Subs Paper Things to Do:

- why are the lin reg of dN, dS and ω NS but the subs graphs are...explain!
- mol clock for my analysis?
- GC content? COG? where do these fit?

Inversions and Gene Expression Letter Things to Do:

- ~~create latex template for paper~~
- confirm inversions with dot plot
- make dot plot of just gene presence and absence matrix (instead of each site) to see if this will go better
- look up inversions and small RNA's paper Marie was talking about at Committee meeting
- write outline for letter
- write Abstract
- ~~write intro~~
- write methods
- compile tables (supplementary)
- write results
- write discussion
- write conclusion
- do same ancestral/phylogenetic analysis that I did in the subs paper

General Things to Do:

- summarize references 40 and 56 from Committee meeting report (Brian was asking)

Last Week

Substitutions Paper:

✓re-did LOO analysis with proper trees

✓look into LOO analysis and the branches that caused a flip in sign

✓finished windowed analysis on dN, dS, and ω values

✓added reference for `Parsnp` (I did mention it briefly in the paper)

Inversions + Gene Expression:

✓Checking over Queenie's dataframes

✓thinking about what results to include in the paper and how to word them

✓added legend to expression and inversions pic

✓figure for H-NS binding and inversions

✓begun extracting H-NS data from PDF tables for Lang 2007 and Oshima 2006

✓finished writing methods for paper (minus DESeq analysis)

✓started writing up the results for paper

Inversions + Gene Expression: I added a legend to the inversions and gene expression figure (Figure). **What are your thoughts?**

I created a figure to show all inversions, H-NS binding, and location of inversions with significant differences in gene expression along the *E. coli* K-12 MG1655 genome (used as a reference since none of the inversions actually occur in this taxa) (Figure). **Please let me know what you think.**

I started extracting the H-NS binding data from the Oshima and Lang datasets using and R program. It is not perfect and requires some manual tweaking, but it is better than typing in 120 pages of tables by hand.

Subst Paper: I looked into the LOO analysis and particularly the results highlighted in red in Tables 2 and 3, where there is a complete switch in sign. I double checked the branch lengths of the trees that I was using and all were correct except *Streptomyces*. After re-doing this analysis with the proper tree, the results did not change. So this did not fix the issue found in pSymB. I also noticed that I forgot to list the result for one of the pSymA strains, so it is not included and unfortunately also swapped in sign.

For *Streptomyces* and pSymA, the taxa that are removed are the ones that are listed as the "outgroup" in the trees. This could potentially explain why there is such a dramatics change in sign. For *B. subtilis* and pSymB (and really for all the bacteria), after diving deeper into the substitutions, it seems as though when a sequence is removed, PAML shifts what branches the substitution is found on. This moves between the tip branch and the ancestor. This intern, changes the genomic location of the substitution (because different branches have different genomic positions associated with them). So there are some regions in the graphs that now have many more substitutions than before, and in the particular LOO cases of pSymB, *B. subtilis*, *Streptomyces* and pSymA, this changes the overall distribution enough that the sign flips. In the case of pSymB, the genome that causes the sign flip happens to be the shortest genome. I am not sure if this has anything to do with anything. **I honestly do not know what to say about this to the reviewer, or what else to do. Maybe we should schedule a meeting so I can show you some of these cases? Maybe I should look into doing my same analysis on previous datasets? Or perhaps re-do this LOO analysis with completely new progressiveMauve alignments and trees where the one strain is truly left out from the analysis? Maybe some sort of permutation test? I would really appreciate any help on what to do.**

This Week

- double check Queenie's final dataframes
- double check new inversion combos with Queenie's new data frames
- finish writing results for inversions paper
- clarify tests used in ↑
- continue working LOO analysis for subst paper
- address reviewers comment on HGT and ori/ter gradient
- address reviewers comment on annotation in outlier bars
- create new cover letter for subst paper

Next Week

- actual analysis on DESeq data
- visualizations/results for ↑
- read papers on H-NS proteins
- double check for HNS and inversions fig that ALL HNS binding sites are shown (not just ones in inversions)
- continue get Lang and Oshima data from PDF to csv formats
- do H-NS analysis on ↑
- final decision on inversion viz (caption that explains it well)
- maybe do inversions in 10kb blocks? (and other sliding windows?)
- dist from ori on DESeq results?
- HGT and HNS binding?

