Genetic Basis of Exploratory Behaviors in Mice (FY1821)

Cynthia, Josh, Kewei, Srinidhi, Zach

With Prof.Andres Bendesky & Data Science Institute

Contents

Abstract

This project explores the link between observable animal behaviors (phenotypes) and their underlying genetic bases (genotypes). We looked at video data of 1500+ mice exploring an environment split across many features. The videos were used to perform genomic analyses and discover correlations between genetic regions and observed behaviors. The mice used were from two sister species and cross-bred to create both first and second generation hybrids. Before our work began, the Bendesky lab had already conducted the experiment on the mice, calculated a few behavioral traits from the data, and linked these traits to the genetics of the mice. Our task was to expand the feature space, conduct significance analysis, and link the significant features we calculated to the genetics of the mice. We were able to expand the feature space to 200+ features, isolate several significant ones, and identify strongly correlated genomic regions that affect them. These findings will help the lab advance and focus its understanding of the relationship between physical behaviors and genetic information.

Intoduction

1 Mice

The experiment was performed using two sister species of mice with known differences in exploratory behavior. *Peromyscus polionotus* (PO, Oldfield mice) are known to be more exploratory and navigate into open spaces, while *Peromyscus maniculatus* (BW, Deer mice) are less prone to explore and prefer closed, hidden spaces. To understand the genetic influence on this difference in behavior, the experiment was performed on both the pure species and their first and second generation hybrid offspring (F1 and F2). These hybrids were created by cross-breeding the mice as shown in *Figure 1*, so each of the F2 hybrids had combinations of the genes from both pure species.

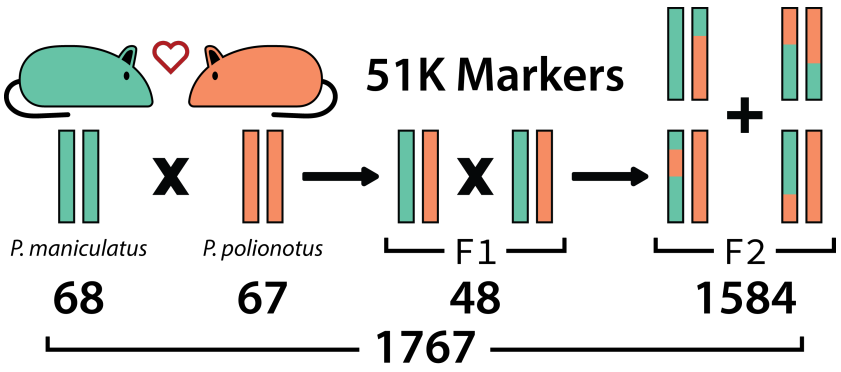


Figure 1: Illustration of genetic cross-breeding to create F1 and F2 hybrids. Numbers indicate how many mice of each type were used for the study.

2 The Experiment

In order to observe and quantify their behaviors, each mouse was placed in an elevated plus maze setup as shown in *Figure 2.* The left and right arms have tall walls, mimicking an enclosed space, which we will refer to as 'closed arms'. The top and bottom arms have no walls, mimicking an open space, which we will refer to as 'open arms'. The intersection region is referred to as the 'middle', which is considered open.

Each mouse was put into the maze via a small mouse elevator. After acclimating, the elevator was removed and the mouse was free to explore the maze. This experiment was run with over 1500 mice (see *Figure 1* for breakdown by pure/F1/F2 animals) and the behavior of each mouse was recorded with a top-down camera for a maximum of 5 minutes. From each video, a processing script determined a bounding box for the location of the mouse at each point in time. We used this location data to derive all the behavioral features for the study.

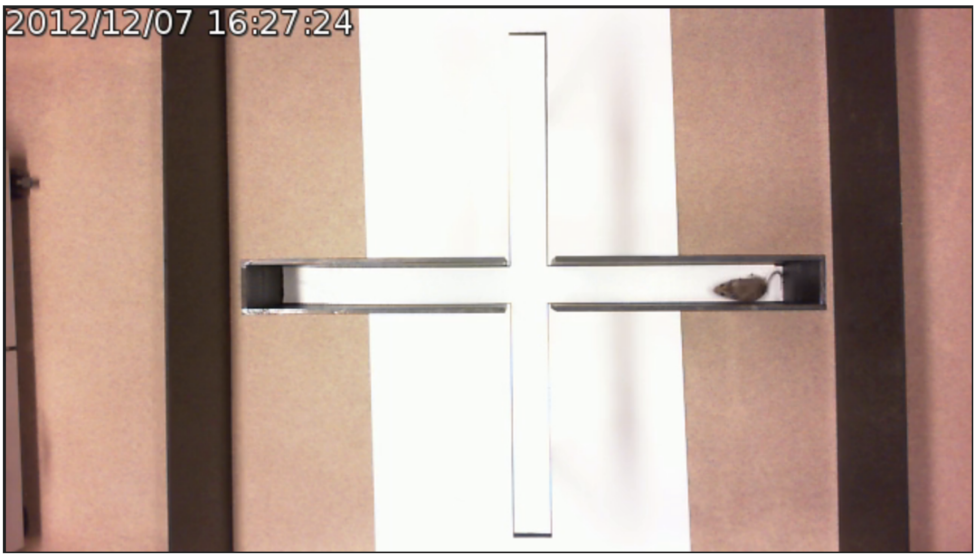


Figure 2: Aerial view of the elevated plus maze. In this frame, the mouse is located in the closed right arm.

3 Data Ethics

The data we used involved raw video footage of mice exploring an elevated plus maze. Considering our experiment did not involve humans, we had minimal ethical concerns with regards to privacy of personal information of the subjects involved. The animals used in the study were bred for the purpose of scientific research and not captured in the wild. All subjects were treated humanely in accordance with the Institutional Animal Care and Use Committee (IACUC) guidelines. The data acquired from the animals was used to further the understanding of the connection between observed behaviors and genetic information. Because this was a scientific research study involving animals, the ethical concern was more on the treatment of the subjects rather than on their consent of divulging private information.

Getting Set Up

1 The Data

From the experiment described above, we were provided with a shared drive of files extracted from the video using tracking software, including frame by frame coordinates for each mouse. Additionally we were given a script that uses frame by frame information to calculate a few mice features. We added features and edited the script provided to migrate all the feature results into a single CSV file.

Figure 3 shows the distribution of the mice by sex and species for the collective CSV file. Most of the observed mice are BWPOF2 and the rest are distributed evenly between the BW, PO and BWPOF1 species. The distribution by sex is roughly even across species.

2 Explore Existing Features

The original features provided include the following:

- Fraction of time in each arm
- Median speed & smoothed median speed in each arm
- Total distance & smoothed total distance travelled in each arm
- Number of entries into each arm

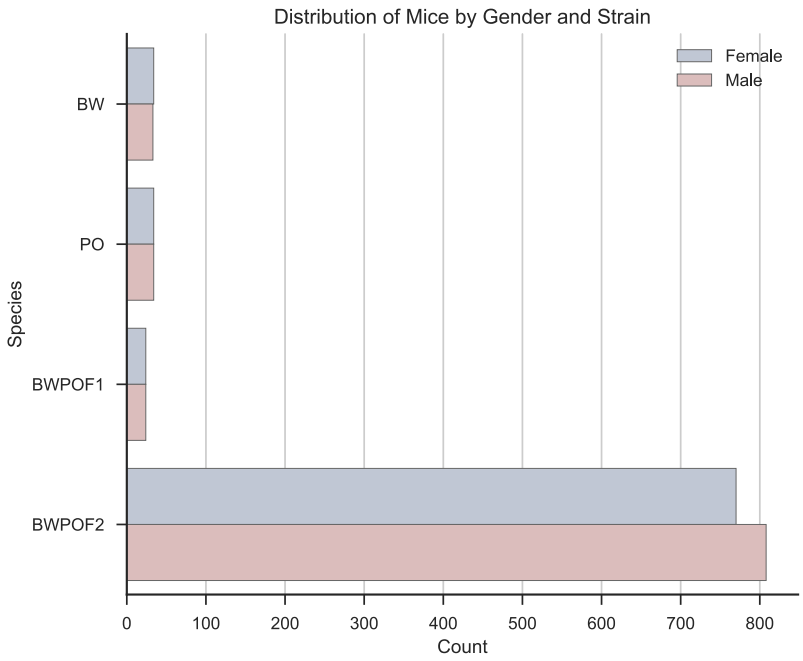


Figure 3: Distribution of mice by gender and Species

The goal is to find behaviors that are demonstrably different between the pure bred mice (BW and PO). As an example of one such behavior, *Figure 4* shows the fraction of time the pure-bred mice spent in closed and middle arms. Notably, we see a significant difference between the two pure breeds in their behavior. The BW mice spend a larger portion of their time in the closed and middle arms while the PO mice spend a larger portion of their time in the open arms. This is consistent with the expected behavior of the mice as the PO mice are more adventurous and the BW mice prefer closed spaces. Now that we know there is a behavioral difference between pure breeds, we can proceed with the analysis to look at correlations between genetic markers in the hybrid mice and these behaviors.

