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.

When bacterial genomes replicate, they begin replication at the origin and continue in both directions until the terminus (in the case of circular genomes), or until the ends of the chromosome arms (in the case of linear genomes). Our data sets have genomic position relative to the origin of replication. This means that values of 0 are located closest to the origin, and the largest values represent areas near the terminus of replication.

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 ω (dN/dS). Since synonymous substitutions do not cause a change in an amino acid sequence, we expect most genes to have a synonymous rate (dS) larger than non-synonymous rate (dN).

The ω ratio allows us to determine if the genes 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 (or close 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). That being said, most amino acids are under some sort of constraint so it is rare that any genes will have an ω value equal to (or close to) 1. 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).

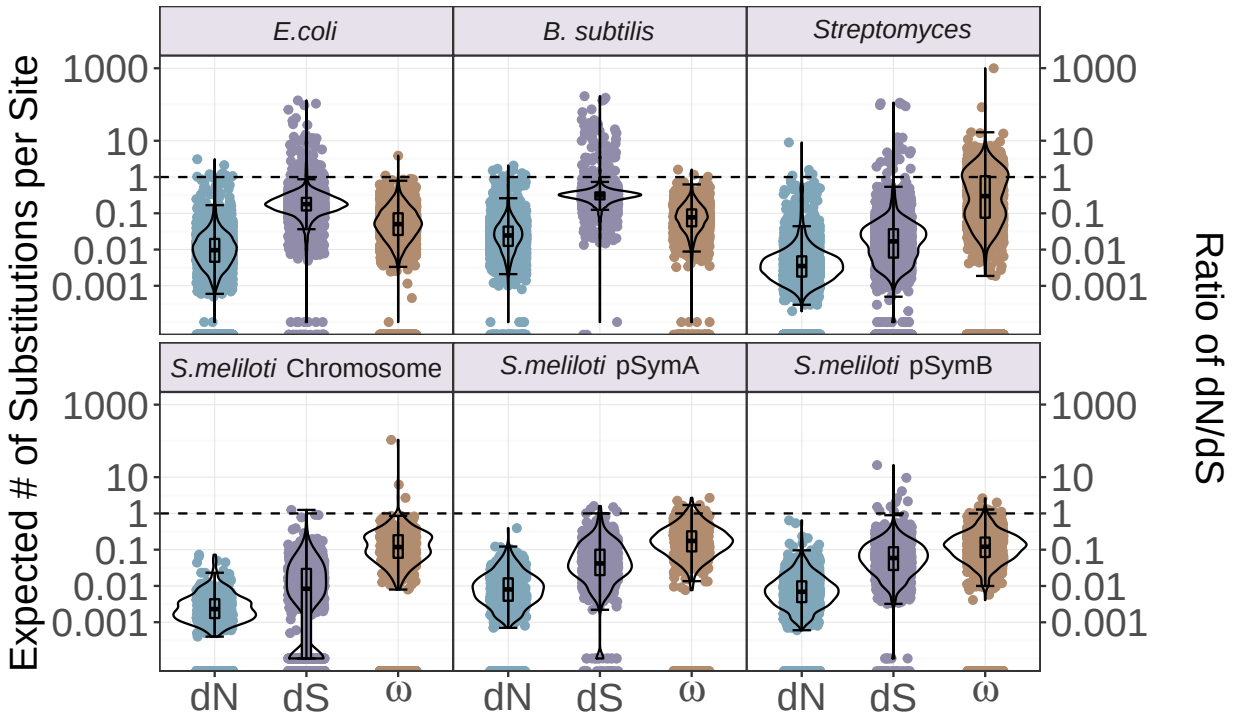
This leaves us with three predictions for our data sets:

1. Gene expression should decrease when moving away from the origin of replication
2. We should have dS > dN for most genes in all bacteria
3. Most genes should be under purifying or neutral 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 for all genes in each of the bacterial replicons. Graphical decisions for this figure are discussed more in the *Graphical Decisions* section.

```
vio_str_box
```



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

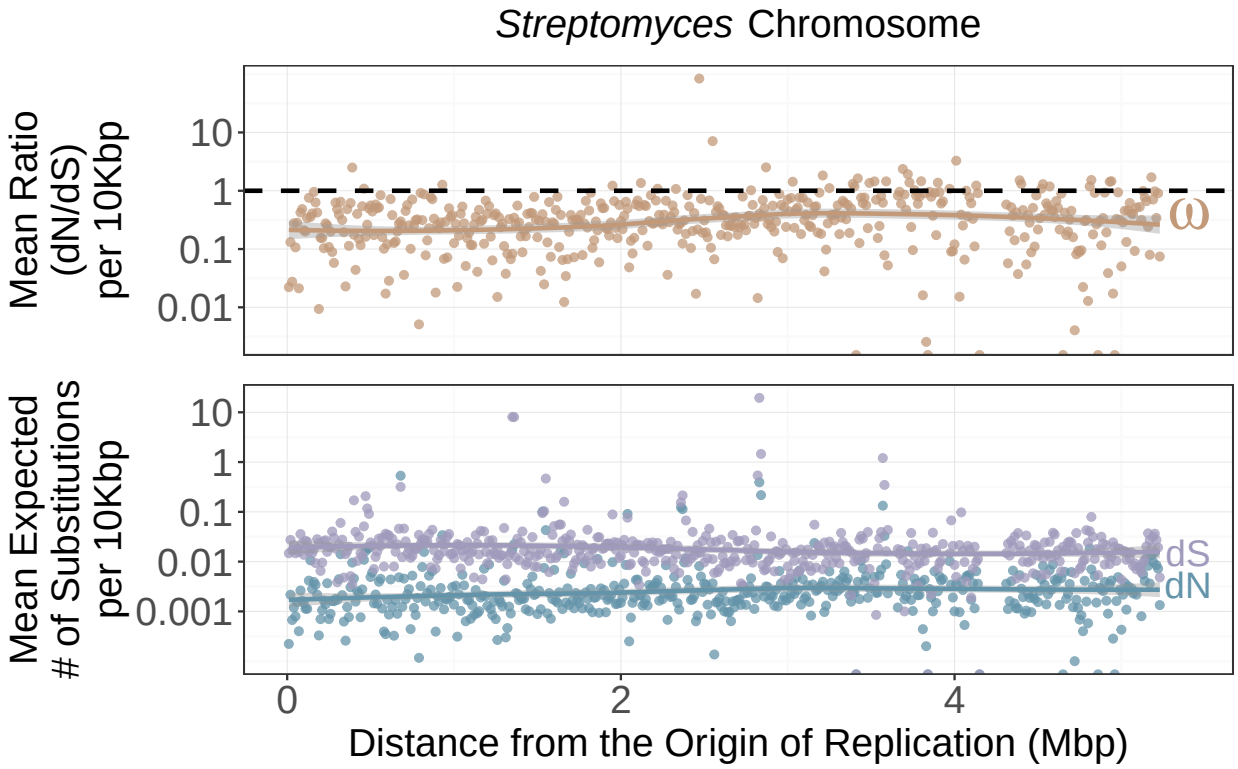
non-functional protein could have catastrophic consequences on the well being of the organism. Across all the bacterial replicons we see that indeed, dS > dN for most genes. In some of the bacterial replicons such as *B. subtilis* and *E. coli*, there appears to be a high number of genes with dS values larger than 1. There is a phylogenetic component to the calculation of dN, dS and ω and the programs are taking an average of all the substitutions that could have occurred along any branch within the phylogentic tree, over the total number of sites in the gene. This means, there depending on how close or distantly related the taxa are, there could have been multiple substitutions at one (or many) sites within the gene. This could cause the value of dS to be estimated to be larger than one substitution per site.

