Subs Paper Things to Do:

- more genomes
- new outgroups? (too distant)
- explain high dS values in *B. subtilis*
- potentially poor alignment and non-orthologous genes (core genome, change methods?)
- non-parametirc analysis for subs
- gap in *Escherichia coli* fig 5
- new methods for trees
- concerned about repeated genes (TEs) and not analyzing core genome
- check if trimming respects coding frame
- clear distinction between mutations and substitutions in intro (separate sections)
- datasets from previous papers (repeat my analysis on them?)
- why would uncharacterized proteins have higher subs rates?
- R^2 values in regression analysis
- update gene exp paper ref
- 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: comparison between blast parameters

✓Queenie: compare blast and gene alignment homologs

✓Queenie: start creating dataframe that is compatible with `limma`

Subst Paper:

✓working out complications in substitutions code with more genomes

✓new outgroups

✓new phylogenetic tree methods

✓new trees

✓updated ↑ all that in the paper and cover letter

Inversions + Gene Expression: Queenie has compared the various blast outputs and found them all to be pretty similar. A few of the various parameters are similar to the default blast parameters (in output) so these were not chosen. The others were very similar in output so we chose the one that had the most genes which is `diamond`'s `-more-sensitive` option. These parameters are what will be used to verify that the alignments are aligning homologous genes.

Queenie has begun comparing the blast output and the alignment homologs, however there are lots of issues with missing gene names/IDs and not being able to have a reliable comparison. We are still working through this.

Substitution Paper

I realized that with more genomes, there are more nuances with my substitutions code (the one that ensures we are comparing the same codon position). I am currently working out these kinks. However, based on what I am seeing so far, it looks like a lot of the sites are trimmed. So the more genomes you have, the worse progressiveMauve is at catching homologous blocks, creating an unreliable alignment. I will be working on coming up with quantitative numbers for this theory.

I have found new outgroups for each of the bacteria and re-done the phylogenetic trees using a `RAxML` pipeline. As expected, even with new outgroups and new methods, the tree topology for all the chromosomes is the same as before. pSymA and pSymB have slightly different topologies, but with more resolution. I have added all this to the paper and cover letter.

As I mentioned at my committee meeting, there are 2 new *S. coelicolor* genomes (published June 2020), so I am now including them in the analysis. I have done the alignment portion of this analysis and will continue to run these genomes through my pipeline.

This Week

- Queenie: compare blast results and alignments
- Queenie: new dataframe for `limma`
- write about TEs and repeated elements in cover letter
- continue working on subst code for more genomes
- fix gene mapping code based on ↑
- re-run inversions mapping based on ↑
- re-run subst analysis for all bac based on ↑
- non-parametric analysis for subs analysis
- quantify trimming loss (more subst genomes)

Next Week

- Queenie: new dataframe for `limma`
- previous subst papers datasets (can I re-do?)
- why do uncharacterized proteins have higher sub rates?
- gap in *E. coli* fig 5
- *B. subtilis* high dS values should not be present

- blast to confirm homologs in subst analysis
- distinction between mutations and substitutions in subst paper intro

	% of Blocks that are		
Datasets: Taxa Per Block	Inverted	Inverted and Differentially Expressed	Increased in Gene Expression in the Inverted Sequences
All 4	68.15	8.66	60.61
At least 3	68.23	8.91	57.14
At least 2	68.02	8.96	58.33

Table 1: 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).

