Spatial Patterns of Gene Expression in Bacterial Genomes

Daniella F Lato[1] and G Brian Golding[1]*

Paper Draft

November 4, 2019

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 25 bacterial genomes of *Escherichia coli*, *Bacillus subtilis*, *Streptomyces*, and *Sinorhizobium meliloti*. We confirmed that gene expression tends to decrease when moving away from the origin of replication in the majority of the replicons analysed in this study. The exception to this was pSymA of *S. meliloti* in which gene expression generally decreased when moving away from the origin of replication. This study sheds light on the impact of genomic location on molecular trends such as gene expression and highlights the importance of account for spatial trends in bacterial molecular analysis.

[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, gene expression, origin of replication, *Escherichia coli*, *Streptomyces*, *Bacillus subtilis*, *Sinorhizobium meliloti*

Introduction

Gene expression in bacteria is complex and highly controlled. The regulation of bacterial gene expression is crucial component of bacterial survival because these organisms can 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 therefore 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 100Kbp 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 700-800Kbp 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 histones and 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 dN/dS) 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 analysed 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 increasing gene expression values as genomic distance increases away from the origin of replication. However, 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 pre-determined locations around the replicon. However, 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.

In this work we look at the overall expression levels of all genes within 25 bacterial genomes from *Escherichia coli*, *Bacillus subtilis*, *Streptomyces*, and *Sinorhizobium 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. Using whole genome expression data obtained from the 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 linearly with increasing distance from the origin. Understanding how the location 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.

Materials and Methods

Expression Data

Gene expression data for each bacteria was downloaded from the Gene Expression Omnibus (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 environments absent of any stress. A complete list of expression data used is found in Supplementary Table S1. Correlation of gene expression across data sets was assessed for each bacteria, for a detailed protocol, see Supplementary files on `GitHub` at `https://github.com/dlato/Spatial_Patterns_of_Gene_Expression.git`.

Normalisation

The raw counts from control populations for each data set was used and normalised using the TMM method (Robinson and Oshlack 2010). Raw counts were normalised to Counts Per Million (CPM) in `R` using the `edgeR` package (Robinson et al. 2010). After normalisation, 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.

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. A single median CPM per 10Kbp section of each bacterial genome was determined. The median CPM was determined for each 10Kbp section of the genome to determine if there was a similarity between the number of substitutions in that same 10Kbp section and gene expression.

The gene expression information was summarised in bar graphs in R using ggplot2 (Wickham 2009) (Figures 2 and 3).

Origin and Bidirectionality of Replication

For each bacteria 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). 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 1).

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 midpoint between the origin of replication and the translated end of the genome was used as the terminus location. 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).

The origin scaling and bidirectional replication transformations were done in R (R Development Core Team 2014) and allows 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 1.

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). These bacteria therefore have approximately symmetrical chromosomal arms and as a result have genomic position labelling in Figures 2 and 3, accounting for bidirectional replication. *Streptomyces* on the other hand, is an acrocentric linear chromosome meaning that one chromosomal arm is much shorter than the other (see Figure 2). The genomic position labelling of *Streptomyces* in Figure 2 has negative numbers to indicate the shorter chromosome arm, and positive numbers indicating the longer chromosome arm.

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 10,000bp increments in each direction from the original origin (Supplementary Table: S2) to a maximum of 100,000bp in each direction. The remainder of the analysis was performed the same way on each of these shuffled origin positions and the linear regression was calculated (Supplementary Table: S4).

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). A normalised CPM expression value was calculated for each 10Kbp region of the genome. not sure if normalised is the correct word...or weighted maybe? This was calculated by taking the sum of all CPM expression values over a 10Kbp region of the genome, and dividing this by the total number of genes present in that 10Kbp segment. A linear regression was performed on these 10Kbp normalised 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-normalised expression values, and total additive expression values were also calculated. These results and methods can be found in the Supplementary Material (Supplementary Tables S3- S5).

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

Results and Discussion

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 where in the genome the farthest point from the origin is located. 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 1). Be that

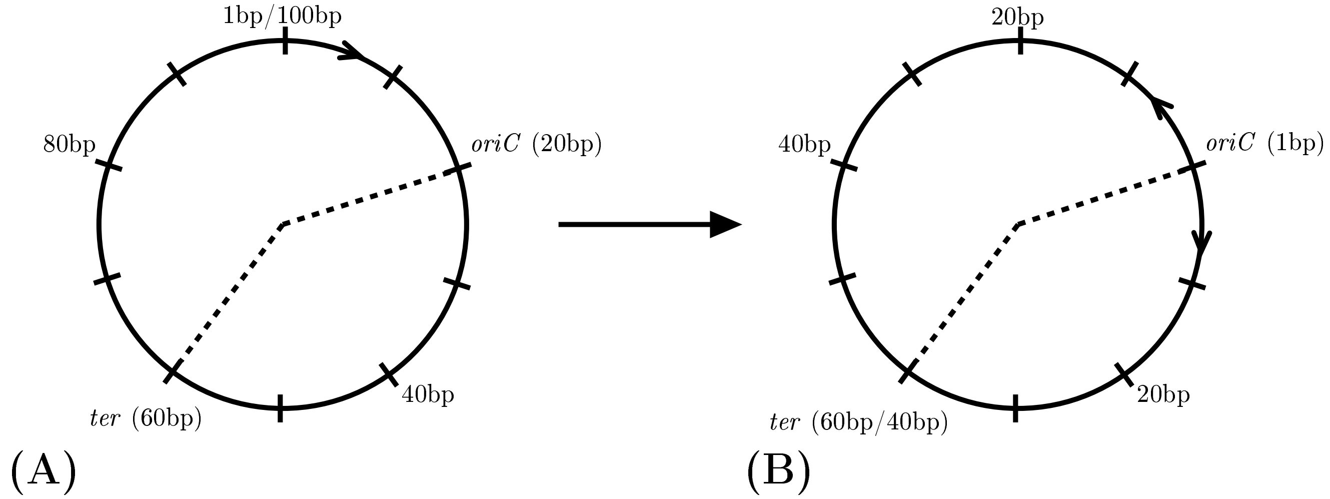


Figure 1: Schematic of the transformation used to scale the positions in the genome to the origin of replication and account for bidirectional replication. Circle (A) represents the original replicon genome without any transformation. Circle (B) represents the same replicon genome after the transformation. 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.

as it may, due to the nearly symmetrical location of the terminus to the origin, this effect in this work 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. These results from our origin of replication permutation tests (data not shown) determined that moving the origin of replication does not affect the trends seen at the end of the analysis, providing a robust check for origin of replication location.

Table 1: Arithmetic gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million. not sure if this is a necessary table or if people care about this? I dont really talk about it at all in the paper. just did it as extra thing

Bacteria and Replicon	Average Expression Value (CPM)
E. coli Chromosome	176.001
B. subtilis Chromosome	186.533
Streptomyces Chromosome	6.453
S. meliloti Chromosome	286.723
S. meliloti pSymA	764.793
S. meliloti pSymB	628.318

Linear Regression

The normalised CPM gene expression values over 10Kbp decreases when moving away from the origin of replication for the chromosomes of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. All secondary replicons of *S. meliloti* did not have significant linear regression values. This negative linear regression trend was expected based on previous work in similar bacterial species (Couturier and Rocha 2006; Morrow and Cooper 2012) looking at highly expressed (Couturier and Rocha 2006) or orthologous genes (Morrow and Cooper 2012) though out the genome, 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 2.

We additionally performed a linear regression on a per gene basis, we found similar results as the linear regression of normalised expression values over 10Kbp regions: *E. coli* and *B. subtilis* had gene expression decrease with increasing distance from the origin of replication (Supplementary Table: S3). We were unable to determine a significant trend between gene expression and genomic

Table 2: Linear regression results of normalised expression and distance from the origin of replication. The normalized 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 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'. A grey row indicates a significant negative trend.