We also notice that most of the ω values for each bacterial replicon are below (or close to) 1, this is what we expected. An ω value below (or close to) one means that the genes are likely neutral or under negative selection, meaning that mutations having deleterious impact on the organism will be removed over time. As mentioned previously, any genes that have an ω value equal to or close to 1 are most likely under neutral selection. It is difficult to tell from the graph which genes have an ω value statistically close enough to 1 to be considered neutral. The program (PAML) used to estimate the values of dN, dS, and ω uses a maximum-likelihood framework to calculate the average value over all sites in the gene. This means that we are quite confident that what we are seeing is the true value because it is impacting all sites over a prolonged period of time. To be more conservative, any ω values that are "close" to one should be considered neutral with caution. The notable exception to this is *Streptomyces*, which appears to have a bi-modal distribution of ω values with a high number of genes with ω 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.

As mentioned previously, genes that are highly conserved and essential to the function of the organism are usually found near the origin of replication and part of the core genome. Genes that are less conserved and non-essential are usually found near the terminus of replication and are used in local environmental adaptation, these are a part of the accessory genome. 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.

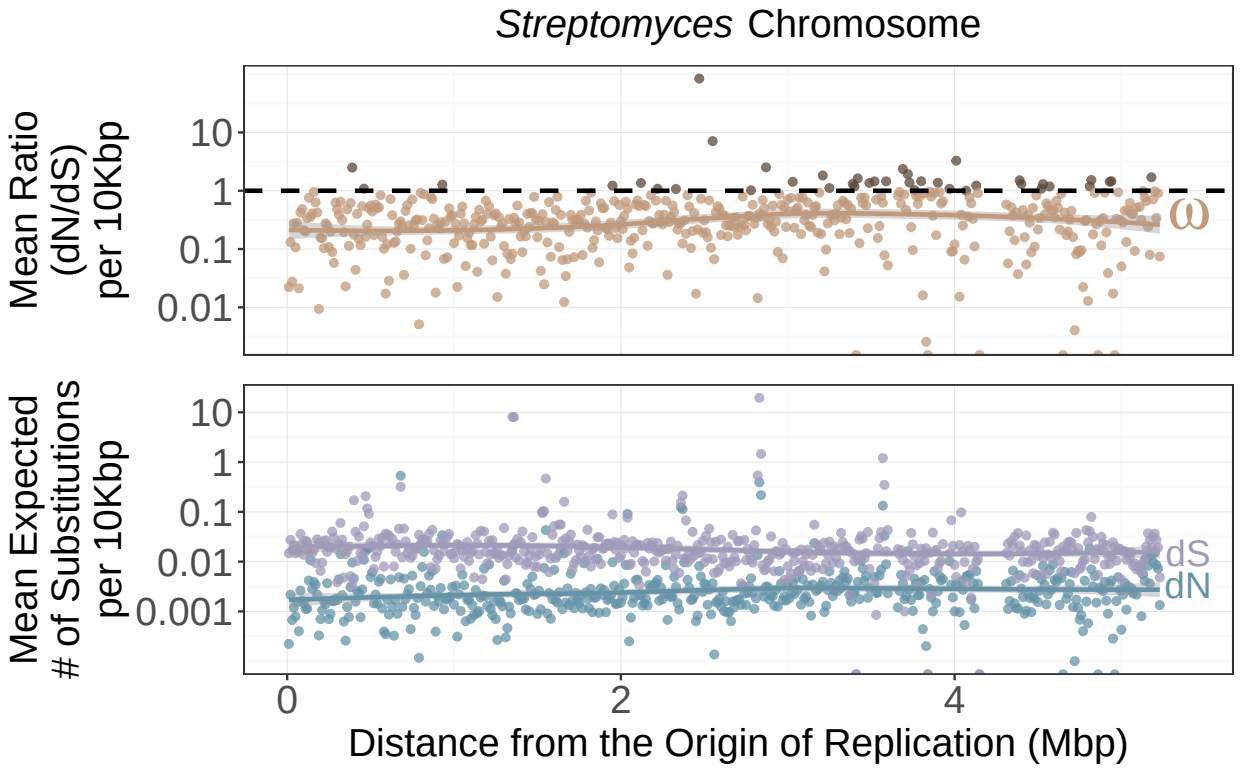
```
# arrange the graphs on one. since facet will not let you re-lable each
# axis in a facet
grid.newpage()
grid.draw(rbind(ggplotGrob(omeg_g), ggplotGrob(rate_g)))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
# arrange the graphs on one. since facet will not let you re-lable each
# axis in a facet
grid.newpage()
grid.draw(rbind(ggplotGrob(omeg_g), ggplotGrob(rate_g)))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The top graph shows the mean ω calculated over each 10Kbp region of the *Streptomyces* genome. This was calculated by averaging the ω values over each 10,000 bp (10Kbp) section of the genome. As mentioned before, this was calculated by averaging the ω values of all genes within each 10Kbp section of the genome. A non-linear trend line with confidence intervals was added to the plot to show an overall trend about how ω changes with distance from the origin of replication. As mentioned previously, we see that most sections of the genome have a mean ω value of less than (or close to) 1, as expected. This means that majority of the genes in the genome are under neutral or negative selection. As mentioned previously, any genes that have an ω value equal to or close to 1 are most likely under neutral selection. It is difficult to tell from the graph which genes have an ω value statistically close enough to 1 to be considered neutral. The program (PAML) used to estimate the values of dN, dS, and ω uses a maximum-likelihood framework to calculate the average value over all sites in the gene. This means that we are quite confident that what we are seeing is the true value because it is impacting all sites over a prolonged period of time. To be more conservative, any ω values that are "close" to one should be considered neutral with caution.

