCenzo, G. C., MacLean, A. M., Milunovic, B, Golding, G. B., and Finan, T. M. (2014). Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet 10, e1004742.

DiCenzo, G. C., Mengoni, A, and Perrin, E (2019). Chromids aid genome expansion and functional diversification in the family *Burkholderiaceae*. Mol Bio Evol 36(3), 562–574.

Dillon, E. W. and Smith, J. A. (2017). Determinants of the match between student ability and college quality. J Labor Econ 35(1), 45–66.

Dillon, M. M., Sung, W, Lynch, M, and Cooper, V. S. (2015). The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. Genetics 200(3), 935–946.

Dillon, M. M., Sung, W, Lynch, M, and Cooper, V. S. (2018). Periodic variation of mutation rates in bacterial genomes associated with replication timing. MBio 9(4), e01371–18.

Dillon, S. C. and Dorman, C. J. (2010). Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol 8, 185–195.

Dorman, C. J. (2004). H-NS: a universal regulator for a dynamic genome. Nat Rev Microbiol 2, 391–400.

Dorman, C. J. (2007). H-NS, the genome sentinel. Nat Rev Microbiol 5, 157–161.

Dryselius, R, Izutsu, K, Honda, T, and Iida, T (2008). Differential replication dynamics for large and small *Vibrio* chromosomes affect gene dosage, expression and location. BMC Genomics 9, 559.

Duggin, I. G. and Bell, S. D. (2009). Termination structures in the *Escherichia coli* chromosome replication fork trap. J Mol Biol 387, 532–539.

Egan, E. S., Fogel, M. A., and Waldor, M. K. (2005). Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol 56(5), 1129–1138.

Eisen, J. A., Heidelberg, J. F., White, O, and Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1(6), research0011–1.

Ely, B, Wilson, K, Ross, K, Ingram, D, Lewter, T, Herring, J, Duncan, D, Aikins, A, and Scott, D (2019). Genome Comparisons of Wild Isolates of *Caulobacter crescentus* Reveal Rates of Inversion and Horizontal Gene Transfer. Curr Microbiol 76(2), 159–167.

Epstein, B, Sadowsky, M. J., and Tiffin, P (2014). Selection on horizontally transferred and duplicated genes in *Sinorizobium* (*Ensifer*), the root-nodule symbionts of *Medicae*. Genome Biol Evol 6(5), 1199–1209.

Esnault, E, Valens, M, Espéli, O, and Boccard, F (2007). Chromosome structuring limits genome plasticity in *Escherichia coli*. PLoS Genet 3(12), e226.

Fang, F. C. and Rimsky, S (2008). New insights into transcriptional regulation by H-NS. Curr Opin Microbiol 11, 113–120.

Farabaugh, P. J., Schmeissner, U, Hofer, M, and Miller, J. H. (1978). Genetic studies of the *lac* repressor: VII. On the molecular nature of spontaneous hotspots in the *lacI* gene of Escherichia coli. J Mol Biol 126(4), 847–863.

Feltz, C. J. and Miller, G. E. (1996). An asymptotic test for the equality of coefficients of variation from k populations. Stat Med 15, 646–658.

Filutowicz, M, Ross, W, Wild, J, and Gourse, R. L. (1992). Involvement of Fis protein in replication of the *Escherichia coli* chromosome. J Bacteriol 174, 398–407.

Flynn, K. M., Vohr, S. H., Hatcher, P. J., and Cooper, V. S. (2010). Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. Genom Biol Evol 2, 859–869.

Foster, P. L., Hanson, A. J., Lee, H, Popodi, E. M., and Tang, H (2013). On the mutational topology of the bacterial genome. G3 3(3), 399–407.

Foster, P. L., Niccum, B. A., Popodi, E, Townes, J. P., Lee, H, MohammedIsmail, W, and Tang, H (2018). Determinants of base-pair substitution patterns revealed by whole-genome sequencing of DNA mismatch repair defective *Escherichia coli*. Genetics 209(4), 1029–1042.

Francino, M. P. and Ochman, H (1997). Strand asymmetries in DNA evolution. Trends Genet 13, 240–245.

Francois, V, Louarn, J, Patte, J, Rebollo, J. E., and Louarn, J.-M. (1990). Constraints in chromosomal inversions in *Escherichia coli* are not explained by replication pausing at inverted terminator-like sequences. Mol Microbiol 4(4), 537–542.

Freeman, J. M., Plasterer, T. N., Smith, T. F., and Mohr, S. C. (1998). Patterns of genome organization in bacteria. Science 279(5358), 1827.

Frimodt-Moller, J, Charbon, G, Krogfelt, K. A., and Lobner-Olesen, A (2015). Control regions for chromosome replication are conserved with respect to sequence and location among *Escherichia coli* strains. Front Microbiol 6, 1011.

Furuta, Y, Abe, K, and Kobayashi, I (2010). Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. Nucleic Acids Res 38, 2428–2443.

Furuta, Y, Kawai, M, Yahara, K, Takahashi, N, Handa, N, Tsuru, T, Oshima, K, Yoshida, M, Azuma, T, and Hattori, M (2011). Birth and death of genes linked to chromosomal inversion. Proc of Natl Acad Sci 108(4), 1501–1506.

Galardini, M, Pini, F, Bazzicalupo, M, Biondi, E. G., and Mengoni, A (2013). Replicon-dependent bacterial genome evolution:the case of *Sinorhizobium meliloti*. Genome Biol Evol 5(3), 542–558.

Gally, D. L., Bogan, J. A., Eisenstein, B. I., and Blomfield, I. C. (1993). Environmental regulation of the fim switch controlling type 1 fimbrial phase variation in *Escherichia coli* K-12: effects of temperature and media. J Bacteriol 175(19), 6186–6193.

Garmendia, E, Brandis, G, and Hughes, D (2018). Transcriptional Regulation Buffers Gene Dosage Effects on a Highly Expressed Operon in *Salmonella*. mBio 9(5), e01446–18.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G, Ravasz, E, Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I, Gelfand, M. S., et al. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli MG1655*. J Bacteriol 185(19), 5673–5684.

Gerganova, V, Berger, M, Zaldastanishvili, E, Sobetzko, P, Lafon, C, Mourez, M, Travers, A, and Muskhelishvili, G (2015). Chromosomal position shift of a regulatory gene alters the bacterial phenotype. Nucleic Acids Res 43(17), 8215–8226.

Glaser, P, Rusniok, C, Buchrieser, C, Chevalier, F, Frangeul, L, Msadek, T, Zouine, M, Couvé, E, Lalioui, L, and Poyart, C (2002). Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol Microbiol 45(6), 1499–1513.

Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3(9), 679.

Goldman, N, Anderson, J. P., and Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. System Biol 49(4), 652–670.

Gordon, A. J. and Halliday, J. A. (1995). Inversions with deletions and duplications. Genetics 140(1), 411–414.

Gouy, M and Gautier, C (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10, 7055–7074.

Gouy, M, Guindon, S, and Gascuel, O (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221–224.

Grainger, D. C. (2016). The unexpected complexity of bacterial genomes. Microbiology (Reading) 162, 1167–1172.

Grainger, D. C., Hurd, D, Goldberg, M. D., and Busby, S. J. (2006). Association of nucleoid proteins with coding and non-coding segments of the *Escherichia coli* genome. Nucleic Acids Res 34, 4642–4652.

Gray, Y. H. M. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet 16(10), 461–468.

Grigoriev, A (1998). Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 26, 2286–2290.

Griswold, A (2008). Genome packaging in prokaryotes: The circular chromosome of *E. coli*. Nat Educat 1(1), 57.

Guijo, M. I., Patte, J, del Mar Campos, M, Louarn, J.-M., and Rebollo, J. E. (2001). Localized remodeling of the *Escherichia coli* chromosome: the patchwork of segments refractory and tolerant to inversion near the replication terminus. Genetics 157(4), 1413–1423.

Guo, F. B. (2011). [Strong strand specific composition bias-a genomic character of some obligate parasites or symbionts]. Yi Chuan 33, 1039–1047.

Gustafsson, C, Govindarajan, S, and Minshull, J (2004). Codon bias and heterologous protein expression. Trends Biotechnol 22(7), 346–353.

Gutman, G. A. and Hatfield, G. W. (1989). Nonrandom utilization of codon pairs in *Escherichia coli*. Proc Natl Acad Sci 86(10), 3699–3703.

Haffter, P and Bickle, T. A. (1987). Purification and DNA-binding properties of FIS and Cin, two proteins required for the bacteriophage P1 site-specific recombination system, cin. J Mol Biol 198, 579–587.

Hanage, W. P. (2016). Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harbor perspectives in biology, a018069.

Hanage, W. P., Fraser, C, Tang, J, Connor, T. R., and Corander, J (2009). Hyper-recombination, diversity, and antibiotic resistance in *pneumococcus*. Science 324(5933), 1454–1457.

Hartman, JL, t., Garvik, B, and Hartwell, L (2001). Principles for the buffering of genetic variation. Science 291, 1001–1004.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L, et al. (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406(6795), 477–483.

Helmstetter, C. E. (1996). Timing of synthetic activities in the cell cycle. Escherichia coli and Salmonella typhimurium: cellular and molecular biology Washington, p. 1627–1649.

Hendrickson, H and Lawrence, J. G. (2006). Selection for chromosome architecture in bacteria. J Mol Evol 62, 615–629.

Hendrickson, H. L., Barbeau, D, Ceschin, R, and Lawrence, J. G. (2018). Chromosome architecture constrains horizontal gene transfer in bacteria. PLoS Genet 14(5), e1007421.

Herskowitz, I and Oshima, Y (1981). The molecular biology of the yeast saccharomyces, vol. 1.

Hicks, J. B. and Herskowitz, I (1976). Interconversion of Yeast Mating Types I. Direct Observations of the Action of the Homothallism (HO) Gene. Genetics 83, 245–258.

Higashi, K, Tobe, T, Kanai, A, Uyar, E, Ishikawa, S, S uzuki, Y, Ogasawara, N, Kurokawa, K, and Oshima, T (2016). H-NS Facilitates Sequence Diversification of Horizontally Transferred DNA s during Their Integration in Host Chromosomes. PLoS Genet 12, e1005796.

Hill, C. W. and Gray, J. A. (1988). Effects of chromosomal inversion on cell fitness in *Escherichia coli K-12*. Genetics 119(4), 771–778.

Ho, S. Y. W., Lanfear, R, Bromham, L, Phillips, M. J., Soubrier, J, Rodrigo, A. G., and Cooper, A (2011). Time-dependent rates of molecular evolution. Mol Ecol 20(15), 3087–3101.

Hommais, F, Krin, E, Laurent-Winter, C, Soutourina, O, Malpertuy, A, LeCaer, J. P., Danchin, A, and Bertin, P (2001). Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. Mol Microbiol 40, 20–36.

Honarvar, S, Choi, B.-K., and Schifferli, D. M. (2003). Phase variation of the 987P-like CS18 fimbriae of human enterotoxigenic Escherichia coli is regulated by site-specific recombinases. Mol Microbiol 48(1), 157–171.

Hudson, R. E., Bergthorsson, U, Roth, J. R., and Ochman, H (2002). Effect of chromosome location on bacterial mutation rates. Mol Biol Evol 19, 85–92.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al. (2000). Functional discovery via a compendium of expression profiles. Cell 102(1), 109–126.

Husnik, F and McCutcheon, J. P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol 16(2), 67.

Huynen, M. A., Snel, B, and Bork, P (2001). Inversions and the dynamics of eukaryotic gene order. Trends Genet 17, 304–306.

Hyrien, O, Rappailles, A, Guilbaud, G, Baker, A, Chen, C.-L., Goldar, A, Petryk, N, Kahli, M, Ma, E, and D'Aubenton-Carafa, Y (2013). From simple bacterial and archaeal replicons to replication N/U-domains. J Mol Bio 425(23), 4673–4689.

Ikeda, H, Ishikawa, J, Hanamoto, A, Shinose, M, Kikuchi, H, Shiba, T, Sakaki, Y, Hattori, M, and Omura, S (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotechnol 21(5), 526–531.

Ikemura, T (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2, 13–34.

Jansen, R and Gerstein, M (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res 28(6), 1481–1488.

Jeong, K. S., Ahn, J, and Khodursky, A. B. (2004). Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. Genome Biol 5(11), R86.

Jiao, J, Ni, M, Zhang, B, Zhang, Z, Young, J. P. W., Chan, T. F., Chen, W. X., Lam, H. M., and Tian, C. F. (2018). Coordinated regulation of core and accessory genes in the multipartite genome of *Sinorhizobium fredii*. PLoS Genet 14, e1007428.

Jinks-Robertson, S and Nomura, M. (1982). Ribosomal protein S4 acts in trans as a translational repressor to regulate expression of the alpha operon in *Escherichia coli*. J Bacteriol 151(1), 193–202.

Johansson, J, Balsalobre, C, Wang, S. Y., Urbonaviciene, J, Jin, D. J., Sonden, B, and Uhlin, B. E. (2000). Nucleoid proteins stimulate stringently controlled bacterial promoters: a link between the cAMP-CRP and the (p)ppGpp regulons in *Escherichia coli*. Cell 102, 475–485.

Johnson, R. C., Bruist, M. F., and Simon, M. I. (1986). Host protein requirements for in vitro site-specific DNA inversion. Cell 46, 531–539.

Johnson, R. C. and Simon, M. I. (1985). Hin-mediated site-specific recombination requires two 26 bp recombination sites and a 60 bp recombinational enhancer. Cell 41, 781–791.

Juhas, M, Van Der Meer, J. R., Gaillard, M, Harding, R. M., Hood, D. W., and Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33(2), 376–393.

Junier, I (2014). Conserved patterns in bacterial genomes: A conundrum physically tailored by evolutionary tinkering. Comp Biol Chem 53, 125–133.

Juurik, T, Ilves, H, Yeras, R, Ilmjarv, T, Tavita, K, Ukkivi, K, Teppo, A, Mikkel, K, and Kivisaar, M (2012). Mutation frequency and spectrum of mutations vary at different chromosomal positions of *Pseudomonas putida*. PLOS 7, e48511.

Kahmann, R, Rudt, F, Koch, C, and Mertens, G (1985). G inversion in bacteriophage Mu DNA is stimulated by a site within the invertase gene and a host factor. Cell 41, 771–780.

Kahramanoglou, C, Seshasayee, A. S., Prieto, A. I., Ibberson, D, Schmidt, S, Zimmermann, J, Benes, V, Fraser, G. M., and Luscombe, N. M. (2011). Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. Nucleic Acids Res 39, 2073–2091.

Kanaya, S, Yamada, Y, Kudo, Y, and Ikemura, T (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.

Karcagi, I, Draskovits, G, Umenhoffer, K, Fekete, G, Kovács, K, Méhi, O, Balikó, G, Szappanos, B, Györfy, Z, Fehér, T, et al. (2016). Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. Mol Bio Evol 33(5), 1257–1269.

Karlin, S (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends in Microbiol 9(7), 335–343.

Katoh, K, Misawa, K, Kuma, K, and Miyata, T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14), 3059–3066.

Kelly, A, Goldberg, M. D., Carroll, R. K., Danino, V, Hinton, J. C. D., and Dorman, C. J. (2004). A global role for Fis in the transcriptional control of metabolism and type III secretion in *Salmonella enterica serovar Typhimurium*. Microbiology (Reading) 150, 2037–2053.

Képes, F (2004). Periodic transcriptional organization of the *E. coli genome*. J Mol Bio 340(5), 957–964.

Khedkar, S. and Seshasayee, A. S. N. (2016). Comparative genomics of interreplichore translocations in bacteria: a measure of chromosome topology? G3 6(6), 1597–1606.

Kirby, R (2011). Chromosome diversity and similarity within the *Actinomycetales*. FEMS Microbiol Lett 319(1), 1–10.

Kono, N, Arakawa, K, and Tomita, M (2011). Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes. BMC Genomics 12, 19.

Koonin, E. V. (2009). Evolution of genome architecture. Int J Biochem Cell Biol 41, 298–306.

Koonin, E. V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. F1000Research 5.

Kopejtka, K, Lin, Y, Jakubovičová, M, Koblížek, M, and Tomasch, J (2019). Clustered core-and pan-genome content on *Rhodobacteraceae* chromosomes. Genome Biol Evol 11(8), 2208–2217.

Korneev, S and O'Shea, M (2002). Evolution of nitric oxide synthase regulatory genes by DNA inversion. Mol Biol Evol 19, 1228–1233.

Kosmidis, K, Jablonski, K. P., Muskhelishvili, G, and Hutt, M. T. (2020). Chromosomal origin of replication coordinates logically distinct types of bacterial genetic regulation. NPJ Syst Biol Appl 6, 5.

Kowalczykowski, S. C. (2015). An overview of the molecular mechanisms of recombinational DNA repair. Cold Spring Harbor Perspec Biol 7(11), a016410.

Kresse, A. U., Dinesh, S. D., Larbig, K, and Römling, U (2003). Impact of large chromosomal inversions on the adaptation and evolution of Pseudomonas aeruginosa chronically colonizing cystic fibrosis lungs. Mol Microbiol 47(1), 145–158.

Krinos, C. M., Coyne, M. J., Weinacht, K. G., Tzianabos, A. O., Kasper, D. L., and Comstock, L. E. (2001). Extensive surface diversity of a commensal microorganism by multiple DNA inversions. Nature 414(6863), 555.

Krishnamoorthy, K. and Lee, M. (2014). Improved tests for the equality of normal coefficients of variation. Computational Statistics 29(1-2), 215–232.

Kunkel, T. A. (2004). DNA replication fidelity. J Biol Chem 279(17), 16895–16898.

Kunst, F, Ogasawara, N, Moszer, I, Albertini, A. M., Alloni, G, Azevedo, V, Bertero, M. G., Bessieres, P, Bolotin, A, Borchert, S, et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

Kuwahara, T, Yamashita, A, Hirakawa, H, Nakayama, H, Toh, H, Okada, N, Kuhara, S, Hattori, M, Hayashi, T, and Ohnishi, Y (2004). Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. Proc Natl Acad Sci 101(41), 14919–14924.

Lamberte, L. E., Baniulyte, G, Singh, S. S., Stringer, A. M., Bonocora, R. P., Stracy, M, Kapanidis, A. N., Wade, J. T., and Grainger, D. C. (2017). Horizontally acquired AT-rich genes in *Escherichia coli* cause toxicity by sequestering RNA polymerase. Nat Microbiol 2, 16249.

Lamont, G. S., Tucker, R. S., and Cross, G. A. (1986). Analysis of antigen switching rates in *Trypanosoma brucei*. Parasitology 92 (Pt 2), 355–367.

Lang, B, Blot, N, Bouffartigues, E, Buckle, M, Geertz, M, Gualerzi, C. O., Mavathur, R, Muskhelishvili, G, Pon, C. L., Rimsky, S, et al. (2007). High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. Nucleic Acids Res 35, 6330–6337.

Lassalle, F, Périan, S, Bataillon, T, Nesme, X, Duret, L, and Daubin, V (2015). GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet 11(2), e1004941.

Lato, D. F. and Golding, G. B. (2020a). Spatial patterns of gene expression in bacterial genomes. J Mol Evol 88, 510–520.

Lato, D. F. and Golding, G. B. (2020b). The Location of Substitutions and Bacterial Genome Arrangements. Genome Biol Evol.

Le, T. B. K. and Laub, M. T. (2014). New approaches to understanding the spatial organization of bacterial genomes. Curr Opin Microbiol 22, 15–21.

Le Bourgeois, P, Lautier, M, Van Den Berghe, L, Gasson, M. J., and Ritzenthaler, P (1995). Physical and genetic map of the *Lactococcus lactis subsp. cremoris MG1363* chromosome: comparison with that of *Lactococcus lactis subsp. lactis IL 1403* reveals a large genome inversion. J Bacteriolo 177(10), 2840–2850.

Lee, H, Doak, T. G., Popodi, E, Foster, P. L., and Tang, H (2016). Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. Nucleic Acids Res 44(15), 7109–7119.

Lee, H., Popodi, E., Tang, H., and Foster, P. L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci 109(41), E2774–E2783.

Li, J. W., Li, J, Wang, J, Li, C, and Zhang, J. R. (2019). Molecular Mechanisms of hsdS Inversions in the cod Locus of *Streptococcus pneumoniae*. J Bacteriol 201.

Li, W. H., Wu, C. I., and Luo, C. C. (1985). A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2, 150–174.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13(5), 660–665.

Loconto, J, Viswanathan, P, Nowak, S. J., Gloudemans, M, and Kroos, L (2005). Identification of the omega4406 regulatory region, a developmental promoter of *Myxococcus xanthus*, and a DNA segment responsible for chromosomal position-dependent inhibition of gene expression. J Bacteriol 187(12), 4149–4162.

Long, H, Miller, S. F., Strauss, C, Zhao, C, Cheng, L, Ye, Z, Griffin, K, Te, R, Lee, H, Chen, C, et al. (2016). Antibiotic treatment enhances the genome-wide mutation rate of target cells. Proc Natl Acad Sci 113(18), E2498–E2505.

Long, H, Sung, W, Miller, S. F., Ackerman, M. S., Doak, T. G., and Lynch, M (2014). Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient *Pseudomonas fluorescens* ATCC948. Genome Biol Evol 7(1), 262–271.

Louarn, J. M., Bouche, J. P., Legendre, F, Louarn, J, and Patte, J (1985). Characterization and properties of very large inversions of the *E. coli* chromosome along the origin-to-terminus axis. Mol Gen Genet MGG 201(3), 467–476.

Lusetti, S. L. and Cox, M. M. (2002). The bacterial RecA protein and the recombinational DNA repair of stalled replication forks. Annu Rev Biochem 71(1), 71–100.

Mackiewicz, P, Gierlik, A, Kowalczuk, M, Dudek, M. R., and Cebrat, S (1999). How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res 9(5), 409–416.

Mackiewicz, P, Mackiewicz, D, Kowalczuk, M, and Cebrat, S (2001). Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol 2(12), interactions1004–1.

Maddamsetti, R and Lenski, R. E. (2018). Analysis of bacterial genomes from an evolution experiment with horizontal gene transfer shows that recombination can sometimes overwhelm selection. PLoS Genet 14(1), e1007199.

Mahan, M. J. and Roth, J. R. (1988). Reciprocality of recombination events that rearrange the chromosome. Genetics 120(1), 23–35.

Mahan, M. J. and Roth, J. R. (1991). Ability of a bacterial chromosome segment to invert is dictated by included material rather than flanking sequence. Genetics 129(4), 1021–1032.

Maharjan, R. P. and Ferenci, T. (2018). The impact of growth rate and environmental factors on mutation rates and spectra in *Escherichia coli*. Enviro Microbiol Reports 10(6), 626–633.

Mao, X, Zhang, H, Yin, Y, and Xu, Y (2012). The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res 40, 8210–8218.

Marczynski, G. T., Rolain, T, and Taylor, J. A. (2015). Redefining bacterial origins of replication as centralized information processors. Fronti Microbiol 6, 610.

Marin, A, Gallardo, M, Kato, Y, Shirahige, K, Gutiérrez, G, Ohta, K, and Aguilera, A (2003). Relationship between G+ C content, ORF-length and mRNA concentration in *Saccharomyces cerevisiae*. Yeast 20(8), 703–711.

Marin, A and Xia, X (2008). GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J Theor Biol 253, 508–513.

Marrs, C. F., Ruehl, W. W., Schoolnik, G. K., and Falkow, S (1988). Pilin-gene phase variation of *Moraxella bovis* is caused by an inversion of the pilin genes. J Bacteriol 170(7), 3032–3039.

Martens, M, Dawyndt, P, Coopman, R, Gillis, M, De Vos, P, and Willems, A (2008). Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). Intern Syst Evol Microbiol 58(1), 200–214.

Martin-Didonet, C. C., Chubatsu, L. S., Souza, E. M., Kleina, M, Rego, F. G., Rigo, L. U., Yates, M. G., and Pedrosa, F. O. (2000). Genome structure of the genus *Azospirillum*. J Bacteriol 182, 4113–4116.

Martina, M. A., Correa, E. M. E., Argaraña, C. E., and Barra, J. L. (2012). *Escherichia coli* frameshift mutation rate depends on the chromosomal context but not on the GATC content near the mutation site. PLoS One 7(3), e33701.

Martinez-Antonio, A, Medina-Rivera, A, and Collado-Vides, J (2009). Structural and functional map of a bacterial nucleoid. Genome Biol 10, 247.

Marwick, B. and Krishnamoorthy, K. (2019). cvequality.

McInerney, J. O., McNally, A, and O'connell, M. J. (2017). Why prokaryotes have pangenomes. Nat Microbiol 2(4), 17040.

McLean, M. J., Wolfe, K. H., and Devine, K. M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47(6), 691–696.

McVean, G, Awadalla, P, and Fearnhead, P (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160(3), 1231–1241.

Meadows, L. A., Chan, Y. S., Roote, J, and Russell, S (2010). Neighbourhood continuity is not required for correct testis gene expression in *Drosophila*. PLoS Biol 8, e1000552.

Medini, D, Donati, C, Tettelin, H, Masignani, V, and Rappuoli, R (2005). The microbial pan-genome. Curr Opin Genet Dev 15, 589–594.

Medini, D, Serruto, D, Parkhill, J, Relman, D. A., Donati, C, Moxon, R, Falkow, S, and Rappuoli, R (2008). Microbiology in the post-genomic era. Nat Rev Microbiol 6(6), 419–430.

Merrikh, C. N. and Merrikh, H (2018). Gene inversion increases evolvability in bacteria. bioRxiv, 293571.

