[1]

Mice Behavior (FY1821): Project Report 2

Cynthia, Josh, Kewei, Srinidhi, Zach

With Prof. Andres Bendesky & Data Science Institute


[2]

Contents

Introduction

Our goal is to explore the connections between mice behavior and genetic markers. For this report, we extend the work done previously by creating the features proposed in the first report, statistically evaluating the relevance of all features in distinguishing the mice across two factors: species—peromyscus polionotus (PO), peromyscus maniculatus (BW)— & sex, and creating plots to show the influence of different sections of the mice's genome on the relevant features.

Features

The first part of our project is to create features representing mouse behaviors. We were provided with a script that reads in location data of the mice over time and calculates some basic features. We expanded on that script to create a total of 210 features. The features are split across different categories of observable behaviors and attributes (see Figure 1 for an illustrated explanation of the categories):

- **Mouse Length:** The length of the mouse.
- **Time Spent in Arms:** This is the fraction of time the animal spends in different arms of the maze. The *closed* arms have high walls around them, while the *open* arms have no walls. We subdivide this into individual arm fractions as well as collections of arms.
- **Time Spent in Safety:** The amount of time spent at the outermost-section of the closed arms. We use the length of the mouse to determine how far from the edge of the arm is *safety*.
- **Time Spent at Rest:** The amount of time spent not moving. We consider the mouse not moving when it's velocity is at a certain threshold for a minimum number of frames.
- **Peeking Behavior:** A peeking behavior is defined as being either in the middle section or close to the middle section within a closed arm. We use the length of the mouse to determine the distance from the middle that we consider peeking. We calculated the average length of all peeks and the total peek count.
- **Velocity:** the speed of the animal in different scenarios. We split this into several features (by individual arms, different directions, only when active, etc.)
- **Turning Preference:** The kinds of turns the animal makes between arms. Specifically, the animal transitions from one arm to the middle and into another arm. We again split this into many subfeatures (Ex. turning right, going from a closed arm to an open one via a right turn, etc.)

The complete list of generated features is available on our github repo.

