Subs Paper Things to Do:

- ~~# of coding and non-coding sites~~
- ~~# of subs in each of †~~
- ~~Look into *Streptomyces* non-coding issue~~
- ~~Look into *E. coli* coding issue~~
- ~~Look into pSymB coding/non-coding trend weirdness~~
- ~~Figure out why *Streptomyces* appears to have tons of coding data missing~~
- ~~Figure out what is going on with cod/non-cod code and why it is still not working!~~
- ~~write up methods for coding/non-coding~~
- ~~write methods and results for clustering~~
- ~~start code to split alignment into multiple alignments of each gene~~
- figure out how to deal with overlapping genes
- figure out how to deal with gaps in gene of ref taxa
- split up the alignment into multiple alignments of each gene
- get dN/dS for coding/non-coding stuff per gene
- Or get 1st, 2nd, 3rd codon pos log regs
- write up coding/non-coding results
- take out gene expression from this paper
- write better intro/methods for distribution of subs graphs
- mol clock for my analysis?
- write discussion for coding/non-coding
- GC content? COG? where do these fit?
- write coding/non-coding into conclusion

Gene Expression Paper Things to Do:

- ~~look for more GEO expression data for *S. meliloti*~~
- ~~look for more GEO expression data for *Streptomyces*~~
- ~~look for more GEO expression data for *B. subtilis*~~

- ~~format paper and put in stuff that is already written~~
- ~~look for more GEO expression data for *E. coli*~~
- ~~Get numbers for how many different strains and multiples of each strain I have for gene expression~~
- ~~re-do gene expression analysis for *B. subtilis*~~
- ~~re-do gene expression analysis for *E. coli*~~
- ~~find papers about what has been done with gene expression~~
- read papers ↑
- put notes from ↑ papers into word doc
- write abstract
- write intro
- add stuff from outline to Data section
- create graphs for expression distribution (no sub data)
- add # of genes to expression graphs (top)
- average gene expression
- write discussion
- write conclusion
- add into methods: filters for Hiseq, RT PCR and growth phases for data collection
- update supplementary figures/file

Inversions and Gene Expression Letter Things to Do:

- ~~get as much GEO data as possible~~
- ~~find papers about inversions and expression~~
- ~~see how many inversions I can identify in these strains of *Escherichia coli* with gene expression data~~
- check if opposite strand in progressiveMauvemeans an inversions (check visual matches the xmfa)
- check if PARSNP and progressiveMauveboth identify the same inversions (check xmfa file)
- create latex template for paper
- read papers ↑

- put notes from papers ↑ into doc
- use large PARSNP alignment to identify inversions
- confirm inversions with dot plot
- 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

Last Week

✓write up methods for coding/non-coding

✓write methods and results for clustering

✓see how many inversions I can identify in these strains of *Escherichia coli* with gene expression data

✓start code to split alignment into multiple alignments of each gene

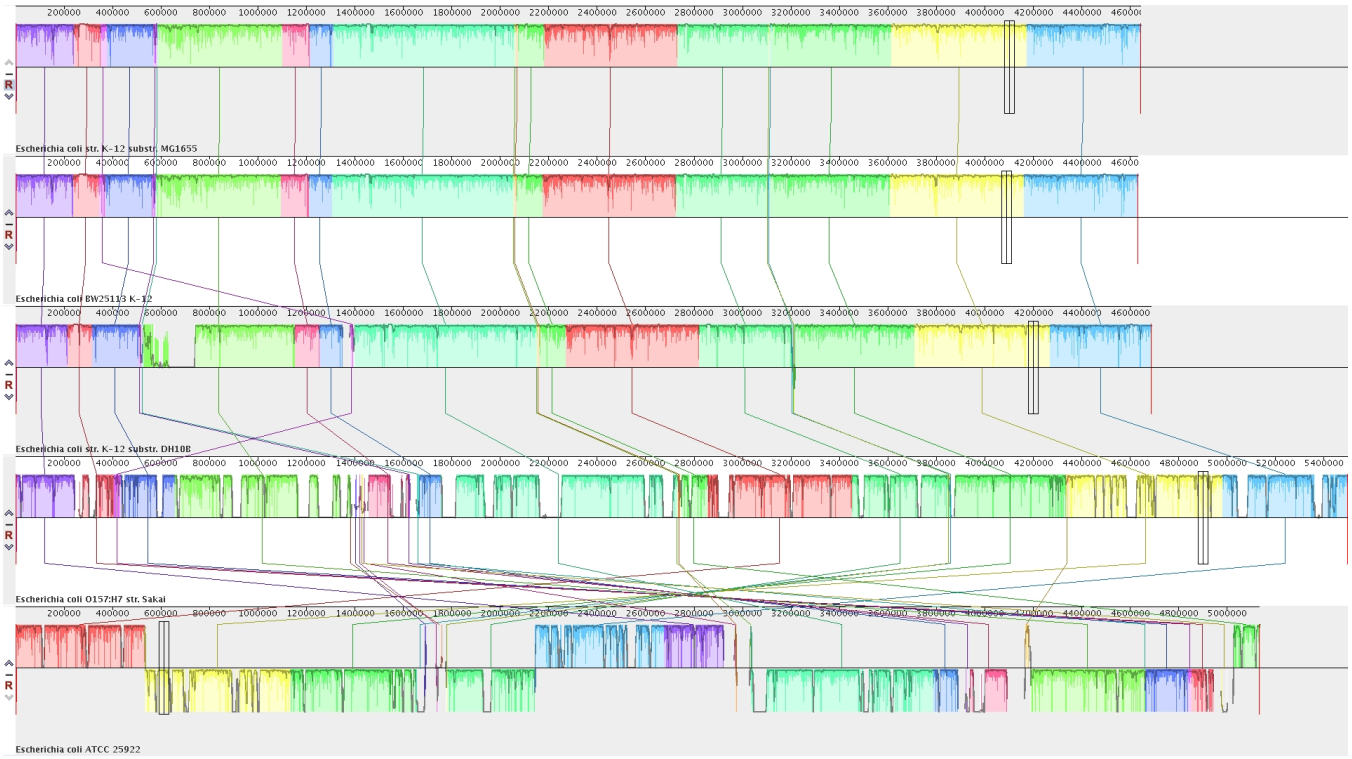
Did some easy writing last week for the substitution paper.

