GENETIC BASIS OF SOCIABILITY WITHIN DROSOPHILA MELANOGASTER

INVESTIGATING THE GENETIC BASIS OF NATURAL VARIATION IN SOCIABILITY WITHIN *DROSOPHILA MELANOGASTER*

By Arteen Torabi-Marashi, B.Sc. (Hons)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2023) Hamilton Ontario (Biology)

TITLE: Investigating the genetic basis of natural variation in sociability within *Drosophila melanogaster*

AUTHOR: Arteen Torabi-Marashi

SUPERVISOR: Dr. Ian Dworkin

Pages: xv, 147

Lay Abstract

Sociability is an individual’s tendency to associate with other individuals of the same species in a non-aggressive manner. Previous work has been conducted to evolve lineages of high and low social fruit flies (*Drosophila melanogaster*) using artificial selection. The main goal of this thesis was to integrate analyses of differential gene expression, transcript usage and population genomics to investigate the genetic architecture of sociability in *Drosophila*. I developed pipelines to analyze the sequencing data and was able to identify genes that are differentially expressed, transcripts that are differential used and regions of the genome under selection.

Abstract

Sociability is an individual's tendency to associate with conspecifics in a non-aggressive manner. Sociability can manifest in the formation of social groups that can reduce predation risk and increase feeding success. Studies of social behaviour in insects are typically through the lens of classically know social insects, however many insect species that have been long thought as non-social have been shown to exhibit social behaviour, in particular *Drosophila*. A previous experiment evolved lineages of high and low sociable fruit flies (*Drosophila melanogaster*) following 25 generations of artificial selection, after which RNA and DNA was extracted and sequenced. The main goal of this thesis was to integrate analyses of differential gene expression, transcript usage and population genomics to investigate the genetic architecture of sociability in *Drosophila*. I developed a pipeline to perform differential gene expression analysis by modelling gene expression using a generalized linear mixed-effect model. Here I found a total of 327 genes differentially expressed and 174 genes differentially expressed between the low and high sociable lineages. Next, I developed a pipeline to perform differential transcript usage analysis using a generalized linear mixed-effect model to model transcript usage. I found 619 genes to have transcripts with differential usage and 190 genes to have transcripts with differential usage between the low and high sociable lineages. Lastly, I developed a pipeline for population genomics to identify regions of the genome under selection. I identified genes that are likely under selection and the overlap between these genes and genes/transcripts found to be differentially expressed/used. Overall, I identified potential genes that are involved in the genetic architecture of sociability and can be further candidate tested.

Acknowledgements

There are numerous people who have helped me and supported me throughout my time at McMaster that I would like to thank. First, I'd like to thank my supervisor, Dr. Ian Dworkin. Thank you for the mentorship and support the past two years. Without your guidance I would not be in this position today, and I am grateful for all the help you have given me the past two years. I'd also like to thank my supervisory committee, Dr. Reuven Dukas and Dr. Ben Evans. Both have been amazing resources whenever I have had questions and have provided constant feedback to help me improve my work.

Thank you to Katie, Tyler, Amanda, and Brandon for welcoming me into the Dworkin lab and being great friends and helping me out whenever I got stuck or needed a colleague to bounce ideas off of.

Thank you to the graduate student community at McMaster, particularly my friends from the graduate office and our trivia group. You have all been amazing friends and have made my time at McMaster a wonderful experience.

I'd also like to thank my friends from back home for constantly checking in, catching up and seeing each other as much as we can. You all know how much I appreciate and value our friendships.

Thank you to my amazing parents for supporting me and being there for me for all these years. Thank you to my brother Arman, although we have lived together for most of our lives, being your roommate for two years was one of my favourite experiences. Thank you to the rest of my loving family as well for all the support.

Finally thank you to Nina. Thank you for all your encouragement and belief in me throughout the years. I wouldn't be here without your support.

Table of Contents

List of Figures and Tables

List of all Abbreviations and Symbols

DGE – Differential Gene Expression

DTU – Differential Transcript Usage

GWAS – Genome Wide Association Study

RDA – Regularized Discriminant Analysis

PCA – Principal Component Analysis

CMH Test – Cochran–Mantel–Haenszel Test

TIN – Transcript Integrity Number

CPM – Counts Per Million

GO – Gene Ontology

DE – Differentially Expressed

MA – Mean-Average

ESTs – Expressed Sequence Tags

MAPQ – Map Quality

SNP – Single Nucleotide Polymorphism

VCF – Variant Call Format

Declaration of academic achievement

The preceding artificial selection experiment (detailed in Scott et al. (2022)) was performed by Andrew Scott and designed by him, Reuven Dukas, and Ian Dworkin. Extraction of DNA and RNA was also performed by Andrew Scott. Figure 1 was created by Andrew Scott and Figure 2 was created by Reuven Dukas (adapted from Scott et al. (2022)). I modified and rewrote code from the pre-existing Dworkin lab population genomics pipeline written by Ian Dworkin, Tyler Audet and Katie Pelletier. All other analyses, statistical techniques, and generation of all other figures from the DNA and RNA sequences were performed by me.

Chapter 1: Social ~~Butter~~Flies: What is sociability and how do we study it?

Sociability

Sociability is an individual's tendency to associate and engage in non-aggressive activities with other individuals of the same species (conspecifics). Sociability manifests in the formation of social groups that can result in reduction of predation risk, reduced time spent to find resources and increased feeding success from cooperative hunting (Krause & Ruxton, 2002; Ward & Webster, 2016). We see the formation of social groups and sociability in various species of insects, fish, birds and primates and their formation of social groups can vary in regard to group size, the timing of when individuals form groups, and the formation of sex or age specific groups (Nordell & Valone, 2014). There are a number of ways to quantify sociability such as social integration (Kajokaite et al., 2022), individuals choosing to join each other at a food patch (Scott et al., 2018), measures of the proportion of time spent close to conspecifics (Ward & Webster, 2016), average distance from conspecifics (Ward & Webster, 2016), or in humans, quantifying a sociability score based on a survey assessment (Bralten et al., 2021).

The evolution of sociability in these various species suggests that an individual has a higher fitness when in a group rather than if it would be alone. We do see positive correlation between social behaviour and fitness in various mammal species (Snyder-Mackler et al., 2020) such as in *Cebus capucinus* (white-faced capuchin monkeys; Kajokaite et al., 2022) and baboons (Silk et al., 2003; Silk et al., 2010). Kajokaite et al. (2022) measured social integration of white-faced capuchin monkey groups by incorporating data related to the behaviours of grooming, joining conflicts, and foraging in proximity of others and modelled survivorship over 18 years as a function of the individual social integration metric. They found that females that engaged more with their group and with other females had a higher survivorship (Kajokaite et al., 2022). Within baboons (Genus: *papio*) we see evidence of this positive correlation in different species. In the wild savannah baboon (*Papio cynocephalus*), an increase in sociality of females led to an increase in infant survival (Silk et al., 2003). The measure of sociality here was a sociality index, calculated based on the amount of time a female was within 5 metres of an adult conspecific, grooming other adults and being groomed by other adults (Silk et al., 2003). Within the female chacma baboons (*Papio hamadryas ursinus*), females that displayed strong social bonds with other females were associated with living longer than females who formed weaker bonds with other females (Silk et al., 2010). Social bonds here were measured by creating a composite sociality index for a given female, which represents the rate of interactions with others, grooming, grooming initiation, and grooming duration, all relative to their averages for all females. While those are not explicitly sociability, we do see sociability positively associated with fitness in masai giraffes (*Giraffa camelopardalis tippelskirchi*). Female giraffes were found to have higher survival when in groups with a higher number of other females than the average (Bond et al., 2021).

Sociability Studies

Sociability is commonly studied in group living species such as mammals and social insects. In humans, many studies examine low sociable individuals and related traits such as loneliness (Gao et al., 2017) and social isolation (Holt-Lunstad et al., 2015) that are associated with low sociable individuals. This reduced level of sociability and social interaction is associated with mental health related disorders such as bipolar disorder (Tiğli Filizer et al., 2016), major depressive disorder (Kupferberg et al., 2016; Saris et al., 2017) and schizophrenia (Green et al., 2015). We also see studies that link low sociability in mice, with autism observed in humans (Brodkin, 2007). Other studies in humans aim to understand the role of the hormone oxytocin, with variation in the oxytocin receptor gene *OXTR* associated with variability in human sociability and social behaviour (Bakermans-Kranenburg & van Ijzendoorn, 2014; Pearce et al., 2017).

Sociability and social behaviour are also commonly studied in insects, as there are numerous social species of insects especially those that demonstrate eusociality. Eusociality refers to a societal group's possession of three traits: cooperative care of the young; a reproductive division of labour; and an overlap of at least two generations in life stages capable of contributing to colony labor (Wilson, 1971). Many of these classically known social species exhibit social behaviour such as in ants (Chapman et al., 2011), termites (Higashi et al., 2000), and bees (Amsalem et al., 2015; Plateaux-Quénu, 2008). We can also look at eusocial insects from an evolutionary perspective. The independent evolution of eusociality across *Hymenoptera* proves a powerful tool to investigate the evolution of this behaviour (Wilson & Hölldobler, 2005). Here we see work that investigates eusocial and non-social species of bees, looking at genes that are potentially involved in this evolution (Jones et al., 2023; Woodard et al., 2011).

Sociability in Drosophila

While a majority of studies in social insects are through the lens of the classically known social species (ants, termites, and bees), recently many insect species that have been long thought as non-social have been shown to exhibit social behaviour (Costa, 2006; Prokopy & Roitberg, 2001). Specifically, we can look at the fruit fly (*Drosophila melanogaster*) which has been previously thought to be non-social. Here we see that they do in fact form social groups (Dukas, 2020; Schneider et al., 2012) as well as other social behaviours like social learning (Sarin & Dukas, 2009) and a group response to danger (Ferreira & Moita, 2020). Sociability is prevalent and important in many species; however we have a limited knowledge in relation to its genetic architecture, or how genes (and what genes) contribute to the observed phenotypic variation of sociability.

How Sociability translates across species

While the definition of sociability is straightforward, it is much more complicated to think how sociability translates across species. Is social or anti-social behaviour in an individual the same when you look at two different species? Or are there subtle differences between the two? Even measuring sociability or social behaviour varies when studying different species. Johnsson et al. (2018) studied an aspect of social behaviour, social reinstatement, by removing a chicken from conspecifics and recording how long it took for the individual to reinstate itself with conspecifics. Another study looked at early life social conditions in *Bombus impatiens*, in which they raised a single generation of bumblebees in three early life conditions (Wang et al., 2022). The three conditions were colony-raised, group-raised or isolated and RNA was extracted for differential gene expression (DGE) analysis following behaviour assays (Wang et al., 2022). In *Apis mellifera*, assays were performed to determine if the behaviour of an individual bee was unresponsive, guard-like or nurse-like after RNA was extracted for DGE (Shpigler et al., 2017). A study in humans surveyed individuals from the UK biobank with questions related to their social behaviour, and then a genome wide association study (GWAS) was performed (Bralten et al., 2021). Are all of these different methods capturing the same latent states in individuals? That is, are we potentially capturing similar genes or gene networks that are underlying this behaviour? Or does capturing different measures of social behaviour make it difficult to compare between species?

Artificial Selection of Sociability

Looking at fruit flies, we can ask what genes and the regulation of which genes are responsible for sociability. While we do observe sociability in fruit flies, we have a limited knowledge in relation to its underlying genetic basis. Previous work by Scott et al. (2022) worked on quantifying and artificially selecting for high and low sociable fruit flies. A sociability arena (Figure 1.1) was designed and constructed where flies would be able to congregate with one another on food patches and their behaviour was documented (Scott et al., 2022). If there was a low number of flies in a section of the arena, they were selected for low sociability in the next generation and conversely, if there were a larger number of flies in a section, they were selected for high sociability in the next generation. The flies were selected upon for a total of 25 generations after which RNA and DNA was extracted from pooled adult heads and sent for sequencing (see detailed methods below). Artificial selection is a beneficial experimental tool that has been used in the past to select on behaviour in various species (Doyle & Talbot, 1986; Dukas et al., 2020; Hämäläinen et al., 2022; Ramos & Gonçalves, 2019; Siegel, 1972; Vega-Trejo et al., 2020). Studies investigating sociability remain underrepresented, and there are little to no studies investigating the behaviour through the lens of artificial selection.

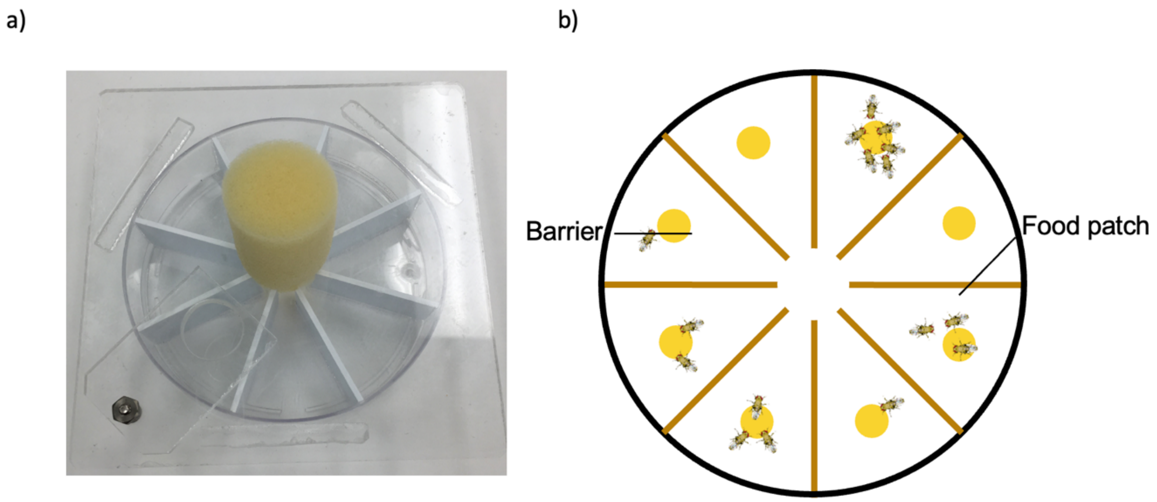

Figure 1.1. Sociability Arena A) Image of the arena. B) Schematic of the arena with flies and indication of barriers and food patches for the flies. Figures created by Andrew Scott for Scott et al. (2022).

How we can use Sociability to address our questions

One of the more interesting aspects from Scott et al. (2022) was that we saw a response to selection (Figure 1.2). Typically, with an artificial selection experiment, say for size, selection occurs on an individual and their individual contributions. However, in this study we are capturing a group response, while the selection is happening on an individual. Given that we see such a clear response to the selection, we are able to study evolutionary changes in this behaviour largely absent of other differences. For example, in studies that look at the differences between solitary and eusocial bees, there are many inherit differences between the species. While the behavioural difference between the two may be captured, other unrelated and confounding factors may be present. The same can be thought of in studies that deal with humans. Many measures of sociability are investigating and capturing effects associated with disorders that may have other contributing factors that don't involve just the genetic component (socioeconomic factors, life history, environment). While it can be argued that this response is not necessarily the same sociability we see in other species and can be potentially group forming or social aggregation, we are capturing a specific aspect of sociability. This allows us to investigate only this aspect of the behaviour, absent of other factors, which can prove to be a powerful tool to study the underlying genetic architecture of sociability and social behaviour.

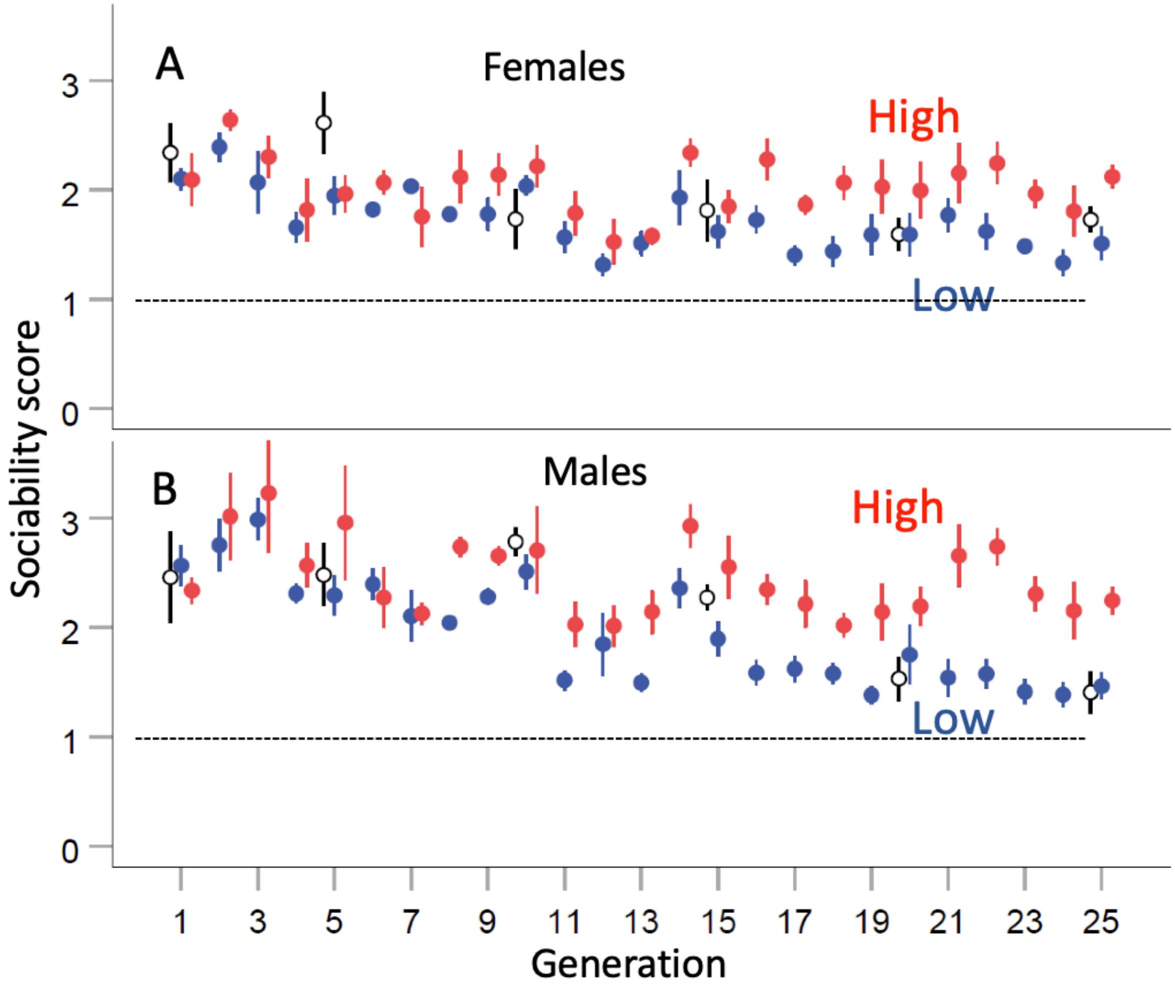

Figure 1.2. Mean ± SEM sociability scores over 25 generation in females (A) and males (B) of the low (blue), high (red) and control (unfilled) lineages (n = 4 lineages for each of the 3 treatments). Note that we quantified sociability in the control lineages only every 5 generations. Values significantly above 1 (dashed lines) indicate significant sociability. Data from Scott et al. (2022).

Motivation

With the sequencing data from the previously mentioned artificial selection experiment by Scott et al. (2022) we can investigate the underlying genetic architecture of the trait and provide valuable insight into a system that remains underrepresented in the literature. In total, we will have RNA and DNA sequences that we can use to begin to start to understand the genetic architecture and underpinnings of social behaviour. Population genomics (with the DNA data) can help us understand evolutionary changes associated with the change in behaviour that may be captured within the genomes, and differential gene expression and transcript usage (with the RNA data) can help us understand if there are any genes or transcripts being differentially expressed or used across selection treatments. This study provides us with a large experimental design and sample size, which allows us to accurately and appropriately model gene expression and transcript usage, which will be covered in depth in Chapter 2.

Thesis Objectives

With this large genomic dataset, we are able to address the questions of what genes, or the regulation of which genes are responsible for the observed variation in sociability. The overall objective is to use genomic approaches to investigate the genetic architecture of sociability and social behaviour in *Drosophila*. With this we can identify potential genes that mediate the observed response to selection. Specifically, we can use the RNA sequencing data to identify the set of genes that are being differentially expressed and transcripts that are being differentially used between the low and high sociable lineages. With the DNA sequencing we can use a population genomics approach and investigate the evolution of the behaviour across the genome and identify potential regions under selection and whether or not these regions overlap with genes previously identified in either the DGE or DTU analysis. Throughout the analyses, potential candidate genes may be identified that can be further tested. We can compare our set of genes with orthologous genes linked to social behaviour in other species and see if there are in fact common genetic underpinnings across these different species or if the underlying genetic architecture is specific to their respective species.

Chapter 2: RNA sequencing analyses of artificially selected lineages

Introduction

As discussed in chapter 1 we have a limited knowledge on the genetic architecture of sociability, especially through the lens of *D. melanogaster*. Other studies have used differential gene expression as a method to investigate gene regulation of social behaviour in bees (Shpigler et al., 2017; Wang et al., 2022). Building on the work of Scott et al. (2022), we have obtained mRNA sequencing reads from 142 samples following artificial selection with these samples coming from pooled heads of 16 adults across artificially selected treatments (low, control, and high), across sex, and across experiential contexts. This study provides us with, to our knowledge, the largest RNA sequencing experiment that investigates sociability in *D. melanogaster*.

RNA sequencing can be a powerful tool to investigate how genes are being expressed differentially in different treatment conditions. We can also utilize transcript level information to identify how transcripts are being expressed differentially between treatments. In this chapter, we perform differential gene expression (DGE) and differential transcript usage (DTU) analysis. DGE lets us identify if a given gene is expressed at higher or lower expression values in differing treatments, and if so, how large of an effect is this difference in expression. A DTU analysis tests for proportional differences in the transcript expression of a gene across conditions (Tekath & Dugas, 2021). DTU lets us identify if a given transcripts proportion of usage is increasing or decreasing across conditions, relative to other transcripts of that gene. We can then identify how many transcripts are being differentially used and how large of a difference in usage there is across treatments. There is another analysis regarding transcript expression, differential transcript expression (DTE) analysis which examines an individual transcripts expression across treatments. Typically, if an individual transcript is exhibiting differential expression across treatments, it is either due to a change in a genes expression across treatments or a change in a genes relative usage of its transcripts, both of which are captured with DGE and DTU (Soneson et al., 2016). Because of this reason, we are performing a DTU analysis alongside our DGE analysis, rather than a DTE analysis.

There are many different pipelines and software tools available for both DGE and DTU analyses. At nearly each step of the pipeline, a different tool can potentially be swapped in and out to achieve similar goals. With the vast number of tools, it may be difficult to initially parse through all the available tools to determine what works best for a given experiment. For a given RNA sequencing analysis pipeline you may encounter different trimming, mapping, DGE and DTU tools as well as different methods of extracting estimates. One example of this can occurring during the transcript mapping stage. There are tools that calculate transcript abundance like salmon (Patro et al., 2017) and kallisto (Bray et al., 2016), that are alignment-free and identify what transcript or loci the sequencing reads originate from, rather than a base to base alignment with a reference genome (Bray et al., 2016; Patro et al., 2017; Srivastava et al., 2016). There are also aligners like STAR (Dobin et al., 2013) that align reads base to base to a reference genome and produce gene-level counts. STAR is able to take a read from mRNA sequencing and match it to a reference genome, by first searching for the longest contiguous sequence in a read that matches with the reference and then searching again with the unmapped portion of the read for the longest contiguous sequence that matches with the reference (Dobin et al., 2013). This process continues until the entire read is accounted for (Dobin et al., 2013). Both have pros and cons, namely that salmon and kallisto are quicker and less memory intensive, and STAR has the ability to identify novel transcripts. Choosing one of the two methods is appropriate for most workflows, but a more inclusive approach is to use both (in separate workflows) and compare the two methods to ensure consistency in the results. Ideally, these two methods would largely agree with one another, but given the fact that salmon and kallisto are not able to identify novel transcripts there may be cases where salmon and kallisto incorrectly map transcripts which would lead to different results than a workflow using STAR.

Once the sequencing data has been processed and mapped to count data, DGE and DTU analyses can be performed. For these steps there are numerous tools available to choose from, with their own use cases. A majority of the tools that perform DGE model gene expression using linear models that are fit for every gene using either a gaussian (limma; Ritchie et al., 2015) or a negative binomial (DESeq2; Love et al., 2014; edgeR; Robinson et al., 2010) probability distribution. However, if a study has a complex experimental design, it may be appropriate to model gene expression using generalized linear mixed-effect models. A larger experimental design might mean that there are certain random effects that need to be fit in order to account for statistical variance in the design that may not be biologically relevant. This is not possible when using linear or generalized linear models, however using linear mixed-effect models solves this problem. Tools like limma-Voom also account for technical variation from sequencing (Robinson et al., 2010), and use an empirical bayes approach to shrink and account for these sources of variation (McCarthy et al., 2012). DESeq2 and edgeR also estimate the dispersion parameter, or the variation between replicates, when fitting the negative binomial distribution (Love et al., 2014; McCarthy et al., 2012). Dispersion can be difficult to estimate in experiments with lower sample sizes and number of replicates, as the estimated dispersion for each gene is highly variable and results in imprecise accuracy of differential expression testing (Love et al., 2014). DESeq2 assumes that genes with similar average expression have similar dispersion and therefore estimates dispersion using a method that shares information across genes (Love et al., 2014). In this approach, first gene-wise dispersion is estimate for each gene, and then a curve is fit between the gene-wise dispersion estimate and average expression (Love et al., 2014). Next, an empirical bayes approach is used to shrink the gene-wise dispersion estimates toward the fitted line, and the result value is used for the dispersion estimates (Love et al., 2014). With large sample sizes this approach is not necessary (Love et al., 2014) as the estimation of gene-wise dispersion parameters will be more accurate, so there is no need to share information across genes with similar average expression levels to estimate dispersion. Similar to DGE, DTU tools also use similar approaches to account for this source of various (Anders et al., 2012; Nowicka & Robinson, 2016).

In our experiment we had a large sample size as well as 4 independent lineages in each selection treatment that should be appropriately fit as random effects as there is variation between lineages which is statistically relevant, but not biologically relevant to the objectives of our study. We decided to implement our own method of DGE and DTU analysis using linear mixed-effect models. To do so, we used glmmTMB (Brooks et al.,

2017) to model gene expression or transcript usage per gene or transcript, for DGE and DTU respectively. Using this approach, we were able to fit random effects and build our models appropriately. With the flexibility of glmmTMB, we were able to implement model runs that uses similar approaches to gene filtering, count normalization and outputting effect sizes of $\log_2$(cpm) as those seen in DESeq2, limma-Voom and DEXseq, but in the framework of a linear mixed-effect model. Importantly, we referred to the vignettes of DESeq2 and limma-Voom and replicated how counts were normalized or variance stabilized, as well as extract normalization factors to input into glmmTMB as offsets (Love et al., 2014; Ritchie et al., 2015). By implementing separate runs of DGE that used either the gaussian or negative binomial probability distribution, we were able have a two pronged approach that is representative of methods used in the literature. With both implementations, we are able to compare the results to ensure that the results we observe are consistent with one another.

In large, a motivation for this implementation is that with the cost of sequencing decreasing, these larger scale and more complex experimental design RNA sequencing studies are becoming more and more common (Metzker, 2010; Pareek et al., 2011). While a lot of the commonly used tools are appropriate and highly useful for studies with simpler experimental designs, they do not hold as true for more complex experimental designs. Here, we present a method that can be implemented and adapted to most workflows with large experimental designs and sample sizes.

Objectives

In this chapter specifically we ask the question of what genes and transcripts are being differentially expressed and used between the high and low sociability lineages previously generated by Scott et al. (2022). We can also compare the set of genes identified with orthologous genes in other species that are associated with social behaviour, to see if there are common genes that arise across the literature, or if the underlying genetic basis is specific to their respective species.

Methods

Artificial Selection

We previously applied artificial selection on sociability (Scott et al., 2022). For each selection treatment, we had 4 independently evolving lineages (4 low sociability lineages, 4 high sociability lineages, and 4 control lineages). Each generation, we quantified sociability in 12 groups of 16 females, and 12 groups of 16 males, from each of the 4 low and 4 high sociability lineages. To quantify sociability, we placed each group of 16 flies inside a sociability arena, which had 8 equal sized compartments, each containing a food disc (Fig. 1 in Scott et al., 2022). Flies could move freely among the compartments for 90 minutes, after which we blocked the fly passage and recorded the number of flies in each compartment. From this record, we calculated the sociability score as the variance over mean number of flies in each compartment (See Fig. 1 in Scott et al., 2018). We then selected 4 flies from each arena. For the low-sociability lineages, we selected flies from compartments with the lowest numbers of individuals, while for the high-sociability lineages, we selected flies from compartment(s) with the highest number of individuals. For the 4 control lineages, we randomly selected 4 flies from each of the 12 groups of 16 same-sex flies per lineage. Owing to time constraints, we quantified sociability in the control lineages only every 5 generations. We used the 48 males and 48 females from each lineage to generate the next generation of individuals. After 25 generations of selection, the high-sociability lineages showed sociability scores about 50% greater compared with low-sociability lineages. (Scott et al., 2022; Fig. 1.2).

In generation 26, we collected adult fly heads for gene expression analysis. We had 2 experience treatments. In the sociability arena treatment, we placed groups of 16 same-sex flies in the sociability arenas prior to their collection for gene expression. This provided the flies with the social dynamics experienced during their evolutionary history under artificial selection. When placed in the sociability arenas, flies initially engage in exploration and frequent contacts with other flies (Scott et al., 2022). We presumed that such social interactions would affect the expression of pertinent genes. In the vial treatment, we just moved groups of 16 same-sex flies into fresh vials, so these flies did not experience the sociability arenas. After 20 min, we rapidly transferred each group of 16 individuals into a 1.5 mL tube and submerged it in liquid nitrogen. We had 3 replicates per lineage x 12 lineages x 2 sexes x 2 experience conditions for a total of 144 samples. We later separated the flies' heads and extracted RNA.

RNA Extraction and Sequencing

We extracted RNA using MagMAX -96 Microarrays Total RNA Isolation Kit (Thermo Fisher). We checked purity using a Nanodrop (Thermo Fisher) spectrophotometer and checked quantification using Denovix Fluorometer (Denovix) and Qubit RNA high sensitivity assay kit (Thermo Fisher). We sent the samples to Génome Québec (Centre d'expertise et de services, Génome Québec) for library preparation and sequencing. Library preparation used NEBNext dual multiplex oligoes, and sequencing was done using an Illumina Novaseq 6000 S4, with a 100 bp paired-end sequencing technology. One sample was rejected in the quality control check for poor quality and another sample was rejected for low quantity of RNA, so 142 samples were sequenced. A total of approximately 6.2 billion read clusters were generated with an average of 44 million read clusters per sample.

Read Processing and Mapping

All computational analysis was performed using the Digital Research Alliance of Canada (formerly Compute Canada; www.alliancecan.ca). We checked quality of reads using FASTQC (v0.11.9, Andrews, 2010) and MultiQC (v1.12, Ewels et al., 2016), which assessed adapter content, per sequence quality scores and GC content. All samples had a mean Phred score value of > 35. We assessed transcript integrity number (TIN) using RSeQC (v4.0.0, Wang et al., 2012), and all but two samples had median TIN scores > 60, with those two having a median TIN score of 49 and 59. As such, we included all samples in the analysis. We trimmed adapters using trimmomatic (v0.36, Bolger et al., 2014), with both leading and trailing set to “3” and run parameters set to “MAXINFO:20:0.2”. We removed reads shorter than 36 bp from the sample. Following trimming, we once again used FASTQC and MultiQC to confirm adapters were trimmed while maintaining high quality sequence. We mapped reads to a reference transcriptome of *D. melanogaster* from Flybase (version r6.38, Gramates et al., 2022) using Salmon (v1.4.0, Patro et al., 2017) with decoys which produced counts of transcripts per sample. To use Salmon, we first generated an index file from a list of decoys, a reference transcriptome, and a reference genome (version r6.38, Gramates et al., 2022), which we then used for mapping. We also separately used the splice-aware aligner STAR (v2.7.9a, Dobin et al., 2013) to map reads to a reference genome, which produced gene level counts (Figure 2.1, Table A1).

