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:

✓completed the "leave one out" analysis

Inversions + Gene Expression:

✓Checking over Queenie's dataframes

✓wrote code to get all possible BLAST names

✓working on DESeq analysis

✓summary of all results for analysis (HNS, gene exp, distance from the ori..etc) except DESeq

✓looking into ATCC rev comp for inversion viz

Inversions + Gene Expression:

Final dataframes: Queenie has finished the final dataframes for this analysis for both the raw data (for DESeq) and normalized expression values (for other analysis). I have been checking these to ensure that they are correct and have found some minor errors in her code. She is fixing these and should be done by end of day Monday. Additionally, she needed more information from the BLAST outputs on all gene names so I have written a short script to extract this information for her. I suspect that the overall results will not change that much because the errors only impact a few genes and if they are included/excluded from the analysis.

DESeq Analysis: As I mentioned to you briefly last week, I was having trouble performing DESeq because my "matrix is not full rank". This means that the combination of levels in each of my experimental design columns (treatment (inversion/non-inversion), strain, experiment) are co-linear. This is because the strain and inversion/non-inversion combinations are similar. For example, DESeq can not tell the difference between the ATCC strain and the inverted treatment, because it impacts the same samples (all the ATCC samples). Therefore, DESeq can not say if the differential expression is due to the strain or the inversion. I took your advice and tried to make my input data as simple as possible, but I am still getting this error. The only way that I will not get this error is if I look at each column in my experimental design matrix separately. i.e. $\sim$ treatment, $\sim$ strain, $\sim$ experiment. However, based on my preliminary exploration of the data, it seems as though the experiment is driving most of the differential expression (which makes sense because the raw data is coming from a number of experiments that although we tried to use control data, they still were done in different labs, at different time...etc). **I am concerned that by doing $\sim$ treatment, and not $\sim$ treatment + experiment (which is what you usually do to combat say batch or lane effects), I am not accounting for the variation in expression between experiments. Do you have any thoughts on how I can ensure that the expression differences I am seeing with $\sim$ treatment are due to treatment along and not confounded by experiment?**

Summary of Inversion and Exp Results: I have included all the results for this analysis (except the DESeq analysis) in Tables 1 - 7.

I performed the H-NS analysis (looking at Pearson correlation between H-NS and Inversion/significant inversion) on each of the H-NS datasets (Table 1). It seems as though within the Higashi 2016 dataset, it does not seem to matter which criteria for H-NS binding I use, they all give me roughly the same answer. This dataset also have multiple criteria for the non-coding H-NS binding so I looked at each non-coding criteria and the coding criteria 1, and again found that there appears to be no difference in the results depending on what criteria you choose. However, when we

look across datasets, Grainger 2006 and Ueda 2013 have no significant correlation between H-NS binding and the inverted/significant inversions. In my data (meaning H-NS binding sites within my data), there are only 10 genes where all the H-NS datasets have the same binding sites. This could be why there is a difference in significance? For this reason, I think that the results from all the datasets need to be included in this paper/supplement. **I am not sure if it makes sense to say that H-NS has a positive correlation between inverted sites compared to non-inverted sites if some of the datasets do not show this. Thoughts?**

In Table 2, there is a clear correlation between inverted blocks or individual inverted sequences and a significant difference in gene expression. When there is a significant difference in gene expression between inverted and non-inverted sequences within a block, inverted sequences have higher expression than non-inverted sequences about 58% of the time (Table 3). Table 4 shows a positive correlation between block length and blocks with a significant difference in expression between inverted and non-inverted sequences. However, I am just looking into this now and I am confused by this. I will get back to you on that.

When looking at distance from the origin of replication, we see that using all genomic positions across all strains (where inversions can be present in different genomic positions), there is no correlation between distance from the origin of replication and blocks with significant differences in gene expression between inverted and non-inverted sequences (Table 5). Table 6 shows multiple logistic regressions on the various values for inversions and distance from the origin of replication. This table again is considering points from all strains. The placement of inversion (all inversions, including ones with significant and non-significant differences in gene expression) appears to be concentrated near the terminus. When looking at only inversions with significant difference in gene expression, we see that these are located closer to the origin of replication.

