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

Inversions + Gene Expression:

✓Queenie: compare blast results and alignments

✓more inversions and position preliminary figures

✓download, read in, and plot H-NS binding data

✓deeply look into H-NS datasets

Inversions + Gene Expression:

H-NS Datasets:

I did a little more work looking into the datasets that have information for H-NS binding sites. It is summarized in Table 1 (please ignore the ugly formatting of this LaTeXTable...I did not have the energy to fix it). All datasets are essentially the same in that they removed signal noise and had additional criteria to determine what a "good" binding match was to H-NS (which is fine). Higashi et al. 2016, used three different criteria for how to define H-NS binding, and still came to the same conclusions for their analysis about sequence diversity, dN, dS...etc. This lead them to believe that the definition of H-NS binding did not matter or alter the results. I think using criteria 3) (Table 1) would give us the most amount of information. **Does this seem like ok logic?** Most of the datasets have information on if genes that are impacted by H-NS have been horizontally acquired (I believe recently). These datasets do differ in which K-12 strain they use. Reading into Oshima et al. 2006, it seems as though the W3110 and MG1655 strains differ by only 8bp, minor insertion sequences, and one small gene inversion. This paper used the annotation from K-12 MG1566 but the bacteria grown for the experiment was W3110. It appears as though these strains are nearly identical so using annotation interchangeably is ok? This makes me think it is ok to combine binding information from all the datasets. **What are your thoughts on this?** I am also unsure of what we think would be best, to take only the intersection of these datasets? To do my analysis 4 different times using each data set separately (and hopefully find the same overall conclusions and therefore can just combine them all)? **What are your thoughts on this?**

The Lang et al. 2007 paper is disappointing. I can not for the life of my find out which sub-strain of K-12 they used! Their data is rich (with binding motifs and incorporating data from the Grainger et al. 2006 paper), but if I do not know what sub-strain they used, I am not sure I can use it. Although since they directly used data from the Grainger et al. 2006 paper, it is likely the K-12 MG1655 sub-strain. **What are your thoughts on this?**

Additionally, some of the datasets have the coding sites and non-coding sites in separate files. **I should combine them and look at all the information correct?** I suspect that the gene expression data is mostly coding sites, so this might not matter.

H-NS and genome pos/Inversions

I mapped the Grainger et al 2006 dataset onto the inversions data for K-12 MG1655 in Figure 1. It looks like H-NS binds across the genome and the inversions we identified only span a portion of the genome. I will be working on doing more analysis to see if the H-NS binding is more associated with inversions or not (since visually inspecting has no trend).

Inversions and Exp graphs:

I have made a few more graphs to help visualize the inversion and gene expression data (Figures

Reference	Strain	Binding Criteria	Format	Other Notes
Higashi et al. 2016	*E. coli* K-12 W3110	Three criteria for binding sites: 1) H-NS Binding 2) H-NS Binding with cutoff (excluding genes that were restricted at the 3′ end and H-NS overlap length covered <10% of total gene length) 3) H-NS Binding and all associated genes in transcription unit **Found the same results using all three criteria for H-NS binding site. Stated that their conclusion did not depend on the definition of H-NS binding.** - Removed noise - Only looked at orthologous genes between 50 strains (binding patterns were only done with 3 wet lab strains, binding pattern was consistent across 3 strains tested in wet lab)	csv Roughly the same H-NS binding sites as 3)	HGT association Least H-NS binding genes/sites Most H-NS binding genes/sites
Grainger et al. 2006	*E. coli* K-12 MG1655	-Removed noise - cutoff criteria for what was a "good" binding site - confirmed 30/36 H-NS binding targets listed by Ecocyc	csv	
Ueda et al. 2013	*E. coli* K-12 W3110	-Removed noise - cutoff criteria for "good" binding site	csv	HGT association
Oshima et al. 2006	*E. coli* K-12 W3110	-Removed noise - cutoff criteria for "good" binding site - assessed binding sites manually	pdf :(	HGT association
Lang et al. 2007	*E. coli* K-12 **SUB-STRAIN NOT LISTED**	-Removed noise - cutoff criteria for "good" binding site	mish mash of files that will take a bit to figure out where/how to get the info	HGT association

Table 1: H-NS binding site data set information.

2 and 3). Again, please ignore the axis labels and aesthetics, these will be fixed eventually. **Let me know what you think of these and if you have any other ideas on how to arrange the data.** The blocks are so close together and there are so many of them that they appear overlapping on the graph, but each point is actually the midpoint of a block and they are all distinct. The genomic positions for each block is represented by the position of k-12 MG1655 strain in each block.