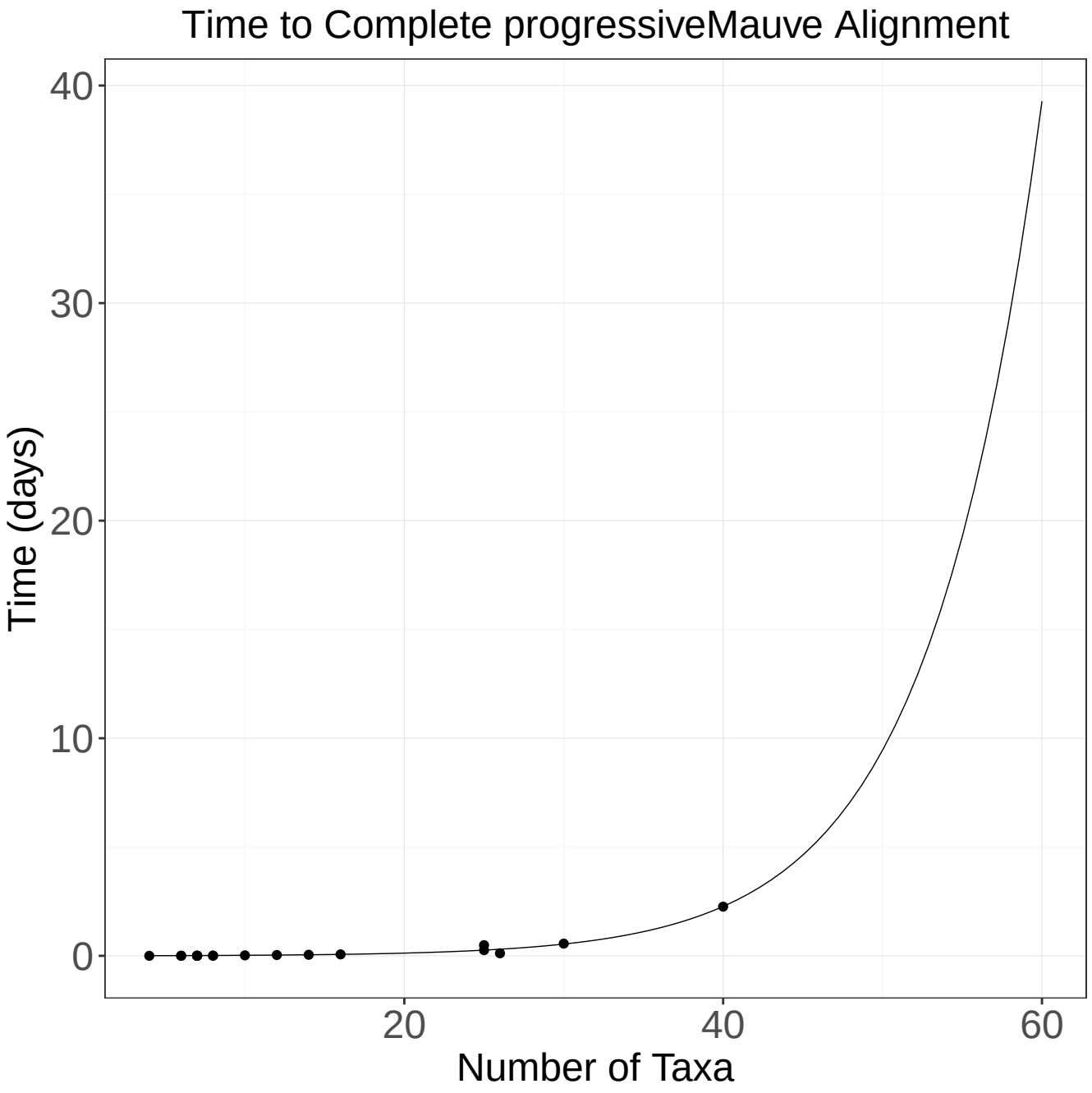

Time to Complete progressiveMauve Alignment
Time (days)
0
10
20
30
40
20
40
60
Number of Taxa

