
# Omics-based Predictive and Causative Modeling of Neurobehavioral Traits

By

John A. Williams

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Centre for Computational Biology
Institute of Cancer and Genomic Sciences
University of Birmingham
September 2020

## ABSTRACT

Neurobehavioral disorders can be phenotypically and genetically complex, and often diagnosed through observational study or subjective assessment alone. Certain neurobehavioral phenotypes, such as those caused by circadian rhythm related behavior, are biochemically well characterized, others, though, do not have yet a well understood genetic aetiology. Furthermore, circadian biology and psychological disorders are often intertwined. To advance our understanding of neurobehavioral trait/gene relationships, I first built a machine learning model that encompasses mouse transcriptomics to predict genes involved in circadian rhythms. Next, I used genome wide association studies to model the causal influence of genetic exposure in humans to an evening chronotype on several mental health and social support traits, from depression to group religious participation. To more accurately model how neurobehaviors relate to one another, I mined psychological assessment instruments to build a species-agnostic psychological neurobehavior ontology encompassing autism and schizophrenia phenotypes. I, then, tested the utility of this ontology in clustering children on the autism spectrum based on phenotypic profiles. Lastly, I annotated genes to behaviors identified among subgroups through genome wide association studies applied to phenotype profiles. This allowed for the gene prioritization of circadian related experimentation results and the discovery of new, potentially, casual relationships between chronotype and neurobehavioral traits. Finally, the semantic representation of schizophrenia endophenotypes in a consistent, ontology framework catered its application for the identification of novel gene-trait associations in humans. These contributions provide new knowledge to the scientific community of the potential novel circadian functions for known genes, of the likely causal influence of chronotype on social and mental health, provide novel robust ways of modeling the complex phenotype of autism and schizophrenia patients, while annotating neurologically active genes to new behavioral traits for the first time.

## DEDICATION

This thesis is dedicated to my parents: Jean, John, and Janice. Thank you for all the love and support.

## ACKNOWLEDGMENTS

First, I would like to thank my supervisors: Dr. Michelle Simon, Dr. Ann-Marie Mallon, and Prof. Georgios Gkoutos. Without their guidance, support, and encouragement this PhD would not be possible. Michy, Annie, and George - y'all have been wonderful! I owe my start in science to Annie and Michy, who accepted my application to do my MSc research at MRC Harwell, kept me on as an intern, and helped me meet George. This PhD was a collaborative effort between one student, two institutions, and three supervisors. Not only have I learned how to be a better scientist from you all, but also thoroughly enjoyed my time with you.

This PhD was funded and co-supervised by the MRC Harwell Institute's doctoral training centre, with financial aid from a grant by the National Human Genome Research Institute of the National Institutes of Health, under award number UM1HG006370. This grant would not be possible without the hard work of all the researchers and support community at MRC Harwell and throughout the International Mouse Phenotyping Consortium.

Drs. Gareth Banks and Patrick Nolan, and no doubt other members of the Nolan group, were responsible for the experimental work described in Chapter 2. I could not have completed this project without their support, and I look forward to ongoing collaborations to keep mice in neurobehavioral genetics. Pat and Gareth, thank you.

My appreciation goes to the entire Gkoutos group, especially my fellow PhD students Luke and Victor who worked closely with George and I in the early days of the group. Luke, your latex troubleshooting skills are legendary. I'm also thankful for the great support

and feedback from the members of the omics group, including Dom, Laura, and Animesh. Animesh and Andreas are wonderful senior leaders in the group and my time in Birmingham has been greatly enriched learning from them. Dom was especially helpful in create our group's first GWAS pipelines during his MSc and later PhD.

I am grateful for all those in my time in the Biocomputing Group at MRC Harwell Institute, especially Sid, Simon, Hugh, Henrik, Neil, Luis, George, Daniel, Peter, Ramon, Anna, Ewan, Habib, Helen, Piia, Ruairidh, and all the former residents of Room 16/4.

Work modeling the ANBO ontology in Chapter 4 was collaborative, and I would like to thank Drs. Martínez-Santiago and García-Viedma for allowing me to collaborate with them on the ANBO project. They were responsible for the genesis and overall experimental design, my role was working with the ANBO and additional ontologies as described.