Interestingly, the sections of the genome with mean ω values larger than one seem to be clustered near the terminus of replication. There appears to be a particular peak at around 3.7 million base pairs (Mbp) from the origin of replication. Genes that have an ω value larger than 1 are thought to be under positive selection and conferring some sort of benefit to the organism, and will therefore likely be maintained over time. Interestingly, the majority of the core and well conserved portion of the *Streptomyces* genome is located in the first ~3Mbp 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 mutations that are beneficial to the organism.

The points on the bottom graph represent the mean expected number of substitutions per 10Kbp for both dN and dS. This was calculated by averaging the dS and dN values respectively over each 10Kbp section of the genome. A non-linear trend line with confidence intervals was added to the plot to show an overall trend about how dN and dS respectively are changing with distance from the origin of replication. As mentioned previously, we see that dS is higher than dN for most of the genes in *Streptomyces*. dS appears to be slightly decreasing with increased distance from the origin of replication, while dN appears to be slightly increasing. We also see that the pattern for dN and dS are non-linear, although only slightly. Biologically, we know that the majority of the core genome for *Streptomyces* (the part of the genome containing functionally important genes), is located about 3Mbp from the origin of replication, while the accessory genome is located approximately 2Mbp from the terminus or replication. It is therefore plausible that core genes should have more synonymous substitutions, so the amino acid sequence is not altered, compared to the accessory genome. It is conceivable that the accessory genome has more non-synonymous substitutions which could provide increased genetic diversity, assisting in for example producing new antibiotics. With the location of the accessory genome coinciding with areas where we see decreased dS and increased dN, we could infer that biologically this is what is happening.

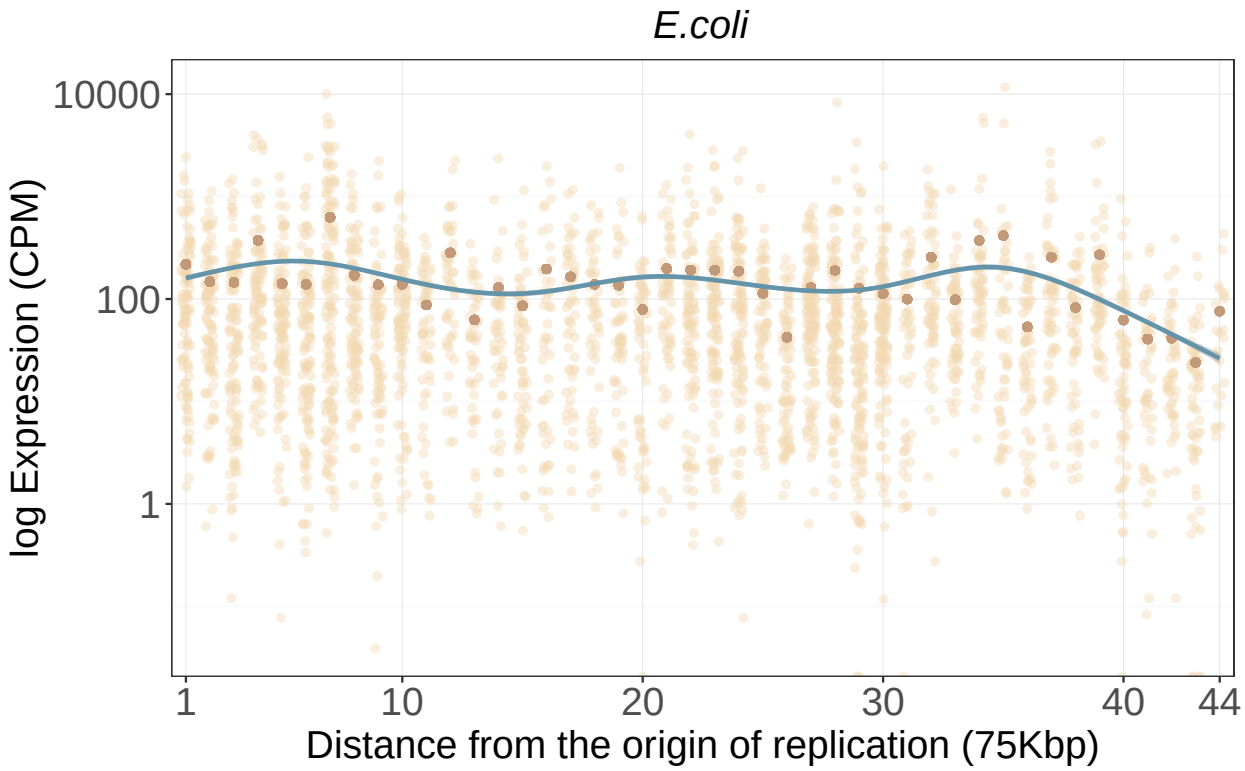
Discussion of our graphical choices for this graph can be found in the *Graphical Decisions* section.

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. We created the same graph for each of the bacterial genomes to explore how gene expression changes with distance from the origin of replication.

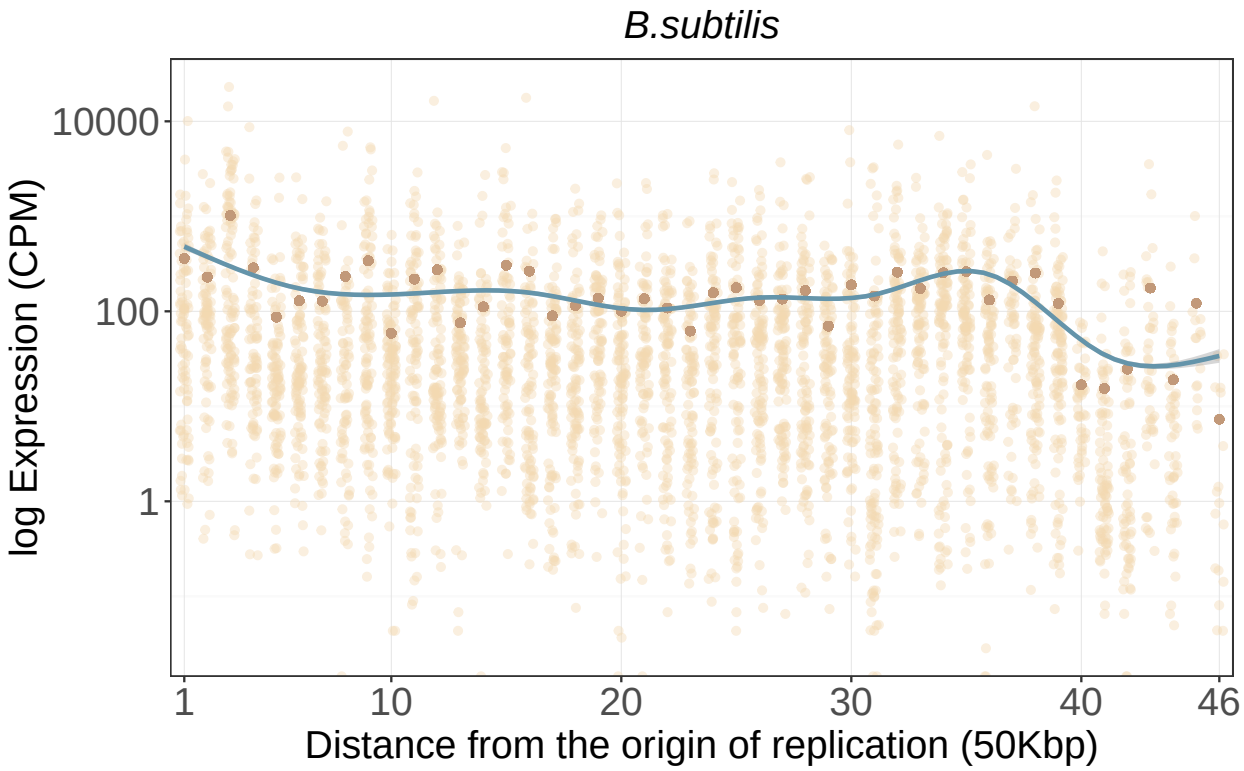
```
g3
```

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