Figure 1

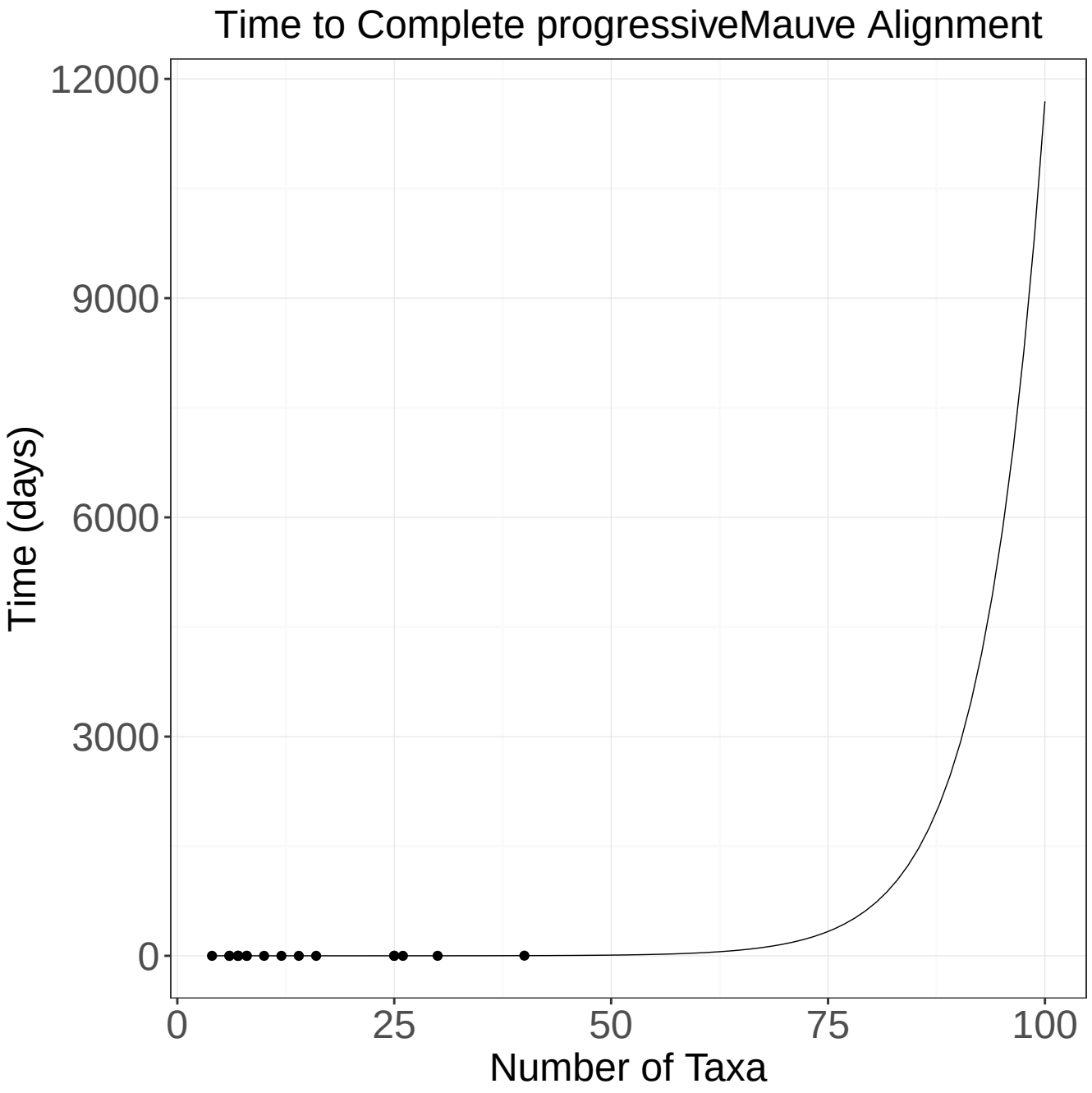
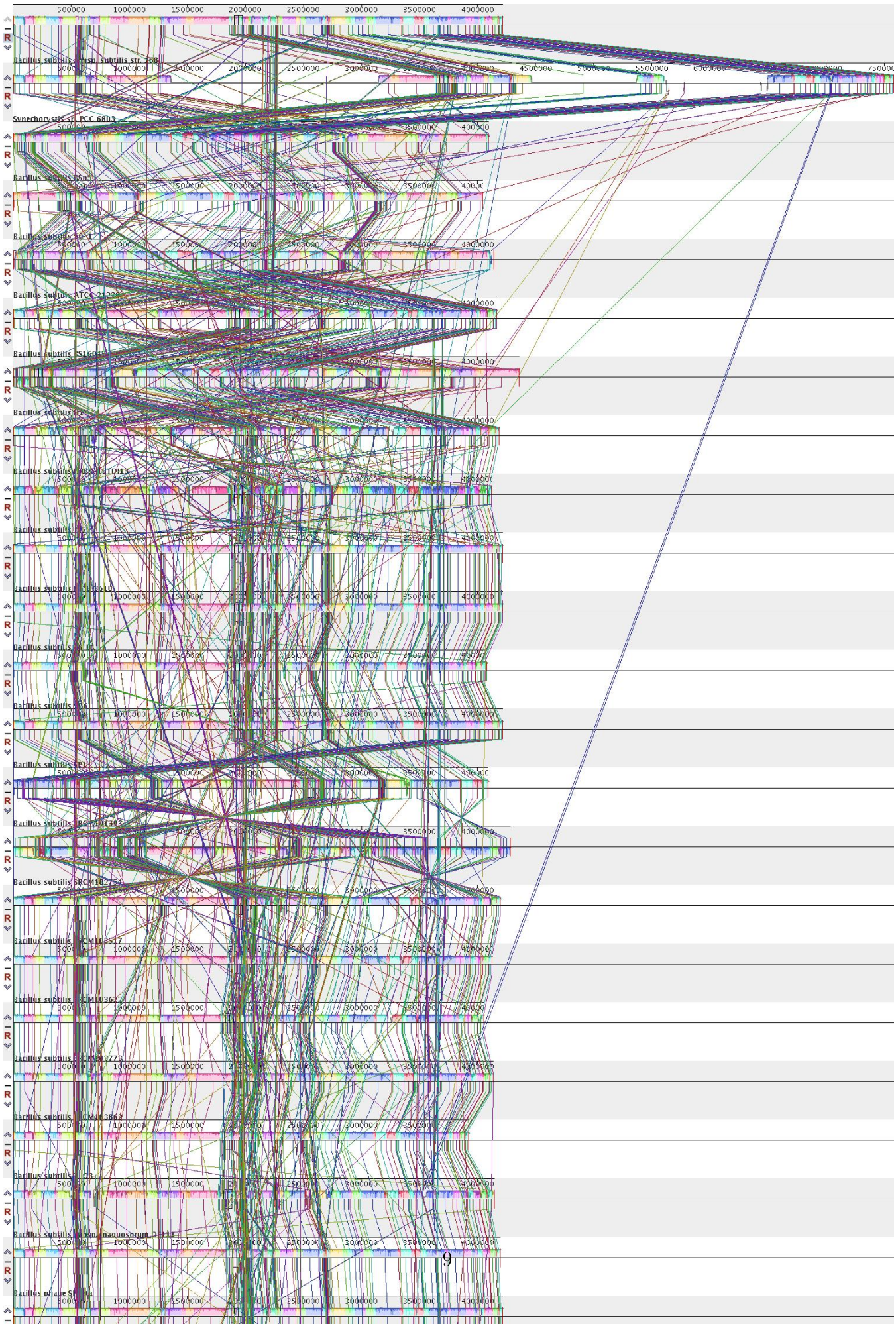