Table 7 shows the logistic regression between the inversion category (for all inversions, including ones that did not have a sig difference in gene expression) and distance from the origin of replication in each strain. The "rev comp" column is when that particular strain is inverted. The "inversion" column is when at least one strain in that block is inverted. Only *E. coli* K-12 DH10B and ATCC have inverted sequences, so really the other strains values are irrelevant. I also think the "inversion" column is irrelevant because it is the same data for each strain just with different genomic positions. We can see that the ATCC strain has most of its inversions concentrated near the origin of replication and less near the terminus. In the K-12 DH10B there was no significant correlation with distance from the origin of replication and placement of inversions.

Aside from the DESeq analysis, this is all the analysis that I have planned. Plus some sort of figure potentially looking at H-NS binding and inversions. Do you have anything else that you think I should look at or look more into?

Inversion Visualization: I have figured out the differences between the images I presented you with last time (where ATCC was reverse complemented or not). Each time you run PARSNP, even if it is on the same sequences but one is reverse complemented, it (obviously) runs the entire algorithm again. However, the output creates subtle difference in the sequences within the blocks. So some blocks that did not have a rearrangement when ATCC was not reverse complemented, have a rearrangement when ATCC is reverse complemented. This is why there appears to be more rearrangements between BW25113 and K-12 DH10B in Figure 2 (ATCC reverse complemented) compared to Figure 1 (ATCC not reverse complemented).

I cheated a bit and did a "manual" reverse complement of the ATCC block locations from the original data (where ATCC is not reverse complemented) by simply reversing the genomic positions (not altering the sequence data or anything else at all). This resulted in a much "cleaner" picture of the inversions (Figure 3). This only reverse complemented the inversions visually in ATCC and did not create new rearrangements between BW25113 and K-12 DH10B.

The reason I wanted to "clean up" the inversion visualization is to combat reviewers comments. I suspect a reviewer will say "it looks like the entire ATCC genome is inverted (in Figure 1), so why didn't you fix this? Are your inversions accurate?" I am unsure if it is "correct" to use the manual reverse complement cleaned version (Figure 3) and say that ATCC was reverse complemented to simplify the diagram? But then I think maybe the inversions should also be manually reversed (so anything that was an inversion when it was not reverse complemented would now not be an inversion, and vice versa). But again, I am not sure if this is "correct". What do you think about all this?

Subst Paper: The results from the "leave one out" analysis can be found in Tables 8 and 9. It looks like the results are mostly consistent, so when you leave a taxa out, the sign remains unchanged. There are a few exceptions which I have highlighted in red. I will be looking into these further. I am wondering if this could be because I used the maximum genome position when scaling my points to the origin of replication as the longest genome out of all the taxa, instead of the taxa that are used in each "leave one out". I will explore this to see if I missed something or messed up along the way. **Thoughts on these results so far?**

This Week

- double check Queenie's final dataframes
- double check new inversion combos with Queenie's new data frames
- check into block lengths and inversions results
- choose log reg values for final table
- actual analysis on DESeq data
- visualizations/results for ↑
- continue working on other edits for subst paper

Next Week

- read papers on H-NS proteins
- check how many HNS binding sites there are for each dataset
- think about how to visualize H-NS and inversions info

- get Lang and Oshima data from PDF to csv formats
- do H-NS analysis on ↑
- final decision on inversion viz
- continue working on other edits for subst paper

