

# 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
<i>E. coli</i>	Chromosome
<i>B. subtilis</i>	Chromosome
<i>Streptomyces</i>	Chromosome
<i>S. meliloti</i>	Chromosome
<i>S. meliloti</i>	pSymA
<i>S. meliloti</i>	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 Million (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  $\omega$  (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  $\omega$  ratio allows us to determine if these changes in the sequence cause beneficial or deleterious traits to arise. If  $\omega$  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  $\omega$  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  $\omega$  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 replication 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 environment adaptation, we could suppose that these genes might be the best candidates for positive selection (increase beneficial 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