Time to Complete progressiveMauve Alignment
12000
9000
6000
3000
0
Time (days)
0
25
50
75
100
Number of Taxa

Figure 2


Escherichia coli str. K-12 substr. MG1655
Escherichia coli BL21(DE3)
Escherichia coli BW25113 K-12
Escherichia coli ECC-1470
Escherichia coli O182:H21-D181
Escherichia coli 108
Escherichia coli 112
Escherichia coli 4/1-1
Escherichia coli 4/1-2
Escherichia coli 4/3
Escherichia coli CA6_6126
Escherichia coli CI5
Escherichia coli EcoI_224
Escherichia coli EcPF16
Escherichia coli O2:K1:ExPEC_XM
Escherichia coli KBN10P04869
Escherichia coli LQ92
Escherichia coli O125:K70:H19-KCHCB63
Escherichia coli RHB26-C06
Escherichia coli RHB34-C08
Escherichia coli RHBSTW-00928


Bacillus subtilis subsp. subtilis str. 168
Synechocystis sp. PCC 6803
Bacillus subtilis BSn5
Bacillus subtilis QB928
Bacillus subtilis ATCC 21332
Bacillus subtilis BS168
Bacillus subtilis QB928 pTTDH13
Bacillus subtilis BSP1
Bacillus subtilis BGSC 1A1
Bacillus subtilis BGSC 1A6
Bacillus subtilis BSP1
Bacillus subtilis BGSC 1A1
Bacillus subtilis BGSC 1A2
Bacillus subtilis BGSC 1A3
Bacillus subtilis BGSC 1A4
Bacillus subtilis BGSC 1A5
Bacillus subtilis BGSC 1A7
Bacillus subtilis subsp. inaquosorum DSM 321
Bacillus phage SPO1