Messer, W, Egan, B, Gille, H, Holz, A, Schaefer, C, and Woelker, B (1991). The complex of oriC DNA with the DnaA initiator protein. Res Microbiol 142, 119–125.

Meyer, S, Reverchon, S, Nasser, W, and Muskhelishvili, G (2018). Chromosomal organization of transcription: in a nutshell. Curr Genet 64, 555–565.

Miller, O. L., Hamkalo, B. A., and Thomas, C. A. (1970). Visualization of bacterial genes in action. Science 169, 392–395.

Miller, W. G. and Simons, R. W. (1993). Chromosomal supercoiling in *Escherichia coli*. Mol Microbiol 10(3), 675–684.

Mira, A, Martin-Cuadrado, A. B., D'Auria, G, and Rodriguez-Valera, F (2010). The bacterial pan-genome:a new paradigm in microbiology. Intl Microbiol 13(2), 45–57.

Mira, A and Ochman, H (2002). Gene location and bacterial sequence divergence. Mol Biol Evol 19, 1350–1358.

Mirkin, E. V. and Mirkin, S. M. (2005). Mechanisms of transcription-replication collisions in bacteria. Mol Cell Biol 25, 888–895.

Moran, N. A. and Mira, A (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol 2, RESEARCH0054.

Morrow, J. D. and Cooper, V. S. (2012). Evolutionary effects of translocations in bacterial genomes. Genom Biol Evol 4(12), 1256–1262.

Nakagawa, I., Kurokawa, K., Yamashita, A., Nakata, M., Tomiyasu, Y., Okahashi, N., Kawabata, S., Yamazaki, K., Shiba, T., Yasunaga, T., et al. (2003). Genome sequence of an M3 strain

of *Streptococcus pyogenes* reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. Genome Res 13(6a), 1042–1055.

Naseeb, S, Carter, Z, Minnis, D, Donaldson, I, Zeef, L, and Delneri, D (2016). Widespread impact of chromosomal inversions on gene expression uncovers robustness via phenotypic buffering. Mol Biol Evol 33(7), 1679–1696.

Navarre, W. W., McClelland, M, Libby, S. J., and Fang, F. C. (2007). Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev 21, 1456–1471.

Niccum, B. A., Lee, H, MohammedIsmail, W, Tang, H, and Foster, P. L. (2019). The Symmetrical Wave Pattern of Base-Pair Substitution Rates across the *Escherichia coli* Chromosome Has Multiple Causes. mBio 10(4), e01226–19.

Novichkov, P. S., Wolf, Y. I., Dubchak, I, and Koonin, E. V. (2009). Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol 191(1), 65–73.

Nowell, R. W., Green, S, Laue, B. E., and Sharp, P. M. (2014). The extent of genome flux and its role in the differentiation of bacterial lineages. Genom Biol Evol 6(6), 1514–1529.

Ochman, H (2003). Neutral mutations and neutral substitutions in bacterial genomes. Mol Biol Evol 20(12), 2091–2096.

Ochman, H, Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784), 299.

Ohta, T (1992). The nearly neutral theory of molecular evolution. Ann Rev Ecol System 23(1), 263–286.

Oliveira, P. H., Touchon, M, Cury, J, and Rocha, E. P. C. (2017). The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun 8(1), 841.

Oshima, T, Ishikawa, S, Kurokawa, K, Aiba, H, and Ogasawara, N (2006). *Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. DNA Res 13, 141–153.

Parkhill, J, Sebaihia, M, Preston, A, Murphy, L. D., Thomson, N, Harris, D. E., Holden, M. T. G., Churcher, C., Bentley, S. D., and Mungall, K. L. (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nature Genet 35(1), 32.

Parkhill, J, Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M, James, K. D., Churcher, C, Mungall, K. L., et al. (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413, 523–527.

Patrick, S, Parkhill, J, McCoy, L. J., Lennard, N, Larkin, M. J., Collins, M, Sczaniecka, M, and Blakely, G (2003). Multiple inverted DNA repeats of *Bacteroides fragilis* that control polysaccharide antigenic variation are similar to the hin region inverted repeats of *Salmonella typhimurium*. Microbiol 149(4), 915–924.

Paul, B. J., Ross, W, Gaal, T, and Gourse, R. L. (2004). rRNA transcription in *Escherichia coli*. Annu Rev Genet 38, 749–770.

Penny, D (2015). Cooperation and selfishness both occur during molecular evolution. Biol Dir 10(1), 26.

Perriere, G, Lobry, J. R., and Thioulouse, J (1996). Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. Comput Appl Biosci 12, 519–524.

Peter, B. J., Arsuaga, J, Breier, A. M., Khodursky, A. B., Brown, P. O., and Cozzarelli, N. R. (2004). Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. Genome Biol 5(11), R87.

Pinto, U. M., Flores-Mireles, A. L., Costa, E. D., and Winans, S. C. (2011). RepC protein of the octopine-type Ti plasmid binds to the probable origin of replication within *repC* and functions only *in cis*. Mol Microbiol 81(6), 1593–1606.

Prescott, D. M. and Kuempel, P. L. (1972). Bidirectional replication of the chromosome in *Escherichia coli*. Proc Natl Acad Sci 69(10), 2842–2845.

Price, M. N., Alm, E. J., and Arkin, A. P. (2005). Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. Nucleic Acids Res 33(10), 3224–3234.

Quax, T. E. F., Claassens, N. J., Söll, D, and Oost, J van der (2015). Codon bias as a means to fine-tune gene expression. Mol Cell 59(2), 149–161.

R Development Core Team (2014). R: a language and environment for statistical computing. Vienna, Austria.

Raeside, C, Gaffé, J, Deatherage, D. E., Tenaillon, O, Briska, M, Ptashkin, R. N., Cruveiller, S, Médigue, C, Lenski, R. E., and Barrick, J. E. (2014). Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. MBio 5(5), e01377–14.

Rangarajan, A. A. and Schnetz, K (2018). Interference of transcription across H-NS binding sites and repression by H-NS. Mol Microbiol 108, 226–239.

Ravenhall, M, Škunca, N, Lassalle, F, and Dessimoz, C (2015). Inferring horizontal gene transfer. PLoS Comp Biol 11(5), e1004095.

Ravichandar, J. D., Bower, A. G., Julius, A. A., and Collins, C. H. (2017). Transcriptional control of motility enables directional movement of *Escherichia coli* in a signal gradient. Sci Rep 7(1), 8959.

Redenbach, M, Kieser, H. M., Denapaite, D, Eichner, A, Cullum, J, Kinashi, H, and Hopwood, D. A. (1996). A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3 (2) chromosome. Mol Microbiol 21(1), 77–96.

Reif, H. J. and Saedler, H (1975). IS1 is involved in deletion formation in the *gal* region of *E. coli K12*. Mol Gen Genet MGG 137(1), 17–28.

Rentschler, A. E., Lovrich, S. D., Fitton, R, Enos-Berlage, J, and Schwan, W. R. (2013). OmpR regulation of the uropathogenic *Escherichia coli* fimB gene in an acidic/high osmolality environment. Microbiol 159(Pt 2), 316.

Repar, J and Warnecke, T (2017). Non-random inversion landscapes in prokaryotic genomes are shaped by heterogeneous selection pressures. Mol Biol Evol 34(8), 1902–1911.

Reynolds, A. E., Felton, J, and Wright, A (1981). Insertion of DNA activates the cryptic bgl operon in *E. coli K12*. Nature 293(5834), 625.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinfor 26(1), 139–140.

Robinson, M. D. and Oshlack, A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11(3), R25.

Rocha, E (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10, 393–395.

Rocha, E. P. and Danchin, A (2002). Base composition bias might result from competition for metabolic resources. Trends Genet 18, 291–294.

Rocha, E. P. and Danchin, A (2003). Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34, 377–378.

Rocha, E. P. and Danchin, A (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21, 108–116.

Rocha, E. P. C. (2004a). Order and disorder in bacterial genomes. Curr Opin Microbiol 7(5), 519–527.

Rocha, E. P. C. (2004b). The replication-related organization of bacterial genomes. Microbiol 150(6), 1609–1627.

Rocha, E. P. C. (2008). The organization of the bacterial genome. Annu Rev Genet 42, 211–233.

Romling, U, Schmidt, K. D., and Tummler, B (1997). Large chromosomal inversions occur in *Pseudomonas aeruginosa* clone C strains isolated from cystic fibrosis patients. FEMS Microbiol Lett 150, 149–156.

Roth, J. R., Benson, N, Galitski, T, Haack, K, Lawrence, J. G., and Miesel, L (1996). Rearrangements of the bacterial chromosome: formation and applications. Escherichia coli and Salmonella: cellular and molecular biology 2, 2256–2276.

Ryan, V. T., Grimwade, J. E., Camara, J. E., Crooke, E, and Leonard, A. C. (2004). *Escherichia coli* prereplication complex assembly is regulated by dynamic interplay among Fis, IHF and DnaA. Mol Microbiol 51, 1347–1359.

Saedler, H, Cornelis, G, Cullum, J, Schumacher, B, and Sommer, H (1981). IS1-mediated DNA rearrangements. In: Cold Spring Harbor symposia on quantitative biology. Vol. 45. Cold Spring Harbor Laboratory Press, 93–98.

Sauer, C, Syvertsson, S, Bohorquez, L. C., Cruz, R, Harwood, C. R., Rij, T van, and Hamoen, L. (2016). Effect of genome position on heterologous gene expression in *Bacillus subtilis*: an unbiased analysis. ACS Syn Biol 5(9), 942–947.

Schmid, M. B. and Roth, J. R. (1983). Selection and endpoint distribution of bacterial inversion mutations. Genetics 105(3), 539–557.

Schmid, M. B. and Roth, J. R. (1987). Gene location affects expression level in *Salmonella typhimurium*. J Bacteriol 169(6), 2872–2875.

Schneider, D, Duperchy, E, Coursange, E, Lenski, R. E., and Blot, M (2000). Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. Genetics 156(2), 477–488.

Schneider, D and Lenski, R. E. (2004). Dynamics of insertion sequence elements during experimental evolution of bacteria. Res Microbiol 155(5), 319–327.

Schneider, R, Lurz, R, Luder, G, Tolksdorf, C, Travers, A, and Muskhelishvili, G (2001). An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. Nucleic Acids Res 29, 5107–5114.

Schneider, R, Travers, A, and Muskhelishvili, G (1997). FIS modulates growth phase-dependent topological transitions of DNA in *Escherichia coli*. Mol Microbiol 26, 519–530.

Scholz, S. A., Diao, R, Wolfe, M. B., Fivenson, E. M., Lin, X. N., and Freddolino, P. L. (2019). High-Resolution Mapping of the *Escherichia coli* Chromosome Reveals Positions of High and Low Transcription. Cell Syst 8, 212–225.e9.

Schrider, D. R., Hourmozdi, J. N., and Hahn, M. W. (2011). Pervasive multinucleotide mutational events in eukaryotes. Curr Biolo 21(12), 1051–1054.

Segall, A, Mahan, M. J., and Roth, J. R. (1988). Rearrangement of the bacterial chromosome: forbidden inversions. Science 241(4871), 1314–1318.

Segall, A. M. and Roth, J. R. (1989). Recombination between homologies in direct and inverse orientation in the chromosome of *Salmonella*: intervals which are nonpermissive for inversion formation. Genetics 122(4), 737–747.

Sekulovic, O, Garrett, E. M., Bourgeois, J, Tamayo, R, Shen, A, and Camilli, A (2018). Genome-wide detection of conservative site-specific recombination in bacteria. PLoS Genet 14(4), e1007332.

Senra, M. V. X., Sung, W, Ackerman, M, Miller, S. F., Lynch, M, and Soares, C. A. G. (2018). An unbiased genome-wide view of the mutation rate and spectrum of the endosymbiotic bacterium *Teredinibacter turnerae*. Genome Biol Evol 10(3), 723–730.

Serkin, C. D. and Seifert, H. S. (2000). Iron availability regulates DNA recombination in *Neisseria gonorrhoeae*. Mol Microbiol 37(5), 1075–1086.

Sharp, P. M., Bailes, E, Grocock, R. J., Peden, J. F., and Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33(4), 1141–1153.

Sharp, P. M. and Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1-2), 28–38.

Sharp, P. M., Shields, D. C., Wolfe, K. H., and Li, W.-H. (1989). Chromosomal location and evolutionary rate variation in *Enterobacterial* genes. Science 246, 808–810.

Shimodaira, H and Hasegawa, M (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16(8), 1114.

Sibley, M. H. and Raleigh, E. A. (2004). Cassette-like variation of restriction enzyme genes in *Escherichia coli* C and relatives. Nucleic Acids Res 32(2), 522–534.

Siguier, P, Gourbeyre, E, and Chandler, M (2014). Bacterial insertion sequences: Their genomic impact and diversity. FEMS Microbiol Rev 38(5).

Singh, K, Milstein, J. N., and Navarre, W. W. (2016). Xenogeneic Silencing and Its Impact on Bacterial Genomes. Annu Rev Microbiol 70, 199–213.

Singh, S. S. and Grainger, D. C. (2013). H-NS can facilitate specific DNA-binding by RNA polymerase in AT-rich gene regulatory regions. PLoS Genet 9, e1003589.

Singh, S. S., Singh, N, Bonocora, R. P., Fitzgerald, D. M., Wade, J. T., and Grainger, D. C. (2014). Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev 28, 214–219.

Somvanshi, V. S., Sloup, R. E., Crawford, J. M., Martin, A. R., Heidt, A. J., Kim, K, Clardy, J, and Ciche, T. A. (2012). A single promoter inversion switches *Photorhabdus* between pathogenic and mutualistic states. Science 337(6090), 88–93.

Soucy, S. M., Huang, J, and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. Nat Rev Genet 16(8), 472.

Sousa, C, Lorenzo, V de, and Cebolla, A (1997). Modulation of gene expression through chromosomal positioning in *Escherichia coli*. Microbiol 143(6), 2071–2078.

Stamatakis, A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinfor 30(9), 1312–1313.

Streisinger, G, Okada, Y, Emrich, J, Newton, J, Tsugita, A, Terzaghi, E, and Inouye, M (1966). Frameshift mutations and the genetic code. In: Cold Spring Harbor Symposia on Quantitative Biology. Vol. 31. Cold Spring Harbor Laboratory Press, 77–84.

Sun, S, Ke, R, Hughes, D, Nilsson, M, and Andersson, D. I. (2012). Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. PloS one 7(8), e42639.

Sung, W, Ackerman, M. S., Gout, J.-F., Miller, S. F., Williams, E, Foster, P. L., and Lynch, M (2015). Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. Mol Bio Evol 32(7), 1672–1683.

Suyama, M. and Bork, P. (2001). Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends in Genet 17(1), 10–13.

Suyama, M., Bork, P., Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O, Hickey, E. K., Peterson, J, et al. (2000). Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends in Genet 28(6), 10–13.

Syvanen, M (2012). Evolutionary implications of horizontal gene transfer. Annu RevGenet 46, 341–358.

Tenaillon, O, Skurnik, D, Picard, B, and Denamur, E (2010). The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol 8(3), 207.

Tettelin, H, Riley, D, Cattuto, C, and Medini, D (2008). Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11, 472–477.

Than, C, Ruths, D, Innan, H, and Nakhleh, L (2007). Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. J Comp Biol 14(4), 517–535.

Thomas, C. M. and Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3(9), 711.

Thompson, J. F., MoitosodeVargas, L, Koch, C, Kahmann, R, and Landy, A (1987). Cellular factors couple recombination with growth phase: characterization of a new component in the lambda site-specific recombination pathway. Cell 50, 901–908.

Tidjani, A.-R., Lorenzi, J.-N., Toussaint, M, Dijk, E van, Naquin, D, Lespinet, O, Bontemps, C, and Leblond, P (2019). Massive gene flux drives genome diversity between sympatric *Streptomyces* conspecifics. mBio 10(5), e01533–19.

Tillier, E. R. and Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. Nat Genet 26(2), 195–197.

Touchon, M, Hoede, C, Tenaillon, O, Barbe, V, Baeriswyl, S, Bidet, P, Bingen, E, Bonacorsi, S, Bouchier, C, Bouvet, O, et al. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5(1), e1000344.

Touzain, F, Petit, M. A., Schbath, S, and ElKaroui, M (2011). DNA motifs that sculpt the bacterial chromosome. Nat Rev Microbiol 9, 15–26.

Trapnell, C, Roberts, A, Goff, L, Pertea, G, Kim, D, Kelley, D. R., Pimentel, H, Salzberg, S. L., Rinn, J. L., and Pachter, L (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578.

Travers, A, Schneider, R, and Muskhelishvili, G (2001). DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. Biochimie 83, 213–217.

Treangen, T. J., Ondov, B. D., Koren, S, and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15, 524.

Tsai, M. Y., Zheng, W, Chen, M, and Wolynes, P. G. (2019). Multiple Binding Configurations of Fis Protein Pairs on DNA: Facilitated Dissociation versus Cooperative Dissociation. J Am Chem Soc 141, 18113–18126.

Ueda, T, Takahashi, H, Uyar, E, Ishikawa, S, Ogasawara, N, and Oshima, T (2013). Functions of the Hha and YdgT proteins in transcriptional silencing by the nucleoid proteins, H-NS and StpA, in *Escherichia coli*. DNA Res 20, 263–271.

Vickerman, K (1978). Antigenic variation in trypanosomes. Nature 273, 613–617.

Wade, J. T. and Grainger, D. C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol 12, 647–653.

Wang, X, Llopis, P. M., and Rudner, D. Z. (2013). Organization and segregation of bacterial chromosomes. Nat Rev Genet 14(3), 191.

Warnecke, T, Supek, F, and Lehner, B (2012). Nucleoid-associated proteins affect mutation dynamics in *E. coli* in a growth phase-specific manner. PLoS Comput Biol 8, e1002846.

Washburn, R. S. and Gottesman, M. E. (2011). Transcription termination maintains chromosome integrity. Proc Natl Acad Sci 108(2), 792–797.

Wei, W, Xiong, L, Ye, Y.-N., Du, M.-Z., Gao, Y.-Z., Zhang, K.-Y., Jin, Y.-T., Yang, Z, Wong, P.-C., Lau, S. K. P., et al. (2018). Mutation Landscape of Base Substitutions, Duplications, and Deletions in the Representative Current *Cholera* Pandemic Strain. Genome Biol Evol 10(8), 2072–2085.

Wickham, H (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the {tidyverse}. Journal of Open Source Software 4(43), 1686.

Wiedenbeck, J and Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35(5), 957–976.

Wold, S, Crooke, E, and Skarstad, K (1996). The *Escherichia coli* Fis protein prevents initiation of DNA replication from oriC in vitro. Nucleic Acids Res 24, 3527–3532.

Wolf, Y. I. and Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. Bioessays 35(9), 829–837.

Wong, S and Wolfe, K. H. (2005). Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet 37, 777–782.

Wright, M. A., Kharchenko, P, Church, G. M., and Segrè, D (2007). Chromosomal periodicity of evolutionarily conserved gene pairs. Proc Natl Acad Sci 104(25), 10559–10564.

Wright, S. I., Lauga, B, and Charlesworth, D (2002). Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol Bio Evol 19(9), 1407–1420.

Wright, S. I., Yau, C. B. K., Looseley, M, and Meyers, B. C. (2004). Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol 21(9), 1719–1726.

Wu, F, Wu, J, Ehley, J, and Filutowicz, M (1996). Preponderance of Fis-binding sites in the R6K gamma origin and the curious effect of the penicillin resistance marker on replication of this origin in the absence of Fis. J Bacteriol 178, 4965–4974.

Yang, J. C., Madupu, R, Durkin, A. S., Ekborg, N. A., Pedamallu, C. S., Hostetler, J. B., Radune, D, Toms, B. S., Henrissat, B, Coutinho, P. M., et al. (2009). The complete genome of *Teredinibacter turnerae T7901*: an intracellular endosymbiont of marine wood-boring *bivalves* (shipworms). PloS one 4(7), e6085.

Yang, Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinfor 13(5), 555–556.

Yang, Z and Dos Reis, M (2010). Statistical properties of the branch-site test of positive selection. Mol Bio Evol 28(3), 1217–1228.

Yates, J. L., Arfsten, A. E., and Nomura, M (1980). In vitro expression of *Escherichia coli* ribosomal protein genes: autogenous inhibition of translation. Proc Natl Acad Sci 77(4), 1837–1841.

Yoshikawa, H, O'Sullivan, A, and Sueoka, N (1964). Sequential Replication of the *Bacillus subtilis* chromosome 3. regulation of initiation. Proc Natl Acad Sci U S A 52, 973–980.

Zeigler, D. R. and Dean, D. H. (1990). Orientation of genes in the *Bacillus subtilis* chromosome. Genetics 125(4), 703–708.

Zheng, W. X., Luo, C. S., Deng, Y. Y., and Guo, F. B. (2015). Essentiality drives the orientation bias of bacterial genes in a continuous manner. Sci Rep 5, 16431.

Zieg, J, Hilmen, M, and Simon, M (1978). Regulation of gene expression by site-specific inversion. Cell 15(1), 237–244.

Zieg, J, Silverman, M, Hilmen, M, and Simon, M (1977). Recombinational switch for gene expression. Science 196, 170–172.

Zimmerman, C.-U. R., Stiedl, T, Rosengarten, R, and Spergser, J (2009). Alternate phase variation in expression of two major surface membrane proteins (MBA and UU376) of *Ureaplasma parvum serovar* 3. FEMS Microbiol Lett 292(2), 187–193.

Zipkas, D and Riley, M (1975). Proposal concerning mechanism of evolution of the genome of *Escherichia coli*. Proc Natl Acad Sci 72(4), 1354–1358.

Zivanovic, Y, Lopez, P, Philippe, H, and Forterre, P (2002). *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res 30, 1902–1910.

Appendix A

Chapter 2 Supplementary Files

Title: The Location of Substitutions and Bacterial Genome Arrangements

Authors: Daniella F. Lato and G. Brian Golding

Journal: Genome Biology and Evolution

Corresponding Author Information:
G. Brian Golding
McMaster University
Department of Biology
1280 Main St. West
Hamilton, ON
Canada
L8S 4K1
Email: golding@mcmaster.ca

Supplementary Material

For the most up to date Supplementary Material, please visit www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements.

Further supplemental information and code are available on GitHub at www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements.

A0.1 Software Version Numbers

Program	Version Number	Build Date
`baseml`	4.9	March 2015
`codeml`	4.9	March 2015
`consense`	3.6b	NA
`dnadist`	3.6b	NA
`dnaml`	3.6b	NA
`MAFFT`	v7.045b	June 5, 2013
`neighbor`	3.6b	NA
`progressiveMauve`	Snap Shot	June 7, 2012
`RAxML`	8.0.25	June 16, 2014
`seqboot`	3.6b	NA
`trimAl`	v1.4.rev15	December 17, 2013

Table S1.1: Version numbers and build dates for each of the programs used.

Bacteria Strain/Species	Accession Number	Date Accessed
Escherichia coli		
E. coli 0104H4	CP003289	September 29, 2016
E. coli 0157H7	BA000007	September 29, 2016
E. coli 083H1	CP001855	September 29, 2016
E. coli IAI39	CU928164	September 26, 2016
E. coli K12 *	U00096	September 26, 2016
E. coli UMN026	CU928163	September 26, 2016
Outgroup: *Escherichia fergusonii* ATCC 35469T	NC_011740	August 26, 2020
Bacillus subtilis		
B. subtilis 168 *	NC_000964	November 10, 2016
B. subtilis BS38	NZ_CP017314	November 11, 2016
B. subtilis BSn5	NC_014976	November 11, 2016
B. subtilis PY79	NC_022898	November 11, 2016
B. subtilis QB928	NC_018520	November 11, 2016
B. subtilis RONN1	NC_017195	November 11, 2016
B. subtilis W23	NC_014479	November 11, 2016
Outgroup: *Bacillus cereus* FDAARGOS_797	NZ_CP053931	August 26, 2020
Streptomyces		
Streptomyces lividans TK24	NZ_GG657756	August 26, 2020
S. lividans 1362	NZ_CM001889	August 26, 2020
Streptomyces coelicolor A3 *	AL645882	November 30, 2016
S. coelicolor A32 CFB NCB	NZ_CP042324	August 26, 2020
S. coelicolor M1154/pAMX4/pGP1416	NZ_CP050522	August 26, 2020
Outgroup: *Streptomyces aureofaciens* DM1	NZ_CP020567	August 26, 2020
Sinorhizobium meliloti Chromosome		
S. meliloti 2011	NC_020528	April 24, 2017
S. meliloti 1021 *	NC_003047	June 3, 2014
S. meliloti AK83	NC_015590	June 3, 2014
S. meliloti BL225C	NC_017322	June 3, 2014
S. meliloti SM11	NC_017325	June 3, 2014
S. meliloti RMO17	NC_CP009144	April 24, 2017
Outgroup: *Rhizobium leguminosarum* trifolii WSM1689 chromosome	NZ_CP007045	August 26, 2020
S. meliloti pSymA		
S. meliloti 2011	NC_020527	April 24, 2017
S. meliloti 1021 *	NC_003037	June 3, 2014
S. meliloti AK83	NC_015591	June 3, 2014
S. meliloti BL225C	NC_017324	June 3, 2014
S. meliloti SM11	NC_017327	June 3, 2014
S. meliloti RMO17	NC_CP009145	April 24, 2017
Outgroup: *R. leguminosarum* trifolii WSM1689 plasmid pRLG202	NC_0113665	August 26, 2020
S. meliloti pSymB		
S. meliloti 2011	NC_020560	April 24, 2017
S. meliloti 1021 *	NC_003078	June 3, 2014
S. meliloti AK83	NC_015596	June 3, 2014
S. meliloti BL225C	NC_017323	June 3, 2014
S. meliloti SM11	NC_017326	June 3, 2014
S. meliloti RMO17	NC_CP009146	April 24, 2017
Outgroup: *R. leguminosarum* trifolii WSM1689 plasmid pRLG201	NC_011368	August 26, 2020

Table S1.2: Strains and species used for each replicon analysis. Accession numbers, date accessed, and outgroups for each replicon are provided. An astrix (*) insicates the strain that was used as the representative strain.