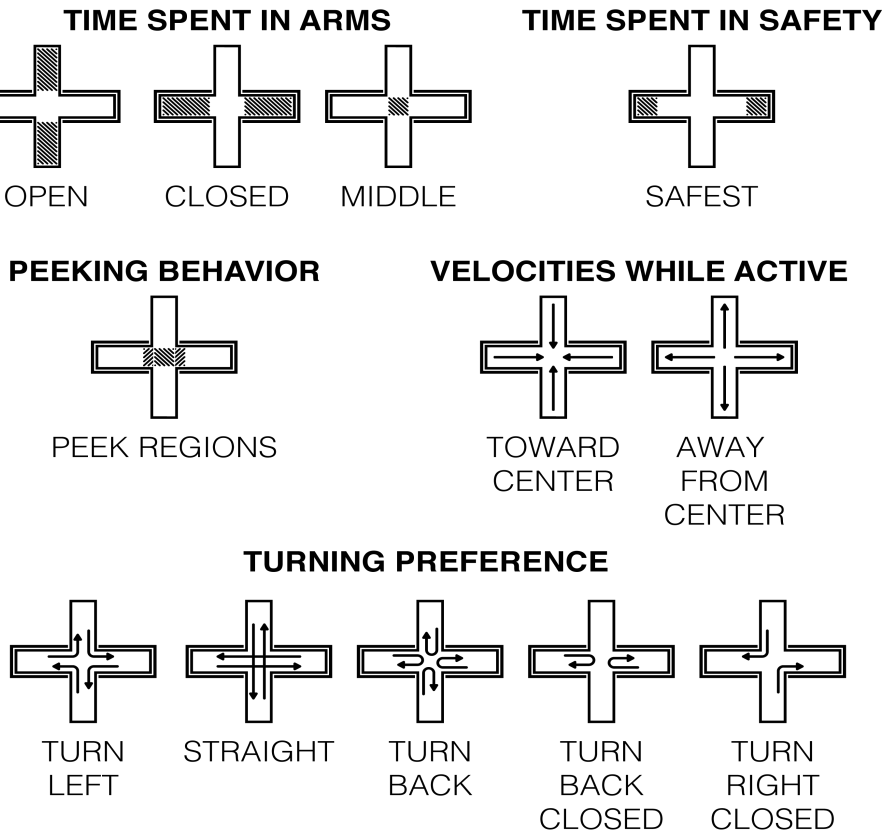

Figure 1: Graphical Explanation of selected feature categories

Data Visualization

As before, we want to explore behaviors that are demonstrably different between the two species and/or the sex of mice. In the first report, we used violin plots to look at the difference in behavior of features that were already provided to us Here we include plots of two newly added features, as examples. In Figure 2 we can see that the fraction of straight turns (a straight turn is defined as leaving one arm and then entering the arm directly across from it) appears to differ between species of mouse. PO mice have a lower mean and are more condensed, while BW mice are more spread out with a higher mean. Likewise, Figure 3 shows that PO mice tend to be smaller than BW mice.

Note: Some of our violin plots have clear outliers. We investigated significant outliers and removed them only if they were the result of an error in the tracking software. If they are genuine, we leave them in. For example, in Figure 3 there was one mouse with a length of 0. We found that this was an error in the tracking software and removed this mouse.

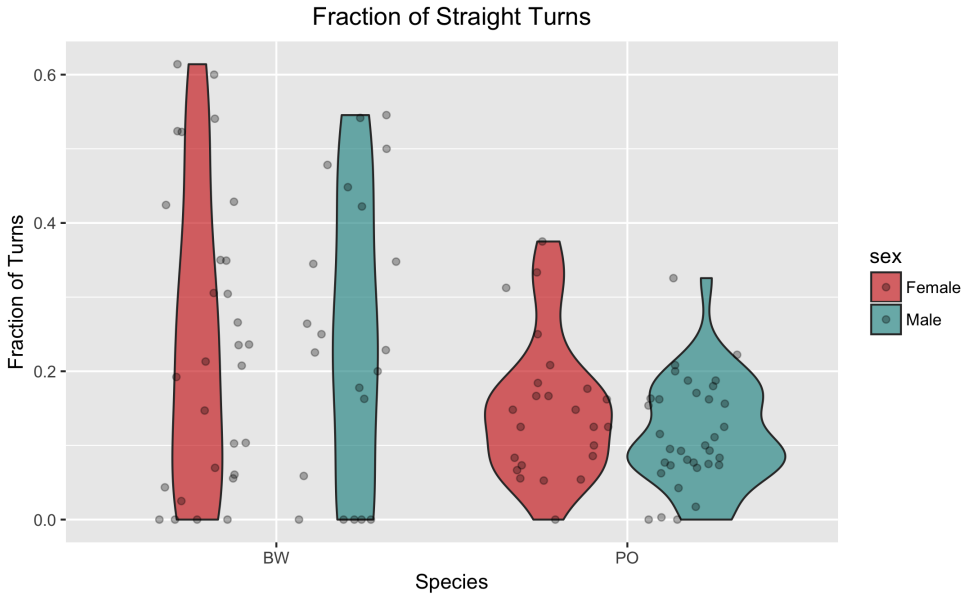

Figure 2: Distribution of the fraction of straight turns out of all turns for each mouse by gender and species

