Final Project Write-Up

Daniella Lato and Jana Taha

December 2019

Visualizing Molecular Trends in Bacterial Genomes

The Data: Our data is biologically based and mostly deals with genome wide trends. We will be looking at gene expression and selection in four bacterial genomes: *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 multi-repliconic bacteria. A multi-repliconic 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. The gene expression data set has information about the average expression value of the gene (averaged across multiple data sets) and the genomic location of that gene relative to the origin of replication. Additionally, we have obtained selection information on a few of these genes from each bacterial genome. This selection information tells us about the synonymous substitution rate (dS, mutations that do not cause a change in the amino acid sequence), the non-synonymous substitution rate (dN, mutations that cause a change in the amino acid sequence), and *omega* (dN/dS). The ω ratio allows us to determine if these substitutions in the sequence will be maintained or deleted over time. 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. This selection data is again linked to the relative distance from the origin of replication.

Both data sets are looking at how the response variables change 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 environmental 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. Gene expression should decrease when moving away from the origin of replication
2. Most genes should be under neutral or purifying selection, any genes that are under

positive selection should be located near the terminus

Selection Data

We first present a graph showing the distributions of dN, dS and ω values in each of the bacterial replicons.

```
vio_str_box
```

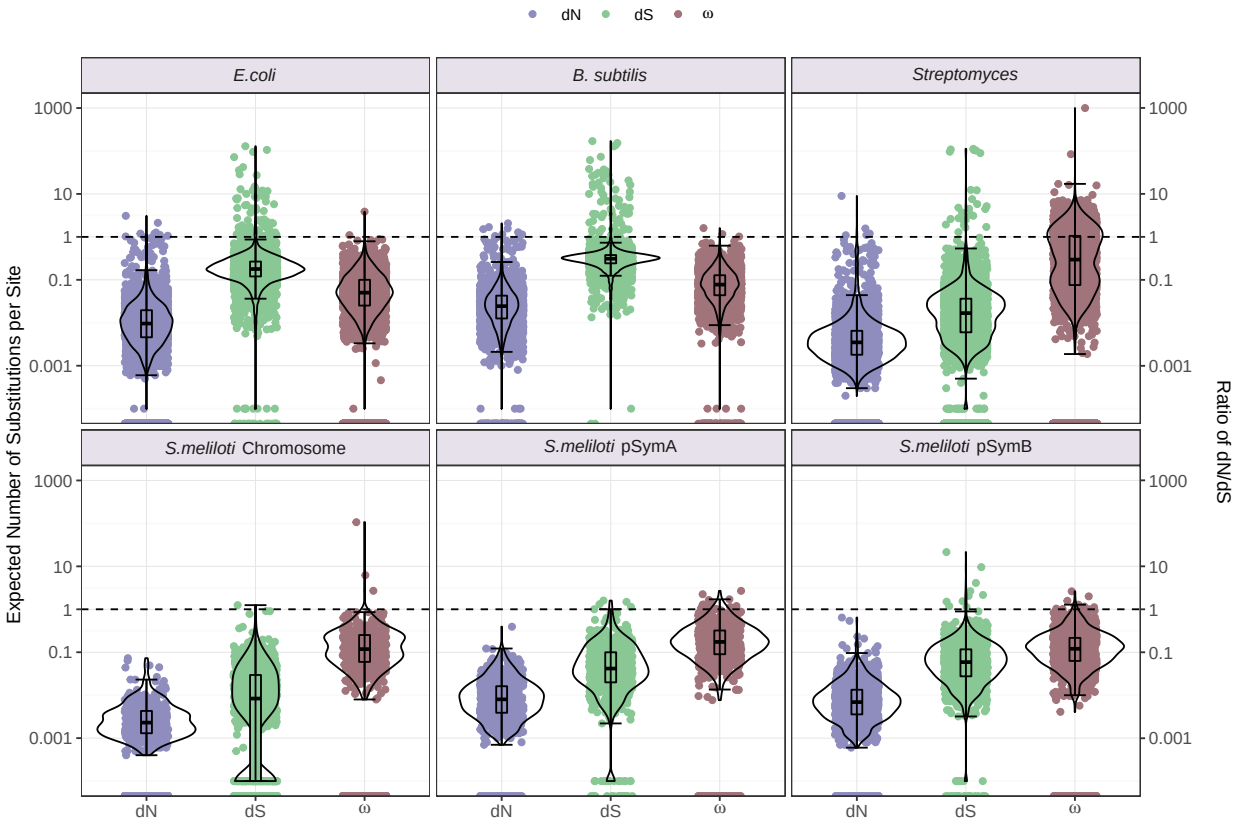
```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 10021 rows containing non-finite values (stat_ydensity).
## Warning: Removed 10021 rows containing non-finite values (stat_boxplot).

## Warning: Removed 10021 rows containing non-finite values (stat_boxplot).
```



When looking at dN and dS substitution rates, we expect that the rate of synonymous sub-

stitutions (dS) should be higher than the rate of non-synonymous substitutions. Biologically, mutations that cause a change in an amino acid are more likely to alter the function of the protein than mutations that do not cause a change in an amino acid. As mentioned, a non-functional protein could have catastrophic consequences on the well being of the organism. Across all the bacterial replicons we see that indeed, dS > dN.