Bacteria and Replicon	Change in Gene Expression with Distance from the Origin of Replication
E. coli Chromosome	-2.29×10^{-5}***
B. subtilis Chromosome	-2.48×10^{-5}**
Streptomyces Chromosome	-1.41×10^{-7}**
S. meliloti Chromosome	-2.54×10^{-5}*
S. meliloti pSymA	NS
S. meliloti pSymB	NS

position in majority of the other bacterial replicons (Supplementary Table: S3). 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 normalised expression values over 10Kbp regions in *E. coli*, where median gene expression decreases with increasing distance from the origin of replication (Supplementary Table: S4). For all of the other bacterial replicons we were unable to determine a significant trend between median gene expression and genomic position (Supplementary Table: S4). 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 normalised expression values over 10Kbp regions in most bacterial replicons where total gene expression decreases with increasing distance from the origin of replication (Supplementary Table: S5). For the two secondary replicons of *S. meliloti*, we were unable to determine a significant trend between total gene expression and genomic position (Supplementary Table: S5).

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 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 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 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 10Kbp decreases with distance from the origin of replication (Table 3). The higher concentration of genes is near the beginning of the genome, where we see increased expression, and the lower concentration of genes is near the terminus, where we observed decreased expression.

Table 3: Linear regression analysis of the total number of protein coding genes 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'.

Bacteria and Replicon	Change in Number of Genes with Distance from the Origin of Replication
E. coli Chromosome	-1.51×10^{-6}***
B. subtilis Chromosome	-3.00×10^{-6}***
Streptomyces Chromosome	NS
S. meliloti Chromosome	NS
S. meliloti pSymA	NS
S. meliloti pSymB	-3.08×10^{-6}*

We were unable to find a significant relationship between gene expression and distance from the origin of replication for the secondary replicons of *S. meliloti* (pSymA and pSymB). This bacteria is not as well studied as the other bacteria in this analysis (Martens et al. 2008). In our search, 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*.

Areas of the bacterial genomes with extremely high gene expression (Supplementary Table S6), are regions that encode proteins involved in things such as ribosomal proteins, DNA repair and replication, RNA synthesis, and metabolism. 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 1.49Mbp away from the origin of replication (in the case of the chromosome of *S. meliloti*, see Supplementary Table: S6 for more details).

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). We are the only study to assess gene expression levels across all protein coding genes within the bacterial genomes of *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*, and determine if 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 analysis 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.

My figures get pushed to the end of the document even though I am using [h]...any help would be appreciated!

Supplementary Material

Supplementary Figures S1- S2 and Tables S1–S6 are available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank CMAS, and JL for comments on the manuscript. We thank NSERC for funding for this project (grant XXXXXX to GBG).

References

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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

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

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.

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.

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.

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

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.

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.

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 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.

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 (1987). Gene location affects expression level in *Salmonella typhimurium*. J Bacteriol 169(6), 2872–2875.

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.

