

Subs Paper Things to Do:

- causes for weird selection and subs results in *Streptomyces*
  - see how often class 4 arises in strep to see what is going on in later portion of the genome (to see if annotation is really a problem)
  - split up the strep data into core and non core and see if results are the same
- ~~make graphs proportional to length of respective cod/non-cod regions~~
- ~~test examples for genes near and far from terminus (robust log reg/results)~~
- ~~linear regression on 10kb regions for weighted and non-weighted substitutions~~
- ~~average number of substitutions in 20kb regions near and far from the origin~~
- ~~figure out why the data is weird for number of cod/non-cod sites~~
- why are the lin reg of  $dN$ ,  $dS$  and  $\omega$  NS but the subs graphs are...explain!
- grey out outliers in subs graphs?
- mol clock for my analysis?
- GC content? COG? where do these fit?

Gene Expression Paper Things to Do:

- ~~linear regression on 10kb regions~~
- ~~put new 10kb lin reg and # of genes over 10kb lin reg into paper~~
- ~~write about  $\uparrow$  in methods and discussion~~
- ~~put expression lin reg and # coding sites log reg into supplement~~
- ~~write about  $\uparrow$  in paper and how results are the same~~
- ~~update supplementary figures/file~~
- ~~correlation of gene expression across strains~~
  - ~~make graphs pretty and more informative with label names~~
  - ~~add them to supplement with a mini write up of what we did and why~~
  - ~~mention this in the actual paper~~
- if necessary add a phylogenetic component to the analysis
- ~~potentially remove genes that have been recently translocated from the analysis~~
- ~~model gene exp + position + number of genes~~

- ~~split up the strep data into core and non-core and see if results are the same~~
- ~~what is going on with *Streptomyces* number of genes changing drastically from core to non-core~~
- codon bias?
- ~~what is going on with really high gene expression bars~~
- edit paper
- submit paper

### Inversions and Gene Expression Letter Things to Do:

- ~~check if opposite strand in progressiveMauve means an inversions (check visual matches the xmfa)~~
- ~~check if PARSNP and progressiveMauve both identify the same inversions (check xmfa file)~~
- create latex template for paper
- ~~put notes from papers into doc~~
- ~~use large PARSNP alignment to identify inversions~~
- 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)
- read and make notes on papers I found for dissertation intro

## Last Week

- ✓pick which *Streptomyces* group of species to use
- ✓make tree for new *Streptomyces* species
- ✓assess and re-align the *A. tumefaciens* strains
- ✓assess *Borrelia burgdorferi* progressiveMauve alignments
- ✓finalise Gblocks pipeline, run *E. coli* through Gblocks pipeline to work out any kinks

***Streptomyces*** I decided that using species for *Streptomyces* that had more rearrangements was better than using more taxa. So I will be using the one *S. coelicolor* and two *S. lividans* genomes for my analysis. I started this by creating a new tree with branchlengths and bootstrap values for this taxa.

**Other Linear Bacteria progressiveMauve Alignments** I assessed the progressiveMauve alignments for *Borrelia burgdorferi* and all of the strains are too similar and progressiveMauve can only find one LCB. So I can not use this bacteria. I also assessed the progressiveMauve alignments for *A. tumefaciens* and found that there were combinations of the strains that yielded between 32-66 LCBs. This could be an option for my analysis. However, if I decide to use *A. tumefaciens* then I think I will have to analyze the circular and linear chromosomes, I can not just pick one. So, we decided to put this portion of the project on hold for now and focus on fixing the *Streptomyces* issues.

**Gblocks pipeline** Last week I finalised incorporating Gblocks into my pipeline and ran it on *E. coli* just to work out any final kinks. I began re-running the substitution analysis on all the bacterial strains last week and will continue to run that this week.