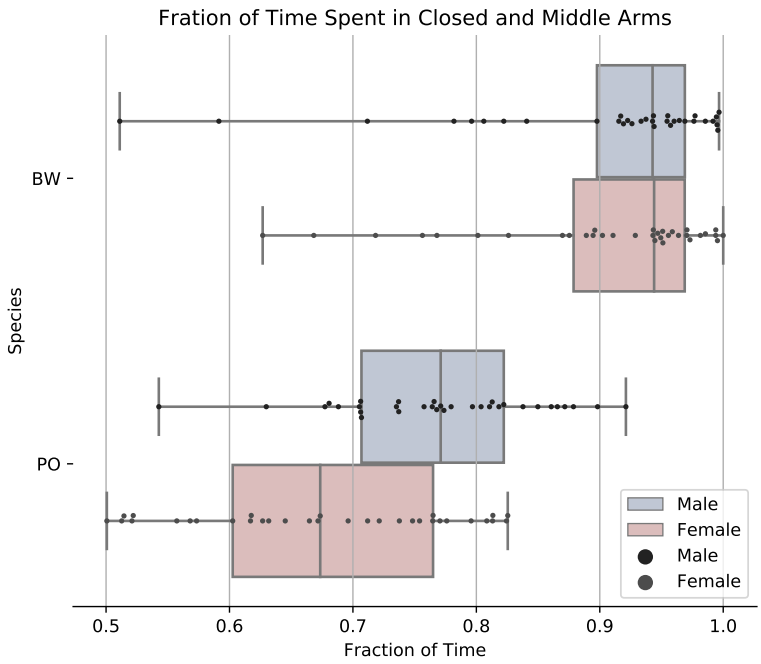


Figure 4: Distribution of the fraction of time spent by pure species in the closed and middle sections of the maze

Feature Engineering

We were provided with a script that reads in location data of the mice over time and calculates some basic features. We expand on that script to create a total of 210 features. These features are split across different categories of observable behaviors and attributes (see *Figure 5* for an illustrated explanation of the categories):

1 Feature Types

- **Mouse Dimensions:** the length, width and size of the mouse.
- **Time Spent in Arms:** the fraction of time the mice 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(end of arm). We use the length of the mouse to determine how far from the edge of the arm is considered *safety*.
- **Time Spent at Rest:** the amount of time spent not moving. We consider the mouse not moving when it's velocity is below a certain threshold for a minimum number of frames.
- **Peeking Behavior:** a peeking behavior is defined as the mice being in rest either in the middle section or close to the middle section within a closed arm. We have observed many mice to stop near the middle zone and stop/consider/peek before moving to the Open arms. We use the length of the mouse to determine the distance from the middle that we consider peeking.
- **Velocity:** the velocity of the mice 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 sub-features (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.

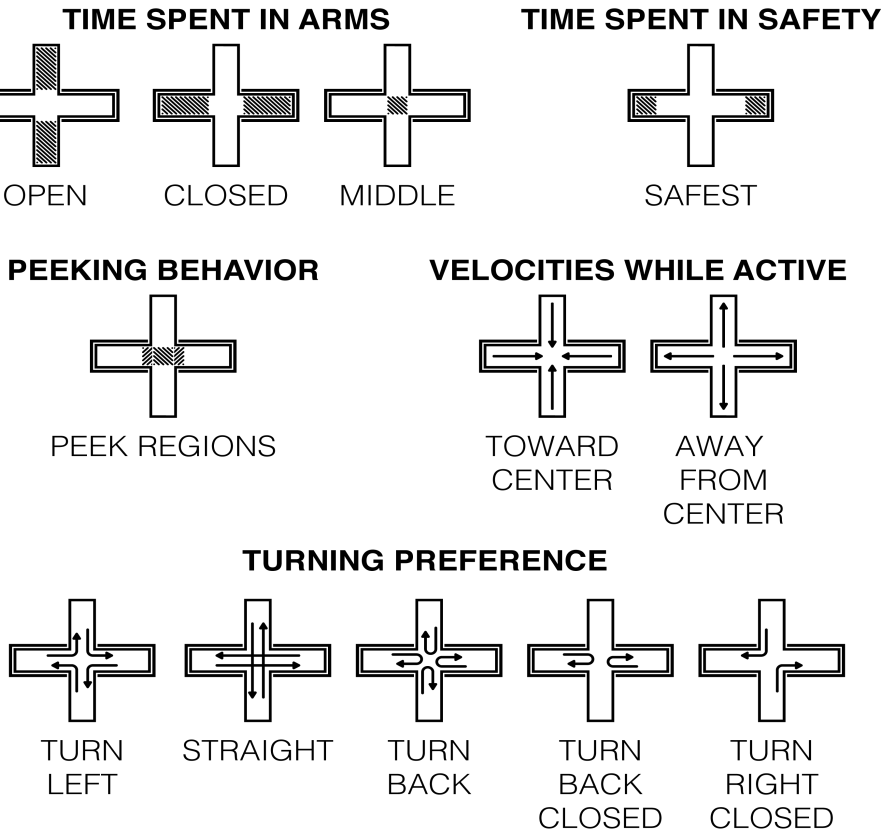


Figure 5: Pictorial representation of selected feature categories

2 Explore New Features

As before, we want to explore behaviors that are demonstrably different between the two species and/or the sex of the mice. We use boxplots to look at the differences in behavior of the features we calculated. In *Figure 6* we can see that the fraction of straight turns (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 7* shows that PO mice tend to be smaller than BW mice. For plots of additional features, please refer to the Appendix.

Note: We used these box plots to find and investigate outliers. If we found that there was an error in our code or the video tracked that was causing an outlier, it was fixed. If not, it was considered as a true outlier and it was left alone.

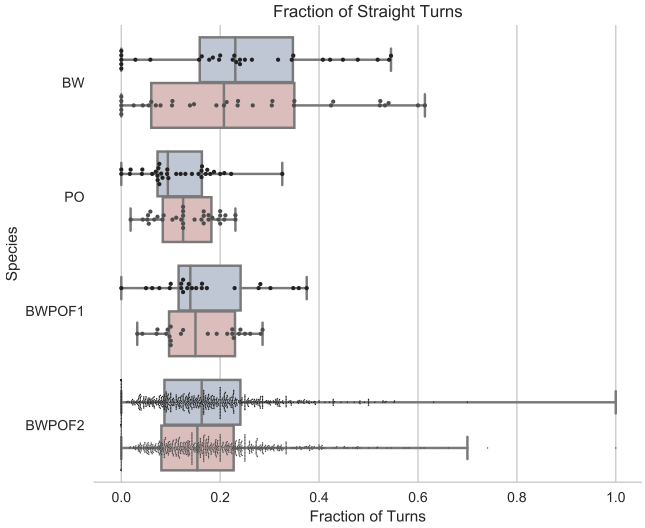


Figure 6: Distribution of the fraction of straight turns out of all turns for each mouse by gender and species

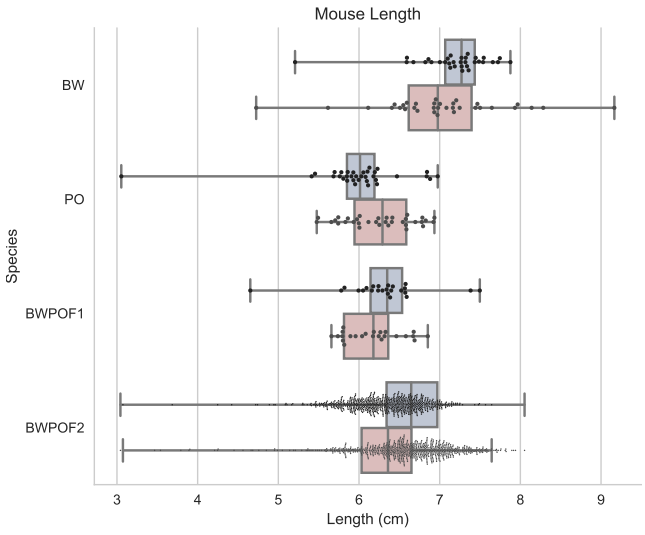


Figure 7: Distribution of mice length by gender and species

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 different dimensions: Species, Sex, and Species*Sex interaction.

1 Significance Tests- ANOVA

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 combinations of each feature. We also generate Q-Q plots to examine the data distributions. 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 8* shows a peeking feature with a heavy right tail. After a log transformation the feature satisfied the normality conditions for the ANOVA test.

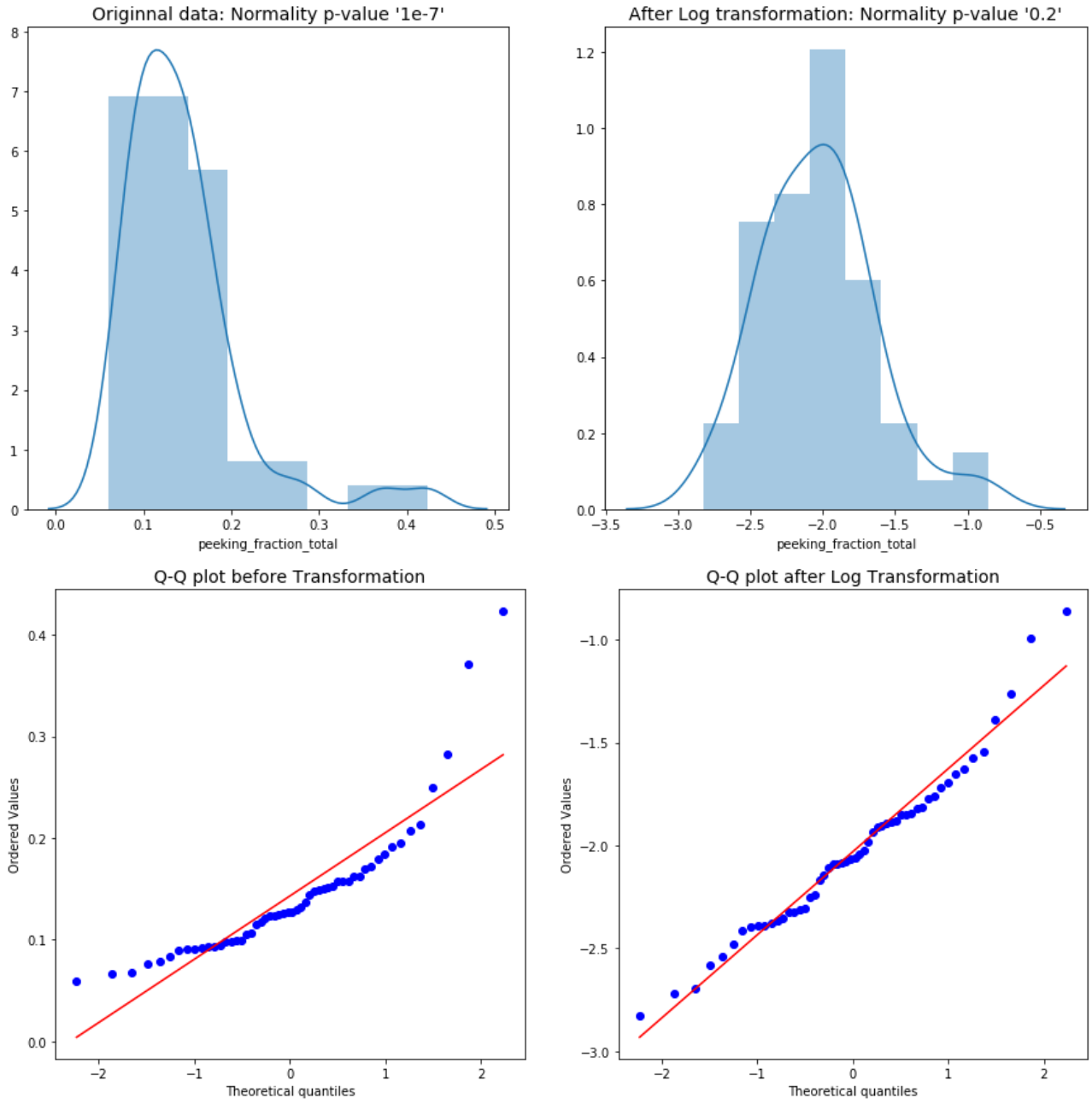


Figure 8: The distributions and Q-Q plots for the peeking feature before(left) and after(right) log transformation with p values from the Shapiro-Wilk test

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 9* shows the impact of outliers on 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.