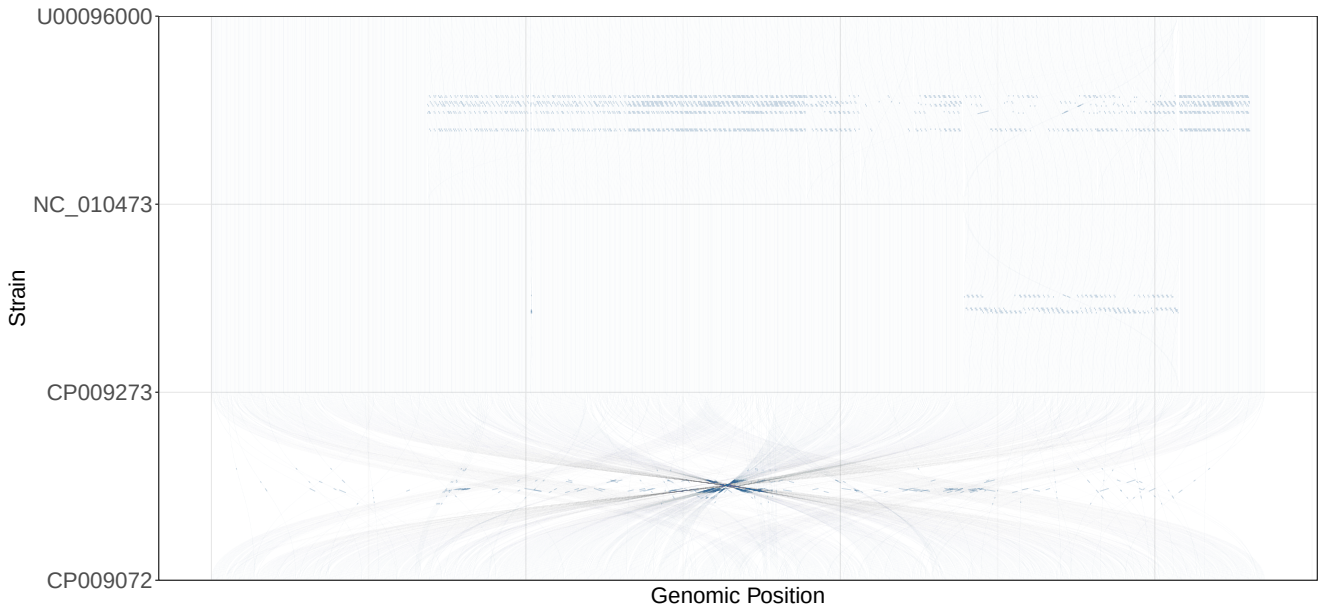

Figure 1: Visualization of rearrangements and inversions in all *E. coli* strains. ATCC is in the GenBank listed orientation.

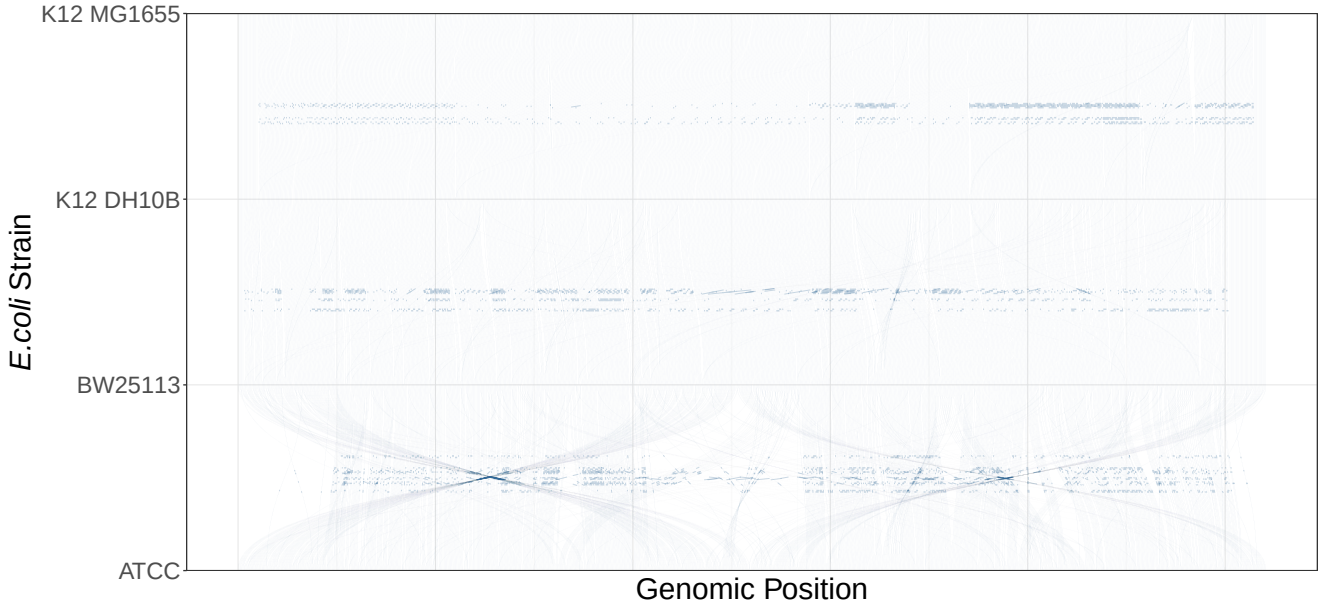

Figure 2: Visualization of rearrangements and inversions in all *E. coli* strains. ATCC is reverse complemented from the GenBank listed orientation.

