Ordination methods

PCA and RDA

***Questions:***

1. Is the temperature and important determinate of strain fitness in our mesocosms? How much variation in your data is explained by the treatments?

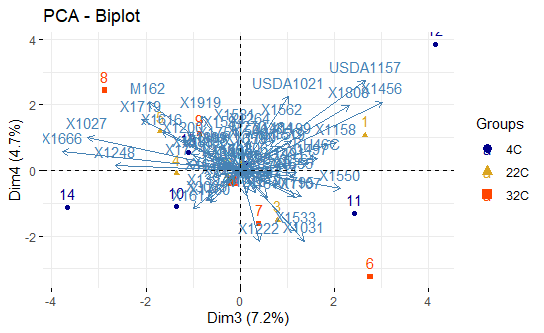
Yes. By the RDA r squared, temperature treatment explains 66% of the variation in the data sets.

1. Which strains are associated with success in 32C? What about 4C?

It’s a little hard to see. But the PCA bi plot, it seems strains 1623, 4819, 1211 have higher abundance at 4c. By filtering the data for 4c and the strains that have positive values, I was able to check that KH35C, X1819, HM006.1, X1790, X1724, X1623, all increase at 4c. By the same method, at 32C, X1724, X1919, X1200, X1215, X1212, X1248, X3085, X1264, 1660, 1170 all increase.

1. How similar are the results from the “exploratory” PCA and the “explanatory” RDA analysis? In what circumstances would we expect them to be more divergent?

The results are similar for us because the treatment does explain most of the variation. Therefore when the RDA fits the data by the treatment we give it first, then does a PCA, not a lot of information is lost, because they explains a lot of it R2 is 66%. The PCA is exploratory because it doesn’t constrain the axis that it fits the data to as a treatment, it just picks the axis that works best. If treatment did not well explain the variation in the data, then an RDA and PCA would be quite different.

1. In the tutorial, I only visualized the first two major axes of variation. How much variation is explained by the third and fourth axes in the PCA? Can you figure out how to plot them?

The 3rd and 4th axes explain 7% and 4.7 percent of the variation in the data set.

1. Which temperature environment is most selective? One way to quantify this is to use the functions range, sd, and skewness to summarise the distribution of strain fitness in each treatment. Can you write some additional code by modifying the summarize function to quantify this? If you want to formally test for statistical differences you can calculate these metrics on every sample and use an ANOVA to test for differences between treatments.

fit%>%group\_by(Trt)%>%summarise("min"=min(fitness), "max"=max(fitness), "sd"=sd(fitness), "mean"=mean(fitness), "median"=median(fitness))%>%

mutate("range"= max-min, "skew"= mean-median)

* 1. **Range** is perhaps the simplest measure of the strength of selection. It is the difference between the fitness of the best and worst strain. The bigger the difference the stronger selection.

4c has the highest range, closely followed by 32C.

* 1. **Standard Deviation** measures the square root of the variance. The larger the variance around the mean the stronger selection.

4c also had the highest standard deviation.

* 1. **Skewness** is the degree to which fitness is asymmetric around the mean (remember in a true Gaussian distribution the mean and the median are the same). Positive (or right) skewness indicates that more strains have fitness lower than the mean then in a true Gaussian distribution (the median is lower than the mean). This means there is selection for a few awesome strains. In contrast if the skew is positive, this means that there are lots of strains with slightly positive fitness and a few strains that basically disappear in the mesocosms entirely.

4c likewise had the highest positive (right) skew (the median is lower than the mean) which shows there are a lot of “losers” strains who’s abundance decreases, and few ‘winners’ or strains that increase a lot in abundance.

1. Write some code to identify “generalist” strains that do well in all temperatures and “specialist” strains that do well in only one.

generalists<-fit%>%filter(fitness>0)%>%ungroup() %>%

group\_by(strain) %>%

filter(n() == 3)

1248, 1562, 1655, 1660, 1724, 1795 are all generalists which increase in all conditions.

specialists<-fit%>%filter(fitness>2)%>%ungroup() %>%

group\_by(strain) %>%

filter(n() == 1)

1724, 1659 and 1562 are all 4C specialists.

1. How do these results differ from the inferences we would make from ordination methods that do not require a normally distributed outcome. What would happen if we used NMDS or PCoA to analyze the raw frequencies?

NMDS and PcoA don’t require normal distributions. They don’t try fit the data to a normal distrubtion.

***Additional Data Challenge Problems:***

1. How does strain fitness change over time within a treatment treatment?
2. Is Time or Temperature a more important driver of strain fitness in our mesocosms?