We also notice that most of the *omega* values for each bacterial replicon are below 1, this is what we expected. An *omega* value below one means that the genes are likely neutral or under negative selection, meaning that mutations having deleterious impact on the organism will be removed. The notable exception to this is *Streptomyces*, which appears to have a bi-modal distribution of *omega* values with a high number of genes with omega values at or above 1. *Streptomyces* creates 80% of the antibiotics that we currently use. This means that the genome of *Streptomyces* would generally benefit from positive selection, where mutations that confer a benefit to the organism are retained.

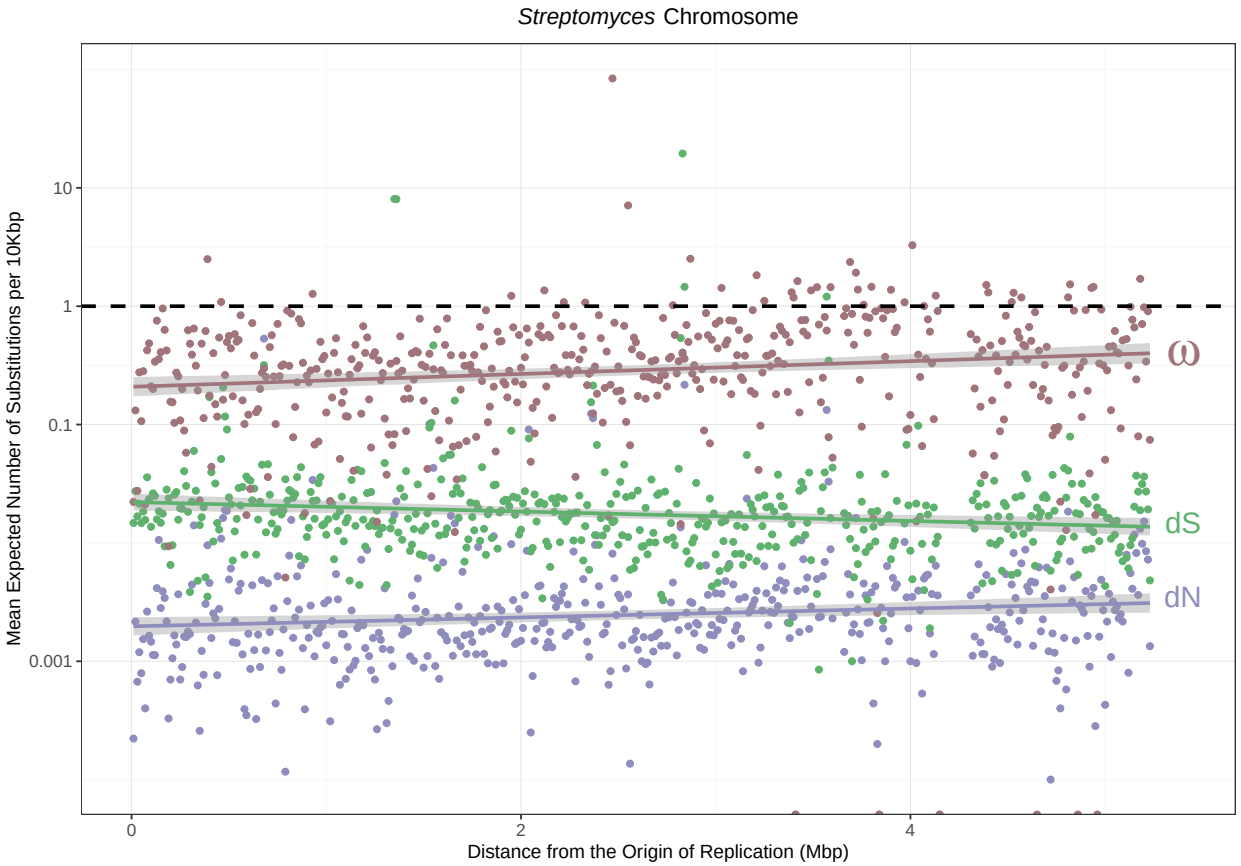
Since we have some theory about how genes are organized on bacterial genomes, we decided to take a closer look at the selection values for *Streptomyces* and see where these genes fall relative to the origin of replication.

```
distg
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 15 rows containing non-finite values (stat_smooth).
```



This graph shows the mean selection value (dN, dS, or *omega*) calculated over each 10,000 base pair (bp) region of the *Streptomyces* genome. We observe again that dS is generally higher than dN and most of the *omega* values are less than 1. However, we see that regions of the genome that have an average *omega* value larger than 1 are mostly concentrated near the terminus of the genome. This is reflected in the trend line which is increasing with increasing distance from the origin of replication. Interestingly, the majority of the core and well conserved portion of the *Streptomyces* genome is located in the first ~2 million base pairs (Mbp) near the origin of replication. The rest of the genome is part of the accessory genome which primarily consists of genes involved in local environmental adaptation and production of antibiotics. It is therefore conceivable that this area of the genome is mostly under positive selection and trying to "hold on" to beneficial mutations.

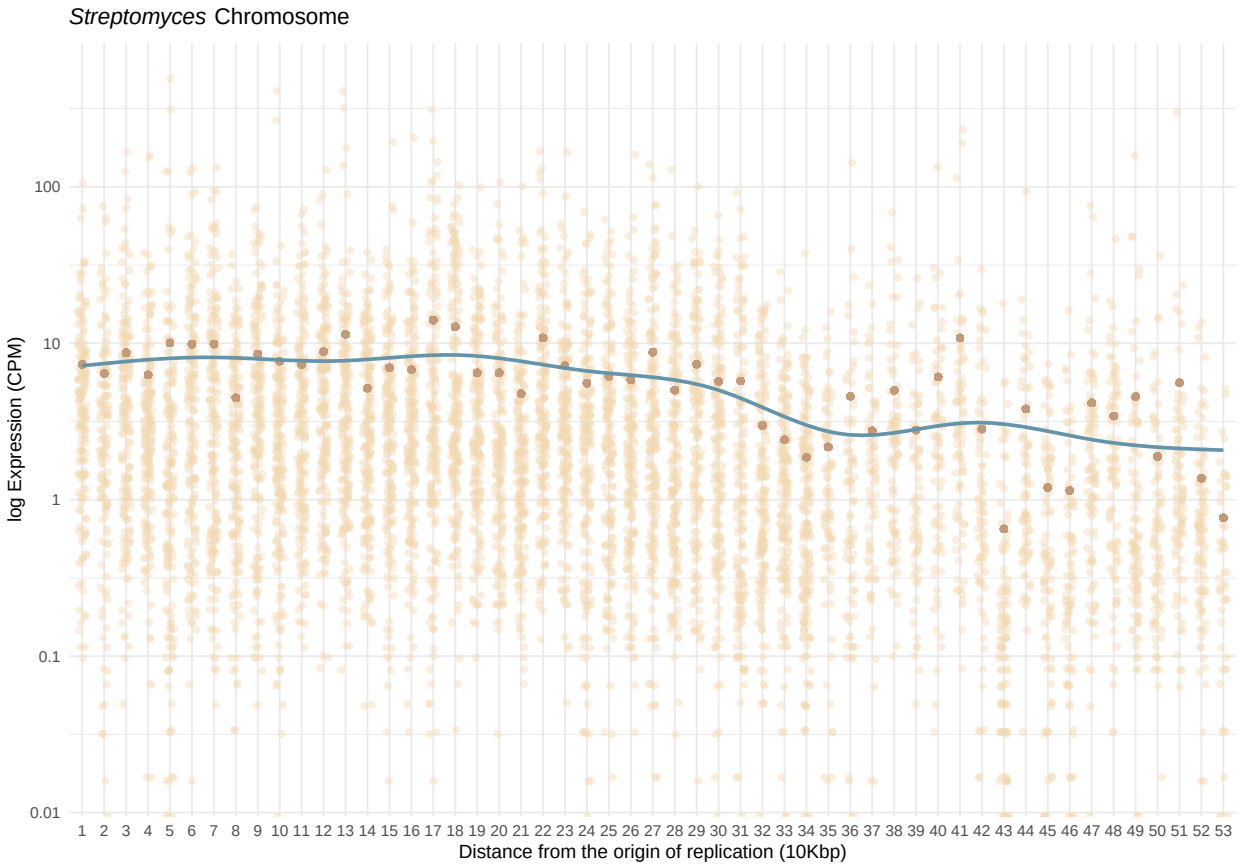
Gene Expression Data

Now we are going to see whether our prediction, that the Gene expression should decease when moving away from the origin of replication, holds. To begin we created the same graph for each of the bacterial genomes to explore how gene expression changes with distance from the origin of replication.

```
g1
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



For most of the exploration graphs we see that gene expression does decrease with increasing distance from the origin of replication for most of the bacteria. However, there are two notable exceptions, pSymB and pSymA. These happen to be the two secondary replicons of *S. meliloti* which in general contain majority of the accessory genome. This could be why these replicons appear to have the opposite gene expression trend from the other bacteria. We will be focusing on *Streptomyces* and analizing the trends we see in that graph in depth.

From the summary of our data set, we saw that the distance from the origin of replication ranges from 219 to 5,247,360. Since we have in total 7,762 total observations, it will be more appropriate to bin our data. We decided to group by each 100,000 base pairs (bp) region of the *Streptomyces* genome, and ended up with 53 bins.

In the plot above, the brown points on the graph shows the mean expression value calculated over each 10,000 base pairs (bp) region of the *Streptomyces* genome.

From the trend line above, we see that in general Expression decreases as we move further away from the origin of replication. This decrease is definitely not linear, we observe some jumps through out the graph. When we are 3,600,000 bp away from the origin, we start to see more of a wave-like pattern.

For *Streptomyces*, the core genome is located within aproximatly 3,000,000 bp from the origin of replication. The accessory genome is generally located within the region approximatly

2,000,000 bp from the terminus of replication. It makes sense that we are seeing expression dipping down at around the 3,000,000 bp point and staying lower for the remainder of the genome. This is where we expect most of the accessory genome to be located.

Graphical Decisions

General:

Since the response variables for both the gene expression and the selection data have a wide range of values, we chose to use a log scale to make it easier to read. When considering genomic distance from the origin of replication we scaled the points by 1 million base pairs to make the values on the x-axis more readable. Additionally, all axis labels are clear and have units where applicable. We ensured that Greek letters and italic bacteria names were used. We utilized trend lines, box-plots, violin plots and reference lines to aid in showing summary statistics and patterns in the data. We also wanted to pick colours that were subjectively pretty, but also dichromat-friendly. Most of our graphs are based around scatter plots, so the colours needed to be fairly saturated to easily identify points/elements of the graph that we wanted to have stand out.

Any graphs that involve the distance from the origin of replication and the response variables, we chose to focus on one bacteria. All the bacterial replicons vary greatly in length from 1Mbp to ~5Mbp. If we had used a facet to show for example, how gene expression changes with distance from the origin of replication in all replicons, some of the replicons would be "squished" on the x-axis and we would be unable to see any of the results. We therefore chose to focus on the most interesting bacteria (*Streptomyces*) for storytelling.

Selection:

With regards to colour, the selection graphs have the same colours for dN, dS, and *omega* in all graphs so that it is easy for the viewer to follow along when switching between graphs. We chose to show the data points as a strip plot with a box plot and violin plot overlayed. This allows for the maximum amount of information about the distribution of the selection values to be shown. We chose to use the facet_wrap() function in R to allow for overarching similarities or differences between the bacterial replicons to be visible.

For the facet selection graph we chose to add in another y-axis to show the values of ω, since the units are different than the units for dN and dS. The arrangement of bacteria in the facet plot was mostly guided by biological relevance. *E. coli* and *B. subtilis* are the "lab rats", and therefore people often care about them the most, so we put them first. *Streptomyces* is similar to *E. coli* and *B. subtilis* because it has it's genome in one chromosome. *S. meliloti* is a multi-repliconic bacteria (has more than one chromosome-like structure), and therefore we wanted to keep the replicons of this bacteria close together so they could be easily compared to one another. In the selection summary graphic we decided to add in a redundant legend to aid viewers in determining what colours were linked to which selection measure. We also used direct labeling when we could to avoid the need for a legend.

Gene Expression:

In our expression plot, we used geom_jitter() and plotted all the observations within each group. We then calculated the mean expression value of each bin and plotted it as a point on top of the observations. We used the same color brown, but in two different shades to represent the observations and the mean value. That is, because both represent the same response variable, the expression value.

We included all of the the observations along with the mean, because taking means alone could ignore some unusual data points that could be of importance to us. We chose the width of our bins and median values in a way that would reduce the overall noise of the graph, but still allow for a signal to be seen. We also used geom_smooth() in our plot, this made it easier for us to see any trends in our plot. For the gene expression graph we wanted to have as much information as possible, without it being too distracting. So, we chose a lighter colour (light brown) for the raw data and a darker colour (dark brown) for the median values for each bin. We additionally chose a bold colour (blue) for the trend line. All these colour components help the viewer focus on the median values, which is what we were trying to highlight, but still have access to all the data points.