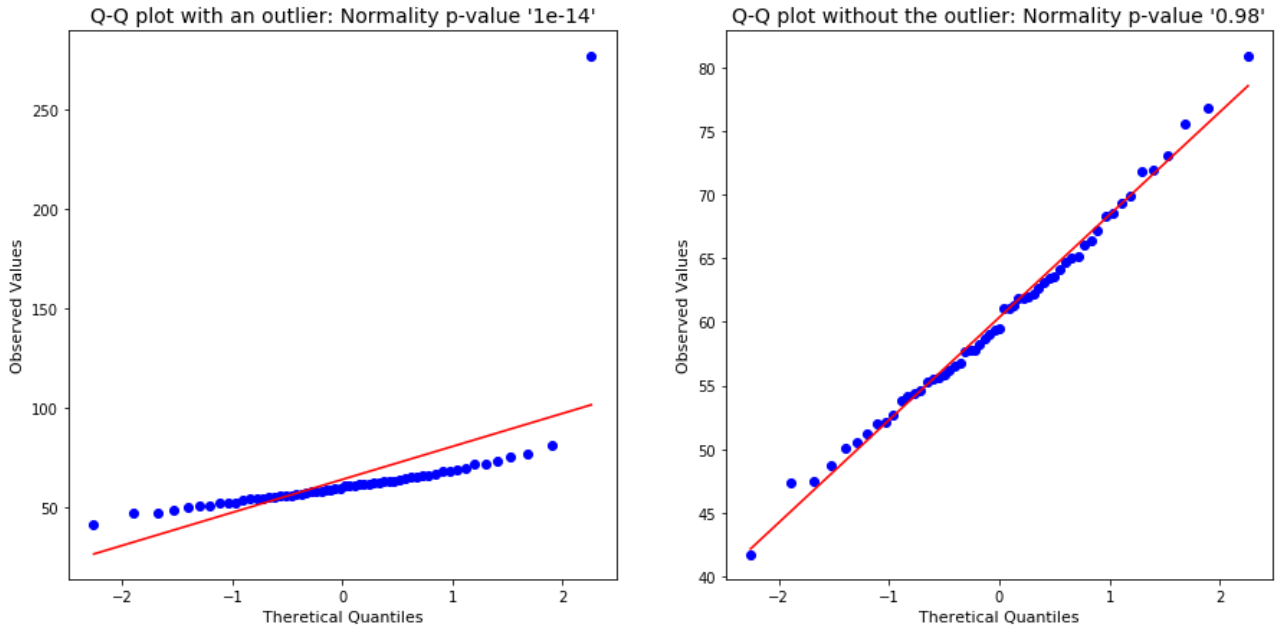


Figure 9: Velocity Feature: Impact of outliers on the Shapiro-Wilk p-values

Through normality tests and visual inspection of the Q-Q plots, we find that $\sim 75\%$ of the features are approximately normal and a few more could be made normal through a log transformation. Once normality has been established, we conduct a two factor ANOVA. For the a small percent of features for which we could not establish normality, we will have to take the results with a pinch of salt.

2 Significance Conclusions

- **Species level significance:** Features below are statistically different across species (with p-values, %d refers to the percentage difference in means between the species: $\frac{BW-PO}{BW}$). An example is shown in *Figure 10*.

Attribute	%d	p	Attribute	%d	p
Mouse Length	13%	6E-18	Turn preference back	33%	1E-4
Mouse Width	5%	2E-15	Fraction of time at rest in closed arms	-8%	2E-2
Mouse Size	17%	9E-26	Fraction of time at rest in open arms	7%	1E-2
Fraction of time in open arms	-177%	4E-17	Fraction of time at rest in all arms	9%	2E-2
Fraction of time in closed arms	22%	2E-13	Average velocity moving from closed away from middle	-13%	4E-5
Turn preference left	-54%	3E-8	Average velocity moving from open arms away from middle	-27%	4E-8
Turn preference right	-48%	3E-8	Average velocity moving from open arms towards middle	-20%	6E-8
Turn preference straight	54%	4E-10			

- **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
Average velocity moving from open arms away from middle	-8%	-15%	2E-2
Average velocity moving from open arms towards middle	0%	-19%	7E-3

- **Species-Sex Interaction significance:** %d_sex(X) refers to the % difference between Male and Female mice of species X. Similarly, %d_species(Y) refers to the % difference between BW and PO mice of sex Y. An example is shown in figure 11.

Attribute	%d_sex(BW)	%d_sex(PO)	%d_species(M)	%d_species(F)	p
Mouse Size	3%	-7%	21%	14%	3E-2
Mouse Width	2.5%	-4%	7%	2.5%	7E-3
Fraction of time in Closed arms	0%	13%	17%	27%	4E-2
Turn Preference Back	-18%	25%	15%	46%	3E-2

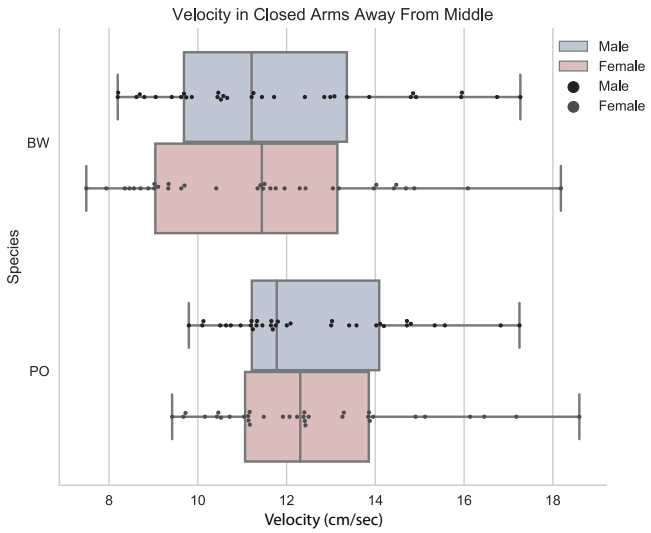


Figure 10: Distribution of attribute 'Velocity closed away from middle', which varies by species (and not significantly by sex)

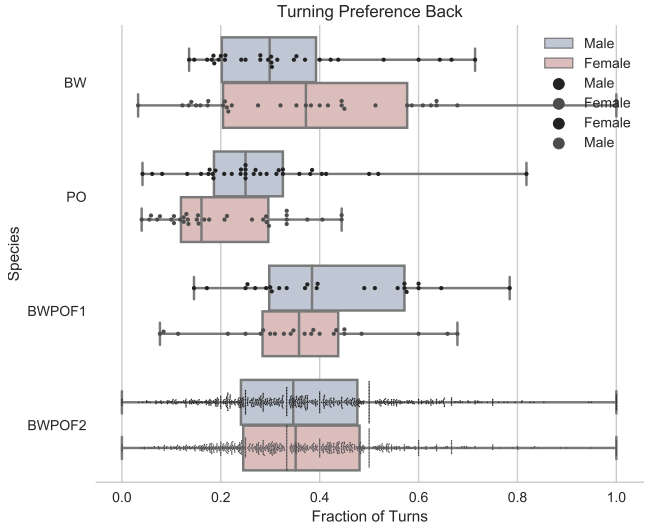


Figure 11: Distribution of feature 'Turn preference back'. Differences between species & sex can be observed

Genomics

1 QTL Process

QTL (quantitative trait loci) analysis serves to show which regions of the mice genomes correlate with different features. The process uses the second generation hybrid mice (F2), which have a combination of genetic information from both species. When we observe behavior in the hybrid mice, we can use QTL to explore which genetic markers the hybrids share with each of the pure species to understand how certain genetic markers influence different mice behaviors.

Here we present plots of the QTL analysis on three selected features found to be significant between species. These plots show the effect of different sections of the genome on the observed behavior. The plots show LOD scores (logarithm of the odds), which measure the correlation between an observed behavior and a genetic region. We chose a LOD score of four as the threshold for significance, as marked by the dotted lines on the charts.

Figure 12 shows Mouse Length, a different feature from the rest as it is a morphology not a behavior. Because of this, we expect many genes heavily influence it. The QTL plot shows multiple peaks far above the threshold, suggesting regions in chromosomes 4, 6, 8, 14, 18, and X are very highly correlated with this morphology.

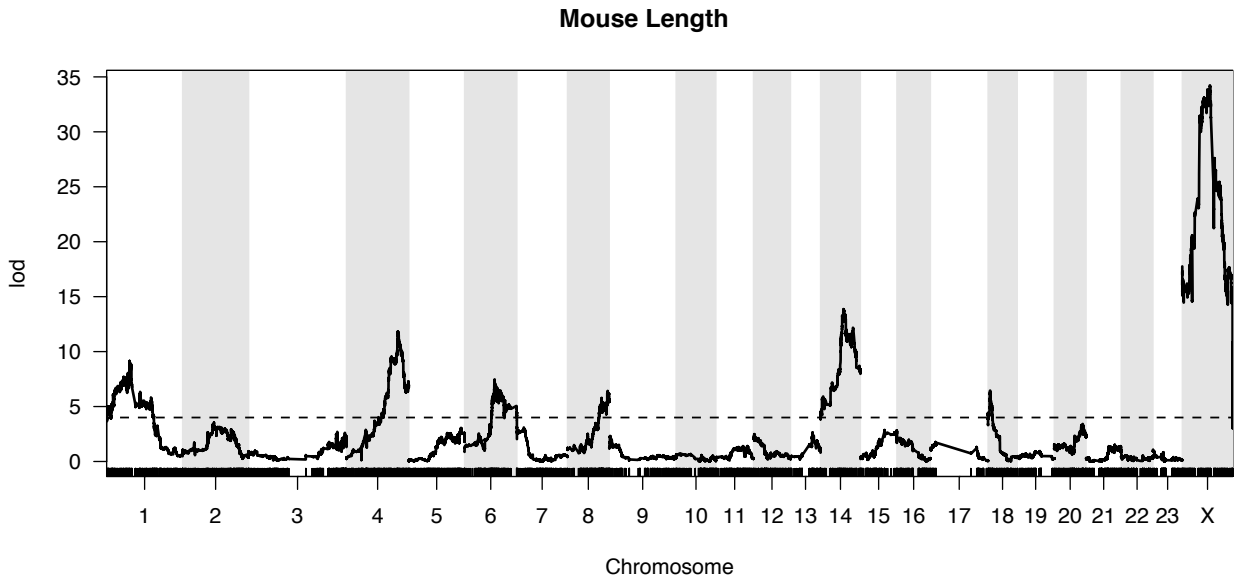


Figure 12: QTL plot of Mouse Length. Chromosome X looks very significant and many other peaks also hit well above the threshold

Figure 13 shows Average Velocity Towards Safety, which describes the average velocity of a mouse only inside the closed arms running away from the center of the maze. We see that chromosome 19 is highly linked to this behavior. It is interesting to observe a QTL that is linked so strongly and narrowly to one behavior. Additionally, it is of note because this shows a genetic region tied to a exploration-adverse behavior.