```
g4
```

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

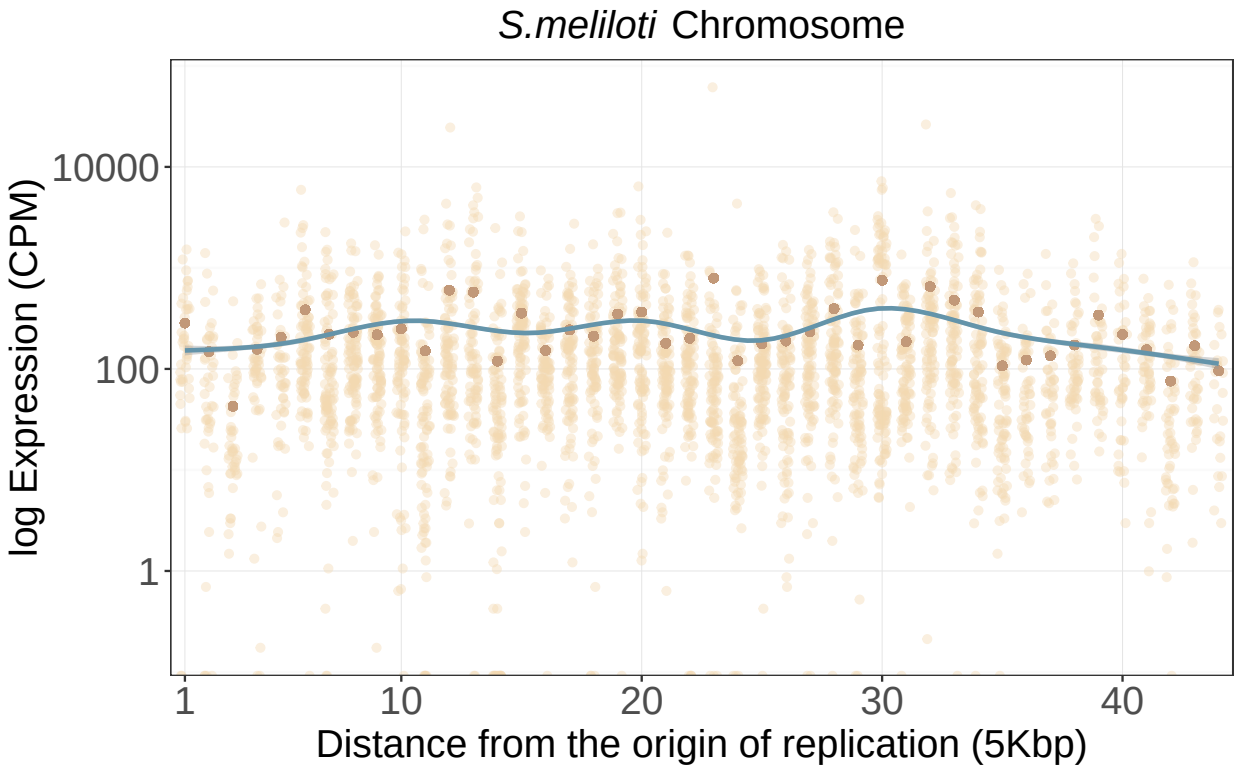



