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

Dissertation:

✓general formatting

✓declaration of authorship

✓final edit for intro and conclusion

✓insert substitutions chapter

✓check citations

1 Dissertation

I spent most of the break working on your edits for the final chapter of my dissertation. I also fixed formatting and citation errors. The substitutions chapter and supplement has been added to the dissertation.1

This Week

- fix any supplementary things that are missing in inversions paper
- Brian's edits for inversions paper
- send Brian new draft of inversions paper (with supplement)

Next Week

- Send dissertation to committee (Jan 18th)
- maybe do inversions in 10kb blocks? (and other sliding windows?)
- dist from ori on DESeq results?

Group	Inversions	Non-Inversions
All Blocks	3.26	3.43
Only ATCC genes	3.24	3.78
Group	Significant Inversions	Non-Significant Inversions
Significant Inversions	4.39	3.08

Table 1: Coefficient of variance in gene expression between different groups. "All Blocks" indicates all identified alignment blocks. "Only ATCC genes" indicates all ATCC genes that are both inverted and non-inverted. "Significant Inversions" indicates all inverted blocks that had a significant difference in gene expression between the inverted and non-inverted sequences. The coefficient variance in this group was calculated for the inversions that were significant inversions and non-significant inversions.

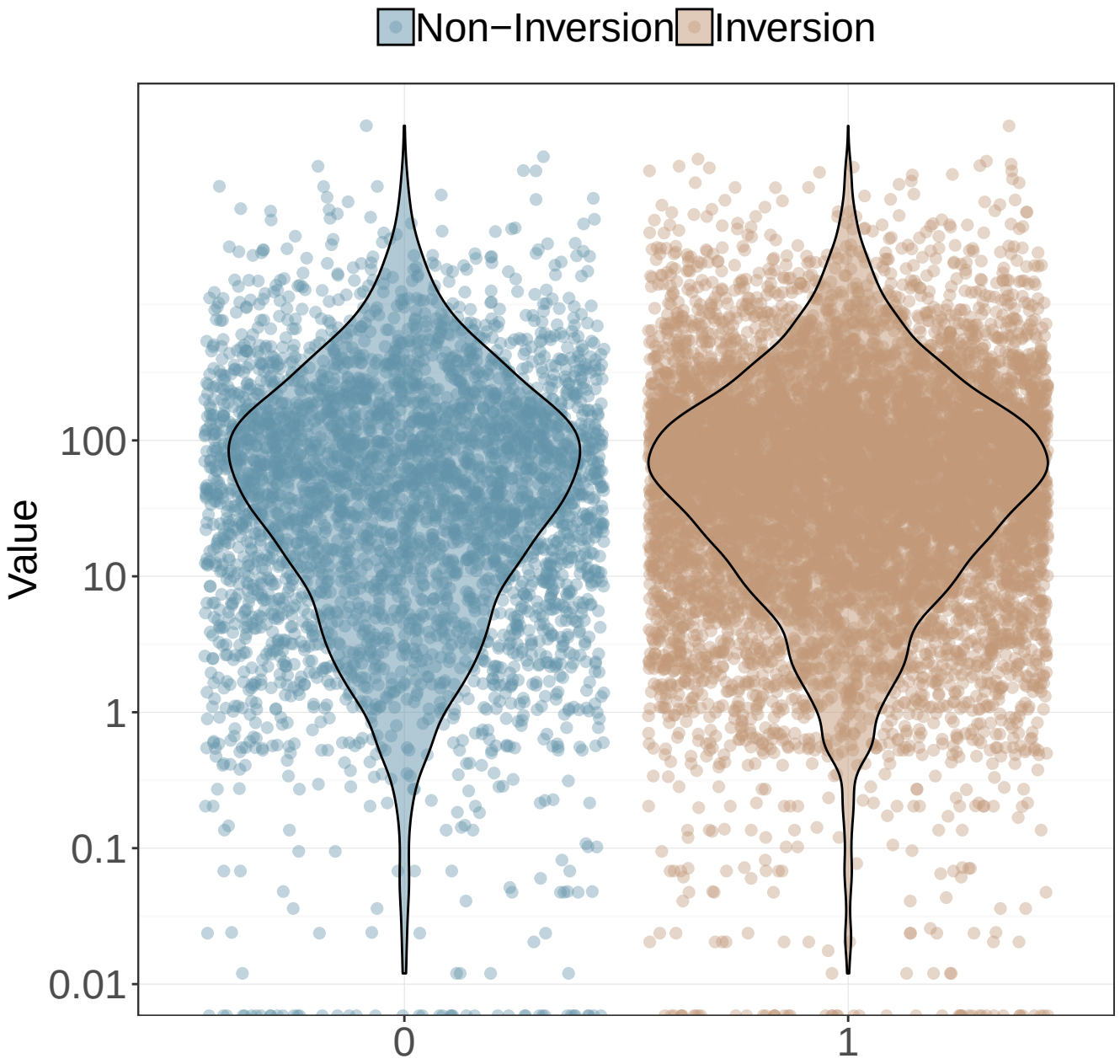

Figure 1: Violin plots of distribution of expression values between inverted and non-inverted alignment blocks.