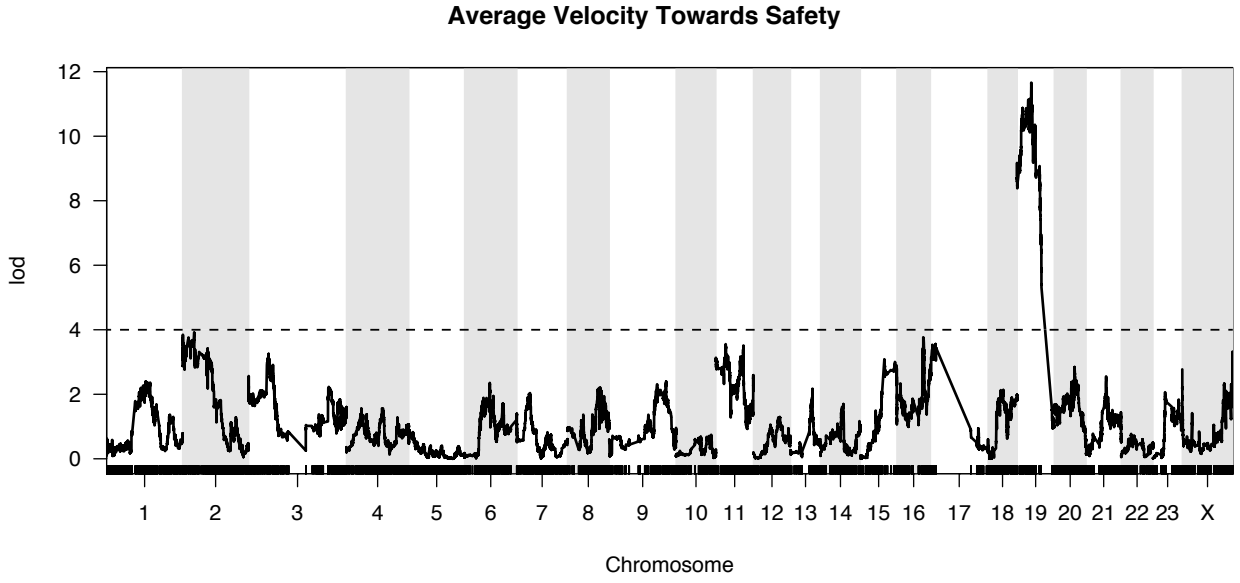


Figure 13: QTL plot of Average Velocity Towards Safety. Chromosome 19 looks to be very significant

Figure 14 shows Fraction in Closed Arms, which is simply the fraction of time a mouse spends in the closed arms. Interestingly, chromosome 19, which was linked to how quickly mice move into safety, is not highly correlated with this feature. This suggests there could be a genetic difference between running into a closed space and spending time in a closed space.

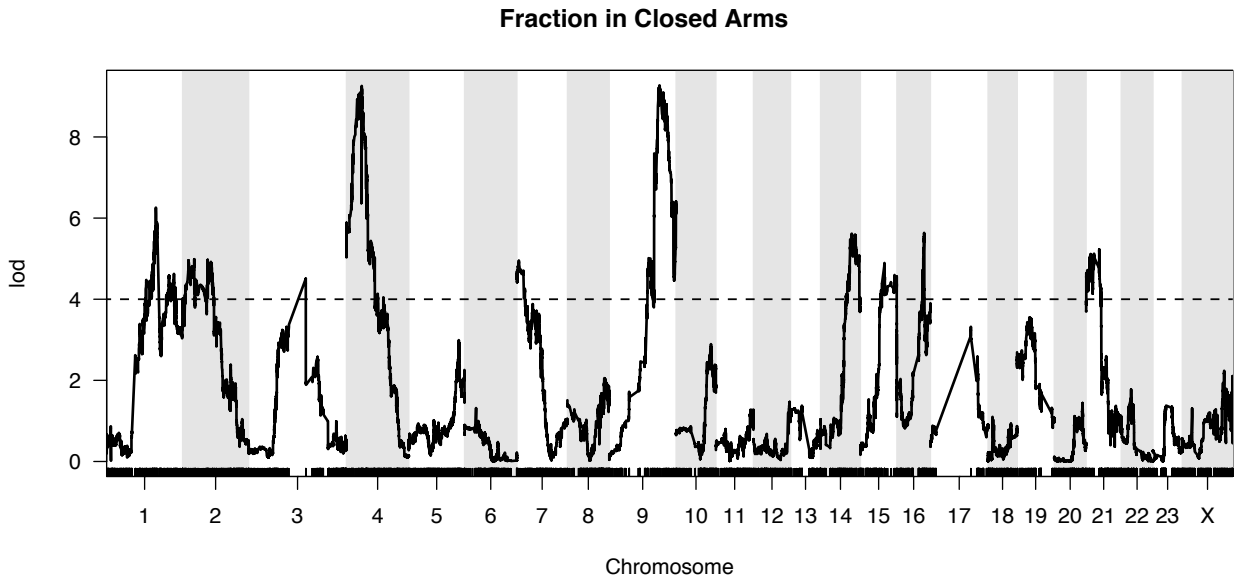


Figure 14: QTL plot of Fraction in Closed Arms. Many chromosomes seem to be significant in this behavior

2 Comparing QTL Between Features

The QTL heat maps in *Figures 15 and 16* are the same as the QTL plots, just using saturation rather than height to show LOD scores. With this map, we can compare how certain chromosomes affect multiple features.

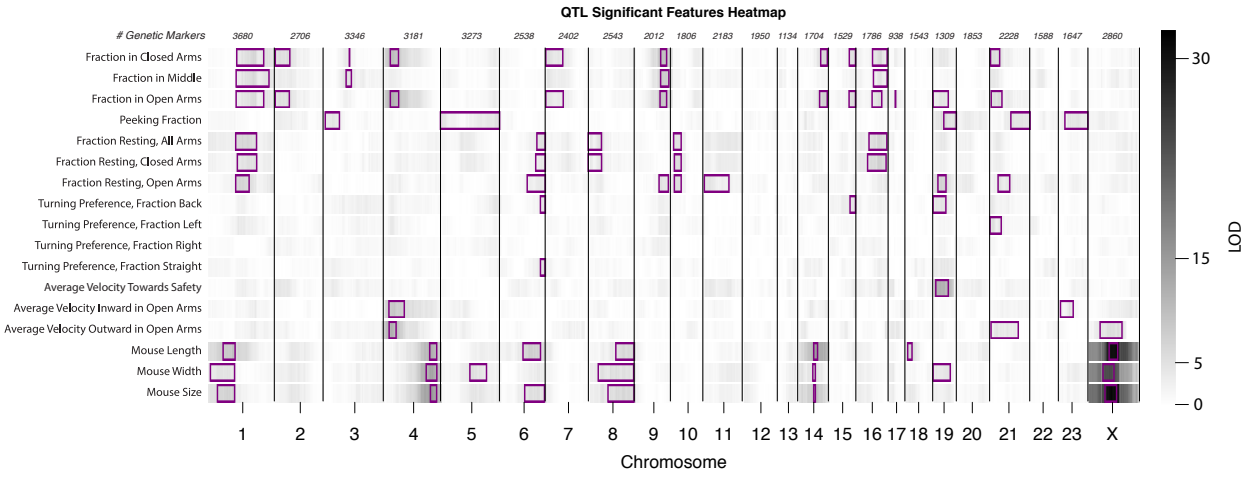


Figure 15: QTL heat map for all significant features

In comparing the QTL plots on this heat map, we can look at interesting interactions between features. Firstly, many of the interactions we see were as expected. All the mouse size features rely on mostly the same genes, and with similar LOD scores. This is also true for the fraction in different arms and the rest times in different arms. This makes sense, as we expect any chromosomes that lead to one of these behaviors also leads to similar behaviors.