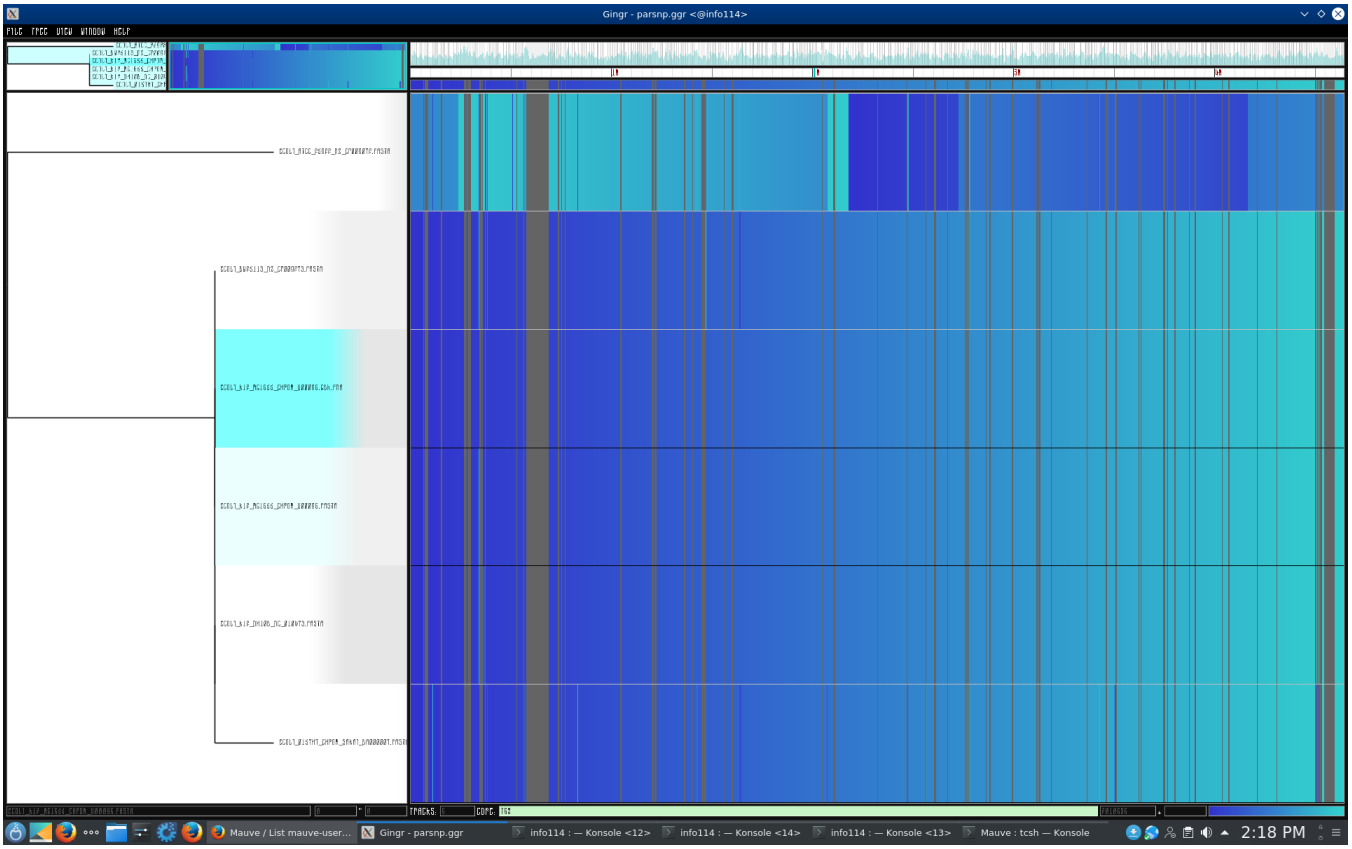
I also looked at the inversions and gene expression for the 5 strains of *E. coli* and there are inversions (pics of alignments below, the taxa with all the weird stuff going on is ATCC)!! Mauve identified 3 large inversions roughly: 1,000,000bp, 800,000bp and 500,000bp. Parsnp identified similar inversions (in similar locations) roughly: 1,350,000bp, 547,000bp and 1,376,300bp. Both of the programs had some sections that were not within an LCB or part of the core genome within these regions so the actual continuous length of the inversions may be smaller. These inversions only exist in one of the bacteria: *Escherichia coli* ATCC. The other 4 strains are all VERY similar with hardly any rearrangements, let alone inversions. So based on this we decided that we can just compaire the inversions between K-12 MG1655 and ATCC. I believe that progressiveMauveis only showing information from one strand so when something is listed on the opposite strand from the reference in progressiveMauvethen it should be an inversion. I obviously need to check this and make sure this is the case. If it is, then it will make coding to identify inversions super simple! Also need to check if the inversions identified by progressiveMauveand PARSNP are the same.