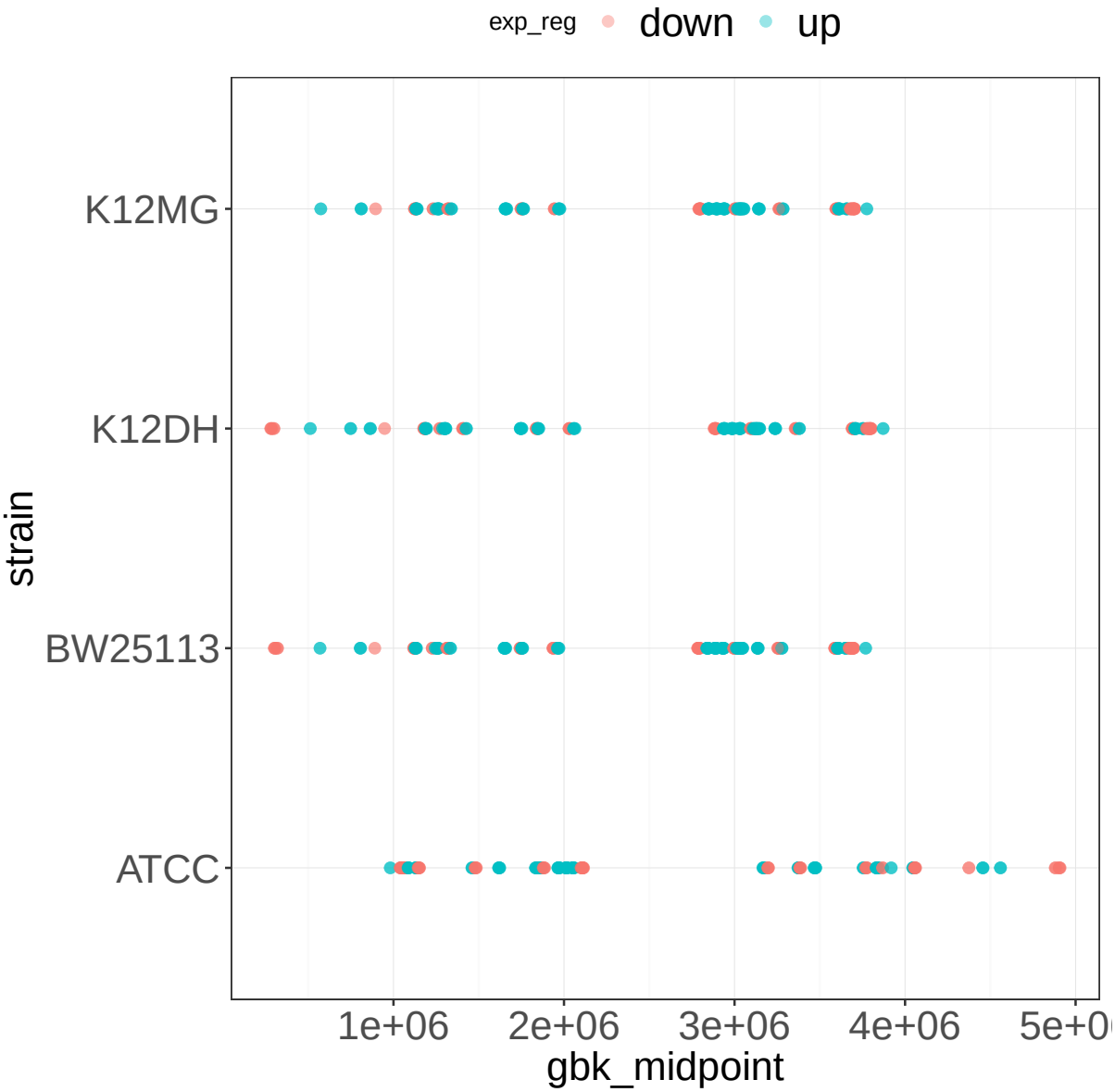


Figure 5: Test graphic for the visualization of inversions and distance from the origin of replication. Each dot represents a gene in a block where there is a significant difference in gene expression between inverted and non-inverted sequences within that block. The points are coloured based on if the inverted sequences have higher expression ("up") or lower expression ("down") compared to the non-inverted sequences. Genomic position is on the x-axis with NO bidirectional replication accounted for.