Though many of the interactions were expected, some were not. For example, the chromosomes that influence velocity in closed arms are different from those that influence velocity in open arms. This is surprising as it might make sense for the velocity of the mice in any arm to be caused by the same chromosomes. We can also see interesting results of some chromosomes in multiple categories of features. The peeking feature

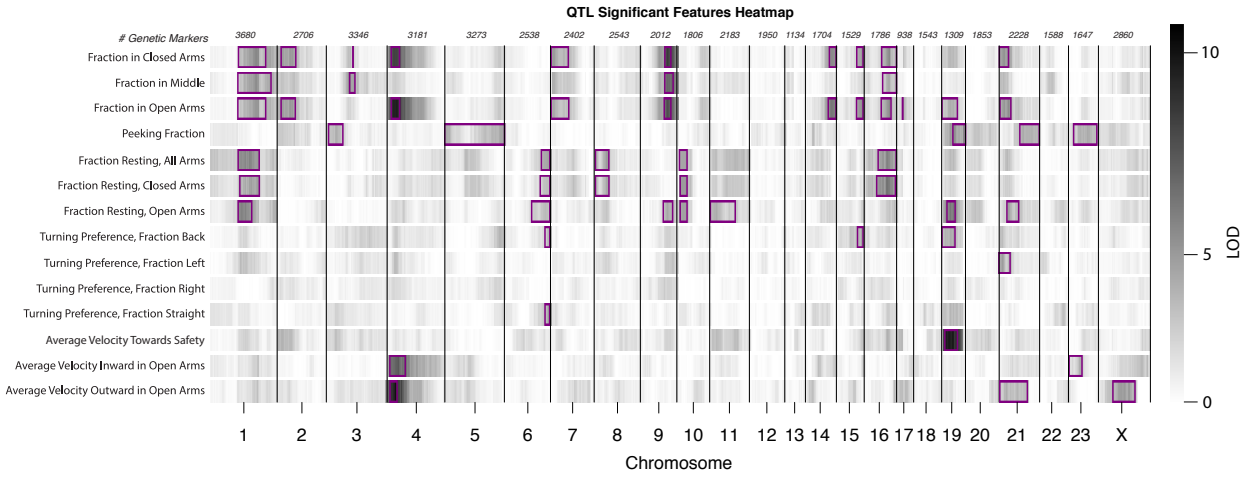


Figure 16: QTL heat map excluding features related to mouse size. These features were so significant they made it hard to read the other features

is influenced by chromosomes that also influence mouse width, fraction in open arms, velocity towards safety in closed arms, and a couple turning features. There are many interactions like this one that would be interesting to study further, investigating which specific gene segments influence these behaviors and if the same segments influence multiple behaviors.

3 Deeper Look into Features: Species× Sex Interaction

Phenotypes are frequently affected by genotype-by-sex interactions, so we consider adding sex as a covariate. Before we do that, we first perform genome scans for each sex separately. *Figure 17* is a plot for average velocity moving from Open arms towards middle, where the blue lines show the scan for male mice and the red lines show the scan for female mice. It seems that the two sexes highly different on chromosomes 13, which have a much stronger connection with this behavior for male mice than female mice.

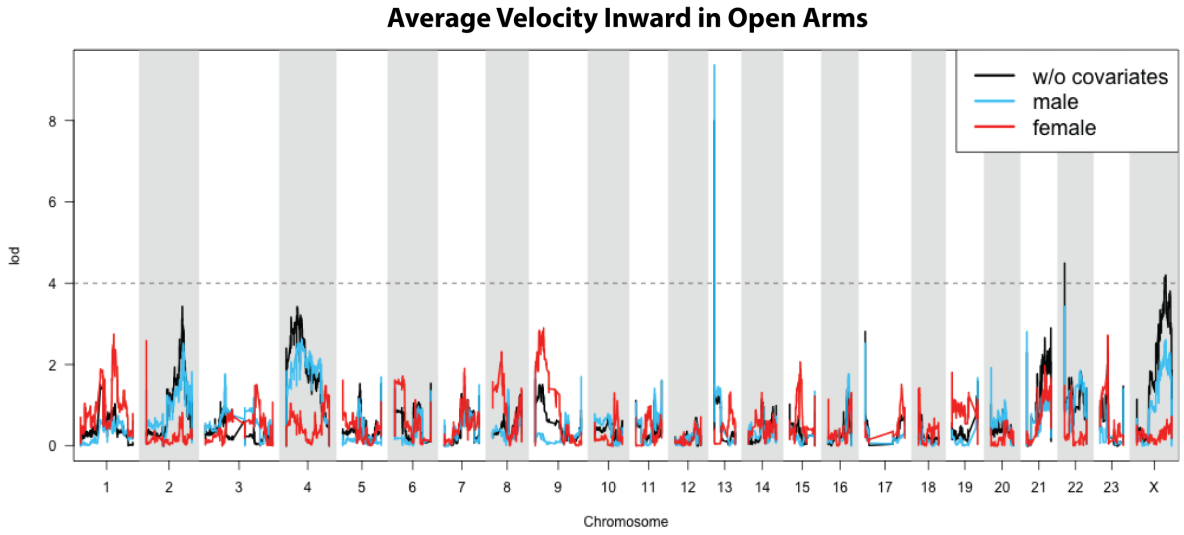


Figure 17: QTL plot of Average Velocity from Open Arms towards Middle, by sexes & without covariates

Note: For all the models in this part, we set parameters model and method to be 'normal' and 'hk'.

Now we combine sex as an covariate in the model. The QTL analysis is performed in two ways:

- Model with sex as an additive covariate
- Model with sex as an additive and interactive covariate

When sex is used as an additive covariate only, the differences between this model and model without covariate are tiny among most the features found significant in the previous analysis. One example of such comparisons is shown in *Figure 18*.

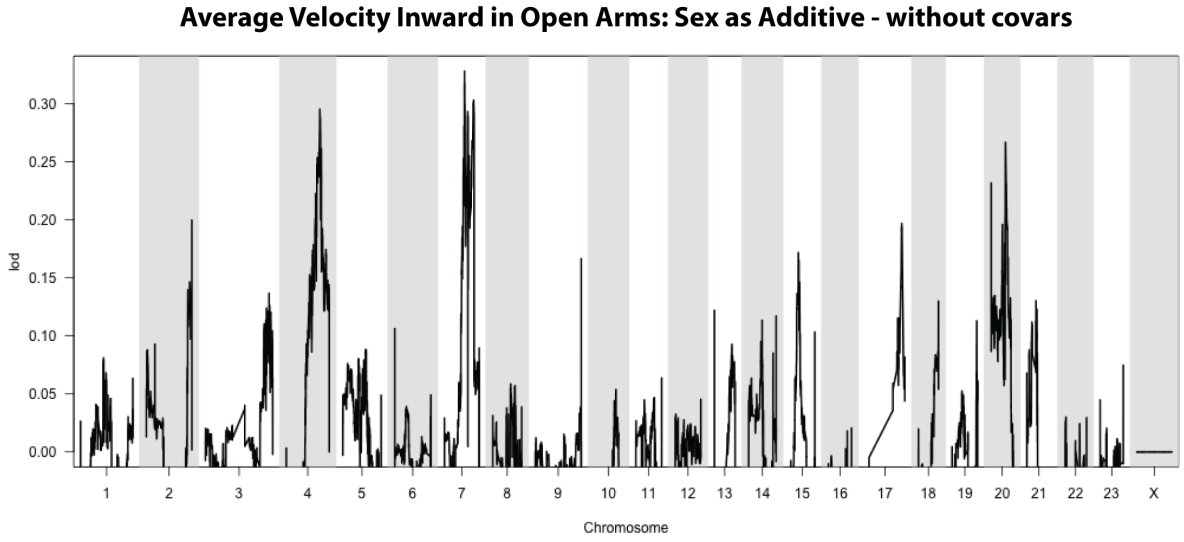


Figure 18: QTL plot of Average Velocity from Open Arms towards Middle, sex as an additive covar — without covars

Only three features (Mouse Length, Mouse Width, and Mouse Size) have comparatively significant changes in LOD scores after adding sex as an additive covariate, as shown in *Figure 19*.

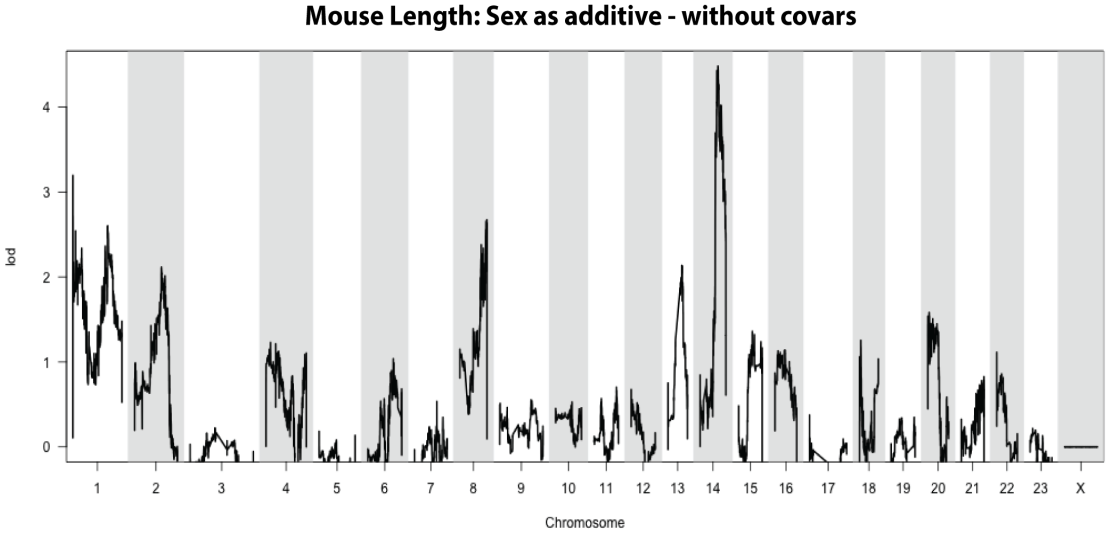


Figure 19: QTL plot of Mouse Length, sex as an additive covar — without covars

The sex as an additive covariate model allows the average phenotype to be different in the two sexes, but assumes the same effect of the QTL in the two sexes. The results show that after eliminating the average phenotype differences in the two sexes, the correlations between behavior and genetics for all the significant features don't change much. Therefore we can allow the QTL to be different in the two sexes. *Figure 20* shows

the difference between model with sex as an additive & interactive covariate and model with sex as an additive covariate only, still for feature Average Velocity from Open Arms towards Middle.