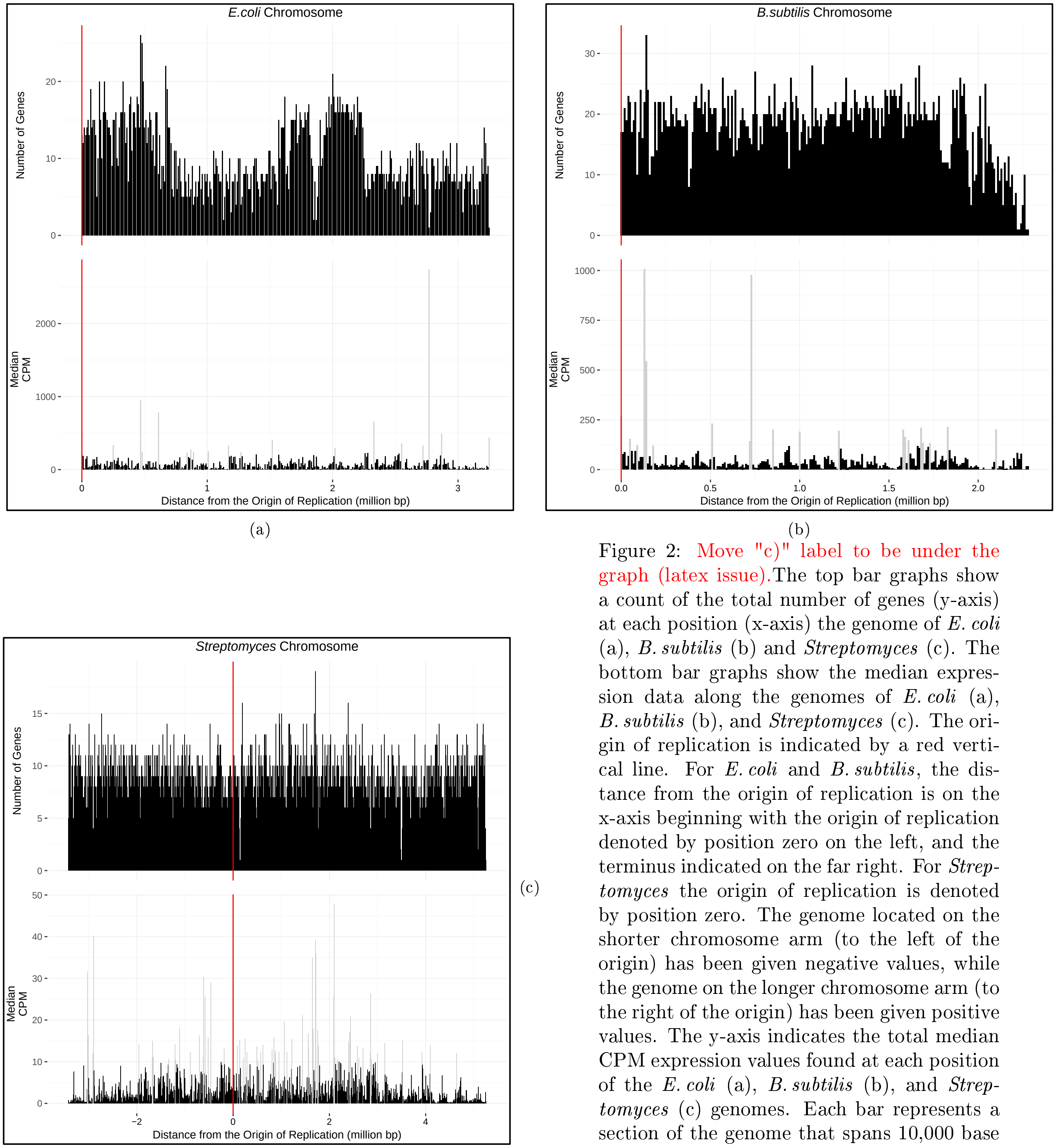
Sharp P M, Shields D C, Wolfe K H, and Li W.-H (1989). Chromosomal location and evolutionary rate variation in *Enterobacterial* genes. Science 246, 808–810.

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

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

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.

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

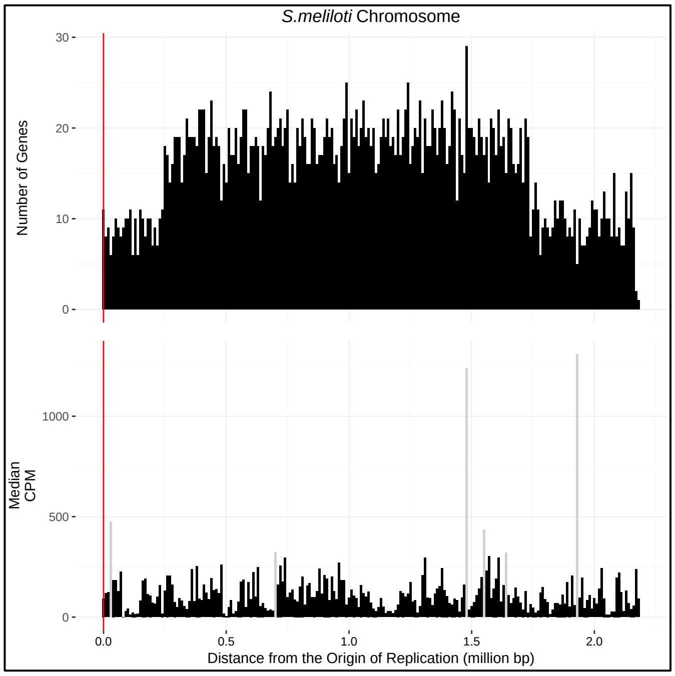


(a)