```
g1
```

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

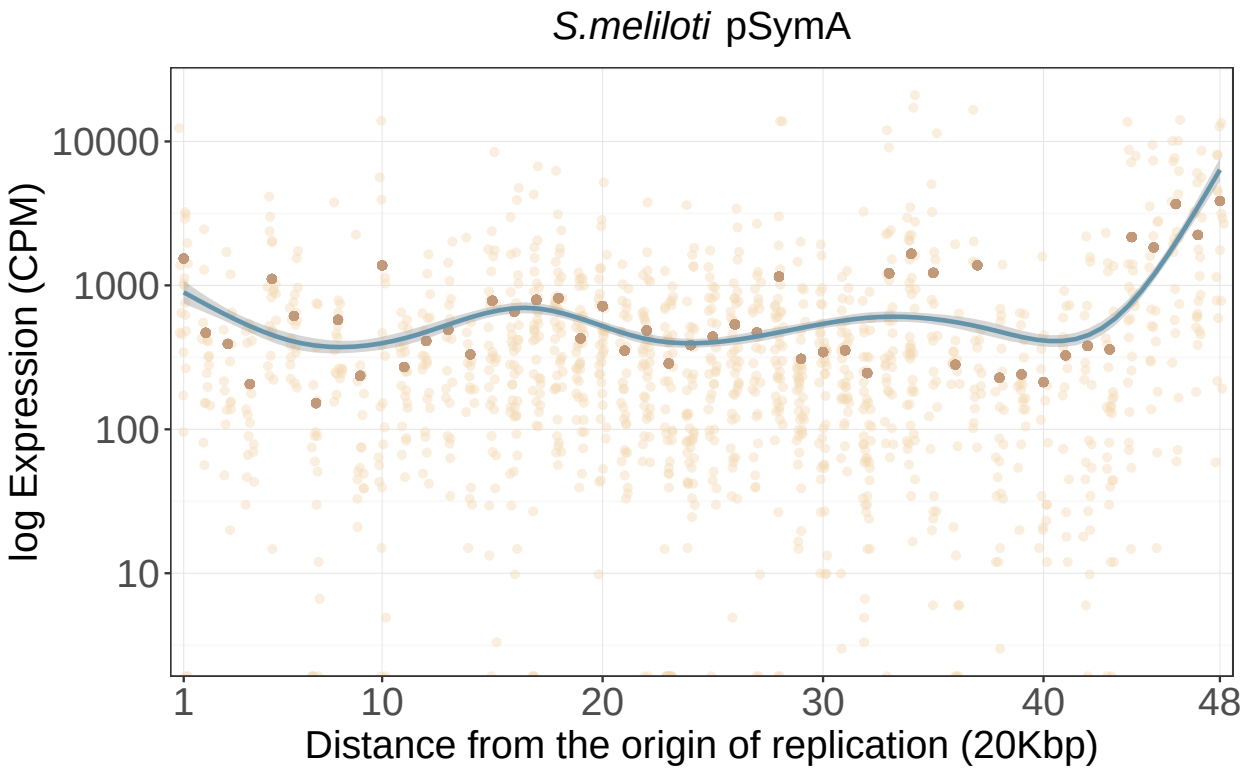
```
g2
```

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



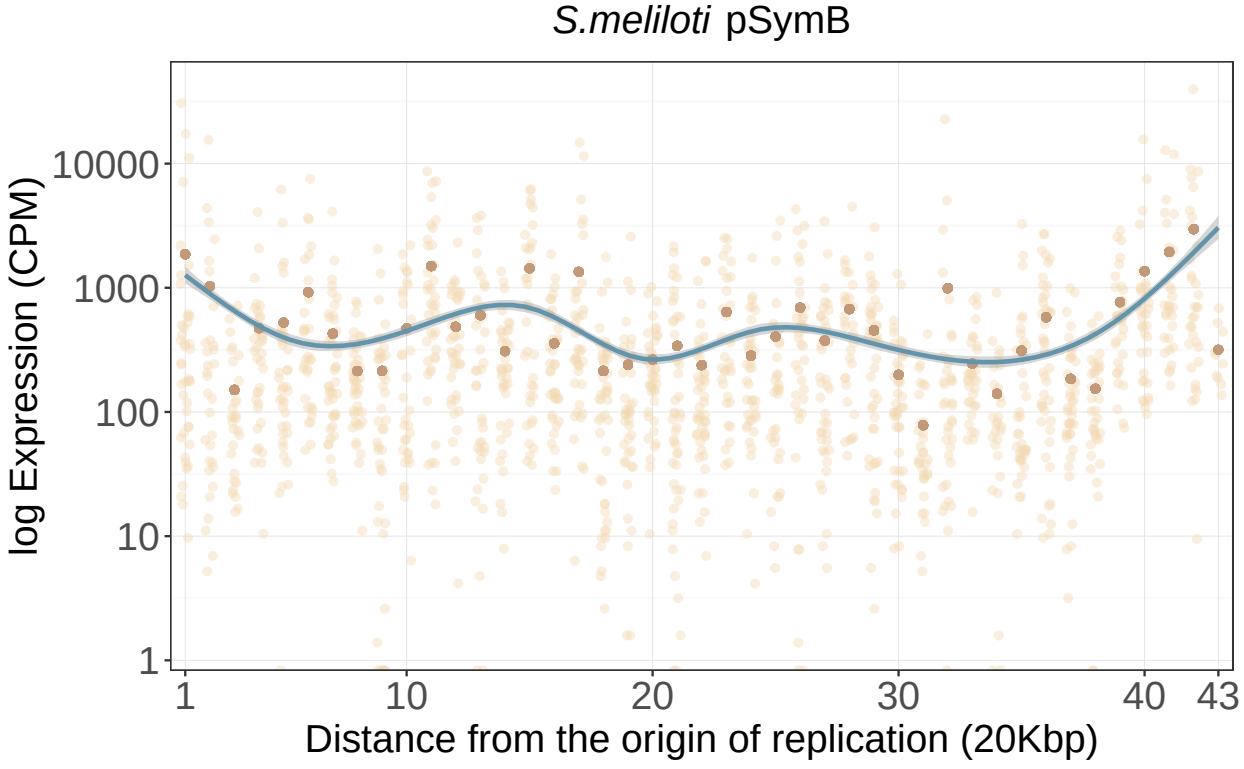