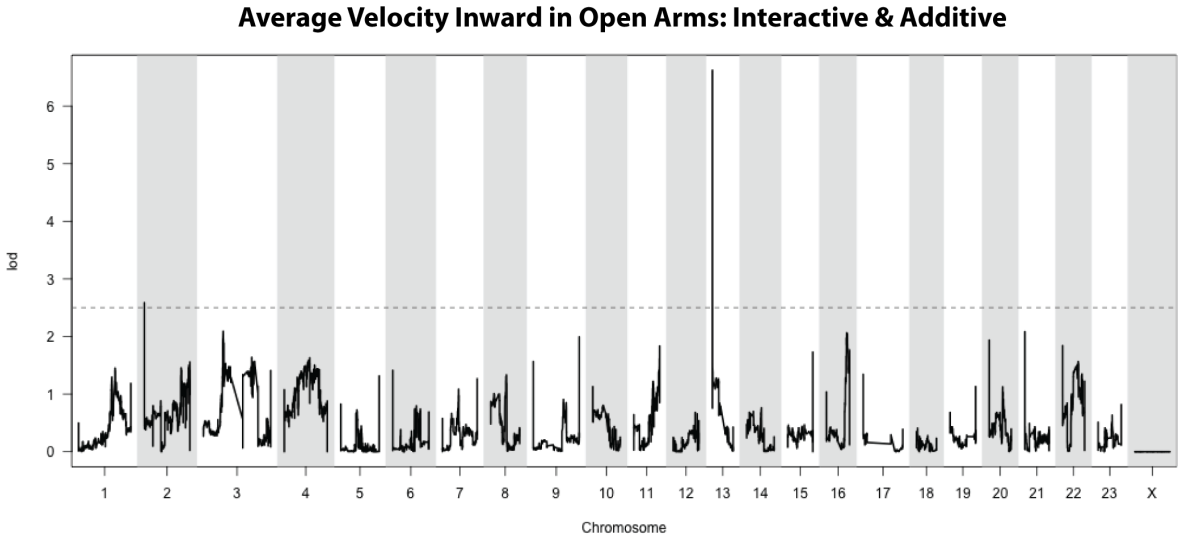


Figure 20: QTL plot of Average Velocity from Open Arms towards Middle, sex as an additive&interactive covar — sex as an additive covar

By choosing a LOD score of 2.5 as the threshold for significance, we can see the difference is significant at chromosomes 2 and 13. This means that the QTL has a different effect on the two sexes for this feature. And with threshold of 2.5 LOD score, the results we get indicate that species*sex interaction is significant for most significant features.

Figures 21 and 22 show a heat map of the QTL plots by sex. We can see that very few QTL that are significant in one sex are also significant in the other. This suggests that the genetic basis of exploratory behavior is different in the two sexes. In other words, the genes responsible for differences in exploration between the species are likely not the same in males and females.

That being said, looking at all the QTL plots in the appendix, none of the QTL that are significant in one sex but not the other have evidence of a significantly different effect in the two sexes. This could be due to lack of power (from not a large enough sample size) or from something else we are missing.

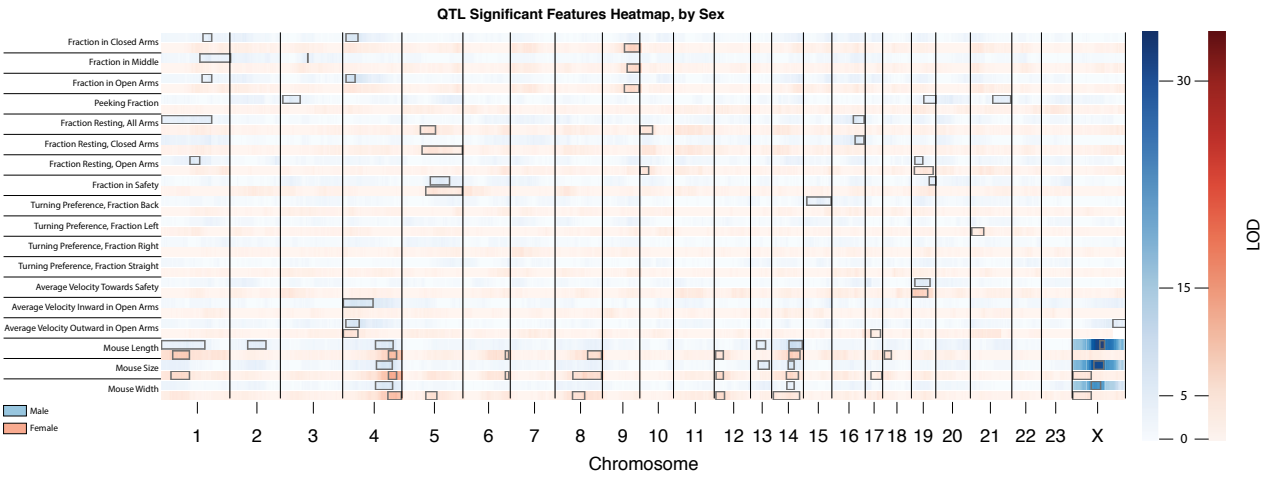


Figure 21: QTL heat map for all significant features separated by gender

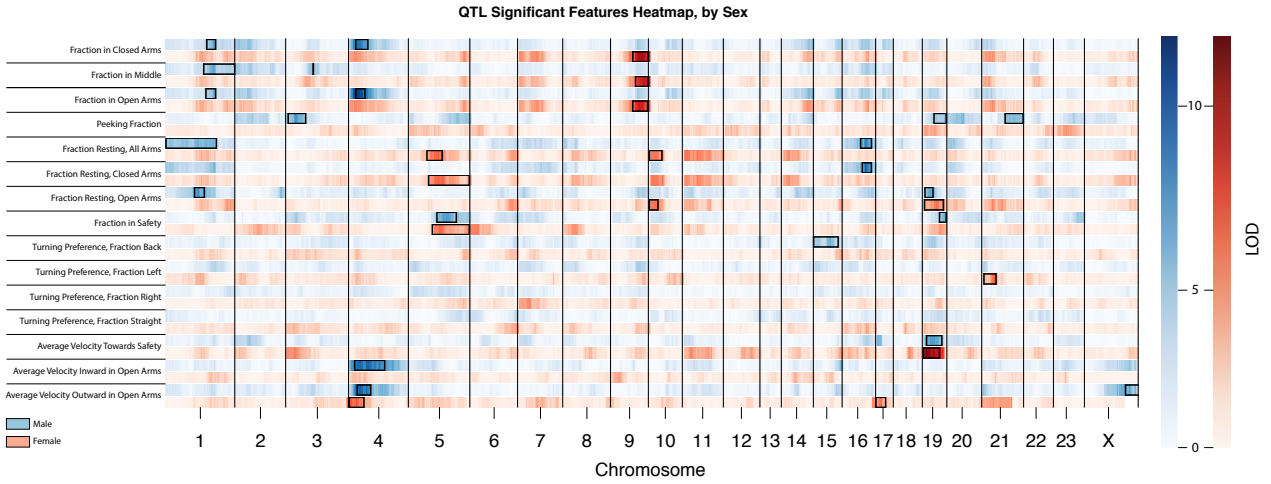


Figure 22: QTL heat map for all significant features separated by gender, excluding features related to mouse size. These features were so significant they made it hard to read the other features

Conclusion & Limitations

We were able to identify genetic regions that correlate with observed exploratory behaviors in the two sister species of mice. The results suggest that exploratory behaviors are complicated and are influenced by several QTLs across different chromosomes. Some limitations in this work involved a data processing bottleneck. Better performance could be achieved by running permutations on the QTL analyses. However, we were still able to achieve meaningful results and pave the way for future advancements by the lab.

Further Steps

Using the regions discovered, the Bendesky Lab can take further steps to explore the connection between observed behaviors and genetic information. Further analyses can be performed to discover which genetic version (BW or PO) at significant QTL regions promotes the observed behaviors. Additionally, the accuracy of the QTL analyses could be improved by performing permutations for each feature, leading to a more accurate determination of statistical significance. Other advances for this work will occur at the biological level. With the QTL regions identified, novel experiments can be designed and performed to isolate and confirm the effect of specific regions on observed behaviors.

Contributions

Below are general roles performed by each member. All members worked closely on the reports and other deliverables. A illustrated depiction of our pipeline is provided for reference (*Figure 23*)

- Zach Bogart: Processing Script, EDA, Design
- Cynthia Clement: EDA, Visualizations, Design
- Josh Feldman: Processing Script, Feature Engineering, QTL
- Kewei Liu: Processing Script, Feature Engineering, QTL
- Srinidhi Murthy: Feature Engineering, Process Script, Statistical Significance

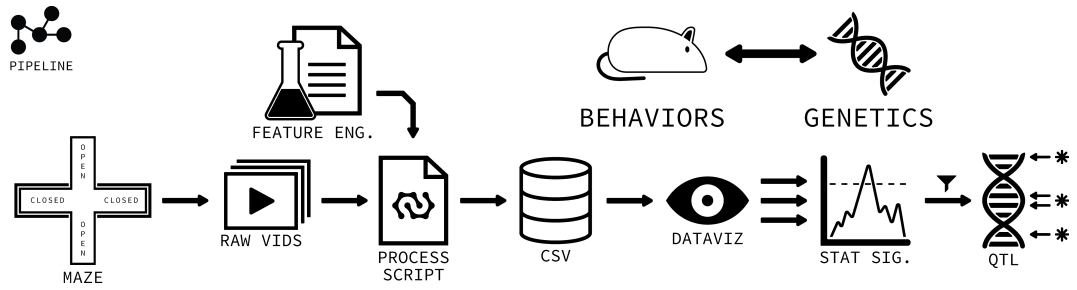


Figure 23: Illustrated pipeline of the project

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).

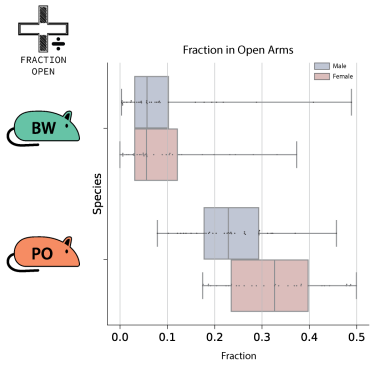
[5] Broman. K. W. "A brief tour of R/qtl". Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison. 2007.

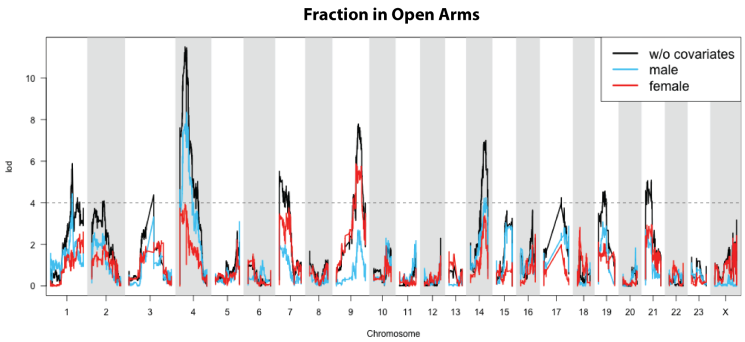
[6] Miles, C. Wayne, M. 2008. "Quantitative trait locus (QTL) analysis". Nature Education 1.1(2008):208

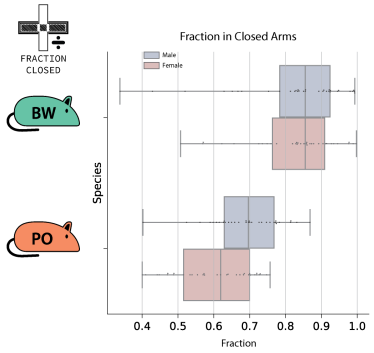
[7] Leduc, M. S., et al. "Integration of QTL and bioinformatic tools to identify candidate genes for triglycerides in mice." The Journal of Lipid Research 52.9(2011):1672-1682.

