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~~
- 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:

✓address reviewers comment on HGT and ori/ter gradient

✓address reviewers comment on annotation in outlier bars (mobile elements)

✓create new cover letter for subst paper

✓do subst analysis with only extant taxa branches (PAML and genome position)

✓other analysis on subst paper

Inversions + Gene Expression:

✓double check Queenie's final dataframes

✓double check new inversion combos with Queenie's new data frames

✓finish writing results for inversions paper (minus DESeq)

✓clarify tests used in ↑

✓attempt DESeq on sample_name ∼ treatment

✓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): Figure **??**

Dissertation:

✓Finished dissertation discussion on gene expression paper

✓Finished 2/3 proposed studies for dissertation

✓Finished prefaces for each chapter

✓Finished overall conclusion

Inversions + Gene Expression: I attempted to use the individual sample names as a component in the DESeq analysis (sample_name ∼ treatment). But it did not work. DESeq still complains that the matrix is not full rank and the column factors (sample_name and treatment) are linear combinations of each other. **I do not know what to do. I need to account for variation in expression between datasets (since they were all done in different labs at different times), while still being able to see what differences are due to the "treatment" = inversions. I would really appreciate your help and insight into this.**

I finished getting the Lang and Oshima data from PDF into a csv and complete the H-NS correlation analysis on these (Table 4). Oshima is NS so it is sort of irrelevant. The Lang data however has a positive correlation with all inversions but a negative correlation with significant inversions (inversions that had a significant difference in gene expression between inverted and non-inverted sequences within the inversion). **This is curious, and maybe related to the small number of data points? Do you have any thoughts on this?**

Subst Paper:

I finished addressing all the other reviewers comments in the paper and in the cover letter (which I will send to you soon). One of the comments was to check what types of genes were present in the outlier bars from the substitution figures, and perhaps if these are mobile elements, this may explain why these bars have high substitutions and are considered outliers. I looked into this and almost all mobile elements or phage-like genes were removed either via annotation (no `mobile_elements`) or due to poor alignment. The reviewer suggested that I could look at the ω values of these mobile genes to see if they are what is driving the high substs in these bars. **I was thinking, should I be looking at all ω values in these outlier bars? I was really only searching for mobile elements as the reviewer suggested. What do you think?**

As for the reviewer who is concerned about robust data, I was thinking about doing my analysis without accounting for rearrangements to show what those trends are (and hopefully they show a negative correlation like previous studies) and then show that when I account for rearrangements, then things change. However, using progressiveMauve inherently accounts for rearrangements so I am not sure if I should be re-aligning everything (and new trees...etc) with a different aligner that does not account for rearrangements? But then I was thinking that you have to account for rearrangements somehow otherwise you would aligning non-homologous sequences. So I looked into what some of the previous papers did (Cooper et. al 2010, Morrow et. al 2012...etc) and they all used basically the same protocol. They identified orthologs via blastp (and often required these orthologs to have similar positions in the genome), aligned them with ClustalW, then used codeml to calculate rates. **These studies also just chucked anything that had $dS > 1$ since these are unreliable. Should I be doing the same? This also got me thinking that I use different datasets for the subs analysis and the selection analysis. Should I be using the selection analysis to identify outlier genes and exclude those from my subst analysis??**

Anyways, this all got me thinking that perhaps my progressiveMauve alignments are ok and I can just choose the extant branches from PAML to use in my logistic regression of subs v.s. distance from the origin of replication. So I did this, where all the data and methods are the same except I only selected the tip branches and their corresponding substitutions and genomic positions. The results are found in Table 1, you can see that they are essentially the same as "with rearrangements". However, based on my LOO we know that PAML can change what branch the substitutions are on based on the tree and what taxa are present. **So maybe for this "no rearrangements" analysis I need to only count what substitutions we can see in the alignment, so use a program other than PAML? I am just concerned that PAML is moving this extant subs to deeper branches and then its not captured in this "no rearrangements" dataset. Thoughts?**

Overall, I do not know what else to do about the robust-ness of the data to satisfy this reviewer. Unless you can think of anything else, I will just craft the cover letter addressing the reviewers concerns and saying that we have done all we can do.

This Week

- finish subst analysis with "no rearrangements"

- write ↑ into polished cover letter and main paper
- submit subst paper
- actual analysis on DESeq data
- visualizations/results for ↑
- read papers on H-NS proteins
- write inversions paper abstract

Next Week

- read papers on H-NS proteins
- write inversions paper discussion
- write conclusion for dissertation
- 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	-3.29×10^{-8}***
B. subtilis Chromosome	8.70×10^{-9}*
Streptomyces Chromosome	NS
S. meliloti Chromosome	-6.80×10^{-7}***
S. meliloti pSymA	4.49×10^{-7}***
S. meliloti pSymB	6.27×10^{-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'.