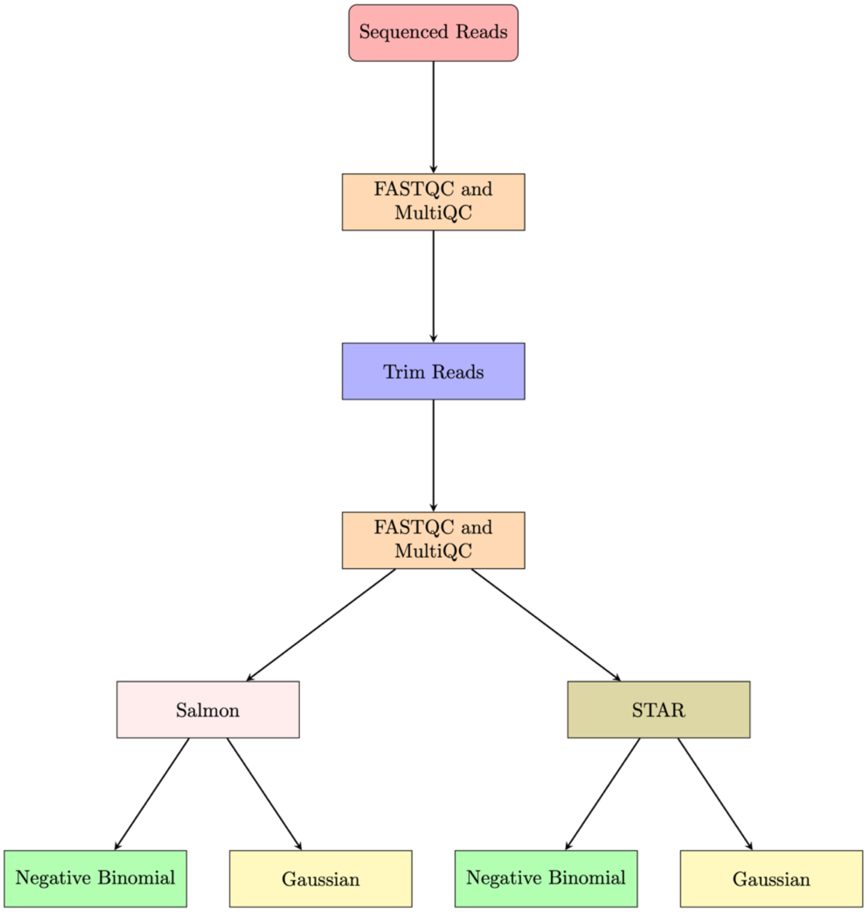

Figure 2.1. Flowchart detailing steps and software used for differential gene expression (DGE) analysis. The flowchart details steps from raw sequenced reads until fitting models onto gene counts. Each branch point after the second "FASTQC and MultiQC" step is independent of one another.

Principal component analysis (PCA)

Counts were imported into R (v4.2.0, R Core Team, 2022) using tximport (v1.24.0, Soneson et al., 2016). Count data was normalized using the "vst()" function from DESeq2 (v1.36.0, Love et al., 2014) which performs a variance stabilizing transformation. In the "vst()" call, we set "blind = FALSE", providing a design matrix consisting of sex, experience, selection and the interaction between selection and sex. We also used “nsub = 5000” to filter for only the top 5,000 most variable genes. To visualize the PCA results, we used the function “plot_pca()” from RNAseqQC (v0.1.4, DeLuca et al., 2012) with “nfeats = 500” to plot the top 500 most variable genes.

Regularized Discriminant Analysis (RDA)

We performed RDA to examine broad, shared features of transcriptome-wide responses. Given the potential influence of small individual changes in gene expression across many genes substantially modulating sociability, we also explored a genome-wide approach to evaluate the degree of shared response. If there is a high degree of shared response (i.e. similar sign of effect across treatments) across genes, then we should be able to predict specific lineages, given a training set of all other lineages. We used the same normalized counts as from our PCA analysis (see above) and we modified the code of the function “plot_pca()” from the RNAseQC package (DeLuca et al., 2012) to extract the top 500 most variable genes. We then split our data into a training set and a test set. The test set contained only samples from the specific lineage we aimed to predict, and the training set contained all other samples. We ran two separate implementations, one with the classifiers set to only selection and the other with both sex and selection included in the classifiers. We then used the function “rda()” from the package klaR (v1.7-2, Weihs et al., 2005) on our training set with “crossval = TRUE” and “fold = 10” to perform a 10-fold cross-validation. Following this we then used the function “predict()” on both our training set and test set.

Differential Gene Expression (DGE) Analysis

To filter out low expressed genes, we used the “filterByExpr()” function from edgeR (v3.38.4, Robinson et al., 2010). We removed genes that had lower than 0.3 counts-per-million (cpm) in at least 8 samples. CPM is used here to avoid overrepresentation of genes that are expressed in larger libraries over genes expressed in smaller libraries, as detailed in Chen et al. (2016). From a total of 13,701 genes, we removed 2,176 genes, leaving 11,525 genes. Counts were used directly for the negative-binomial GLMM. For the gaussian mixed models, counts were normalized and variance stabilized in the form of $\log_2$(cpm) using the “voom()” function in the limma package (v3.52.4, Ritchie et al., 2015). As we needed to incorporate random effects into our generalized linear models, we used glmmTMB (v 1.1.4, Brooks et al., 2017), which allowed us to fit the appropriate model. The package let us model our data similarly to limma-Voom using a Gaussian distribution and modelled $\log_2$(cpm) per gene.

The full model used lane, sex, experience (sociability arena or vial), and selection as main effect terms (Equation 1). We included all 2nd order interactions between selection, sex, and experience. Lineage was modelled as a random effect nested within selection treatments. Variation for sex and experience was allowed to vary by lineage (i.e. random “slopes” for sex by lineage). The full model in glmmTMB syntax is:

$$\begin{aligned} log_2 cpm \sim lane + sex + expMatched + selection + sex\!:\!expMatched \\ + selection\!:\!expMatched + selection\!:\!sex \\ + diag(0 + sex + expMatched \mid selection\!:\!lineage) \end{aligned}$$

$$y_i \sim N(\mu = \beta_{0[j]} + \beta_1 x_{1i} + \beta_{2[j]} x_{2i} + \beta_{3[j]} x_{3i} + \beta_4 x_{4i} + \beta_5 x_{2i} x_{3i} + \beta_6 x_{4i} x_{3i} + \beta_7 x_{4i} x_{2i}, \sigma^2)$$

With (co)variance due to lineage:

$$\begin{pmatrix} \beta_{0[j]} \\ \beta_{2[j]} \\ \beta_{3[j]} \end{pmatrix} = MVN \begin{pmatrix} 0 & {\sigma^2}_{\beta_0} & 0 & 0 \\ 0, & 0 & {\sigma^2}_{\beta_2} & 0 \\ 0 & 0 & 0 & {\sigma^2}_{\beta_3} \end{pmatrix}$$

where y_i = $\log_2$(cpm), x_1 = lane, x_2 = sex, x_3 = expMatched and x_4 = selection.

If the model failed to converge for a given gene, we adjusted the model to fit a slightly less complex random effect, while keeping all other terms identical. The adjustment to the random effect was to drop experience, such that the random effect was now: $diag(0 + sex \mid selection{:}\, lineage)$. We chose to drop experience because it had minimal influences on model fits for the reduced model. Importantly, checks with fitting models where covariance between the sexes across lineages was set to 0 had minimal impacts on model estimates (and associated uncertainties) for coefficients of interest (selection treatment and sex) for this study. For our specific contrasts and downstream analyses, we utilized estimated marginal means (emmeans) and associated contrasts from model fits using the emmeans package (v1.8.1, Lenth, 2022). Given that most studies of differential gene expression use modeling tools like limma-Voom and DESeq2, we confirmed that our model estimates were similar for the simplified parameterized models. Comparison of contrasts from our glmmTMB model fit to limma-Voom computed estimates (lowest r = 0.896, CI = 0.892 – 0.899; Figure 2.2, Table 1). While the approach we use is computationally slower than linear model fits using limma-Voom or DESeq2 (albeit still quite fast), it has the advantages and flexibility enabled by modern mixed modeling approaches (Brooks et al. 2017).

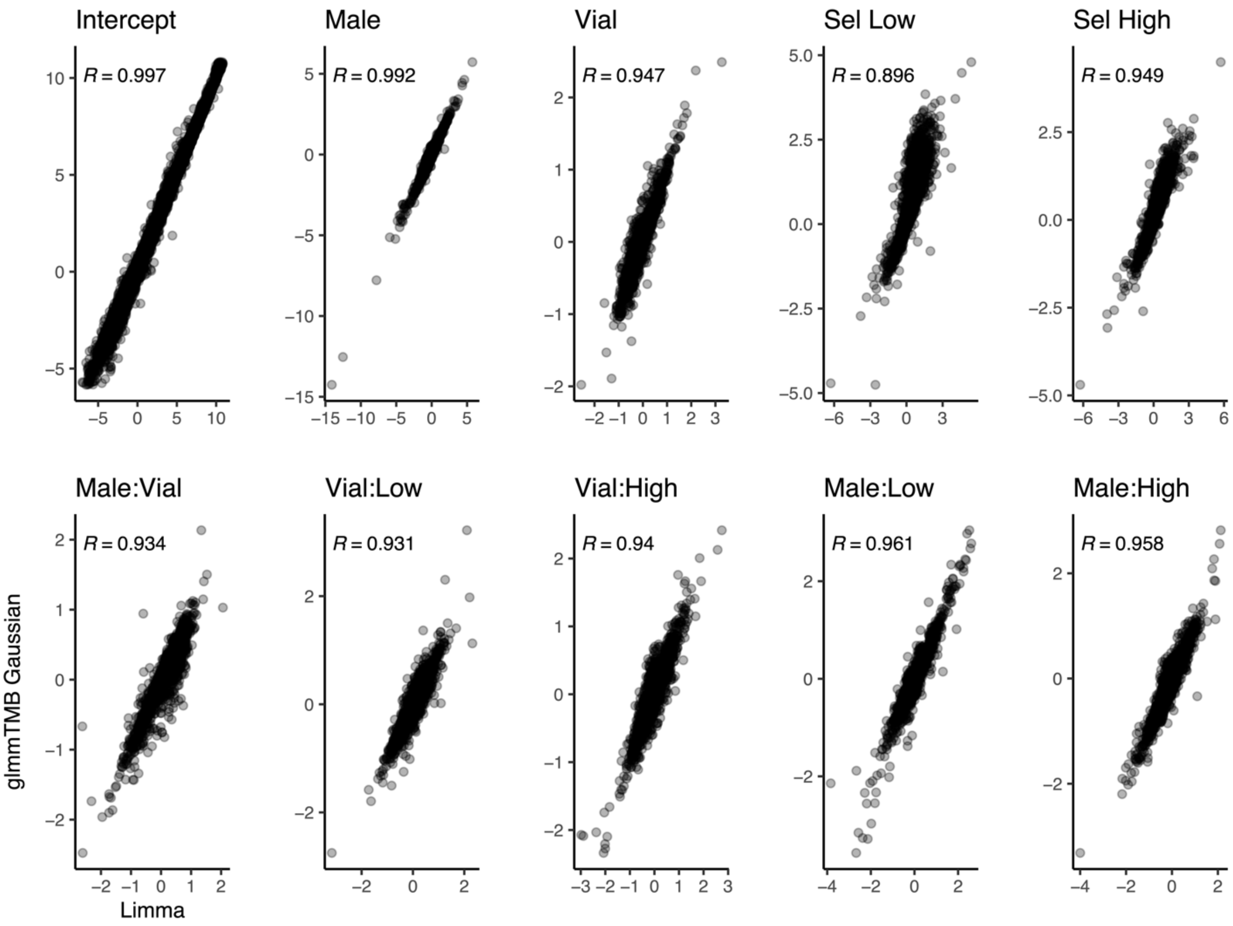

Figure 2.2. Estimate comparisons between limma-Voom and glmmTMB (following a gaussian distribution) estimates for various coefficients of their respective model. Both the x and y axis estimates are in $\log_2$(cpm). The difference between the two models is that lineage is a main effect in limma-Voom while it is a random effect in the glmmTMB run. Both estimates are $\log_2$(cpm). From top to bottom and left to right the terms we have are intercept (R = 0.99), sex male (R = 0.99), experience vial (R = 0.95),

selection low (R = 0.90), selection high (R = 0.95), male:vial interaction (R = 0.93), vial:low interaction (R = 0.93), vial:high interaction (R = 0.94), male:low interaction (R = 0.96) and male:high interaction (R = 0.96).

Table 2.1. Table of Pearson correlation coefficient and corresponding confidence intervals of estimates between limma-Voom and glmmTMB (Gaussian Distribution) by model term.

Model Term	r	Lower	Upper
Intercept	0.99699	0.99688	0.99709
Male	0.99192	0.99162	0.99220
Vial	0.94699	0.94507	0.94884
Selection Low	0.89554	0.89187	0.89910
Selection High	0.94873	0.94688	0.95053
Male:Vial	0.93431	0.93195	0.93659
Vial:Low	0.93121	0.92874	0.93359
Vial:High	0.93999	0.93783	0.94208
Male:Low	0.96065	0.95922	0.96204
Male:High	0.95805	0.95652	0.95953

The approach used by limma-Voom has been shown to generate comparable results to approaches based on the negative binominal distribution on “raw” count data, especially for experiments with sufficient sample sizes. However, as we re-implemented such modeling strategies, but in the context of generalized linear mixed models, we wanted to confirm this.

As such, we also fit models in glmmTMB using the raw counts as response variables, utilizing a natural log link and the negative binomial distribution. We obtained the normalization factors from DESeq2 package by using the function “estimateSizeFactors()” and extracted them using the function “normalizationFactors()”. We then used those extracted normalization factors as model offsets. To account for over-dispersion, we used the quadratic parameterization for the variance, “family = nbinom2()”, which specifies the variance as $V = \mu(1 + {}^{\mu}/_{\phi})$, with μ, the predicted mean, and ϕ, the dispersion parameter. For purposes of comparisons, it is worth noting that DeSeq2 uses log2 while glmmTMB uses a natural log scale for the link function. Otherwise, model specification was identical. These different approaches produced comparable estimates (lowest r = 0.729, CI = 0.720 – 0.737; Figure 2.3, Table 2), although as expected the standard errors of the estimates differ substantially.

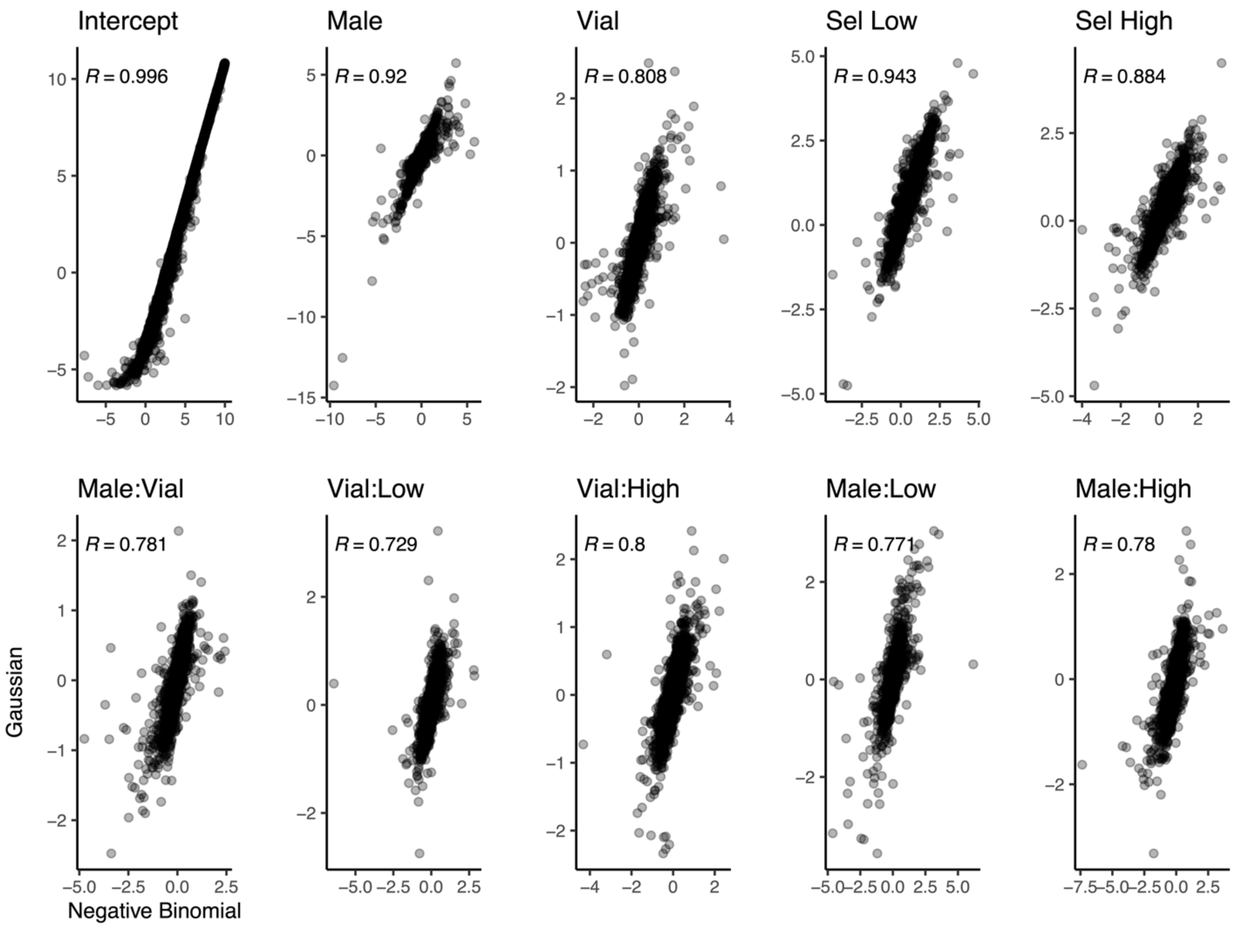

Figure 2.3. Estimate comparisons between negative binomial and gaussian implementation estimates produced from glmmTMB for various coefficients of their respective model. The gaussian is fitting $\log_2$(cpm) while the negative binomial is fitting $\log_2$Fold-Change. Both are fitting the same model, the difference between the two models is that one uses the negative binomial distribution and the other uses gaussian. Both estimates are $\log_2$(cpm). From top to bottom and left to right the terms we have are intercept (R = 0.99), sex male (R = 0.92), experience vial (R = 0.81), selection high (R = 0.87), selection low (R = 0.94), male:vial interaction (R = 0.78), vial:low interaction (R = 0.73), vial:high interaction (R = 0.80), male:low interaction (R = 0.76) and male:high interaction (R = 0.77).

Table 2.2. Table of Pearson correlation coefficient and corresponding confidence intervals of estimates between gaussian and negative binomial implementations from glmmTMB by model term.

Model Term	r	Lower	Upper
Intercept	0.99594	0.99579	0.99608
Male	0.91959	0.91672	0.92236
Vial	0.80832	0.80189	0.81455
Selection Low	0.94311	0.94106	0.94510
Selection High	0.88382	0.87976	0.88776
Male:Vial	0.78077	0.77354	0.78780
Vial:Low	0.72860	0.71992	0.73706
Vial:High	0.79972	0.79305	0.80621
Male:Low	0.77100	0.76349	0.77831
Male:High	0.77962	0.77235	0.78668

Gene Curation

After fitting the models, we used the "Anova()" function from the R package car (Fox & Weisberg, 2019) to perform a Wald test on our fitted model, which we then extracted the associated gene and p-value from. From here we ended up with a list of genes and their p-values, which we then applied a p-value adjustment to, using the R function "p.adjust()" with "method = 'BY' (Benjamini & Yekutieli, 2001) for controlling false discovery rate. After adjustment, we filtered out genes with FDR < 0.05, and subsetted emmeans contrast list to include only those that fit this criterion. From here, we split emmeans contrast lists into three lists, which corresponded to our three contrasts, low versus high, low versus control and control versus high. In each of these three lists, we pulled out genes with a p-value < 0.05 to obtain a list of genes in each of the three contrasts that potentially mediate sociability. We investigated these genes using the *Drosophila* specific database, Flybase (vFB2023_01), focusing on whether previous work indicated expression in the head tissue, links to social behaviour, and orthologous human genes.

Differential Transcript Usage (DTU) Analysis

We followed the recommendations for DTU analysis as outlined in Love et al. (2018). We generated transcript level counts from Salmon and imported them into R. We normalized the counts to scale to library size during import. We filtered transcripts using the “dmFilter()” function from DRIMseq (v1.24.0, Nowicka & Robinson, 2016). For a gene to be retained through filtering, the gene had to be expressed in a minimum of 28 samples (out of 142 total) with a minimum expression of 10 counts per sample in those samples. For a given transcript to be retained, it had to be expressed in a minimum of 20 samples, with the transcript representing at least 5% of the gene’s total expression in those samples. This removes rare transcripts (within sample) or genes with limited expression (across samples), as *a priori* we would not expect sufficient statistical power to estimate these coefficients with sufficient precision. Prior to filtering, there were 6,559 genes with at least two transcripts and 21,143 transcripts representing those genes. Post filtering, we had 4,761 genes and 12,335 transcripts representing those genes (Figure 2.4).

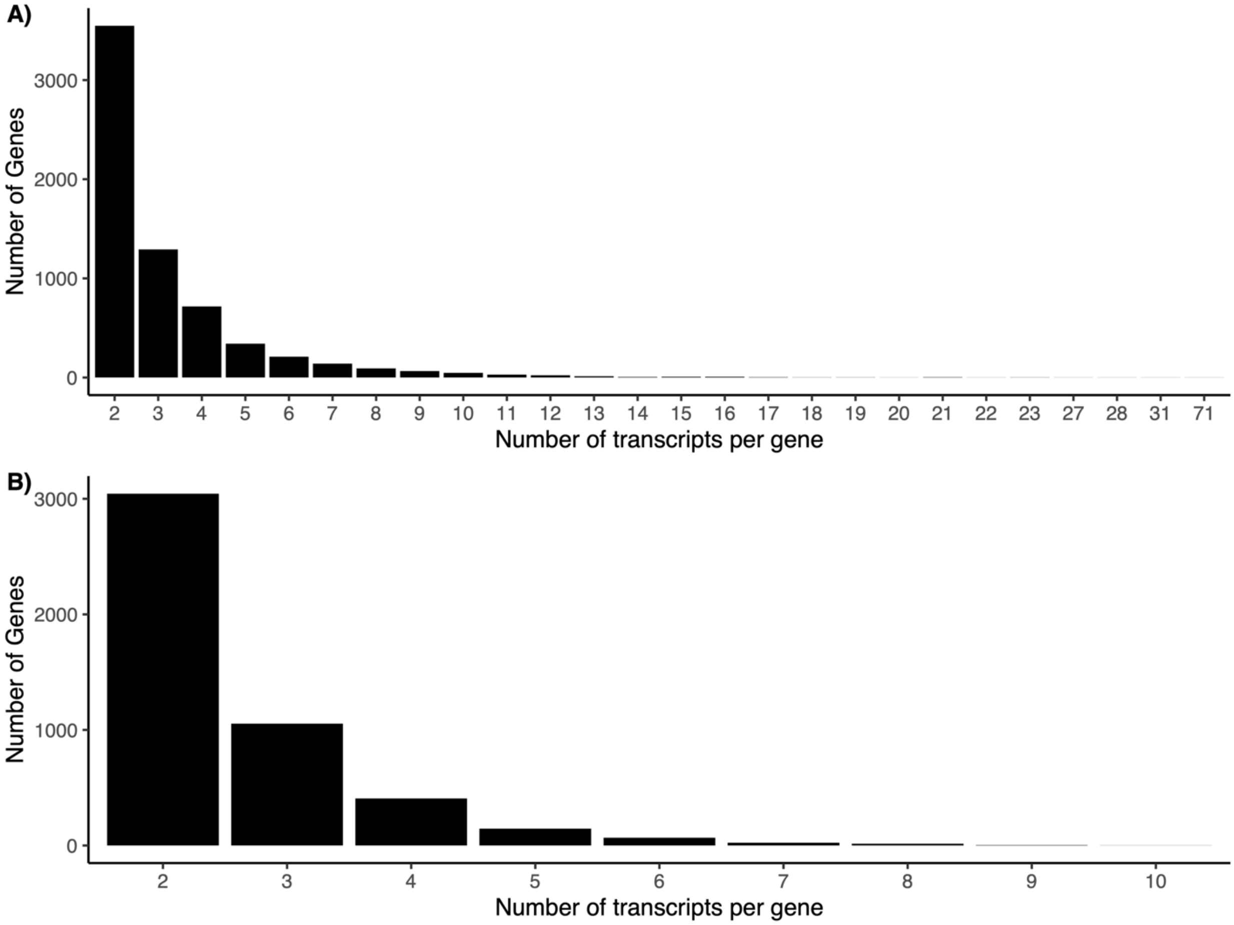

Figure 2.4. Bar plot of the distribution of the number of transcripts per gene. A) Pre filtering where all genes with at least two transcripts are represented. **B)** Post filtering of genes and their transcript. For a gene to be retained through filtering, the gene had to be expressed in a minimum of 28 samples (out of 142 total) with a minimum expression of 10 counts per sample in those samples. For a given transcript to be retained, it had to be expressed in a minimum of 20 samples, with the transcript representing at least 5% of the gene's total expression in those samples.

For model fitting, we used an approach analogous to that used in DEXseq (Anders et al., 2012). However, as we did for total gene expression, we modified this approach to allow for the inclusion of random effects in the model. Importantly, this approach focuses on examining transcript-treatment interactions to assess DTUs (Love et al., 2018). For computational efficiency, DEXseq (v1.5.3) implemented a change (relative to earlier versions of the software) in how the design matrix is coded (Reyes et al., 2013), and thus how contrasts between transcripts are estimated (Reyes et al., 2013). This was done to deal with the computational overhead for situations where the number of exons (or transcripts) per gene was very high (Anders et al., 2012). However, for our purposes this was not a constraint, as such we retained treatment contrast coding for our design matrix during estimation, and, as discussed below, used emmeans to extract specific estimates and contrasts.

For all genes that passed filtering, we used glmmTMB to fit a model that predicted counts for each individual transcript of a gene. For a given gene, we modelled the counts of all transcripts against each other. For a few genes, we observed complete separation (where a transcript was completely absent in one treatment but varying in others). To account for this, we added a count of one to all transcripts. Thus, changes in transcript usage will be slightly underestimated. We fit both full and “null” models using the negative binomial distribution with glmmTMB. The full model was:

$$\begin{aligned} counts \sim 1 &+ transcript + transcript{:}sex + transcript{:}selection \\ &+ transcript{:}sex{:}selection \\ &+ diag(sex + transcript\ |selection{:}lineage) + (1\ |\ sample\_id) \end{aligned}$$

The null model was:

$$counts \sim 1 + transcript + transcript{:}\, sex + diag(sex|selection{:}\, lineage) \\ + (1 \mid sample\_id)$$

The full reduced model, in case the full model failed to converge was:

$$counts \sim 1 + transcript + transcript{:}\, sex + transcript{:}\, selection \\ + transcript{:}\, sex{:}\, selection + (1 \mid sample\_id)$$

The null reduced model was:

$$counts \sim 1 + transcript + transcript{:}\, sex + (1 \mid sample\_id)$$

Following this, we again used the “Anova()” function from the car package to perform a Wald test on our fitted model. We also obtained the emmeans contrast for low versus high sociability. We then p adjusted our p-values using the BY method, and subsetted the anova results to only include transcripts with an adjusted p-value < 0.05. From here we subsetted our emmeans list to only include the genes that passed the previous cut-off for the anova.

Go Analysis

We performed gene ontology (GO) analysis using topGO (v2.48.0, Alexa & Rahnenfuhrer, 2022) on the set of genes extracted from the linear mixed models. We separately performed the GO analysis on the total set of genes from the DGE analysis and from the total set of genes from the DTU analysis. Here we wished to identify enriched GO terms in our gene list produced. We set the number of genes per GO term to 5 and used Fisher's exact test. We then adjusted the resulting p-values for multiple comparisons using the Benjamini-Hochberg method of multiple comparisons correction.

Simulations to assess overlap

As we are comparing and identifying the intersection between sets of gene lists, we also set up a simulation to assess whether the number of overlapping genes identified between two sets of gene lists were more or less common than expected by random chance alone. We did this for the sets of genes that overlapped between the 4 glmmTMB implementations, the comparison between DGE and DTU gene lists, between our lists of DE genes and when we compare to orthologs in different species. Each run was between two contrasts and consisted of 1,000 simulations in which we randomly sampled two sets of genes (one for each contrast or species) without replacement and independently of each other that corresponded with the total number of genes identified in each contrast. We sampled from the total number of genes that we mapped to (or were shared between two species) and observed how many genes would be common between the two. For the simulations where we looked at the number of genes between species we had an estimate of the number of shared genes, but we also performed the simulation with +/- 10% of shared genes (which can be found in the script sharedDEGenes.R at the GitHub repository linked in the appendix). From the simulations we report the maximum number of genes overlapping and the number of overlapping genes in the 95th percent quantile.

Results

Principal Component Analysis (PCA)

To examine broad scale patterns of variation in gene expression, we used PCA on the samples (Figure 2.5). Sexually dimorphic gene expression accounts for much of the (co)variation that loads on the second principal component accounting for ~25% of the variation in gene expression, consistent with large scale sex-biased gene expression in the adult head (Arbeitman et al., 2016; Khodursky et al., 2020; Nanni et al., 2023). Interestingly, PC1 (accounting for ~36% of the variation), shows that the lineages artificially selected for reduced sociability (low) tend to have positively valued scores on PC1, while the samples from the control and high lineages are more variable along PC1. This variation (for high and control treatments) in gene expression is a result of lineage specific effects (i.e., replicate lineages within each selection treatment; Figure 2.6).

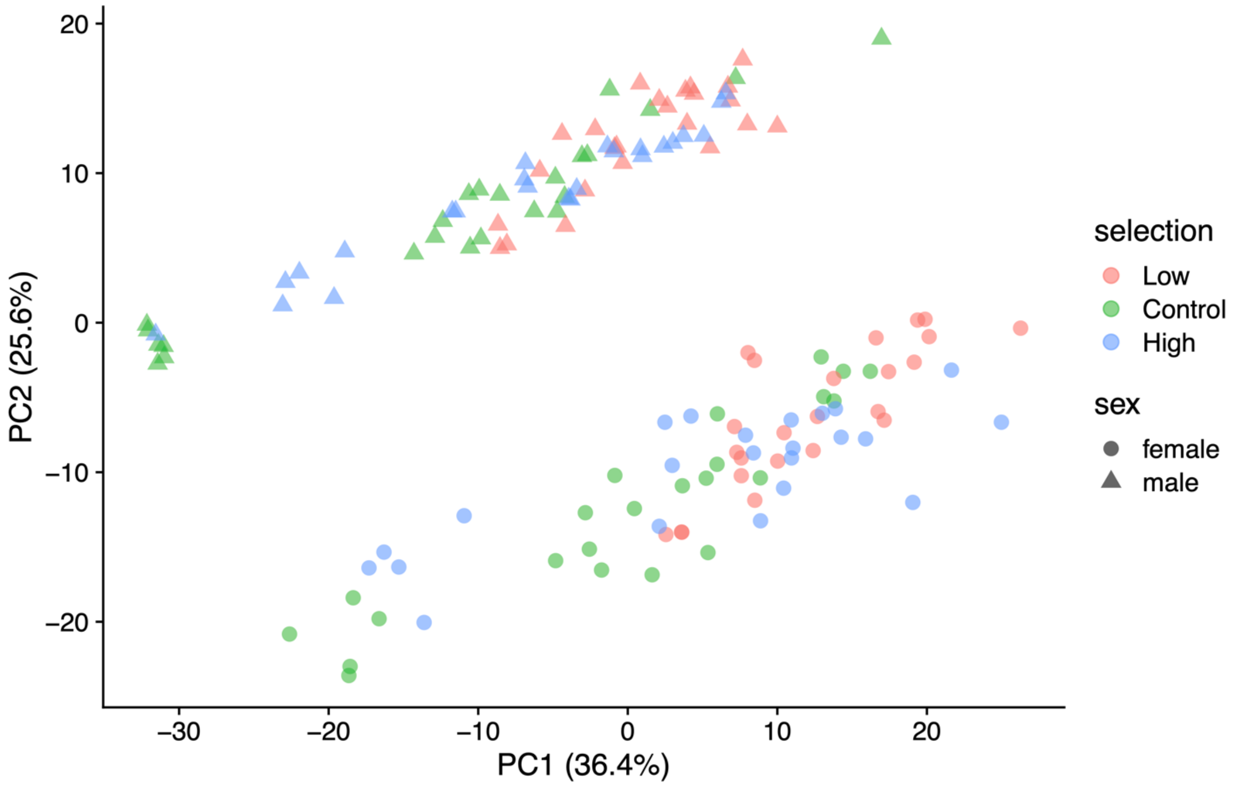

Figure 2.5. Principal component analysis (PCA) plots showing the variance associated with samples. Points on each plot are coloured by selection treatment with low in red, control in green, and high in blue. Different shaped points represent sex. PC1 (36.4% of variance) on the x axis and PC2 (25.6% of variance) on the y axis.

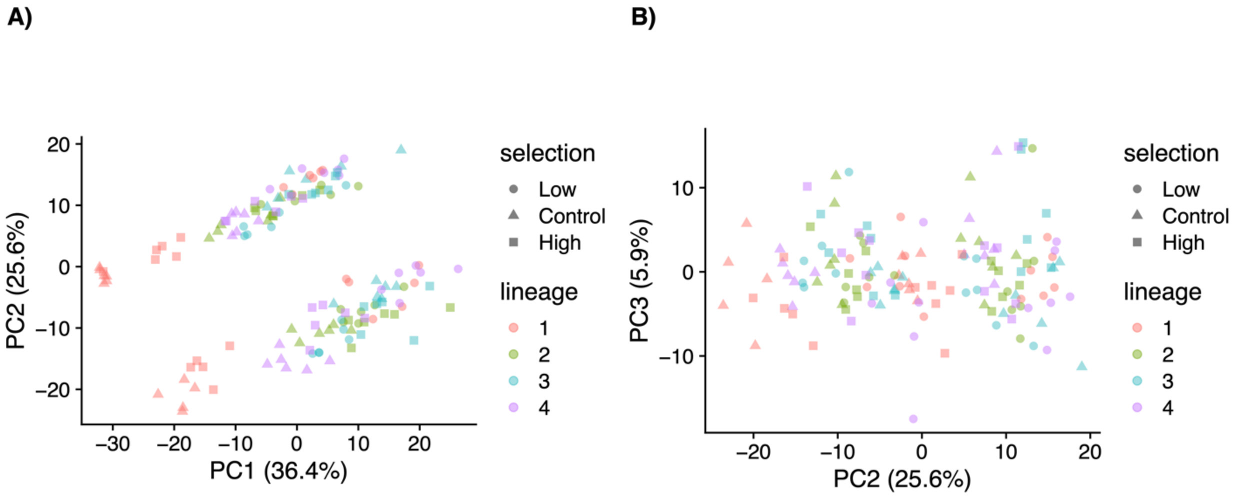

Figure 2.6. Principal component analysis (PCA) plots showing the variance associated with samples. Points on each plot are coloured by lineage with lineage 1 in red, lineage 2 in green, lineage 3 in blue and lineage 4 in purple. Different shaped points represent different selection treatments associated with the sample. **A)** PC1 (36.4% of variance) on the x axis and PC2 (25.6% of variance) on the y axis. **B)** PC2 on the x axis and PC3 (5.9% of variance) on the y axis.

Regularized Discriminant Analysis (RDA)

The first implementation of RDA was to predict lineages with classifiers including both sex and selection information. For all lineages we predicted, we found that we were able to predict the training set as well as the sex of the test set every time. However, we often failed to predict selection treatment. To ensure that our results were not due to sex effects, we split our dataset by sex and ran RDA separately. Again here, we found that we were not able to predict selection. We also extracted the top 1,000 most variable genes and performed the same analysis. We saw that we were unable to predict even the training set, let alone the test set.

Differential Gene Expression (DGE) Analysis

Following DGE analysis, we had results between all four possible combinations of using Salmon or STAR and using the gaussian ($\log_2$ cpm) and negative binomial distributions (counts with offsets). Results presented will be in reference to the Salmon-mapped counts fit with the gaussian distribution (All other contrasts and gene lists are provided in the Appendix; Figure 2.7). As a check, we first extracted differentially expressed (DE) genes between females and males. Previous studies have shown that within *D. melanogaster*, there are a large number of genes that show sex-biased gene expression differences (Parisi et al., 2004; Ranz et al., 2003) and relevant to our study, within the head (Arbeitman et al., 2016; Khodursky et al., 2020; Nanni et al., 2023). We found 5,331 genes that are DE between females and males (Figure 2.8).

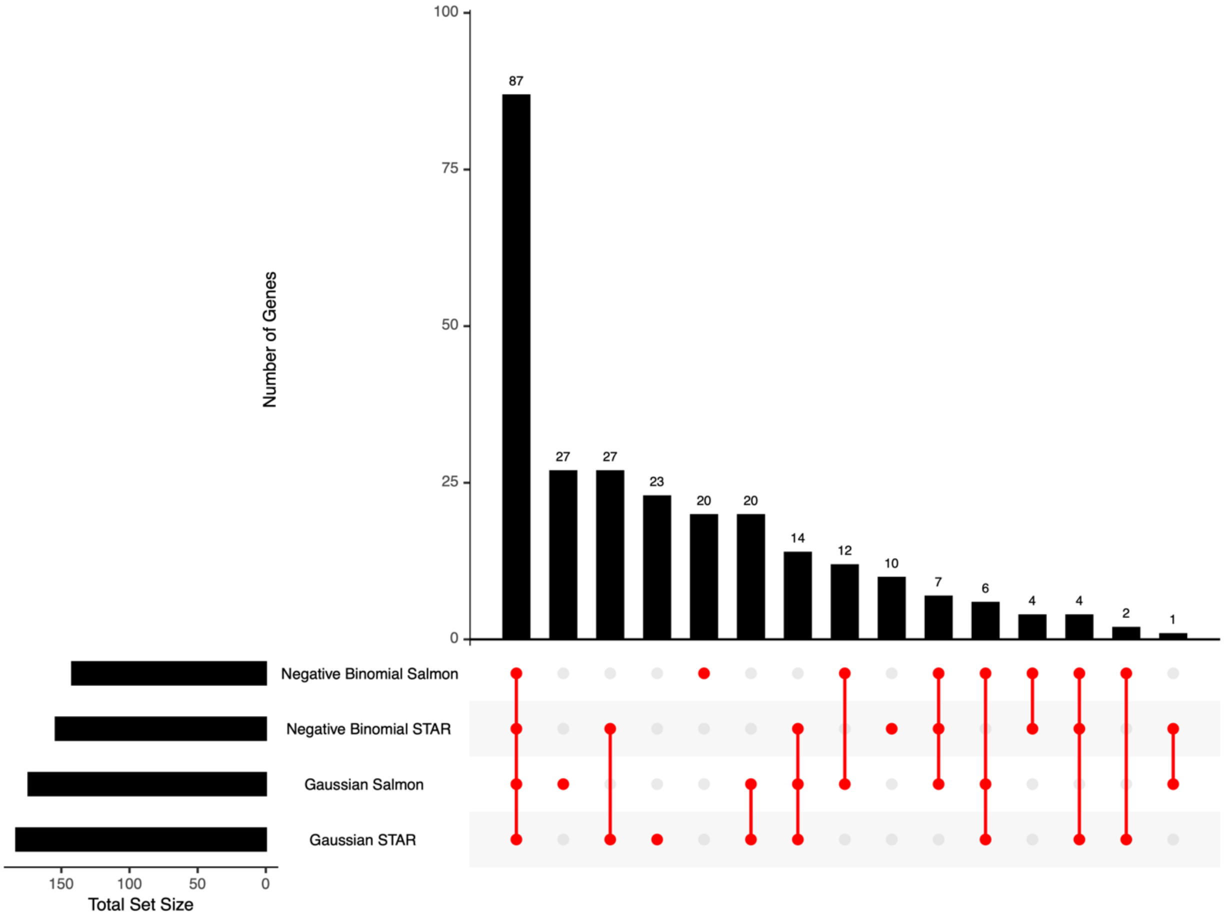

Figure 2.7. Upset plot showing overlapping genes between differentially expressed gene lists for the selection contrast of low versus high sociability. The four gene lists here are the four different ways we modelled differential gene expression. There are two mapping approaches of using STAR or salmon. There are two probability distributions we used to model gene expression with glmmTMB, those being the gaussian and negative binomial distribution. The points below the bars indicate which of the sets of genes are overlapping. Between the gaussian salmon and gaussian STAR we see 127 genes overlapping (simulation max 10 overlapping genes, 95% quantile of 6 genes). Between the gaussian salmon and negative binomial salmon we see 112 overlapping genes (simulation max 9 overlapping genes, 95% quantile of 5 genes).

Between the gaussian salmon and negative binomial STAR we see 109 overlapping genes (simulation max 9 overlapping genes, 95% quantile of 5 genes). Between the gaussian STAR and negative binomial salmon we see 99 overlapping genes (simulation max 10 overlapping genes, 95% quantile of 5 genes). Between the gaussian STAR and negative binomial STAR we see 132 overlapping genes (simulation max 11 overlapping genes, 95% quantile of 5 genes). Between the negative binomial salmon and negative binomial STAR, we see 102 overlapping genes (simulation max 9 overlapping genes, 95% quantile of 4 genes).

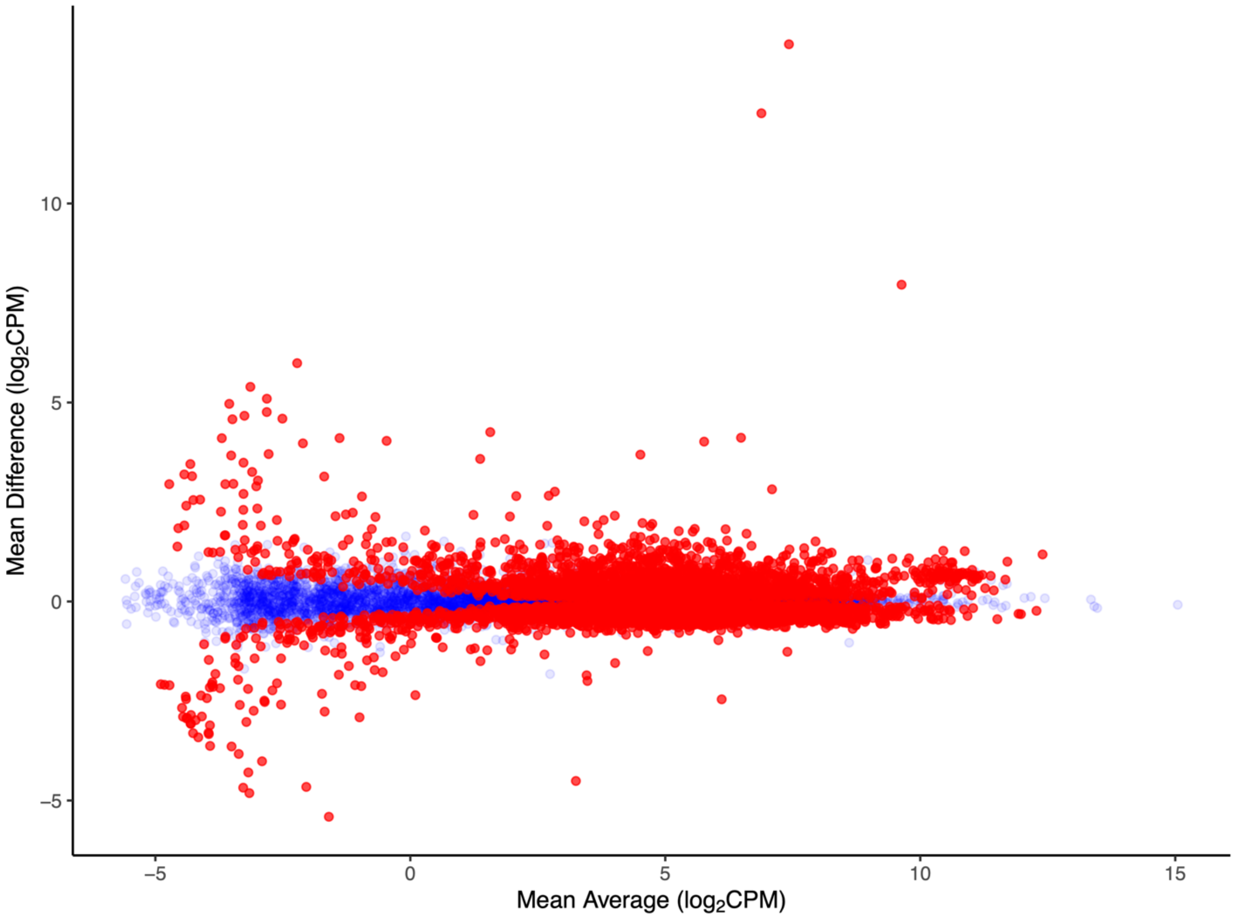

Figure 2.8. Mean-Average (MA) Plot for female versus male emmeans contrast. X axis is the mean average in $\log_2$(cpm) for each gene obtained from emmeans. Y axis is the mean difference between the female and male expression in $\log_2$(cpm), also obtained from emmeans. Red points are differentially expressed genes that have an adjusted p value < 0.05 when looking at the contrast (low versus high) estimate, and blue points are genes that have an adjusted p value > 0.05.

Within the three selection contrasts, we examined the distribution of effects and observed that a majority of them fell between a $\log_2$(cpm) of -1 to 1 (Figure 2.9). In the low versus control selection contrast, we saw 271 DE genes (Figure 2.10A). In the low versus high selection contrast, we saw 174 DE genes (Figure 2.10B). In the control versus high selection contrast, we saw 194 DE genes (Figure 2.10C). We saw a total of 327 DE genes across the three selection contrasts. A subset of the 12 genes with the largest effect size in the low versus high selection contrast were observed and visualized (Figure 2.11). The maximum and minimum change in gene expression was 3.98 and -6.59 respectively in $\log_2$(cpm), the majority (159/174) of differentially expressed genes show more modest changes between -1 to 1 $\log_2$(cpm) (Figure 2.12A). For full gene lists and plots of all genes see the Appendix section. Additionally, we found 213 genes differentially expressed between the vial and social arena experience contrast. We also examined if either sex or experience had an interacting effect with selection and found no evidence of genes altering gene expression in either the sex by selection interaction or the experience by selection interaction.

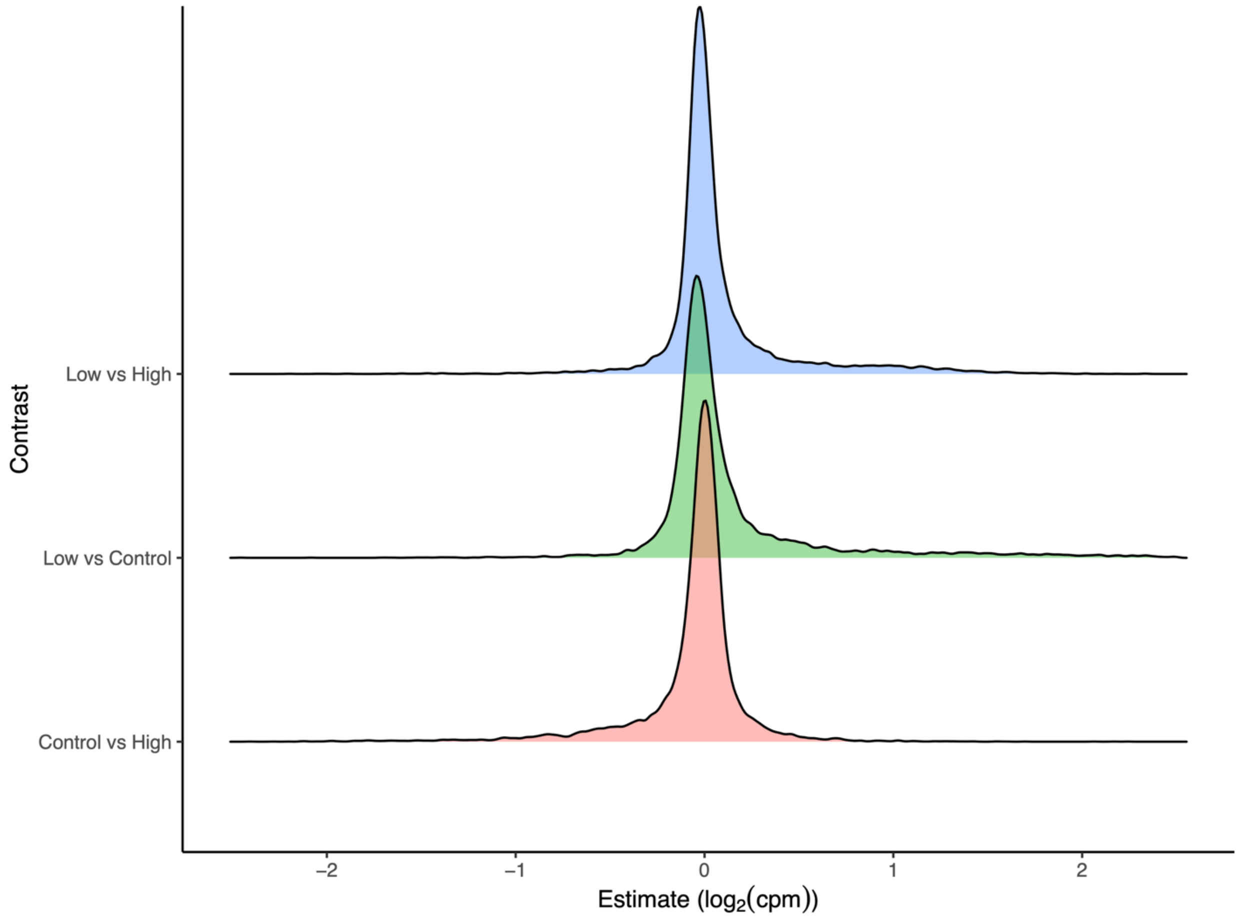

Figure 2.9. Density plots for three selection contrasts. The x axis represents the estimate of gene expression in $\log_2$(cpm). The y axis represents the sociability contrast. In blue there is low versus high sociability, in green we have low versus control sociability, and in red we have control versus high sociability. Note that this figure only shows the bounds of -2 to 2 in order to observe overlap of estimates between the three contrasts, as a majority of the estimates fall between a $\log_2$(cpm) of -1 to 1.

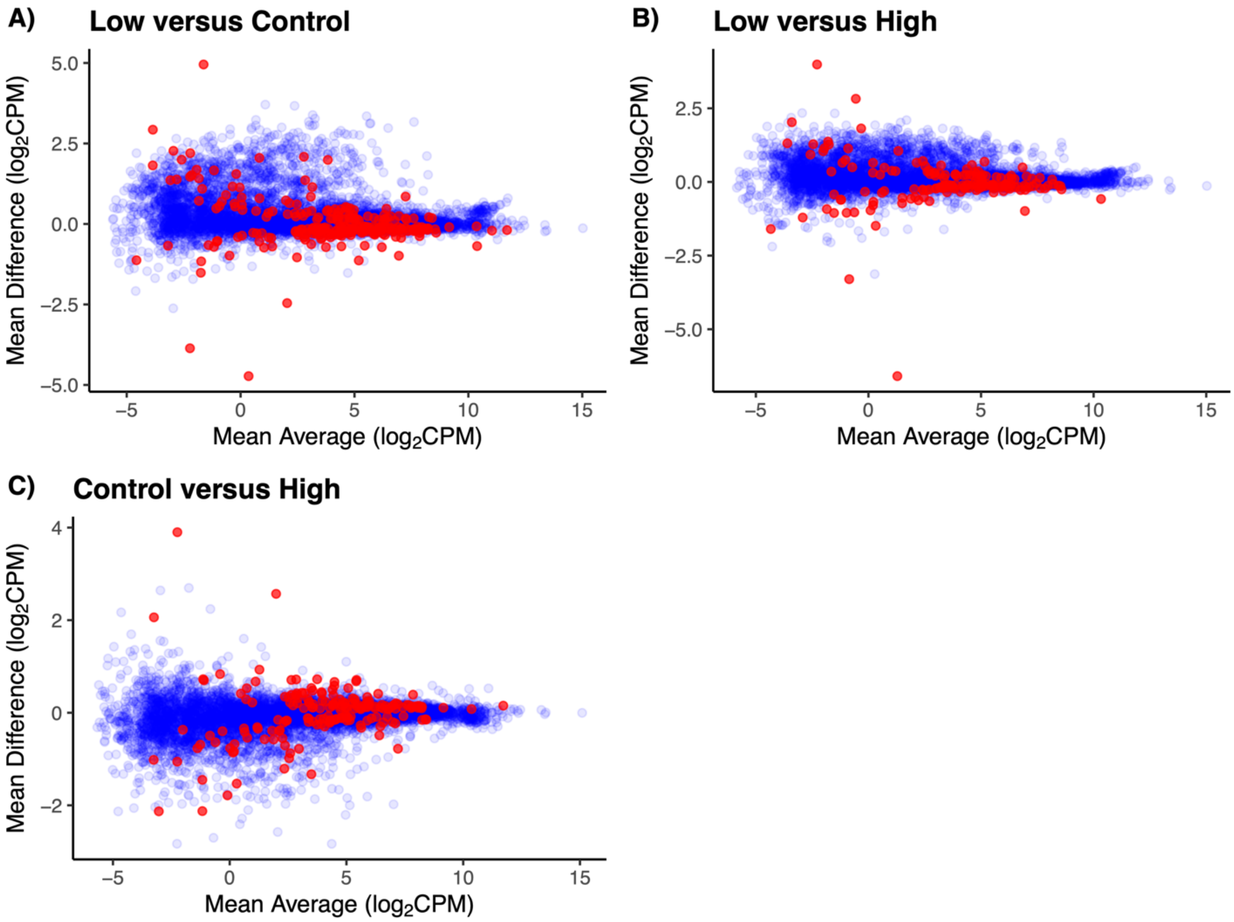

Figure 2.10. Mean-Average (MA) Plot for a given selection contrast. X axis is the mean average in $\log_2$(cpm) for each gene obtained from emmeans. Y axis is the mean difference between the low and high sociability expression in $\log_2$(cpm), also obtained from emmeans. Red points are differentially expressed genes that have an adjusted p value < 0.05 when looking at the given contrast, and blue points are genes that have an adjusted p value > 0.05. **A)** Low versus Control sociability emmeans contrast. **B)** Low versus High sociability emmeans contrast. **C)** Control versus High sociability emmeans contrast.

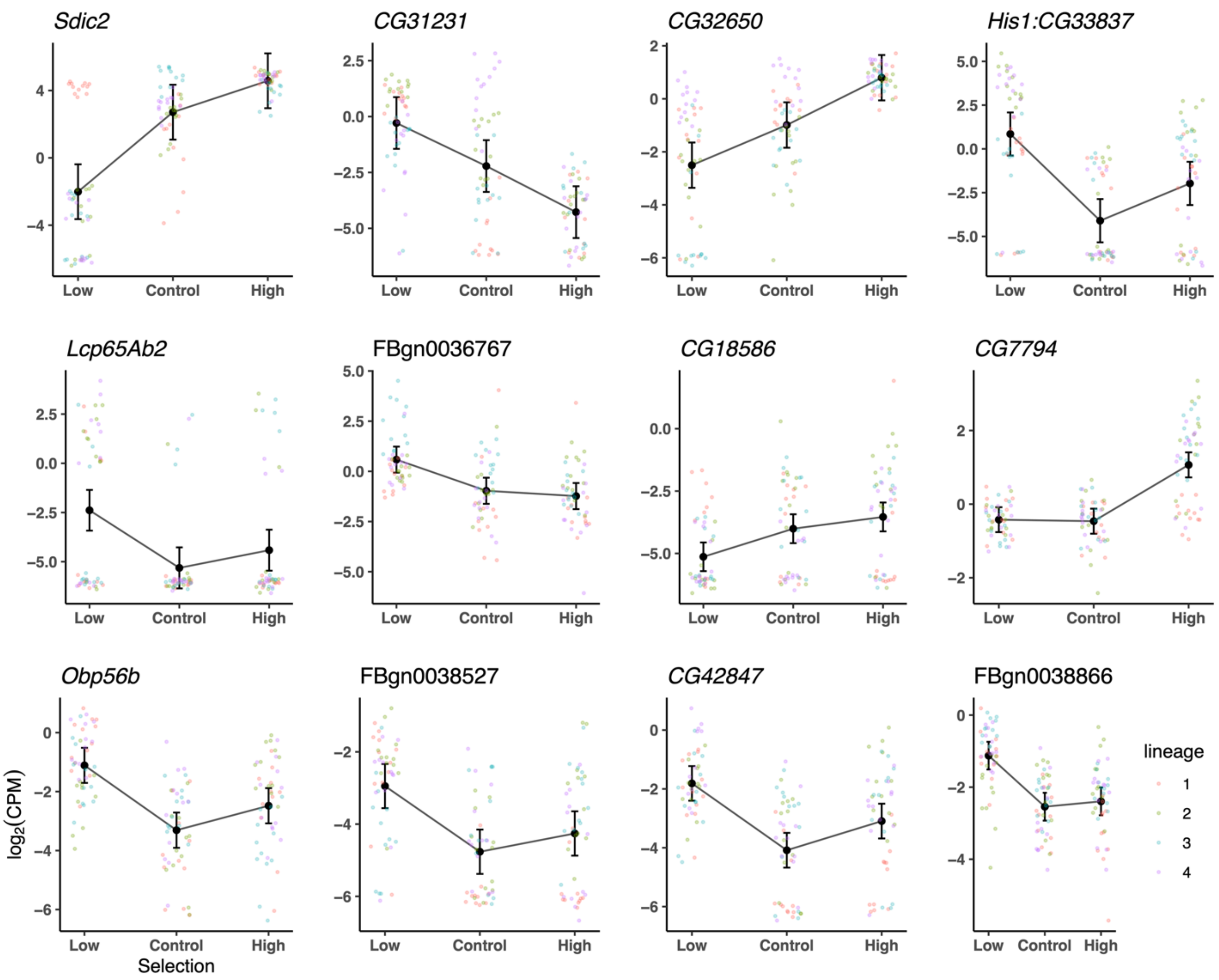

Figure 2.11. Reaction norms of the top 12 genes by log$_2$(cpm) difference in the low versus high sociability contrast. Each plot shows fitted gene expression in log$_2$(cpm) as obtained by emmeans with their 95% confidence interval. The individual points indicate the log$_2$(cpm) of each sample, where the 4 colours are the 4 lineages of each treatment.

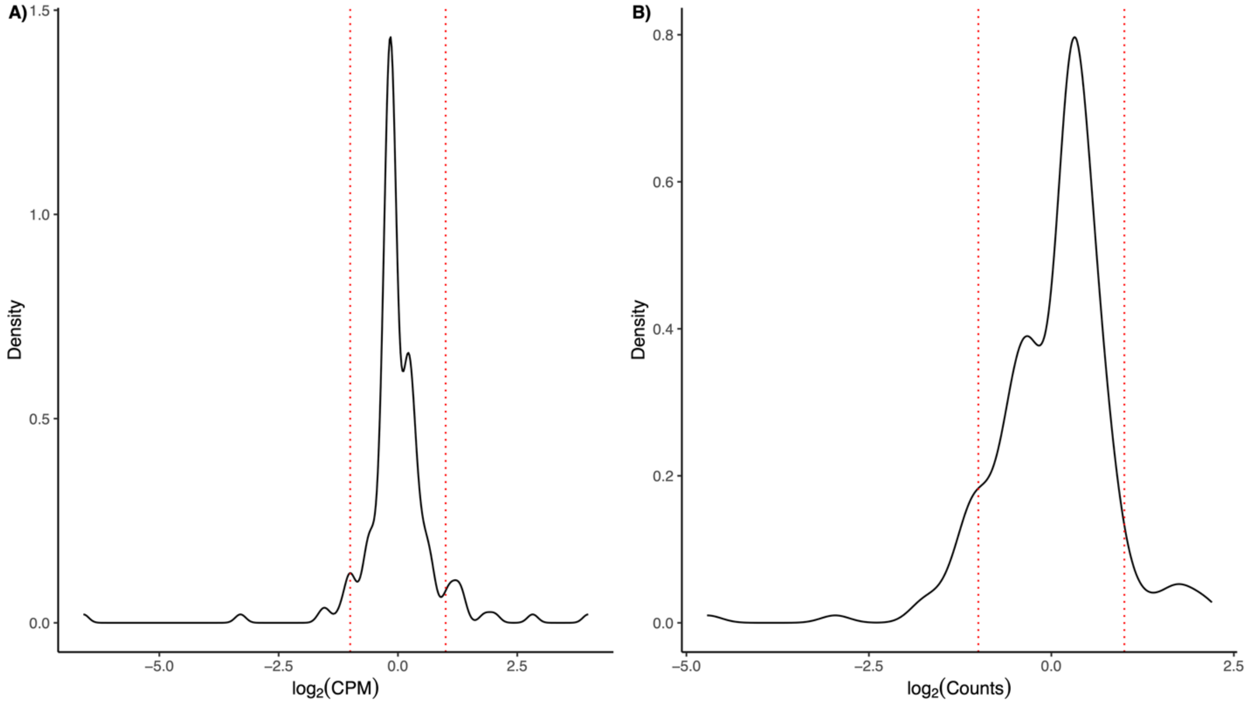

Figure 2.12. Density plots for the low versus high selection contrasts following differential gene expression (DGE) analysis or differential transcript usage (DTU) analysis. Estimates here are only those with an adjusted p value < 0.05. The x axis represents the estimated effect size ($\log_2$(cpm) for DGE and $\log_2$(Counts) for DTU). The y axis is the density of genes at a given estimate. Vertical lines are at an estimate of -1 and 1 of the respective axis units. **A)** Density plot for genes with DGE. **B)** Density plot for genes with DTU.

Gene Curation

Using the low versus high selection contrast gene list, we manually curated the genes. We found 33 genes that had supporting evidence of expression in the adult head and had a relevant phenotype associated. The relevant phenotypes included anything involved in abnormal neuroanatomy, neurophysiology, locomotor behaviour or circadian rhythm (Table 1). A subset of the 12 genes with largest effect size were observed and visualized (Figure 2.13).

Table 2.3. Manually curated list of genes with relevant phenotypes. The list contains Flybase ID (FBgnID), gene name, low versus high sociability contrast estimate, p value and a brief description of phenotypes reported on Flybase for different alleles of the gene.

FBgnID	Gene	Estimate	P-value	Phenotype
FBgn0010222	*Nmdmc*	0.68608	1.45×10^{-5}	Abnormal locomotor behaviour and stress response
FBgn0015773	*NetA*	0.65412	0.0011	Abnormal neuroanatomy and involved in axon guidance
FBgn0033885	*DJ-1α*	-0.62460	1.47×10^{-4}	Abnormal locomotor behaviour, neuroanatomy, and dopaminergic neuron
FBgn0036150	*Ir68a*	-0.51839	0.0021	Abnormal behaviour and involved in sensory neurons
FBgn0030795	*ppk28*	0.38051	5.3×10^{-4}	Abnormal memory, neurophysiology, taste perception
FBgn0037217	*CG14636*	0.34868	0.0016	Abnormal auditory perception
FBgn0031435	*Elba2*	0.31115	0.0488	Abnormal locomotor behaviour
FBgn0027783	*SMC2*	-0.28518	0.0015	Abnormal neuroanatomy

FBgn0016672	*Ipp*	-0.25391	8.82 x 10^{-6}	Abnormal learning in males, and abnormal neurophysiology
FBgn0261563	*wb*	0.24136	0.0093	Abnormal Neuroanatomy
FBgn0003174	*pwn*	-0.22274	1.4 x 10^{-4}	Abnormal neurophysiology
FBgn0266670	*Sec5*	-0.21897	1.5 x 10^{-4}	Abnormal developmental rate, neuroanatomy and size
FBgn0000565	*MsrA*	0.21552	0.0170	Involved in neuron projection
FBgn0032701	*CG10341*	-0.19039	0.0120	Abnormal neuroanatomy
FBgn0003654	*sw*	0.18939	0.0007	Abnormal neuroanatomy, paralytic, dendritic arborizing neuron
FBgn0035464	*PIG-B*	-0.18849	1.22 x 10^{-5}	Abnormal locomotor behaviour
FBgn0030932	*Ggt-1*	-0.18690	0.0018	Abnormal Behaviour
FBgn0030969	*Usp39*	-0.17637	2.95 x 10^{-5}	Abnormal locomotor behaviour
FBgn0266418	*wake*	0.16675	0.0017	Abnormal locomotor, courtship behaviour, abnormal sleep
FBgn0003301	*rut*	-0.14431	0.0208	Abnormal Behaviour, locomotor behaviour, neurophysiology, and neuroanatomy
FBgn0034585	*Rbpn-5*	-0.14072	2.36 x 10^{-5}	Abnormal Developmental rate and locomotor behaviour

FBgn0052982	*CG32982*	0.13876	9.9 x 10^{-4}	Abnormal locomotor behaviour
FBgn0029992	*Upf2*	0.13765	0.0011	Abnormal neurophysiology
FBgn0260635	*Diap1*	-0.13093	5.37 x 10^{-7}	Abnormal neuroanatomy, oxidative stress response, larval neurons, peptidergic neurons, abnormal size, and cell death
FBgn0030352	*sicily*	-0.13038	1.27 x 10^{-4}	Abnormal neuroanatomy and neurophysiology
FBgn0026083	*tyf*	-0.1136	1.28 x 10^{-5}	Abnormal Circadian behaviour and rhythm, abnormal locomotor rhythm
FBgn0001316	klar	0.10995	0.0153	Abnormal locomotor
FBgn0023095	*caps*	0.10796	0.0201	Abnormal neuroanatomy and axon guidance
FBgn0039861	*pasha*	-0.09879	0.0088	Abnormal neuroanatomy and neurophysiology
FBgn0037574	*Coq2*	-0.09574	0.0016	Abnormal locomotor rhythm
FBgn0024179	*wit*	-0.08577	0.0303	Abnormal neurophysiology and neuroanatomy
FBgn0032222	*Cox10*	-0.07619	0.0183	Abnormal locomotor behaviour
FBgn0039635	*Pdhb*	-0.07284	0.0232	Abnormal locomotor behaviour

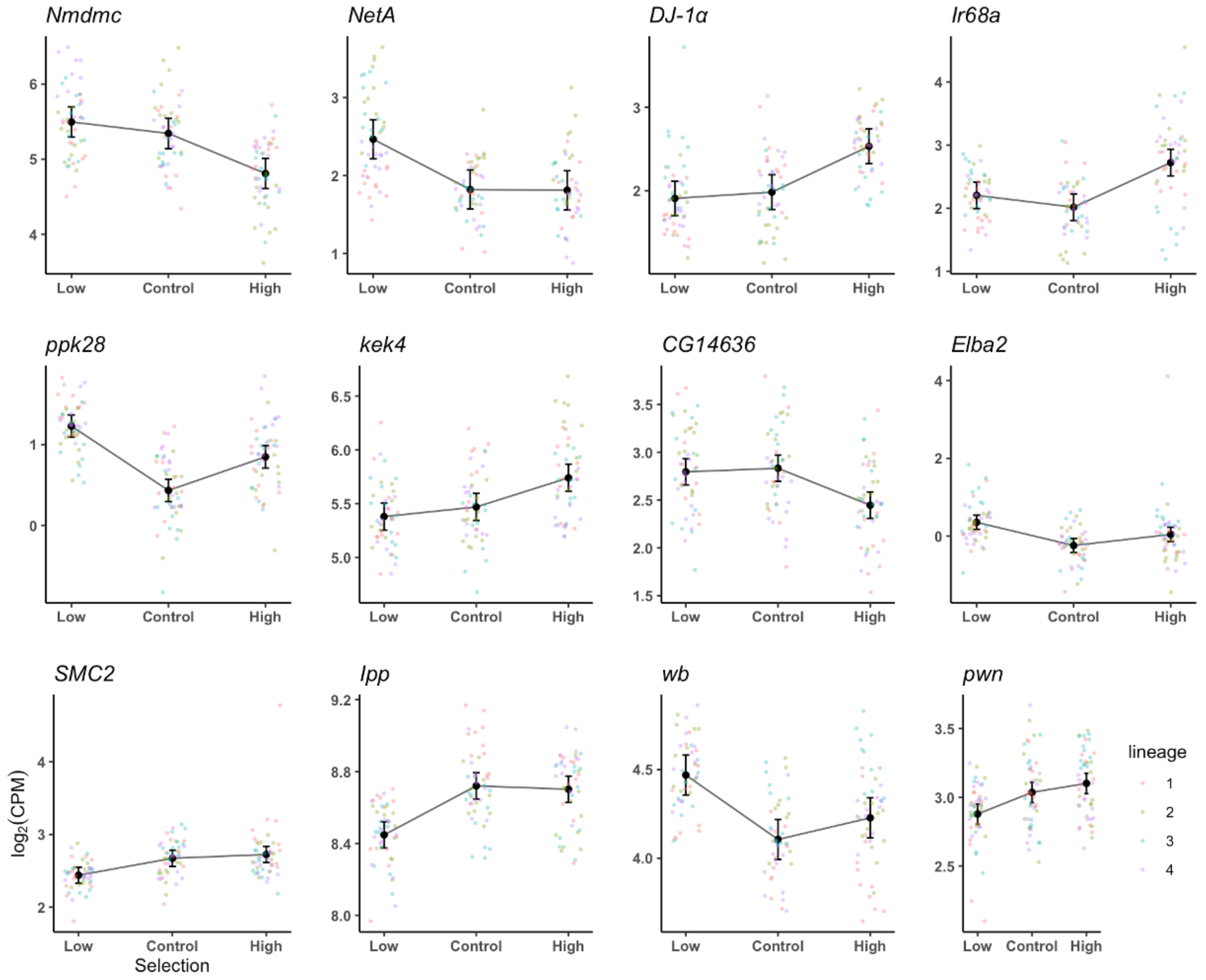

Figure 2.13. Reaction norms of the top 12 genes with relevant phenotypes by log2(cpm) difference in the low versus high sociability contrast. See table 2.3 for phenotypes. Each plot shows fitted gene expression in $\log_2$(cpm) as obtained by emmeans with their 95% confidence interval. The individual points indicate the $\log_2$(cpm) of each sample, where the 4 colours are the 4 lineages of each treatment.

Differential Transcript Usage (DTU) Analysis

Following DTU analysis we obtained gene list for genes (and their corresponding transcripts) that were differentially transcribed. As a check, we looked at DTU between females and males and saw 2,631 genes with differential transcript usage. In the low versus high selection contrast, we saw 190 genes with differential transcript usage (Figure 2.14; Figure 2.15). In the low versus control selection contrast, we saw 384 genes with differential transcript usage and in the control versus high selection contrast we saw 252 genes with differential transcript usage. In total we saw 619 genes overlap between all three selection contrasts. When looking at a sex by selection interaction we found 14 genes with differential transcript usage. When comparing our DTU results back to the DGE results, we see 39 genes that appear in both the overall DGE list and overall DTU list (Simulation maximum of 32 overlapping genes, 95% quantile of 21 genes; Figure 2.16). For full gene lists and plots of all genes that have transcripts with DTU in the selection contrasts, see the Appendix.

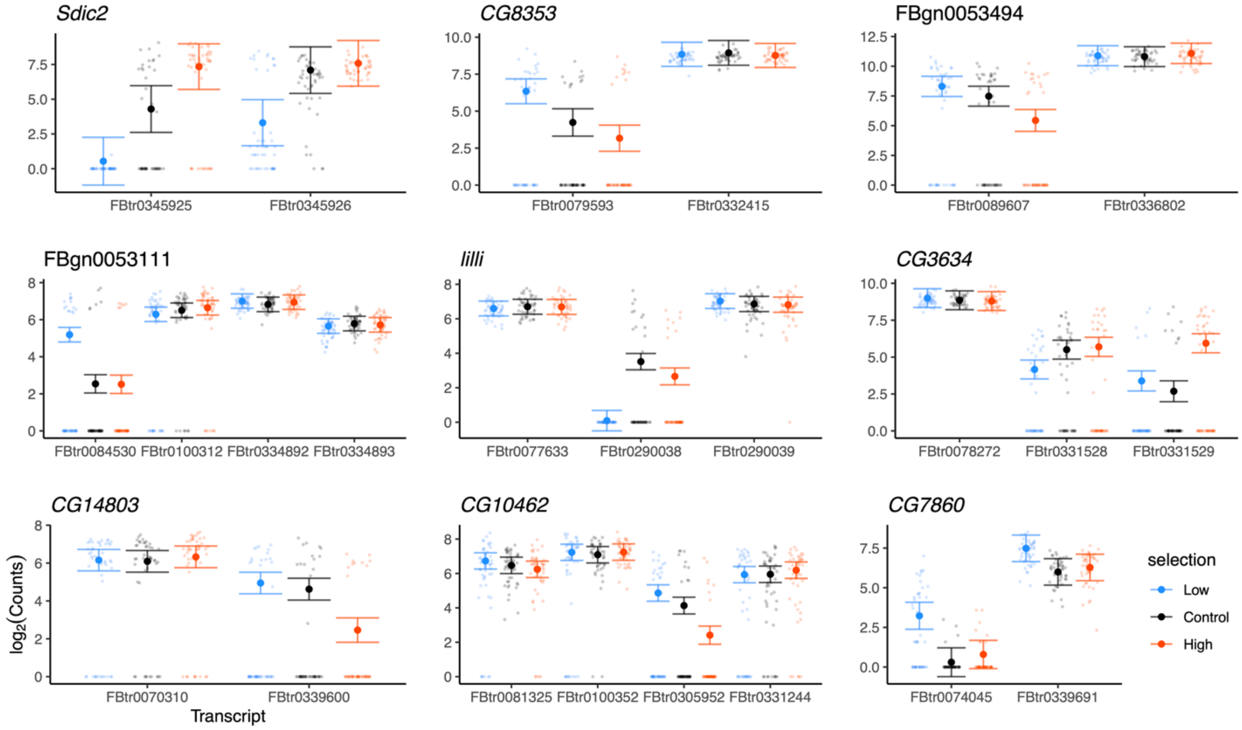

Figure 2.14. Reaction norms of the top 9 genes by $\log_2$(cpm) difference in the low versus high sociability contrast within the differential transcript usage gene list. Three genes with a large number of transcripts are depicted in Fig. 2.15. Each plot shows expression of the given gene and its associated transcripts. Along the x axis is the transcript of a given gene and the y axis is $\log_2$(counts). Each transcript is coloured by selection, with blue representing low sociability, black representing control, and red representing high sociability. The large points are the fitted expression values of a transcript as obtained by emmeans with their 95% confidence intervals. The small points are the $\log_2$(counts) of each sample.

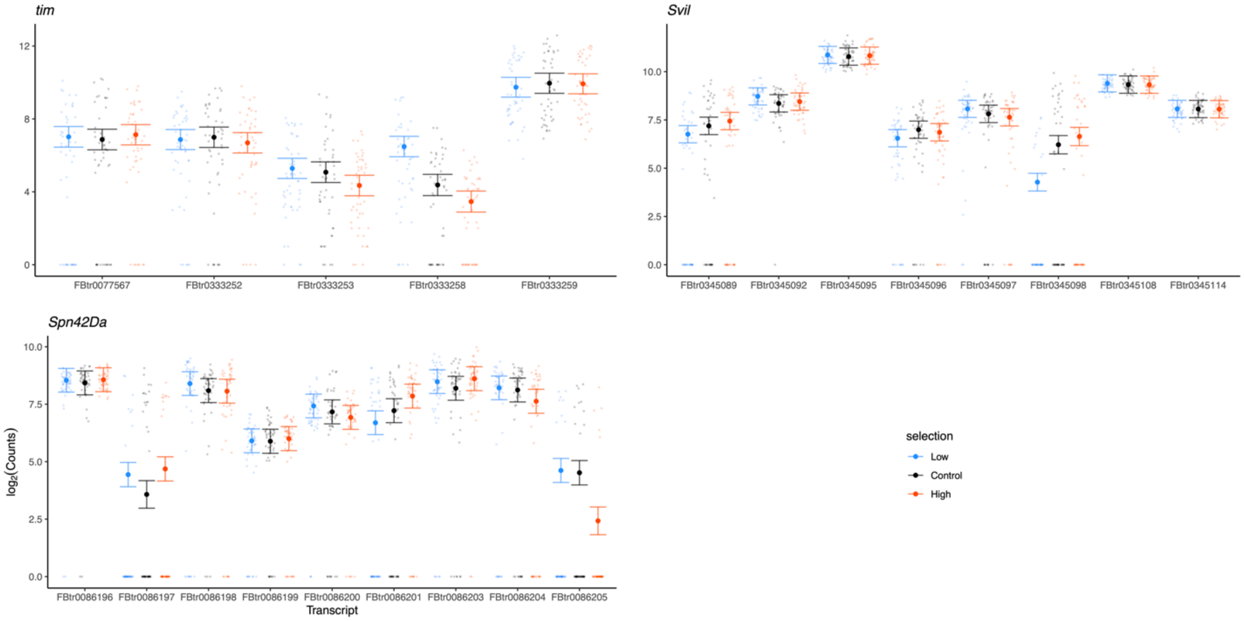

Figure 2.15. Reaction norms of three of the top 12 genes by $\log_2$(cpm) difference in the low versus high sociability contrast within the differential transcript usage gene list. These genes have a large number of transcripts associated with them (See figure 2.14 for the rest of the top 12). Each plot shows expression of the given gene and its associated transcripts. Along the x axis is the transcript of a given gene and the y axis is $\log_2$(counts). Each transcript is coloured by selection, with blue representing low sociability, black representing control, and red representing high sociability. The larger points are the fitted expression values of a transcript as obtained by emmeans with their 95% confidence intervals. The more transparent points behind the fitted values are the $\log_2$(counts) of raw expression for the given transcript and selection combination.

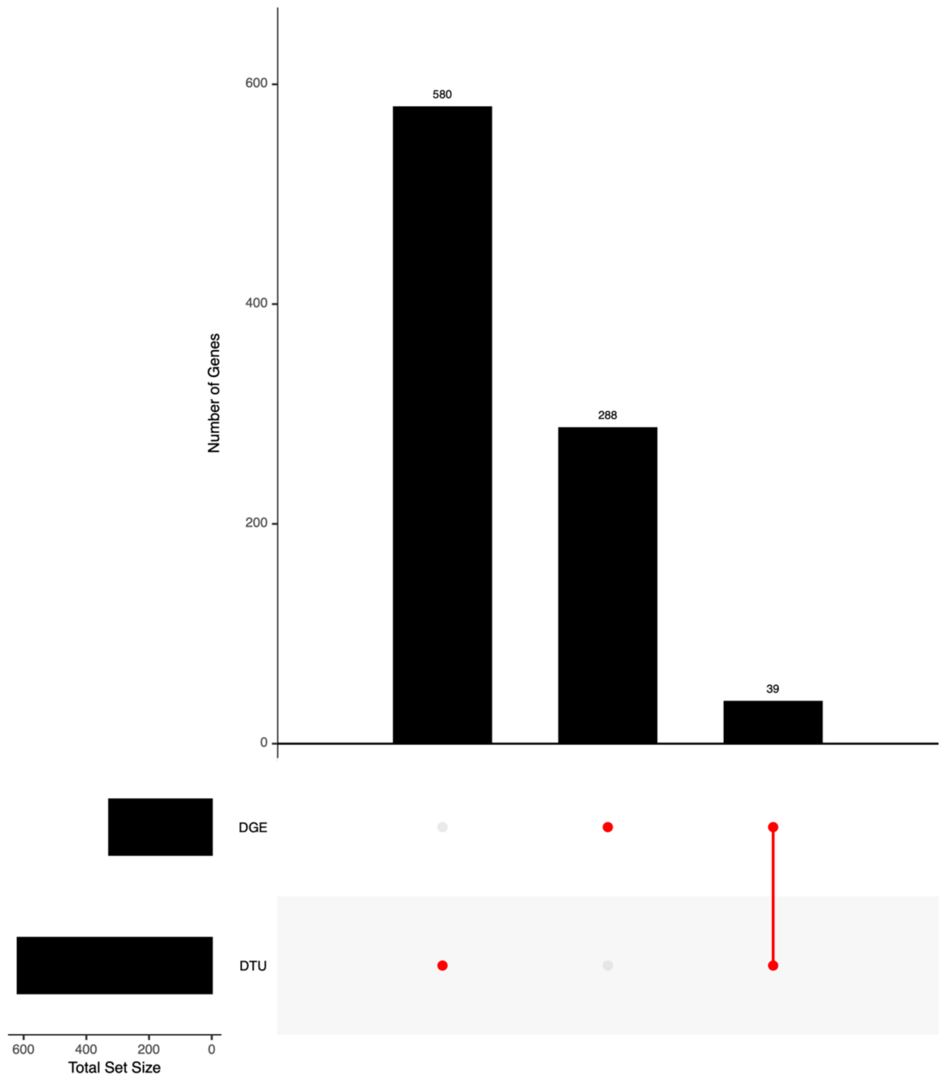

Figure 2.16. Upset plot showing overlapping genes between differentially expressed gene lists and the list of genes corresponding to transcripts that were differentially used. We see 39 genes overlap between the two lists (simulation max 32 overlapping genes, 95% quantile of 21 genes).

GO Analysis

Following GO analysis, we identified GO terms that are deemed as significantly overrepresented in our gene list of differentially expressed and differentially transcribed genes. When looking at all DE genes in our gene list, we found 60 GO terms overrepresented (Table A2). These terms included sensory perception of mechanical stimulus and synaptic assembly at neuromuscular junction. When looking at the DTU gene set, we found 43 GO terms overrepresented (Table A3) including photoreceptor cell axon guidance, regulation of neuron synaptic plasticity and regulation of compound eye photoreceptor development.

Comparison to Other Social Behaviour Studies

We compared differentially expressed genes from our study with those represented in the literature. Bralten et al. (2021) performed a GWAS study using 342,461 people from the UK Biobank and identified 56 genes and 18 independent loci associated with sociability. We took their list of 56 genes and identified orthologous genes in *Drosophila*. While no specific *Drosophila* orthologs appeared in our list of “differentially expressed” genes, the family of Solute Carrier genes did appear in both lists. We also took the orthologs of the 18 independent loci and identified the corresponding 8 orthologs (as some were SNP locations with no corresponding *Drosophila* orthologs). We did not find any of them in our gene list, and the expression of these 8 orthologs did not appear to be changing across selection conditions in our data (Figure 2.17). In our simulations we found a maximum overlap of 15 genes with a 95% quantile of 7 overlapping genes.

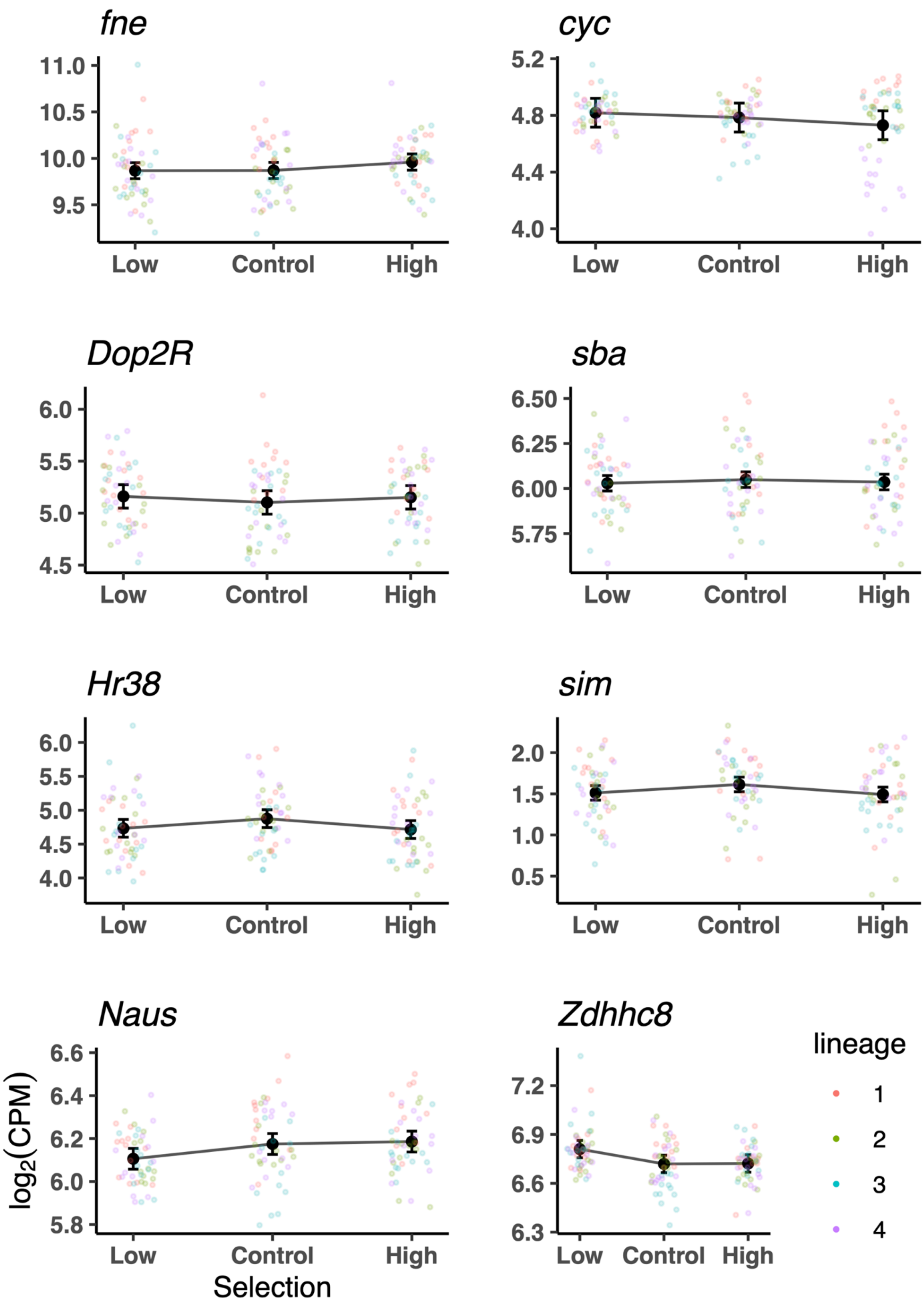
fne
11.0
10.5
10.0
9.5
Low
Control
High
cyc
5.2
4.8
4.4
4.0
Low
Control
High
Dop2R
6.0
5.5
5.0
4.5
Low
Control
High
sba
6.50
6.25
6.00
5.75
Low
Control
High
Hr38
6.0
5.5
5.0
4.5
4.0
Low
Control
High
sim
2.0
1.5
1.0
0.5
Low
Control
High
Naus
6.6
6.4
6.2
6.0
5.8
Low
Control
High
Zdhhc8
7.2
6.9
6.6
6.3
Low
Control
High
$\log_2(\text{CPM})$
Selection
lineage
1
2
3
4

Figure 2.17. Reaction norms of a subset of orthologous genes representing the gene list in Bralten et al. (2021) of independent loci in humans associated with social behaviour. The first three plots represent the orthologs to the human genes *ELAVL2, ARNTL, DRD2,* which were the most associated genes in their study. Each plot shows fitted gene expression in $\log_2$(cpm) as obtained by emmeans with their 95% confidence interval. The individual points indicate the $\log_2$(cpm) of each sample, where the 4 colours are the 4 lineages of each treatment. We identified 0 overlapping genes with our study (simulation max 15 overlapping genes, 95% quantile of 7 genes).

Another study, by Wang et al. (2022) examined early life social experience in the bumblebee, *Bombus impatiens*. They performed RNA sequencing to look for genes differentially expressed between three separate early life conditions, colony-housed, group-housed (with others but outside of the colony) and isolation (Wang et al., 2022). They ended up with a list of 94 DE genes between isolated and colony reared bees and 27 DE genes between isolated and group-housed bees, with six genes overlapping between the two contrasts (Wang et al., 2022). We identified orthologs of 68 out of the 115 total genes in *Drosophila* and found two genes that appeared in our DE gene list, *yellow-c* and *CG43066* (Figure 2.18). In our simulations we found a maximum overlap of 17 genes with a 95% quantile of 10 overlapping genes.

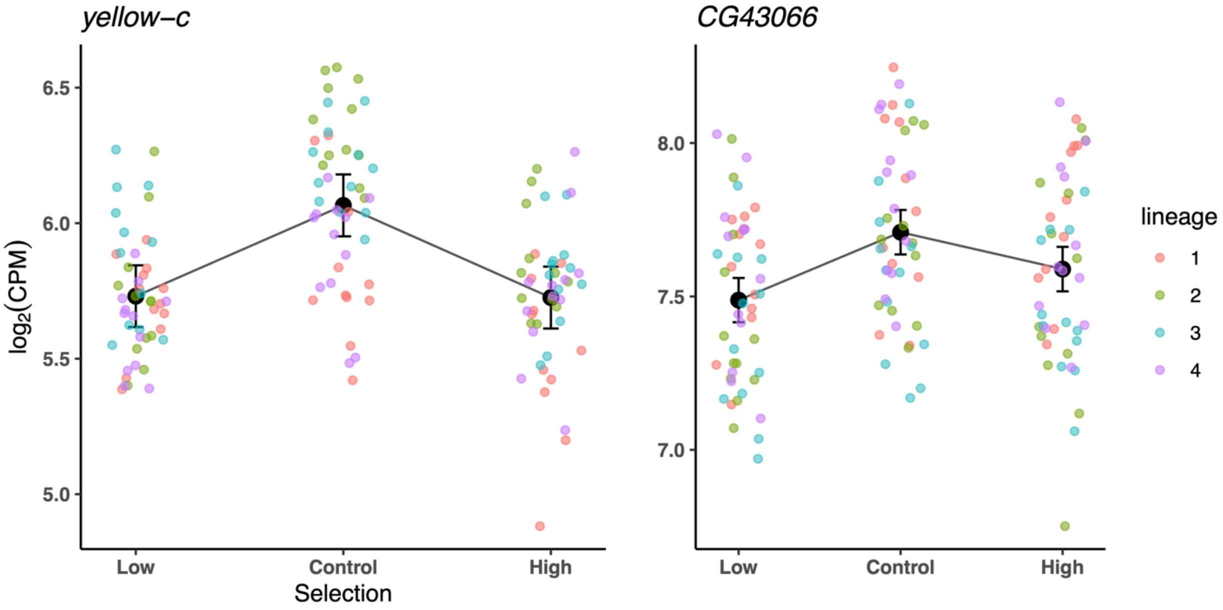

Figure 2.18. Reaction norms of orthologous genes in *Drosophila* that overlapped between the differentially expressed genes found in Wang et al. (2022) and differentially expressed genes found in our study. Each plot shows fitted gene expression in $\log_2$(cpm) as obtained by emmeans with their 95% confidence interval. The individual points indicate the $\log_2$(cpm) of each sample, where the 4 colours are the 4 lineages of each treatment. We identified 2 overlapping genes with our study (simulation max 17 overlapping genes, 95% quantile of 10 genes).

A study by Woodard et al. (2011) examined the convergent evolution of eusociality across bee species. They looked across nine socially diverse bee species which included eusocial and non-eusocial bees and identified 212 genes that evolved more rapidly in eusocial lineages compared to non-eusocial lineages (Woodard et al., 2011). Of the 212 genes they identified, we found four orthologs in *Drosophila* that appeared to be differentially expressed among our contrasts (Figure 2.19). In our simulations we found a maximum overlap of 25 genes with a 95% quantile of 16 overlapping genes.

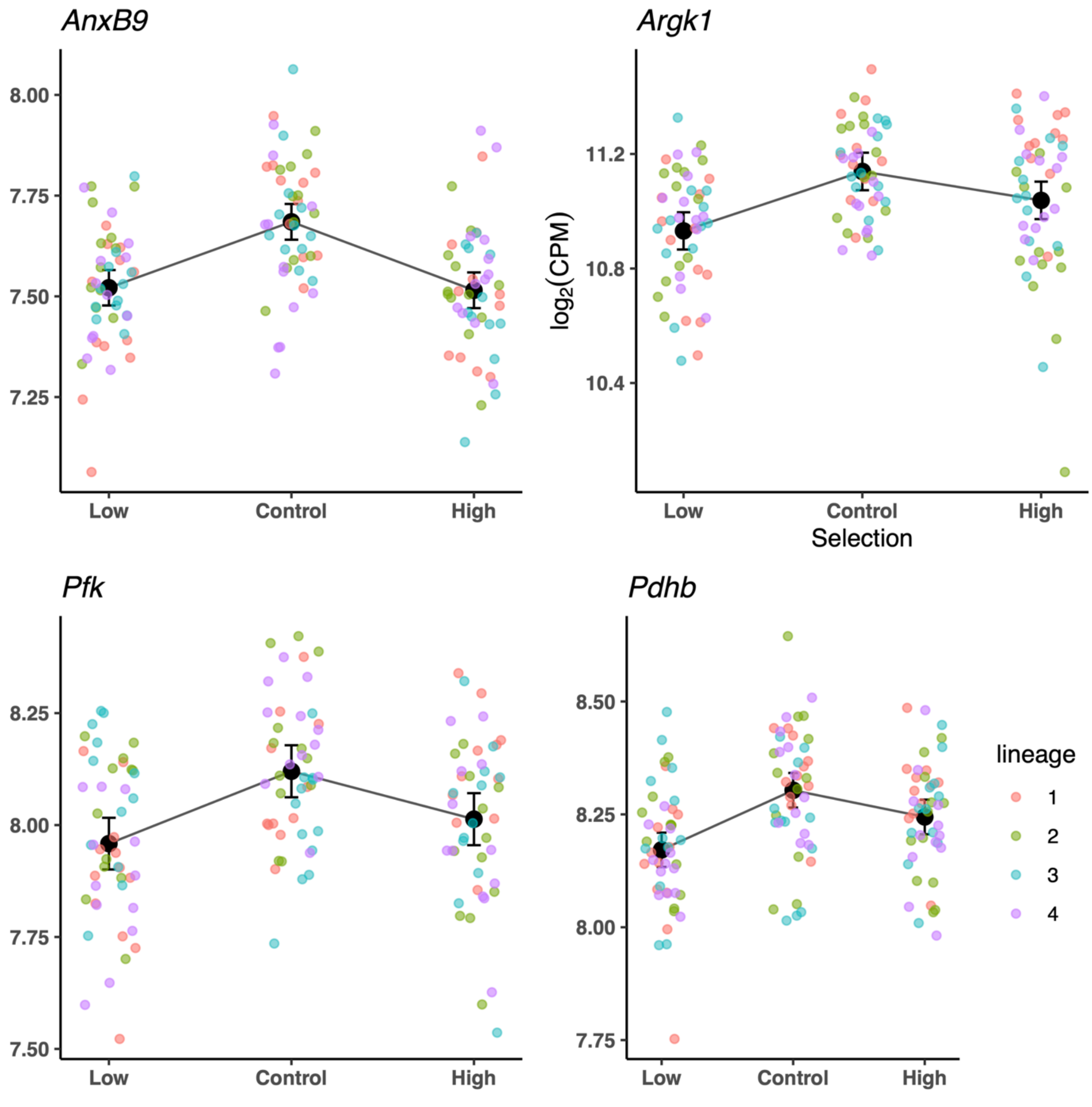

Figure 2.19. Reaction norms of orthologous genes in *Drosophila* that overlapped between the differentially expressed genes found in Woodard et al. (2011) and differentially expressed genes found in our study. Each plot shows fitted gene

expression in $\log_2$(cpm) as obtained by emmeans with their 95% confidence interval. The individual points indicate the $\log_2$(cpm) of each sample, where the 4 colours are the 4 lineages of each treatment. We identified 4 overlapping genes with our study (simulation max 25 overlapping genes, 95% quantile of 16 genes).

Shpigler et al. (2017) performed differential gene expression analysis on *Apis mellifera*. They performed assays and classified individuals as either guards, nurses or unresponsive and then performed DGE analysis on RNA obtained from the mushroom body of the brain (Shpigler et al., 2017). They identified 1,057 DE genes between all three groups of social responsiveness (Shpigler et al., 2017). From that list of 1,057, we identified 14 orthologs in *Drosophila* that appeared in any of our Low versus High, Low versus control or control versus high DE gene lists (Figure 2.20). In our simulations we found a maximum overlap of 81 genes with a 95% quantile of 66 overlapping genes.

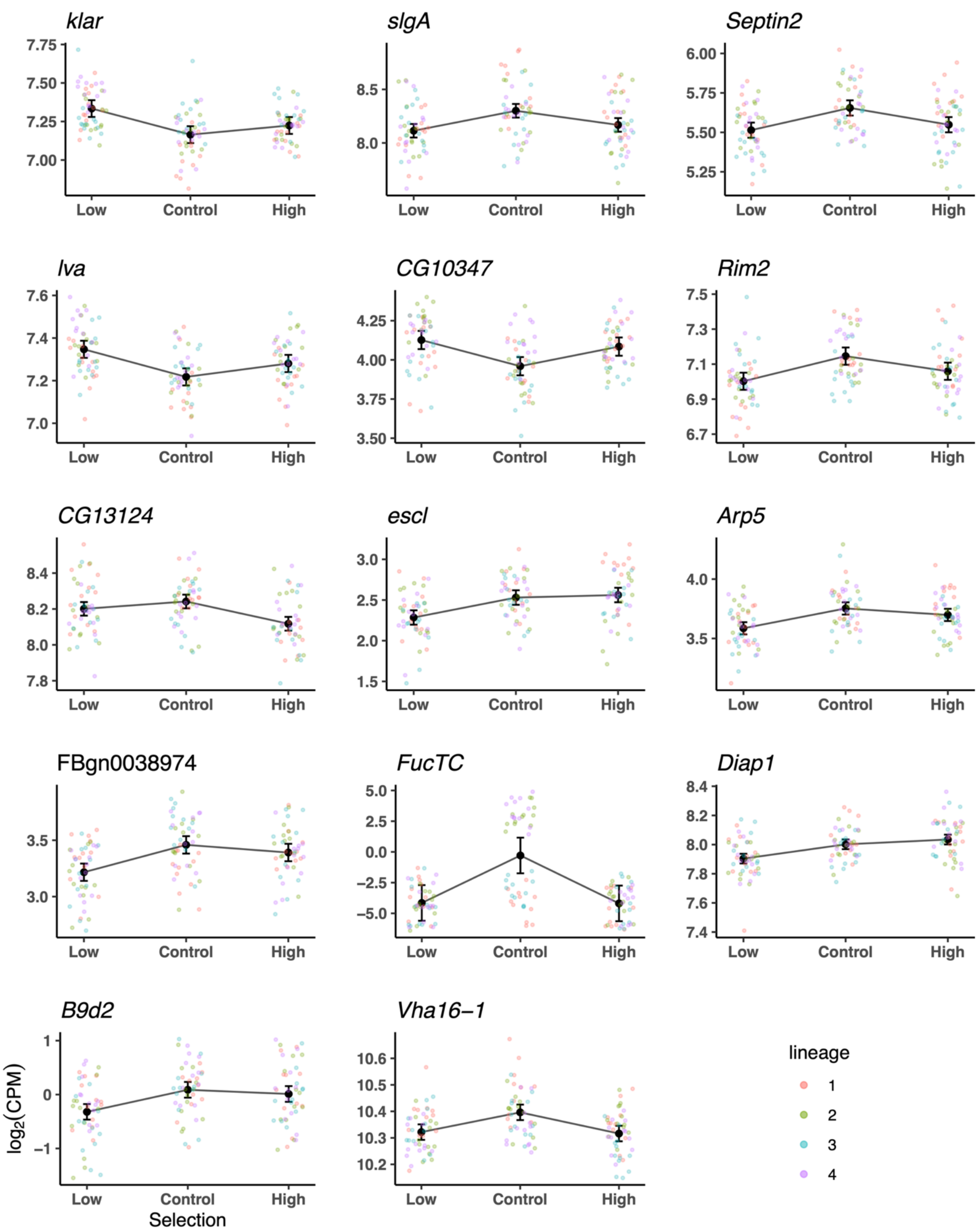

Figure 2.20. Reaction norms of orthologous genes in Drosophila that overlapped between the differentially expressed genes found in Shpigler et al. (2017) and

differentially expressed genes found in our study. Each plot shows fitted gene expression in $\log_2$(cpm) as obtained by emmeans with their 95% confidence interval. The individual points indicate the $\log_2$(cpm) of each sample, where the 4 colours are the 4 lineages of each treatment. We identified 14 overlapping genes with our study (simulation max 81 overlapping genes, 95% quantile of 66 genes).

When comparing to other study's findings, we did not necessarily see the same genes appear. Perhaps because this is a group response, it is more difficult to "capture" or assess the genomics of the trait consistently across studies. We measure sociability within *Drosophila* in a specific manner (Methods in Scott et al., 2022) and while we are capturing some form of group or social behaviour, it may not be the exact same sociability or behaviour that other studies are capturing. Different studies assess social behaviour in different manners which make interpretation and comparisons to the literature more difficult. For example, Bralten et al. (2021) performed a genome wide association study in humans from the UK biobank database who answered survey questions. In Woodard et al. (2011) expressed sequence tags (ESTs) were mapped to transcriptomes to identify regions of the genome evolving at higher rates. Also, in studies look at social behaviour in A. mellifera and B. impatiens, behavioural assays were performed and then RNA was extracted from a single generation for RNA sequencing (Shpigler et al., 2017; Wang et al., 2022). Those four studies produced lists of 56, 115, 212 and 1,057 genes that played a role in influencing social behaviour (Bralten et al., 2021; Shpigler et al., 2017; Wang et al., 2022; Woodard et al., 2011). Between the four of those lists, we see 22 genes that overlap with the genes we identified. While we are observing sets of genes influencing social behaviour across these species, we are not seeing a consistent influence of certain genes at a larger scale. It seems that different species use different genes to evolve this behaviour. Or perhaps, because of the different approaches to capturing social behaviour, we are seeing different genes come up in these gene lists.

Discussion

The main objective of this chapter was to identify what genes and transcripts are being differentially expressed and used between the high and low sociability lineages. Following our DGE and DTU analyses we have identified lists of genes and transcripts between the high and low sociable lineages that are differentially expressed and used. We have also identified lists of genes and transcripts in other contrasts such as between the low versus control and between the control versus high that are differentially expressed and used. We also asked the question of if there are genes that influence sociability in *Drosophila* that are shared in other species as well. Here, we did not find evidence of a shared response or a common underlying genetic basis.

In total we identified 327 genes showing differential expressed across all three selection contrast, with 174 genes differentially expressed between the low versus high selection contrast. While the maximum and minimum change in gene expression was 3.98 and (-6.59) $\log_2$(cpm), the majority (159 out of 174) of differentially expressed genes showed more modest changes between (-1) to 1 $\log_2$(cpm) (Figure 2.12A). We also see overlap between the sets of contrasts, with 146 genes overlapping between the control versus high and low versus control DE gene list (simulation max 14 overlapping genes, 95% quantile of 8 genes) as well as 122 genes overlapping between the low versus high and low versus control gene lists (simulation max 13 overlapping genes, 95% quantile of 8 genes; Figure 2.21). Given the consistent changes in gene expression, association with distinct biological functions, the size of each contrasts' gene lists and the number of overlapping genes between contrasts, the genetic response we observe may be polygenic. Unfortunately, this is only a suggestion because we cannot be certain that this response in gene expression is not from an overarching regulatory gene or gene pathway.

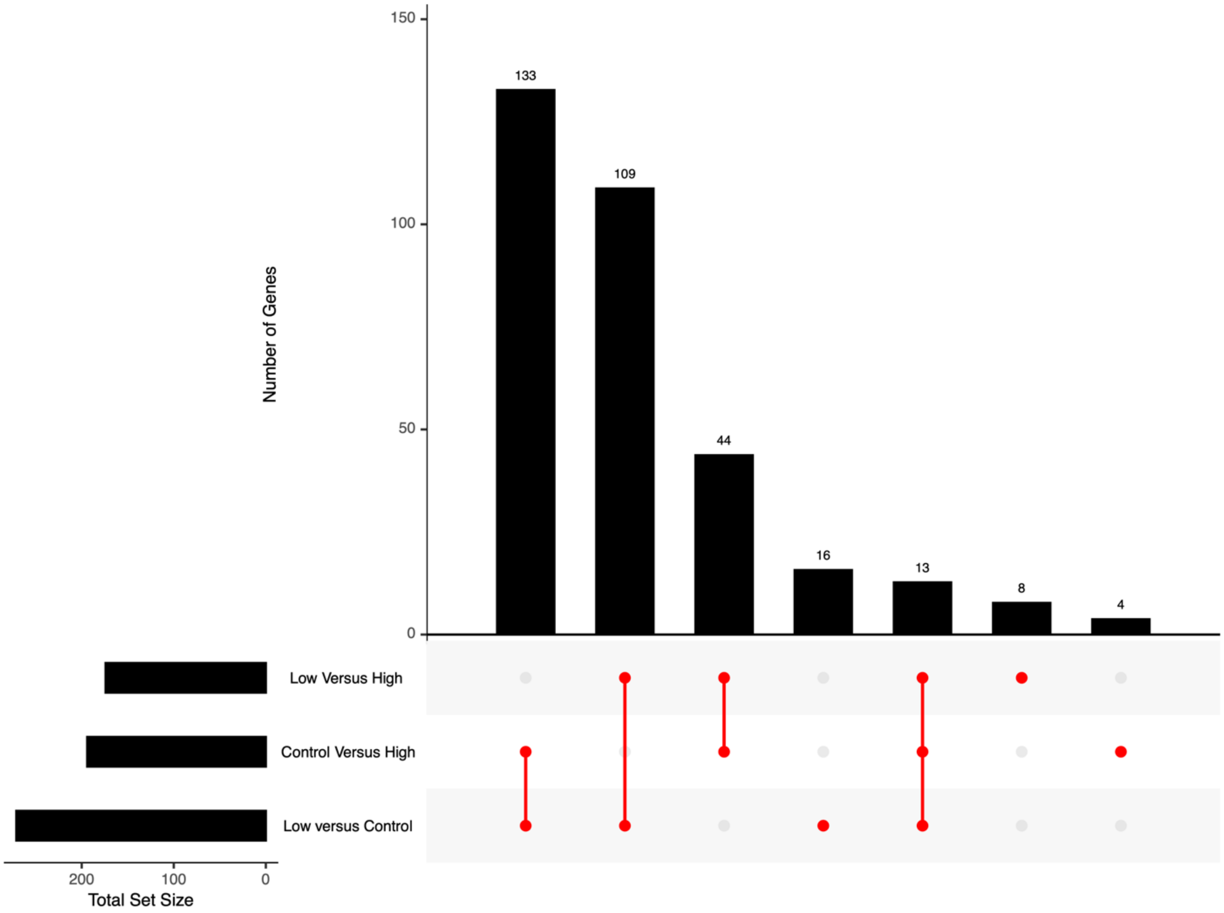

Figure 2.21. Upset plot showing the overlap between the gene lists obtained from differential gene expression analysis. Gene lists are selection contrasts of Low Versus High, Control Versus High and Low versus Control. Between the low versus high and low versus control we see 122 overlapping genes (simulation max 13 overlapping genes, 95% quantile of 8 genes). Between the low versus high and control versus high we see 57 overlapping genes (simulation max 11 overlapping genes, 95% quantile of 6

genes). Between the low versus control and control versus high we see 146 overlapping genes (simulation max 14 overlapping genes, 95% quantile of 8 genes).

We identified transcripts showing differential transcript usage across all three selection contrasts. In total, we identified 619 genes that coded for transcripts that were differentially used across all three selection contrasts, with 190 genes found between the low versus high contrast. Similar to the DE gene list, we see a majority (159 out of 190) of estimates falling between the range of -1 to 1 $\log_2$(cpm) (Figure 2.12B). In this case it was 159 genes of the total set of 190 genes falling between that range. The smallest effect size in this list was a $\log_2$(cpm) difference of -4.71 and the largest was a $\log_2$(cpm) difference of 2.19. This again is similar to the response we observed with the DE genes, which gives us evidence that the genetic response following artificial selection acts both on the regulation of gene expression, but also on the regulation of transcript usage of certain genes.

We also compared the list of DE genes that we identified with orthologs of previously identified genes associated with *Hymenoptera* sociality and human sociability. When looking across studies in *Hymenoptera* we saw few genes overlap with our gene list (Figures 2.18 – 2.20). Of the 20 genes that overlapped, only 7 genes were from the low versus high contrast, while the rest were from the low versus control or control versus high contrast. From the 8 *Drosophila* orthologs we could identify from the set of genes associated with changes in sociability in humans, we found that none of them showed a substantial change in gene expression across our selection treatments (Figure 2.17).

There are many reasons we see little overlap between the studies, with one being that we are simply capturing different aspects of social behaviour. We are comparing measures of social aggregation (our study), the evolution of eusociality (Woodard et al., 2011), early life experience (Wang et al., 2022) and responsiveness to social stimuli (Shpigler et al., 2017). While they are all social behaviours, there may not be too much overlap between them and our study. An interesting, yet unlikely reason that we do not see a shared response of gene expression differences across species may be due to a potential overarching regulatory gene or gene pathway that is conserved across species. The genetic underpinnings of the behaviour could also be extremely polygenic or perhaps different species use different genes to regulate social behaviour, but neither can be for certain without further research.

This chapter is not without caveats. The experiment was 25 generations of artificial selection from an initial population. It is possible that given the relatively low number of generations, the response we see from selecting upon sociability is largely from the segregating genetic variation that was present in the initial population, rather than new mutations arising in the populations. This is further compounded with the constrained population size as well as the effects of lab domestication and genetic drift. Alleles may have been lost or fixed in populations because of the stochastic effects of genetic drift. Larger population sizes can better deal with genetic drift as there is more genetic variation in the population. Additionally, the RNA sequencing samples consist of pooled heads of 16 individuals across all levels of our experimental design (selection treatments, sex, and experience). However, the gene expression we observe is an aggregated response across all 16 of these individuals and not at an individual-level resolution. We also sampled heads of adults, and we are not capturing developmental timepoints. There may be differential expression between the selection treatments earlier in development that we cannot capture but may be playing a large role in sociability.

Future work of the results in this chapter would be testing candidate genes identified in the DGE or DTU analyses, and this is currently taking place within the Dukas lab with promising results so far. If the work in candidate testing results in potential genes that, when knocked out result in a large reduction in sociability, there are potential experiments that can use artificial selection to restore the level of sociability found in the ancestral population. Given the evidence that there are many genes and transcript that show differential expression and usage, it would be interesting to see what the genetic response would be to go back to the ancestral state, and if the same genes were involved.

Overall, we identified genes and transcripts that are differentially expressed and used across selection treatments. We compared the set of genes identified with orthologous genes in other species associated with social behaviour and found little overlap. We also developed flexible pipelines to perform DGE and DTU analyses using linear mixed-effect models that can be customized for RNA sequencing experiments with complex experimental designs.

Chapter 3: Population Genomic analysis of artificially selected lineages

Introduction

In chapter 1 we discussed sociability and studies related to social behaviour, and in chapter 2 we examined the genetics of sociability and differences in gene regulation between treatments following artificial selection. While this may explain some of the underlying genetics associated with the evolution of sociability, it may not be the entire picture. Gene expression experiments and analyses provide information at a specific timepoint. In our experiment we sequenced adult heads, and there may be key developmental genes that are differentially expressed earlier in development which would not be captured in our sequencing. Additionally, there are cases where a new mutation arises in one population causing a change in protein function related to sociability, but the expression of the gene remains similar to across populations. A previously posed question by Robinson and Ben-Shahar (2002) is whether the changes in social behaviour is the result of changes in gene sequence, changes in gene expression, or both. Here we can ask similar questions in relation to *D. melanogaster* and the evolution of sociability as a result of artificial selection focusing on intraspecific variation. Is the response we see here largely due to differences in expression, differences in alleles or are they both contributing? It seems likely that they are both contributing as both can work together, for instance in a situation where alleles are changing in the population resulting in gene expression differences between the populations. In chapter 2, we have found evidence of differential gene expression and in this chapter, we wish to investigate if there are alleles in the population that are under selection in response to artificial selection. Building on the work of Scott et al. (2022), we obtained DNA sequencing reads of 16 samples following artificial selection and are using them to understand the genetic architecture of sociability, through a population genomics lens.

The evolution of social behaviour has been studied for decades (Hamilton, 1964a, 1964b; Queller, 1985), and as discussed in chapter 1, social behaviour is studied in various known social species. Some of these studies use population genomics approaches to better understand the evolution of social behaviour (Woodard et al., 2011). Here, Woodard et al. (2011) examined the rate of nonsynonymous to synonymous nucleotide substitutions between species of bees to identify genes that have a likely accelerated evolution in eusocial species relative to non-eusocial species. Another approach that we can take to understand the evolution of sociability between populations of fruit flies is looking at F_{ST}. F_{ST} is a measure of genomic differentiation between two populations that describes the reduction in heterozygosity due to differences between subpopulations (Cutter, 2019). We can use F_{ST} to measure genomic differentiation between two populations from the level of a single nucleotide polymorphism (SNP) all the way to entire genomes. If we were looking at a single polymorphic position, an F_{ST} measure of 1 indicates that the populations are diverged, where one of the two populations has entirely one nucleotide at the position, and the other population does not have that specific nucleotide at that position. Conversely, a value of 0 indicates that both populations have the exact same nucleotide at that position. While F_{ST} can be calculated at the level of individual SNP all the way to entire genomes, to identify positions in the genome under selection a genomic scan approach of calculating F_{ST} across predefined window sizes may be more appropriate as you are either losing too much information (at the level of the genome) or introducing noise to the results (at the level of individual SNP). There are also different approaches to defining what a region with high F_{ST} is. For example, high F_{ST} can be defined as values greater than three standard deviations higher than the mean F_{ST} for that population (Porto-Neto et al., 2013), or choosing a more outlier based approach and extracting the list of the highest 5% of F_{ST} values.

There are other population based statistics that can be considered to be analyzed for this chapter. One potential statistic that can be calculated is nucleotide diversity (θ_π or π) which is defined as the average proportion of nucleotide differences between all possible pairs of sequences (Hartl & Clark, 2007). Nucleotide diversity can help identify hard selective sweeps and regions under strong selection. Low nucleotide diversity in a region can indicate a hard selective sweep, as an allele rising fast to fixation will bring along nearby nucleotides and alleles, result in a lower nucleotide diversity. However, in our experiment, given the low number of generations and relatively low effective population size, it is unlikely that a new mutation or rare allele in the ancestral population has a large enough selection coefficient to fix rapidly in the population, and so an analysis of π may not be meaningful.

Other factors can also play a role in our analyses such as genetic drift. Drift is a stochastic force that can cause alleles to be lost or fixed, regardless of selective pressures. With an artificial selection experiment with relatively low populations sizes, drift can be a large factor. To combat that we have sequenced ancestral populations that we can compare with control lineages to determine sites that are affected by drift and omit those from sites under selection.

There are other approaches to account for drift in downstream analyses. For example the package ACER (Spitzer et al., 2020), implements a modified Cochran–Mantel–Haenszel (CMH) test to account for drift by adjusting assumptions in the null hypothesis (where no selection is occurring). Typically, it is assumed that the probability of sampling a given allele is the same when comparing allele frequencies between populations (Spitzer et al., 2020). However genetic drift can also cause allele frequency changes when selection is not happening, leading to alleles to be deemed as under selection when this is not the case, and so Spitzer et al. (2020) incorporate this increase in variation due to genetic drift in their test. To be cost effective, the sequencing for our study was done via pooled sequencing, where individuals are pooled together and then sequenced (Anand et al., 2016; Cutler & Jensen, 2010; Futschik & Schlötterer, 2010). Many downstream tools and programs (SNP callers) are designed for sequencing of individuals, so specific tools must be used to account for pooled sequencing samples. There are many different commonly used tools, but for our analyses we chose to use a new implementation for examining F_{ST}, called grenedalf (Czech et al., 2023). Grenedalf uses the same functionality and input data structures that are well established and used in the field but corrects for a number of software bugs that likely produced erroneous estimates of F_{ST} in common tools like popoolation2 (Kofler et al., 2011). Further, it is implemented in C++, greatly reducing the speed and computational cost required. On top of identifying potential regions of the genome under selection, a large motivation for this chapter is to provide a framework for performing population genomics analyses using pooled sequencing data. With the tools in this workflow (as well as scripts written by us) we present a flexible pipeline that can be easily adapted to include further downstream analyses.

Objectives

The objectives of this chapter are to identify genes that have responded to artificial selection for sociability and represent candidate genes influencing sociability. Specifically, we want to identify the genes and genomic regions that are under selection following artificial selection between the low and high sociable lineages. We will also compare the set of genes identified in this chapter to the list of differentially used transcripts and expressed genes to see if there are overlapping genes between the studies. With this, we can answer the question of whether the evolution of sociability is due to differential expression or differences in alleles?

Methods

Artificial Selection

Flies from the low, control and high sociable lineages were also sequenced after 25 generations of artificial selection, as well as ancestral flies that were not subjected to the arena or selection. DNA was extracted from the heads of flies and samples were pooled, consisting of 96 individuals per sample that were then sent to Génome Québec for sequencing. There were four lineages in each of the four groups (including four distinct replicates of the ancestor), for a total of 16 unique samples representing a total of 1,536 individuals. Each of the lineages were independent of one another, meaning that the first high lineage had no relation with the first low lineage (neither with control nor ancestor). We used four distinct replicates of the ancestor to increase our sequencing depth, in order to better represent the variation in our ancestral population. This allows us to capture any potential rare alleles that exist naturally in the population, and if that allele was found to be influencing sociability, we would know that it existed in the population and was not a new mutation.

Quality Checking

All the following computational analyses were performed using the Digital Research Alliance of Canada (formerly Compute Canada; www.alliancecan.ca). Please refer to table A4 (Appendices) for a list of software, scripts and parameters used for this chapter. From the sequencing, raw reads were obtained that underwent processing and alignment. We first used FASTQC (v0.11.9, Andrews, 2010) and MultiQC (v1.12, Ewels et al., 2016), to check quality of reads, and ensured that all samples had a mean Phred quality score of > 35. Adapters were trimmed using trimmomatic (v0.36, Bolger et al., 2014), with the parameters of leading and trailing set to “3” and run parameters set to “MAXINFO:20:0.2”. After trimming, samples were once again run through FASTQC via MultiQC to check for quality.

Read Mapping and Processing

Next, in preparation for aligning reads we indexed our reference *D. melanogaster* genome (version r6.38, Gramates et al., 2022) with bwa (v0.7.17, Li & Durbin, 2009), and proceeded to map our reads with bwa-mem (Heng, 2013). Following this step, we were left with 16 aligned samples in SAM file format. From the resulting SAM files, we convert them to the BAM file format and filtered for only reads with a MAPQ (map quality) score of > 30 with the command “samtools view”, using samtools (v1.15, Danecek et al., 2021). Using awk and samtools (v1.15, Danecek et al., 2021), the core genome was extracted (chromosomes 2L, 2R, 3L, 3R, 4 and X) and then duplicate alignments were marked and removed using the commands “samtools fixmate” and “samtools markdup”. Next, we used Picard (v2.26.3, Broad Institute, 2019) to add read groups to our samples and then used GATK (v3.8, McKenna et al., 2010) to mark and realign around indels. We merged the replicate lineages of each treatment into a single file with the command “samtools merge”, which resulted in 4 total files (Ancestor, Low, Control and High). From here we created a single mpileup containing all the samples using the command “samtools mpileup”. The full pipeline of steps up until the creation of the mpileup can be visualized in figure 3.1. The full samtools pipeline can be visualized in figure 3.2.

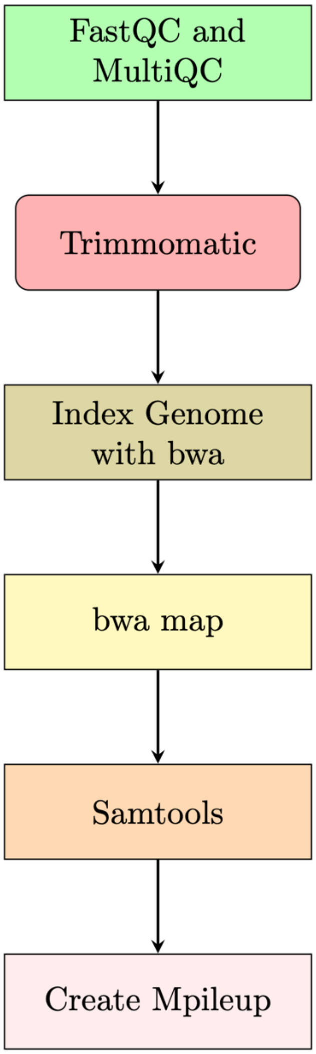

Figure 3.1. Flowchart detailing steps and software used for population genomics analysis. The flowchart details steps from raw sequenced reads until the mpileup is made. Samtools refers to figure 3.2, detailing all samtools commands used.

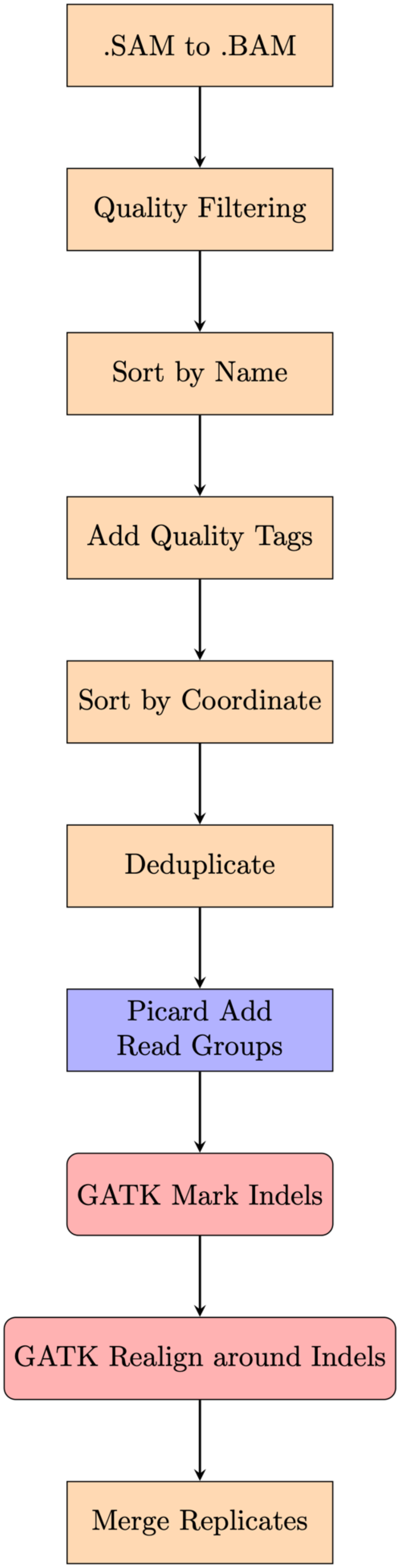
.SAM to .BAM
Quality Filtering
Sort by Name
Add Quality Tags
Sort by Coordinate
Deduplicate
Picard Add Read Groups
GATK Mark Indels
GATK Realign around Indels
Merge Replicates

Figure 3.2. Flowchart detailing steps and software used for population genomics analysis using the samtools package. The flowchart details steps from mapped .SAM files until replicates are merged. The full pipeline can be observed in figures 3.1 and 3.3.

SNP Calling

With the mpileup, single nucleotide polymorphisms (SNPs) were called using PoolSNP (Kapun et al., 2020), which is a heuristic SNP caller designed to be used with pooled sequencing samples. In order for a position to be called, the position had to have a minimum coverage of 25, a maximum coverage within the 98% percentile of coverage for a given sample (to account for repetitive regions), a minimum cumulative minor allele count of 10 and minimum minor allele frequency of 1%. Following SNP calling, we were left with a VCF (variant call format) file. At this stage we filtered out repetitive regions and indels from the VCF. We did this by using RepeatMasker to identify repetitive regions in the genome from a reference genome and list of known transposons, and then used a script used in Kapun et al. (2020) to identify indels from the mpileup. Then, we filtered out the indels and repetitive regions using another script from Kapun et al. (2020). We filter out indels as there may be gaps in the alignment around indels which can cause false-positive SNPs to be called (Li & Homer, 2010) and repetitive regions can result in incorrect mapping to the reference genome, creating false-positive SNPs (Shen et al., 2010). The last step of filtering was to remove problematic regions of the genome known as the ENCODE blacklist (Amemiya et al., 2019) with bedtools (v2.30.0, Quinlan & Hall, 2010). The blacklist includes regions of the *Drosophila* genome that are potentially unannotated repeats in the genome that can bias results towards these regions (Amemiya et al., 2019). If these sites were not filtered out, results from these sites may have been attributed to biological variation (Amemiya et al., 2019).

F_{ST}

Next, we went to calculate genomic differentiation across the genomes using the statistics F_{ST}. We used the command “grenedalf sync” to convert our unfiltered mpileup into a sync file. A sync file (synchronized file) is a file that essentially contains positional information, the reference allele at that position and allele frequencies of each group. In order to filter out indels and repetitive regions that we had previously filtered out, we wrote a script that subsets a sync file based on positional information from a VCF. We used this with the above mentioned sync file and filtered VCF to end up with a sync file that has filtered out indels and repetitive regions and only has identified SNPs. From here we used the command “grenedalf fst” to calculate pairwise F_{ST} for all contrasts of ancestor, low, control and high in 5,000 basepair sliding windows. For a given contrast of interest (Low versus Control and Low versus High) we chose an outlier based approach of extracting the windows with the top 5% of F_{ST} values within the contrast. We did this separately for windows in the X chromosome and for windows in the autosomes. We did this to account for the increased variation in the X chromosome results due to sampling (as we are sampling 3/4 of the X chromosomes compared to the autosomes) which can influence and bias the outliers to the X chromosome. Following this, we merged the outliers for the X chromosome and autosomes back together in each contrast. This outlier approach is less stringent than looking for F_{ST} values greater than 3 standard deviations from the mean, and we chose this because we are also comparing these regions to SNPs identified by a CMH test and do not want to potentially exclude important SNPs. Given that the there was no artificial selection acting upon the control lineages, the F_{ST} between the ancestor and control is the combination of genetic drift and lab domestication (selection occurring as a result of being in the lab versus in the wild) which we can use to account for lab domestication in our low versus control and control versus high comparisons. To account for this lab domestication, we identified the windows with the top 5% of F_{ST} values in the Ancestor versus Control contrast and subsetted those windows out of initial list of windows in the low versus control and control versus high lineages. Given that the comparisons between the ancestor and low/high are not entirely due to lab domestication and genetic drift, we cannot take a similar approach for the low versus high contrast.

Cochran–Mantel–Haenszel (CMH) Test

To identify positions that are potentially under selection, we utilized a modified CMH test. The CMH test is an extension of the χ^2 test (Cochran, 1954; Mantel & Haenszel, 1959) that is used in population genomics to compare allele frequencies and identify positions in the genome under selection between two populations (Wiberg et al., 2017). However, in pooled sequencing experiments one of the assumptions that allele counts are independent draws is violated (Wiberg et al., 2017), and due to genetic drift, the assumption that the probability of sampling an allele is the same between populations is violated (Spitzer et al., 2020). To account for these violations of assumptions, Spitzer et al. (2020) implemented a modified CMH test that accounts for genetic drift and pooled sequencing in the R package that we used, ACER. Rather than using the sync file obtained previously from the merged replicates, we needed to generate a new sync file where replicates are not merged. To do so, we went through our pipeline as usual but omitted the merging step. For a given contrast, we first split our dataset into two groups, allele frequencies from the X chromosome and allele frequencies from the autosomes. We did this to account for the increased variation in the X chromosome results due to sampling (as we are sampling 3/4 of the X chromosome compared to the autosome) as well as since we are calculating effective population size, this value will change between the X chromosome and autosomes. For both the dataset with the X chromosome allele frequencies and the autosomal allele frequencies we used ACER to identify positions in the genome that are under selection between the low versus control, control versus high and low versus high sociability contrasts. Each of the contrasts were run separately and the output was p-values associated with positions along the genome. We then applied a p-value adjustment, using the R function “p.adjust()” with “method = ‘BY’ ”, referring to the Benjamini and Yekutieli method of controlling false discovery rate (Benjamini & Yekutieli, 2001). We then subsetted the list for the lowest 1% of adjusted p-values, which left us with positions of the genome under selection. Following this subset, we then merged our results for the X chromosome and autosomal chromosomes back together. We chose the lowest 1% of adjusted p-values as this provided us with a large list (> 20,000) of outlier positions that are potentially under selection that we could compare back to our windows with the top 5% of F_{ST} values.

SNP annotation and extraction

We chose to compare the list of top 5% F_{ST} values with the positions from the CMH test that corresponded to the lowest 1% of adjusted p-values. In both cases we chose an outlier based approach to identify regions/positions, which if used exclusively, may not be the most sensible approach as there is a chance you are introducing noise in the list by solely choosing the highest (or lowest) values. Instead, we chose to see what positions are identified by both analyses (F_{ST} and CMH) as the intersection between the two methods should in theory result in largely positions that are under selection. We created a list of SNPs that overlapped between the two lists of high F_{ST} and statistically significant CMH. To do so, we took the previously generated lists of F_{ST} and CMH and manually converted them into bed files. Then, using the command “bedtools intersect” (Quinlan & Hall, 2010), we generated a bed file that included only the regions of the genome where the F_{ST} window and CMH position overlapped. We then annotated this list using SnpEff (Cingolani et al., 2012). In order to use SnpEff, we needed to input a VCF file, so similarly to subsetting our sync file from a VCF, we created a script that subsetted our initial VCF with only positions from our bed file. Another benefit of performing this subsetting is to filter out positions identified in the CMH test that do not appear in the merged replicate VCF, as the positions in the CMH list come from the unmerged VCF. SnpEff also provides an assessment of impact of the variant identified which are low, moderate, and high impact. We extracted the SNPs that had both high and moderate impact variants. If a given contrast had no SNPs labelled as high effect variants, we used the list of moderate and vice versa. We then created a list of the genes associated with the overlapping SNPs in a given contrast and compared that list to the list of differentially used transcripts or expressed genes (see Chapter 2) of the same contrast to see if any genes overlapped between the two analyses. We also created a list of genes associated with the SNPs with predicted high and moderate impact variants for a given contrast and compared that list to the list of differentially used transcripts or expressed genes of the same contrast. The full pipeline of all steps following the creation of the mpileup can be observed in figure 3.3.

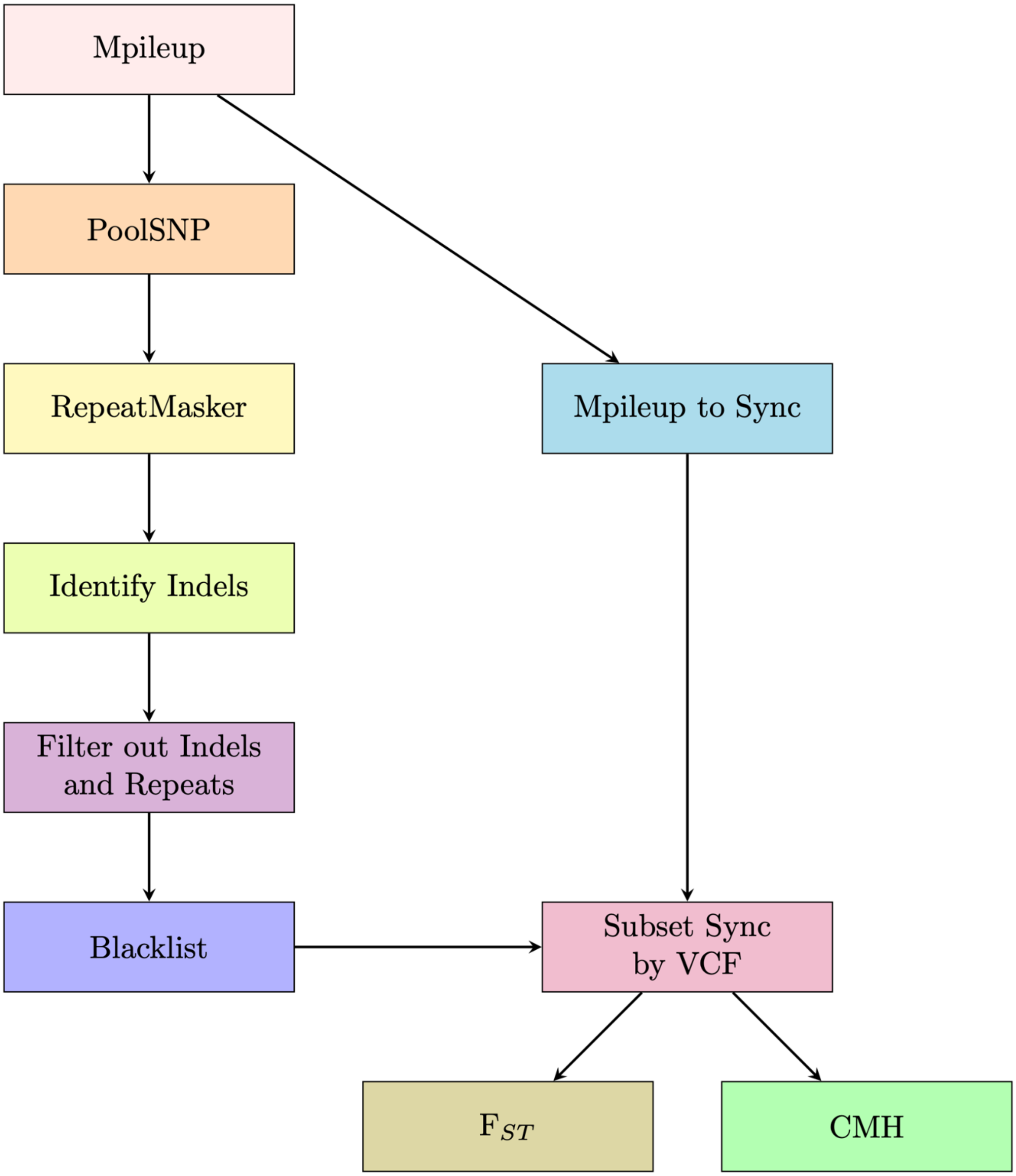

Figure 3.3. Flowchart detailing steps and software used for population genomics analysis following the creation of the mpileup. The flowchart details steps from

mpileup until F_{ST} and CMH are calculated. Branch points are independent of one another.

We also compared the overlapping genes between three contrasts (low versus high, low versus control, and control versus high) and set up a simulation to evaluate whether the number of overlapping genes between our lists of genes from the three contrasts were more or less common than expected by random chance alone. Each run was between two contrasts and consisted of 1,000 simulations in which we randomly sampled two sets of genes (one for each contrast) without replacement and independently of each other that corresponded with the total number of genes identified in each contrast. We sampled from the total number of genes that we mapped to, and observed how many genes would be common between the two. The three runs we performed were between the low versus high and low versus control contrast, the low versus high and control versus high contrast, as well as the low versus control and control versus high contrast.

We extracted the list of genes associated with the positions identified following both of the F_{ST} and CMH analyses for all contrasts. We used the previously mentioned bed file containing F_{ST} and CMH positions and intersected each with the initial VCF, using the bedtools command “intersect”. Following this, we annotated the resulting VCF using SnpEff to identify the genes associated with the positions in the VCF.

GO Analysis

We performed a gene ontology (GO) analysis using the package topGO (v2.48.0, Alexa & Rahnenfuhrer, 2022) on the set of genes identified from SNPs that overlapped between the list of high F_{ST} and significant CMH. We performed this gene enrichment separately for all three sociability contrasts (low versus high, low versus control and control versus high). Here, we set the minimum number of genes per GO term to 5 and used Fisher's exact test. We then adjusted the resulting p-values for multiple comparisons by using the function “p.adjust” with “method = ‘BH’”, for the Benjamini-Hochberg method of multiple comparisons correction.

Results

F_{ST} Results

We generated Manhattan plots to visualize the genomics patterns observed F_{ST} in the Ancestor versus Control (Figure 3.4), Ancestor versus Low (Figure 3.5), Ancestor versus High (Figure 3.6), Low versus High (Figure 3.7), Low versus Control (Figure 3.8) and Control versus High (Figure 3.9) contrasts. When accounting for lab domestication and drift by subsetting out the windows that overlap in the ancestor versus control with either of the low versus control and control versus high contrasts, we see a resulting 826 and 776 windows remain, respectively. Given that there are the same number of windows in all of our contrasts, we see that the top 5% of F_{ST} values represent 1,236 windows (221 windows for the X and 1,015 windows for the rest of the autosomes) in all contrasts. Using solely these windows by themselves may not be meaningful, as there will always be a set amount of outlier windows even if there is nothing meaningful about the outliers. However, when these windows are intersected with positions of the genome under selection (CMH results; see below) we are able to potentially identify SNPs and genes in the genome under selection. The full list of genes identified to be associated with the top 5% of F_{ST} values for all contrasts can be found in the GitHub repository that is linked in the appendices.

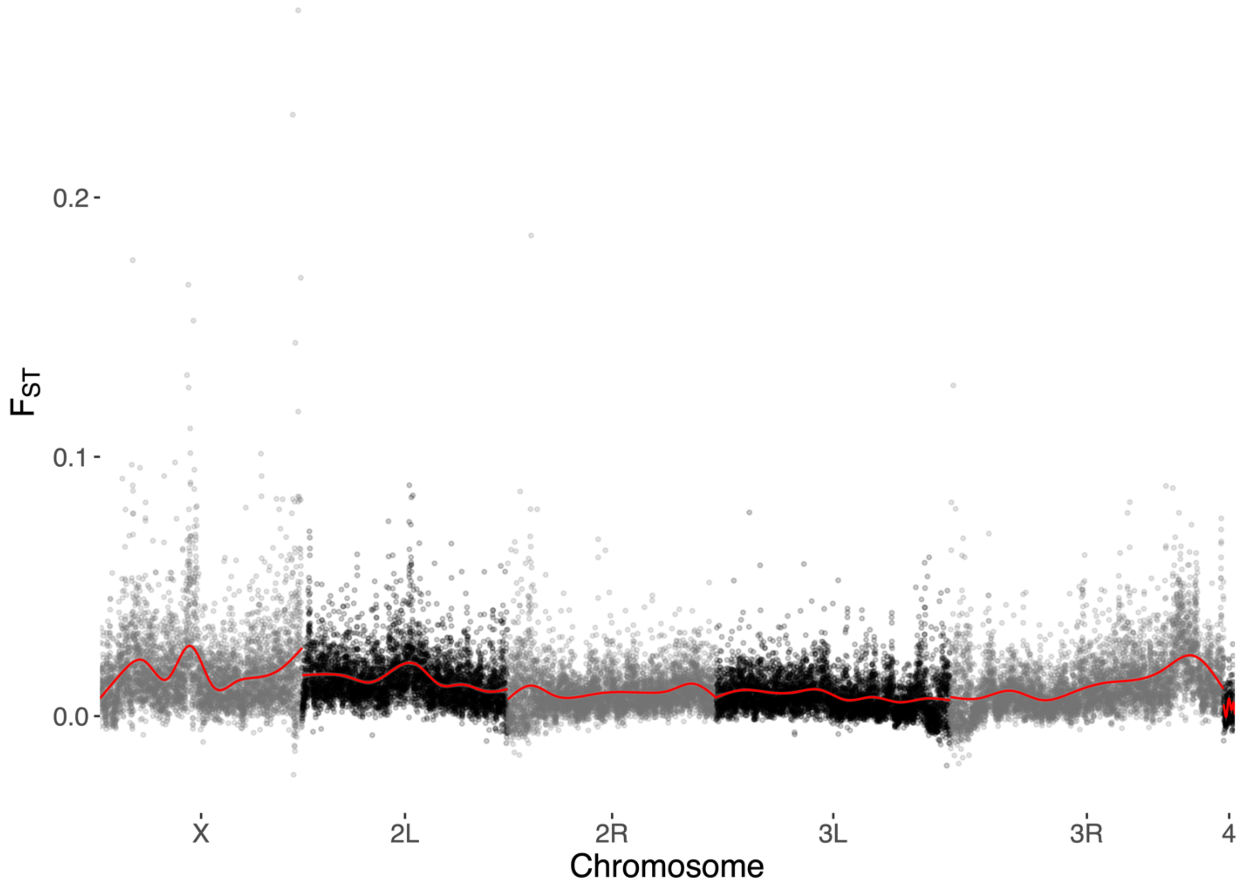

Figure 3.4. Manhattan plot of genomic differentiation (F_{ST}) in 5,000 basepair windows across the genome for Ancestor versus Control contrast. X axis is the genomic position with chromosome denoted. Y axis is genomic differentiation (F_{ST}) in 5,000 basepair windows. Red lines are smoothed trendlines using a GAM of F_{ST} for a given chromosome.

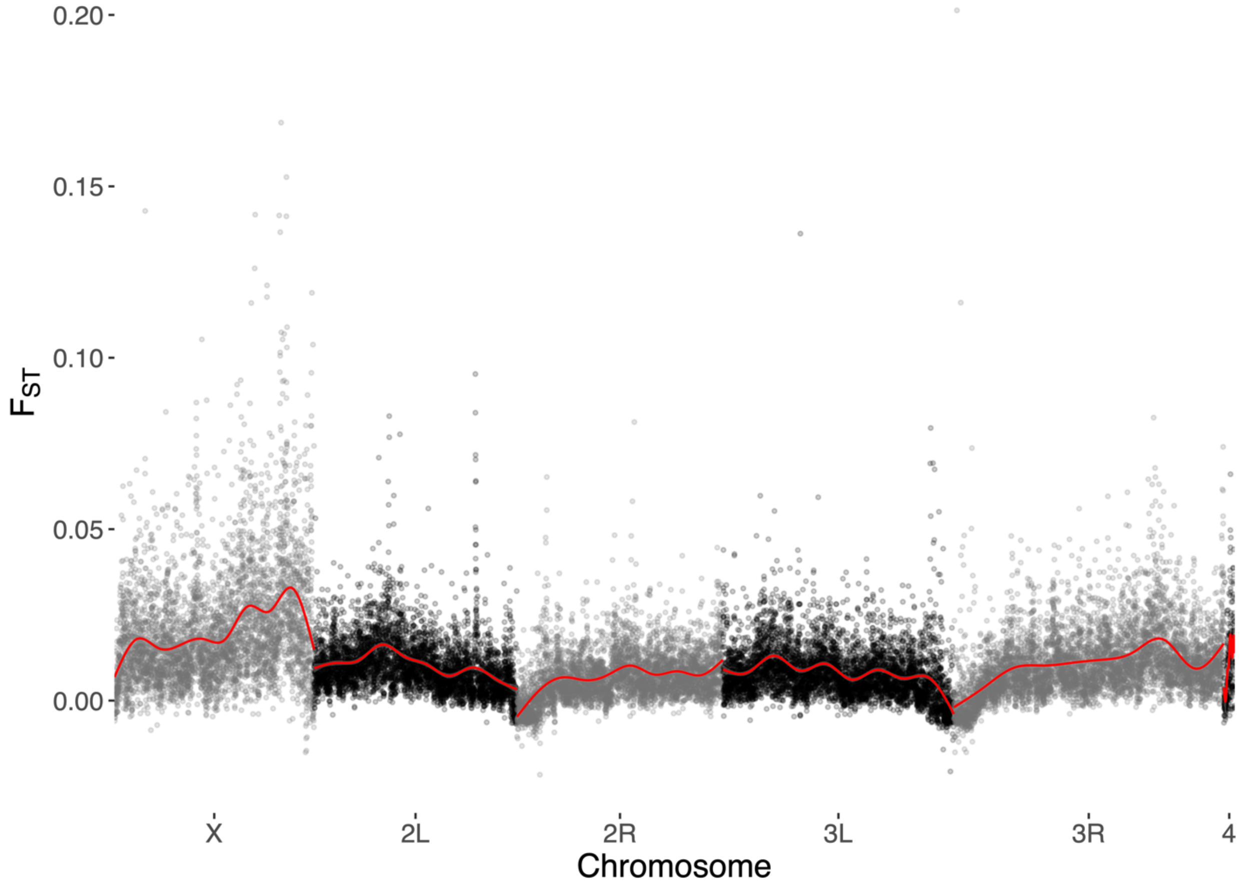

Figure 3.5. Manhattan plot of genomic differentiation (F_{ST}) in 5,000 basepair windows across the genome for Ancestor versus Low contrast. X axis is the genomic position with chromosome denoted. Y axis is genomic differentiation (F_{ST}) in 5,000 basepair windows. Red lines are smoothed trendlines using a GAM of F_{ST} for a given chromosome.

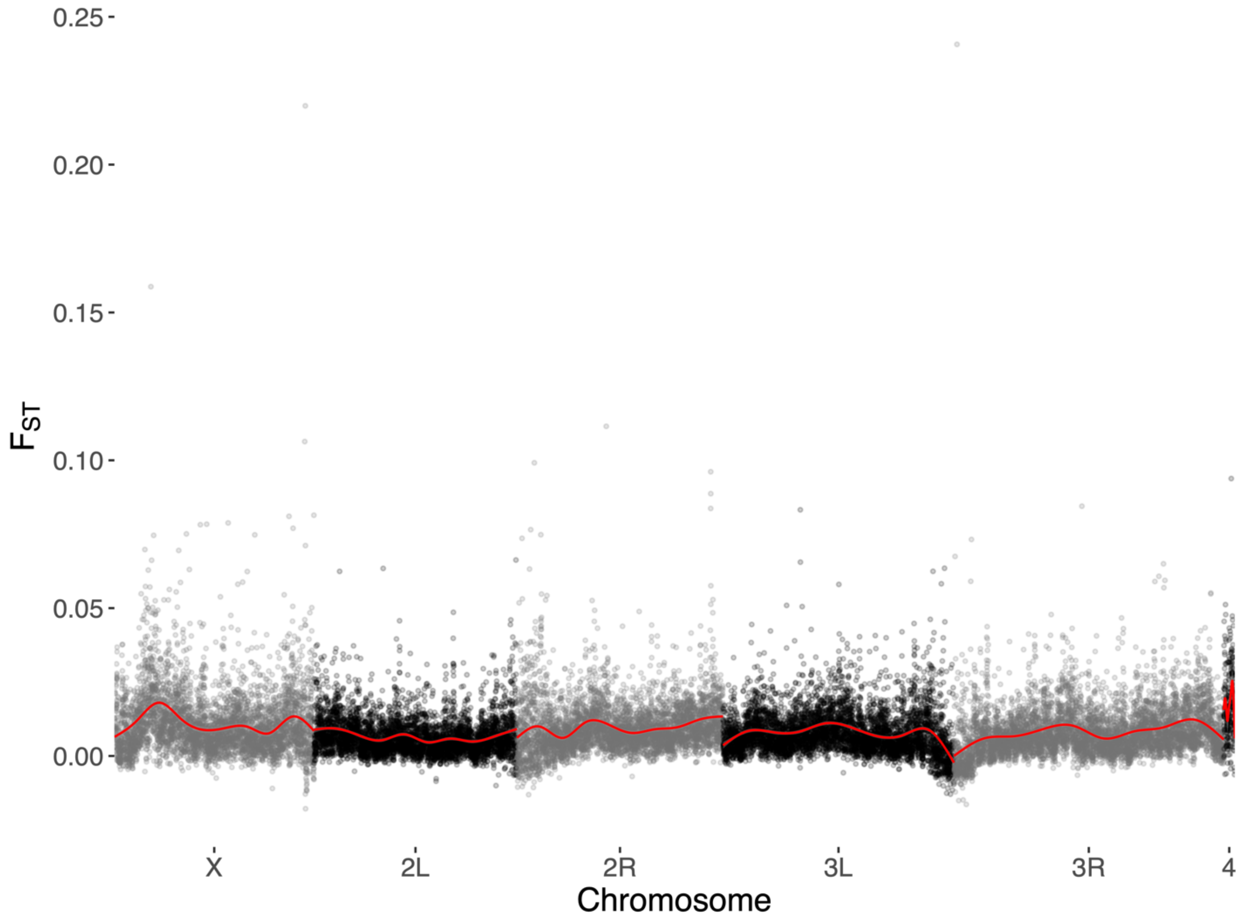

Figure 3.6. Manhattan plot of genomic differentiation (F_{ST}) in 5,000 basepair windows across the genome for Ancestor versus High contrast. X axis is the genomic position with chromosome denoted. Y axis is genomic differentiation (F_{ST}) in 5,000 basepair windows. Red lines are smoothed trendlines using a GAM of F_{ST} for a given chromosome.

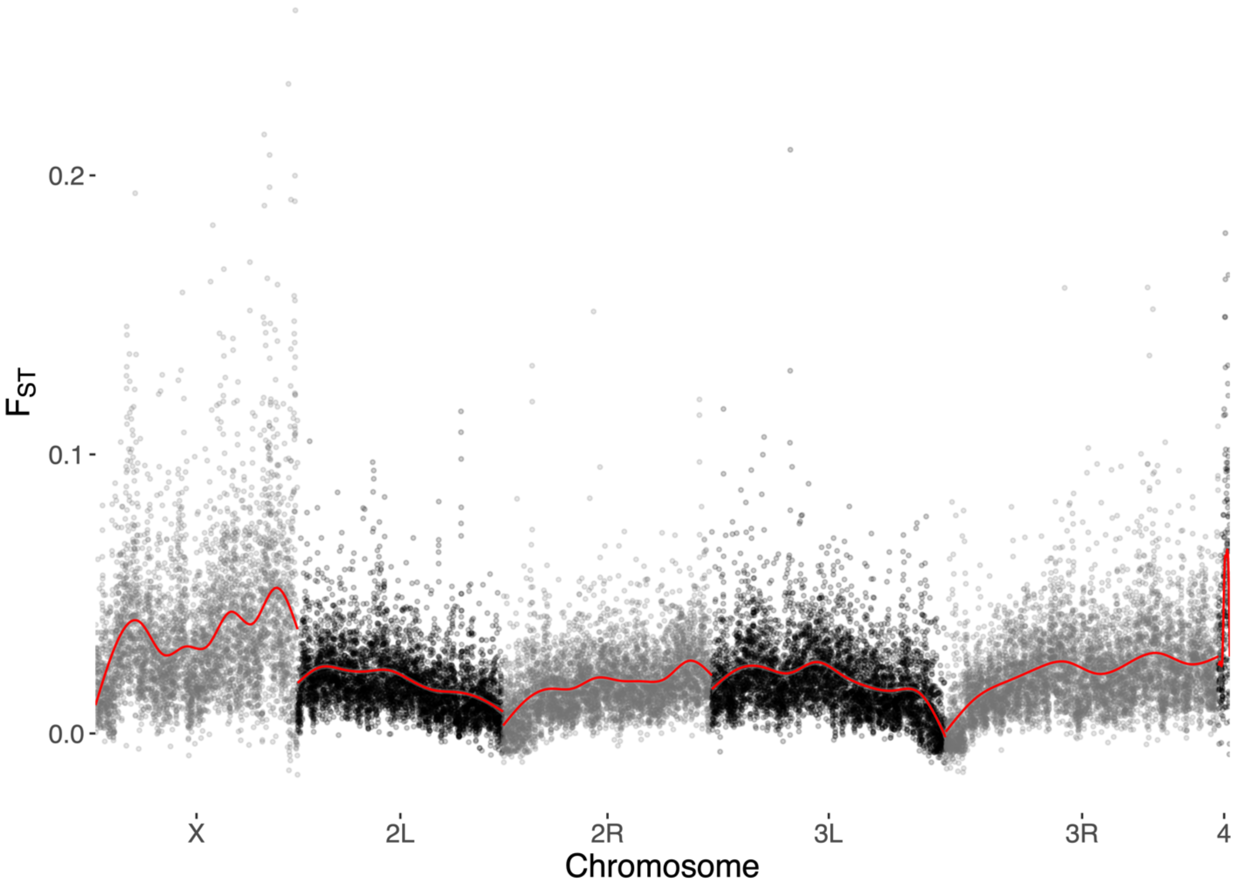

Figure 3.7. Manhattan plot of genomic differentiation (F_{ST}) in 5,000 basepair windows across the genome for Low versus High contrast. X axis is the genomic position with chromosome denoted. Y axis is genomic differentiation (F_{ST}) in 5,000 basepair windows. Red lines are smoothed trendlines using a GAM of F_{ST} for a given chromosome.

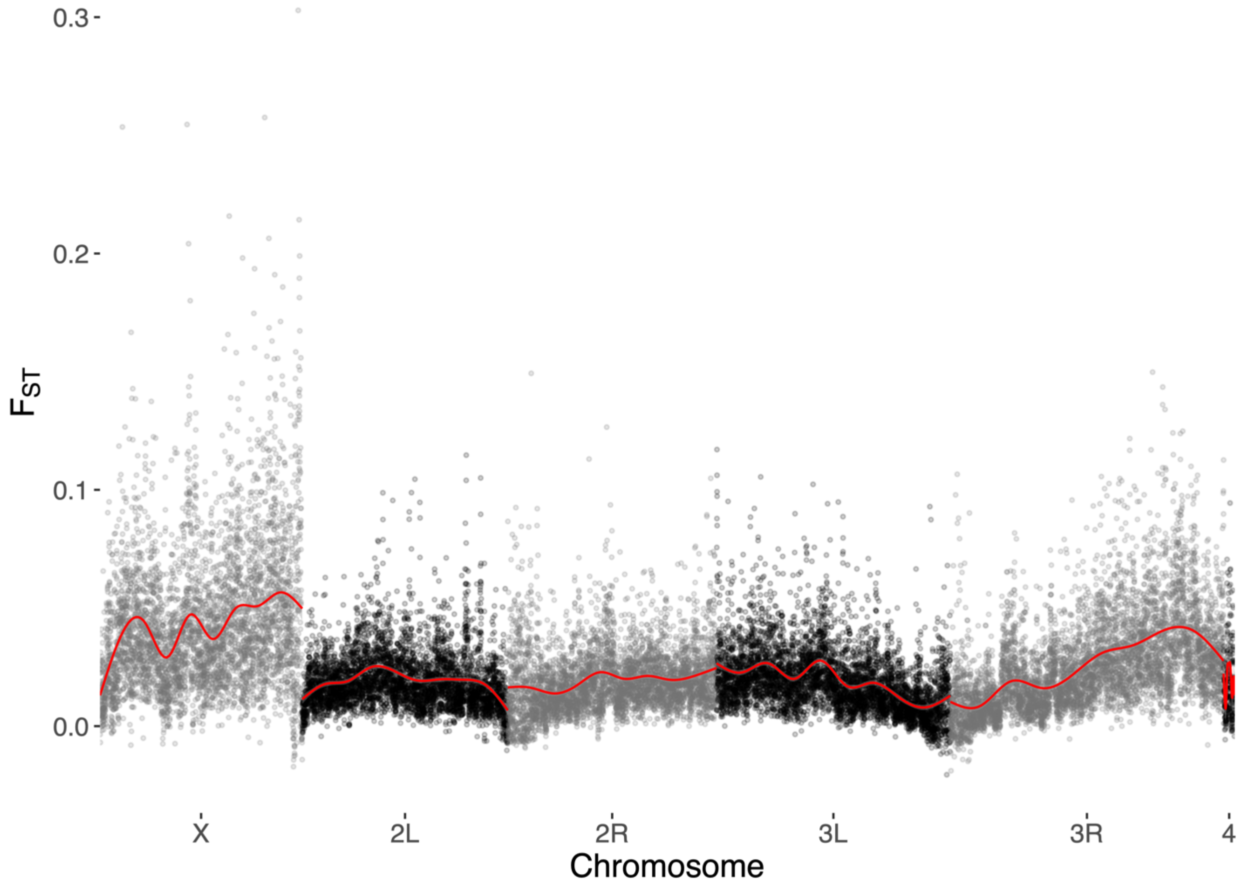

Figure 3.8. Manhattan plot of genomic differentiation (F_{ST}) in 5,000 basepair windows across the genome for Low versus Control contrast. X axis is the genomic position with chromosome denoted. Y axis is genomic differentiation (F_{ST}) in 5,000 basepair windows. Red lines are smoothed trendlines using a GAM of F_{ST} for a given chromosome.

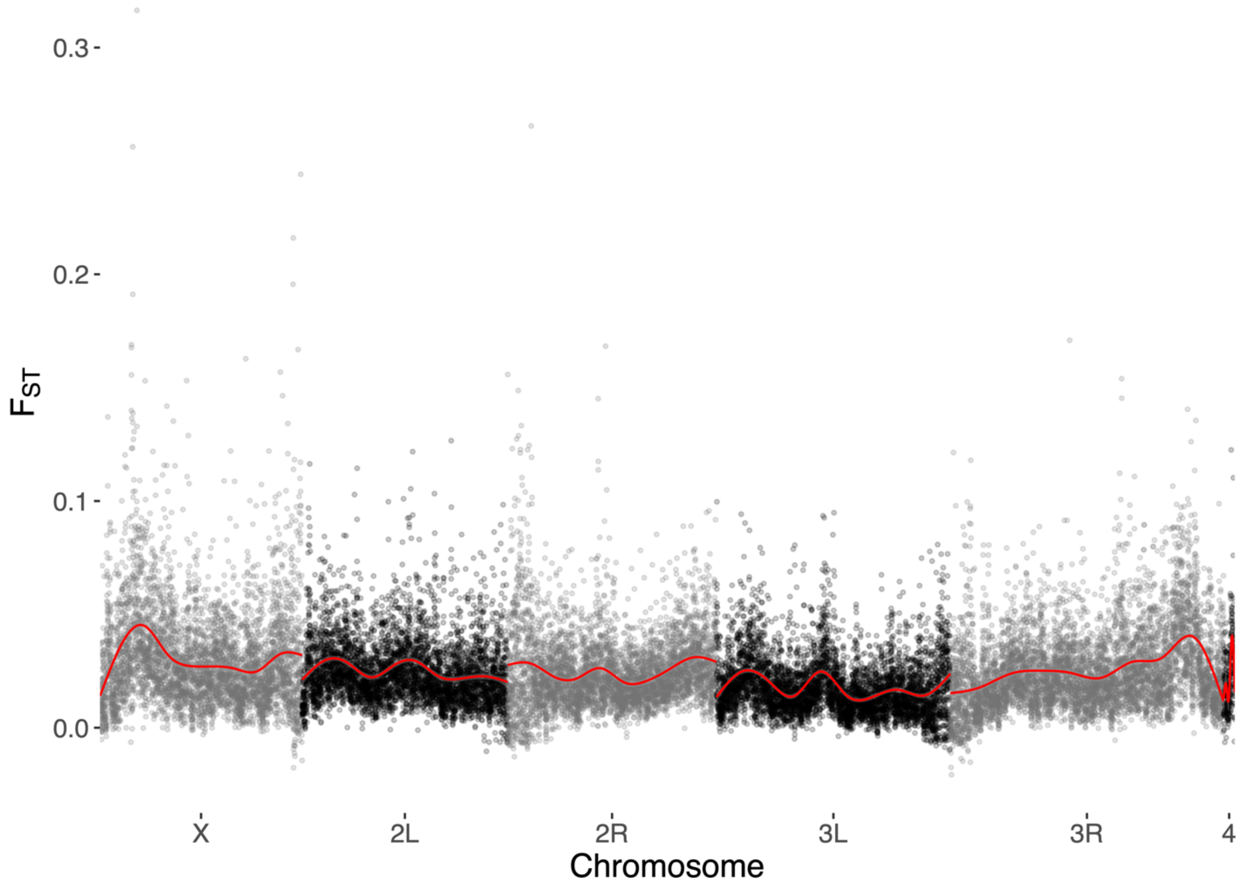

Figure 3.9. Manhattan plot of genomic differentiation (F_{ST}) in 5,000 basepair windows across the genome for Control versus High contrast. X axis is the genomic position with chromosome denoted. Y axis is genomic differentiation (F_{ST}) in 5,000 basepair windows. Red lines are smoothed trendlines using a GAM of F_{ST} for a given chromosome.

CMH Results

Following the CMH test between the low versus high contrast, we found 21,479 base positions (2,661 on the X chromosome and 18,818 on the autosomes) in the top 1% of adjusted p-values. In the control versus high contrast, we found 21,779 positions (2,702 on the X chromosome and 19,077 on the autosomes) in the top 1%, and in the low versus control contrast we found 21,541 positions (2,639 on the X chromosome and 18,902 on the autosomes) in the top 1%. The full list of genes identified to be associated with the top 1% of adjusted p-values for all contrasts can be found in the GitHub repository that is linked in the appendices.

Overlapping Results

Full comparisons of all overlapping results can be found in Table 3.1. After intersecting with bedtools for the low versus high comparison, we found 1,132 overlapping positions between the high F_{ST} list and lowest 1% of adjusted p-values (CMH test). After subsetting our merged VCF with the intersecting positions, we were left with 1,084 SNPs that were identified in our VCF. Following annotation of the SNPs we found 64 SNPs with a predicted moderate effect variant and 4 SNPs with a predicted high effect variant. In total we found 333 genes associated with the 1,084 positions and of those 333 genes, 9 genes were found to overlap with our DE genes and 0 genes with an associated moderate effect variant were found in the DE gene list. When comparing the transcripts associated with the 1,084 SNPs, we found that 13 transcripts overlap with our list of differentially used transcripts. We also plotted the non-reference allele frequency changes across selection treatments for two genes with a SNP with a predicted moderate effect, *CG34049* and *stv* (Figure 3.10). In *CG34049*, we see a missense mutation causing an amino acid change from an asparagine to lysine. In *stv*, we also see a missense mutation causing an amino acid change from a valine to an alanine.

Table 3.1. Overlapping results follow F_{ST} and CMH analysis. For each contrast of treatments, we report the number of positions overlapping between F_{ST} and CMH lists, the number of moderate and high predicted effect variants, number of genes associated with overlapping positions, the number of genes found to overlap with the appropriate DE gene list, the number of transcripts found to overlap with the appropriate DTU transcript list.

Contrast	Overlap between F_{ST} and CMH	Moderate and (high) effect SNPs	Genes	Overlapping DE genes (moderate or high)	Overlapping transcripts
Low vs High	1,084	64 (4)	333	9 (0)	13
Low vs Control	1,027	55 (0)	244	4 (2)	16
Control vs High	523	58 (1)	179	2 (0)	6
Ancestor vs Low	1,418	83 (1)	379	8 (1)	13
Ancestor vs High	956	80 (0)	316	6 (1)	13

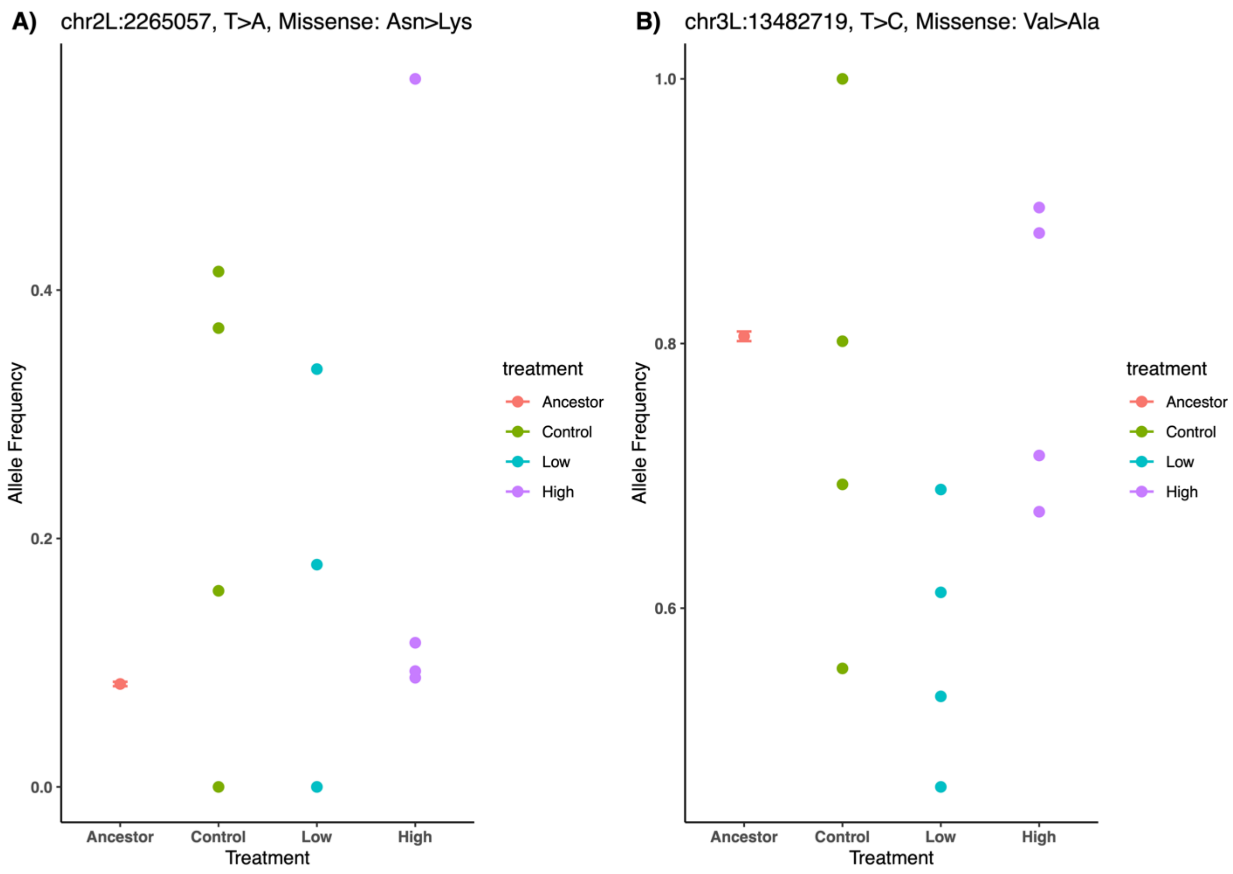

FIGURE 3.10. Non-reference allele frequency plot for two SNPs associated with a predicted moderate effect variant in the low versus high contrast. A) The SNP associated with the gene *CG34049*, which causes a missense mutation changing an asparagine to a lysine. **B)** the SNP associated with the gene *stv*, which causes a missense mutation changing a valine to an alanine. The X axis represents the selection treatment of either Ancestor, Control, Low or High. The Y axis is the non-reference allele frequency. The value for the ancestor is the mean non-reference allele frequency for all 4 ancestral pools, with the standard error plotted as well. The values for the Control, Low and High are the non-reference allele frequency for each of the 4 independent lineages associated with the selection treatment. Each plot title includes the location of the SNP, the nucleotide substitution, and the resulting amino acid change.

For the low versus control comparison, we found 1,027 overlapping positions between the high F_{ST} list and lowest 1% of adjusted p-values (CMH test), with 977 SNPs remaining after subsetting the merged VCF with the overlapping positions. Following SNP annotation, 55 SNPs were predicted to be moderate effect variants and 0 SNPs were predicted to be a high predicted effect variant. In total 244 genes were associated with the 977 positions, and 4 were found to overlap with our DE genes and 2 genes were associated with a moderate effect variant that were found in the DE list. When comparing the transcripts associated with the 977 SNPs, we found that 16 transcripts overlap with our list of differentially used transcripts.

For the control versus high comparison, we found 545 overlapping positions between the high F_{ST} list and lowest 1% of adjusted p-values (CMH test), with 523 SNPs remaining after subsetting the merged VCF with the overlapping positions. Following SNP annotation, 58 SNPs were predicted to be moderate effect variants and 1 SNP was predicted to be a high predicted effect variant. In total 179 genes were associated with the 523 positions, and 2 were found to overlap with our DE genes and no genes that were associated with either moderate or high effect variants were found in the DE list. When comparing the transcripts associated with the 523 SNPs, we found that 6 transcripts overlap with our list of differentially used transcripts.

For the ancestor versus low comparison, we found 1,486 overlapping positions between the high F_{ST} list and lowest 1% of adjusted p-values (CMH test), with 1,418 SNPs remaining after subsetting the merged VCF with the overlapping positions. Following SNP annotation, 83 SNPs were predicted to be moderate effect variants and

1 SNP was predicted to be a high predicted effect variant. In total 379 genes were associated with the 1,418 positions, and 8 were found to overlap with our DE genes and 1 gene was associated with moderate effect variants that was also found in the DE list, *CG14200*. When comparing the transcripts associated with the 1,418 SNPs, we found that 18 transcripts overlap with our list of differentially used transcripts.

For the ancestor versus high comparison, we found 989 overlapping positions between the high F_{ST} list and lowest 1% of adjusted p-values (CMH test), with 956 SNPs remaining after subsetting the merged VCF with the overlapping positions. Following SNP annotation, 80 SNPs were predicted to be moderate effect variants and no SNPs were predicted to be a high effect variant. In total 316 genes were associated with the 956 positions, and 6 were found to overlap with our DE genes and 1 gene that was associated with a moderate effect variant that was also found in the DE list, *rdog*. When comparing the transcripts associated with the 1,266 SNPs, we found that 13 transcripts overlap with our list of differentially used transcripts.

We also performed simulations to see if the number of overlapping genes between the three gene lists of low versus high, low versus control, and control versus high were more or less common than by chance alone. For the overlapping genes between the low versus high and low versus control contrast, we randomly selected 333 and 244 genes and found that after 1,000 simulations, a maximum of 17 overlapping genes (95% quantile of 10). Between the low versus high and control versus high we randomly sampled 333 and 179 genes and found a maximum of 14 overlapping genes (95% quantile of 8). Between the low versus control and control versus high we randomly sampled 244 and 179 genes and found a maximum of 12 overlapping genes (95% quantile of 6).

GO Analysis

Following GO analysis, we found 95 GO terms that are deemed significantly overrepresented in our gene list from SNPs identified between the high F_{ST} and significant CMH lists in the low versus high sociability contrast (Table A5), including terms like motor neuron axon guidance, olfactory learning, visual perception as well as long and short-term memory. Between the low versus control sociability contrast, we found 85 GO terms that were overrepresented (Table A6), with enrichment of terms such as motor neuron axon guidance, visual behaviour, and olfactory learning. Between the control versus high sociability contrast, we found 33 GO terms overrepresented (Table A7) with terms such as synapse organization, synaptic signaling and regulation of axonogenesis being enriched.

Discussion

The main objective of this chapter was to identify genes and regions of the genome that are under selection following artificial selection of sociability. Following our analyses, we found no regions of the genome with genes of large effect that contribute to the change in sociability (Figures 3.4 – 3.9). Specifically, we found no genomic windows with a large enough F_{ST} that would suggest there is a gene or region of large effect contributing to the divergence in sociability. While the next logical step may be to suggest the response that we see is polygenic, we cannot claim that for certain, as the response we see can also be attributed to low population size coupled with genetic drift and lab domestication. We may only be seeing segregating alleles from the initial population rather than new mutations and genetic drift can be acting upon those alleles, resulting in the response that we see.

In total, we identified lists of genes associated with SNPs under selection in our contrasts between artificial selection treatments. In the low versus high contrast, we identified 324 genes, in the low versus control we found 245 genes, and in the control versus high we found 184 genes. Between the three lists of genes there was a total 646 genes represented, but only 14 genes overlapping between the three lists (Figure 3.11). The largest overlap between any two lists was 51 genes, which was between the low versus high and the control versus low contrasts (Figure 3.11). When comparing the number of overlapping genes to the simulation we performed to see the number of overlapping genes by random chance alone, we see that the overlapping genes we identified are all higher than the 95% quantile. This is in line with (Scott et al., 2022) showing that there was a larger response to selection in the low sociable lineages compared to the high sociable lineages (Figure 1.2), which would explain why we see more shared genes between the low versus high and low versus control, compared to the control versus high contrast.

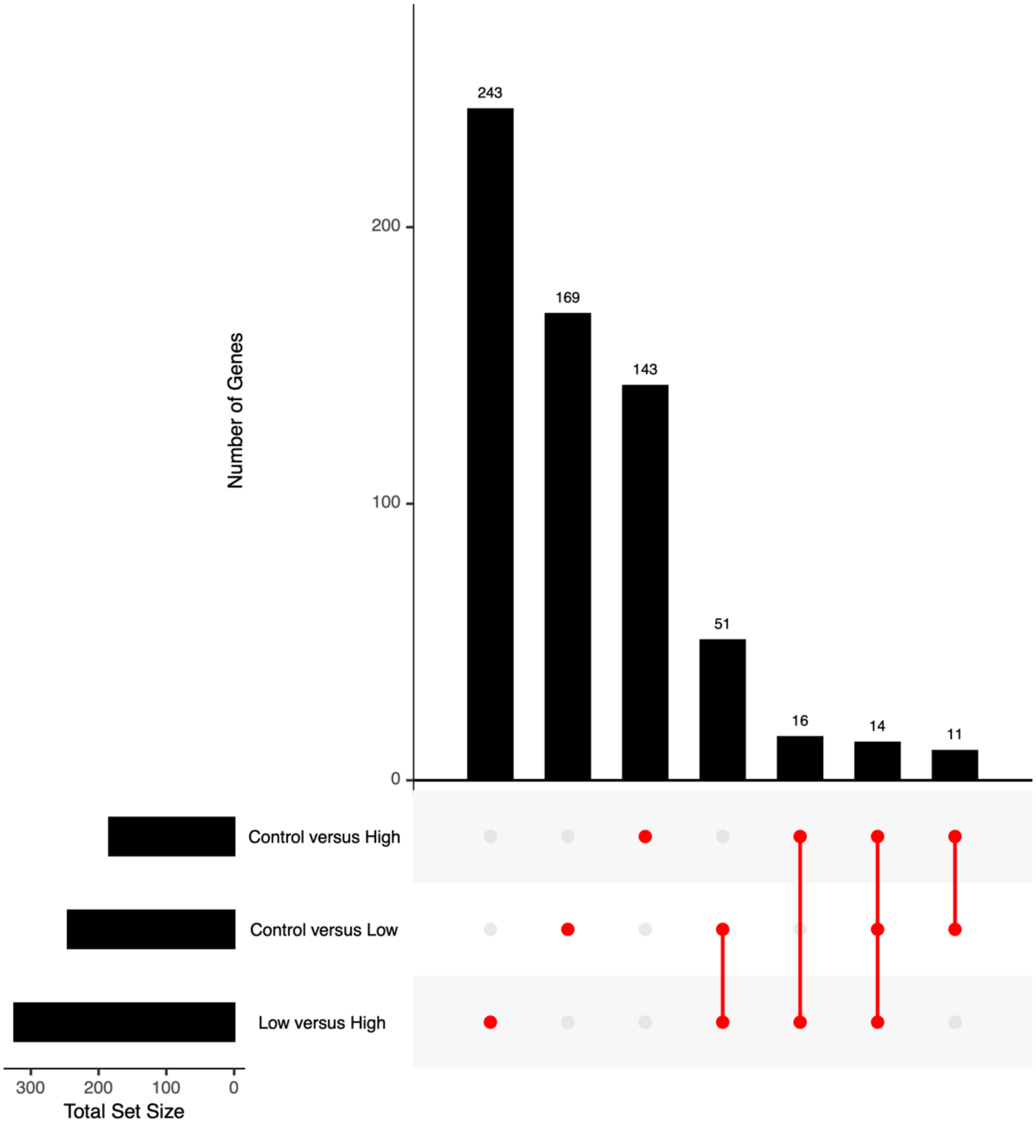

Figure 3.11. Upset plot showing the overlap between the gene lists obtained from the overlap between high F_{ST} and statistically significant CMH within a selection contrast of interest. Gene lists are selection contrasts of Low Versus High, Control Versus High and Low versus Control.

In addition to the fact that we see more genes in common between our lists than by random chance, we see evidence that we are identifying genes involved in sociability in the list of enriched gene ontology terms. We see many terms related to the neuron and behaviour, such as visual behaviour, circadian behaviour, motor neuron axon guidance and synaptic signalling.

We also compared our list of differentially used transcripts and expressed genes to the gene lists of genes associated with SNPs under selection for all three of our contrasts. Between the list of differentially expressed genes and genes associated with SNPs under selection, we identified 11, 8 and 3 genes overlapping between the low versus high, low versus control and control versus high contrasts, respectively.

To compare between the lists of genes and differentially used transcripts, we compared between the genes that the transcripts came from, and the list of genes associated with SNPs under selection. Between those two lists, we identified 9, 17 and 7 genes overlapping between the low versus high, low versus control and control versus high contrasts, respectively.

To answer the previously posed question on whether the evolution of sociability is due to differential expression or differences in alleles, we see both. We see little overlap between the two gene lists, indicating that the response we see is likely due to both. It is not surprising that we see genes that are under selection that were not observed our list of differentially expressed genes. The RNA sequencing is of adult heads, and we are not capturing developmental timepoints. There may be genes that are differentially expressed between sociability lineages earlier in development that are playing a large role in sociability but are not being captured. However, this information can still be preserved in the population genomics analysis as there may be different allele frequencies between the sociability lineages for developmental genes. We can see evidence of this, as some of the enriched gene ontology terms identified in the population genomics analysis are developmental, such as regulation of axonogenesis, nervous system development, central nervous system development and positive regulation of neuron differentiation.

There are some caveats with the results obtained in this section. As we mentioned in chapter 1, we are capturing a specific aspect of sociability. This aspect of sociability may not translate to other species or even other studies that look at sociability or social behaviour in *Drosophila* as different studies use different experimental protocols that can capture other aspects of the behaviour. However, the response to selection that we see (Figure 1.2) indicates that we are still capturing at least an aspect of sociability, which is supported by the results of this chapter. Another caveat is that compared to the DGE/DTU analyses in chapter 2, we lose lineage specific resolution as we merge our replicates. With this, we are unable to observe potential parallel evolution of lineages within either the low or high sociable treatments. We are also unable to know if potential genes or regions that were identified to be under selection showed similar results in all lineages, or were present in only one, two or three of the lineages, and was a strong enough effect to be observed in the merged samples.

The work in this chapter can easily be expanded upon. As previously mentioned, nucleotide diversity (θ_π or π) may not be the best indicator to identify selection in our experiment, given the fact that we do not see regions of the genome under strong selection from our F_{ST} results. However, investigating regions of low π may still provide information on regions under purifying selection (Cvijović et al., 2018).

There are also supervised learning approaches that can take advantage of both the SNP data and the sociability scores of the lineages as recorded in Scott et al. (2022). Methods such as random forest algorithm can be used with the SNP data and sociability scores of lineages to predict loci that best explain the variation in the trait, as described in Brieuc et al. (2018) and implemented in Brieuc et al. (2015).

Overall, we found no evidence of gene regions under strong selection but did find genes that are associated with SNPs that are most likely under weak selection and show genomic differentiation between selection contrasts. We also developed a pipeline that can be adapted to future studies that wish to perform population genomics analyses with pooled sequencing data, as well as a pipeline to process and SNP call pooled sequencing data. The pipeline is flexible and can be used in combination with other downstream analyses that are appropriate for a given study.

Chapter 4: Conclusions and Final Remarks

As evident throughout this thesis, sociability is a complex phenotype. As previously mentioned in chapter 1, we have limited knowledge of the underlying genetic architecture of the trait, particularly in *Drosophila melanogaster*. In this thesis, I investigated the genetic architecture of sociability in *D. melanogaster* following artificial selection of lineages evolved for low and high sociability. I also developed pipelines to perform differential gene expression, transcript usage and population genomics analyses from pooled sequencing data.

In chapter 2, I investigated the genetic architecture of sociability using DGE and DTU analyses. Here, I developed a method to perform differential gene expression and transcript usage analysis in an uncommon manner, by fitting gene expression and transcript usage with linear mixed-effect models. The main goal in this chapter was to identify genes that showed differential transcript usage or gene expression between the low and high sociable lineages. I found a total of 190 genes with transcripts indicating differential usage and 174 genes that showed differential expression, and with the latter I manually curated to identify genes with phenotypes associated with sociability that can be further tested.

In chapter 3, we investigated what genes are under selection between the sociability lineages, using a population genomics approach. Using downstream tools designed for pooled sequencing data, I developed a pipeline to process and analyze the genomes as well as identify SNPs under selection. The main goal in this chapter was to find genes associated with SNPs under selection between the low and high sociable lineages. While I did not observe strong selection, I still identified 324 genes that we compared back to the results of chapter 2. When looking at the results from both chapter 2 and 3, we see that the variation and change in the sociability phenotype is most likely a result of differential gene expression, differential transcript usage and differences in alleles between selection treatments.

The work in this thesis does not come without specific caveats. The artificial selection experiment to evolve the lineages of high and low sociable flies most likely captures a specific aspect of sociability. Given that experimental protocols and methodologies to assess sociability vary between species and even within studies of the same species, the results here may not perfectly translate comparatively to other studies. When comparing our results to the literature in chapter 2, we did not see a shared list of differentially expressed genes between studies which can be for a number of reasons. Perhaps it is because we are capturing a specific aspect of sociability that is not being captured in other social behaviour studies for a number of methodological reasons. In that case, it would not be a surprise that we do not see similar genes arise in different studies. Or perhaps the evolution of sociability and social behaviour is polygenic in nature and involves different genes in different studies. Regardless, we are supplementing the field with additional knowledge of the underlying genetic architecture of the sociability phenotype within *Drosophila*.

This thesis helps build our knowledge of the underlying genetics of sociability within *Drosophila*. The identified gene lists, as well as the pipeline can be easily integrated in future *Drosophila* sociability studies. The gene lists produced in this thesis can serve as useful candidate genes for further study. Given the noise and lack of evidence for strong selection in the DNA analysis, perhaps testing candidate genes from the list of differentially expressed genes or genes that code for the differentially used transcripts may serve as more powerful candidates. In fact, there are some promising results coming from testing candidate genes identified in the differential gene expression analysis from the Dukas lab. Additionally, a long term artificial selection experiment that goes longer than this study would be intriguing, perhaps with more generations we might observe a stronger phenotypic response and evidence of stronger selection in the analyses.

I also developed pipelines for the DGE, DTU and population genomics analyses. The DGE and DTU pipelines offers a unique method of modelling gene expression and transcript usage, by fitting a linear mixed-effect model. The majority of studies that perform differential gene expression analysis use linear models to model gene expression, and our pipeline offers an alternative method for studies with large enough sample sizes and experimental designs. The population genomics pipeline provides tools to process, analyze and variant call pooled sequencing data, as well as incorporating the new tool grenedalf (Czech et al., 2023) which resolves bugs and issues with other population genomics analyses software. Both pipelines can be adapted in other studies to incorporate more downstream analyses or modifications to parameters to fit their need. Overall, we are adding to the body of work to elucidate the genetic architecture of sociability.

References

Alexa, A., & Rahnenfuhrer, J. (2022). *topGO: Enrichment Analysis for Gene Ontology*. In R Package version 2.48.0. https://bioconductor.org/packages/release/bioc/html/topGO.html

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), 9354. https://doi.org/10.1038/s41598-019-45839-z

Amsalem, E., Grozinger, C. M., Padilla, M., & Hefetz, A. (2015). Chapter Two - The Physiological and Genomic Bases of Bumble Bee Social Behaviour. In A. Zayed & C. F. Kent (Eds.), *Advances in Insect Physiology* (Vol. 48, pp. 37-93). Academic Press. https://doi.org/https://doi.org/10.1016/bs.aiip.2015.01.001

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Martinelli Boneschi, F., D'Alfonso, S., & De Bellis, G. (2016). Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. *Scientific Reports*, *6*(1), 33735. https://doi.org/10.1038/srep33735

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008-2017. https://doi.org/10.1101/gr.133744.111

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* In https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arbeitman, M. N., New, F. N., Fear, J. M., Howard, T. S., Dalton, J. E., & Graze, R. M. (2016). Sex Differences in Drosophila Somatic Gene Expression: Variation and Regulation by doublesex. *G3 Genes|Genomes|Genetics*, *6*(7), 1799-1808. https://doi.org/10.1534/g3.116.027961

Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2014). A sociability gene? Meta-analysis of oxytocin receptor genotype effects in humans. *Psychiatric Genetics*, *24*(2). https://journals.lww.com/psychgenetics/Fulltext/2014/04000/A_sociability_gene__Meta_analysis_of_oxytocin.1.aspx

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165-1188. https://doi.org/10.1214/aos/1013699998

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bond, M. L., Lee, D. E., Farine, D. R., Ozgul, A., & König, B. (2021). Sociability increases survival of adult female giraffes. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1944), 20202770. https://doi.org/10.1098/rspb.2020.2770

Bralten, J., Mota, N. R., Klemann, C. J. H. M., De Witte, W., Laing, E., Collier, D. A., De Kluiver, H., Bauduin, S. E. E. C., Arango, C., Ayuso-Mateos, J. L., Fabbri, C., Kas, M. J., Van Der Wee, N., Penninx, B. W. J. H., Serretti, A., Franke, B., & Poelmans, G. (2021). Genetic underpinnings of sociability in the general population. *Neuropsychopharmacology*, *46*(9), 1627-1634. https://doi.org/10.1038/s41386-021-01044-z

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525-527. https://doi.org/10.1038/nbt.3519

Brieuc, M. S. O., Ono, K., Drinan, D. P., & Naish, K. A. (2015). Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (Oncorhynchus tshawytscha). *Molecular Ecology*, *24*(11), 2729-2746. https://doi.org/https://doi.org/10.1111/mec.13211

Brieuc, M. S. O., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, *18*(4), 755-766. https://doi.org/https://doi.org/10.1111/1755-0998.12773

Broad Institute. (2019). *Picard toolkit*. In https://broadinstitute.github.io/picard/

Brodkin, E. S. (2007). BALB/c mice: Low sociability and other phenotypes that may be relevant to autism. *Behavioural Brain Research*, *176*(1), 53-65. https://doi.org/https://doi.org/10.1016/j.bbr.2006.06.025

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R journal*, *9*, 378--400. https://doi.org/10.32614/RJ-2017-066

Chapman, B. B., Thain, H., Coughlin, J., & Hughes, W. O. H. (2011). Behavioural syndromes at multiple scales in Myrmica ants. *Animal Behaviour*, *82*(2), 391-397. https://doi.org/https://doi.org/10.1016/j.anbehav.2011.05.019

Chen, Y., Lun, A., & Smyth, G. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research*, *5*(1438). https://doi.org/10.12688/f1000research.8987.2

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*(2), 80-92. https://doi.org/10.4161/fly.19695

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ2 Tests. *Biometrics*, *10*(4), 417-451. https://doi.org/10.2307/3001616

Costa, J. T. (2006). *The Other Insect Societies*. Harvard University Press.

Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, *186*(1), 41-43. https://doi.org/10.1534/genetics.110.121012

Cutter, A. D. (2019). *A primer of molecular population genetics*. Oxford University Press, USA.

Cvijović, I., Good, B. H., & Desai, M. M. (2018). The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics*, *209*(4), 1235-1278. https://doi.org/10.1534/genetics.118.301058

Czech, L., Spence, J. P., & Expósito-Alonso, M. (2023). grenedalf: population genetic statistics for the next generation of pool sequencing. *arXiv*. https://doi.org/10.48550/arXiv.2306.11622

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

DeLuca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., & Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, *28*(11), 1530-1532. https://doi.org/10.1093/bioinformatics/bts196

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.*, *29*(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Doyle, R. W., & Talbot, A. J. (1986). Artificial Selection on Growth and Correlated Selection on Competitive Behaviour in Fish. *Canadian Journal of Fisheries and Aquatic Sciences*, *43*(5), 1059-1064. https://doi.org/10.1139/f86-132

Dukas, R. (2020). Natural history of social and sexual behavior in fruit flies. *Scientific Reports*, *10*, 21932.

Dukas, R., Yan, J. L., Scott, A. M., Sivaratnam, S., & Baxter, C. M. (2020). Artificial selection on sexual aggression: Correlated traits and possible trade-offs. *Evolution*, *74*(6), 1112-1123. https://doi.org/10.1111/evo.13993

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.*, *32*(19), 3047-3048. https://doi.org/10.1093/bioinformatics/btw354

Ferreira, C. H., & Moita, M. A. (2020). Behavioral and neuronal underpinnings of safety in numbers in fruit flies. *Nature Communications*, *11*(1), 4182. https://doi.org/10.1038/s41467-020-17856-4

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression, Third Edition*. Sage Publications. https://us.sagepub.com/en-us/nam/an-r-companion-to-applied-regression/book246125

Futschik, A., & Schlötterer, C. (2010). The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, *186*(1), 207-218. https://doi.org/10.1534/genetics.110.114397

Gao, J., Davis, L. K., Hart, A. B., Sanchez-Roige, S., Han, L., Cacioppo, J. T., & Palmer, A. A. (2017). Genome-Wide Association Study of Loneliness Demonstrates a Role for Common Variation. *Neuropsychopharmacology*, *42*(4), 811-821. https://doi.org/10.1038/npp.2016.197

Gramates, L. S., Agapite, J., Attrill, H., Calvi, B. R., Crosby, M. A., dos Santos, G., Goodman, J. L., Goutte-Gattat, D., Jenkins, V. K., Kaufman, T., Larkin, A., Matthews, B. B., Millburn, G., Strelets, V. B., Perrimon, N., Gelbart, S. R., Broll, K., Crosby, L., & dos Santos, G. (2022). FlyBase: a guided tour of highlighted features. *Genetics.*, *220*(4). https://doi.org/10.1093/genetics/iyac035

Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, *16*(10), 620-631. https://doi.org/10.1038/nrn4005

Hämäläinen, A., Kiljunen, M., Koskela, E., Koteja, P., Mappes, T., Rajala, M., & Tiainen, K. (2022). Artificial selection for predatory behaviour results in dietary niche differentiation in an omnivorous mammal. *Proceedings of the Royal Society B:*

Biological Sciences, *289*(1970), 20212510. https://doi.org/10.1098/rspb.2021.2510

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, *7*(1), 1-16. https://doi.org/https://doi.org/10.1016/0022-5193(64)90038-4

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, *7*(1), 17-52. https://doi.org/https://doi.org/10.1016/0022-5193(64)90039-6

Hartl, D., & Clark, A. (2007). *Principles of population genetics fourth edition*. Sinauer and Associates Publishing, Sunderland MA, USA.

Heng, L. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. https://doi.org/10.48550/arXiv.1303.3997

Higashi, M., Yamamura, N., & Abe, T. (2000). Theories on the Sociality of Termites. In T. Abe, D. E. Bignell, & M. Higashi (Eds.), *Termites: Evolution, Sociality, Symbioses, Ecology* (pp. 169-187). Springer Netherlands. https://doi.org/10.1007/978-94-017-3223-9_8

Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T., & Stephenson, D. (2015). Loneliness and Social Isolation as Risk Factors for Mortality:A Meta-Analytic Review. *Perspectives on Psychological Science*, *10*(2), 227-237. https://doi.org/10.1177/1745691614568352

Johnsson, M., Henriksen, R., Fogelholm, J., Höglund, A., Jensen, P., & Wright, D. (2018). Genetics and Genomics of Social Behavior in a Chicken Model. *Genetics*, *209*(1), 209-221. https://doi.org/10.1534/genetics.118.300810

Jones, B. M., Rubin, B. E. R., Dudchenko, O., Kingwell, C. J., Traniello, I. M., Wang, Z. Y., Kapheim, K. M., Wyman, E. S., Adastra, P. A., Liu, W., Parsons, L. R., Jackson, S. R., Goodwin, K., Davidson, S. M., McBride, M. J., Webb, A. E., Omufwoko, K. S., Van Dorp, N., Otárola, M. F., . . . Kocher, S. D. (2023). Convergent and complementary selection shaped gains and losses of eusociality in sweat bees. *Nature Ecology & Evolution*, *7*(4), 557-569. https://doi.org/10.1038/s41559-023-02001-3

Kajokaite, K., Whalen, A., Koster, J., & Perry, S. (2022). Social integration predicts survival in female white-faced capuchin monkeys. *Behavioral Ecology*, *33*(4), 807-815. https://doi.org/10.1093/beheco/arac043

Kapun, M., Barrón, M. G., Staubach, F., Obbard, D. J., Wiberg, R. A. W., Vieira, J., Goubert, C., Rota-Stabelli, O., Kankare, M., Bogaerts-Márquez, M., Haudry, A., Waidele, L., Kozeretska, I., Pasyukova, E. G., Loeschcke, V., Pascual, M., Vieira, C. P., Serga, S., Montchamp-Moreau, C., . . . González, J. (2020). Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Molecular Biology and Evolution*, *37*(9), 2661-2678. https://doi.org/10.1093/molbev/msaa120

Khodursky, S., Svetec, N., Durkin, S. M., & Zhao, L. (2020). The evolution of sex-biased gene expression in the Drosophila brain. *Genome Res*, *30*(6), 874-884. https://doi.org/10.1101/gr.259069.119

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples

(Pool-Seq). *Bioinformatics*, *27*(24), 3435-3436. https://doi.org/10.1093/bioinformatics/btr589

Krause, J., & Ruxton, G. (2002). *Living in Groups*. Oxford University Press.

Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, *69*, 313-332. https://doi.org/https://doi.org/10.1016/j.neubiorev.2016.07.002

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. In R package version 1.8.1-1. https://CRAN.R-project.org/package=emmeans

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473-483. https://doi.org/10.1093/bib/bbq015

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Love, M. I., Soneson, C., & Patro, R. (2018). Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research*, *7*, 952. https://doi.org/10.12688/f1000research.15398.3

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719-748.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*, *40*(10), 4288-4297. https://doi.org/10.1093/nar/gks042

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & Depristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31-46. https://doi.org/10.1038/nrg2626

Nanni, A. V., Martinez, N., Graze, R., Morse, A., Newman, J. R. B., Jain, V., Vlaho, S., Signor, S., Nuzhdin, S. V., Renne, R., & McIntyre, L. M. (2023). Sex-Biased Expression Is Associated With Chromatin State in Drosophila melanogaster and Drosophila simulans. *Molecular Biology and Evolution*, *40*(5). https://doi.org/10.1093/molbev/msad078

Nordell, S. E., & Valone, T. J. (2014). *Animal Behavior: Concepts, Methods, and Applications*. Oxford University Press.

Nowicka, M., & Robinson, M. D. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, *5*, 1356. https://doi.org/10.12688/f1000research.8900.2

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413-435. https://doi.org/10.1007/s13353-011-0057-x

Parisi, M., Nuttall, R., Edwards, P., Minor, J., Naiman, D., Lü, J., Doctolero, M., Vainer, M., Chan, C., Malley, J., Eastman, S., & Oliver, B. (2004). A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogasteradults. *Genome Biology*, *5*(6), R40. https://doi.org/10.1186/gb-2004-5-6-r40

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417-419. https://doi.org/10.1038/nmeth.4197

Pearce, E., Wlodarski, R., Machin, A., & Dunbar, R. I. M. (2017). Variation in the β-endorphin, oxytocin, and dopamine receptor genes is associated with different dimensions of human sociality. *Proceedings of the National Academy of Sciences*, *114*(20), 5300-5305. https://doi.org/doi:10.1073/pnas.1700712114

Plateaux-Quénu, C. (2008). Subsociality in halictine bees. *Insectes Sociaux*, *55*(4), 335-346. https://doi.org/10.1007/s00040-008-1028-z

Porto-Neto, L. R., Lee, S. H., Lee, H. K., & Gondro, C. (2013). Detection of Signatures of Selection Using FST. In C. Gondro, J. van der Werf, & B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction* (pp. 423-436). Humana Press. https://doi.org/10.1007/978-1-62703-447-0_19

Prokopy, R. J., & Roitberg, B. D. (2001). Joining and avoidance behavior in nonsocial insects. *Annual Review of Entomology*, *46*(1), 631-665. https://doi.org/doi:10.1146/annurev.ento.46.1.631

Queller, D. C. (1985). Kinship, reciprocity and synergism in the evolution of social behaviour. *Nature*, *318*(6044), 366-367. https://doi.org/10.1038/318366a0

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2022). *R: A language and environment for statistical computing.* In (Version 4.2.0) R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramos, A., & Gonçalves, D. (2019). Artificial selection for male winners in the Siamese fighting fish Betta splendens correlates with high female aggression. *Frontiers in Zoology*, *16*(1), 34. https://doi.org/10.1186/s12983-019-0333-x

Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D., & Hartl, D. L. (2003). Sex-Dependent Gene Expression and Evolution of the <i>Drosophila</i> Transcriptome. *Science*, *300*(5626), 1742-1745. https://doi.org/doi:10.1126/science.1085881

Reyes, A., Anders, S., & Huber, W. (2013). *Inferring differential exon usage in RNA-Seq data with the DEXSeq package*. https://bioconductor.org/packages/release/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.html

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47-e47. https://doi.org/10.1093/nar/gkv007

Robinson, G. E., & Ben-Shahar, Y. (2002). Social behavior and comparative genomics: new genes or new gene regulation? *Genes Brain Behav*, *1*(4), 197-203. https://doi.org/10.1034/j.1601-183x.2002.10401.x

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Sarin, S., & Dukas, R. (2009). Social learning about egg laying substrates in fruit flies. *Proceedings of the Royal Society of London B-Biological Sciences*, *276*, 4323-4328.

Saris, I. M. J., Aghajani, M., van der Werff, S. J. A., van der Wee, N. J. A., & Penninx, B. W. J. H. (2017). Social functioning in patients with depressive and anxiety disorders. *Acta Psychiatrica Scandinavica*, *136*(4), 352-361. https://doi.org/https://doi.org/10.1111/acps.12774

Schneider, J., Dickinson, M. H., & Levine, J. D. (2012). Social structures depend on innate determinants and chemosensory processing in *Drosophila*. *Proceedings of the National Academy of Sciences*, *109*(Supplement 2), 17174-17179. https://doi.org/10.1073/pnas.1121252109

Scott, A. M., Dworkin, I., & Dukas, R. (2018). Sociability in Fruit Flies: Genetic Variation, Heritability and Plasticity. *Behavior Genetics*, *48*(3), 247-258. https://doi.org/10.1007/s10519-018-9901-7

Scott, A. M., Dworkin, I., & Dukas, R. (2022). Evolution of sociability by artificial selection. *Evolution*, *76*(3), 541-553. https://doi.org/10.1111/evo.14370

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A., & Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research*, *20*(2), 273-280. https://doi.org/10.1101/gr.096388.109

Shpigler, H. Y., Saul, M. C., Corona, F., Block, L., Cash Ahmed, A., Zhao, S. D., & Robinson, G. E. (2017). Deep evolutionary conservation of autism-related genes. *Proceedings of the National Academy of Sciences*, *114*, 9653–9658. https://doi.org/10.1073/pnas.1708127114

Siegel, P. B. (1972). Genetic analysis of male mating behaviour in chickens (Gallus domesticus). I. Artificial selection. *Animal Behaviour*, *20*(3), 564-570. https://doi.org/https://doi.org/10.1016/S0003-3472(72)80021-6

Silk, J. B., Alberts, S. C., & Altmann, J. (2003). Social bonds of female baboons enhance infant survival. *Science*, *302*(5648), 1231-1234. http://www.sciencemag.org/cgi/content/abstract/302/5648/1231

Silk, J. B., Beehner, J. C., Bergman, T. J., Crockford, C., Engh, A. L., Moscovice, L. R., Wittig, R. M., Seyfarth, R. M., & Cheney, D. L. (2010). Strong and consistent social bonds enhance the longevity of female baboons. *Current Biology*, *20*(15), 1359-1361. https://doi.org/http://dx.doi.org/10.1016/j.cub.2010.05.067

Smit, A., Hubley, R., & Green, P. (2013-2015). Repeatmasker Open-4.0. http://www.repeatmasker.org

Snyder-Mackler, N., Burger, J. R., Gaydosh, L., Belsky, D. W., Noppert, G. A., Campos, F. A., Bartolomucci, A., Yang, Y. C., Aiello, A. E., O'Rand, A., Harris, K. M., Shively, C. A., Alberts, S. C., & Tung, J. (2020). Social determinants of health and survival in humans and other animals. *Science*, *368*(6493), eaax9553. https://doi.org/10.1126/science.aax9553

Soneson, C., Love, M. I., & Robinson, M. D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. https://doi.org/10.12688/f1000research.7563.2

Spitzer, K., Pelizzola, M., & Futschik, A. (2020). Modifying the Chi-square and the CMH test for population genetic inference: Adapting to overdispersion. *The Annals of Applied Statistics*, *14*(1), 202-220, 219. https://doi.org/10.1214/19-AOAS1301

Srivastava, A., Sarkar, H., Gupta, N., & Patro, R. (2016). RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, *32*(12), i192-i200. https://doi.org/10.1093/bioinformatics/btw277

Tekath, T., & Dugas, M. (2021). Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle. *Bioinformatics*, *37*(21), 3781-3787. https://doi.org/10.1093/bioinformatics/btab629

Tiğli Filizer, A., Cerit, C., Tüzün, B., & Aker, A. T. (2016). Social Aspect of Functioning Deteriorates More Than Individual Aspect in Patients with Remitted Bipolar Disorder. *Noro Psikiyatr Ars*, *53*(2), 158-162. https://doi.org/10.5152/npa.2015.10106

Vega-Trejo, R., Boussard, A., Wallander, L., Estival, E., Buechel, S. D., Kotrschal, A., & Kolm, N. (2020). Artificial selection for schooling behaviour and its effects on associative learning abilities. *Journal of Experimental Biology*, *223*(23), jeb235093. https://doi.org/10.1242/jeb.235093

Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics.*, *28*(16), 2184-2185. https://doi.org/10.1093/bioinformatics/bts356

Wang, Z. Y., Mckenzie-Smith, G. C., Liu, W., Cho, H. J., Pereira, T., Dhanerawala, Z., Shaevitz, J. W., & Kocher, S. D. (2022). Isolation disrupts social interactions and destabilizes brain development in bumblebees. *Current Biology*, *32*(12), 2754-2764.e2755. https://doi.org/10.1016/j.cub.2022.04.066

Ward, A., & Webster, M. (2016). Sociality: The Behaviour of Group Living Animals. In. Basel, Switzerland: Springer.

Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klaR Analyzing German Business Cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data Analysis and Decision Support* (pp. 335-343). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-28397-8_36

Wiberg, R. A. W., Gaggiotti, O. E., Morrissey, M. B., & Ritchie, M. G. (2017). Identifying consistent allele frequency differences in studies of stratified populations. *Methods in Ecology and Evolution*, *8*(12), 1899-1909. https://doi.org/https://doi.org/10.1111/2041-210X.12810

Wilson, E. O. (1971). *The Insect Societies*. Belknap Press of Harvard University Press.

Wilson, E. O., & Hölldobler, B. (2005). Eusociality: Origin and consequences. *Proceedings of the National Academy of Sciences*, *102*(38), 13367-13371. https://doi.org/10.1073/pnas.0505858102

Woodard, S. H., Fischman, B. J., Venkat, A., Hudson, M. E., Varala, K., Cameron, S. A., Clark, A. G., & Robinson, G. E. (2011). Genes involved in convergent evolution of eusociality in bees. *Proceedings of the National Academy of Sciences*, *108*(18), 7472-7477. https://doi.org/10.1073/pnas.1103457108

Appendix

All links to supplemental figures, gene lists and code can be found at:

https://github.com/ArteenTorabiMarashi/ThesisLinks

Supplemental figures on the GitHub include plots of all differentially expressed genes and differentially used transcripts in the low versus high sociability contrast. Full gene lists for both differentially expressed genes and differentially used transcripts for all three selection contrasts can also be found in the above GitHub repository. Full GO tables of all terms can also be found in the GitHub repository.

Table A1. Table detailing all software used during analyses with corresponding version numbers and references for the DGE and DTU analysis. The Notable parameters column provides certain parameters that were different than the default settings with either a code chunk, or the entire code that was used with the software. Default parameters denotes that no additional parameters were used besides input/output flags and those detailed in their respective documentation's base use.

Software / Step	Version	Notable parameters	Reference
FASTQC	0.11.9	Default parameters	(Andrews, 2010)
MultiQC	1.12	Default parameters	(Ewels et al., 2016)

trimmomatic	0.36	ILLUMINACLIP:/adapter_dir/:2:30:10 \ LEADING:3 TRAILING:3 MAXINFO:20:0.2 MINLEN:36	(Bolger et al., 2014)
Salmon Index	1.4.0	Default parameters Transcriptome version r6.38 Using decoys.txt in order to account for decoys	(Patro et al., 2017)
Salmon Mapping	1.4.0	salmon quant -i /index_dir/ -l A \ -1 /trimmed/dir/R1_PE.fastq.gz \ -2 /trimmed/dir/R2_PE.fastq.gz \ -p 6 --validateMappings --rangeFactorizationBins 4 \ --seqBias --gcBias \ -o /counts_dir/sample_quant	(Patro et al., 2017)
STAR Index	2.7.9a	Default parameters Genome version r6.38	(Dobin et al., 2013)
STAR Mapping	2.7.9a	STAR --runThreadN 16 \ --quantMode TranscriptomeSAM	(Dobin et al., 2013)

		GeneCounts \ --genomeDir /genome/dir \ --readFilesIn /trim/sample_name/R1_PE.fastq.gz \ /trim/sample_name/ _R2_PE.fastq.gz \ --readFilesCommand zcat \ --outFileNamePrefix /out_dir/sample_name \ --outSAMtype BAM SortedByCoordinate	

edgeR:: FilterByExpr()	v3.38.4	d0 <- DGEList(raw.counts)
		keep2 <- filterByExpr(d0, design)
		d0 <- d0[keep2,]
		# with keep2 acting as indices of counts to keep
			(Robinson et al., 2010)

DRIMseq:: dmFilter()	1.24.0	dex_filtered_subset <- dmFilter(dex_pre_filter,
		min_samps_feature_prop = 20,
		min_feature_prop = 0.05,
			(Nowicka & Robinson, 2016)

		min_samps_gene_expr = 28,
		min_gene_expr = 10)	

Table A2. Biological GO terms that were found to be significantly enriched in the differential gene expression analysis as determined by topGO. We extracted all the genes and p values from the anova for the term of selection. We chose to cut off the table for only GO terms with a p value < 0.03, the full list can be found in the GitHub repository linked in the appendix.

GO.ID	Term	Annot.	Sig.	Exp.	p.val	p.adj
GO:0002181	cytoplasmic translation	123	74	37.11	4.6e-16	1.90578e-12
GO:0040003	chitin-based cuticle development	189	86	57.03	4.2e-06	0.0087003
GO:0003333	amino acid transmembrane transport	41	23	12.37	0.00019	0.26239
GO:0006171	cAMP biosynthetic process	5	5	1.51	0.00249	1
GO:0018345	protein palmitoylation	16	8	4.83	0.00250	1
GO:0009065	glutamine family amino acid catabolic process	15	10	4.53	0.00380	1

GO:0090254	cell elongation involved in imaginal disc-derived wing morphogensis	11	8	3.32	0.00444	1
GO:0006486	protein glycosylation	87	29	26.25	0.00459	1
GO:0001737	establishment of imaginal disc-derived wing hair orientation	25	14	7.54	0.00628	1
GO:0065003	protein-containing complex assembly	444	145	133.97	0.00744	1
GO:0007450	dorsal/ventral pattern formation imaginal disc	54	22	16.29	0.00824	1
GO:0006030	chitin metabolic process	106	46	31.98	0.00925	1
GO:0051028	mRNA transport	43	16	12.97	0.01094	1
GO:0060232	delamination	11	7	3.32	0.01120	1

GO:0051036	regulation of endosome size	6	5	1.81	0.01121	1
GO:0006627	protein processing involved in protein targeting to mitochondrion	6	5	1.81	0.01121	1
GO:0050714	positive regulation of protein secretion	21	8	6.34	0.01124	1
GO:0045571	negative regulation of imaginal disc growth	17	10	5.13	0.01318	1
GO:0035317	imaginal disc-derived wing hair organization	47	26	14.18	0.01434	1
GO:0072659	protein localization to plasma membrane	46	21	13.88	0.01454	1
GO:0018401	peptidyl-proline hydroxylation to	13	8	3.92	0.01882	1

	4-hydroxy-L-proline					
GO:0035002	liquid clearance open tracheal system	18	10	5.43	0.02174	1
GO:0001666	response to hypoxia	62	24	18.71	0.02497	1
GO:0007349	cellularization	98	27	29.57	0.02561	1
GO:0009952	anterior/posterior pattern specification	168	58	50.69	0.02578	1
GO:0042659	regulation of cell fate specification	27	14	8.15	0.02585	1
GO:0043984	histone H4-K16 acetylation	9	6	2.72	0.02597	1
GO:0006414	translational elongation	30	15	9.05	0.02729	1
GO:0034982	mitochondrial protein processing	9	8	2.72	0.02736	1
GO:0014902	myotube differentiation	55	19	16.6	0.02744	1

GO:1904747	positive regulation of apoptotic process involved in development	8	4	2.41	0.02746	1
GO:0050000	chromosome localization	29	9	8.75	0.02750	1
GO:0000398	mRNA splicing via spliceosome	244	81	73.62	0.02927	1
GO:0001706	endoderm formation	7	5	2.11	0.02947	1
GO:0006465	signal peptide processing	7	5	2.11	0.02947	1
GO:0006528	asparagine metabolic process	7	5	2.11	0.02947	1
GO:0090251	protein localization involved in establishment of planar polarity	7	5	2.11	0.02947	1
GO:0050884	neuromuscular process	7	5	2.11	0.02947	1

	controlling posture					
GO:0051124	synaptic assembly at neuromuscular junction	136	46	41.04	0.030	1

Table A3. Biological GO terms that were found to be significantly enriched in the differential transcript usage analysis as determined by topGO. We extracted all the genes and p values from the anova for the term of selection. We chose to cut off the table for only GO terms with a p value < 0.03, the full list can be found in the GitHub repository linked in the appendix.

GO.ID	**Term**	**Annot.**	**Sig.**	**Exp.**	**p.val**	**p.adj**
GO:0006413	translational initiation	26	19	9.65	0.0023	1
GO:0001736	establishment of planar polarity	63	31	23.39	0.0026	1
GO:0072499	photoreceptor cell axon guidance	21	14	7.8	0.0056	1
GO:0007297	ovarian follicle cell migration	75	36	27.84	0.0069	1
GO:0045132	meiotic chromosome segregation	27	14	10.02	0.0070	1

GO:0046660	female sex differentiation	10	7	3.71	0.0070	1
GO:0016226	iron-sulfur cluster assembly	5	5	1.86	0.0070	1
GO:0001731	formation of translation preinitiation complex	5	5	1.86	0.0070	1
GO:0033500	carbohydrate homeostasis	26	12	9.65	0.0070	1
GO:0016325	oocyte microtubule cytoskeleton organization	20	11	7.42	0.0073	1
GO:0046843	dorsal appendage formation	28	17	10.39	0.0093	1
GO:0035317	imaginal disc-derived wing hair organization	32	19	11.88	0.0102	1
GO:0048168	regulation of neuronal synaptic plasticity	7	6	2.6	0.0124	1
GO:0006605	protein targeting	34	15	12.62	0.0125	1
GO:0006672	ceramide metabolic process	9	7	3.34	0.0159	1
GO:0050803	regulation of synapse structure or activity	117	46	43.43	0.0160	1

GO:0035147	branch fusion open tracheal system	11	8	4.08	0.0179	1
GO:0006633	fatty acid biosynthetic process	21	13	7.8	0.0181	1
GO:0045451	pole plasm oskar mRNA localization	26	17	9.65	0.0187	1
GO:0007392	initiation of dorsal closure	10	7	3.71	0.0189	1
GO:0032007	negative regulation of TOR signaling	11	6	4.08	0.0190	1
GO:0051453	regulation of intracellular pH	12	6	4.45	0.0190	1
GO:0007298	border follicle cell migration	70	31	25.98	0.0262	1
GO:0045498	sex comb development	6	5	2.23	0.0291	1
GO:0060232	delamination	6	5	2.23	0.0291	1
GO:1902410	mitotic cytokinetic process	6	5	2.23	0.0291	1
GO:0006941	striated muscle contraction	6	5	2.23	0.0291	1
GO:0032233	positive regulation of actin filament bundle assembly	6	5	2.23	0.0291	1

GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	6	5	2.23	0.0291	1
GO:0007362	terminal region determination	6	5	2.23	0.0291	1
GO:0046716	muscle cell cellular homeostasis	22	13	8.17	0.0295	1

Table A4. Table detailing all software used during analyses with corresponding version numbers and references for the population genomics analysis. The Notable parameters column provides certain parameters that were different than the default settings with either a code chunk, or the entire code that was used with the software. Default parameters denotes that no additional parameters were used besides input/output flags and those detailed in their respective documentation's base use.

Software / Step	Version	Notable Parameters	Reference
FastQC	0.11.9	Default parameters	(Andrews, 2010)
MultiQC	1.12	Default parameters	(Ewels et al., 2016)
trimmomatic	0.36	ILLUMINACLIP:/adapter_dir/:2:30:10 \	(Bolger et al., 2014)

		LEADING:3 TRAILING:3 MAXINFO:20:0.2 MINLEN:36	
bwa indexing	0.7.17	Default parameters	(Li & Durbin, 2009)
bwa mapping	0.7.17	Default parameters	(Heng, 2013)
samtools sam to bam	1.15	Default parameters	(Danecek et al., 2021)
samtools quality filtering	1.15	samtools view -b -q 30 -@ 8	(Danecek et al., 2021)
samtools sort by name	1.15	Default parameters	(Danecek et al., 2021)
samtools add quality tags	1.15	samtools fixmate -m -u -@ 12	(Danecek et al., 2021)
samtools sort by coordinate	1.15	Default parameters	(Danecek et al., 2021)
samtools deduplicate	1.15	samtools markdup -l 150 -r -s -f stats.txt -d 2500\ -@ 12	(Danecek et al., 2021)
picard add read groups	2.26.3	Default parameters of AddOrReplaceReadGroups function	(Broad Institute, 2019)

GATK mark indels	3.8	Default parameters	(McKenna et al., 2010)
GATK realign indels	3.8	Default parameters	(McKenna et al., 2010)
samtools merge	1.15	Default parameters	(Danecek et al., 2021)
samtools create mpileup	1.15	samtools mpileup -Q 20 -q 20 -d 1200	(Danecek et al., 2021)
grenedalf mpileup to sync	0.2.0	Default parameters	(Czech et al., 2023)
poolSNP	1.0	max-cov=0.98 \ min-cov=25 \ min-count=10 \ min-freq=0.01 \ miss-frac=0.2	(Kapun et al., 2020)
Identify indels	1.0	--minimum-count 20 \ --mask 5	(Kapun et al., 2020)
RepeatMasker ID repeats	4.1.1	Default parameters	(Smit et al., 2013-2015)
Filter VCF for indel and	1.0	Default parameters	(Kapun et al., 2020)

repeat positions			
bedtools blacklist	2.30.0	Default parameters	(Quinlan & Hall, 2010)
Filter sync for VCF positions	Personal script	Default parameters	Personally written
Grenedalf F_{ST}	0.2.0	--window-type sliding \ --window-sliding-width 5000 \ --method unbiased-nei \ --pool-sizes 96	(Czech et al., 2023)

Table A5. Biological GO terms that were found to be significantly enriched of genes identified from SNPs between high F_{ST} and significant CMH, in the low versus high contrast. All genes were extracted following SnpEff variant prediction. We chose to cut off the table for only GO terms with a p value < 0.0222, the full list can be found in the GitHub repository linked in the appendix.

GO.ID	Term	Annot.	Sig.	Exp.	p.val.	p.adj
GO:2000289	regulation of photoreceptor cell axon guidance	5	3	0.12	0.00012	0.27365
GO:0007528	neuromuscular junction development	176	12	4.12	0.00013	0.27365
GO:0045892	negative regulation of transcription DNA-templated transcription	291	17	6.81	0.00039	0.49584
GO:0007615	anesthesia-resistant memory	17	4	0.4	0.00055	0.49584
GO:0046339	diacylglycerol metabolic process	8	3	0.19	0.00065	0.49584
GO:0022008	neurogenesis	894	46	20.93	0.00072	0.49584

GO:0008016	regulation of heart contraction	29	6	0.68	0.00093	0.49584
GO:0051017	actin filament bundle assembly	34	7	0.8	0.00102	0.49584
GO:0006812	cation transport	286	12	6.69	0.00106	0.49584
GO:0034332	adherens junction organization	39	5	0.91	0.00179	0.75359
GO:0007632	visual behavior	24	4	0.56	0.00215	0.82286
GO:0030708	germarium-derived female germ-line cyst encapsulation	12	3	0.28	0.00238	0.83498
GO:0007043	cell-cell junction assembly	63	6	1.47	0.00311	1
GO:0007476	imaginal disc-derived wing morphogenesis	273	23	6.39	0.00341	1
GO:0007190	activation of adenylate cyclase activity	5	2	0.12	0.00521	1
GO:0097688	glutamate receptor clustering	5	2	0.12	0.00521	1

GO:0042066	perineurial glial growth	5	2	0.12	0.00521	1
GO:0007212	dopamine receptor signaling pathway	5	2	0.12	0.00521	1
GO:0007216	G protein-coupled glutamate receptor signaling pathway	5	2	0.12	0.00521	1
GO:0045494	photoreceptor cell maintenance	16	3	0.37	0.00566	1
GO:0007291	sperm individualization	52	5	1.22	0.00717	1
GO:0019344	cysteine biosynthetic process	6	2	0.14	0.00769	1
GO:0006751	glutathione catabolic process	6	2	0.14	0.00769	1
GO:0031532	actin cytoskeleton reorganization	34	4	0.8	0.00783	1
GO:0007479	leg disc proximal/distal pattern formation	18	3	0.42	0.00797	1

GO:0008355	olfactory learning	45	5	1.05	0.00859	1
GO:0045214	sarcomere organization	35	4	0.82	0.00868	1
GO:0007601	visual perception	19	3	0.44	0.00931	1
GO:0035002	liquid clearance open tracheal system	19	3	0.44	0.00931	1
GO:0007630	jump response	7	2	0.16	0.01060	1
GO:0045823	positive regulation of heart contraction	7	2	0.16	0.01060	1
GO:0007475	apposition of dorsal and ventral imaginal disc-derived wing surfaces	20	3	0.47	0.01076	1
GO:0042461	photoreceptor cell development	111	6	2.6	0.01376	1
GO:0046661	male sex differentiation	31	3	0.73	0.01387	1
GO:0007415	defasciculation of motor neuron axon	8	2	0.19	0.01392	1

GO:0097320	plasma membrane tubulation	8	2	0.19	0.01392	1
GO:0071880	adenylate cyclase-activating adrenergic receptor signaling pathway	8	2	0.19	0.01392	1
GO:0090278	negative regulation of peptide hormone secretion	8	2	0.19	0.01392	1
GO:0016332	establishment or maintenance of polarity of embryonic epithelium	8	2	0.19	0.01392	1
GO:0035317	imaginal disc-derived wing hair organization	48	4	1.12	0.01399	1
GO:0007435	salivary gland morphogenesis	98	6	2.29	0.01673	1

GO:0008078	mesodermal cell migration	20	3	0.47	0.01753	1
GO:0032958	inositol phosphate biosynthetic process	9	2	0.21	0.01763	1
GO:0010496	intercellular transport	9	2	0.21	0.01763	1
GO:0016204	determination of muscle attachment site	9	2	0.21	0.01763	1
GO:0048047	mating behavior sex discrimination	9	2	0.21	0.01763	1
GO:0007155	cell adhesion	211	10	4.94	0.01861	1
GO:0061320	pericardial nephrocyte differentiation	10	2	0.23	0.02170	1
GO:0043153	entrainment of circadian clock by photoperiod	10	2	0.23	0.02170	1
GO:0045186	zonula adherens assembly	10	2	0.23	0.02170	1

GO:0045937	positive regulation of phosphate metabolic process	72	3	1.69	0.02177	1
GO:0048736	appendage development	340	26	7.96	0.02180	1
GO:0044719	regulation of imaginal disc-derived wing size	26	3	0.61	0.02215	1

Table A6. Biological GO terms that were found to be significantly enriched of genes identified from SNPs between high F_{ST} and significant CMH, in the low versus control contrast. All genes were extracted following SnpEff variant prediction. We chose to cut off the table for only GO terms with a p value < 0. 0124, the full list can be found in the GitHub repository linked in the appendix.

GO.ID	Term	Annot.	Sig.	Exp.	p.val	p.adj
GO:0090630	activation of GTPase activity	28	6	0.49	7.2e-06	0.03031
GO:0043113	receptor clustering	12	4	0.21	1.0e-04	0.19647
GO:0007630	jump response	7	3	0.12	0.00017	0.19647
GO:0042059	negative regulation of epidermal growth	33	5	0.58	0.00025	0.19647

	factor receptor signaling pathway					
GO:0031623	receptor internalization	8	3	0.14	0.00028	0.19647
GO:0097120	receptor localization to synapse	8	3	0.14	0.00028	0.19647
GO:0007520	myoblast fusion	53	6	0.93	0.00085	0.46836
GO:0097090	presynaptic membrane organization	11	3	0.19	0.00089	0.46836
GO:0008016	regulation of heart contraction	29	4	0.51	0.00164	0.65144
GO:0045433	male courtship behavior veined wing generated song production	14	3	0.24	0.00166	0.65144
GO:0016203	muscle attachment	41	5	0.72	0.00219	0.65144
GO:0045434	negative regulation of female	16	3	0.28	0.00249	0.65144

	receptivity, post-mating					
GO:0007608	sensory perception of smell	128	8	2.24	0.00277	0.65144
GO:0007391	dorsal closure	108	7	1.89	0.00283	0.65144
GO:2000289	regulation of photoreceptor cell axon guidance	5	2	0.09	0.00294	0.65144
GO:0051496	positive regulation of stress fiber assembly	5	2	0.09	0.00294	0.65144
GO:0007205	protein kinase C-activating G protein-coupled receptor signaling pathway	5	2	0.09	0.00294	0.65144
GO:0051928	positive regulation of calcium ion transport	5	2	0.09	0.00294	0.65144

GO:0051899	membrane depolarization	5	2	0.09	0.00294	0.65144
GO:0045214	sarcomere organization	35	4	0.61	0.00311	0.65466
GO:0007271	synaptic transmission cholinergic	18	3	0.31	0.00354	0.69973
GO:0016200	synaptic target attraction	6	2	0.1	0.00436	0.69973
GO:0046834	lipid phosphorylation	6	2	0.1	0.00436	0.69973
GO:1900242	regulation of synaptic vesicle endocytosis	6	2	0.1	0.00436	0.69973
GO:0007298	border follicle cell migration	124	8	2.17	0.00446	0.69973
GO:0006807	nitrogen compound metabolic process	4660	81	81.49	0.00453	0.69973
GO:0051017	actin filament bundle assembly	34	6	0.59	0.00463	0.69973

GO:0035556	intracellular signal transduction	648	21	11.33	0.00470	0.69973
GO:0000132	establishment of mitotic spindle orientation	20	3	0.35	0.00482	0.69973
GO:0007266	Rho protein signal transduction	53	5	0.93	0.00544	0.74666
GO:0001738	morphogenesis of a polarized epithelium	118	5	2.06	0.00596	0.74666
GO:0001941	postsynaptic membrane organization	14	3	0.24	0.00598	0.74666
GO:0048172	regulation of short-term neuronal synaptic plasticity	7	2	0.12	0.00603	0.74666
GO:0006359	regulation of transcription by RNA polymerase III	7	2	0.12	0.00603	0.74666

GO:0008586	imaginal disc-derived wing vein morphogenesis	44	4	0.77	0.00713	0.85568
GO:0008355	olfactory learning	45	4	0.79	0.00772	0.85568
GO:0046339	diacylglycerol metabolic process	8	2	0.14	0.00795	0.85568
GO:0016332	establishment or maintenance of polarity of embryonic epithelium	8	2	0.14	0.00795	0.85568
GO:0007632	visual behavior	24	3	0.42	0.00813	0.85568
GO:0070983	dendrite guidance	24	3	0.42	0.00813	0.85568
GO:0046330	positive regulation of JNK cascade	25	3	0.44	0.00912	0.91417
GO:2000331	regulation of terminal button organization	25	3	0.44	0.00912	0.91417
GO:0008045	motor neuron axon guidance	75	5	1.31	0.00997	0.96734
GO:0007602	phototransduction	57	4	1	0.01011	0.96734

GO:0030866	cortical actin cytoskeleton organization	61	4	1.07	0.01124	1
GO:0007480	imaginal disc-derived leg morphogenesis	79	5	1.38	0.01232	1

Table A7. Biological GO terms that were found to be significantly enriched of genes identified from SNPs between high F_{ST} and significant CMH, in the control versus high contrast. All genes were extracted following SnpEff variant prediction.

GO.ID	**Term**	**Annot.**	**Sig.**	**Exp.**	**p.val.**	**p.adj**
GO:0061541	rhabdomere morphogenesis	5	2	0.06	0.0014	1
GO:0051017	actin filament bundle assembly	34	4	0.41	0.0016	1
GO:0007475	apposition of dorsal and ventral imaginal disc-derived wing surfaces	20	3	0.24	0.0017	1
GO:0016199	axon midline choice point recognition	25	3	0.3	0.0032	1

GO:0050808	synapse organization	293	8	3.52	0.0036	1
GO:0016332	establishment or maintenance of polarity of embryonic epithelium	8	2	0.1	0.0038	1
GO:0007157	heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	27	3	0.32	0.0040	1
GO:0018401	peptidyl-proline hydroxylation to 4-hydroxy-L-proline	27	3	0.32	0.0040	1
GO:0043087	regulation of GTPase activity	89	7	1.07	0.0042	1
GO:0090630	activation of GTPase activity	28	3	0.34	0.0045	1
GO:0045186	zonula adherens assembly	10	2	0.12	0.0061	1

GO:0045214	sarcomere organization	35	3	0.42	0.0084	1
GO:0030036	actin cytoskeleton organization	281	16	3.38	0.0097	1
GO:0035023	regulation of Rho protein signal transduction	32	3	0.38	0.0152	1
GO:0045886	negative regulation of synaptic assembly at neuromuscular junction	44	3	0.53	0.0157	1
GO:0045176	apical protein localization	17	2	0.2	0.0173	1
GO:0099536	synaptic signaling	295	11	3.55	0.0227	1
GO:0008078	mesodermal cell migration	20	2	0.24	0.0237	1
GO:0046390	ribose phosphate	86	2	1.03	0.0239	1

	biosynthetic process					
GO:0048489	synaptic vesicle transport	21	2	0.25	0.0259	1
GO:0051491	positive regulation of filopodium assembly	23	2	0.28	0.0308	1
GO:0006816	calcium ion transport	66	3	0.79	0.0331	1
GO:0098813	nuclear chromosome segregation	183	6	2.2	0.0349	1
GO:0060078	regulation of postsynaptic membrane potential	9	2	0.11	0.0354	1
GO:0050802	circadian sleep/wake cycle sleep	38	2	0.46	0.0355	1
GO:0030866	cortical actin cytoskeleton organization	61	3	0.73	0.0369	1

GO:0007605	sensory perception of sound	63	3	0.76	0.0400	1
GO:0007455	eye-antennal disc morphogenesis	42	3	0.5	0.0437	1
GO:0050770	regulation of axonogenesis	46	3	0.55	0.0438	1
GO:0008064	regulation of actin polymerization or depolymerization	45	2	0.54	0.0471	1
GO:0035149	lumen formation open tracheal system	29	2	0.35	0.0471	1
GO:0007614	short-term memory	29	2	0.35	0.0471	1
GO:0010883	regulation of lipid storage	41	3	0.49	0.0495	1