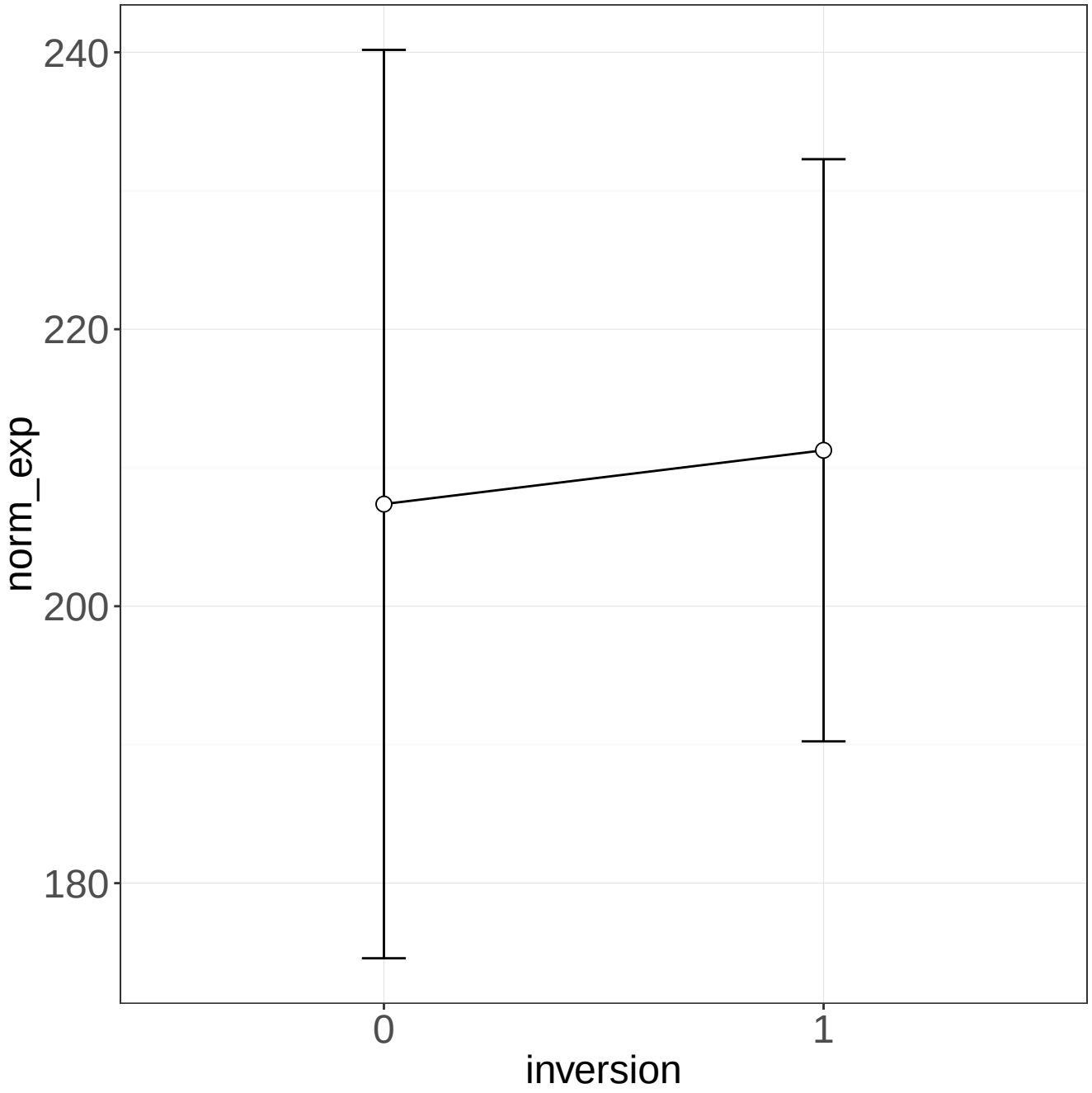

Figure 2: Plots of mean expression values between inverted and non-inverted alignment blocks with 95% confidence intervals.