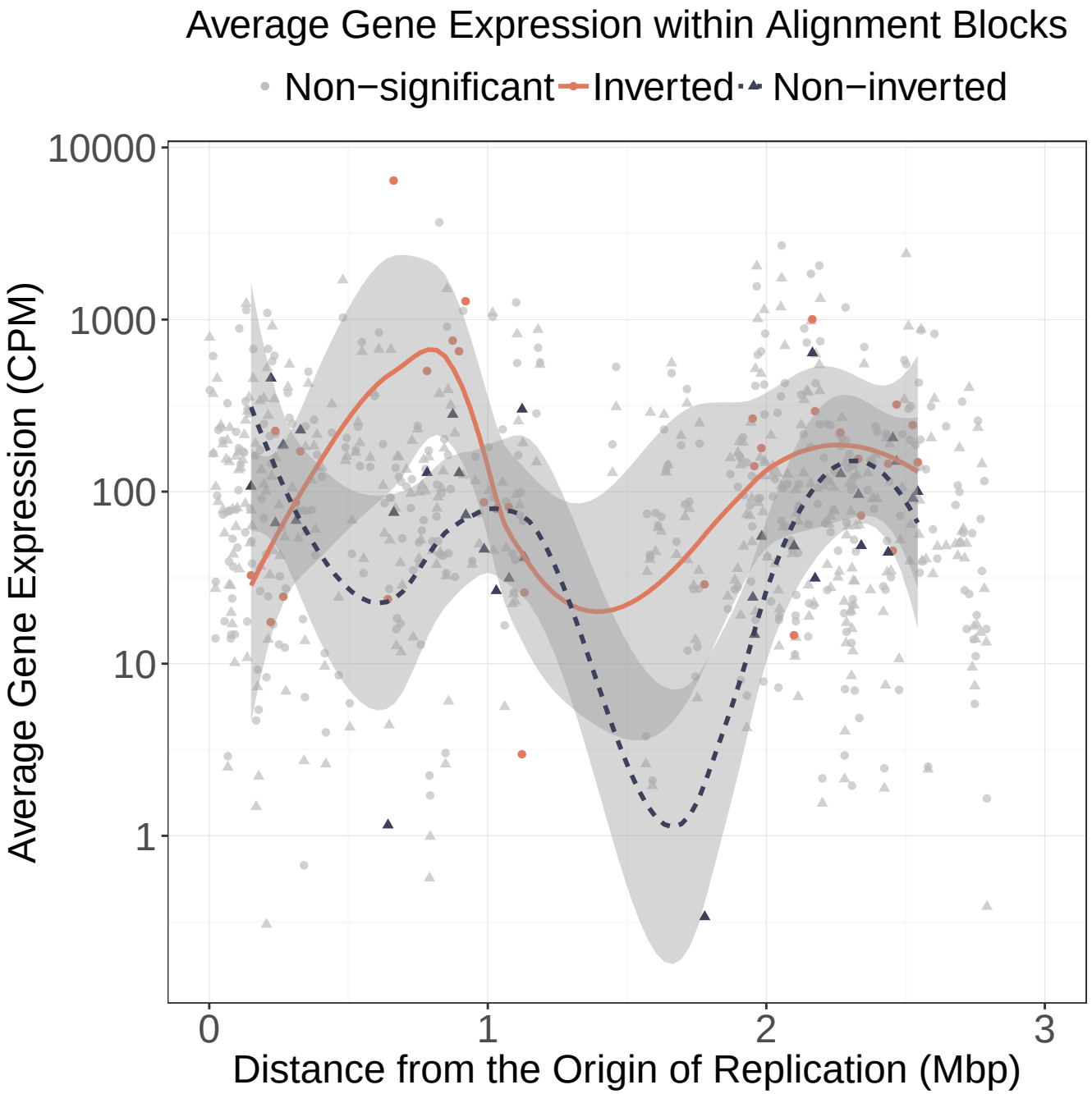

Figure 1: Visualization of the difference in gene expression between inverted and non-inverted sequences within alignment blocks. Each alignment block represents homologous sequences between the *Escherichia coli* strains insert table ref here. *E. coli* K-12 MG1655 was used as the reference genome for 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.