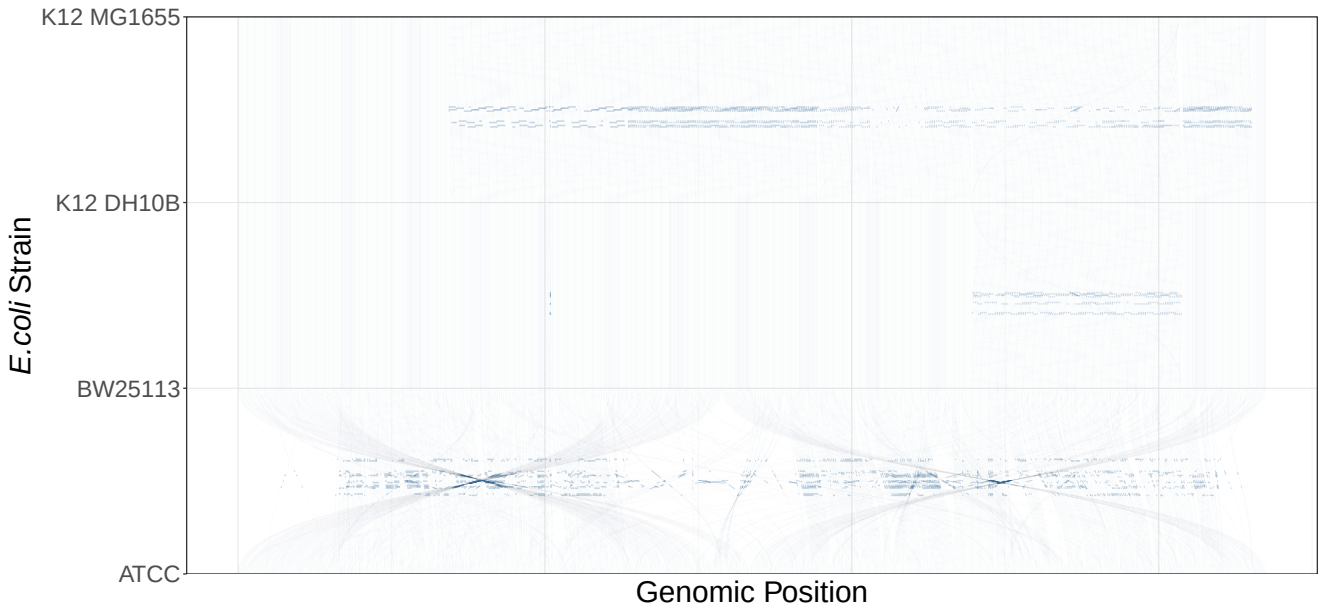

Figure 3: Visualization of rearrangements and inversions in all *E. coli* strains. ATCC is "manually" reverse complemented (only genomic position).

H-NS Binding Study	All Inversions and H-NS Binding	Significant Inversions and H-NS Binding
Grainger 2006	NS	NS
Ueda 2013	NS	NS
Higashi 2016: coding criteria 1	0.102*	0.101***
Higashi 2016: coding criteria 1 and non-coding criteria 1	0.101*	0.089***
Higashi 2016: coding criteria 1 and non-coding criteria 2	0.101*	0.089***
Higashi 2016: coding criteria 1 and non-coding criteria 3	0.101*	0.089***
Higashi 2016: coding criteria 2	0.104*	0.090***
Higashi 2016: coding criteria 3	0.104*	0.090***

Table 1: 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'.