I are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). I appreciate obtaining access to phenotypic and genetic data on SFARI Base.

Lastly, thank you to my friends and family both here in the UK and back home in the US. Working through the COVID-19 pandemic while living abroad has been an interesting and at times trying experience, but having calls with y'all has always brightened my day. Thank you Daniel for all of your support these last two years!

## Published Work

During the course of this PhD thesis, I was involved in the following publications:

- Brown, L.A., **Williams, J.**, Taylor, L., Thomson, R.J., Nolan, P.M., Foster, R.G., Peirson, S.N., 2017. Meta-analysis of transcriptomic datasets identifies genes enriched in the mammalian circadian pacemaker. Nucleic Acids Res 45, 9860-9873. https://doi.org/10.1093/nar/gkx714

- **Williams, John A.**, George Powell, Ann-Marie Mallon, and Michelle M. Simon. 'Genomic Mutation Identification in Mice Using Illumina Sequencing and Linux-Based Computational Methods'. Current Protocols in Mouse Biology 9, no. 3 (2019): e64. https://doi.org/10.1002/cpmo.64.

- Bhattacharjee, D., S. Vracar, R. A. Round, P. G. Nightingale, **J. A. Williams**, G. V. Gkoutos, I. M. Stratton, et al. 'Utility of HbA1c Assessment in People with Diabetes Awaiting Liver Transplantation'. Diabetic Medicine 36, no. 11 (2019): 1444–52. https://doi.org/10.1111/dme.13870.

- Bravo-Merodio L, **Williams JA**, Gkoutos GV, Acharjee A. 2019. -Omics biomarker identification pipeline for translational medicine. J Transl Med. 17(1):155. doi:10.1186/s12967-019-1912-5

- Martínez-Santiago, Fernando, M. Rosario García-Viedma, **John A. Williams**, Luke

T. Slater, and Georgios V. Gkoutos. 'Aging Neuro-Behavior Ontology'. Applied Ontology 15, no. 2 (1 January 2020): 219–39. https://doi.org/10.3233/AO-200229.

- Ghosh, Sandip, Susan E. Manley, Peter G. Nightingale, **John A. Williams**, Radhika Susarla, Irene Alonso-Perez, Irene M. Stratton, et al. Prevalence of Admission Plasma Glucose in diabetes or at Risk Ranges in Hospital Emergencies with No Prior Diagnosis of Diabetes by Gender, Age and Ethnicity. Endocrinology, Diabetes & Metabolism, e00140. https://doi.org/10.1002/edm2.140

- Jahangiri L, Tsaprouni L, Trigg RM, **Williams JA**, Gkoutos GV, Turner SD, et al. 2020. Core regulatory circuitries in defining cancer cell identity across the malignant spectrum. Open Biology. 10(7):200121.

The following publications are in submission or in preparation:

- Sethi, S., Vorontsov, I.E., Kulakovskiy,I.V., Greenaway, S., **Williams, J.A.**, Makeev, V.J., Brown, S.D.M., Simon, M.M., Mallon, A-M. 2019. Deciphering the impact of enhancer architecture on gene function and mouse phenotypes. Nature Communications. In Submission.

- **Williams, J.A.**, Banks, G., Sethi, S, Greenaway, S., Nolan, P.M., Gkoutos, G.V., Simon, M.M., Mallon A-M., Machine learning predicts abnormal circadian rhythm phenotype-contributing genes in mouse. In preparation.

- **Williams, J.A.**, Slater, L., Simon, M., Mallon, A-M, Gkoutos, G.V. The Neuro Behavior Ontology as a tool for stratifying collections of autism probands. In preparation.

- **Williams, J.A.**, Russ, D., Simon, M., Mallon, A-M, Gkoutos, G.V. Harnessing deep phenotyping reveals novel epistatic interaction networks of risk loci for autism-spectrum related traits. In preparation.

- **Williams, J.A.**, Russ, D., Simon, M., Mallon, A-M, Gkoutos, G.V. Evidence for causal relationships between chronotype and psychosocial traits through Mendelian randomization. In preparation.