A0.2 Constraints to Number of Sequence Chosen

Computational time constraints and the nature of the data were limiting factors for the number of strains that were chosen for each bacterial species. `progressiveMauve` is a multiple sequence alignment program which is useful for accounting for local and large scale genomic rearrangements. Some of the bacterial strains are very similar and therefore there was no issue finding a sufficient number of Locally Colinear Blocks (LCBs) without having the genomes broken into an overwhelming number of blocks. We had to strike a balance between having as many genomes in the analysis as possible, and comparing correct homologous sequences. The more distantly related the taxa are, the resulting `progressiveMauve` alignment contained shorter LCBs and many blocks that compared sequences of poor homology. This can be seen in an example of six *Streptomyces* genomes resulting in the genome being split into 521 LCBs (Supplementary Figure S1.1). Consequently, we had to reduce the number of genomes used for this analysis and after many iterations of genome combinations, we settled on the genomes listed in Table S1.2. This allowed for the correct comparison of homologous sequences, while also accounting for recombination.

The computational time required to run `progressiveMauve` was an additional constraint that needed to be considered. `progressiveMauve` can align multiple whole genomes and identify regions that have been rearranged within the taxa provided. This process happens in relatively quick computational time, however, like most other programs, the addition of more data increased the amount of time required to complete the process. We ran multiple instances of `progressiveMauve` with varying numbers of *E. coli* genomes. These data points were connected using a locally estimated scatter plot smoothing method and confidence intervals. From this data, we determined that increasing the number of genomes exponentially increases the run time of `progressiveMauve`. It becomes impractical to align more than 27 genomes with `progressiveMauve`, as anything over that would take more than 24h to run. The estimated computational run time to complete the alignment of 100 genomes would take over a month. The total computational time additionally depends on the divergence of the sequences. 26 divergent *Streptomyces* genomes (Table S1.5) took just under a month to complete the `progressiveMauve` alignment. This information combined with `progressiveMauve`'s inability to pair homologous sequences in LCBs of distantly related taxa, has limited the total number of genomes we can use per taxa to a maximum of 7. This provides the most accurate data and the most reasonable analysis duration.

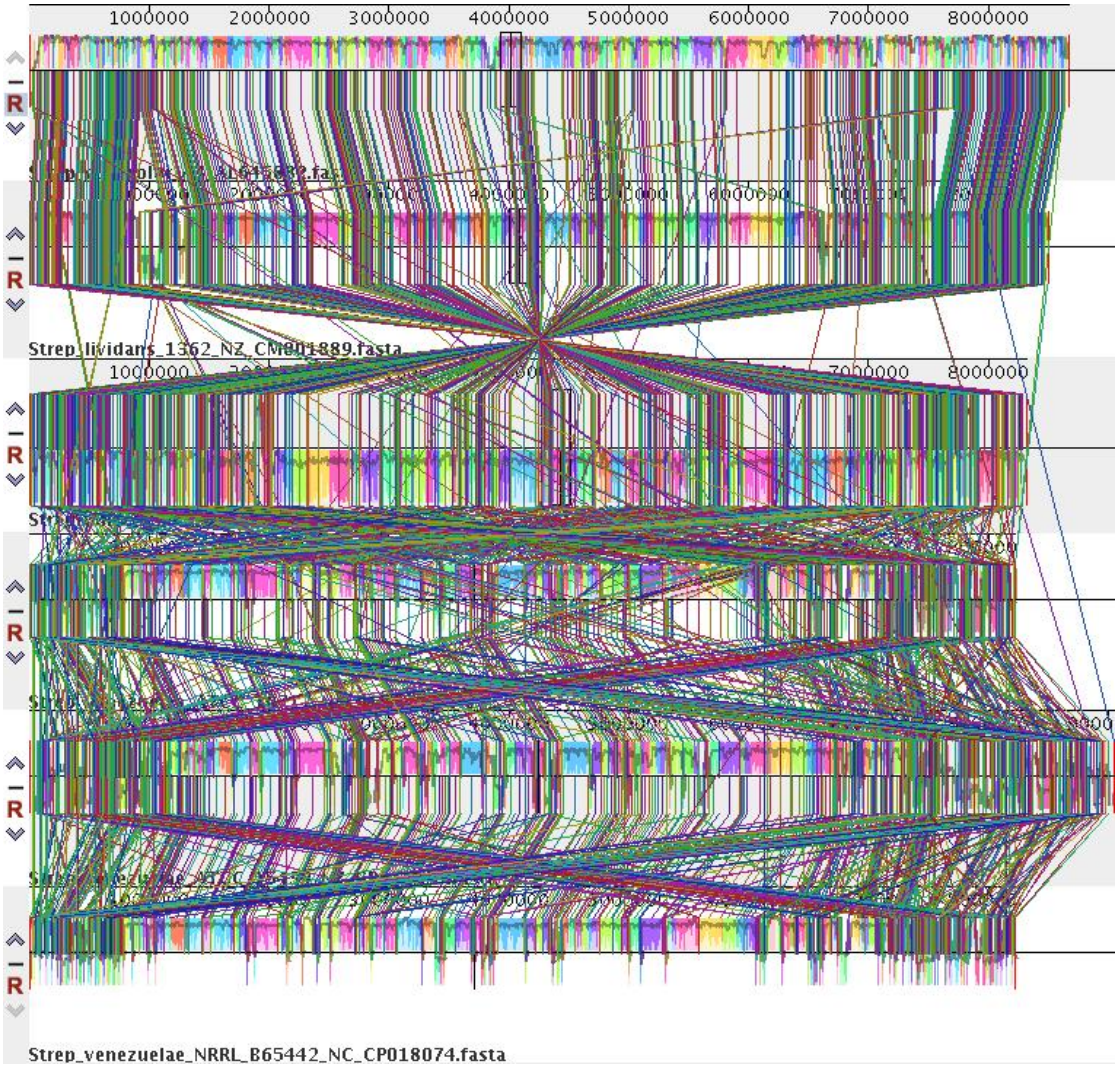


FIGURE S1.1: Visualization of the `progressiveMauve` alignment of 6 *Streptomyces* genomes (from top to bottom): *S. coelicolor* AL645882, *S. lividans* NZ_CM001889, *S. lividans* NZ_GG657756, *S. venezuelae* NC_018750, *S. venezuelae* NZ_CP013129, and *S. venezuelae* NC_CP018074. Each coloured block represents a different LCB. Coloured lines connect LCBs that are similar between taxa. The black lines underneath each LCB represent the whole genome sequence of each of the *Streptomyces* taxa. Each LCB can be treated as a rearrangement, there have therefore been 521 rearrangements between these *Streptomyces* genomes.

A0.3 `progressiveMauve` Alignment

FIGURE S1.2: Visualization of the `progressiveMauve` alignment of the *B. subtilis* genomes. Each coloured block represents a different LCB. Coloured lines connect LCBs that are similar between taxa. The black lines underneath each LCB represent the whole genome sequence of each of the *B. subtilis* taxa. From top to bottom the taxa are: *B. subtilis* PY79, *B. subtilis* QB928, *B. subtilis* 168, *B. subtilis* BSn5, *B. subtilis* BS38, *B. subtilis* RONN1, *B. subtilis* W23. Each LCB can be treated as a rearrangement, there have therefore been 12 rearrangements between these *B. subtilis* genomes.

Strain	NCBI Accession Number	Date Accessed
	E. coli	
K12 MG1655	U00096	June 1, 2020
BL21DE3	CP053602	August 3, 2020
BW25113	CP009273	August 3, 2020
ECC-1470	CP010344	August 3, 2020
O182:H21 D181	CP024252	August 3, 2020
108	CP028693	August 3, 2020
112	CP028683	August 3, 2020
13P460A	CP019271	August 3, 2020
4/1-1	CP023844	August 3, 2020
4/4	CP023826	August 3, 2020
7/2	CP023820	August 3, 2020
CAU16175	CP047378	August 3, 2020
CI5	CP011018	August 3, 2020
Ecol_224	CP018948	August 3, 2020
EcPF16	CP054224	August 3, 2020
ExPEC XM	CP025328	August 3, 2020
KBN10P04869	CP026473	August 3, 2020
LD93-1	CP047662	August 3, 2020
NCTC8623	LR134234	August 3, 2020
RHB26-C18	CP057450	August 3, 2020
RHB34-C08	CP057175	August 3, 2020
RHBSTW-00176	CP056800	August 3, 2020
RM9088	CP042298	August 3, 2020
SCU-479	CP054317	August 3, 2020
SQ37	CP011320	August 3, 2020
tolC-	CP018801	August 3, 2020

Table S1.3: Bacterial strain, NCBI accession number, and date accessed for the whole genome alignments of different strains of *E. coli*. The same alignment and trimming methods described in Chapter 2 was used.

Strain	NCBI Accession Number	Date Accessed
	B. subtilis	
168	AL009126	August 3, 2020
BEST7613	AP012495	August 3, 2020
BSn5	CP002468	August 3, 2020
50-1	CP020915	August 3, 2020
ATCC 21228	CP020023	August 3, 2020
BS16045	CP017112	August 3, 2020
H1	CP026662	August 3, 2020
HRBS-10TDI13	CP015222	August 3, 2020
J-5	CP018295	August 3, 2020
3610	CP020102	August 3, 2020
P8_B1	CP045922	August 3, 2020
SG6	CP009796	August 3, 2020
SP1	CP058242	August 3, 2020
SRCM101393	CP031693	August 3, 2020
SRCM102754	CP028202	August 3, 2020
SRCM103517	CP035226	August 3, 2020
SRCM103622	CP035411	August 3, 2020
SRCM103773	CP035397	August 3, 2020
SRCM103862	CP035161	August 3, 2020
TLO3	CP021169	August 3, 2020
inaquosorum DE111	CP013984	August 3, 2020
AG1839	CP008698	August 3, 2020
168G	CP016852	August 3, 2020
SRCM100761	CP021889	August 3, 2020
BSP1	CP003695	August 3, 2020

Table S1.4: Bacterial strain, NCBI accession number, and date accessed for the whole genome alignments of different strains of *B. subtilis*. The same alignment and trimming methods described in Chapter 2 was used.

Strain/Species	NCBI Accession Number	Date Accessed
	Streptomyces	
S. coelicolor A3	AL645882	August 24, 2017
4F	NZ_CP013142	August 3, 2020
604F	NZ_CP026490	August 3, 2020
769	NZ_CP003987	August 3, 2020
ADI95-16	NZ_CP033581	August 3, 2020
CCM_MD2014	NZ_CP009754	August 3, 2020
CLI2509 strain CLI2905	NZ_CP021118	August 3, 2020
GS7	NZ_CP047146	August 3, 2020
HF10	NZ_CP047144	August 3, 2020
M2	NZ_CP028834	August 3, 2020
NA02536	NZ_CP054939	August 3, 2020
QMT-28	NZ_CP045643	August 3, 2020
RLB1-8	NZ_CP041650	August 3, 2020
RLB1-9	NZ_CP041654	August 3, 2020
RLB3-17	NZ_CP041610	August 3, 2020
RPA4-2	NZ_CP050975	August 3, 2020
RTd22	NZ_CP015726	August 3, 2020
S1D4-14	NZ_CP041607	August 3, 2020
SM17	NZ_CP029338	August 3, 2020
SM18	NZ_CP029342	August 3, 2020
VN1	NZ_CP036534	August 3, 2020
WAC 01438	NZ_CP029601	August 3, 2020
WAC 06738	NZ_CP029618	August 3, 2020
Z022	NZ_CP033073	August 3, 2020
Z423-1	NZ_CP053109	August 3, 2020
ZFG47	NZ_CP030073	August 3, 2020

Table S1.5: Bacterial strain, NCBI accession number, and date accessed for the whole genome alignments of different strains and species of *Streptomyces*. The same alignment and trimming methods described in Chapter 2 was used.

A0.4 Poor Sequence Alignment

After a re-alignment of `progressiveMauve` LCBs with `MAFFT` there were still regions of the alignment that were visibly poor. This prompted the additional alignment quality trimming using a custom `Python` script and `trimAl` (Capella-Gutiérrez et al. 2009). An example of what a "poor" alignment would look like can be found in Figure S1.3. The `FASTA` format of this segment of the alignment can be found on `GitHub` labelled as file `"poor_ecoli_alignment_example.fna"`.

This segment of `MAFFT` alignment (Figure S1.3) appears to have completely misaligned the second

sequence (*E. coli* O157H7). When we look at the genes that these regions of DNA are found within (Table S1.6), we see that the second sequence (*E. coli* O157H7) does not have the same protein sequence as the other bacteria genes. Poor sequence alignments like this, as well as other non-homologous alignment regions were removed from the analysis. Please see the main paper for more detailed methods.

FIGURE S1.3: Visualization of a section of `MAFFT` alignment between the six strains of *E. coli*. This alignment was visualized with the `SeaView` graphical interface (Gouy et al. 2010).

E. coli Strain	NCBI Accession Number	Alignment Gene Id
0104H4	CP003289	O3K_04155
O157H7	BA000007	ECs3861
083H1	CP001855	NRG857_18350
IAI39	CU928164	yghE
K12	U00096	yghE
UMN026	CU928163	yghE

TABLE S1.6: *E. coli* strain, NCBI accession number, and Gene Id for the genes in the poor alignment example (Figure S1.3).

A0.5 Phylogenetic Trees

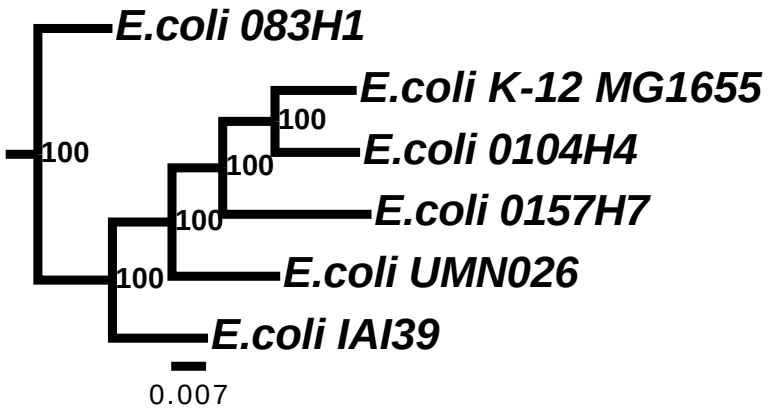


Figure S1.4: Phylogenetic tree of *E. coli* genomes. *E. fergusonii* ATCC 35469T was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 1000.

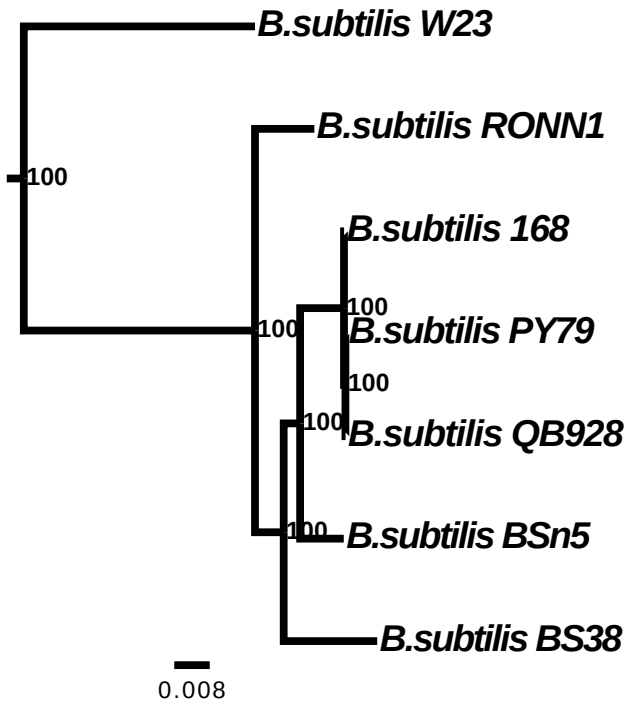


Figure S1.5: Phylogenetic tree of *B. subtilis* genomes. *B. cereus* FDAARGOS_797 was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 1000.

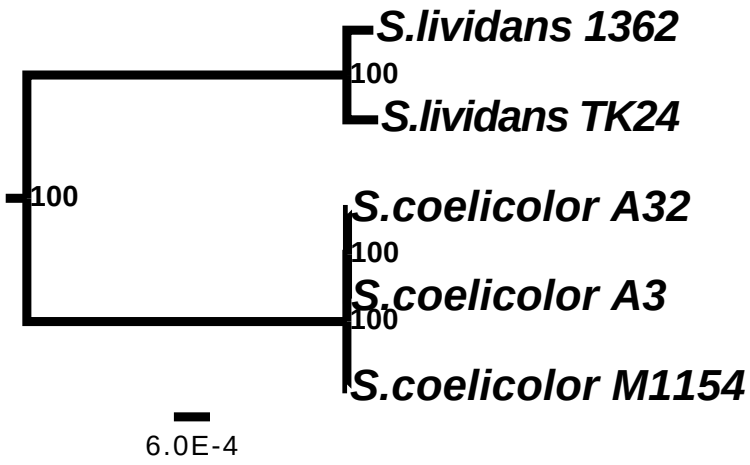


Figure S1.6: Phylogenetic tree of *Streptomyces* genomes. *S. aureofaciens* DM1 was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 1000.

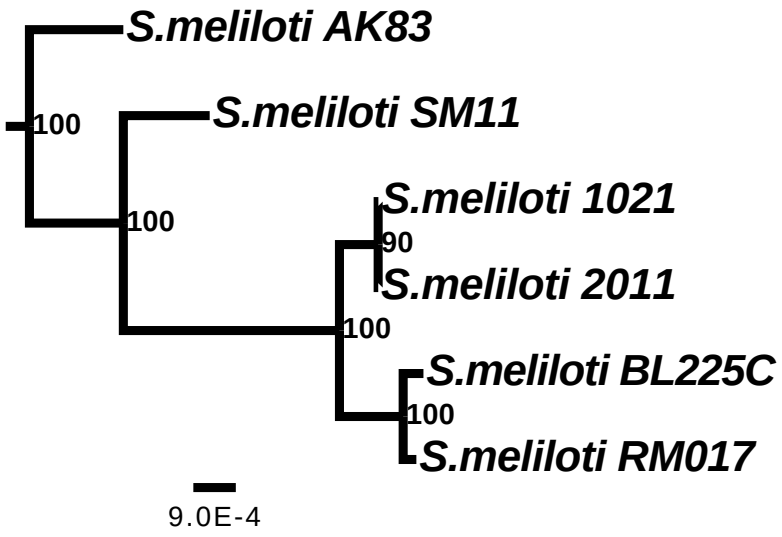


Figure S1.7: Phylogenetic tree using only the chromosomes of *S. meliloti*. *R. leguminosarum* trifolii WSM1689 chromosome was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 1000.

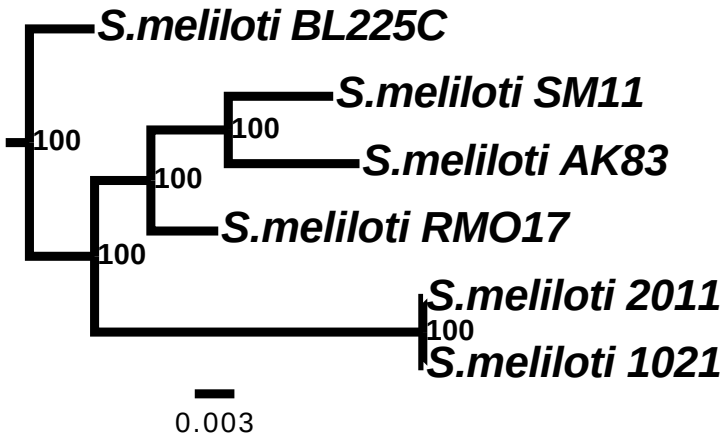


Figure S1.8: Phylogenetic tree using only pSymA of *S. meliloti*. *R. leguminosarum* trifolii WSM1689 plasmid pRLG202 was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 1000.

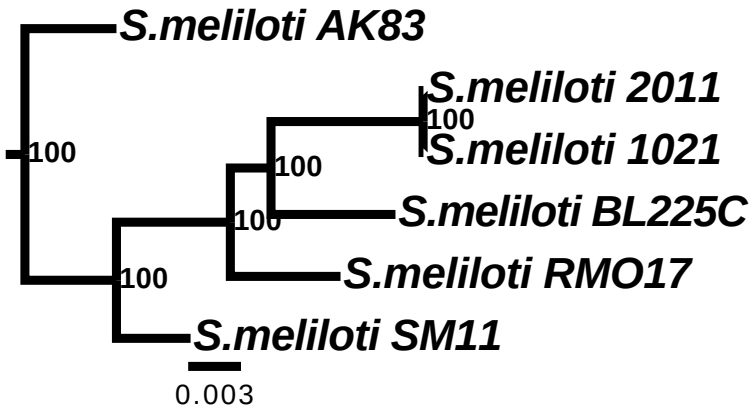


Figure S1.9: Phylogenetic tree using only pSymB of *S. meliloti*. *R. leguminosarum* trifolii WSM1689 plasmid pRLG201 was used as an outgroup to root the tree. Branch lengths are to scale. The numbers at each node indicate the bootstrap value as a percentage. The number of bootstrapped trees was 1000.

A0.6 Origin and Terminus Locations

Each of the bacterial strains used in this analysis vary in total genomic length, in some cases this difference is up to 856Kilobase Pair (Kbp) like in *E. coli* (Table S1.7). This will cause the farthest point from the origin of replication to appear larger because of the increased genome size of some strains.

Bacteria	Origin of Replication	Terminus of Replication	Length of Longest Genome (bp)
E. coli	3925744	1588773	5498450
B. subtilis	1	1942542	4215606
Streptomyces	3419363	1 & 8667664	8667664
S. meliloti Chromosome	1	1735626	3908022
S. meliloti pSymA	1350001	672888	1633319
S. meliloti pSymB	55090	896756	1690594

Table S1.7: Origin of replication and terminus of replication positions in replicons of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. The origin and terminus of replication are values from the representative strain of each bacteria, which can be found in Supplementary Table S1.2. The linear nature of *Streptomyces* chromosome gives it two termini, one at each end of the chromosome. The length of the longest genome is the longest genome length from all strains/species of each bacteria. This is not necessarily the same as the genome length of the representative strain.

