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 ω 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 ↑ in methods and discussion~~
- ~~put expression lin reg and # coding sites log reg into supplement~~
- ~~write about ↑ 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

✓gene expression edits/wrote more

✓take into account number of genes (normalized gene expression based on this)

✓figure out what is going on with *Streptomyces* number of genes drastically increasing after core region

✓investigate really high gene expression bars in all bacteria

✓start new graph with negative positions (subs and gene expression)

Gene Expression Paper I did a lot of work over the weekend editing my gene expression paper and adding in more information. I have another few sections ready for you to edit, so let me know if you want me to send those to you.

***Streptomyces* uneven number of gene** Trying to figure out why there seemed to be such a jump in the total number of genes for my gene expression data at around 2260000 from the origin (bidirectionally). Looking into this I realised that it is because this is where the end of the short arm of the chromosome ends. So, because I am calculating my positions bidirectionally, they are "folding" the two chromosome arms together, artificially increasing the total number of genes in that section of the graph. I separated this out into each of the replicon halves, and still got the same answer, where the total number of substitutions were larger near the origin of replication and smaller near the terminus. This also sparked discussion about how I should be visualising this. There are some papers that use negative numbers to indicate the left half of the replichore and positive numbers to indicate the right half of the replichore, leaving the origin of replication somewhere in the middle of the graph. I am fine with presenting the data this way, but then I think that I should be doing my regression analysis on both replichore halves separately, because I can't really treat it all as one if I am showing the graph with negative and positive numbers. Thoughts?

High Gene expression bars in all bacteria? I looked into the high gene expression bars in all the bacteria and found that they are mostly full of ribosomal proteins, DNA replication and repair proteins, and translation modification (see below Table 1). This is consistent across all the bacteria! Which makes sense why these bars are so much higher than the others! It is super interesting that these are not always found near the origin of replication! I added this explanation into my paper draft.

Normalising by number of genes I have been thinking about the number of genes in each 10kb section and how this may potentially be impacting/biasing my gene expression results. So I decided to do a better normalisation of the gene expression and the number of genes. I took the total expression value per 10kb chunk of the genome (added up all gene expression values for each 10kb region), then divided this by the total number of protein coding genes in that section. I then did a linear regression on these values, and still obtained the same result (Table coming soon). Not sure which of these tables should be the "main result" found in the paper and which should be supplemental? thoughts? I was also thinking that I should probably have one graph with these normalised by gene number values, to avoid having two graphs ontop of eachother (one for gene expression and one for the total number of genes). Thoughts?

This Week

Back to the substitutions project:

1. pick which group of *Streptomyces* genomes to use
2. what to do about length cut off for the non-coding regions? am I putting this whole section in the supplement?
3. assess the *Borrelia burgdorferi* mauve plots for similar seqs
4. align the *A. tumefaciens* taxa into more specific phylo groups with mauve
5. assess the agro mauve plots
6. run *E. coli* through the new pipeline to work out any kinks

Next Week

Substitutions project continued:

1. run either *Streptomyces* or one of the other linear bacteria through the substitutions pipeline
2. run the rest of the bacteria through the new pipeline
3. new graphs for gene expression? (negative positions)
4. new graphs for substitutions? (negative positions), write new (and better) code for this

Bacteria and Replicon	Genomic Position (bp)	Protein/Gene Examples
E. coli 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
B. subtilis Chromosome	0 - 10000	tRNA modification
		Ribosomal proteins
		DNA gyrase
		rRNA small subunit methylation
	130000 - 140000	Ribosomal proteins
		Elongation factor
	730000 - 740000	tRNA subunit
		Transcription regulation
		Glycolysis
	1220000 - 1230000	Pyruvate kinase
		Sporulation membrane proteins
		ATP-binding
		Regulation protein
Streptomyces Chromosome	1590000 - 1600000	Ribosomal proteins
		Hypothetical proteins
	1690000 - 1700000	Ribosomal proteins
S. meliloti Chromosome	1480000 - 1490000	Ribosomal proteins
		Structural elements
		Transmembrane proteins
S. meliloti pSymA	660000 - 680000	Hypothetical proteins
		Unknown proteins
		Small molecule metabolism
S. meliloti 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
E. coli Chromosome	-5.29×10^{-5}	1.66×10^{-5}	$<2\times10^{-16}$
B. subtilis Chromosome	-9.8×10^{-5}	2.4×10^{-5}	6.2×10^{-4}
Streptomyces Chromosome	-1.307×10^{-6}	1.72×10^{-7}	1.3×10^{-13}
S. meliloti Chromosome	8.81×10^{-6}	4.06×10^{-5}	NS (8.3×10^{-1})
S. meliloti pSymA	1.33×10^{-3}	4.3×10^{-4}	3×10^{-3}
S. meliloti pSymB	9.55×10^{-5}	2.1×10^{-4}	NS (7.5×10^{-1})