# Contents

# List of Figures

# List of Tables

# Chapter One

# Introduction

The burden that mental health and behavioral disorders place on society is immense, with one in four people on the planet estimated to experience mental health problems during their lifetime, according to the World Health Organization (Murray, Lopez, and others, 2002). Among adolescents, at least 4% will be diagnosed with major depression (Costello et al., 2004). Recently, advances in molecular biology and genetics have transformed our study of mental health. Genetic epidemiology has grown from a discipline that studied small families and performed genetic association analyses within families or twins, to a discipline studying large cohorts to leverage the increased power of population-level genetic associations (Merikangas and Merikangas, 2016). When applying genetic associations to neurobehavioral disorders, however, these approaches often fail to report significant findings or to be repeatable (Jones et al., 2013). In the model organism community, studying genetic associations with the power of a large human GWAS study remains a challenge, especially for the mouse, as ethics, financial cost, and time are key limiting factors to large gene/trait associating studies. The mouse, however, provides an invaluable tool for investigating mechanism behind behavior through gene knockdown, knockout, and even wild-type behavioral and experimental assays (Mandillo et al., 2008; Barnard and Nolan, 2008). High throughput reverse genetics screens in mouse have been useful in elucidating a range of physiological and

behavioral phenotypes linked to gene perturbation, and have been essential in prioritizing deleterious mutations in humans (Dickinson et al., 2016; Boudellioua et al., 2017; Smedley and Robinson, 2015). Phenotypes investigated in these studies, and related conditions, have ranged from rare diseases and behavioral processes which affect every human on earth.

## 1.1 Introduction to circadian biology

Nearly every cell in every organism on the planet has a circadian rhythm, a biological process which oscillates over a 24 hour period. In mammals, the sleep/wake cycle largely regulated by an endogenous circadian rhythm influenced by external cues called zeitgebers, or timekeepers (Albrecht, 2012). Light, a zeitgeber, enters the retina and entrains the core mammalian clock in the suprachiasmatic nucleus (SCN) via the optic nerve and the circadian entrainment molecular pathway. The SCN is entrained primary by the light/dark cycle, termed photoentrainment: rods and cones (photoreceptors) and the retinal ganglion are all sensitive to light. This input is transmitted to the SCN via the retino-hypothalamic tract to entrain the clock on a daily cycle (Hughes et al., 2015). The core clock exhibits a transcriptional-translational feedback loop. Translated clock proteins (CLOCK, BMAL1) act as transcription factors for additional clock genes (*Per1/2/3,Cry1/2*). These genes, once translated and dimerized, negatively inhibit the action of CLOCK and BMAL1 before being phosphorylated and degraded. This process takes roughly twenty-four hours (Takahashi, 2017). This central clock acts as a pacemaker tissue-specific clocks throughout the rest of the body, and influences every organ system. There have been several computational methods developed to identify rhythmically cycling genes involved in circadian biology, most focusing on statistical approaches (Hughes et al., 2017). While many core circadian clock genes are known, contributors to circadian phenotypes remain to be predicted and cataloged. In the most wide-ranging mouse phenotype project attempted, the International Mouse Phenotyp-

ing Consortium, circadian studies are not part of the mandatory testing pipeline (Brown and Moore, 2012a). Lack of high-throughput experimental circadian phenotyping, which by its very nature takes days to perform, has led to recent attempts to predict genes involved in circadian physiology through machine learning (Anafi et al., 2014a), and meta-analytic (Brown et al., 2017c) methods. While many studies have used *in silico* methods to characterize circadian genes (Hughes et al., 2017), predictive models have been lacking in the literature. Abnormalities of circadian rhythms are associated with a myriad of neurobehavioral disorders in human, from schizophrenia and autism to major depressive disorder (Karatsoreos, 2014a; Takumi et al., 2020).

## 1.2 Psychological traits and the circadian clock

The relationship between circadian biology and psychological, neurodevelopmental, and mental health traits is well known. Mood disorders, such as depression, bi-polar disorder, and anxiety are diagnosed by behavioral assessment; no current clinical biomarkers are widely used in practice (Watmuff et al., 2016). Unlike observing a centrally controlled trait such as delayed circadian period, the presentations of many psychological disorders are heterogeneous (Wray et al., 2018b). Recent evidence supports the involvement of circadian rhythms in psychological disorders. For instance, actigraphy data has been used to produce measures of chronotype in persons with depression (Burton et al., 2013). Previous studies accessing the genetic basis for circadian biology and manic behavior, for example, have often focused on genes involved in the core pacemaker (Moon et al., 2016). With the availability of deep phenotyping and massive genomic coverage in the UK Biobank (Allen et al., 2014a), opportunities exist for surveying mental health behavior outcomes en mass, and exploring the connection between circadian biology and symptoms of neurobehavioral disorders. When attempting to survey an array of heterogeneous behavioral disorders, it may be beneficial to

ask how behaviors are delineated.