Origin Location	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
Moved 100kb Left	$-1.445\times10^{-7***}$	$4.374\times10^{-9*}$	$6.909\times10^{-9***}$	$-1.316\times10^{-6***}$	$-1.058\times10^{-6***}$	$-2.009\times10^{-7***}$
Moved 90kb Left	$-1.544\times10^{-7***}$	$-1.036\times10^{-7***}$	$5.677\times10^{-9***}$	$-1.32\times10^{-6***}$	$-1.246\times10^{-6***}$	$-1.357\times10^{-7***}$
Moved 80kb Left	$-1.65\times10^{-7***}$	$-1.072\times10^{-7***}$	$8.11\times10^{-9***}$	$-1.338\times10^{-6***}$	$-1.398\times10^{-6***}$	$-6.57\times10^{-8***}$
Moved 70kb Left	$-1.667\times10^{-7***}$	$-1.102\times10^{-7***}$	$6.716\times10^{-9***}$	$-1.363\times10^{-6***}$	$-1.405\times10^{-6***}$	9.83×10^{-8}
Moved 60kb Left	$-1.64\times10^{-7***}$	$-1.19\times10^{-7***}$	$8.7\times10^{-9***}$	$-1.324\times10^{-6***}$	$-1.394\times10^{-6***}$	$1.129\times10^{-7***}$
Moved 50kb Left	$-1.446\times10^{-7***}$	$-1.211\times10^{-7***}$	$1.045\times10^{-8***}$	$-1.36\times10^{-6***}$	$-1.403\times10^{-6***}$	$1.521\times10^{-7***}$
Moved 40kb Left	$-1.4\times10^{-7***}$	$-1.299\times10^{-7***}$	$1.214\times10^{-8***}$	$-1.255\times10^{-6***}$	$-1.422\times10^{-6***}$	$1.543\times10^{-7***}$
Moved 30kb Left	$-1.498\times10^{-7***}$	$-1.292\times10^{-7***}$	$1.24\times10^{-8***}$	$-1.26\times10^{-6***}$	$-1.392\times10^{-6***}$	$1.63\times10^{-7***}$
Moved 20kb Left	$-1.51\times10^{-7***}$	$-1.1\times10^{-7***}$	$1.395\times10^{-8***}$	$-1.525\times10^{-6***}$	$-1.412\times10^{-6***}$	$1.603\times10^{-7***}$
Moved 10kb Left	$-1.262\times10^{-7***}$	-2.602×10^{-9}	$1.563\times10^{-8***}$	$-1.599\times10^{-6***}$	$-9.499\times10^{-7***}$	$2.973\times10^{-7***}$
Moved 10kb Right	$-1.305\times10^{-7***}$	$-2.045\times10^{-8***}$	$1.578\times10^{-8***}$	$1.614\times10^{-6***}$	$-1.026\times10^{-6***}$	$3.505\times10^{-7***}$
Moved 20kb Right	$-1.454\times10^{-7***}$	$-1.006\times10^{-7***}$	$1.903\times10^{-8***}$	$-1.634\times10^{-6***}$	$-1.475\times10^{-6***}$	$1.649\times10^{-7***}$
Moved 30kb Right	$-1.548\times10^{-7***}$	$-8.596\times10^{-8***}$	$2.046\times10^{-8***}$	$-1.698\times10^{-6***}$	$-1.417\times10^{-6***}$	$1.526\times10^{-7***}$
Moved 40kb Right	$-1.632\times10^{-7***}$	$-8.378\times10^{-8***}$	$2.125\times10^{-8***}$	$-1.719\times10^{-6***}$	$-1.367\times10^{-6***}$	$1.589\times10^{-7***}$
Moved 50kb Right	$-1.856\times10^{-7***}$	$-7.879\times10^{-8***}$	$1.957\times10^{-8***}$	$-1.735\times10^{-6***}$	$-1.277\times10^{-6***}$	$1.654\times10^{-7***}$
Moved 60kb Right	$-1.91\times10^{-7***}$	$-6.98\times10^{-8***}$	$1.974\times10^{-8***}$	$-1.788\times10^{-6***}$	$-1.169\times10^{-6***}$	$1.645\times10^{-7***}$
Moved 70kb Right	$-1.892\times10^{-7***}$	$-6.634\times10^{-8***}$	$1.934\times10^{-8***}$	$-1.854\times10^{-6***}$	$-1.059\times10^{-6***}$	$1.843\times10^{-7***}$
Moved 80kb Right	$-1.879\times10^{-7**}$	$-5.814\times10^{-8***}$	$2.313\times10^{-8***}$	$-1.891\times10^{-6***}$	$-9.07\times10^{-7***}$	$1.90\times10^{-7***}$
Moved 90kb Right	$-1.862\times10^{-7***}$	$-4.314\times10^{-8***}$	$2.304\times10^{-8***}$	$-1.865\times10^{-6***}$	$-7.171\times10^{-7***}$	$2.415\times10^{-7***}$
Moved 100kb Right	$-1.799\times10^{-7***}$	$-2.597\times10^{-8***}$	$1.945\times10^{-8***}$	$-1.525\times10^{-6***}$	$-6.572\times10^{-7***}$	$3.095\times10^{-7***}$

Table S1.8: Logistic regression analysis of the number of substitutions along the genome of the respective bacterial replicons after the origin location was moved by the specified increments from the original origin of replication position (listed in Table S1.7). All results are marked with significance codes as followed: $< 0.001 =$ '***', $0.001 < 0.01 =$ '**', $0.01 < 0.05 =$ '*', $0.05 < 0.1 =$ '.', $> 0.1 =$ ' '. Logistic regression was calculated after the origin of replication was moved to the new location in the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication.

Bacteria Strain	Accession Number	Date Accessed
E. coli K12 Chromosome	U00096	September 26, 2016
B. subtilis 168 Chromosome	NC_000964	November 10, 2016
S. coelicolor A3 Chromosome	AL645882	November 30, 2016
S. meliloti Chromosome 1021	NC_003047	June 3, 2014
S. meliloti pSymA 1021	NC_003037	June 3, 2014
S. meliloti pSymB 1021	NC_003078	June 3, 2014

Table S1.9: Strains and species used for determining the protein coding regions of each bacterial replicon. GenBank reference annotation was used to determine all protein coding sections of the replicons. NCBI accession numbers and date accessed are provided.

Genomic Position Clustering

A custom `R` script was used to cluster genomic positions together based on a user specified genetic distance using single-link clustering. An illustration of the clustering method used in this supplemental test can be found in Figure S1.10. This clustering was done for genomic distances beginning at 1bp and increasing by one order of magnitude until 1,000,000Base Pair (bp) difference exists between the taxa genomic positions. These newly clustered genomic positions were then put into the same substitution analysis as mentioned previously to determine the impact of this position clustering on the spatial substitution trends through a linear regression. A complete table of the statistical results from the clustering assessment are found in Table S1.10. The results from this analysis indicate that genomic positions up to 1,000,000bp apart can be considered a singular genomic position without altering the overall spatial substitution analysis.

Figure S1.10: Visualization of the genomic position clustering method. In this example, the user specified the genetic distance to be 2, all genomic positions within 2 base pairs would be clustered together. In this example we are looking at 6 taxa with genomic positions 10, 14, 12, 25, 22, and 20. Based on the clustering algorithm, positions 10, 14 and 12 would be grouped into a cluster (outlined in green), position 25 would be its own cluster (outlined in pink), and positions 22 and 20 would be grouped into another cluster (outlined in blue). Once the clusters are determined, a new genomic position for each of the clusters is calculated using the average of all positions within that cluster. In this example, the green cluster would have a new genomic position of 12 (the average between those three positions), the pink cluster would have the sane genomic position of 25, and the blue cluster would have a new genomic position of 21. The new list of genomic positions for the 4 taxa would be: 12, 12, 12, 25, 21 and 21.

Position Difference	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
1bp	-1.394×10^{-7}**	-2.538×10^{-8}**	1.736×10^{-8}**	-1.541×10^{-6}**	-9.130×10^{-7}**	2.488×10^{-7}***
10bp	-1.394×10^{-7}***	-2.518×10^{-8}***	-4.484×10^{-9}***	-1.627×10^{-6}***	-9.13×10^{-7}***	3.487×10^{-7}***
100bp	-1.764×10^{-7}***	-1.417×10^{-8}***	1.448×10^{-8}***	-1.605×10^{-6}***	-1.166×10^{-6}***	4.021×10^{-7}***
1000bp	-1.784×10^{-7}***	-1.417×10^{-8}***	1.505×10^{-8}***	-1.605×10^{-6}***	-1.153×10^{-6}***	4.021×10^{-7}***
10000bp	-1.712×10^{-7}***	-3.496×10^{-8}***	4.790×10^{-8}***	-1.605×10^{-6}***	-3.570×10^{-8}*	3.784×10^{-7}***
100000bp	-2.061×10^{-7}***	-3.561×10^{-8}***	4.167×10^{-9}***	-1.605×10^{-6}***	-4.676×10^{-7}***	3.784×10^{-7}***
1000000bp	4.229×10^{-8}***	-7.710×10^{-9}***	6.083×10^{-8}***	-1.605×10^{-6}***	4.285×10^{-6}***	-8.888×10^{-7}***

Table S1.10: Results from the position clustering analysis. Logistic regression analysis of the number of substitutions along the genome of the respective bacteria replicons to test position differences. The "Position Difference" column denotes different base pair distances that the positions in the genome were clustered together as. All results are marked with significance codes as followed: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', 0.05 $<$ 0.1 = '.', $>$ 0.1 = ' '. Logistic regression was calculated after the positions in the genome were determined to be the same at each position difference listed in the first column.

A0.7 High Substitutions Gene Example

Throughout this analysis there are a few genes/gene segments in all the bacterial replicons that have relatively high numbers of substitutions when compared to other genes or gene segments. These high numbers of substitutions are indeed real changes seen in homologous genes. To illustrate this, we have chosen a segment of alignment from *B. subtilis*. Information about the genes involved

Bacteria and Replicon	Average Replicon Length	Number of Sites	Number of Substitutions
E. coli Chromosome	5082529	3032961	200477
B. subtilis Chromosome	4077077	2411673	218843
Streptomyces Chromosome	8494093	5266854	20929
S. meliloti Chromosome	3426881	2125845	6420
S. meliloti pSymA	1455940	451314	10055
S. meliloti pSymB	1664597	1200129	28233

Table S1.11: Total number of protein coding sites in each replicon for this analysis and the number of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

in this segment can be found in Table S1.12. A protein alignment for these genes can be found on `GitHub` (`www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`) under the file name "`Bacillus_high_substitutions_gene_example.txt`".

Despite this high sequence identity and almost identical protein alignment (Figures S1.11 and S1.12), there are a total of 205 substitutions (across all nodes of the phylogenetic tree, Figure S1.5) within this short stretch of sequence. It is segments like these that are resulting in the appearance of extremely high numbers of substitutions in sections of all the bacterial replicon genomes.

Species	NCBI Accession Number	Gene Id
B. subtilis 168	NC_000964	BSU17380
B. subtilis BS38	NZ_CP017314	BSBS38_RS09695
B. subtilis BSn5	NC_014976	BSN5_RS21150
B. subtilis PY79	NC_022898	U712_RS08990
B. subtilis QB928	NC_018520	B657_RS09460
B. subtilis RONN1	NC_017195	I33_RS09040
B. subtilis W23	NC_014479	BSUW23_RS09220

Table S1.12: Information about the example gene segment from *B. subtilis* alignment with high number of substitutions.

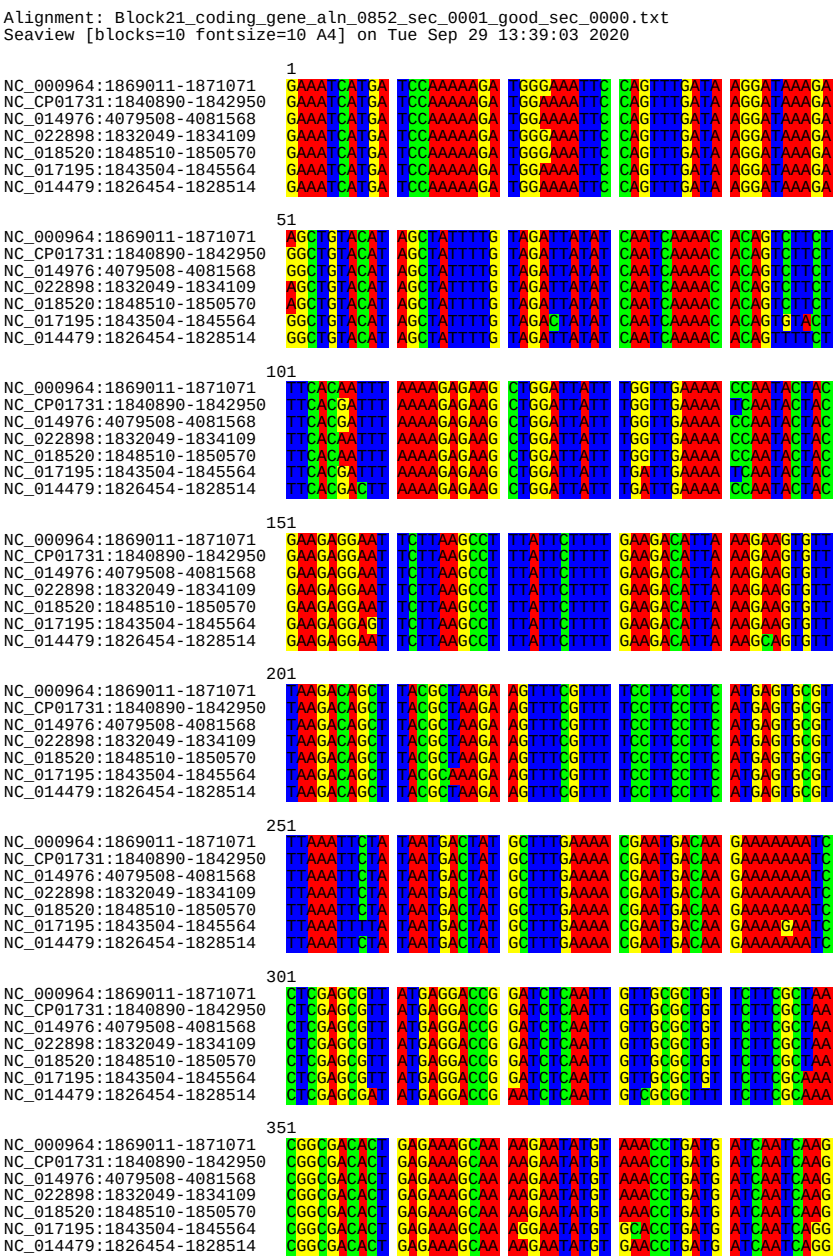


FIGURE S1.11: Visualization of a portion of the nucleotide alignment of *B. subtilis* genes with high numbers of substitutions. Alignment visualization was performed with `SeaView` (Gouy et al. 2010)

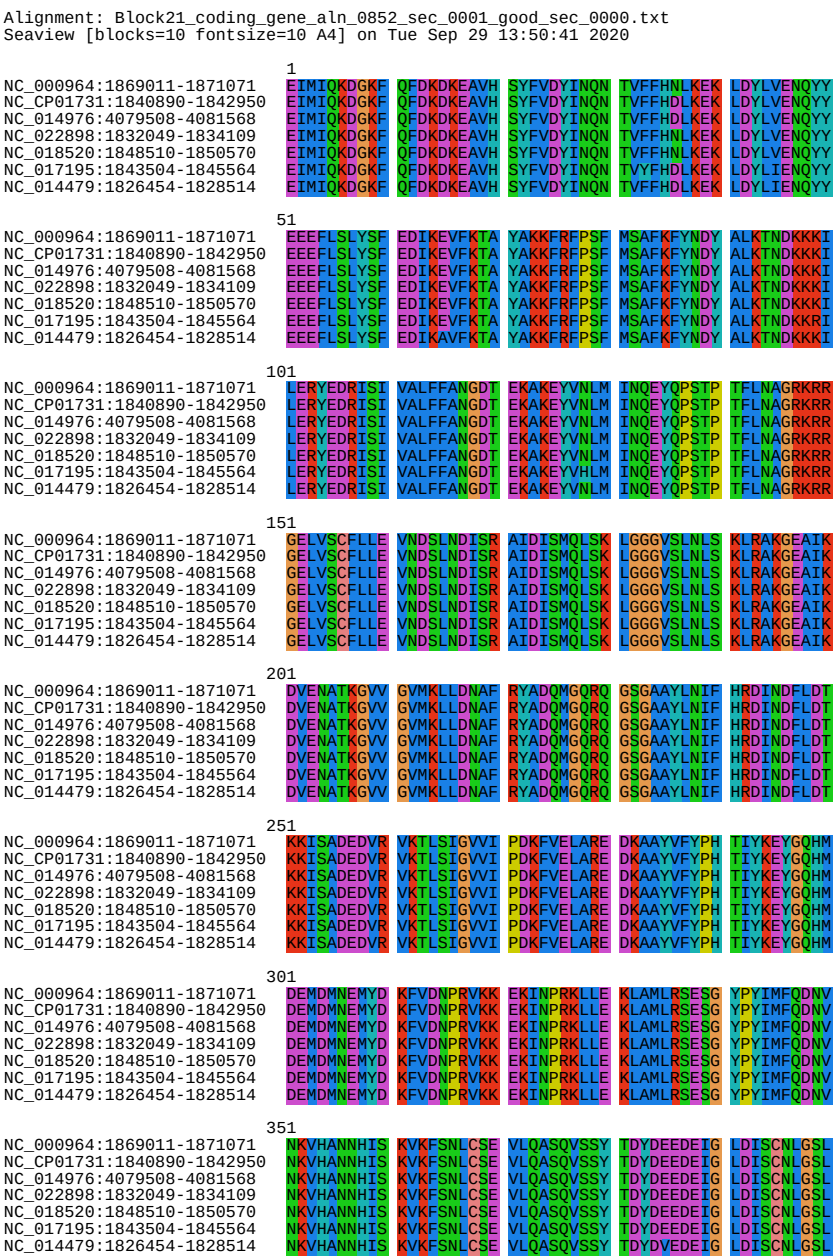


FIGURE S1.12: Visualization of a portion of the protein alignment of *B. subtilis* genes with high numbers of substitutions. Alignment visualization was performed with `SeaView` (Gouy et al. 2010)

A0.8 High Substitution Distribution

Bacteria and Replicon	Bidirectional Genomic Position (bp)	Protein/Gene Examples
E. coli Chromosome	1130000 - 1140000	Uncharacterized proteins
		Hypothetical proteins
		Lipoprotein
		Transcriptional activator
	1720000 - 1740000	Hypothetical proteins
		Predicted protein
		Small toxic polypeptide
B. subtilis Chromosome	1990000 - 2000000	Hypothetical proteins
		Unknown function
Streptomyces Chromosome	3550000 - 3570000	Hypothetical proteins
		Derived by automated computational analysis
		Putative integral membrane protein
		Reductase
S. meliloti Chromosome	180000 - 200000	Hypothetical proteins
		Small molecule metabolism
S. meliloti pSymA	790000 - 800000	Hypothetical proteins
		Transposase
		Small molecule metabolism
S. meliloti pSymB	610000 - 620000	Hypothetical proteins
		Transposon related functions
		Predicted membrane protein

Table S1.13: Table of high number of substitutions per 10Kbp genomic regions for each bacterial replicon and examples of the associated proteins/gene functions found in that region. The genomic position begins at the origin of replication and continues in both directions until the terminus of replication (bidirectional replication).

A0.9 Weighted, Non-weighted, and 20Kbp Near and Far From the Origin Substitution Linear Regression Analysis

Multiple linear regressions were performed to determine if there was any correlation between number of substitutions and distance from the origin of replication. A linear regression to determine how the weighted and non-weighted total number of substitutions in various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) changes with genomic position was performed (Tables S1.14 and S1.15). All additional linear regression results (Tables S1.14 and S1.15) mirror the results from the logistic regression on presence or absence of substitutions and changes in genomic position (see the Main Paper results section for more information). The results from these supplemental tests are consistent with the results from the linear regression found in the Main Paper, most bacterial replicons have a decreasing number of substitutions when moving away from the origin of replication.

To calculate the non-weighted values of the total number of substitutions per 10Kbp region of the genome, the total number of substitutions was summed up over each region of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp), while accounting for bidirectional replication (see Main Paper for details). A linear regression on these total number of substitutions in each section of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) was performed to see how the number of substitutions changes with distance from the origin of replication (Table S1.15). The weighted values of the total number of substitutions per various region of the genome, the total number of substitutions was summed up over each region of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) while accounting for bidirectional replication (see Main Paper for details). These summed values were then divided by the total number of protein coding sites in each region to obtain the weighted value. A linear regression on these weighted total number of substitutions in each section of the genome was performed to see how the number of substitutions changes with distance from the origin of replication (Table S1.14).

The Non-Significant (NS) linear regression results from Tables S1.15 and S1.14 are likely due to a decrease in the number of data points due to the nature of the methods for this supplemental analysis. In the windowed analysis (Tables S1.15 and S1.14) the total number of substitutions per various window size (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) were summed. This reduces the total number of data points used in the linear regressions, resulting in NS coefficient estimates. For example, the replicon of pSymA in *S. meliloti* only has a total length of 1.63Megabase Pair (Mbp) and roughly 16.3 million data points including all ancestral and extant substitutions/genomic positions. When the total number of substitutions is summed over each region of the genome, these data points are collapsed to summarize what is happening in each local window. Lets take the 400Kbp window for example, when the total number of substitutions is summed over each 400Kbp region of the genome, the number of data points is drastically reduced to about 40. It is therefore unlikely that 40 data points provide enough information to detect a significant trend between the number of substitutions and distance from the origin of replication. This same logic can be applied to the other bacteria and window sizes. We therefore conclude that the lack of detection of a significant trend (NS) in Tables S1.15 and S1.14 is due to the decreased number of data points.

We took a closer look at 20Kbp regions of the replicons close and far from the origin of replication. We performed a logistic regression on the presence or absence of a substitution with distance from the origin of replication. Data points from the 20Kbp regions closest to the origin of replication and data points from the 20Kbp regions closest to the terminus of replication were used for this portion of the analysis. Outliers were removed from this analysis. The number of substitutions per site was also calculated in each of these 20Kbp regions for each bacterial replicon. We were unable to determine a consistent spatial substitution trend when considering only the 20Kbp near and far from the origin of replication in all bacterial replicons. Some bacterial replicons had a positive correlation coefficient, indicating that the number of substitutions increases with increasing distance from the origin of replication (Table S1.16). Other replicons had a negative correlation coefficient, suggesting that the number of substitutions decreases with increasing distance from the origin of replication (Table S1.16). Additionally, it was unclear if the number of substitutions per site locally were higher near the origin of replication or near the terminus. Some bacteria had higher number of substitutions per site near the origin (*Streptomyces*, *S. meliloti* chromosome and pSymB), while other replicons has the opposite trend (*E. coli*, *B. subtilis*,

and *S. meliloti* pSymA) (Table S1.16). These results suggest that on a small local scale, there are varying patterns of substitutions with respect to distance from the origin of replication. This varies between bacteria, and in some cases even within the same bacteria (*S. meliloti* pSymB). This variation locally does not allow us to make any overarching statements about the local distribution of substitutions in bacterial genomes. It is therefore more useful to consider the global (genome wide) pattern of substitutions when making overarching statements about genomic substitution arrangements.

	Protein Coding Window Size					
Bacteria and Replicon	10Kbp	25Kbp	50Kbp	100Kbp	200Kbp	400Kbp
E. coli Chromosome	-2.27×10^{-10}*** (0.038)	-2.54×10^{-10}** (0.078)	-2.32×10^{-10}** (0.112)	-2.36×10^{-10}* (0.133)	NS (0.200)	NS (0.362)
B. subtilis Chromosome	NS (0.009)	NS (0.001)	NS (0.0002)	NS (0.002)	NS (0.019)	NS (0.484)
Streptomyces Chromosome	NS (2.49×10^{-5})	NS (2.12×10^{-5})	NS (0.004)	NS (0.0002)	3.68×10^{-11}* (0.126)	NS (0.182)
S. meliloti Chromosome	-1.21×10^{-10}** (0.076)	-1.71×10^{-10}*** (0.137)	-1.86×10^{-10}** (0.126)	-2.78×10^{-10}** (0.350)	NS (0.150)	NS (0.397)
S. meliloti pSymA	NS (0.032)	NS (0019)	NS (0.135)	NS (0.0124)	NS (0.034)	NS (1.42×10^{-30})
S. meliloti pSymB	NS (0.001)	NS (0.003)	NS (0.008)	NS (0.006)	NS (2.12×10^{-8})	NS (0.043)

Table S1.14: Linear regression on various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) with increasing distance from the origin of replication after accounting for bidirectional replication. The total number of substitutions in each section of the genome was divided by the total number of protein coding sites in that genomic region (weighted). All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'. The R^2 value for each coefficient estimate is found below the value in brackets ().

Bacteria and Replicon	Protein Coding Window Size					
	10Kbp	25Kbp	50Kbp	100Kbp	200Kbp	400Kbp
E. coli Chromosome	-1.66×10^{-4}*** (0.398)	-4.12×10^{-4}*** (0.476)	-8.64×10^{-4}*** (0.563)	-1.71×10^{-3}*** (0.509)	-3.42×10^{-3}** (0.534)	-6.71×10^{-3}* (0.592)
B. subtilis Chromosome	NS (0.004)	NS (0.004)	NS (0.001)	NS (0.001)	NS (0.145)	NS (0.027)
Streptomyces Chromosome	NS (0.002)	NS (0.007)	NS (0.014)	NS (0.025)	NS (0.073)	NS (0.074)
S. meliloti Chromosome	-8.97×10^{-6}*** (0.040)	-3.72×10^{-5}** (0.098)	-7.76×10^{-5}* (0.126)	-1.64×10^{-4}* (0.188)	NS (0.082)	NS (0.427)
S. meliloti pSymA	NS (0.027)	NS (0.001)	NS (0.006)	NS (0.193)	NS (0.050)	NS (1.59×10^{-31})
S. meliloti pSymB	NS (0.035)	NS (0.053)	NS (0.010)	NS (0.002)	NS (0.495)	NS (0.491)

Table S1.15: Linear regression on various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) with increasing distance from the origin of replication after accounting for bidirectional replication. The linear regression was performed on the total number of substitutions in each section of the genome without accounting for the number of sites in each genomic region (non-weighted). All results are marked with significance codes as followed: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', $>$ 0.05 = 'NS'. The R^2 value for each coefficient estimate is found below the value in brackets ().

Bacteria and Replicon	Protein Coding			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
	Origin	Terminus	Origin	Terminus
E. coli Chromosome	NS	6.16×10^{-6}**	5.85×10^{-3}	6.47×10^{-3}
B. subtilis Chromosome	1.18×10^{-6}*	1.57×10^{-5}***	4.23×10^{-3}	5.01×10^{-3}
Streptomyces Chromosome	NS	NS	2.36×10^{-4}	2.05×10^{-5}
S. meliloti Chromosome	7.11×10^{-6}***	NS	1.51×10^{-3}	3.86×10^{-5}
S. meliloti pSymA	-6.94×10^{-5}***	NS	2.03×10^{-3}	3.27×10^{-3}
S. meliloti pSymB	1.58×10^{-5}***	-7.10×10^{-5}***	3.06×10^{-3}	1.25×10^{-3}

Table S1.16: Logistic regression on 20Kbp closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. Number of substitutions was calculated by taking the total number of substitutions in each of the 20Kbp regions and dividing by the total number of sites in those regions. All results are marked with significance codes as followed: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', $>$ 0.05 = 'NS'.

A0.10 Non-linear Analysis of Number of Substitutions and Distance From the Origin of Replication

Using a simple smoothed conditional means method (`geom_smooth()` function in `R`), a non-linear trend analysis was performed on all bacterial replicons. The previous mentioned weighted data (see the previous subsection), was used in this analysis. The weighted data represents the total number of substitutions divided by the total number of protein-coding sites in 10Kbp segments of the genomes. Outliers were

removed. The results from this non-linear analysis can be seen in Figures S1.13 - S1.18. The visual results from this analysis mirror the findings from the main paper, the total number of substitutions varies with distance from the origin of replication, but the direction of this trend is unclear and inconsistent between bacterial replicons.