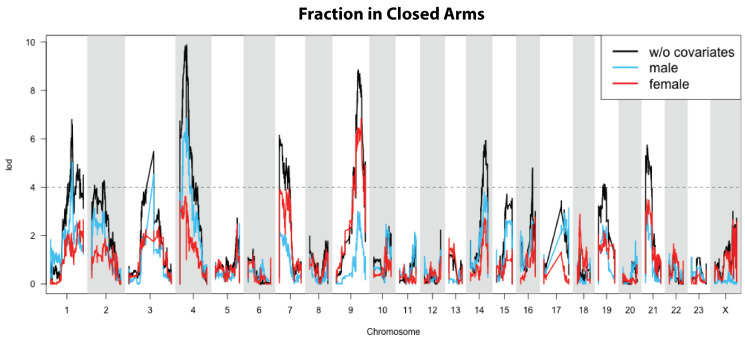
Appendix

Fraction in Arms

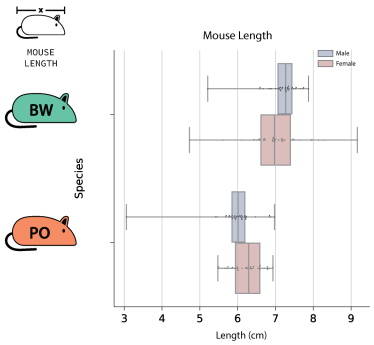


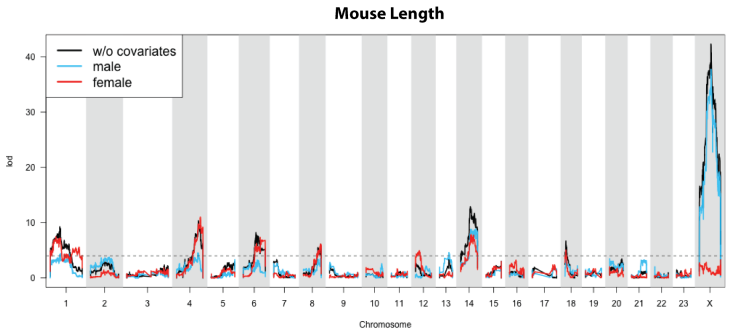


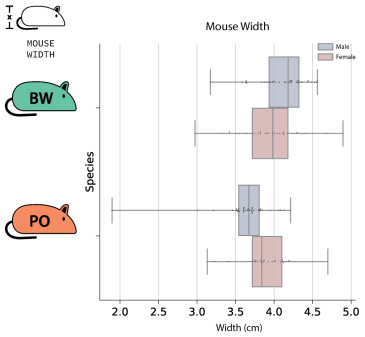


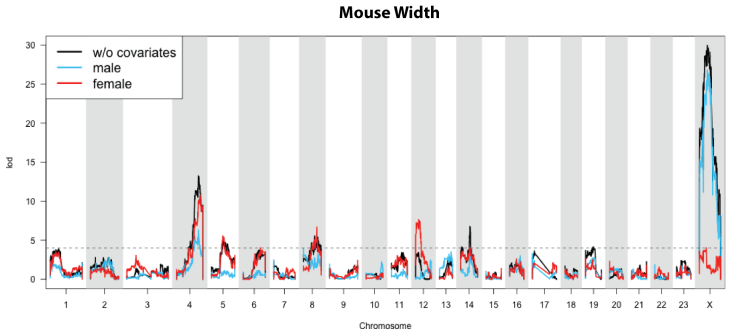


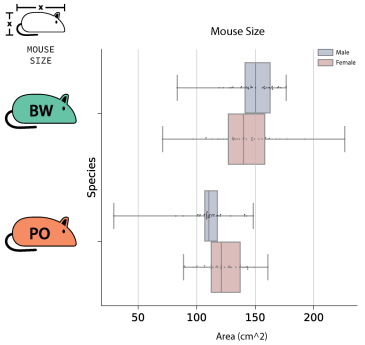
Mouse Details

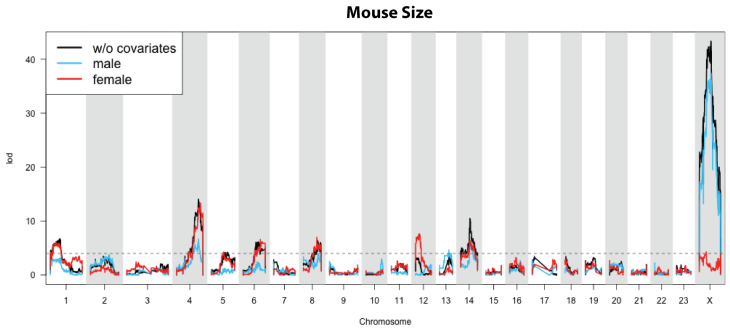




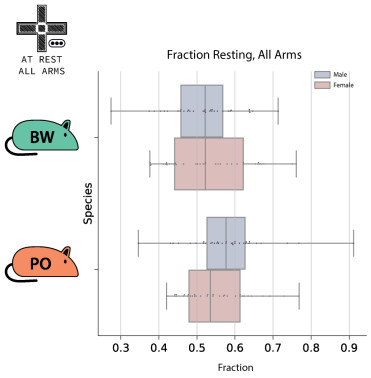


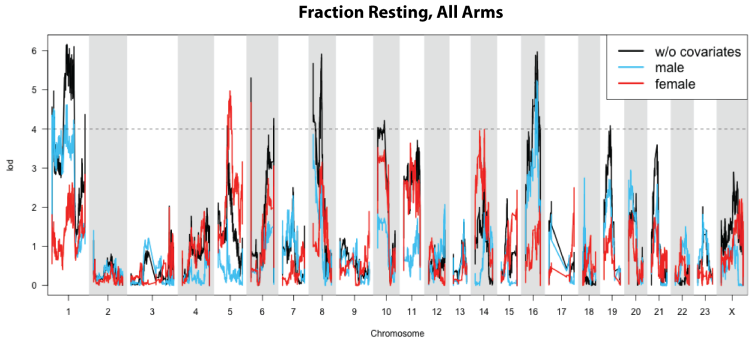


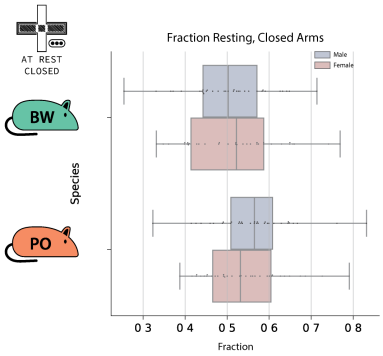


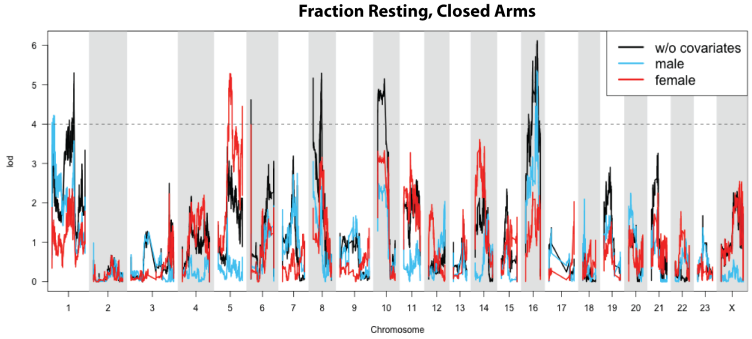


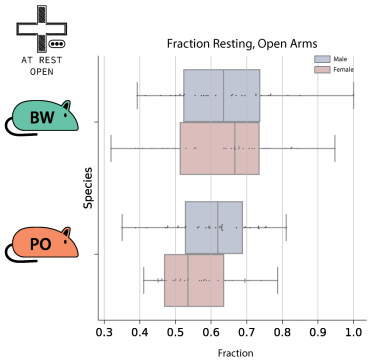
Resting

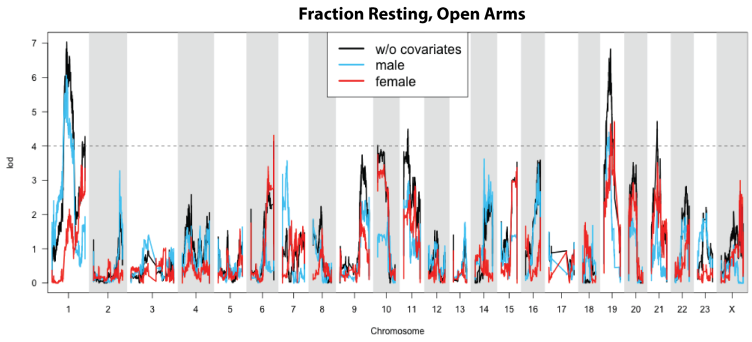

AT REST
ALL ARMS
BW
PO
Fraction Resting, All Arms
Male
Female
Species
0.3 0.4 0.5 0.6 0.7 0.8 0.9
Fraction


Fraction Resting, All Arms
w/o covariates
male
female
lod
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 X
Chromosome

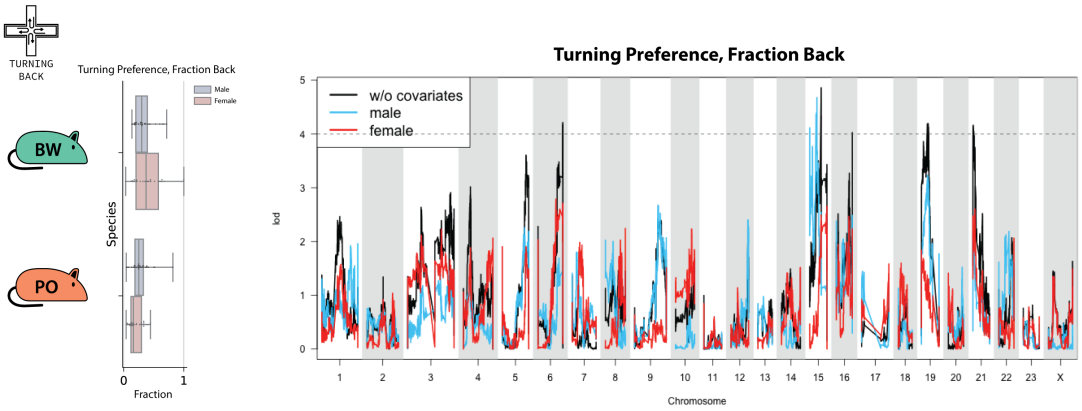

AT REST
CLOSED
BW
PO
Fraction Resting, Closed Arms
Male
Female
Species
0.3 0.4 0.5 0.6 0.7 0.8
Fraction

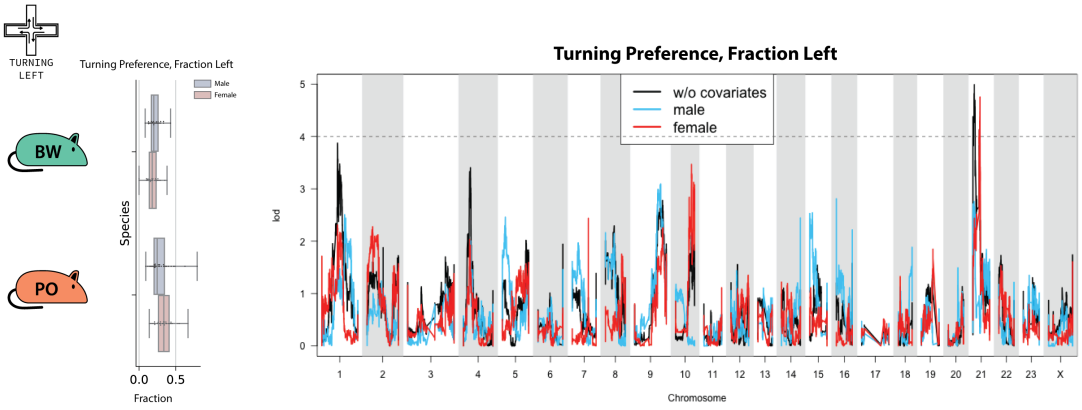

Fraction Resting, Closed Arms
w/o covariates
male
female
lod
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 X
Chromosome

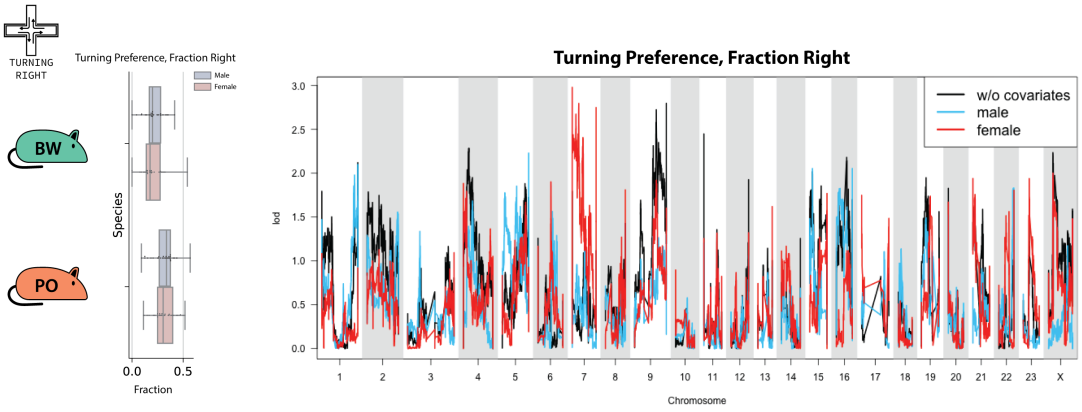

AT REST
OPEN
BW
PO
Fraction Resting, Open Arms
Male
Female
Species
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Fraction

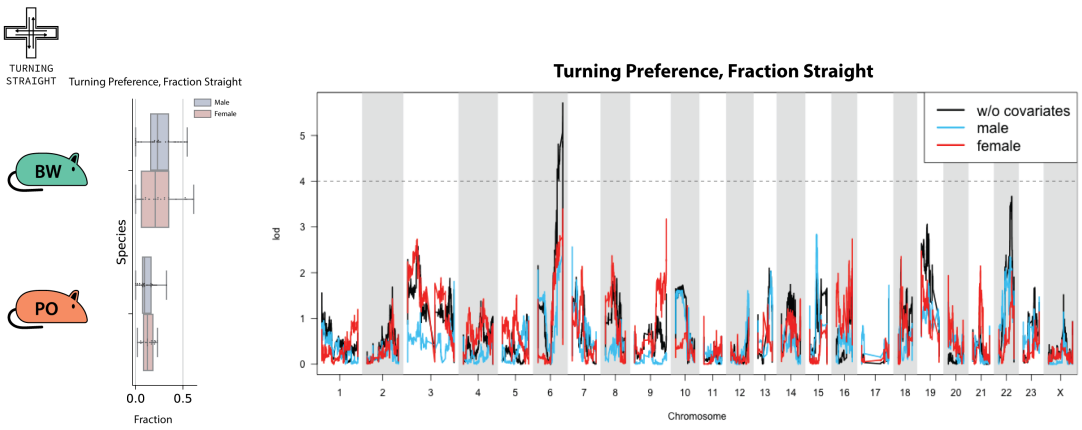

Fraction Resting, Open Arms
w/o covariates
male
female
lod
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 X
Chromosome

Turning Preferences

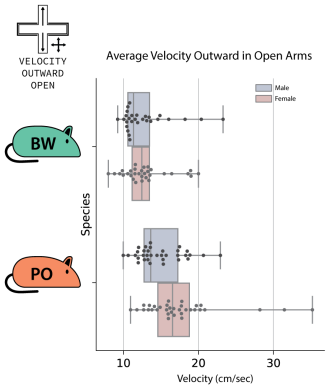


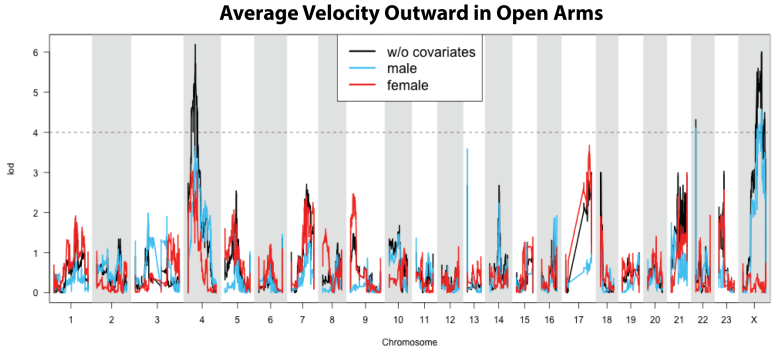


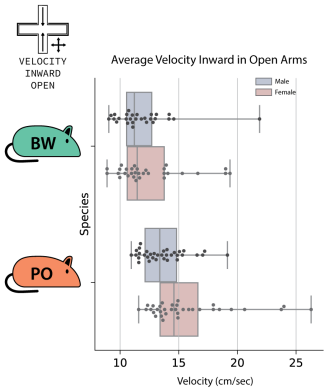


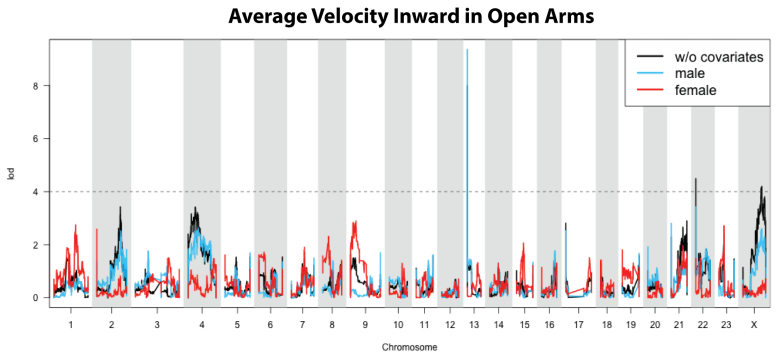


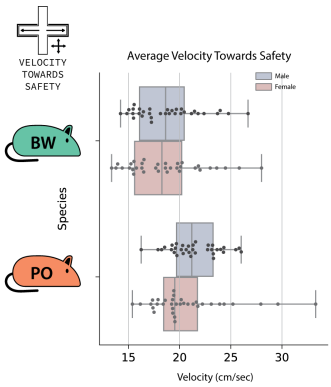
Velocity

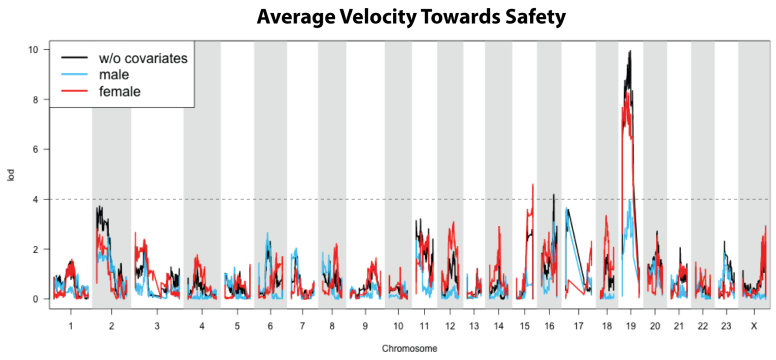

VELOCITY
OUTWARD
OPEN
Average Velocity Outward in Open Arms
Male
Female
BW
PO
Species
10
20
30
Velocity (cm/sec)


Average Velocity Outward in Open Arms
w/o covariates
male
female
lod
Chromosome


VELOCITY
INWARD
OPEN
Average Velocity Inward in Open Arms
Male
Female
BW
PO
Species
10
15
20
25
Velocity (cm/sec)


Average Velocity Inward in Open Arms
w/o covariates
male
female
lod
Chromosome


VELOCITY
TOWARDS
SAFETY
Average Velocity Towards Safety
Male
Female
BW
PO
Species
15
20
25
30
Velocity (cm/sec)


Average Velocity Towards Safety
w/o covariates
male
female
lod
Chromosome