Bacteria and Replicon	Protein Coding Sequences Coefficient Estimate
E. coli Chromosome	$-3.29\times10^{-8***}$
B. subtilis Chromosome	$8.70\times10^{-9*}$
Streptomyces Chromosome	NS
S. meliloti Chromosome	$-6.80\times10^{-7***}$
S. meliloti pSymA	$4.49\times10^{-7***}$
S. meliloti pSymB	$6.27\times10^{-8*}$

Table 1: Logistic regression analysis of the number of substitutions along all protein coding positions of the genome of the respective bacteria replicons. ONLY EXTANT BRANCHES. 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 codes as followed: $< 0.001 =$ '***', $0.001 < 0.01 =$ '**', $0.01 < 0.05 =$ '*', $> 0.05 =$ 'NS'.

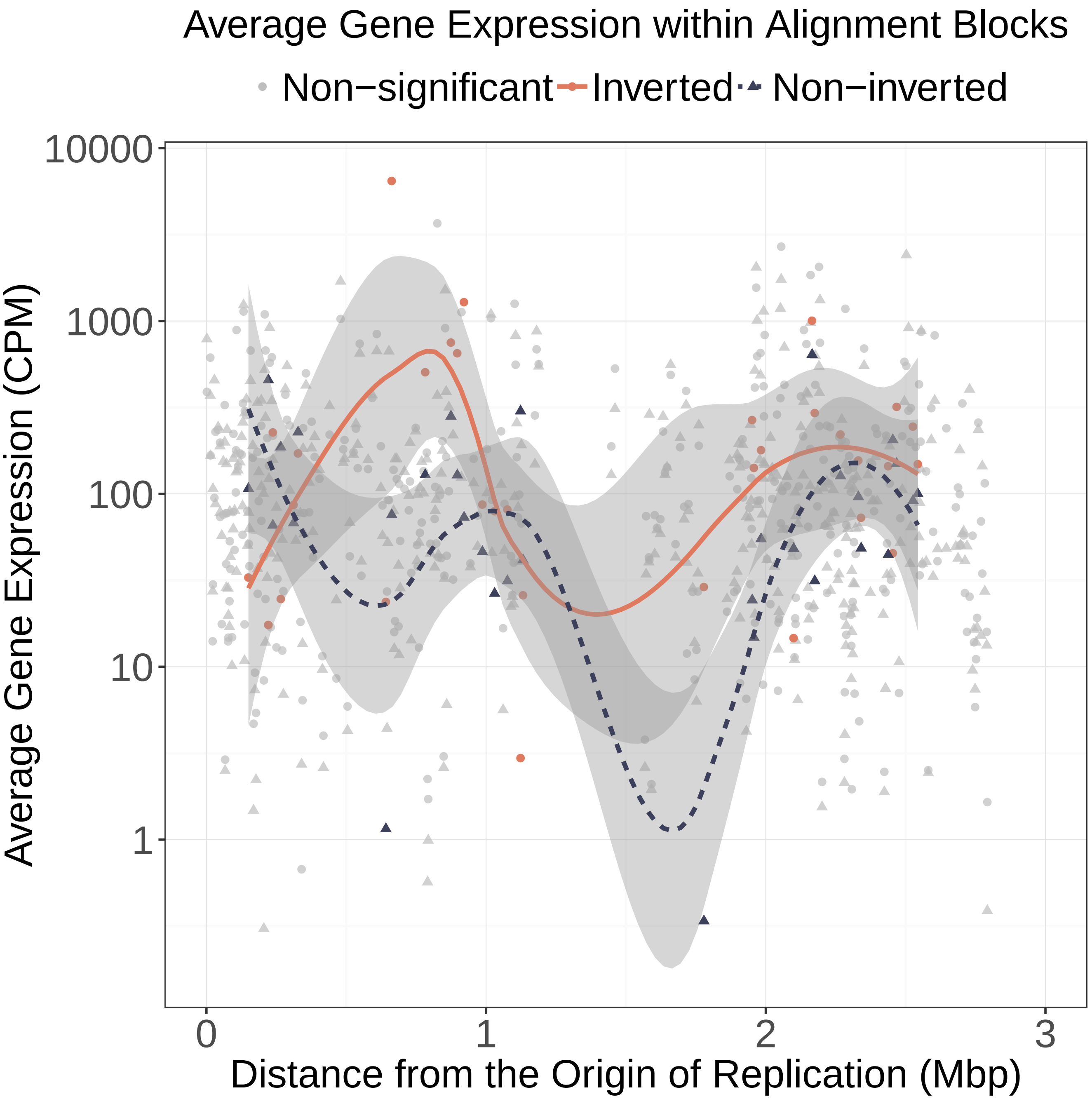

Figure 1: Visualization of the difference in gene expression between inverted and non-inverted sequences within alignment blocks. Each alignment block represents homologous sequences between the *Escherichia coli* strains insert table ref here. *E. coli* K-12 MG1655 was used as the reference genome for genomic position for each alignment block. The midpoint of each alignment block was calculated to be the genomic distance from the *E. coli* K-12 MG1655 origin of replication. 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 pink circles and purple triangles respectively. A smoothing line (`loewss`) was added to link the average gene expression values for the inverted (pink 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). All blocks that did not have a significant difference in average gene expression between inverted and non-inverted sequences within alignment blocks have the average inversion (circles) and non-inversion (triangles) gene expression values coloured in light grey.

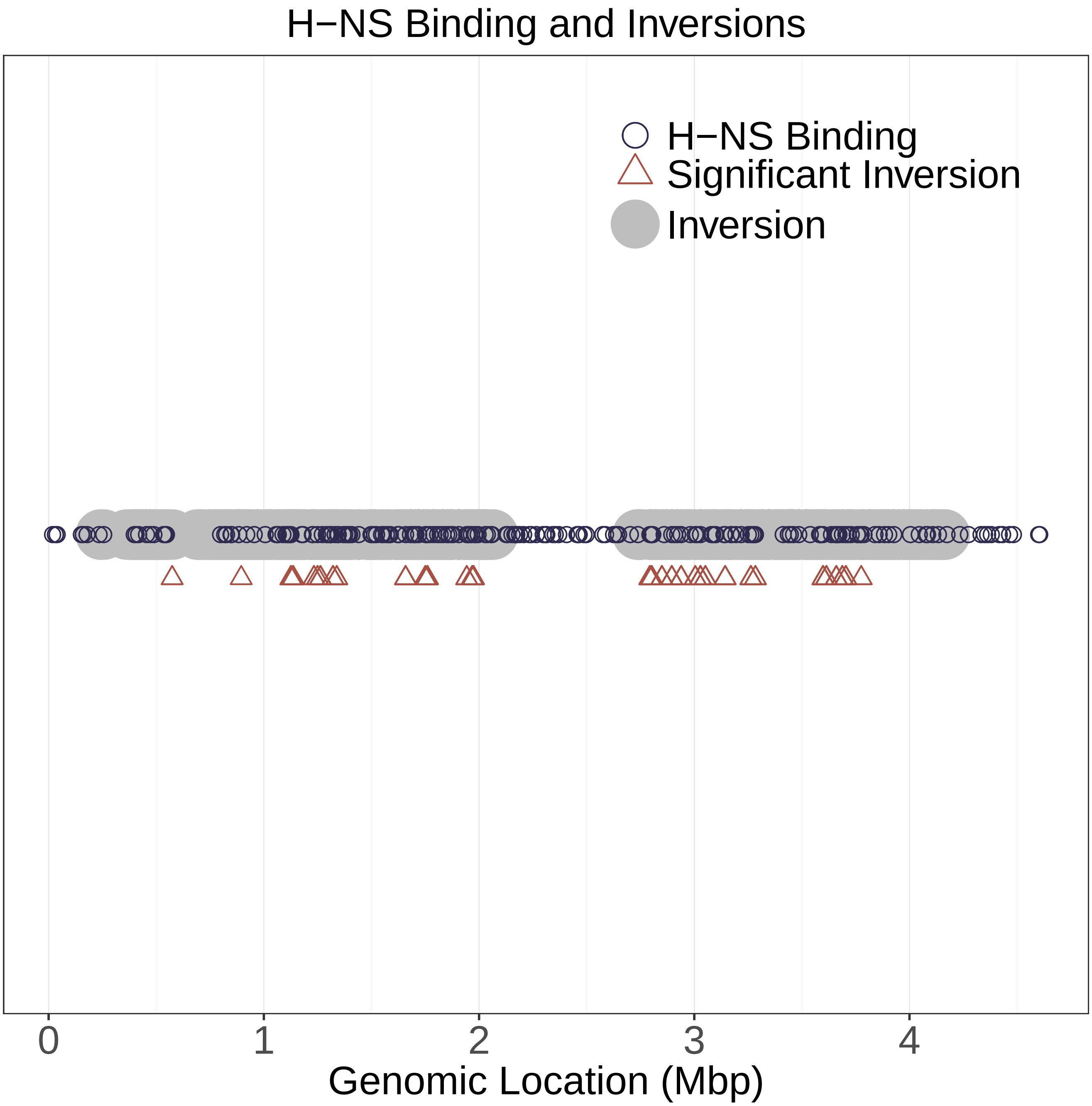

Figure 2: Visualization of the genomic locations of all inversion alignment blocks (light grey filled circles) identified between *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113, and *E. coli* ATCC. The data points are plotted on the genome of *E. coli* K-12 MG1655 which is used as a reference. Each inversion alignment block has a single genomic location chosen to be the midpoint of the inverted region calculated to be the genomic distance from the *E. coli* K-12 MG1655 origin of replication. **H**istone-like **N**ucleoid-**S**tructuring (H-NS) protein binding sites in the *E. coli* K-12 MG1655 are overlaid on top of the inversion alignment blocks (circles outlined in dark purple). Data for the H-NS binding information is from Higashi insert citation here. Inversion alignment blocks that 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 marked below the inverted alignment blocks with dark pink outlined triangles.

Strain 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}***

Table 2: 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'.

Strain Removed	Coefficient Estimate
S. meliloti Chromosome	
None	$-6.57\times10^{-7***}$
NC_015590	$-3.18\times10^{-7***}$
NC_003047	$-6.01\times10^{-7***}$
CP004140	$-6.00\times10^{-7***}$
CP009144	$-6.67\times10^{-7***}$
NC_017322	$-7.19\times10^{-7***}$
NC_017325	$-5.01\times10^{-7***}$
S. meliloti pSymA	
None	$2.74\times10^{-7***}$
NC_017327	$6.98\times10^{-7***}$
CP009145	$1.78\times10^{-7***}$
NC_003037	$2.09\times10^{-7***}$
CP004138	$2.08\times10^{-7***}$
NC_015591	NS
NC_017324	$-1.52\times10^{-6***}$
S. meliloti pSymB	
None	$1.10\times10^{-7***}$
NC_015596	$6.78\times10^{-7***}$
NC_017326	$1.67\times10^{-7***}$
NC_017323	NS
CP009146	$-2.57\times10^{-7***}$
CP004139	$1.04\times10^{-7***}$
NC_003078	$1.04\times10^{-7***}$

Table 3: 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'.

H-NS Binding Study	All Inversions H-NS Binding	Significant Inversions and H-NS Binding	Total Number of H-NS Binding Sites Within All Alignment Blocks
Grainger 2006	NS	NS	37
Ueda 2013	NS	NS	165
Higashi 2016: coding criteria 1	0.103*	0.113***	206
Higashi 2016: coding criteria 1 and non-coding criteria 1	0.102*	0.104***	189
Higashi 2016: coding criteria 1 and non-coding criteria 2	0.102*	0.104***	189
Higashi 2016: coding criteria 1 and non-coding criteria 3	0.102*	0.104***	189
Higashi 2016: coding criteria 2	0.105*	0.104***	187
Higashi 2016: coding criteria 3	0.105*	0.104***	187
Lang 2007: composite data	0.101*	-0.056*	80
Oshima 2006	NS	NS	277

Table 4: are there any other stats related to correlation that people like to have in these tables that I should also be including? 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 taxa: *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113, and *E. coli* 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. All results are marked with significance codes as followed: < 0.001 = '***', 0.001 < 0.01 = '**', 0.01 < 0.05 = '*', > 0.05 = 'NS'.

Datasets:	Correlation Coefficient (W)
Inverted Blocks	15218699**
Inverted Sequences	11436344***

Table 5: 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: < 0.001 = '***', 0.001 < 0.01 = '**', 0.01 < 0.05 = '*', > 0.05 = 'NS'.

% of Blocks that are		
Inverted	Inverted with Differences in Gene Expression	Increased in Gene Expression in Inverted Sequences
68.29	8.22	58.06

Table 6: Percent of blocks in categories for various datasets (blocks with all 4 taxa, at least 3 taxa, or at least 2 taxa). The second column is any block that had at least one sequences that was inverted. The last column only deals with blocks that had at least one inverted sequence and had a significant difference in gene expression (column 3).

Block Length Correlation Coefficient (W)
4060729.5***

Table 7: Correlation coefficients for Wilcoxon signed-rank test in alignment blocks. The correlation coefficient represents a correlation between alignment block length and blocks with a significant/non-significant difference in normalized gene expression between inverted and non-inverted sequences within the block. All results are marked with significance codes as followed: < 0.001 = '***', 0.001 < 0.01 = '**', 0.01 < 0.05 = '*', > 0.05 = 'NS'.

Genomic Position Correlation Coefficient (W)
NS

Table 8: Correlation coefficients for Wilcoxon signed-rank test in alignment blocks with a significant difference in normalized gene expression between inverted and non-inverted sequences within the block. The correlation coefficient between the significant blocks and the genomic position of the alignment blocks. All results are marked with significance codes as followed: < 0.001 = '***', 0.001 < 0.01 = '**', 0.01 < 0.05 = '*', > 0.05 = 'NS'.

Inversion Category	Correlation Coefficient
rev comp	NS
inversion	$2.20\times10^{-7***}$
sig rev comp	$-1.89\times10^{-7*}$
sig $\sim$ midpoint all blocks	NS
sig $\sim$ midpoint inverted blocks	NS

Table 9: Logistic regression between various inversion categories and distance from the origin of replication for all strains. rev comp = individual sequences inverted, inversion = block that has at least one inverted sequence, midpoint = block midpoint, sig = blocks with significant difference in normalized gene expression between inverted and non-inverted sequences within the block. All results are marked with significance codes as followed: $<$ 0.001 = ‘***’, 0.001 $<$ 0.01 = ‘**’, 0.01 $<$ 0.05 = ‘*’, $>$ 0.05 = ‘NS’.

Strain	rev comp	inversion
E. coli K-12 MG1655	NS	$3.55\times10^{-7***}$
E. coli K-12 DH10B		$3.45\times10^{-7***}$
E. coli BW25113		$3.73\times10^{-7***}$
E. coli ATCC	$-1.92\times10^{-7***}$	$-1.92\times10^{-7***}$

Table 10: Logistic regression between various inversion categories and distance from the origin of replication for each strain. rev comp = individual sequences inverted, inversion = block that has at least one inverted sequence, sig = blocks with significant difference in normalized gene expression between inverted and non-inverted sequences within the block. All results are marked with significance codes as followed: $<$ 0.001 = ‘***’, 0.001 $<$ 0.01 = ‘**’, 0.01 $<$ 0.05 = ‘*’, $>$ 0.05 = ‘NS’.