```
g5
```

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




```
g6
```

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



For the expression plots above,we took every c Kbp region of the genome and grouped them into a bin. The c differed for each bacteria, as each one of the data sets contains different number of observations and different range of values for the distance from the origin of replication. The dark brown points on the graph shows the mean expression value calculated over cKbp section of the genome. We included all the raw data along with the mean and plotted it in a lighter color. Each plot contained a non-linear trend line with confidence intervals to show an overall trend about how expression value changes with distance from the origin of replication. In *Streptomyces*, *B. subtilis*, and *E. coli*, we see that expression decreases as we move away from the origin, just like what we predicted. 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. As for the chromosome, we see that expression neither increases nor decreases as we move away from the origin of replication.

For *Streptomyces* , the core genome is located within approximately 3,000,000 bp from the origin of replication. The accessory genome is generally located within the region approximately 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. We also see that the the expression decreases in a non-linear way, then we start to see more of a wave-like pattern at the terminus.

For *B. subtilis* , as we move away from the origin, we barely see any significant changes in the expression values. Until we are around 35 (50Kbp) away from the origin, that is when expression values starts to decrease. Which is expected, as that is where we expect most of our accessory genomes to be located.

For *E. coli* , we see more of a wave-like pattern, until we are near the terminus, that is when we start to see a decrease in expression values. Again, this is expected, because most of accessory genomes are located near the terminus.

For pSymB and pSymA, we see the same trend in both. We see a wave-like pattern through out the plot until we are near the terminus, that is when we start to see an increase in the expression values.

Gene expression of *Streptomyces*, *B. subtilis*, and*E. coli* decreases with increasing distance from the origin of replication in a **non-linear** manner.

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. Additionally, all axis labels are clear, in a large font size, and have units where applicable. Trailing zeros were removed from the axis labels to again make it more readable. 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 and that went well together so all the graphs looked cohesive. 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. We chose to apply transparency and "jittering" (where applicable) to ensure that overlapping points were identifiable and to maximize the amount of points shown. As we are dealing with multiple bacterial replicons in our graphs we decided to have titles on all graphs where we are showing values for just one replicon. This is to avoid confusion about which bacterial replicon viewers are looking at.

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 create multiple graphs for each bacterial replicon, or focus on only one particularly interesting replicon.

Selection:

General:

With regards to colour, the selection graphs have the same colours for dN, dS, and ω in all graphs so that it is easy for the viewer to follow along when switching between graphs. 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.

When choosing colours for this graph, we wanted colours that also matched with the gene expression data set, but also worked well for these different types of graphs. Since ω provides important information about the selective pressures acting on a gene, we wanted to choose a colour that was different from the dN and dS colours so it could stand out in the facet graph. We also wanted dN and dS to be similar colours (blue and purple) because they have the same units and both represent substitution rates.

Although most of the scales are log base 10, we decided to retain zero values because these provide valuable biological information. Genes with $dS = 0$ means that there are no synonymous substitutions which greatly influences our predictions on what evolutionary trajectories and past events are acting on that gene. The same could be said for dN values equal to 0, or ω values equal to 0.

Facet Graph:

For this graph we wanted to be able to compare the dS, dN and ω values within each bacterial replicon, and across all bacterial replicons, so we decided on a facet graph. This allows us to see differences between dN, dS and ω within each bacterial replicon while also highlighting overarching similarities or differences between the bacterial replicons. We decided to facet

based on bacteria because primarily we wanted to show differences within bacterial replicons rather than between all bacterial replicons. 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.

As mentioned previously, we are interested in ω values that are larger than zero (see intro on positive selection). We decided to add a reference line at 1 to help remind viewers about the differences between ω values that lie above and below this line.

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 chose to arrange the selection parameters (dN, dS and ω) in that particular order because ω is the ratio of dN/dS, we thought that it would be appropriate to put dN first in an attempt to match the ratio order. Likewise, ω was chosen as the right most value because it is a combination of the first two values. We removed the legend as we thought it to be redundant. We also removed any space between the facets to ensure there was less necessary white space.

We experimented with many many versions of this graph, prominent versions can be found in the *Appendix* section. Although ω is "unit less", dN and dS are not. It was therefore necessary to have multiple labels for both sets of values. We experimented with having facets for both the bacteria and selection variables (see Figures A1 and A2 in *Appendix*). These graphs allow for the scales to be "tighter" around the values for the rates (dN and dS) and ω, since both facets are allowed to have varying scales. These graphs also make it clear which units correspond to which values, which could be confusing in the figure presented in the main part of this write up. Since there are more sections to this graph, we decided to separate the bacteria into two groups so that the bacterial name and replicon would be readable without having to use abbreviations. However, we believe that the version of this figure found in the main part of the write up represents the best figure for biologists. Since the selection values (dN, dS and ω) are very similar and based on one another, we wanted to have them all in one graph so we could compare their values to one another. When they are separated as in Figures A1 and A2, it is difficult to see how ω relates to dN and dS. Additionally, we wanted to have all bacterial replicons in the same figure so we could look at any similarities or differences between the bacteria. As mentioned previously, this would not be possible with the supplemental graphs (A1 and A2) because we would be sacrificing readability. The choice for faceting by bacteria or by selection value is difficult because most biologists will want to look at what is happening within each bacterial replicon first, and then look at over arching trends between bacterial replicons. This again is difficult to determine with Figures A1 and A2. Therefore, we believe that the figure presented in the main part of the document provides a compact, accurate, and preferable depiction of the selection values.

Example Selection Values Graph:

Since the range of values and units differ between the dN, dS and ω values, we decided to "facet" the values to allow for two scales. The top graph shows the mean ω values over 10Kbp distances from the origin of replication, and the bottom graph depicts the mean dN and dS values over 10Kbp distances from the origin of replication. Separating the data into two graphs allows for a clear separation of units and allows for a "zoomed in" picture of the values. Since the ω values are so much higher than the dN or dS values, having all selection parameters (dN, dS and ω) on the same graph would obscure any subtle changes in the parameters with respect to distance from the origin of replication. By using separate graphs, we are able to see these changes more clearly.

We chose to put ω on the top part of the figure and the rates (dS and dN) on the bottom part. This is partially because we are interested primarily in the distribution of ω as distance from the origin of replication increases, and partially because the values of ω are generally larger than the values for the rates. It therefore is logical to have ω at a higher "height" in the graphic than dN and dS.

We chose to include a non-linear trend line to help show what overall patterns are happening as the selection parameters change with distance from the origin of replication. This allows us to see peaks and valleys that we would not necessarily see with a simple linear model. We allowed for confidence intervals on the line to help show the fit of the geom_smooth() line.

Again, we are primarily interested in ω values that are larger than zero (see *The Data* section on positive selection). We decided to add a reference line at 1 to help remind viewers about the differences between ω values that lie above and below this line.

We also used direct labeling when we could to avoid the need for a legend. Although the y-axis titles are lengthy, they are biologically accurate and convey the proper units for each of the selection values. We chose to put them on separate lines so they are as readable as possible.

For this graphic we also had many iterations (see *Appendix* Figure A3 for the most prominent one). In general, humans are better at determining the slope of a line when it is near a 45 degree angle. By changing the arrangement of these figures on the graphic, we can adjust the aspect ratio so the trend lines appear to be at an angle closer to 45 degrees. This can be seen in Figure A3 in the *Appendix*. Indeed, this figure does accentuate the trend lines and it is easier to see the peaks and valleys. However, from a genomic perspective it is difficult to see any patterns. Someone that is interested in *Streptomyces* would have issues determining which portions of the genome (relative to the origin of replication) have certain values for dN, dS and ω. Additionally, it is difficult to establish which portions of the genome (relative to the origin of replication) fall within a peak or valley of high or low selection values. We believe that this supplemental graph A3 does not capture this information. In the graph presented in this main portion of this document, it is clear which portions of the genome have high or low selection values. It is particularly easy to determine the locations of the core (~3Mbp from the origin of replication) and accessory genome (~ the last 2Mbp of the graph) in *Streptomyces*, which carry important predication about how selection might be

acting upon those regions of the genome. Showing the graphs in a horizontal manner gives a more in depth picture about the distribution of the selection values across the genome (relative to the origin of replication). A vertical layout (Figure A3) of these graphs loses information about the genomic location of the selection values. We therefore believe that the graph presented in the main text is the best option for conveying the most amount of information and staying biologically relevant.

Gene Expression:

We decided to include all six different gene expression plots, as we see different trends for each bacteria. We did not go with facets, because the range of values of the distance from the origin of replication differs for each bacteria. But the order in which we decided to show the plots, follows the same order we did for the facets in the selection plots (for the same reasoning too).

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 wanted to have as much information as possible, without it being too distracting. So, we chose a lighter colour and included all of the the raw data.

For each expression plot, we considered more than one possible size for the width of our bins. We decided to go with the ones 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 plots, this made it easier for us to see what overall patterns are happening as we move away from the origin of replication. We assessed more than one different span value, but did not see much of a difference in different span values, so we abided by the value 0.5. We allowed for confidence intervals, but we do not see much of it for most of the bacteria, except for pSymA and pSymB.

We chose a lighter colour (light brown) for the raw data and a darker colour (dark brown) for the mean 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 mean values, which is what we were trying to highlight, but still have access to all the data points.

Appendix:

Figure A1:

```
# arrange the graphs for non-multi rep bac
grid.newpage()
grid.draw(rbind(ggplotGrob(single_rates_summary), ggplotGrob(single_omega_summary)))
```

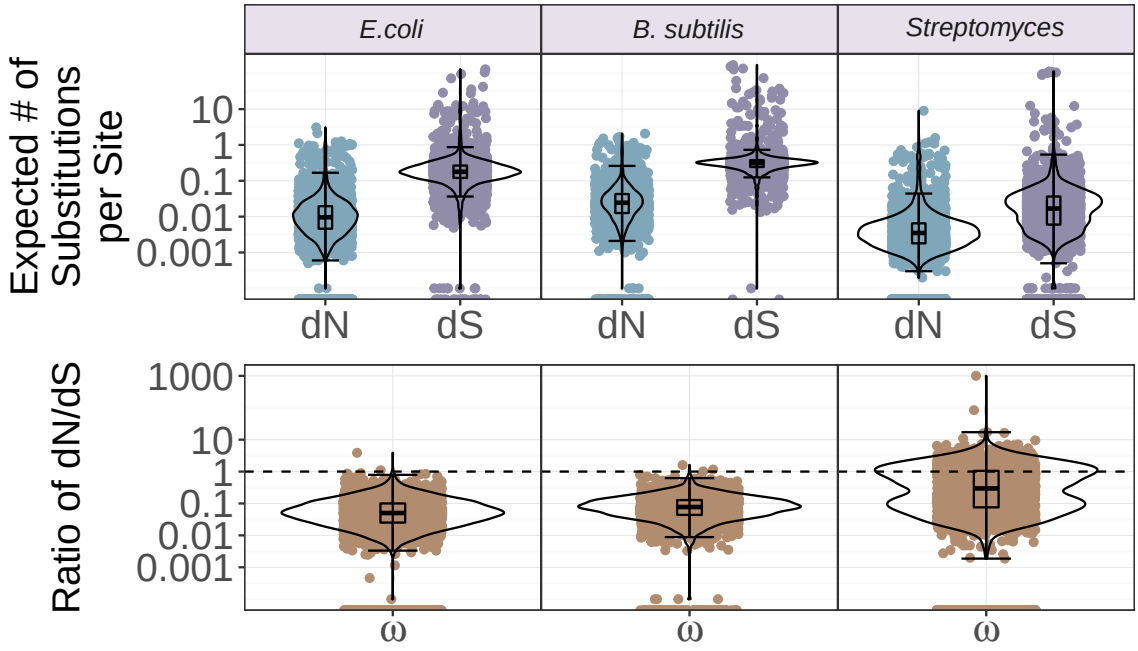


Figure A2:

```
# arrange the graphs for non-multi rep bac
grid.newpage()
grid.draw(rbind(ggplotGrob(multi_rates_summary), ggplotGrob(multi_omega_summary)))
```

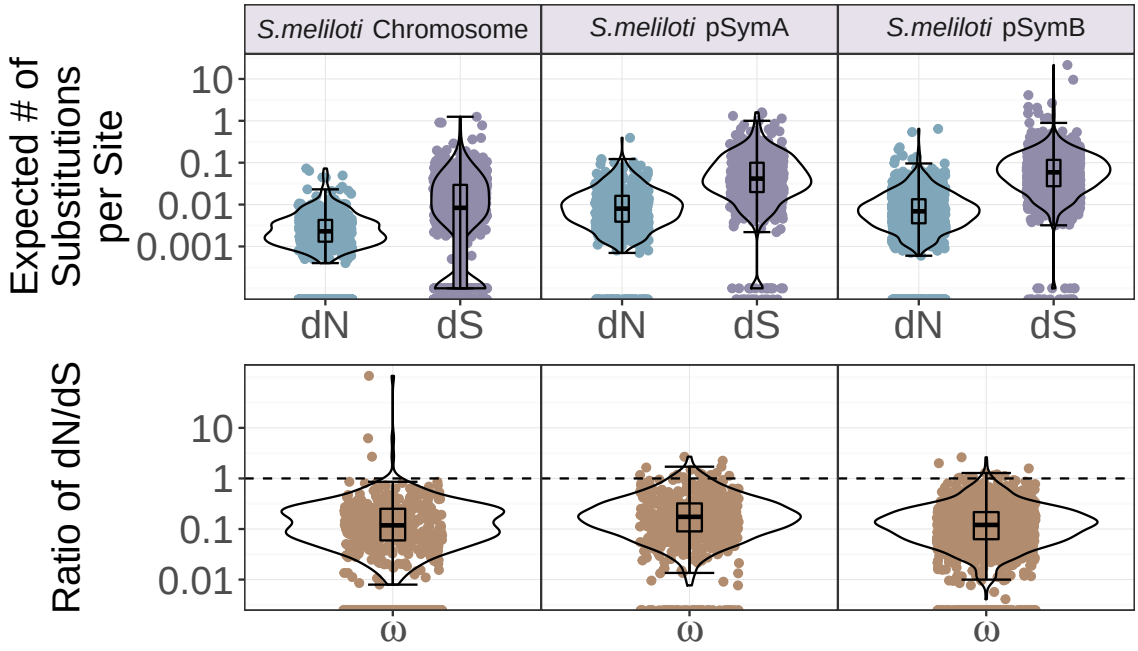
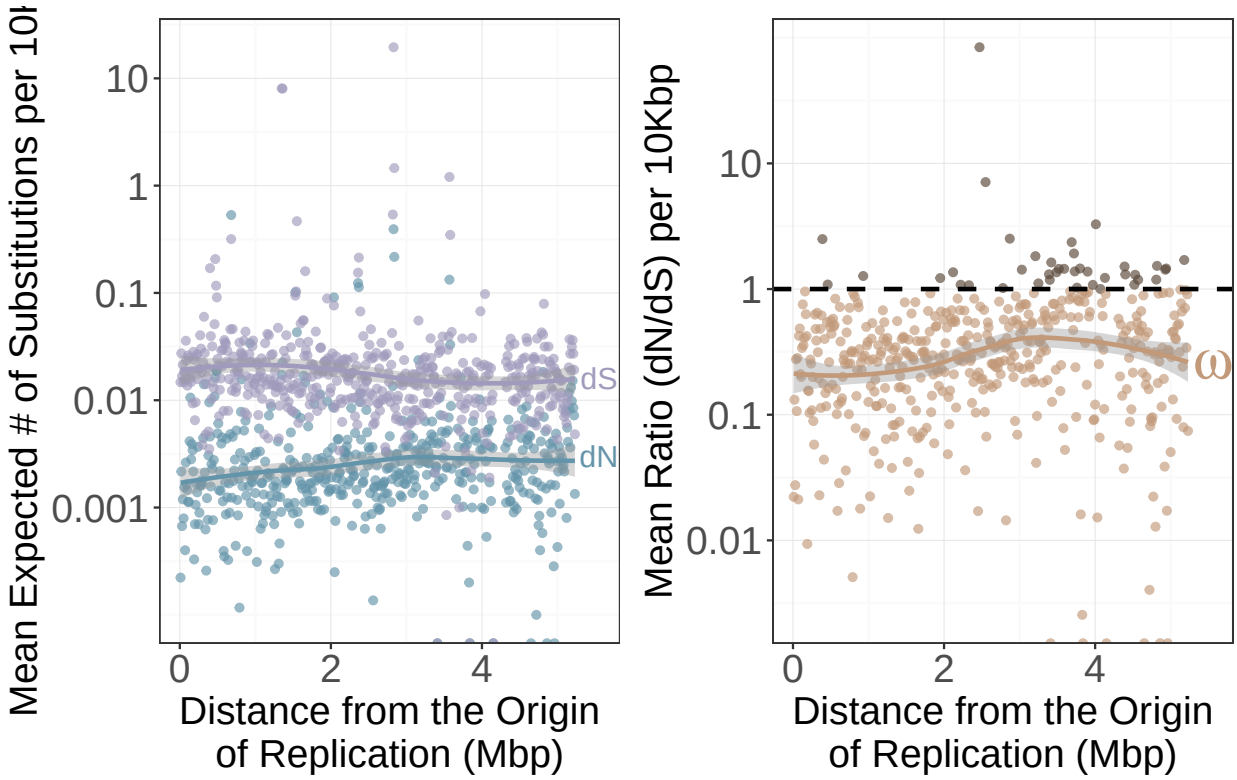


Figure A3:

```
# arrange the graphs on one. since facet will not let you re-lable each
# axis in a facet
grid.newpage()
grid.draw(cbind(ggplotGrob(rate_g2), ggplotGrob(omeg_g2)))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


Mean Expected # of Substitutions per 10
10
1
0.1
0.01
0.001
dS
dN
0
2
4
Distance from the Origin
of Replication (Mbp)
Mean Ratio (dN/dS) per 10Kbp
10
1
0.1
0.01
ω
0
2
4
Distance from the Origin
of Replication (Mbp)