## This Week

Substitutions project:

1. what to do about length cut off for the non-coding regions? am I putting this whole section in the supplement?
2. get new tree picture for ne *Streptomyces* genomes
3. finish re-running subs analysis on all bacteria
4. re-run dN/dS analysis on all bacteria
5. new graphs for gene expression? (negative positions)
6. new graphs for substitutions? (negative positions), write new (and better) code for this

## Next Week

Gene Expression:

1. non-linear model for gene expression to look at periodicity?
2. new graphs for gene expression? (negative positions)
3. new graphs for substitutions? (negative positions), write new (and better) code for this
4. look into journal requirements for submission
5. write cover letter for gene expression paper

Bacteria and Replicon	Genomic Position (bp)	Protein/Gene Examples
<i>E. coli</i> Chromosome	0 - 10000	DNA replication and repair ATP-proton motive force ATP biosynthesis transport
	470000 - 480000	DNA replication and repair tRNA synthesis Ribosomal proteins Putative transport
	610000 - 620000	Ribosomal protein Translation modification tRNA modification RNA synthesis
<i>B. subtilis</i> Chromosome	0 - 10000	tRNA modification Ribosomal proteins DNA gyrase
	130000 - 140000	rRNA small subunit methylation Ribosomal proteins Elongation factor
	730000 - 740000	tRNA subunit Transcription regulation Glycolysis
	1220000 - 1230000	Pyruvate kinase Sporulation membrane proteins ATP-binding Regulation protein
<i>Streptomyces</i> Chromosome	1590000 - 1600000	Ribosomal proteins Hypothetical proteins
	1690000 - 1700000	Ribosomal proteins
<i>S. meliloti</i> Chromosome	1480000 - 1490000	Ribosomal proteins Structural elements Transmembrane proteins
<i>S. meliloti</i> pSymA	660000 - 680000	Hypothetical proteins Unknown proteins Small molecule metabolism
<i>S. meliloti</i> pSymB	290000 - 300000	Cell Division Small molecule metabolism Cell processes
	800000 - 810000	Small molecule metabolism

Table 1: Table of high median CPM (Counts per Million) gene expression over 10kb genomic regions for each bacterial replicon and the associated proteins/gene functions found in that region. The genomic position begins at the origin of replication and continues in both directions until the terminus of replication (bidirectional replication).

Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
<i>E. coli</i> Chromosome	$-5.29 \times 10^{-5}$	$1.66 \times 10^{-5}$	$< 2 \times 10^{-16}$
<i>B. subtilis</i> Chromosome	$-9.8 \times 10^{-5}$	$2.4 \times 10^{-5}$	$6.2 \times 10^{-4}$
<i>Streptomyces</i> Chromosome	$-1.307 \times 10^{-6}$	$1.72 \times 10^{-7}$	$1.3 \times 10^{-13}$
<i>S. meliloti</i> Chromosome	$8.81 \times 10^{-6}$	$4.06 \times 10^{-5}$	NS ( $8.3 \times 10^{-1}$ )
<i>S. meliloti</i> pSymA	$1.33 \times 10^{-3}$	$4.3 \times 10^{-4}$	$3 \times 10^{-3}$
<i>S. meliloti</i> pSymB	$9.55 \times 10^{-5}$	$2.1 \times 10^{-4}$	NS ( $7.5 \times 10^{-1}$ )

Table 2: Linear regression analysis of normalized expression and distance from the origin of replication. The normalized expression values were calculated by dividing the total counts per million expression value per 10kb section of the genome by the total number of genes in the respective 10kb section. Linear 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 bidirectionality of replication. NS indicates Not Significant at  $P \leq 0.05$ .