I have mostly been working on getting started on the dN/dS stuff so that this can get done and I can submit this paper! We agreed that it is best to calculate dN/dS per gene using annotation from the reference genome. My first task with this was to chop up the alignment of each block into the alignment of each gene. This is proving to be more difficult than I thought. At the moment I have and array that tells me where in the alignment there is a gap in the reference or not and if there is not then it assigns the actual genomic position to the array at that spot in the alignment (so I know where the alignment falls within the genome at each column of the aln). I also have an array that will classify each site in the genome as coding or non-coding. If it is coding, it further classifies based on the gene name. So non-coding = 0 and coding = gene name. **I have not figured out what to do when one gene resides in another gene or genes overlap.** Ideally each site can be apart of multiple genes and I would sort of subset the alignment for each gene and print that out. However, I have not quite figured out how to do this yet. I also have an array that then combines the two above ones that will at each site in the alignment tell me if it is non-coding or in a gene and which gene. The last thing I am struggling with is how to actually print the alignment into a proper format. BioPython makes it really easy to read in the alignment and print out columns and manipulate, but I am not sure if it can print out sections of the alignment with proper headers and such. This would probably require me to have a list of start and stop alignment positions, which I can easily get now. So I need to think about this more. If you have any suggestions of things in python or another program like samtools that can print out sections of the alignment I would love to hear about it! I am also unsure of what to do when there is a gap in the middle of the alignment of the ref taxa. Do I just not include any gaps? (they get cut out eventually from the analysis, but I do not know if they will mess up the dN/dS calculation PAML does...)

Next Week/Holiday Break

I have found about 30 articles on both gene expression and/or inversions in bacteria. My plan is to read and make notes on all of these over the break and apply for a few more scholarships


Escherichia coli str. K-12 substr. MG1655
Escherichia coli BW25113 K-12
Escherichia coli str. K-12 substr. DH10B
Escherichia coli O157:H7 str. Sakai
Escherichia coli ATCC 25922


Gingr - parsnp.ggr <@info114>
FILE TREE VIEW WINDOW HELP
Mauve / List mauve-user...
Gingr - parsnp.ggr
info114 : — Konsole <12>
info114 : — Konsole <14>
info114 : — Konsole <13>
Mauve : tcsh — Konsole
2:18 PM

Bacteria and Replicon	% of Coding Sequences	% of Non-Coding Sequences	% of Subs Coding	% of Subs Non-Coding
E. coli Chromosome	86.47%	13.53%	5.00%	8.96%
B. subtilis Chromosome	87.49%	12.51%	7.31%	6.42%
Streptomyces Chromosome	89.03%	10.97%	13.74%	14.91%
S. meliloti Chromosome	86.27%	13.73%	0.19%	0.22%
S. meliloti pSymA	83.34%	16.66%	2.84%	4.58%
S. meliloti pSymB	88.81%	11.19%	2.78%	3.44%

Table 1: Total proportion of coding and non-coding sites in each replicon and the percentage of those sites that have a substitution (multiple substitutions at one site are counted as two substitutions).

Bacteria and Replicon	Coding Sequences	Non-Coding Sequences
E. coli Chromosome	-5.938×10^{-8}***	-9.237×10^{-8}***
B. subtilis Chromosome	-7.584×10^{-8}***	NS
Streptomyces Chromosome	5.483×10^{-7}***	9.182×10^{-9}***
S. meliloti Chromosome	-1.448×10^{-6}***	-7.037×10^{-7}***
S. meliloti pSymA	-9.704×10^{-7}***	-1.464×10^{-7}***
S. meliloti pSymB	5.007×10^{-7}***	NS

Table 2: 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 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 3: Summary of strains and species found for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided.