Final Project Proposal

Daniella Lato and Jana Taha

The Data:

We have data from 6 replicons from 4 different species of bacteria: *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. All of the bacteria have their genome contained in one chromosome except *S. meliloti* which is a multirepliconic bacteria. A multirepliconic bacteria means that the genome is made up of multiple replicons or chromosome like structures. For this reason, each replicon of *S. meliloti* (chromosome, pSymA, pSymB) will be analyzed separately. So we effectively have 6 bacterial categories:

Bacteria	Replicon Name
E. coli	Chromosome
B. subtilis	Chromosome
Streptomyces	Chromosome
S. meliloti	Chromosome
S. meliloti	pSymA
S. meliloti	pSymB

There are three main datasets associated with these bacteria: Substitutions, Gene Expression, and Selection.

Substitutions Data:

The substitutions data set gives information about the number of substitutions (effectively mutations) and the distance from the origin of replication. This data is binary in nature: at each base pair in the genome, there is a substitution present (1) or there is not (0). The data has a phylogenetic component to the analysis and accounts for any substitutions that may also be present in the ancestor of the bacterial strains. Therefore, multiple substitutions may have occurred at a particular base in the genome. The genomic positions in this data set have been scaled to represent base pair distance from the origin of replication, with the furthest distance from the origin of replication being the terminus of replication.

Gene Expression Data:

The gene expression data set has median Counts Per Milltion (CPM) expression values for each gene in the genome. The expression data sets for this analysis were only RNA-seq data sets for control data, where this was defined as the bacteria being grown in environments absent of any stress. Each gene has an associated genomic position (the midpoint between the protein coding start and protein coding end of the gene) which was also scaled to represent base pair distance from the origin of replication.

Selection Data:

The selection data set has information on the non-synonymous (dN) substitution rate, synonymous (dS) substitution rate, and ω (dN / dS) for most genes in the bacterial replicons and the relation of these to distance from the origin of replication. This information allows us to make inferences about the selective pressures acting on a gene. Non-synonymous substitutions cause a change to the amino acid sequence of a gene, which could alter the function of the gene. Synonymous substitutions do not alter the amino acid sequence of a gene, and therefore are not expected to significantly impact the function of an organism. The ω ratio allows us to determine if these changes in the sequence cause beneficial or deleterious traits to arise. If ω for a gene is larger than 1, the gene is under positive selection and therefore is beneficial to the organism and will likely be maintained in the genome over time. If ω is less than 1, the gene is under purifying or negative selection, and therefore is deleterious to the organism and will likely not be maintained in the genome over time. If ω is equal to 1, the gene is under neutral selection, and is neither beneficial nor deleterious to the organism.

Biological Background

All of the datasets are looking at how the response variables change with distance from the origin of replication. There are certain properties that are believed to be associated with distance from the origin of replication. Near the origin of replicaion we expect genes to be more conserved and encoding for essential functions than genes located near the terminus of replication. Genes near the origin typically therefore, have higher gene expression and less mutations or substitutions, because they are important to the function of the organism. We expect that most genes (in any genome) are under neutral or purifying selection (removing deleterious traits), regardless of their genomic location (neutral theory or nearly neutral theory). Since genes near the terminus are changing often (mutations) and involved in local envoronment adaptation, we could suppose that these genes might be the best candidates for positive selection (increase benificial traits).

This leaves us with three predictions for our data sets:

1. The number of substitutions should increase when moving away from the origin of replication
2. Gene expression should decrease when moving away from the origin of replication
3. Most genes should be under neutral or purifying selection, any genes that are under positive selection should be located near the terminus

The Plots:

Selection Data: For this data set, we are going to fit three linear models that tests the effect of the distance from the origin of replication of genes on non-synonymous (dN) substitution rate, synonymous (dS) substitution rate, and ω (dN / dS). We are going to include a coefficient plot that will compare the three models. This plot can confirm to us whether the third

prediction stated above holds. We might also consider fitting other models on this data set. Exploratory graphics will also be included that will help us visualize how our attribute effects our three response variables. Then comment on how our exploratory plot(s) compare to our inferential plot(s).

Gene Expression: For this data set, we are also going to fit a model. This model will tests how the distance from the origin of replication of genes effects its expression value. Inferential plots will be used to help us interpret the model results. We will also have some exploratory graphics to go along the inferential plots that will visualize that effect. The inferential and exploratory plots should help us confirm the second prediction stated above.

Since the Selection data set and the Gene data set have a column in common, the gene ID, we could consider including a plot that contains variables from both data sets and see if genes that are highly expressed are also under a certain selective pressure.

Substitution Data Sets (Might not include in the final project): We have six data sets for each Bacterial replicon,and each data set contains over 30 million observations. We are not going to fit a model for these data sets, but we will use exploratory plots to help us see whether the position from origin of replication of different sites effects whether there is a substitution at that site or not. For each site, there is information from 10 different branches of the phylogenetic tree of the bacterial strains. Maybe we can use box plots, where our y-axis is the position and x-axis is our 10 branches,and then color the boxplots by whether there is a substitution or not. We could also consider facets. This plot will help us see whether the number of substitutions increase when moving away from the origin of replication (The first prediction stated above), and the effect that each branch have on a site.