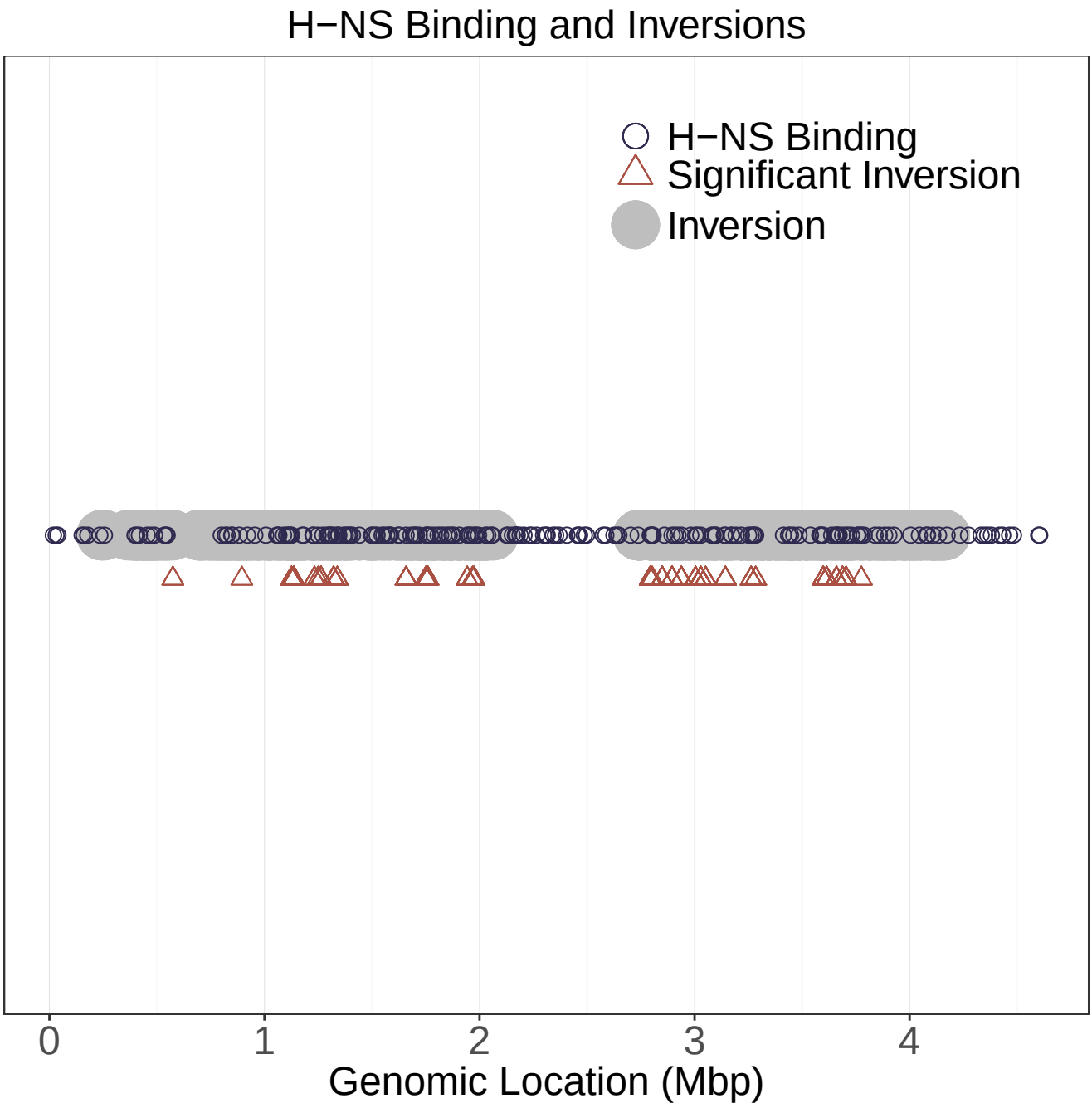

Figure 2: Visualization of the genomic locations of all inversion alignment blocks (light grey filled circles) identified between *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113, and *E. coli* ATCC. The data points are plotted on the genome of *E. coli* K-12 MG1655 which is used as a reference. 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	-6.57×10^{-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 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	53
Ueda 2013 [?]	NS	NS	275
Higashi 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 2007 [?]	0.0574**	NS	115
Oshima 2006 [?]	0.0390*	NS	664

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. The ref Higashi data set had multiple criteria to define H-NS binding sites. 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 cite Higashi), 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. 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	2.20×10^{-7}***
sig rev comp	-1.89×10^{-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		3.55×10^{-7}***
E. coli K-12 DH10B	NS	3.45×10^{-7}***
E. coli BW25113		3.73×10^{-7}***
E. coli ATCC	-1.92×10^{-7}***	-1.92×10^{-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'.