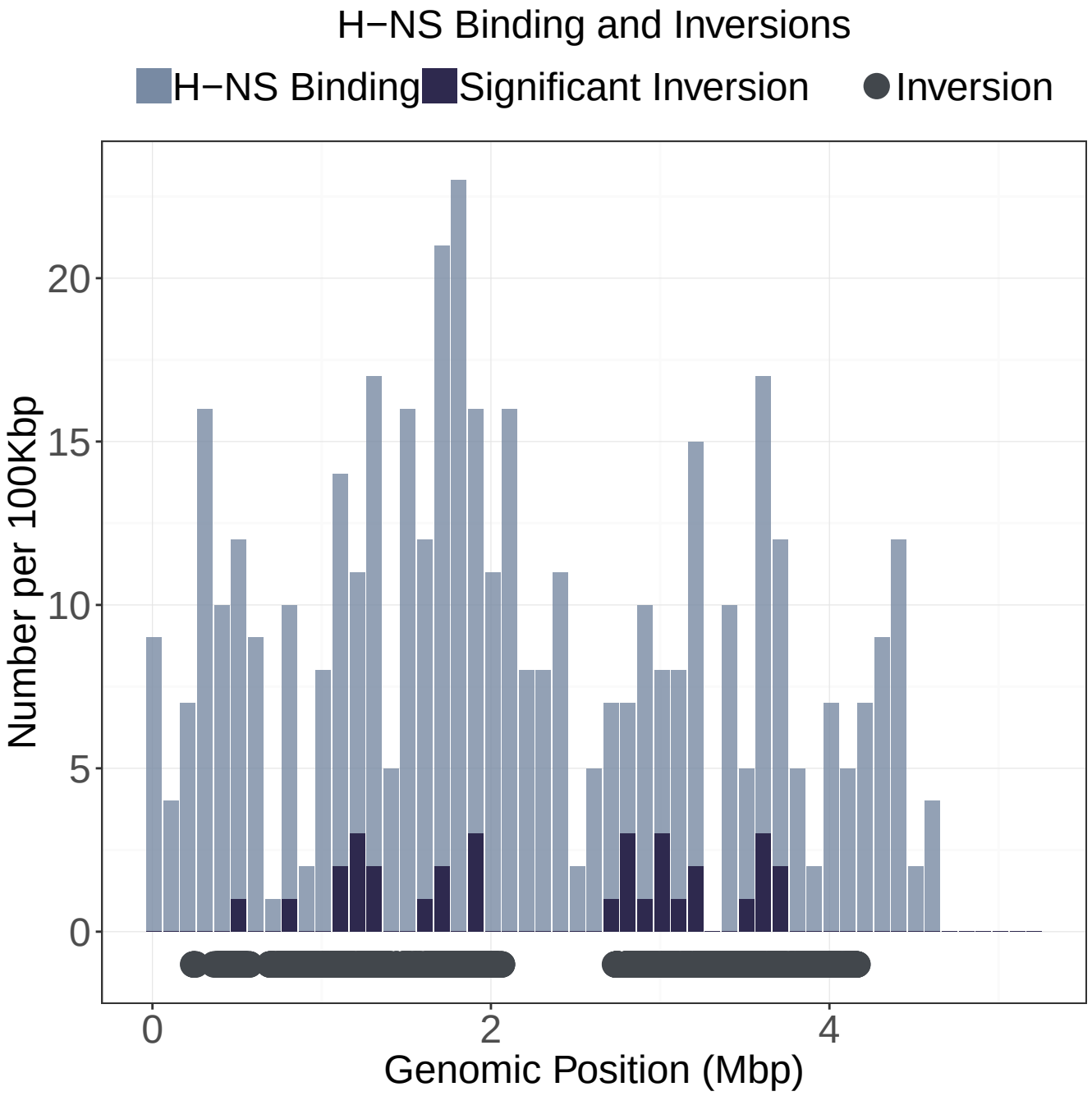

Figure 3: Visualization of the genomic locations of all inversion alignment blocks (light grey filled circles) identified between *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113, and *E. coli* ATCC. The data points are plotted on the genome of *E. coli* K-12 MG1655 which is used as a reference. 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.