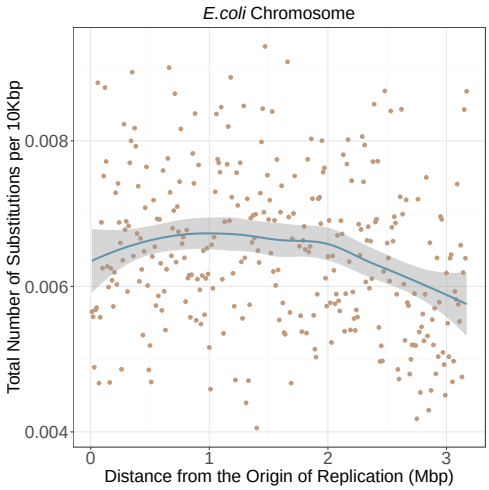


Figure S1.13: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *E. coli* genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

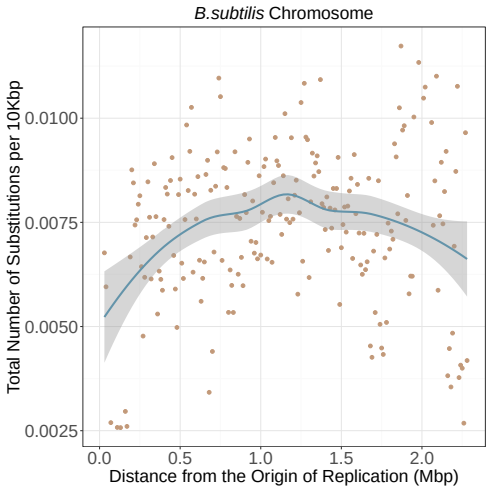


Figure S1.14: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *B. subtilis* genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

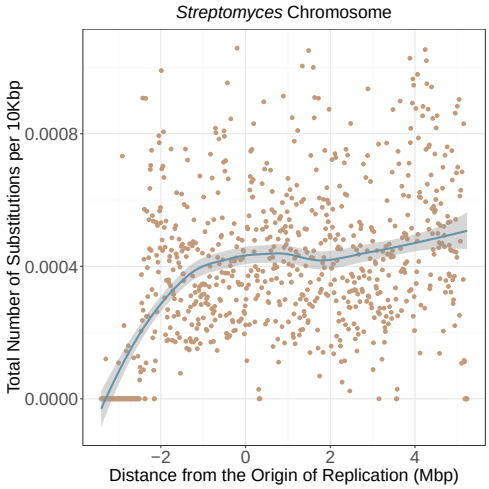


Figure S1.15: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *Streptomyces* genome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

Figure S1.16: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *S. meliloti* Chromosome. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

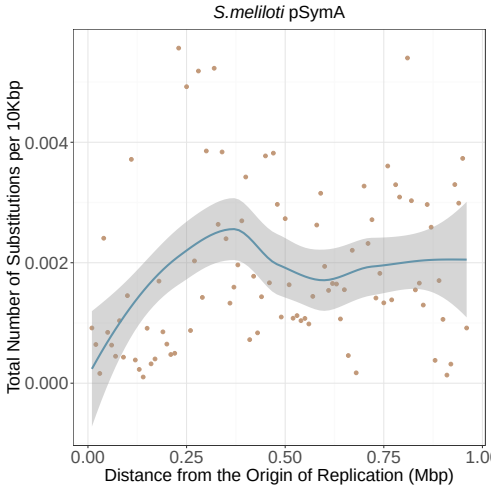


Figure S1.17: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *S. meliloti* pSymA replicon. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

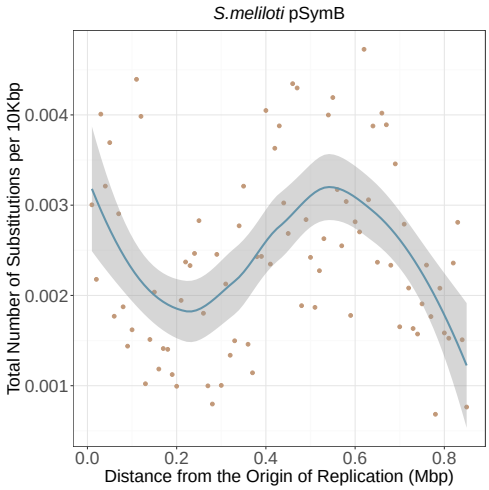


FIGURE S1.18: The graph shows the total number of substitutions weighted by the total number of protein-coding sites per 10Kbp segments of the *S. meliloti* pSymB replicon. Each of these individual values are represented by beige coloured circles. A non-linear trend line (using the `geom_smooth()` function in `R`), was fit to these average values and the associated 95% confidence intervals for this line is represented by the grey ribbon around the blue trend line. Outliers were removed from this graph.

A0.11 Total Number of Sites Linear Regression

We performed a linear regression on the total number of protein coding sites and distance from the origin of replication (Table S1.17). We found that the total number protein coding sites decreases with distance from the origin of replication in majority of the bacterial replicons in this analysis. We were unable to detect a significant relationship between the number of protein coding sites and distance from the origin of replication in *B. subtilis*, the chromosome, and pSymA of *S. meliloti*.

Bacteria and Replicon	Coefficient Estimate	R^2
E. coli Chromosome	-2.33×10^{-2}***	0.423
B. subtilis Chromosome	NS	0.001
Streptomyces Chromosome	-4.09×10^{-3}***	0.095
S. meliloti Chromosome	NS	0.013
S. meliloti pSymA	NS	0.002
S. meliloti pSymB	2.69×10^{-2}**	0.081

TABLE S1.17: Linear regression analysis of the total number of protein coding sites per 10Kbp along the genome of the respective bacteria replicons. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

A0.12 Robust "Leave One Out" Analysis on Substitution Data

Due to the computational and data availability limitations in the quantity of genomes chosen for each bacteria, we have performed an additional test to determine the robustness of our results. We have systematically removed/left out each taxa from the original substitutions analysis (as described in the Main Paper) this is a Leave One Out (LOO) analysis. The goal of this analysis is to see if the overall results, that the number of substitutions significantly varies with distance from the origin of replication but the sign of this correlation is inconsistent, changes when any one taxa is removed. We want to ensure that our particular data sets are not influencing our conclusions. The original whole genome alignments specified by `progressiveMauve` and re-aligned with `MAFFT` following our various alignment quality criteria (see Methods) was used for this LOO analysis. The sequences from each taxa were systematically removed/left out from these alignment blocks. The original phylogenetic trees and corresponding branch lengths (Figures S1.4 - S1.9) were altered so that the same taxa that was remove from the alignment blocks was also removed from the phylogenetic tree, while maintaining correct branch lengths. These LOO alignment blocks and trees were then subject to the same methods for the substitution analysis (see Methods in Main Paper) where the ancestral nucleotide and genomic position was determined for each protein-coding site in the alignment blocks. A logistic regression was performed to determine the relationship between the number of substitutions and distance from the origin of replication (see Methods). A summary of these logistic regression results with each taxa removed can be found in Table S1.18.

The results from the chromosomes of *E. coli* and *S. meliloti* (Table S1.18) indicate that removing any of the taxa from these analysis, results in the same overall conclusion, that the number of substitutions decreases with increasing distance from the origin of replication. For the remaining replicons (*B. subtilis*, *Streptomyces*, pSymA and pSymB of *S. meliloti*), majority of the LOO results mirror what was found in the main paper when all taxa were present in the analysis (Table S1.18). However, there are some specific taxa that cause a reversal in the sign of the coefficient estimate. In the case of *Streptomyces* and pSymA in *S. meliloti*, the taxa which causes a reversal in sign when removed (*S. lividans* 1362 CM001889 and *S. meliloti* BL225C NC_017324 respectively) alter the location of the "outgroup" on the phylogenetic trees (Figures S1.6 and S1.8 respectively). When *S. lividans* 1362 CM001889 is removed, the new outgroup for the phylogenetic tree (Figure S1.6) becomes *S. lividans* TK24. When pSymA from *S. meliloti* BL225C is removed, the clade containing *S. meliloti* 2011 and 1021 is now in the outgroup position. This shift in the outgroup is the cause for the coefficient estimate changing sign when these particular taxa (*S. lividans* 1362 CM001889 and *S. meliloti* BL225C NC_017324) are removed from the analysis. For *B. subtilis* and pSymB of *S. meliloti*, the taxa which caused a reversal in sign when removed (*B. subtilis* BSn5 NC_014976 and *S. meliloti* RMO17 CP009146) are located more in the inner parts of the phylogentic trees (Figures S1.5 and S1.9). These particular taxa heavily influence the ancestral genomic positions present throughout the phylogenetic tree. When these particular taxa are removed, this changes the ancestral genomic positions and alters ancestrally where the substitutions are located. This then influences the distribution of substitutions along the replicons enough to cause a change in the sign of the coefficient estimate (Table S1.18).

Since most of the results from the LOO analysis are the same as what was found in the main paper for each respective replicon, we maintain that our findings are robust even with the systematic removal of

each taxa. The number of substitutions significantly varies with distance from the origin of replication, but the sign of this correlation is inconsistent.

Taxa Removed	Coefficient Estimate
E. coli	
None	-2.66×10^{-8}***
U00096	-3.12×10^{-8}***
CP0032890	-3.07×10^{-8}***
CU9281640	-2.95×10^{-8}***
CP0018550	-1.50×10^{-8}***
BA0000070	-2.63×10^{-8}***
CU9281630	-2.49×10^{-8}***
B. subtilis	
None	2.76×10^{-8}***
NC_000964	2.96×10^{-8}***
NC_018520	3.57×10^{-8}***
NC_017195	1.00×10^{-7}***
NC_022898	5.17×10^{-8}***
NC_014976	-4.02×10^{-8}***
CP01731	5.43×10^{-8}***
NC_014479	NS
Streptomyces	
None	7.21×10^{-8}***
CP050522	8.37×10^{-8}***
GG657756	3.62×10^{-8}***
CP042324	7.72×10^{-8}***
AL645882	7.65×10^{-8}***
CM001889	-2.46×10^{-7}***

Taxa Removed	Coefficient Estimate
S. meliloti Chromosome	
None	-6.57×10^{-7}***
NC_015590	-3.18×10^{-7}***
NC_003047	-6.01×10^{-7}***
CP004140	-6.00×10^{-7}***
CP009144	-6.67×10^{-7}***
NC_017322	-7.19×10^{-7}***
NC_017325	-5.01×10^{-7}***
S. meliloti pSymA	
None	2.74×10^{-7}***
NC_017327	6.98×10^{-7}***
CP009145	1.78×10^{-7}***
NC_003037	2.09×10^{-7}***
CP004138	2.08×10^{-7}***
NC_015591	NS
NC_017324	-1.52×10^{-6}***
S. meliloti pSymB	
None	1.10×10^{-7}***
NC_015596	6.78×10^{-7}***
NC_017326	1.67×10^{-7}***
NC_017323	NS
CP009146	-2.57×10^{-7}***
CP004139	1.04×10^{-7}***
NC_003078	1.04×10^{-7}***

Table S1.18: Logistic regression on the presence or absence of a substitution and distance from the origin of replication. Each strain was systematically removed and the entire analysis was repeated. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

A0.13 High Synonymous Substitution Rate (dS) Values

Throughout this analysis there are a few genes/gene segments in all the bacterial replicons that have relatively high dS values. This is particularly evident in *B. subtilis* near 0.5Mbp from the origin of replication. Although we have rigorous and conservative methods for our sequence alignment and trimming, there appear to be some genes that are well aligned and similar for portions of the gene, but are quite

divergent for other regions of the same gene. To illustrate this, we have chosen a gene alignment from this high dS region in *B. subtilis* located around 0.5Mbp from the origin of replication. The genes in this alignment can be found in Table S1.19. A simple `Clustal Omega` protein alignment of these genes can be found below. In this example it is evident that some portions of the gene have almost 100% sequence identity, while others are drastically divergent. These divergent regions are typically have a length close to our minimum 100bp trimming length, and are retained in our analysis in some cases. These divergent regions are what is driving the high dS values in our analysis.

Species	NCBI Accession Number	Gene Id
B. subtilis 168 *	NC_000964	BSU12750
B. subtilis BS38	NZ_CP017314	BSBS38_RS07215
B. subtilis BSn5	NC_014976	BSN5_RS18735
B. subtilis PY79	NC_022898	U712_RS06655
B. subtilis QB928	NC_018520	B657_RS07025
B. subtilis RONN1	NC_017195	I33_RS06720
B. subtilis W23	NC_014479	BSUW23_RS06800

Table S1.19: Information about the example gene alignment from *B. subtilis* with a high dS value.

```
CLUSTAL O(1.2.4) multiple sequence alignment


NC_014479       MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAAHKNDMNNPHNTTKAQVGLGNVDNV     60
NC_000964       MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAAHKNDMNNPHNTTKAQIGLGNVDNV     60
NC_022898       MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAAHKNDMNNPHNTTKAQIGLGNVDNV     60
NC_018520       MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAAHKNDMNNPHNTTKAQIGLGNVDNV     60
NZ_CP017314     MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAAHKNDMNNPHNTTKAQIGLGNVDNV     60
NC_014976       MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAAHKNDMNNPHNTTKAQIGLGNVDNV     60
NC_017195       MAYEEKTDWLPDDPINEDDVNRWEKGIKDAHTDLAVHKNDMNNPHNTTKAQIGLGNVDNV     60
                ********************************.***************:********

NC_014479       KQAAKKDFDQHDQDQVRHIAEEEREKWNGGQLSKITKDDGSVFITID-NGQDFNEVAAQQ     119
NC_000964       QQASKTEFNEHNHDSTRHITSVERDEWNAKETPAGAQYKADQ------------------     102
NC_022898       QQASKTEFNEHNHDSTRHITSVERDEWNAKETPAGAQYKADQ------------------     102
NC_018520       QQASKTEFNEHNHDSTRHITSVERDEWNAKETPAGAQYKADQ------------------     102
NZ_CP017314     QQAAKKDFDKHEQDQVRHITSTERENWNAKETPGEAQNKADQ------------------     102
NC_014976       QQAAKKDFEKHVNDGTIHITAAERSKWNNAQLSKISGDDGRVFYKSVTEITDYNDL----     116
NC_017195       QQAAKKDFDKHISDETIHISSSERTKWNNAQLTKLTDEKGKYLASIQN-GLDFHKIVEEL     119
                :**:*.:*::*  * . **:  ** :**  :    :  ..

NC_014479       KKSFTFYTVKTGLNTPPQPTKGIYLYSSENDGEAIAMTNDGG-----IWR-KTLTSGEWS     173
NC_000964       ----------A--------------------EANAKAYTD------NFAAR--------     117
NC_022898       ----------A--------------------EANAKAYTD------NFAAR--------     117
NC_018520       ----------A--------------------EANAKAYTD------NFAAR--------     117
NZ_CP017314     ----------A--------------------EANAKAYTD------SFAAR--------     117
NC_014976       TDTGMYLIYNDGLNGPGLNQCFLLVMSYKN--TLVQIAYDGIKGEQSFFRIRKNDSTTWT     174
NC_017195       GQTFFFYTDKTGINTPPFATRGL-YIGYKSYGEALAMDYEGG-----TWR-KSLNDSGWT     172
                                                 .       :  :

NC_014479       EWASFETEAGSKSKAAQ-------------------------------------------     190
NC_000964       ------------------------------------------RD---------------     119
NC_022898       ------------------------------------------RD---------------     119
NC_018520       ------------------------------------------RD---------------     119
NZ_CP017314     ------------------------------------------RD---------------     119
NC_014976       AWIESETTEGSQKKIDAHANKTDIHVTKSDKDKWNDSQLFKITQDNGLAKYCEDA--DFN     232
NC_017195       DWVQLETSEGAQFKVRSHEEKTEIHVNKSDKDKWNSGQLFKVTADNGTQKINLSSGSFYD     232


NC_014479       ------------------------------------------------------------     190
NC_000964       ------------------NPNQVT-KA-----Q-----------VGLGNV----------     134
NC_022898       ------------------NPNQVT-KA-----Q-----------VGLGNV----------     134
NC_018520       ------------------NPNQVT-KA-----Q-----------VGLGNV----------     134
NZ_CP017314     ------------------NPNQVT-KA-----Q-----------VGLGNV----------     134
NC_014976       TVIETGFYYMSGATTTLNAPVNN--NGYLMVYNFSTYAYQEYTSYSSSDTISTGRRKFMR     290
NC_017195       SLKDVGTVTFYGTNAVTDNPSNTSLRGMQLVGQLG---------IGMGYAVDVGGNAWWF     283


NC_014479       ----------------AEKNAKNYIDNHTDNSSIHITNDERVKWNGAQLTKLTKDNGRRT     234
NC_000964       --------------ENVKQASLADFDAHLSNSKVHVSEGERNKWNAAQLIKLTGDDGKRI     180
NC_022898       --------------ENVKQASLADFDAHLSNSKVHVSEGERNKWNAAQLIKLTGDDGKRI     180
NC_018520       --------------ENVKQASLADFDAHLSNSKVHVSEGERNKWNAAQLIKLTGDDGKRI     180
NZ_CP017314     --------------ENVKQASQADFDAHLSNTKVHVSEGERNKWNAAQLIKLTGDDGKRI     180
NC_014976       NKVANSDVWTSWREIESVEGSQIKVDAHANKTDIHVTTSDKDKWNNAQLYRLTDTQGCRT     350
NC_017195      -FYNANDSAINWYQIESITGAQSKIDAHANKTDIHVTTSDKDKWNNAQLYRLTDTQGCRT     342
```

```
                                :   .* * .::.:*:: .:: *** *** :**  :* *

NC_014479       WVPDGTDILSLSTGFYYGVGKYVVNNPVDDDNAWYNYDVIE-GESGRKTIVAYQSFEVTM     293
NC_000964       QLQDGTDILTLSSGFYCAVGQSVVNNPVEGDAAWYNYDIVE-GGSGRKTIVAYQSWGSMM     239
NC_022898       QLQDGTDILTLSSGFYCAVGQSVVNNPVEGDAAWYNYDIVE-GGSGRKTIVAYQSWGSMM     239
NC_018520       QLQDGTDILTLSSGFYCAVGQSVVNNPVEGDAAWYNYDIVE-GGSGRKTIVAYQSWGSMM     239
NZ_CP017314     QLQDGTDILTLSSGFYCAVGQSVVNNPVEGDATWYNYDIVE-GGSGRKTIVAYQSWGSMM     239
NC_014976       KIPDGTDLLTLPSGFYYALGNVITNNPVSGDGSWYNYDVIETEGGGRKTILASRSYDGTF     410
NC_017195       KIPDGTDLLTLPSGFYYAVGNVIINNPVLGDGSWYNYDVIETGGGGRKTIFASRSFDGTF     402
                 : ****:*:* :*** .:*: : **** .* :*****::*   .*****.* :*:   :

NC_014479       WIGMVHTDGKFRGWKRLVTSEELNSENINKITDESLYQDAAYSGNNYPIGITTVAILQGS     353
NC_000964       WIGMVHTDGEFRGWKQIATTDFIDRVQTELDLH--------ENDKTNPHSVTK-------     284
NC_022898       WIGMVHTDGEFRGWKQIATTDFIDRVQTELDLH--------ENDKTNPHSVTK-------     284
NC_018520       WIGMVHTDGEFRGWKQIATTDFIDRVQTELDLH--------ENDKTNPHSVTK-------     284
NZ_CP017314     WIGMVHTDGKFRGWKQIATTDFIDRVQSELDIH--------KNDKTNPHSVTK-------     284
NC_014976       WTATIHTDGVFKGWNKIETE-----------------------------------------     430
NC_017195       WMATIHTDGVFKGWNKIETE-----------------------------------------     422
                * . :**** *:**::: *

NC_014479       TGYPYELGEVLNIKSSKYRFAQFFFYAGNTGQKKVFIRHWYDTVGWTDFITIPSSEELES     413
NC_000964       --QQVGLGNVENVKQETPDGAQ--------KKADTALNQSKDYTNSTAFITRPLNS----     330
NC_022898       --QQVGLGNVENVKQETPDGAQ--------KKADTALNQSKDYTNSTAFITRPLNS----     330
NC_018520       --QQVGLGNVENVKQETPDGAQ--------KKADTALNQSKDYTNSTAFITRPLNS----     330
NZ_CP017314     --QQVGLGNVENVKQETPDGAQ--------KKADTALNQSKDYTNSTAFITRPLNS----     330
NC_014976       -----------------------------------------------------------     430
NC_017195       -----------------------------------------------------------     422


NC_014479       VLNTAKLYTDSHANNTEIHVTQNDKTKWNNSQIFKLTQDDGTLGKFYNEDLNNITKTGFY     473
NC_000964       ITDANDL---------------------------NLP--PGT----YRLDTNYMNAN---     354
NC_022898       ITDANDL---------------------------NLP--PGT----YRLDTNYMNAN---     354
NC_018520       ITDANDL---------------------------NLP--PGT----YRLDTNYMNAN---     354
NZ_CP017314     ITDANDL---------------------------NLP--PGT----YRLDTNYMNAN---     354
NC_014976       -----------------------------------------------------------     430
NC_017195       -----------------------------------------------------------     422


NC_014479       YIYSSTTELNAPINRNGYLLVYNVETYPYQEFTSYSGYTDSIPDNRRKFIRNKKQDSEEW     533
NC_000964       --PVLQNQFPLNDNRTGLLIIYPSANK-------WATRQDWFSISTKTLYTRVAVNGTDY     405
NC_022898       --PVLQNQFPLNDNRTGLLIIYPSANK-------WATRQDWFSISTKTLYTRVAVNGTDY     405
NC_018520       --PVLQNQFPLNDNRTGLLIIYPSANK-------WATRQDWFSISTKTLYTRVAVNGTDY     405
NZ_CP017314     --PELQNQFPLNDNRTGLLIIYPSANK-------WATRQDWFSISTKTLYTRVAVNGTEY     405
NC_014976       -----------------------------------------------------------     430
NC_017195       -----------------------------------------------------------     422


NC_014479       TPWMEIEYSQGAQAKADKALADAKNYVDTNYTNQKLTKLTGSNAIQDARTGGDEYPQGLT     593
NC_000964       SGWYILENSEGSQNKADKALADAKNYVETNYTNQKLTVLTGSNAIQDARISGNDYKYGIT     465
NC_022898       SGWYILENSEGSQNKADKALADAKNYVETNYTNQKLTVLTGSNAIQDARISGNDYKYGIT     465
NC_018520       SGWYILENSEGSQNKADKALADAKNYVETNYTNQKLTVLTGSNAIQDARISGNDYKYGIT     465
NZ_CP017314     TDWYILETSEGSQSKADKALADAKNYVDSNYTNNKLTVLTGSNAIQDARTSGNEYPAGLT     465
NC_014976       ---------VSAQTKADKALADAKNYVDSNYTNNKLTVLTGSNAIQDARTGGNEYPPGLT     481
NC_017195       ---------ASAQTKADKALSDAKNYVETNYTNQKLTVLTGSNAIQDARISGNDYKYGIT     473
                         .:* ******:******::****:*** ********** .*::*  *:*
```

```
NC_014479       LIDIGQGNNTGYPLRYGFVKNEKYSDFRFAQYFYGTGNESGSYIDSTGTWIRHWWSGSGW     653
NC_000964       FMDIGANNTTGYPLTYGFVKNEKHSNYRFTQYFYGNADTTSGSYDHVGTWIRHWWADSGW     525
NC_022898       FMDIGANNTTGYPLTYGFVKNEKHSNYRFTQYFYGNADTTSGSYDHVGTWIRHWWADSGW     525
NC_018520       FMDIGANNTTGYPLTYGFVKNEKHSNYRFTQYFYGNADTTSGSYDHVGTWIRHWWADSGW     525
NZ_CP017314     FMDIGANNTTGYPLTYGIVKNEKYSNYRFAQYFYGTGNESNSYFTSTGSWIRHWWSDSGW     525
NC_014976       LMDIGQGNTTGYPLGYGIVKNEKYSDFRFTQYFYGTGNESNSYIDSTGTWVRHWWSGSGW     541
NC_017195       FMDIGANNTTGYPLTYGFVKNEKHSNYRFTQYFYGNADTTSGSYDHVGTWIRHWWADSGW     533
                ::*** .*.***** **:*****:*::**:*****..: :..    .*:*:****:.***

NC_014479       TAWHKISGFAHAYIRTTGIQYLDKAAHTKIQFNRKIKDSHNAFDTKNSRFVAPNDGMFLV     713
NC_000964       TAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDTKNNRFVASHAGMHLV     585
NC_022898       TAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDTKNNRFVASHAGMHLV     585
NC_018520       TAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDTKNNRFVASHAGMHLV     585
NZ_CP017314     TAWHKISGFAHANIGTTGKQQLIKGELQKVKYNRKIKDSHNTFDTKNNRFIVPNDGMFLV     585
NC_014976       TAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDTKNNRFVASHAGMHLV     601
NC_017195       TAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDAKNNRFVASHAGMHLV     593
                ***:******** * *** * * *.   *:::** *****: **:**.**:. : **.**

NC_014479       GVGLYMINTPAYINFHLKLYLNGSLYKPIDHKRGDFV--DKENEMNLDLNGNVTVPMNKG     771
NC_000964       SASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTVPLDAG     645
NC_022898       SASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTVPLDAG     645
NC_018520       SASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTVPLDAG     645
NZ_CP017314     NAGLYIENYQRYVNYELDIYLNGVRYKNIAHYRANPGDQSDTTEINVGLYGAATVPANQG     645
NC_014976       SASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTVPLDAG     661
NC_017195       SASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTVPLDAG     653
                ...**: *   * *:.* :*:**  ** : : *      .. .*:*  : * .*** : *

NC_014479       DYIEIYCYCNYHGTDRRGVSDYNEVYNYIDIQELGGLNYPTV       813
NC_000964       DYVEIYVYVGYSGDVTRYVTDSNGALNYFDVLELGGRNYPRV       687
NC_022898       DYVEIYVYVGYSGDVTRYVTDSNGALNYFDVLELGGRNYPRV       687
NC_018520       DYVEIYVYVGYSGDVTRYVTDSNGALNYFDVLELGGRNYPRV       687
NZ_CP017314     DYIEIYLYVGYNGGTTRYTTESSGWYNYFDITEIGGRNYPRT       687
NC_014976       DYVEIYVYVGYSGDVTRYVTDSNGALNYFDVLELGGRNYPRV       703
NC_017195       DYVEIYVYVGYSGDITRYVTDSNGVLNYFDVLELGGRNYPRV       695
                **:*** * .* *   * .:: .   **:*: *:** *** .
```