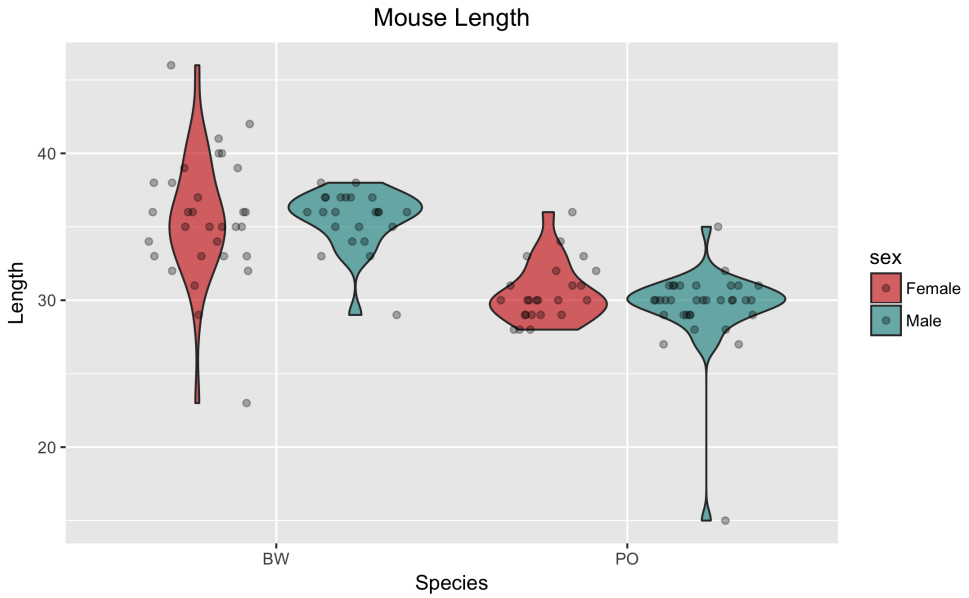

Figure 3: Distribution of mice length by gender and species

Feature Significance

At this point, we have many features across seven attribute groups that may differ between the two mouse species and/or sexes of the mice. We want to run statistical tests to find all of the differences that are significant. To accomplish this, we run a two-way ANOVA test. This will tell us which features are statistically significant across dimensions: Species, Sex, and Species*Sex interaction. The process and results are discussed below.

To check for the underlying required conditions for ANOVA, we perform a Shapiro-Wilk test (for Normality) and a Levene's test (for Homogeneity of Variance) for each of the four species-sex combination of each feature. We also generate Q-Q plots to examine the data distribution. If we find the features to be skewed, we apply appropriate transformations (log, square etc.) to force normality. Most of the features passed the Levene's test.

Few features clearly had skewed data (all right tail heavy). Figure 4 shows a peeking feature with a heavy right tail. After a log transformation the feature satisfied the normality conditions for the ANOVA test.

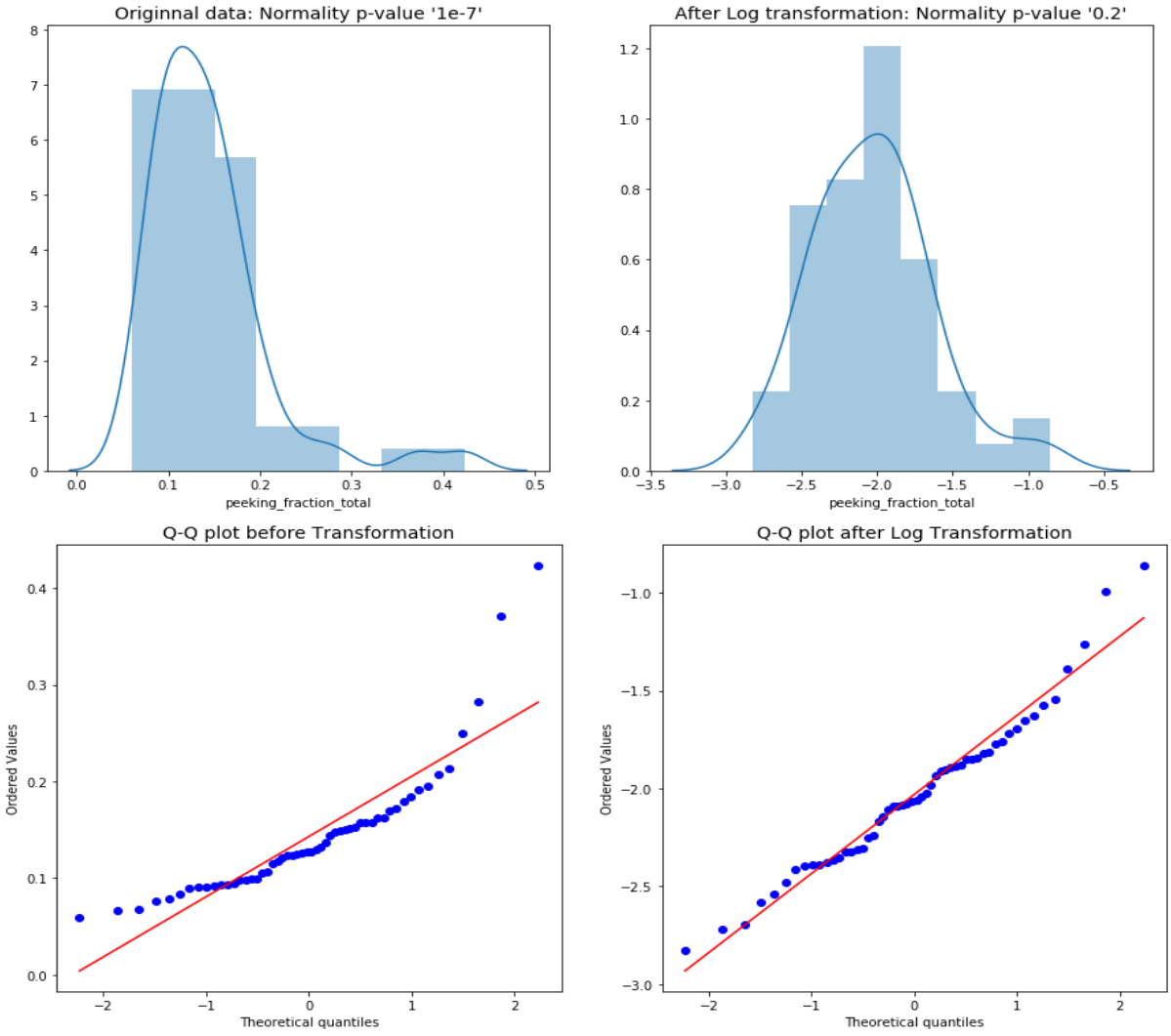

Figure 4: The distributions and Q-Q plots for the peeking feature before(left) and after(right) log transformation with p values from the Shapiro-Wilk test

The Shapiro-Wilk test is sensitive to small skews in the data (We have > 50 data points for both species). That is, just by looking at the p-values of the test, it is not possible to declare non-normality of an underlying distribution. In the example below, removing one outlier drastically changes the result of the test. Figure 5 shows the impact of outliers of one of the velocity features: average velocity open towards middle. With the outlier, the Shapiro-Wilk test yields a low p-value of 1e-14, which rejects the hypothesis of normality. But after removing the outlier, the p-value shot up to 0.98.