Table 2: Linear regression analysis of normalized expression and distance from the origin of replication. The noramlized 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 $P \leq 0.05$.

	Near Origin			Near Terminus		
Bacteria and Replicon	dN	dS	ω	dN	dS	ω
E. coli Chromosome	NS	NS	NS	NS	NS	NS
B. subtilis Chromosome	NS	NS	NS	NS	NS	NS
Streptomyces Chromosome	—	—	—	—	—	—
S. meliloti Chromosome	3.77×10^{-8}**	3.54×10^{-7}**	1.23×10^{-6}**	NS	NS	NS
S. meliloti pSymA	NS	NS	3.42×10^{-5}*	NS	NS	NS
S. meliloti pSymB	NS	NS	NS	-3.24×10^{-7}**	8.33×10^{-6}***	NS

Table 3: Linear regression for dN, dS, and ω 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
E. coli Chromosome	-4.308×10^{-8}***	NS
B. subtilis Chromosome	-4.971×10^{-8}***	-1.055×10^{-7}***
Streptomyces Chromosome		
S. meliloti Chromosome	-1.903×10^{-7}***	-2.900×10^{-7}***
S. meliloti pSymA	-6.642×10^{-7}***	-1.263×10^{-6}***
S. meliloti pSymB	1.769×10^{-7}***	4.771×10^{-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’.

	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	
Bacteria and Replicon	Origin	Terminus	Origin	Terminus	Origin	Terminus	Origin	Terminus
E. coli Chromosome	$-2.889\times10^{-5*}$	NS	2.87×10^{-2}	4.24×10^{-2}	$-4.316\times10^{-5**}$	$-8.209\times10^{-5*}$	1.095×10^{-2}	4.45×10^{-3}
B. subtilis Chromosome	NS	$1.863\times10^{-5*}$	4.8×10^{-3}	3.06×10^{-2}	$1.017\times10^{-4*}$	$5.823\times10^{-5***}$	8×10^{-4}	6.75×10^{-3}
Streptomyces Chromosome								
S. meliloti Chromosome	NS	NS	4.05×10^{-3}	2×10^{-4}	NS	NS	9×10^{-4}	1.5×10^{-4}
S. meliloti pSymA	NS	NS	6.15×10^{-3}	1.9×10^{-3}	$1.403\times10^{-4***}$	$-2.220\times10^{-4**}$	2.8×10^{-3}	5.5×10^{-4}
S. meliloti pSymB	$-1.553\times10^{-5*}$	$-4.908\times10^{-5***}$	3.23×10^{-2}	2.36×10^{-2}	NS	$-4.557\times10^{-5**}$	5.1×10^{-3}	5.4×10^{-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'.

	Protein Coding		Non-Protein Coding	
Bacteria and Replicon	Weighted	Non-Weighted	Weighted	Non-Weighted
E. coli Chromosome	$-4.87\times10^{-10**}$	$-1.839\times10^{-4***}$	NS	$-2.244\times10^{-5***}$
B. subtilis Chromosome	NS	$-2.031\times10^{-4**}$	NS	$-2.885\times10^{-5**}$
Streptomyces Chromosome				
S. meliloti Chromosome	$-1.341\times10^{-10**}$	$-1.461\times10^{-5**}$	$-3.490\times10^{-10*}$	NS
S. meliloti pSymA	NS	NS	$-1.144\times10^{-8**}$	$-6.74\times10^{-5**}$
S. meliloti 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	Gene Expression 10kb
E. coli Chromosome	-2.742×10^{-5}**
B. subtilis Chromosome	-2.198×10^{-5}*
Streptomyces Chromosome	-5.230×10^{-7}***
S. meliloti Chromosome	NS
S. meliloti pSymA	NS
S. meliloti pSymB	NS

Table 7: 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
E. coli Chromosome	-6.03×10^{-5}	1.28×10^{-5}	2.8×10^{-6}
B. subtilis Chromosome	-9.7×10^{-5}	2.0×10^{-5}	1.2×10^{-6}
Streptomyces Chromosome	-1.17×10^{-6}	1.04×10^{-7}	$<2\times10^{-16}$
S. meliloti Chromosome	3.97×10^{-5}	4.25×10^{-5}	NS (3.5×10^{-1})
S. meliloti pSymA	1.39×10^{-3}	2.53×10^{-4}	4.9×10^{-8}
S. meliloti pSymB	1.46×10^{-4}	2.03×10^{-4}	NS ($5.34.7\times10^{-1}$)

Table 8: 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	Coefficient Estimate
E. coli Chromosome	NS
B. subtilis Chromosome	-2.682×10^{-6}***
Streptomyces Chromosome	-2.360×10^{-6}***
S. meliloti Chromosome	-2.074×10^{-6}***
S. meliloti pSymA	NS
S. meliloti pSymB	-4.19×10^{-6}*

Table 9: 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	dN	dS	ω
E. coli Chromosome	NS	NS	NS
B. subtilis Chromosome	NS	NS	-9.08×10^{-6}*
Streptomyces Chromosome	NS	NS	NS
S. meliloti Chromoeom	NS	NS	NS
S. meliloti pSymA	NS	NS	NS
S. meliloti pSymB	NS	NS	1.163×10^{-5}*

Table 10: Linear regression for dN, dS, and ω calculated for each bacterial replicon on a per genome basis. 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	Average Expression Value (CPM)
E. coli Chromosome	160.500
B. subtilis Chromosome	176.400
Streptomyces Chromosome	6.084
S. meliloti Chromosome	271.400
S. meliloti pSymA	690.100
S. meliloti pSymB	595.700

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

	Gene Average			Genome Average		
Bacteria and Replicon	dS	dN	ω	dS	dN	ω
E. coli Chromosome	1.0468	0.1330	1.3183	0.6491	0.0364	0.2432
B. subtilis Chromosome	4.652	0.2333	2.4200	1.0879	0.0703	0.3852
Streptomyces Chromosome	13.4950	2.0973	21.0423	5.1256	0.8911	8.9146
S. meliloti Chromosome	0.0184	0.0012	0.1069	0.0187	0.0013	0.0962
S. meliloti pSymA	1.0602	0.7451	5.1290	0.4100	0.0863	0.8311
S. meliloti 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
E. coli K12 MG1655	GSE60522	December 20, 2017
E. coli K12 MG1655	GSE73673	December 19, 2017
E. coli K12 MG1655	GSE85914	December 19, 2017
E. coli K12 MG1655	GSE40313	November 21, 2018
E. coli K12 MG1655	GSE114917	November 22, 2018
E. coli K12 MG1655	GSE54199	November 26, 2018
E. coli K12 DH10B	GSE98890	December 19, 2017
E. coli BW25113	GSE73673	December 19, 2017
E. coli BW25113	GSE85914	December 19, 2017
E. coli O157:H7	GSE46120	August 28, 2018
E. coli ATCC 25922	GSE94978	November 23, 2018
B. subtilis 168	GSE104816	December 14, 2017
B. subtilis 168	GSE67058	December 16, 2017
B. subtilis 168	GSE93894	December 15, 2017
B. subtilis 168	GSE80786	November 16, 2018
S. coelicolor A3	GSE57268	March 16, 2018
S. natalensis HW-2	GSE112559	November 15, 2018
S. meliloti 1021 Chromosome	GSE69880	December 12, 2017
S. meliloti 2011 pSymA	NC_020527 (Dr. Finan)	April 4, 2018
S. meliloti 1021 pSymA	GSE69880	November 15, 18
S. meliloti 2011 pSymB	NC_020560 (Dr. Finan)	April 4, 2018
S. meliloti 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.