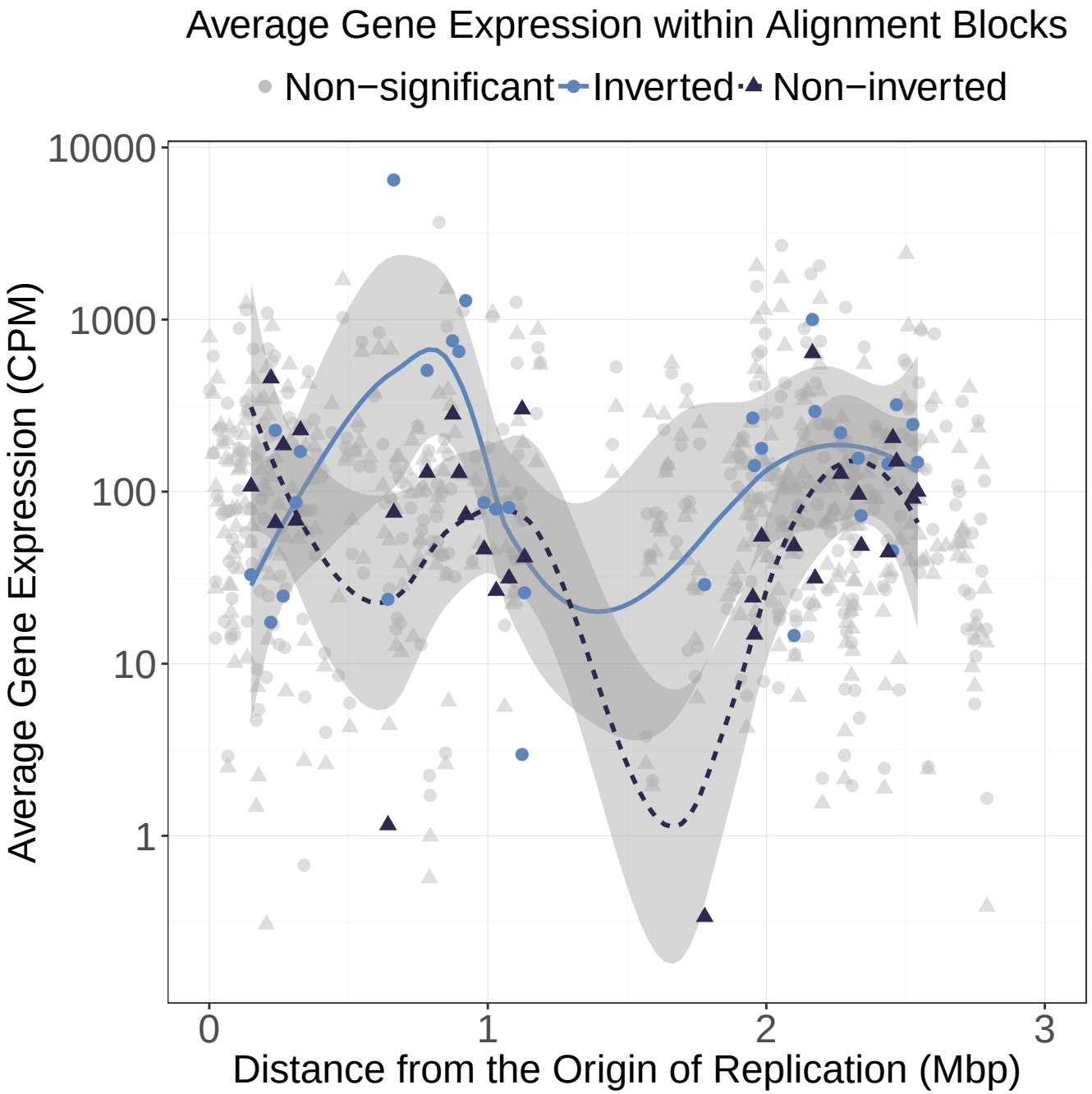

Figure 4: Visualization of the difference in gene expression between inverted and non-inverted sequences within alignment blocks. Each alignment block represents homologous sequences between the *Escherichia coli* strains insert table ref here. *E. coli* K-12 MG1655 was used as the reference genome for 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.

Group	chi-squared
All Blocks	6.005*
Only ATCC genes	6.000*
Significant Inversions	13.904***

Table 2: Fligner-Killeen test for homogeneity of variances in gene expression between different groups. "All Blocks" indicates all identified alignment blocks. "Only ATCC genes" indicates all ATCC genes that are both inverted and non-inverted. "Significant Inversions" indicates all inverted blocks that had a significant difference in gene expression between the inverted and non-inverted sequences. The coefficient variance in this group was calculated for the inversions that were significant inversions and non-significant inversions. All results are marked with significance codes as followed: < 0.001 = '***', $0.001 < 0.01$ = '**', $0.01 < 0.05$ = '*', > 0.05 = 'NS'.

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 3: 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'.