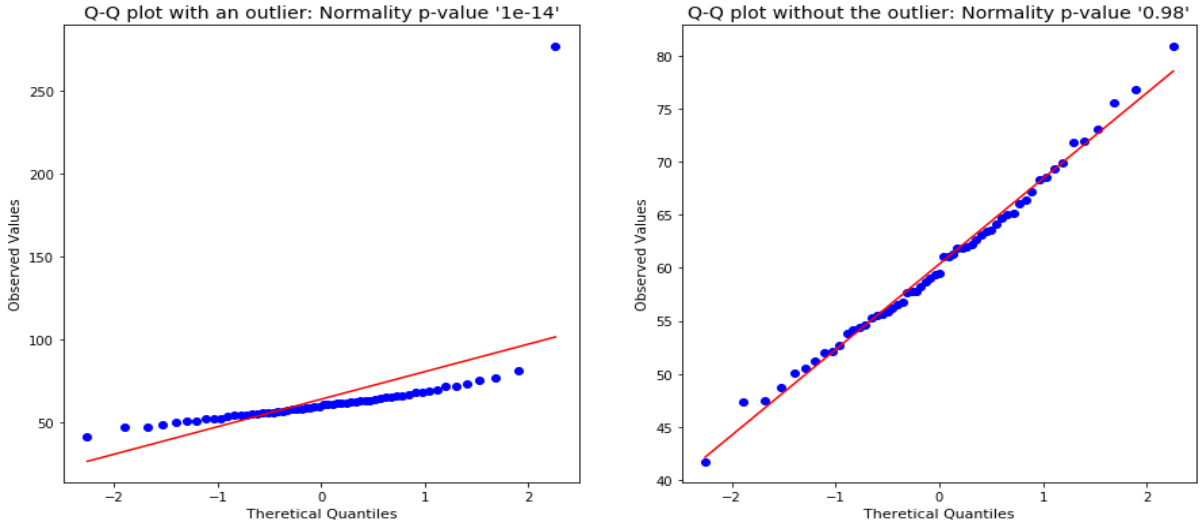

Figure 5: Velocity Feature: Impact of outliers on the Shapiro-Wilk p-values

Through normality test and visual inspection of the Q-Q plot, we find that $\sim 75\%$ of the features are normal(or, close to), and a few more could be made normal through a log transformation. Then, we go ahead and conduct a two factor ANOVA. For features where the conditions were satisfied before or after transformation, we can believe in the veracity of the results. For the others (a small percent of features) we will have to take the results with a pinch of salt.

Results

- **Species level significance:** Features below are statistically different across species (with p-values, %d refers to the percentage difference in means between the species: $\frac{BW-PO}{BW}$).

Attribute	%d	p	Attribute	%d	p
Mouse Length	-19%	7E-16	Turn preference back	-30%	2E-2
Fraction of time in open arms	70%	4E-20	Fraction of time at rest in closed arms	10%	7E-3
Fraction of time in closed arms	-26%	1E-17	Fraction of time at rest in open arms	-11%	9E-3
Fraction of time in middle arms	16%	2E-2	Fraction of time at rest in all arms	10%	9E-3
Turn preference left	31%	6E-6	Average velocity moving from closed away from middle	14%	7E-6
Turn preference right	35%	4E-6	Average velocity moving from open arms away from middle	21%	1E-5
Turn preference straight	-116%	4E-8	Average velocity moving from open arms towards middle	21%	1E-3

An example is shown in Figure 6.

Figure 6: Distributions of attribute 'Velocity closed away from middle', which varies by species (and not much by sex)

- **Sex level significance:** %d(X) indicates the %difference in means of two sexes for each species X: $(\frac{Male-Female}{Male})$

Attribute	%d(BW)	%d(PO)	p
Fraction of time in Open arms	-83%	-28%	1E-3
Fraction of time in Closed arms	6%	7%	4E-3

- **Species-Sex Interaction significance:** %d_sex(X) refers to the % difference between Male and Female mice of species X. Similarly, %d_species(G) refers to the % difference between BW and PO mice of sex G.

Attribute	%d_sex(BW)	%d_sex(PO)	%d_species(M)	%d_species(F)	p
Fraction of time in Safety region	10%	-25%	18%	-13%	2E-4
Turn Preference Left	10%	-24%	-24%	-72%	3E-2

Note: The attribute 'Fraction of time in Safety region' is not significant at either 'Species' or 'Sex' level, but is significant when the two factors combine. This is so because in 'BW', female Mice prefer more 'safety' when compared to Male and vice-versa in 'PO', effectively cancelling out each other at one factor level. This is shown in Figure 7.

Genomics

Now that we have features, each representing a different behavior, with statistically significant effect of Species, Sex, or Species*Sex interaction, we want to see which regions of the mice genomes correlate with which features. To do this, we use quantitative trait loci (QTL) analysis on the second generation hybrid mice. These hybrid mice have a combination of genetic information from both species, meaning if we observe behavior in the hybrid mice, we can use QTL to see which genetic markers the hybrids share with the pure species to understand how certain genetic markers influence different mice behaviors.