This Week

- Queenie: new dataframe for DESeq (combining blast results and raw expression data)
- Keep working on position and inversion visualization (finalize)
- determine what H-NS datasets we will be using (combine, intersect, separate..etc)
- play with parallel sets diagrams in R for inversion visualization btwn strains

Next Week

- actual analysis on DESeq data
- visualizations/results for ↑
- formal stats on inversions and H-NS association (look at Higashi 2016 methods for HGT association)
- read papers on H-NS proteins

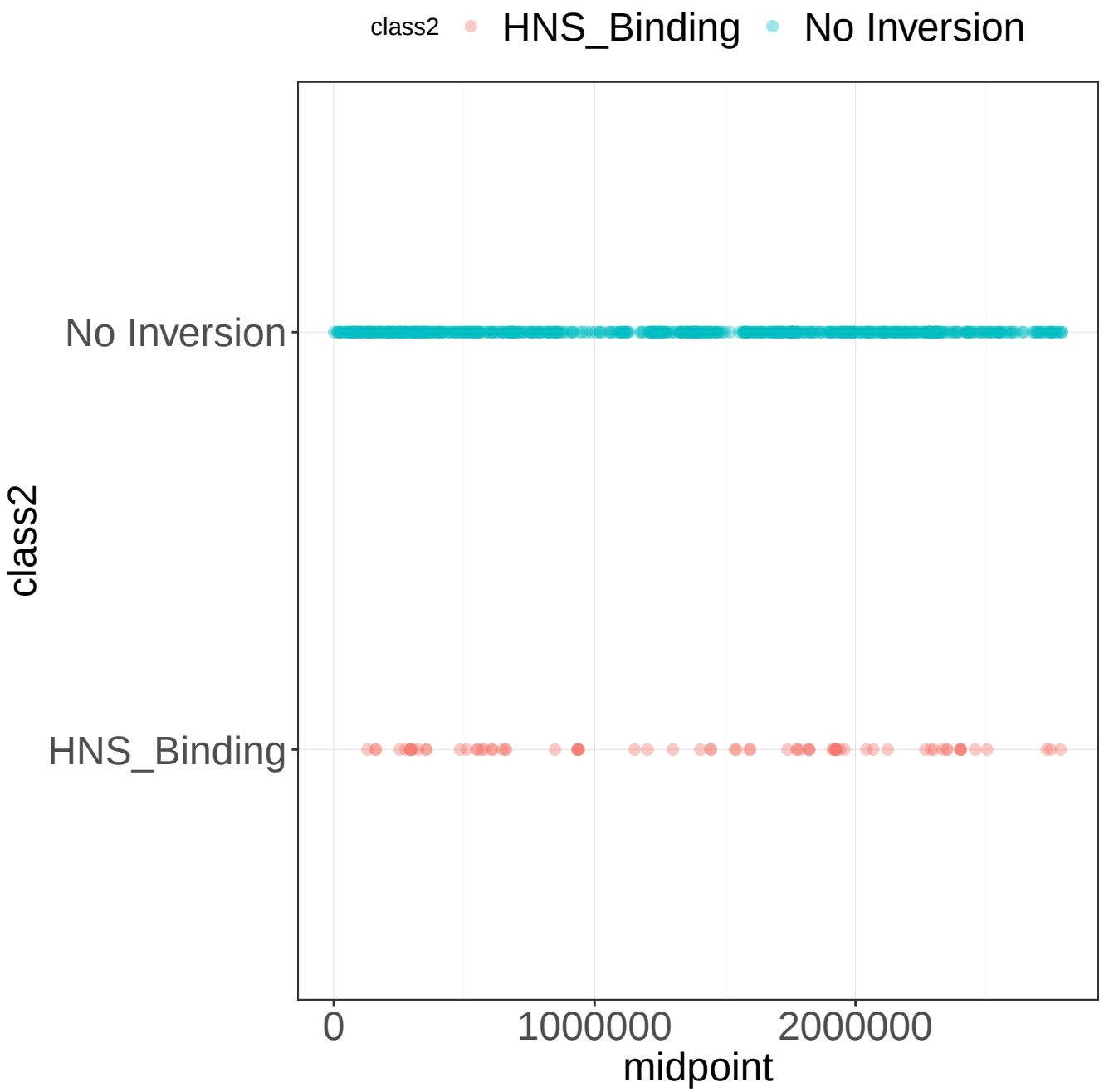

Figure 1: Regions with an inversion (ignore the fact that the label says "No Inversion") along genome (bidirectional replication) and H-NS binding sites from Grainger et al. 2006 along the genome (bidirectional replication).