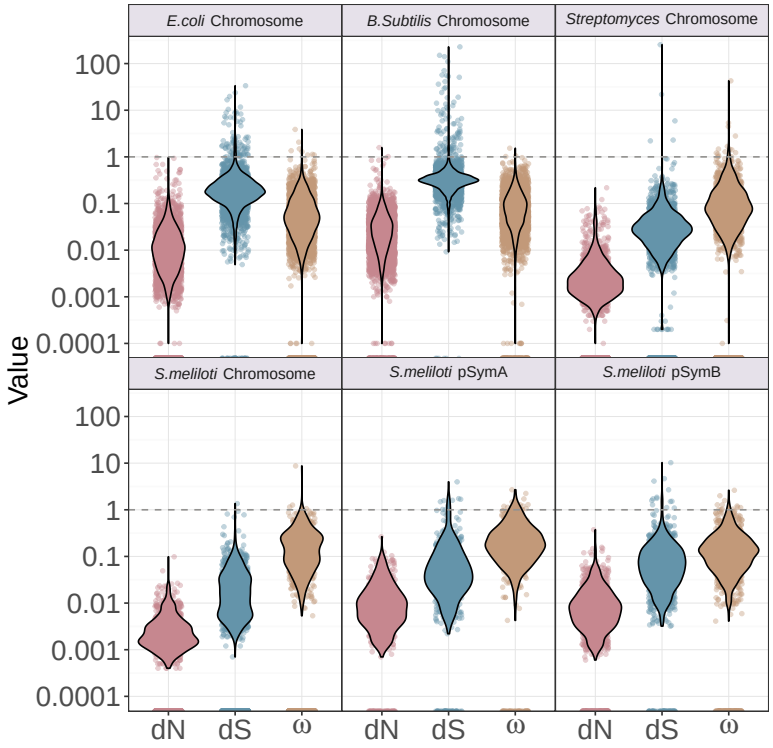


Figure S1.19: Distribution of all Non-synonymous Substitution Rate (dN), dS, and ω values on a log base 10 scale for each replicon. Individual points are shown as a strip chart (which has been jittered in the x-direction in `R` (Wickham et al. 2019)), and the density of these selection values is shown in the overlaid violin plot. All points are included in this graphic including outliers. For more information on how outliers were calculated, please see the main paper. Any dN, dS, or ω values that had a value of zero is pushed to the bottom of the x-axis. Since these values will not appear on a log base 10 scale, they are not included in the violin portions of this graphic. For a complete list of zero values in each of the selection categories please refer to Table S1.20. In these graphs there is a horizontal line of values at 0.0001 for most of the selection coefficients in most of the bacterial replicons. This is due to rounding practices when `codeml` (Yang 1997) calculates dN, dS, and ω values.

Bacteria and Replicon	Outliers (%)	Zero Value (%)		
		dN	dS	ω
E. coli Chromosome	7.49	13.82	1.05	13.82
B. subtilis Chromosome	5.41	4.40	0.16	4.40
Streptomyces Chromosome	4.74	25.70	14.48	25.70
S. meliloti Chromosome	17.05	61.21	59.26	61.21
S. meliloti pSymA	6.69	11.28	9.75	11.28
S. meliloti pSymB	6.13	13.20	5.20	13.20

Table S1.20: Percent of data that was calculated to be an outlier or had a selection variable (dN, dS, and ω) value of zero.

A0.14 Average dN, dS, and ω per Gene Values

The average dN, dS, and ω values per gene were calculated. For genes that were split into multiple parts (due to the presence of gaps or poor homology in the alignment), the dN, dS, and ω values for each gene part were averaged to obtain a single average value per gene. A complete list of these values can be found on `GitHub` (`www.github.com/dlato/Location_of_Substitutions_and_Bacterial_Arrangements`) under the file name "`Supplementary_table_per_gene_dN_dS_omega.pdf`".

A0.15 Window Analysis for dN, dS, and ω

Multiple linear regressions were performed to determine if there was any correlation between the average dN, dS, and ω values and distance from the origin of replication. A linear regression to determine how the average dN, dS, and ω values in various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) changes with genomic position was performed (Table S1.21). The results from these supplemental tests are consistent with the results from the linear regression found in the Main Paper, most bacterial replicons do not have a significant correlation between dN, dS, and ω values and distance from the origin of replication. Linear regressions that were significant, were inconsistent in sign.

A0.16 20Kbp Near and Far From Origin Selection Linear Regression Analysis

We additionally took a closer look at 20 genes close and far from the origin of replication. We performed a linear regression on the change in selection values (dN, dS, and ω) with distance from the origin of replication in these genes (Table S1.22). For majority of the bacterial replicons we failed to find a trend, which is not surprising since there was no evidence of an overall genomic trend when looking at these values (see Main Paper for results). Again, we are unable to conclude that there is a consistent overall trend for any of the selection values, dN, dS, and ω.

	Near Origin			Near Terminus		
Bacteria and Replicon	dN	dS	ω	dN	dS	ω
E. coli Chromosome	NS	NS	NS	NS	NS	NS
B. subtilis Chromosome	NS	NS	NS	NS	NS	NS
Streptomyces Chromosome	NS	NS	-9.36×10^{-7}* (0.328)	NS	NS	NS
S. meliloti Chromosome	NS	NS	NS	NS	NS	NS
S. meliloti pSymA	NS	NS	NS	-2.53×10^{-7}* (0.238)	NS	NS
S. meliloti pSymB	NS	6.19×10^{-6}** (0.372)	NS	NS	4.92×10^{-6}* (0.232)	NS

TABLE S1.22: Linear regression for dN, dS, and ω calculated for each bacterial replicon for the 20 genes closest and 20 genes farthest from the origin of replication. All results are marked with significance codes as followed: p: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', $>$ 0.05 = 'NS'. The R^2 values for each estimate are in brackets.

Bacteria and Replicon	Protein Coding Window Size					
	10Kbp	25Kbp	50Kbp	100Kbp	200Kbp	400Kbp
			d*S*			
E. coli Chromosome	NS (0.008)	NS (0.0168)	NS (0.0194)	NS (0.0332)	NS (0.0713)	NS (0.165)
B. subtilis Chromosome	NS (0.0057)	NS (0.0105)	NS (0.0198)	NS (0.0254)	NS (0.0743)	NS (0.113)
Streptomyces Chromosome	NS (0.002)	NS (0.00105)	NS (0.00139)	NS (0.00245)	NS (0.00401)	NS (0.00645)
S. meliloti Chromosome	NS (0.0143)	NS (0.0216)	NS (0.0293)	NS (0.0299)	NS (0.0676)	NS (0.111)
S. meliloti pSymA	NS (0.00775)	NS (0.0108)	NS (0.0177)	NS (0.0243)	NS (0.0315)	NS (0.912)
S. meliloti pSymB	NS (0.00582)	NS (0.0136)	NS (0.0164)	NS (0.0731)	NS (0.476)	NS (0.701)
			d*N*			
E. coli Chromosome	NS (0.0004)	NS (0.0001)	NS (0.0002)	-NS (0.0001)	NS (1.88×10^{-5})	NS (0.0132)
B. subtilis Chromosome	NS (0.0164)	NS (0.0365)	NS (0.0614)	NS (0.0685)	NS (0.127)	NS (0.15)
Streptomyces Chromosome	NS (0.00376)	NS (0.00196)	NS (0.00454)	NS (0.0005)	NS (0.00385)	NS (0.0154)
S. meliloti Chromosome	NS (0.0178)	NS (0.0213)	NS (0.0247)	NS (0.0245)	NS (0.0565)	NS (0.0836)
S. meliloti pSymA	NS (0.00671)	NS (0.00433)	NS (0.0128)	NS (0.0599)	NS (0.0329)	NS (0.736)
S. meliloti pSymB	NS (0.0001)	NS (2.4×10^{-6})	NS (0.0005)	NS (0.00311)	NS (0.128)	NS (0.24)
			ω			
E. coli Chromosome	5.22×10^{-9}*** (0.061)	4.62×10^{-9}*** (0.11)	5.62×10^{-9}*** (0.174)	4.96×10^{-9}** (0.296)	4.8×10^{-9}* (0.363)	3.51×10^{-9}* (0.51)
B. subtilis Chromosome	NS (0.0084)	NS (0.0281)	NS (0.0348)	NS (0.0185)	NS (0.0255)	NS (0.0179)
Streptomyces Chromosome	2.12×10^{-9}** (0.0104)	NS (0.0115)	1.98×10^{-9}* (0.0312)	NS (0.0308)	NS (0.0654)	NS (0.144)
S. meliloti Chromosome	-1.66×10^{-9}* (0.0278)	NS (0.0327)	NS (0.0337)	NS (0.0238)	NS (0.0383)	NS (0.0416)
S. meliloti pSymA	NS (0.00218)	NS (0.00326)	NS (0.00657)	NS (0.426)	NS (0.511)	NS (0.607)
S. meliloti pSymB	NS (0.0239)	NS (0.0662)	NS (0.098)	NS (0.002)	NS (0.634)	2.74×10^{-8}** (1)

Table S1.21: Linear regression on various sections of the genome (10Kbp, 25Kbp, 50Kbp, 100Kbp, 200Kbp, and 400Kbp) with increasing distance from the origin of replication after accounting for bidirectional replication. The linear regression was performed on the average dN, dS, and ω values in each section of the genome. All results are marked with significance codes as followed: $<$ 0.001 = '***', 0.001 $<$ 0.01 = '**', 0.01 $<$ 0.05 = '*', $>$ 0.05 = 'NS'. The R^2 value for each coefficient estimate is found below the value in brackets ().

Appendix B

Chapter 3 Supplementary Files

Title: Spatial Patterns of Gene Expression in Bacterial Genomes

Authors: Daniella F Lato and G Brian Golding

Journal: Journal of Molecular Evolution

DOI: https://doi.org/10.1007/s00239-020-09951-3

Corresponding Author Information:
G. Brian Golding
McMaster Univeristy
Department of Biology
1280 Main St. West
Hamilton, ON
Canada
L8S 4K1
Email: golding@mcmaster.ca

All supplemental information including interactive graphs of the expression data, and pdf versions of all figures from the paper and supplement can be found on GitHub at https://github.com/dlato/Spatial_Patterns_of_Gene_Expression.git.

A0.1 Interactive Graphs

The normalized gene expression data is available as a interactive graph for each bacterial replicon. The user can use their mouse to hover over gene expression points to determine the National Centre for Biotechnology Information (NCBI) gene Id. This Id can be searched in the NCBI website (https://www.ncbi.nlm.nih.gov/) to obtain more information on that particular gene. These interactive graphs are listed on GitHub as files:

- *Escherichia coli*: "ecoli_gene_exp_interactive_graph.html"
- *Bacillus subtilis*: "bsubtilis_gene_exp_interactive_graph.html"
- *Streptomyces*: "streptomyces_gene_exp_interactive_graph.html"
- *Sinorhizobium meliloti* Chromosome : "smeliloti_chromosome_gene_exp_interactive_graph.html"
- *S. meliloti* pSymA: "smeliloti_pSymA_gene_exp_interactive_graph.html"
- *S. meliloti* pSymB "smeliloti_pSymB_gene_exp_interactive_graph.html"

A0.2 Gene Expression Data

Bacteria Strain/Species	GEO Accession Number	Date Accessed	NCBI Accession Genome Used For Gene Position
E. coli K12 MG1655	GSE60522	December 20, 2017	U00096
E. coli K12 MG1655	GSE73673	December 19, 2017	
E. coli K12 MG1655	GSE85914	December 19, 2017	
E. coli K12 DH10B	GSE98890	December 19, 2017	
B. subtilis 168	GSE104816	December 14, 2017	NC_000964
B. subtilis 168	GSE67058	December 16, 2017	
B. subtilis 168	GSE93894	December 15, 2017	
Streptomyces coelicolor A3	GSE57268	March 16, 2018	AL645882
S. meliloti 1021 Chromosome	GSE69880	December 12, 2017	NC_003047
S. meliloti 1021 pSymA	GSE69880	December 12, 2017	NC_003037
S. meliloti 1021 pSymB	GSE69880	December 12, 2017	NC_003078

Supplemental Table S2.1: Strains and species used for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided. NCBI genome accession numbers are listed for which genome was used to determine the gene position.

A0.3 Origin and Terminus Locations

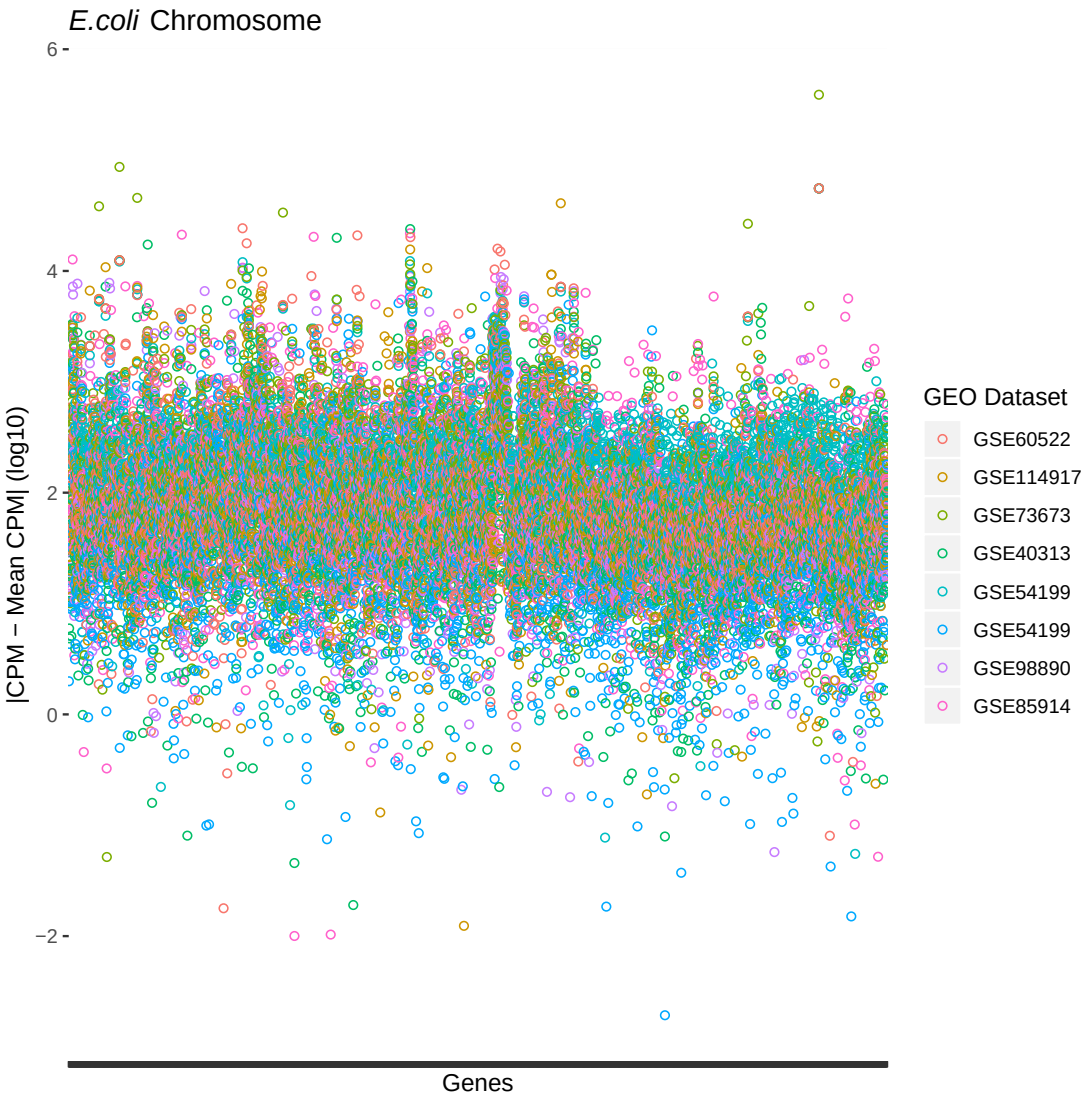
Bacteria	Origin of Replication	Terminus of Replication	Replicon Length (bp)
E. coli	3925744	1588773	4641652
B. subtilis	1	1942542	4215606
Streptomyces	3419363	1 & 8667507	8667507
S. meliloti Chromosome	1	1735626	3654135
S. meliloti pSymA	1350001	672888	1354226
S. meliloti pSymB	55090	896756	1683333

Supplemental Table S2.2: Origin of replication and terminus of replication positions in replicons of representative strains of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. The linear nature of *Streptomyces* chromosome gives it two termini, one at each end of the chromosome. The total base pair length for each bacterial replicon is additionally listed. Representative strain NCBI Accession Number can be found in Supplementary Table S2.1.

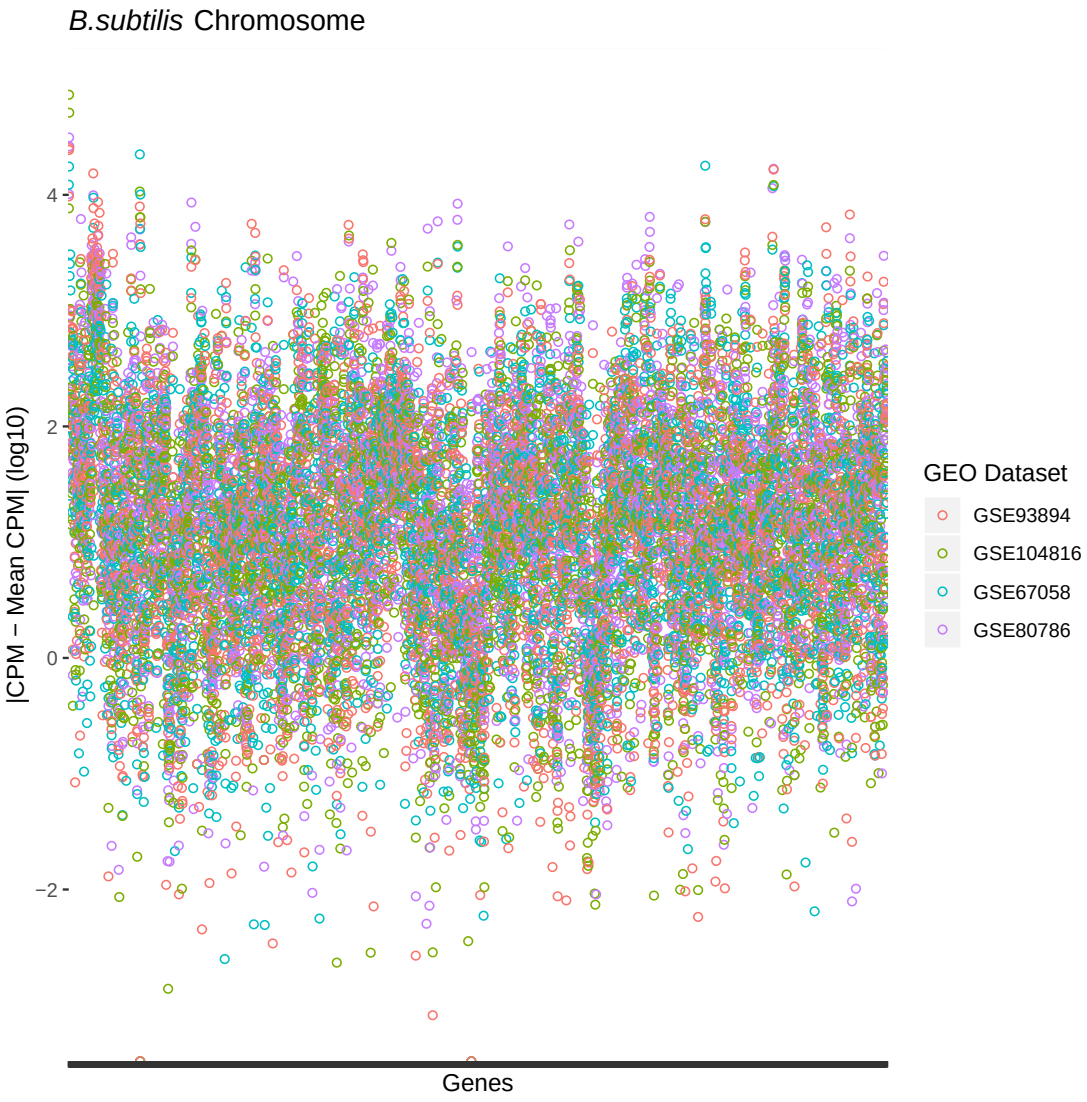
A0.4 Correlation of Gene Expression Over Datasets

To assess uniform expression over bacteria with multiple data sets we looked at the mean normalized expression values. Multiple replicates from a data set were combined by finding the median normalized Counts Per Million (CPM) expression value for each gene. This was done for any data sets that had

multiple replicates. For each gene (x_i) the mean normalized expression value was calculated across all data sets ($\bar{x}_{ij}$). Then the normalized median expression value for each data set was subtracted from the mean across all expression values ($|x_{ij} - \bar{x}_{ij}|$). The distribution of these $|x_{ij} - \bar{x}_{ij}|$ across all genes are found in Figures S2.1 and S2.2. All data sets are well mixed, implying that the expression levels are consistent across all data sets. Only *E. coli* and *B. subtilis* had multiple expression datasets available so they are the only ones that were analyzed. *Streptomyces* and all replicons of *S. meliloti* had only one data set each and therefore were not analyzed.



Supplemental Figure S2.1: Dot plot distribution of the median expression value for each *E. coli* data set minus the mean expression value for that gene across all data sets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis. The values are coloured by GEO data set.



Supplemental Figure S2.2: Dot plot distribution of the median expression value for each *B. subtilis* data set minus the mean expression value for that gene across all data sets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis. The values are coloured by GEO data set.

A0.5 Additional Linear Regression Tests

Multiple more detailed linear regressions are performed to determine if there is any correlation between gene expression per gene and distance from the origin of replication. A linear regression to determine how the median CPM expression values per gene changes with genomic position was performed. Additionally, a linear regression to determine how the median CPM expression value for each 10Kbp section of the genome changes with genomic position was performed. Finally, a linear regression to determine how the total added expression over each 10Kbp region of the genome changes with genomic position was performed. All linear regression results mirror the results from the linear regression on the median gene expression CPM value per gene. Most bacteria have a negative correlation, implying that gene expression tends to decrease with distance from the origin of replication.

We additionally performed a linear regression on a per gene basis. We found similar results as the linear regression of average expression values over 10Kbp regions: *E. coli* and *B. subtilis* had gene expression decrease with increasing distance from the origin of replication (Supplementary Table: S2.3). We were unable to detect a significant trend between gene expression and genomic position in the majority of the other bacterial replicons (Supplementary Table: S2.3). We performed a further linear regression tests on the median CPM gene expression value per 10Kbp region of the genome. This was calculated by determining the median CPM expression value across all genes in 10Kbp regions of the genome. We were able to detect similar results as the linear regression of average expression values over 10Kbp regions in *E. coli*, where median gene expression decreases with increasing distance from the origin of replication (Supplementary Table: S2.4). For all of the other bacterial replicons we were unable to determine a significant trend between median gene expression and genomic position (Supplementary Table: S2.4). Finally, we performed a linear regression test on the total additive CPM gene expression value per 10Kbp region of the genome. This was calculated by summing all gene CPM expression values across 10Kbp regions of the genome. We were able to detect similar results as the linear regression of average expression values over 10Kbp regions in most bacterial replicons where total gene expression decreases with increasing distance from the origin of replication (Supplementary Table: S2.5). For the two secondary replicons of *S. meliloti*, we were unable to detect a significant trend between total gene expression and genomic position (Supplementary Table: S2.5).

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-2.95×10^{-5}	1.29×10^{-5}	3.00×10^{-6}
B. subtilis Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-6}
Streptomyces Chromosome	-1.15×10^{-6}	8.12×10^{-8}	NS
S. meliloti Chromosome	2.85×10^{-5}	4.09×10^{-5}	NS
S. meliloti pSymA	1.39×10^{-3}	2.54×10^{-4}	5.48×10^{-8}
S. meliloti pSymB	1.47×10^{-4}	2.03×10^{-4}	NS

Supplemental Table S2.3: Linear regression analysis of the median counts per million expression values per gene along the genome of the respective bacteria replicons. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. NS indicates Not Significant at $P \leq 0.05$. A grey row indicates a significant negative trend.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-1.53×10^{-5}	3.91×10^{-6}	1.21×10^{-4}
B. subtilis Chromosome	-4.04×10^{-6}	2.82×10^{-6}	NS
Streptomyces Chromosome	-6.29×10^{-7}	3.27×10^{-8}	NS
S. meliloti Chromosome	2.19×10^{-6}	8.05×10^{-6}	NS
S. meliloti pSymA	-1.92×10^{-6}	1.03×10^{-4}	NS
S. meliloti pSymB	7.46×10^{-5}	7.03×10^{-5}	NS

Supplemental Table S2.4: Linear regression analysis of the median counts per million expression data for 10Kbp segments of the genome of the respective bacteria replicons. Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. Statistical outliers were removed from this linear regression calculation. NS indicates Not Significant at $P \leq 0.05$. A grey row indicates a significant negative trend.

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
E. coli Chromosome	-3.41×10^{-4}	1.11×10^{-4}	2.47×10^{-3}
B. subtilis Chromosome	-5.63×10^{-4}	1.79×10^{-4}	1.87×10^{-3}
Streptomyces Chromosome	-1.37×10^{-6}	4.59×10^{-7}	2.88×10^{-3}
S. meliloti Chromosome	-6.97×10^{-4}	2.70×10^{-4}	1.08×10^{-2}
S. meliloti pSymA	9.04×10^{-3}	5.93×10^{-3}	NS
S. meliloti pSymB	-1.72×10^{-3}	2.30×10^{-3}	NS

Supplemental Table S2.5: Linear regression analysis of total added expression and distance from the origin of replication. The total added expression values were calculated by summing the total counts per million expression value per 10Kbp section of the genome . Linear regression was calculated after the origin of replication was moved to the beginning of the genome and all subsequent positions were scaled around the origin accounting for bidirectional replication. NS indicates Not Significant at $P \leq 0.05$. A grey row indicates a significant negative trend.

A0.6 Leading and Lagging Strand

A two-sample Wilcox test was computed to compare expression of genes on the leading strand and the lagging strand. We found that there was no significant difference between gene expression on the leading and lagging strand of any of the bacterial replicons.

Bacteria and Replicon	W	P-value	% of Genes on Leading Strand
E. coli Chromosome	1398352	0.9356	55.0
B. subtilis Chromosome	1678990.5	0.5736	73.8
Streptomyces Chromosome	7920836.5	1.75×10^{-5}	53.9
S. meliloti Chromosome	1462420	0.0124	55.6
S. meliloti pSymA	194005	0.3266	59.5
S. meliloti pSymB	297056.5	0.4736	55.9

Supplemental Table S2.6: Two-sample Wilcox test results to determine if gene expression is significantly different between the leading and lagging strands of each bacterial replicon. The percentage of genes on the leading strand was also computed.

A0.7 COG Analysis

A supplementary analysis of the spatial distribution of Clusters of Orthologous Groups of proteins (COG) categories for each bacterial replicon was performed. For a full list of COG categories, please refer to Table S2.7.

This supplementary analysis shows that there appears to be no clear COG categories that are universally increasing or decreasing among the bacterial replicons in this analysis.

COG Data

Whole genomes of different strains and species of *E. coli*, *B. subtilis*, *Streptomyces* and *S. meliloti* were downloaded (Table S2.8). The analysis was performed on each replicon of multi-repliconic bacteria. For *S. meliloti* the analysis was performed on each of its replicons separately. The COG database information was downloaded on February 27, 2017 and spans the years 2003-2014. This data can be found on `GitHub` at (`https://github.com/dlato/Spatial_Patterns_of_Gene_Expression.git`) The only available data in the COG database for *Streptomyces* was for *Streptomyces bingchenggensis* and not *S. coelicolor*. We were therefore limited to using the annotation for *S. bingchenggensis*.

Using simple Python scripts, the COG protein ID and functional category was obtained for each known protein of each bacterial replicon in this analysis. This information was combined with the GenBank acession number and protein genome location to obtain the functional category of each protein and its midpoint location in the genome. The midpoint of each protein was calculated to be the singular point between the start and the end of the protein. This calculation was done to simplify the statistical calculations to verify the spatial trends of each COG category.

The origin and terminus of replication location, and bidirectional nature of bacterial replication were accounted for using the same methods as in the Gene Expression analysis. See "The Spatial Patters of Gene Expression in Bacterial Genomes" main paper for detailed methods.

COG Statistical Analysis

To determine if each COG category increased or decreased with increasing distance from the origin, a logistic regression was performed on each COG category for each replicon. Each of the proteins was considered present (1) or absent (0) in each COG category. Proteins that were classified under more

COG Abbreviation	COG Category
A	RNA Processing and Modification
B	Chromatin Structure and Dynamics
C	Energy Production and Conversion
D	Cell Cycle Control and Mitosis
E	Amino Acid Transport and Metabolism
F	Nucleotide Transport and Metabolism
G	Carbohydrate Transport and Metabolism
H	Coenzyme Metabolis
I	Lipid Metabolism
J	Translation
K	Transcription
L	Replication and Repair
M	Cell Wall/Membrane/Envelope Biogenesis
N	Cell Motility
O	Post-translational Modification, Protein Turnover, Chaperone Functions
P	Inorganic Ion Transport and Metabolism
Q	Secondary Structure
T	Signal Transduction
U	Intracellular Trafficking and Secretion
V	Defence Mechanisms
W	Extracellular Structures
X	Mobilome: Prophages, Transposons
Y	Nuclear Structure
Z	Cytoskeleton
R	General Function Prediction Only
S	Function Unknown

Supplemental Table S2.7: List of COG category letter abbreviation and full name of COG functional protein category.

than one COG category had a present (1) data point for each COG category. The binary nature of the COG data allowed for a simple logistic regression to be performed for each COG category using R. Logistic regression results are found in Table S2.9.

A visualization of the proportional distribution of the COG categories for each replicon can be seen in Figures:S2.3-S2.8.

A0.8 COG Logistic Regression Results

Bacteria Strain	Accession Number	Date Accessed
E. coli K12	U00096	September 26, 2016
B. subtilis 168	NC_000964	November 10, 2016
S. bingchenggensis BCW1	CP002047	June 7, 2017
S. meliloti 1021	NC_003047	June 3, 2014

Supplemental Table S2.8: List of bacteria genomes used for the COG category information. This includes the accession number and date accessed.

COG Category	*E. coli* Chromosome	*B. subtilis* Chromosome	*Streptomyces* Chromosome	*S. meliloti* Chromosome	*S. meliloti* pSymA	*S. meliloti* pSymB
RNA Processing and Modification	NS	NS	NS	NS	NS	NS
Chromatin Structure and Dynamics	NS	NS	NS	NS	NS	NS
Energy Production and Conversion	2.40×10^{-7}	4.10×10^{-7}	-2.94×10^{-7}	NS	NS	NS
Cell Cycle Control and Mitosis	NS	NS	NS	NS	NS	NS
Amino Acid Transport and Metabolism	4.53×10^{-7}	NS	-1.97×10^{-7}	2.66×10^{-7}	-2.08×10^{-6}	-9.45×10^{-7}
Nucleotide Transport and Metabolism	NS	-7.49×10^{-7}	-1.86×10^{-7}	-6.68×10^{-7}	-1.15×10^{-6}	NS
Carbohydrate Transport and Metabolism	NS	NS	NS	-2.53×10^{-7}	9.78×10^{-7}	2.05×10^{-6}
Coenzyme Metabolis	NS	-4.07×10^{-7}	-1.11×10^{-7}	-1.20×10^{-6}	-9.83×10^{-7}	-1.45×10^{-6}
Lipid Metabolism	4.51×10^{-7}	3.74×10^{-7}	-2.01×10^{-7}	NS	2.01×10^{-6}	1.84×10^{-6}
Translation	NS	-7.13×10^{-7}	-1.36×10^{-7}	1.23×10^{-6}	-1.51×10^{-6}	-1.15×10^{-6}
Transcription	2.22×10^{-7}	7.62×10^{-7}	NS	NS	NS	-4.17×10^{-6}
Replication and Repair	2.95×10^{-7}	NS	-1.17×10^{-7}	NS	1.42×10^{-6}	NS
Cell Wall/Membrane/Envelope Biogenesis	NS	5.18×10^{-7}	-8.05×10^{-8}	4.59×10^{-7}	1.63×10^{-6}	5.41×10^{-6}
Cell Motility	-7.74×10^{-7}	1.01×10^{-6}	-2.04×10^{-7}	NS	NS	NS
Post-translational Modification, Protein Turnover, Chaperone Functions	3.37×10^{-7}	3.51×10^{-7}	-7.75×10^{-8}	3.47×10^{-7}	NS	1.08×10^{-6}
Inorganic Ion Transport and Metabolism	NS	NS	-1.68×10^{-7}	5.36×10^{-7}	NS	-2.05×10^{-6}
Secondary Structure	NS	NS	NS	NS	4.28×10^{-6}	3.81×10^{-6}
Signal Transduction	NS	NS	1.52×10^{-7}	1.85×10^{-6}	NS	NS
Intracellular Trafficking and Secretion	NS	NS	NS	8.62×10^{-7}	NS	NS
Defence Mechanisms	3.75×10^{-7}	7.15×10^{-7}	-1.21×10^{-7}	4.24×10^{-7}	NS	NS
Extracellular Structures	-3.23×10^{-6}	NS	9.06×10^{-7}	NS	NS	NS
Mobilome: Prophages, Transposons	-1.09×10^{-6}	-1.81×10^{-6}	4.32×10^{-7}	1.67×10^{-6}	NS	-3.35×10^{-6}
Nuclear Structure	NS	NS	NS	NS	NS	NS
Cytoskeleton	NS	NS	NS	NS	NS	NS
General Function Prediction Only	2.61×10^{-7}	3.20×10^{-7}	-6.91×10^{-8}	9.49×10^{-7}	NS	NS
Function Unknown	-1.43×10^{-6}	-1.19×10^{-6}	4.62×10^{-7}	-7.44×10^{-7}	2.53×10^{-5}	4.14×10^{-5}

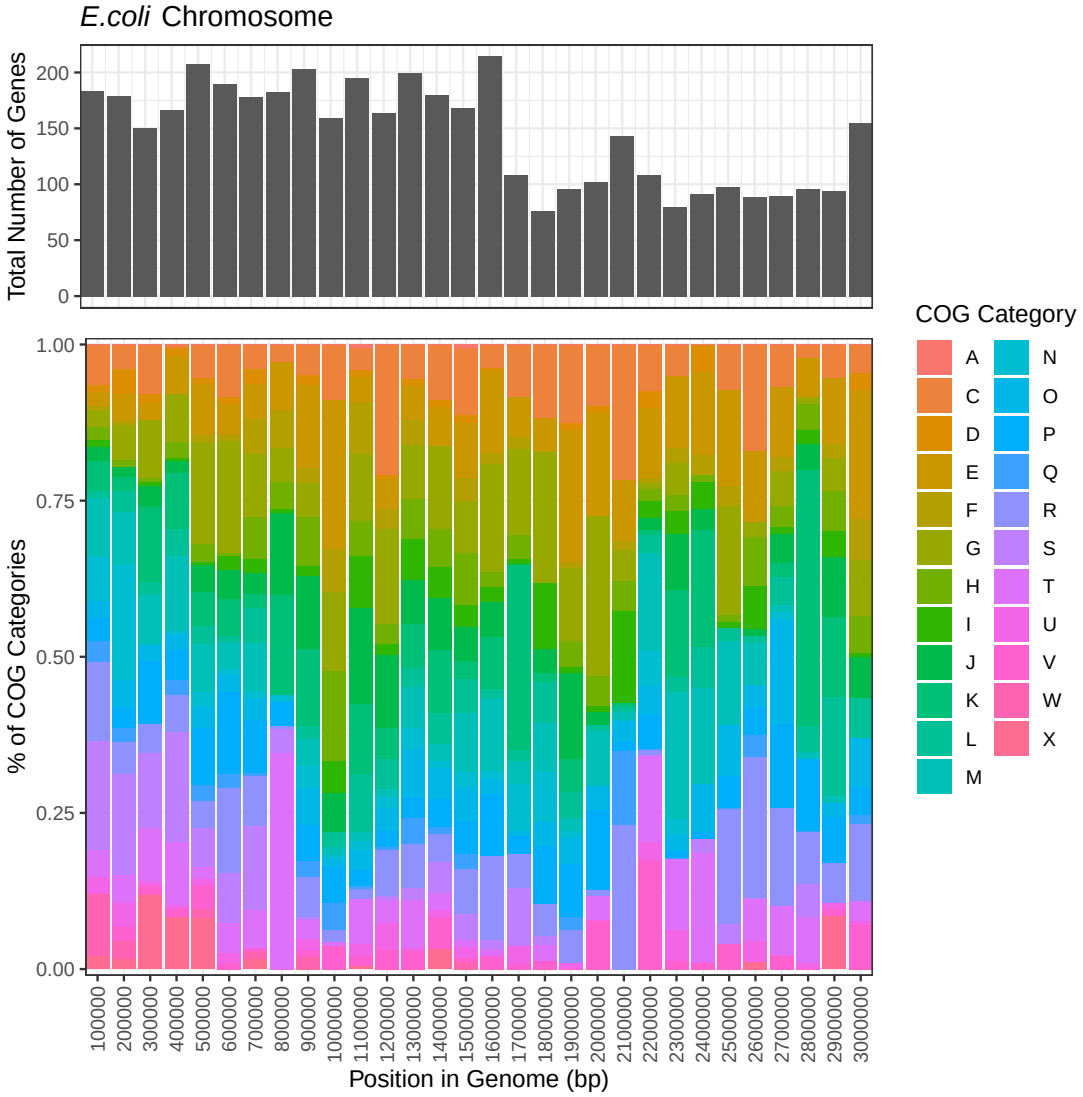
Supplemental Table S2.9: Logistic regression coefficients for each bacterial replicon analysis showing the change in each COG category with increasing distance from the origin of replication. Only statistically significant ($p < 0.05$) coefficient estimates are shown in the table. Any values of NS did not have a statistically significant p-value. Grey cells indicate logistic regression coefficients that were negative.

A0.9 High Gene Expression Distribution

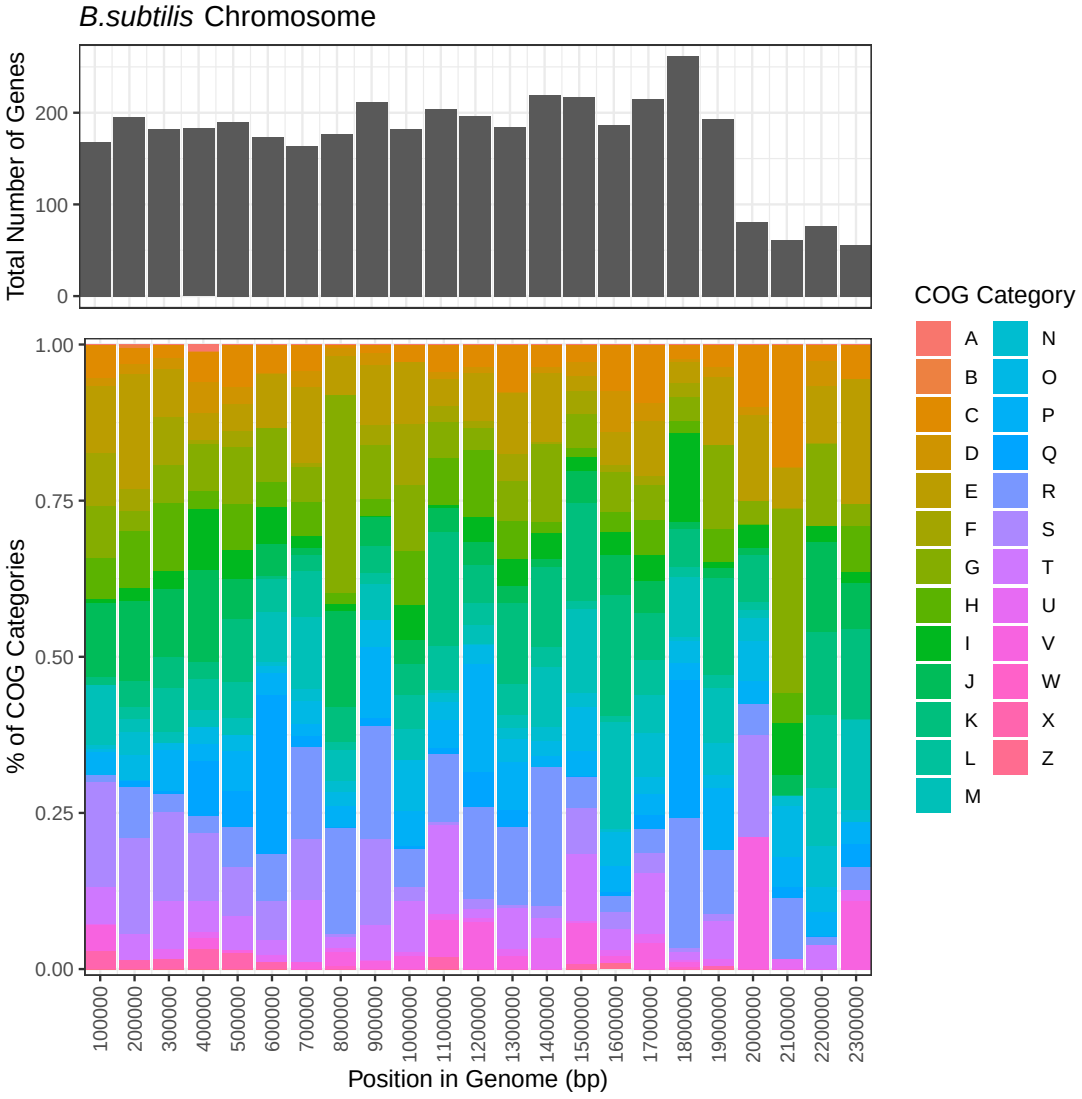
Bacteria and Replicon	Bidirectional Genomic Position (bp)	Protein/Gene Examples
E. coli Chromosome	0 - 10000	DNA replication and repair
		ATP-proton motive force
		ATP biosynthesis
		transport
	470000 - 480000	DNA replication and repair
		tRNA synthesis
		Ribosomal proteins
		Putative transport
	610000 - 620000	Ribosomal protein
		Translation modification
		tRNA modification
		RNA synthesis
	840000 - 850000	Energy metabolism
	1170000 - 1180000	Cell division
		Protein synthesis modification
B. subtilis Chromosome	0 - 10000	tRNA modification
		Ribosomal proteins
		DNA gyrase
		rRNA small subunit methylation
	130000 - 150000	Ribosomal proteins
		Elongation factor
	730000 - 740000	tRNA subunit
		Transcription regulation
		Glycolysis
Streptomyces Chromosome	1700000 - 1720000	Ribosomal proteins
		RNA Polymerase alpha chain
	1200000 - 1210000	Possible ATP-binding proteins
		Putative oxidoreductase
		Integral membrane proteins
	-2900000 - -2890000	Putative peptide synthetase
S. meliloti Chromosome	630000 - 640000	Cell processes
		Structural Elements
	1480000 - 1490000	Ribosomal proteins
		Structural elements
		Transmembrane proteins
S. meliloti pSymA	0 - 20000	Cell processes
		Hypothetical proteins
	660000 - 680000	Small molecule metabolism
		Not classified regulator
		Glimmer prediction

		Hypothetical protein
S. meliloti pSymB	210000 - 220000	Unknown proteins
		Cell processes
		Hypothetical proteins
	290000 - 300000	Cell Division
		Small molecule metabolism
		Cell processes
	790000 - 820000	Small molecule metabolism
		Cell processes

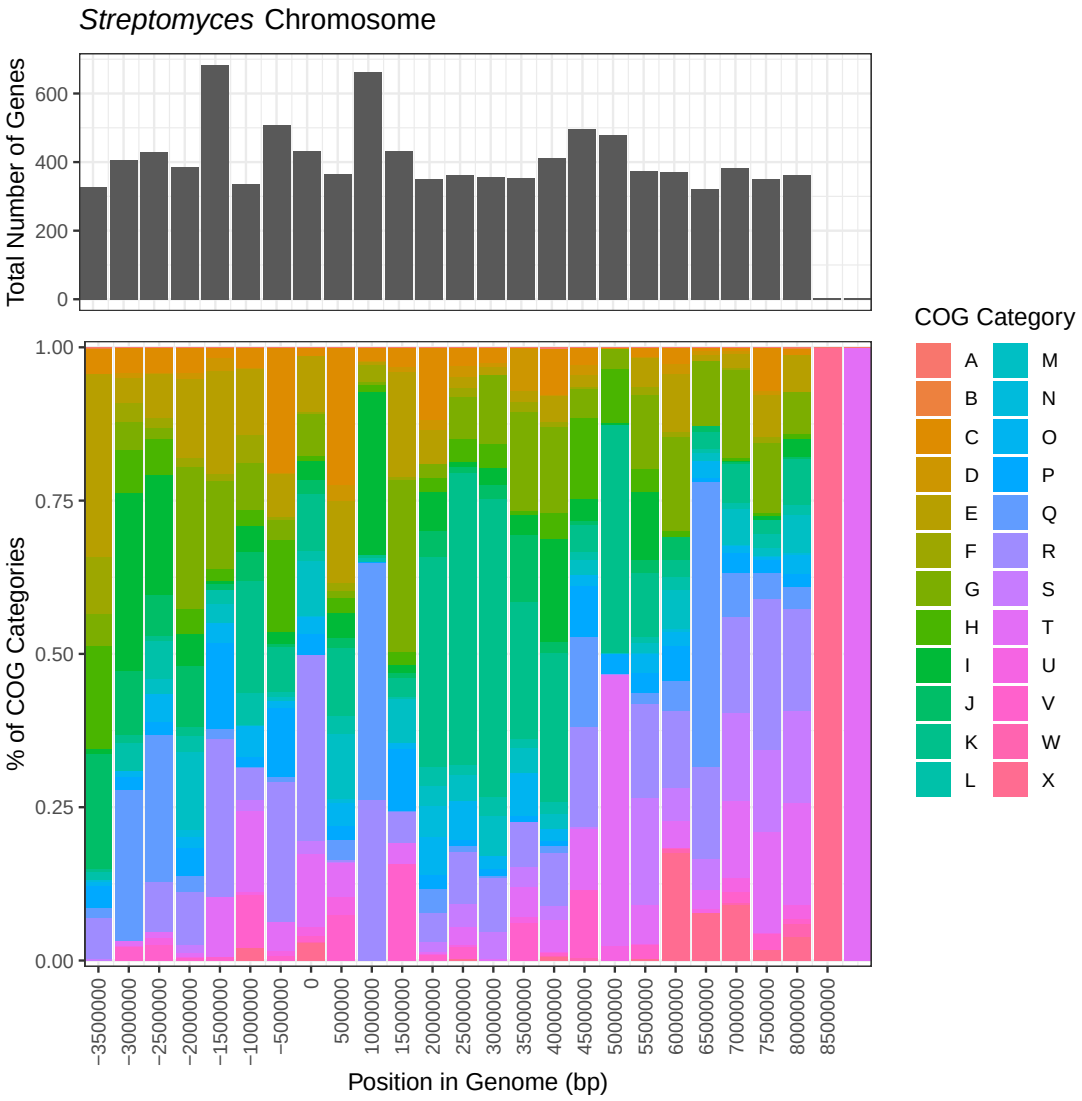
Supplemental Table S2.10: Table of high median CPM gene expression over 10Kbp genomic regions for each bacterial replicon and the associated proteins/gene functions found in that region. The genomic position begins at the origin of replication and continues in both directions until the terminus of replication (bidirectional replication).



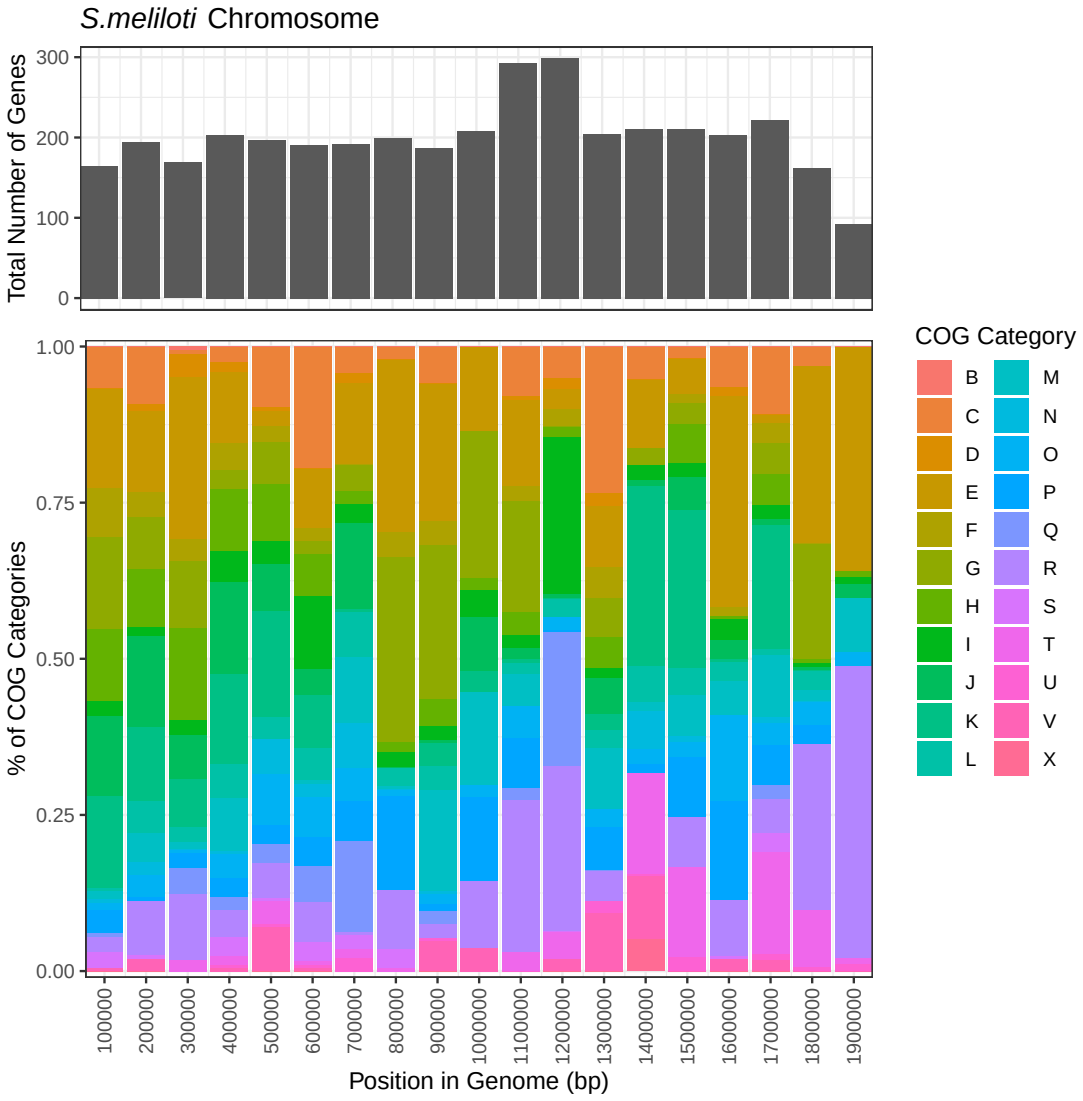
Supplemental Figure S2.3: Graphical representation of COG categories across the chromosome of *E. coli*. Bidirectional distance from the origin of replication is along the x-axis. Each bar represents a 100Kbp segment of the genome. The grey graph represents the total number of genes in each 100Kbp section of the genome. The colourful graph represents the percentage of COG categories in each 100Kbp section of the genome. The full name for each COG category can be found in Table S2.7.



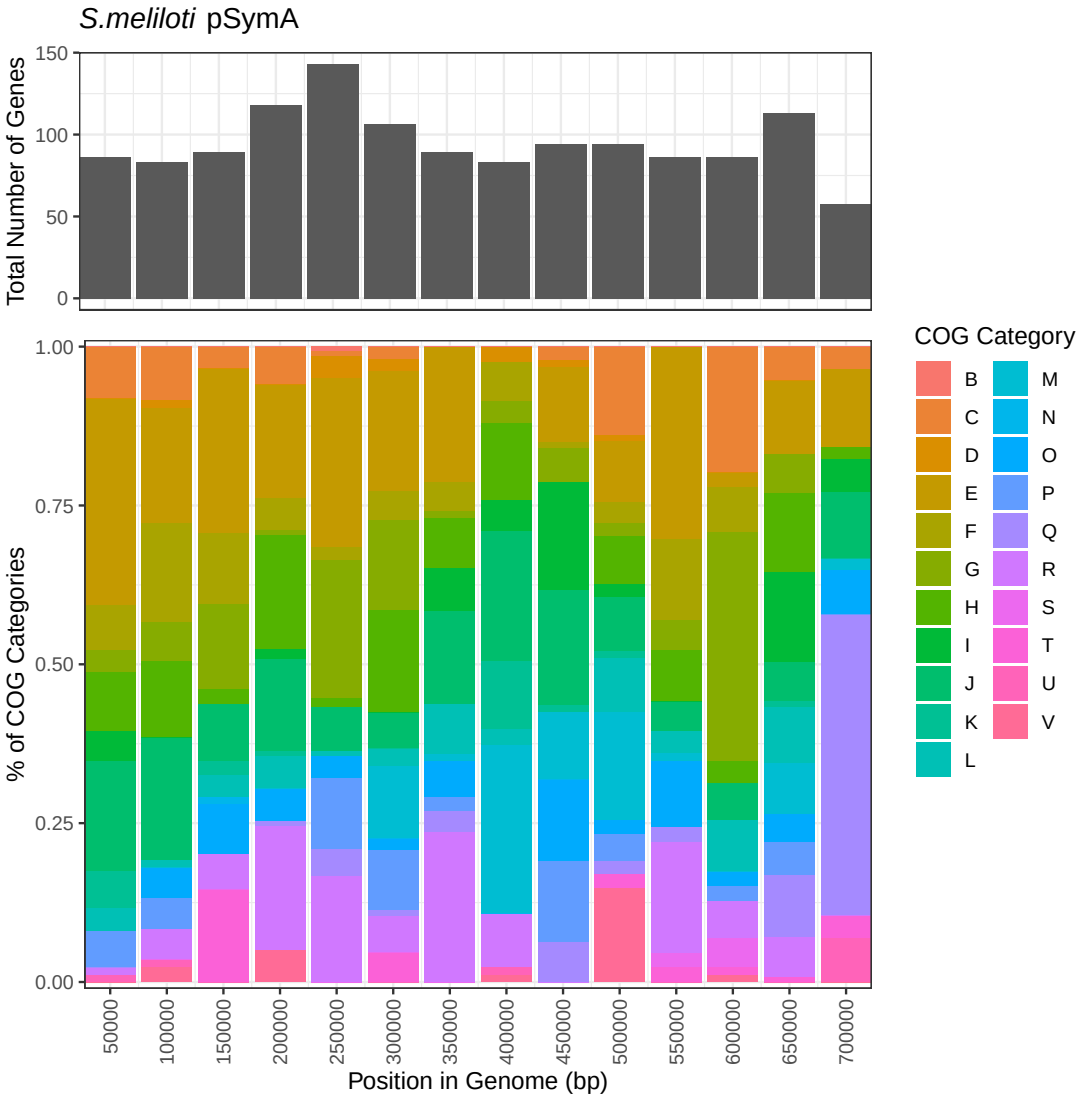
Supplemental Figure S2.4: Histogram of COG categories across the chromosome of *B. subtilis*. Bidirectional distance from the origin of replication is along the x-axis. Each bar represents a 100Kbp segment of the genome. The grey graph represents the total number of genes in each 100Kbp section of the genome. The colourful graph represents the percentage of COG categories in each 100Kbp section of the genome. The full name for each COG category can be found in Table S2.7.



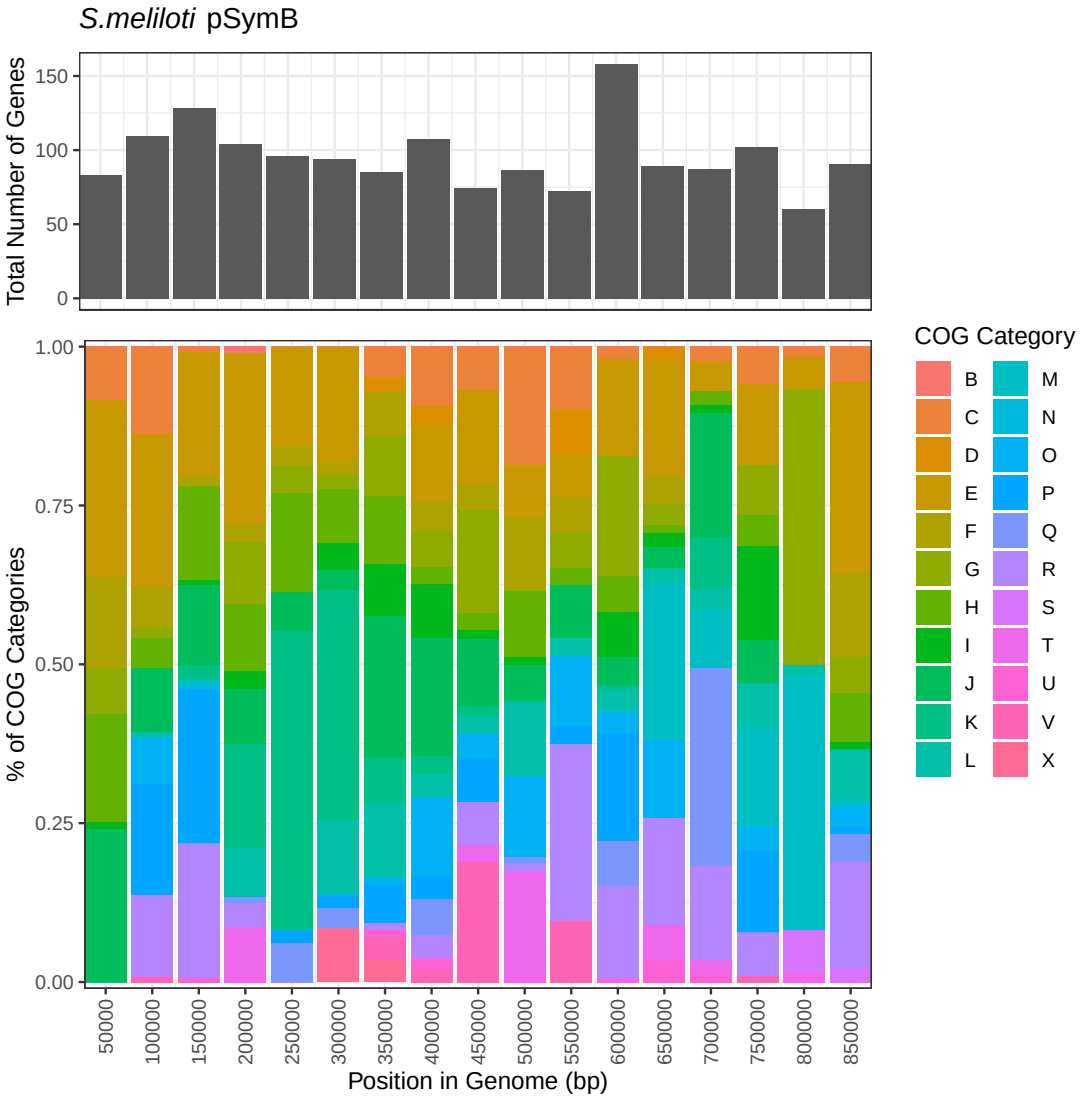
Supplemental Figure S2.5: Histogram of COG categories across the chromosome of *Streptomyces*. Distance from the origin of replication is along the x-axis with the origin of replication denoted by position 0. The genome located on the shorter chromosome arm (to the left of the origin) has been given negative values, while the genome on the longer chromosome arm (to the right of the origin) has been given positive values. Each bar represents a 500Kbp segment of the genome. The grey graph represents the total number of genes in each 500Kbp section of the genome. The colourful graph represents the percentage of COG categories in each 500Kbp section of the genome. The two bars on the far right side of the graph have only one COG category present due to under representation of annotated genes in those sections of the genome. The full name for each COG category can be found in Table S2.7.



Supplemental Figure S2.6: Histogram of COG categories across the chromosome of *S. meliloti*. Bidirectional distance from the origin of replication is along the x-axis. Each bar represents a 100Kbp segment of the genome. The grey graph represents the total number of genes in each 100Kbp section of the genome. The colourful graph represents the percentage of COG categories in each 100Kbp section of the genome. The full name for each COG category can be found in Table S2.7.



Supplemental Figure S2.7: Histogram of COG categories across pSymA of *S. meliloti*. Bidirectional distance from the origin of replication is along the x-axis. Each bar represents a 50Kbp segment of the genome. The grey graph represents the total number of genes in each 50Kbp section of the genome. The colourful graph represents the percentage of COG categories in each 50Kbp section of the genome. The full name for each COG category can be found in Table S2.7.



Supplemental Figure S2.8: Histogram of COG categories across pSymA of *S. meliloti*. Bidirectional distance from the origin of replication is along the x-axis. Each bar represents a 50Kbp segment of the genome. The grey graph represents the total number of genes in each 50Kbp section of the genome. The colourful graph represents the percentage of COG categories in each 50Kbp section of the genome. The full name for each COG category can be found in Table S2.7.

Appendix C

Chapter 4 Supplementary Files

Title: Genomic Inversions in *Escherichia coli* Alter Gene Expression

Authors: Daniella F. Lato, Qing Zeng and G. Brian Golding

Journal: Formatted for submission to GENOME

Corresponding Author Information:
G. Brian Golding
McMaster University
Department of Biology
1280 Main St. West
Hamilton, ON
Canada
L8S 4K1
Email: golding@mcmaster.ca

Supplementary Material

For the most up to date Supplementary Material, please visit GitHub at https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/.

Further supplemental information and code are available on GitHub at https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/.

A1 Gene Expression Data

Strain	GEO Accession Number	Date Accessed	NCBI Accession Genome Used For Gene Position
E. coli K12 MG1655	GSE60522	December 20, 2017	U00096
	GSE114917	November 26, 2018	
	GSE54199	December 18, 2019	
	GSE40313	November 21, 2018	
E. coli K12 DH10B	GSE98890	March 13, 2018	NC_010473
E. coli BW25113	GSE73673	December 19, 2017	NZ_CP009273
	GSE85914	December 19, 2017	
E. coli ATCC 25922	GSE94978	November 23, 2018	NZ_CP009072 BA000007

Supplemental Table S3.1: Strains and species used for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided. NCBI genome accession numbers are listed for which genome was used to determine the gene position. Strains with multiple NCBI genome accession numbers had multiple genome versions/builds used to determine the genomic position.

A2 Sequences

Strain	Accession Number	Date(s) Accessed
E. coli K-12 MG1655 *	U00096	September 26, 2016
E. coli K-12 DH10B	NC_010473	February 13, 2020
E. coli BW25113	NZ_CP009273	October 3, 2018
E. coli ATCC 25922	NZ_CP009072	December 18, 2018

Supplemental Table S3.2: *E. coli* strains used for the analysis. Accession numbers and date accessed for each genome are provided. Multiple dates and accession numbers for one strain denote updated versions of the genome. An astrix (*) insicates the strain that was used as the representative strain.

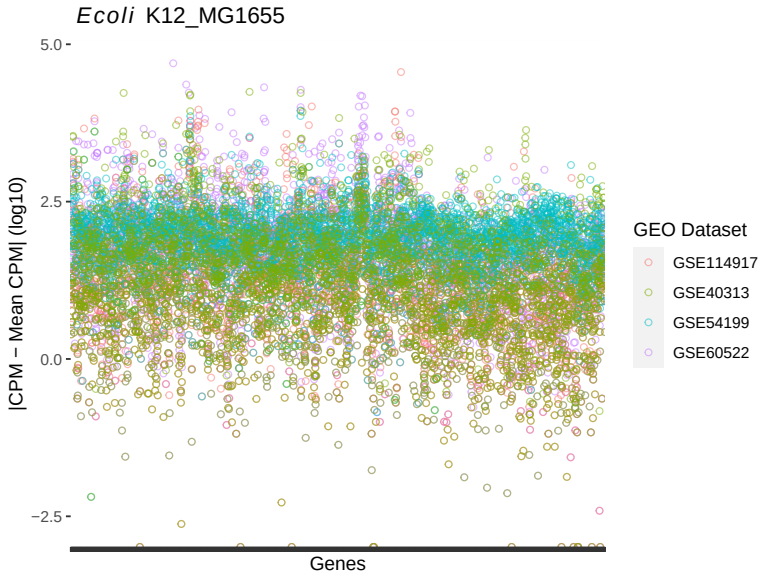
A3 Proteomes

Strain	UniProt Accession Number	NCBI Accession Number	Date(s) Accessed
E. coli K-12 MG1655	UP000000625	U00096	May 4, 2020
E. coli K-12 DH10B	UP000001689	NC_010473	May 4, 2020
E. coli BW25113	UP000029103	NZ_CP009273	May 4, 2020
E. coli ATCC 25922	UP000001410	NZ_CP009072	May 4, 2020

Supplemental Table S3.3: Proteomes used for the *E. coli* analysis were downloaded from `UniProt`. Accession numbers for both `UniProt` and NCBI as well as date accessed are provided.

A4 Correlation of Gene Expression Over Datasets

To assess uniform expression over *E. coli* strains with multiple data sets we looked at the mean normalized expression values. Multiple replicates from a data set were combined by finding the median normalized CPM expression value for each gene. This was done for any data sets that had multiple replicates. For each gene (x_i) the mean normalized expression value was calculated across all data sets ($\bar{x_i}_j$). Then the normalized median expression value for each data set was subtracted from the mean across all expression values ($|x_{ij} - \bar{x_i}_j|$). The distribution of these $|x_{ij} - \bar{x_i}_j|$ across all genes are found in Figures S3.1. All data sets are well mixed, implying that the expression levels are consistent across all data sets. Only the *E. coli* K-12 MG1655 strain had multiple expression datasets available so this is the only one that were analyzed. *E. coli* ATCC 25922, *E. coli* BW25113, and *E. coli* K-12 DH10B had only one data set each and therefore were not analyzed.



Supplemental Figure S3.1: Dot plot distribution of the median expression value for each *E. coli* K-12 MG1655 data set minus the mean expression value for that gene across all data sets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis. The values are coloured by GEO data set.

A5 `DIAMOND/BLAST` Test Parameters

Command
`diamond blastp -query-cover 90 -evalue 1e6 -outfmt 6`
`diamond blastp -query-cover 95 -evalue 1e6 -outfmt 6`
`diamond blastp -sensitive -query-cover 95 -evalue 1e6 -outfmt "6"`
`diamond blastp -more-sensitive -query-cover 95 -evalue 1e6 -outfmt "6"`
`blastp -qcov_hsp_perc 90 -evalue 0.001 -outfmt "6" -use_sw_tback`
`blastp -qcov_hsp_perc 95 -evalue 0.001 -outfmt "6" -use_sw_tback`

Supplemental Table S3.4: Commands used for testing appropriate `DIAMOND` and `BLAST` parameters. Only relevant parameters are shown. The command that yielded the best results and was used for the analysis is indicated in **bold** (`diamond blastp -more-sensitive`).

A6 Length of Inverted Alignment Blocks

A Wilcoxon signed-rank test was used to determine if there was a difference in alignment block length between significant inverted alignment blocks and non-significant inverted alignment blocks. A significant correlation was determined (Wilcoxon signed-rank test: W=4293794.5, p-value < 0.001), indicating that there is a significant difference in the length of significant inverted alignment blocks and non-significant inverted alignment blocks. Significant inverted alignment blocks (mean = 12079bp, median = 10297bp) are on average longer than non-significant inverted alignment blocks (mean = 11310bp, median = 9662bp).

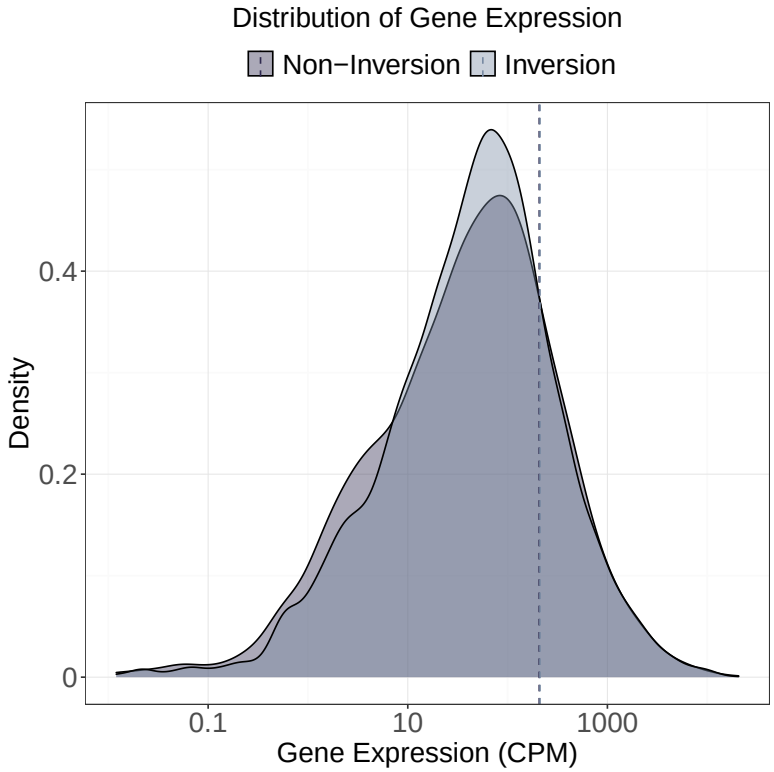
A7 Higashi et al. (2016) H-NS Binding Criteria

The Higashi et al. (2016) data set had multiple criteria to define H-NS binding sites (see Table 4.3). They are listed as follows: A: Genes whose coding regions overlap with the H-NS binding regions, B: Genes whose coding regions overlap with the H-NS binding regions and intergenic regions that were bound by H-NS, C: Genes whose coding regions overlap with the H-NS binding regions and intergenic regions that are "class I " (see Higashi et al. (2016)), D: Genes whose coding regions overlap with the H-NS binding regions and intergenic regions that contain known promoter sequences, E: Same as A, but genes on which H-NS binding is restricted to the 3' end and the length overlapping with H-NS-bound regions is <10% of the total gene length were excluded from H-NS-bound genes, F: When genes included in transcriptional units whose upstream regions or first coding regions overlapped with H-NS bound regions, all genes in the transcriptional units were judged as genes affected by H-NS binding.

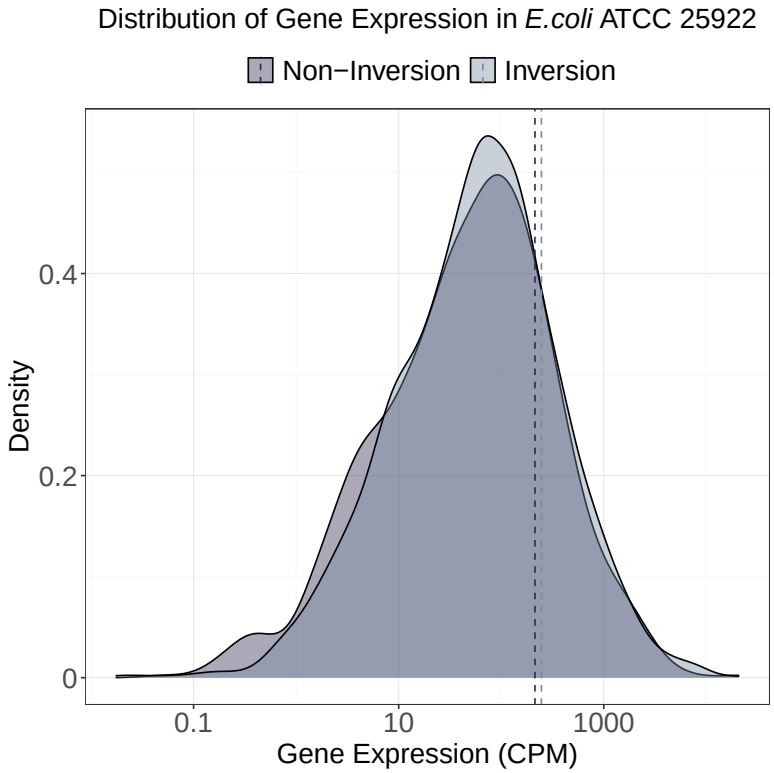
A8 Variation in Expression

	Test Statistic		Coefficient of Variation	
Group	Asymptotic	M-SLRT	Inversion	Non-Inversion
All Blocks	NS	NS	3.26	3.43
Only ATCC genes	NS	NS	3.24	3.78
Group	Asymptotic	M-SLRT	Significant Inversion	Non-Significant Inversion
Significant Inversions	8.738**	13.600***	4.39	3.08

Supplemental Table S3.5: Tests for equality of coefficient of variances in gene expression. The "Asymptotic" test refers to the Feltz and Miller (1996) asymptotic test. The "M-SLRT" test refers to the Modified Signed-Likelihood Ratio Test (M-SLRT) from Krishnamoorthy and Lee (2014). "All Blocks" indicates all identified alignment blocks. "Only ATCC genes" indicates all ATCC genes that are both inverted and non-inverted. "Significant Inversions" indicates all inverted blocks that had a significant difference in gene expression between the inverted and non-inverted sequences. The coefficient variance in this group was calculated for the inversions that were significant inversions and non-significant inversions. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.



Supplemental Figure S3.2: Distribution of gene expression values (CPM) for all genes in Inverted (light grey) and Non-inverted (dark purple) regions of the genomes of *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113 and *E. coli* ATCC 25922. The expression value in CPM is on the x-axis on a $\log_{10}$ and the density of expression values is on the y-axis. The mean expression values for genes in the Inverted (light grey) and Non-inverted (dark purple) regions are denoted by vertical dashed lines. The means for the Inverted and Non-inverted groups are very similar, and nearly overlapping.



Supplemental Figure S3.3: Distribution of gene expression values (CPM) for all genes in Inverted (light grey) and Non-inverted (dark purple) regions of the *E. coli* ATCC 25922 genome. The expression value in CPM is on the x-axis on a $\log_{10}$ and the density of expression values is on the y-axis. The mean expression values for genes in the Inverted (light grey) and Non-inverted (dark purple) regions are denoted by vertical dashed lines.