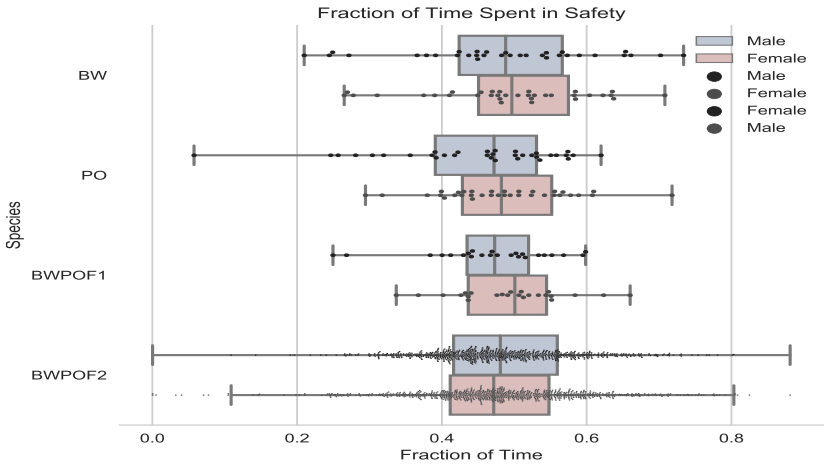

Figure 7: Distributions of attribute 'Time spent in safety'

Results

Here we present plots of the QTL analysis on two selected features found to be significant between species. These plots show the effect of different sections of the genome on the observed behavior.

Figure 8 is an example of a basic feature, $fraction\_in\_arms\_closed$, which is simply tracking when the mouse spends time in the closed/walled-off regions of the maze. The fact that this feature is so generic means we expect the QTL plot will be more volatile, because many genes might influence the behavior of staying in the closed regions. Sure enough, we see that many genes across multiple chromosomes play a role; however, there are localized sections where the influence is greater such as in chromosomes four, nine, and fifteen.

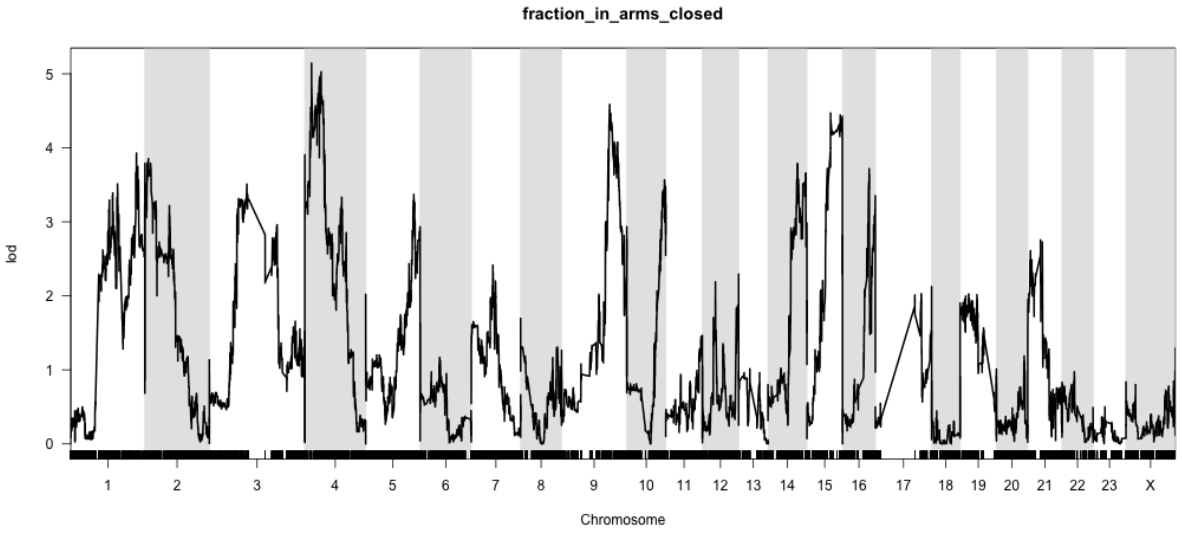

Figure 8: QTL plot of $fraction\_in\_arms\_closed$. If the score on the y axis is higher, this means that chromosome has a greater influence on the behavior

Moving to a more specific feature, let's look at $velocity\_open\_towards\_middle\_average\_speed\_active$

(Figure 9), which measures the average speed of the mouse as it travels from an open arm back toward the center of the maze. Interestingly, the influence of chromosomes on this feature is far more localized, showing chromosome four has a much stronger correlation with this behavior above all other regions.

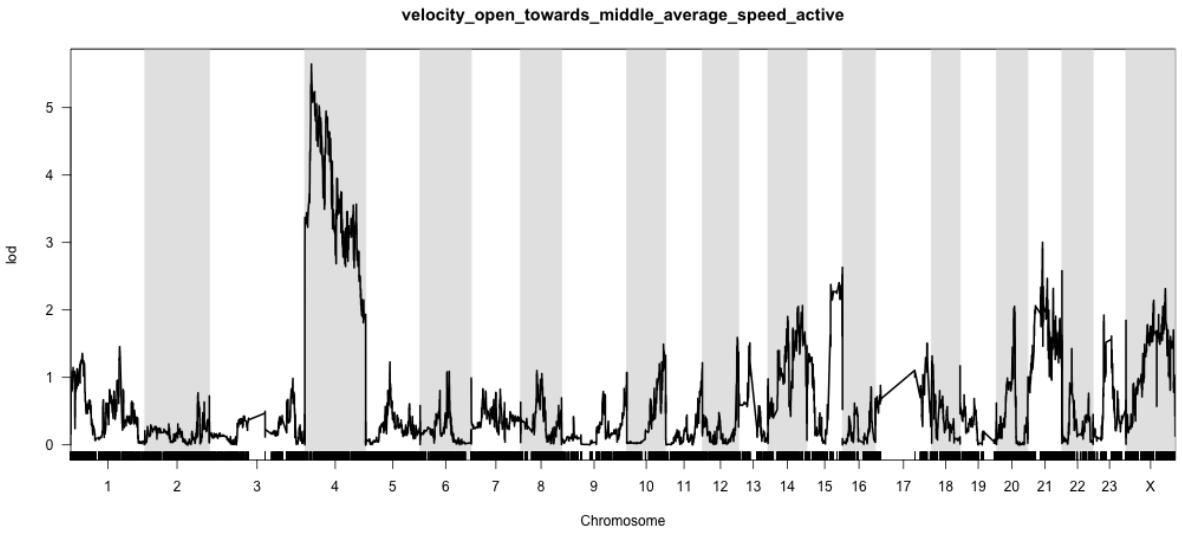

Figure 9: QTL plot of $velocity\_open\_towards\_middle\_average\_speed\_active$. If the score on the y axis is higher, this means that chromosome has a greater influence on the behavior

Note: As a caveat, we expect that many of these plots will be rather volatile. QTL analysis can show you whether a specific phenotype is correlated with many genes in minute amounts or strongly correlated with a few genes.

Further Steps

Now that we have the significant features, we will extend the QTL analysis to explore the genetic connections across all features. We will also be running the QTL analysis using multiple permutations and create a fuller picture of genetic regions that correlate strongly with observed behaviors.

References

[1] DNA - What is DNA? - Basics of DNA, https://www.youtube.com/watch?v=uXdzuz5Q-hs

[2] ”Oldfield_Mouse”, http://www.wildflorida.com/wildlife/mammals/Oldfield_Mouse.php

[3] Jones E, Oliphant E, Peterson P, et al. ”SciPy: Open Source Scientific Tools for Python”, 2001-, http://www.scipy.org/ [Online; accessed 2018-11-27].

[4] Seabold, Skipper, and Josef Perktold. ”Statsmodels: Econometric and statistical modeling with python.” *Proceedings of the 9th Python in Science Conference*. 2010.