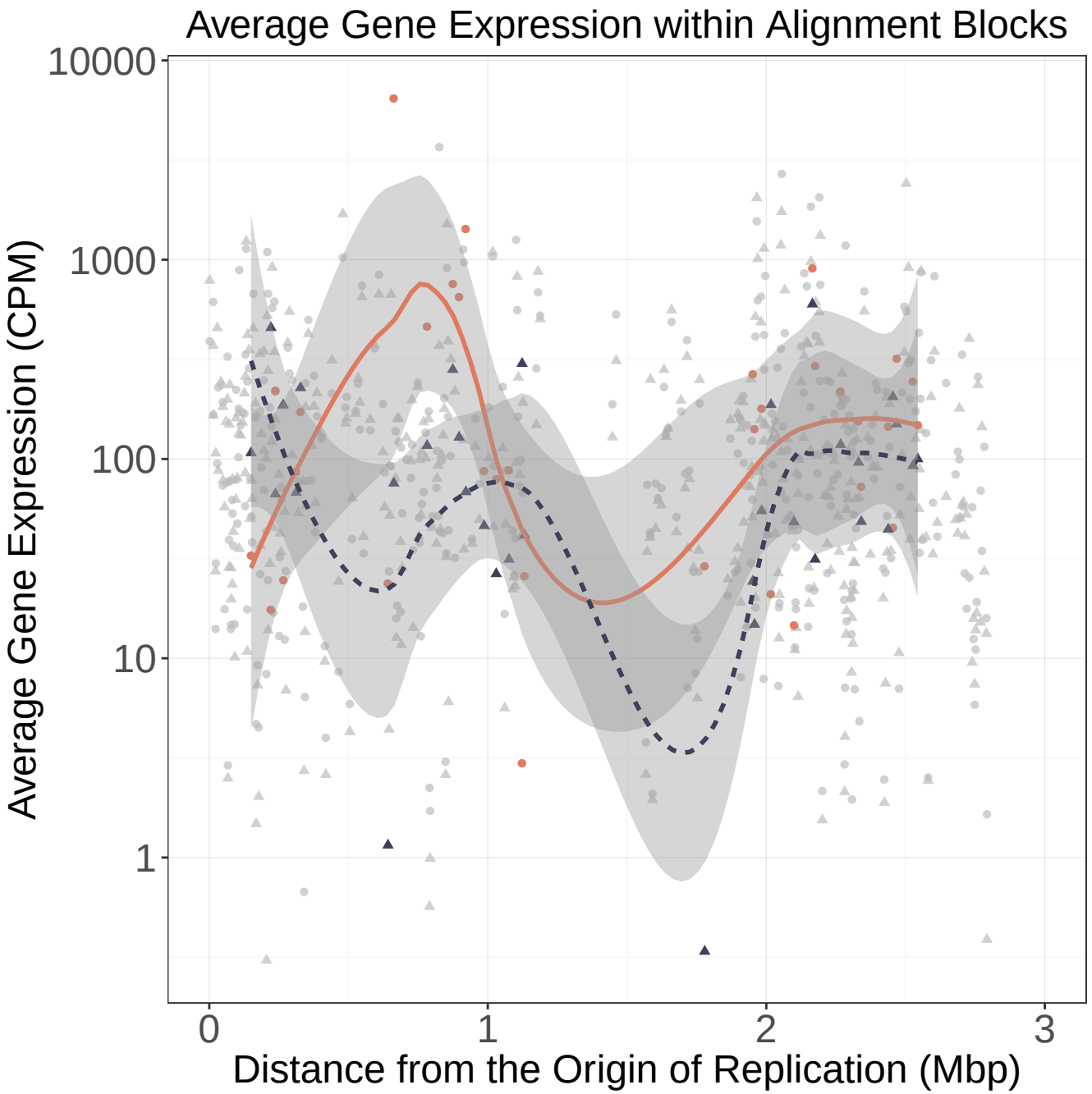

Figure 4: Visualization of the difference in gene expression between inverted and non-inverted sequences within alignment blocks. Each alignment block represents homologous sequences between the *Escherichia coli* strains insert table ref here. Each alignment block has one point on the graph to represent 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.

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

Table 2: 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 3: 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 4: 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 5: 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 6: 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 7: 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'.

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	
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	8.40×10^{-8}***
GG657756	3.62×10^{-8}***
CP042324	7.72×10^{-8}***
AL645882	7.71×10^{-8}***
CM001889	-2.46×10^{-7}***

Table 8: 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}***
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
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}***

Table 9: 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'.