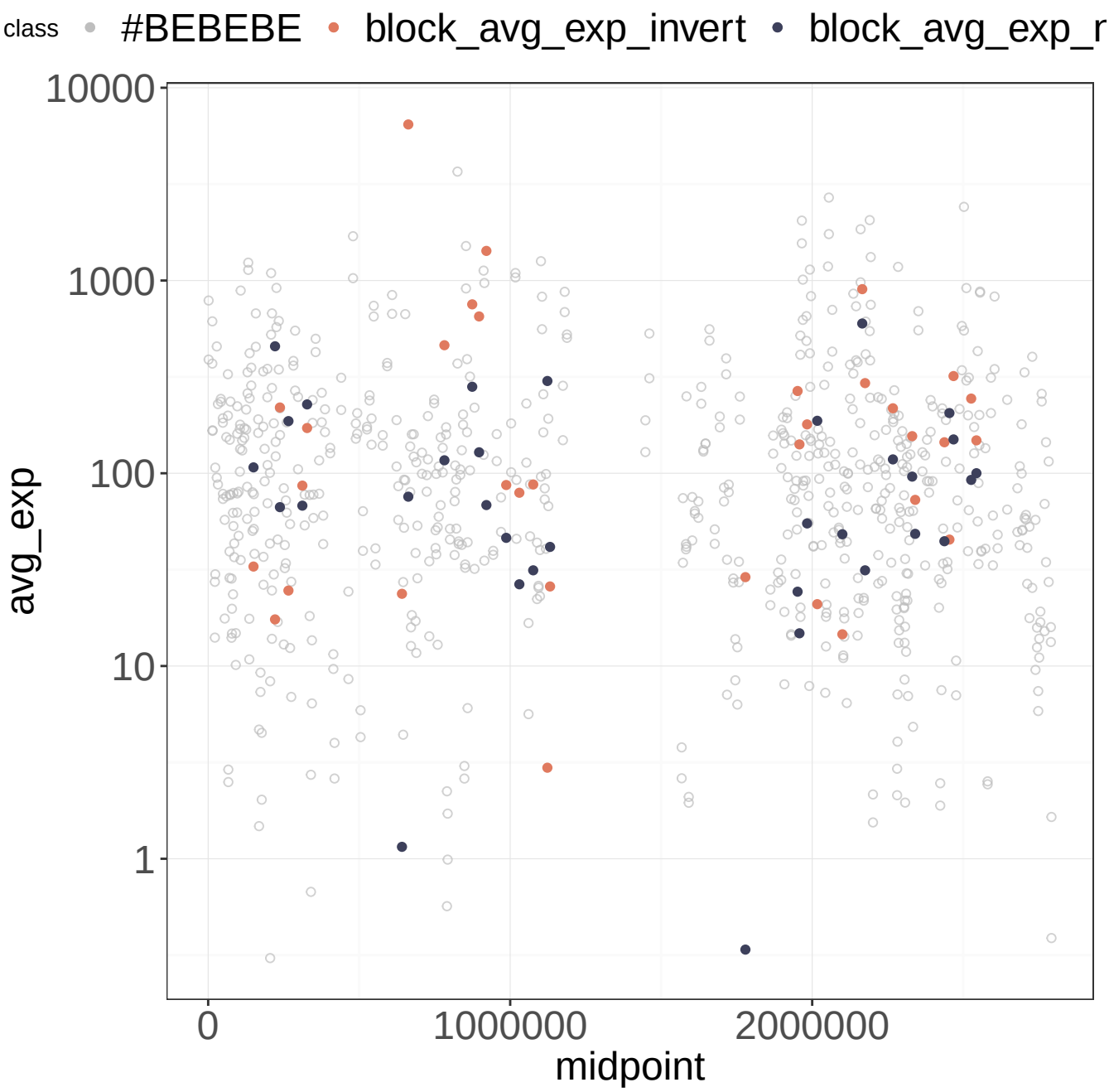

Figure 2: Two points per block that show the average inverted and non-inverted gene expression for sequences in each block. Blocks with a significant difference between inverted and non-inverted gene expression (via wilcox sign-ranked test) are highlighted in pink and dark purple respectively. Non-significant differences in gene expression blocks are represented by grey circles. x-axis is the genomic position with bidirectional replication accounted for.