Bacteria and Replicon	Near Origin			Near Terminus		
	$dN$	$dS$	$\omega$	$dN$	$dS$	$\omega$
<i>E. coli</i> Chromosome	NS	NS	NS	NS	NS	NS
<i>B. subtilis</i> Chromosome	NS	NS	NS	NS	NS	NS
<i>Streptomyces</i> Chromosome	—	—	—	—	—	—
<i>S. meliloti</i> Chromosome	$3.77 \times 10^{-8**}$	$3.54 \times 10^{-7**}$	$1.23 \times 10^{-6**}$	NS	NS	NS
<i>S. meliloti</i> pSymA	NS	NS	$3.42 \times 10^{-5*}$	NS	NS	NS
<i>S. meliloti</i> pSymB	NS	NS	NS	$-3.24 \times 10^{-7**}$	$8.33 \times 10^{-6***}$	NS

Table 3: Linear regression for  $dN$ ,  $dS$ , and  $\omega$  calculated for each bacterial replicon for the 20 genes closest and 20 genes farthest from the origin of replication. All results are marked with significance codes as followed:  $p < 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

Bacteria and Replicon	Protein Coding Sequences	Non-Protein Coding Sequences
<i>E. coli</i> Chromosome	$-1.887 \times 10^{-8***}$	$6.462 \times 10^{-8***}$
<i>B. subtilis</i> Chromosome	$-7.200 \times 10^{-8***}$	$-1.296 \times 10^{-7***}$
<i>Streptomyces</i> Chromosome		
<i>S. meliloti</i> Chromosome	$-2.024 \times 10^{-7***}$	$-1.594 \times 10^{-7**}$
<i>S. meliloti</i> pSymA	$-5.894 \times 10^{-7***}$	$-6.904 \times 10^{-7***}$
<i>S. meliloti</i> pSymB	$1.361 \times 10^{-7***}$	$4.475 \times 10^{-7***}$

Table 4: Logistic regression analysis of the number of substitutions along all positions of the genome of the respective bacteria replicons. These genomic positions were split up into the coding and non-coding regions of the genome. 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 bidirectionality of replication. All results are marked with significance codes as followed:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

Bacteria and Replicon	Protein Coding				Non-Protein Coding			
	Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near		Correlation Coefficient 20kb Near		Number of Substitutions per 20kb Near	
	Origin	Terminus	Origin	Terminus	Origin	Terminus	Origin	Terminus
<i>E. coli</i> Chromosome	-1.343 $\times 10^{-5**}$	NS	6.26 $\times 10^{-2}$	4.25 $\times 10^{-2}$	NS	-9.231 $\times 10^{-5**}$	4.95 $\times 10^{-3}$	2.25 $\times 10^{-3}$
<i>B. subtilis</i> Chromosome	NS	3.114 $\times 10^{-5***}$	4.7 $\times 10^{-3}$	3.025 $\times 10^{-2}$	NS	7.906 $\times 10^{-5*}$	2 $\times 10^{-4}$	2 $\times 10^{-3}$
<i>Streptomyces</i> Chromosome								
<i>S. meliloti</i> Chromosome	NS	NS	4.1 $\times 10^{-3}$	2 $\times 10^{-4}$	NS	NS	7.5 $\times 10^{-4}$	1.5 $\times 10^{-4}$
<i>S. meliloti</i> pSymA	NS	NS	5.9 $\times 10^{-3}$	2.15 $\times 10^{-3}$	1.431 $\times 10^{-4**}$	NS	1.35 $\times 10^{-3}$	2.5 $\times 10^{-4}$
<i>S. meliloti</i> pSymB	-2.375 $\times 10^{-5***}$	-6.976 $\times 10^{-5***}$	3.27 $\times 10^{-2}$	2.47 $\times 10^{-2}$	NS	-6.134 $\times 10^{-5**}$	2.25 $\times 10^{-3}$	3.45 $\times 10^{-3}$

Table 5: Logistic regression on 20kb closest and farthest from the origin of replication after accounting for bidirectional replication and outliers. All results are marked with significance codes as followed:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