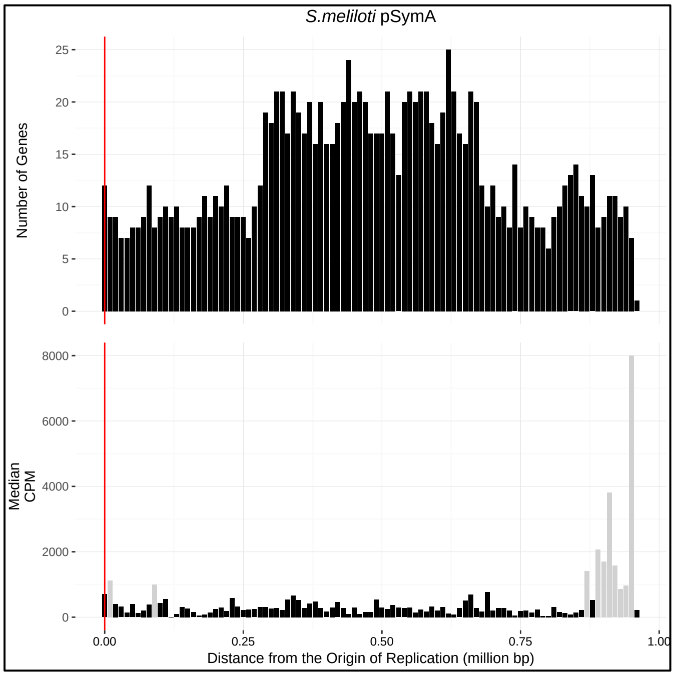
(b)

(c)

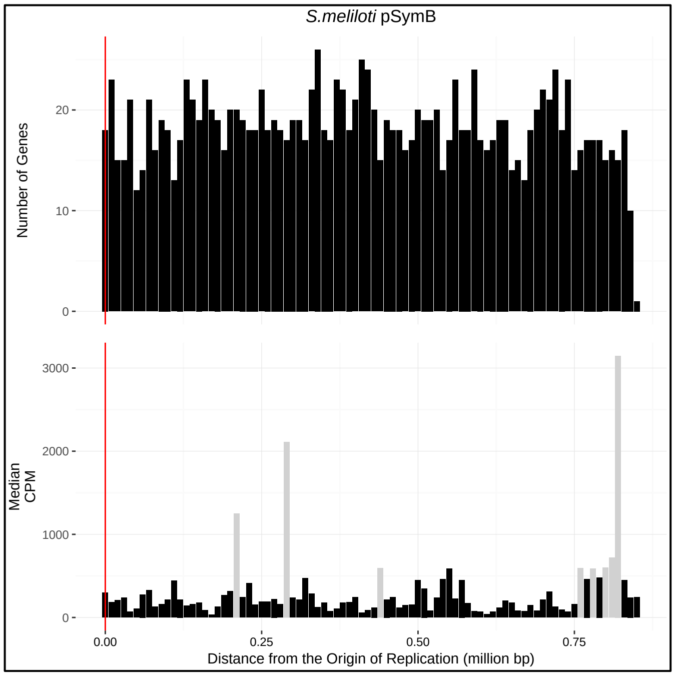
Figure 2: Move "c)" label to be under the graph (latex issue).The top bar graphs show a count of the total number of genes (y-axis) at each position (x-axis) the genome of *E. coli* (a), *B. subtilis* (b) and *Streptomyces* (c). The bottom bar graphs show the median expression data along the genomes of *E. coli* (a), *B. subtilis* (b), and *Streptomyces* (c). The origin of replication is indicated by a red 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 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.



(a)



(b)



(c)

Figure 3: Move "c)" label to be under the graph (latex issue).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 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 indicates the total median CPM expression values found at each position of *S. meliloti* chromosome (a), pSymA (b), and pSymB (c) replicons. Each bar represents a section of the genome that spans 10,000 base pairs.