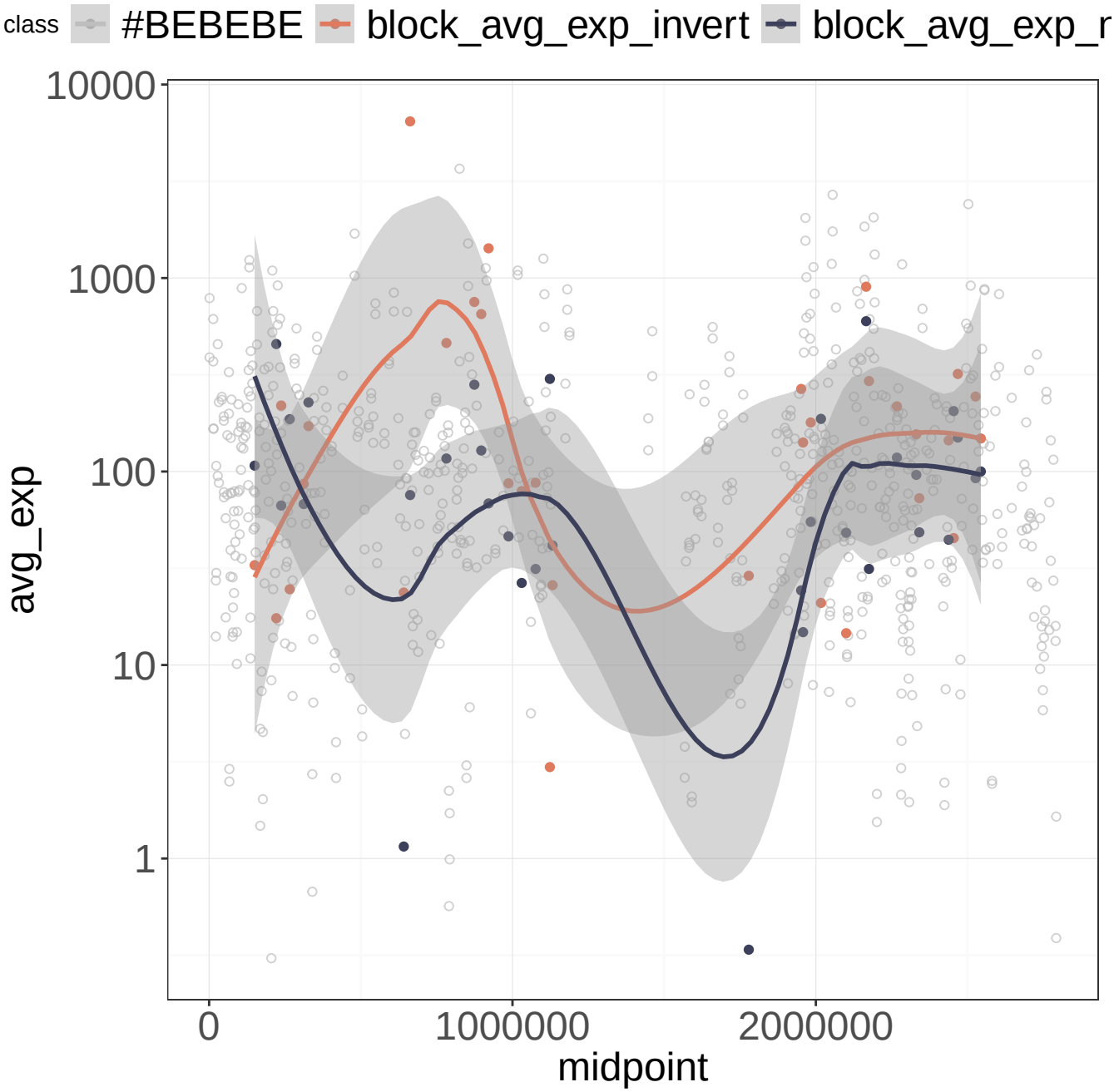

Figure 3: Two points per block that show the average inverted and non-inverted gene expression for sequences in each block. Blocks with a significant difference between inverted and non-inverted gene expression (via wilcox sign-ranked test) are highlighted in pink and dark purple respectively. Smoothing lines for significant blocks have been added with a 95% confidence interval. Non-significant differences in gene expression blocks are represented by grey circles. x-axis is the genomic position with bidirectional replication accounted for.