Bacteria and Replicon	Protein Coding		Non-Protein Coding	
	Weighted	Non-Weighted	Weighted	Non-Weighted
<i>E. coli</i> Chromosome	-4.87 $\times 10^{-10**}$	-1.839 $\times 10^{-4***}$	NS	-2.244 $\times 10^{-5***}$
<i>B. subtilis</i> Chromosome	-1.150 $\times 10^{-9**}$	-1.993 $\times 10^{-4**}$	NS	-8.24 $\times 10^{-6**}$
<i>Streptomyces</i> Chromosome				
<i>S. meliloti</i> Chromosome	-1.341 $\times 10^{-10**}$	-1.461 $\times 10^{-5**}$	-3.490 $\times 10^{-10*}$	NS
<i>S. meliloti</i> pSymA	-2.01 $\times 10^{-9*}$	-1.06 $\times 10^{-4**}$	-3.95 $\times 10^{-9**}$	NS
<i>S. meliloti</i> pSymB	NS	NS	NS	NS

Table 6: Linear regression on 10kb sections of the genome with increasing distance from the origin of replication after accounting for bidirectional replication. Weighted columns have the total number of substitutions in each 10kb section of the genome divided by the total number of protein coding and non-protein coding sites in the genome. Non-weighted columns are performing a linear regression on the total number of substitutions in each 10kb section of the genome. All results are marked with significance codes as followed:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

Bacteria and Replicon	Coefficient Estimate
<i>E. coli</i> Chromosome	NS
<i>B. subtilis</i> Chromosome	$-2.239 \times 10^{-8}***$
<i>Streptomyces</i> Chromosome	$-2.360 \times 10^{-6}***$
<i>S. meliloti</i> Chromosome	$-2.074 \times 10^{-6}***$
<i>S. meliloti</i> pSymA	$3.084 \times 10^{-7}888$
<i>S. meliloti</i> pSymB	$-2.172 \times 10^{-7}***$

Table 7: Linear regression analysis of the total number of protein coding genes per 10kb along the genome of the respective bacteria replicons. Linear 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 bidirectionality of replication. All results are marked with significance codes as followed:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .

Bacteria and Replicon	Gene Expression 10kb
<i>E. coli</i> Chromosome	$-2.742 \times 10^{-5}**$
<i>B. subtilis</i> Chromosome	$-2.198 \times 10^{-5}*$
<i>Streptomyces</i> Chromosome	$-5.230 \times 10^{-7}***$
<i>S. meliloti</i> Chromosome	NS
<i>S. meliloti</i> pSymA	NS
<i>S. meliloti</i> pSymB	NS

Table 8: Linear regression analysis of the median counts per million expression data for 10kb segments of the genome of the respective bacteria replicons. Linear 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 bidirectionality of replication. All results are marked with significance codes as followed:  $< 0.001 = '***'$ ,  $0.001 < 0.01 = '**'$ ,  $0.01 < 0.05 = '*'$ ,  $> 0.05 = 'NS'$ .



Bacteria and Replicon	Coefficient Estimate	Standard Error	P-value
<i>E. coli</i> Chromosome	$-6.03 \times 10^{-5}$	$1.28 \times 10^{-5}$	$2.8 \times 10^{-6}$
<i>B. subtilis</i> Chromosome	$-9.7 \times 10^{-5}$	$2.0 \times 10^{-5}$	$1.2 \times 10^{-6}$
<i>Streptomyces</i> Chromosome	$-1.17 \times 10^{-6}$	$1.04 \times 10^{-7}$	$< 2 \times 10^{-16}$
<i>S. meliloti</i> Chromosome	$3.97 \times 10^{-5}$	$4.25 \times 10^{-5}$	NS ( $3.5 \times 10^{-1}$ )
<i>S. meliloti</i> pSymA	$1.39 \times 10^{-3}$	$2.53 \times 10^{-4}$	$4.9 \times 10^{-8}$
<i>S. meliloti</i> pSymB	$1.46 \times 10^{-4}$	$2.03 \times 10^{-4}$	NS ( $5.34.7 \times 10^{-1}$ )

Table 9: Linear regression analysis of the median counts per million expression data along the genome of the respective bacteria replicons. Linear 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 bidirectionality of replication.

Bacteria and Replicon	$dN$	$dS$	$\omega$
<i>E. coli</i> Chromosome	NS	NS	NS
<i>B. subtilis</i> Chromosome	NS	NS	$-9.08 \times 10^{-6*}$
<i>Streptomyces</i> Chromosome	NS	NS	NS
<i>S. meliloti</i> Chromosome	NS	NS	NS
<i>S. meliloti</i> pSymA	NS	NS	NS
<i>S. meliloti</i> pSymB	NS	NS	$1.163 \times 10^{-5*}$

Table 10: Linear regression for  $dN$ ,  $dS$ , and  $\omega$  calculated for each bacterial replicon on a per genome basis. All results are marked with significance codes as followed:  $p < 0.001 = '***'$ ,  $0.001 < p < 0.01 = '**'$ ,  $0.01 < p < 0.05 = '*'$ ,  $p > 0.05 = 'NS'$ .

Bacteria and Replicon	Average Expression Value (CPM)
<i>E. coli</i> Chromosome	160.500
<i>B. subtilis</i> Chromosome	176.400
<i>Streptomyces</i> Chromosome	6.084
<i>S. meliloti</i> Chromosome	271.400
<i>S. meliloti</i> pSymA	690.100
<i>S. meliloti</i> pSymB	595.700

Table 11: Arithmetic gene expression calculated across all genes in each replicon. Expression values are represented in Counts Per Million.

Bacteria and Replicon	Gene Average			Genome Average		
	dS	dN	$\omega$	dS	dN	$\omega$
<i>E. coli</i> Chromosome	1.0468	0.1330	1.3183	0.6491	0.0364	0.2432
<i>B. subtilis</i> Chromosome	4.652	0.2333	2.4200	1.0879	0.0703	0.3852
<i>Streptomyces</i> Chromosome	13.4950	2.0973	21.0423	5.1256	0.8911	8.9146
<i>S. meliloti</i> Chromosome	0.0184	0.0012	0.1069	0.0187	0.0013	0.0962
<i>S. meliloti</i> pSymA	1.0602	0.7451	5.1290	0.4100	0.0863	0.8311
<i>S. meliloti</i> pSymB	3.2602	0.0256	0.3878	0.1436	0.0100	0.1943

Table 12: Weighted averages calculated for each bacterial replicon on a per genome basis using the gene length as the weight. Arithmetic mean calculated for the per gene averages for each bacterial replicon.

Bacteria Strain/Species	GEO Accession Number	Date Accessed
<i>E. coli</i> K12 MG1655	GSE60522	December 20, 2017
<i>E. coli</i> K12 MG1655	GSE73673	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE85914	December 19, 2017
<i>E. coli</i> K12 MG1655	GSE40313	November 21, 2018
<i>E. coli</i> K12 MG1655	GSE114917	November 22, 2018
<i>E. coli</i> K12 MG1655	GSE54199	November 26, 2018
<i>E. coli</i> K12 DH10B	GSE98890	December 19, 2017
<i>E. coli</i> BW25113	GSE73673	December 19, 2017
<i>E. coli</i> BW25113	GSE85914	December 19, 2017
<i>E. coli</i> O157:H7	GSE46120	August 28, 2018
<i>E. coli</i> ATCC 25922	GSE94978	November 23, 2018
<i>B. subtilis</i> 168	GSE104816	December 14, 2017
<i>B. subtilis</i> 168	GSE67058	December 16, 2017
<i>B. subtilis</i> 168	GSE93894	December 15, 2017
<i>B. subtilis</i> 168	GSE80786	November 16, 2018
<i>S. coelicolor</i> A3	GSE57268	March 16, 2018
<i>S. natalensis</i> HW-2	GSE112559	November 15, 2018
<i>S. meliloti</i> 1021 Chromosome	GSE69880	December 12, 2017
<i>S. meliloti</i> 2011 pSymA	NC_020527 (Dr. Finan)	April 4, 2018
<i>S. meliloti</i> 1021 pSymA	GSE69880	November 15, 18
<i>S. meliloti</i> 2011 pSymB	NC_020560 (Dr. Finan)	April 4, 2018
<i>S. meliloti</i> 1021 pSymB	GSE69880	November 15, 18

Table 13: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.