## 1.3 Characterising behavior computationally

One cannot meaningfully study the genetics of neurobehavior without first characterising what are the expected behaviors and how to assay their alterations. To this end, computational biologists have developed biomedical ontologies to organize and structure biological knowledge. The largest and most well known ontology is the Gene Ontology, (Gene Ontology Consortium, 2015), which characterizes gene function and the dynamics of gene products through three domains of knowledge: biological progresses, which are made up of molecular functions, and take place in particular cellular components. In the domain of human health and disease, the Human Phenotype Ontology (Robinson et al., 2008) depicts phenotypic abnormalities, including behavioral traits from sleep-wake disorders to autism. Biomedical ontologies have been designed for mammalian species as well as humans, facilitating the transfer of gene function annotations between homologous human and mouse genes (Smith and Eppig, 2009). The AberOWL repository has catalogued hundreds of ontologies, from the Alzheimer's Disease Ontology to the Zebrafish Phenotype ontology (Hoehndorf et al., 2015). Perhaps the most widely used ontology for describing behavior is the Neuro Behavior Ontology (Gkoutos, Schofield, and Hoehndorf, 2012), which provides the foundational description of behavior in both mouse and human specific phenotype ontologies.

## 1.4 An overview of omics data for behavioral trait modeling

This study uses two main modes of multi-omics data: DNA microarrays for studying the genome and RNA-sequencing assays for studying the transcriptome.

### 1.4.1 DNA Microarray

Genomic medicine historically starts at the level of the protein. Over a century ago, the ABO blood group system was characterized. This provided a biomarker that reflected underlying genetic variation before the discovery of genetic structure (Landsteiner, 1901; Landsteiner, 1961), making blood transfusions safe because of precise biomarker detection. In recent decades precision medicine has moved from isolated biomarker identification to holistic, genome-wide investigations of coding and non-coding DNA. This allows testing for several genomic variations: single-nucleotide polymorphisms (SNPs) involved in mis- or nonsense mutations, copy number variations (CNVs), structural transpositions/rearrangements, and splice-cite substitutions which affect the translated protein. Common technologies to detect these variations include DNA microarrays for CNV and SNP detection, and next-generation sequencing (NGS) based DNA-sequencing. In this study, I used data generated from DNA microarrays to genotype the Simons Simplex Collection and the UK Biobank (Fischbach and Lord, 2010a; Collins, 2007). SNP-detecting microarrays probe for thousands of known SNPs via included oligos, are relatively cheap and allow a wide array of potential SNPs and disease-associated loci to be probed. As a cost-effective and easily portable standard of assaying for thousands of potentially deleterious variants, SNP arrays have been manufactured by several companies including Illumina, Agilent, and Affymetrix. Studies have shown a high degree of correlation between different platforms, bolstering trust in use of the technology for accurate

patient assessment (Li et al., 2015). DNA microarrays have several limitations for SNP/CNV detection, including an a-priori decision of probes to be mapped, and the poor resolution of SNPs within of SNP deserts, genomic regions of low SNP density (Surrey et al., 2016). The cost effectiveness of DNA microarrays lend them to be used by biobanks, which facilitate deep phenotyping of participants useful for characterizing behavioral disorders. The study of behavioral genomics presented here incorporates data from human genomics as well as mouse transcriptomics.

### 1.4.2 RNA-Seq

RNA-Sequencing revolutionized transcriptome studies by providing quantitative estimates of gene product abundance (Wang, Gerstein, and Snyder, 2009). In contrast to microarray experiments, most RNA-Seq protocols involve capturing mRNA from lysed cells via the mRNA poly-A tail and converting the mRNA to a more stable cDNA library. The library is then fragmented, and after adapters are added the library is sequenced via one of many next-generation sequencing (NGS) technologies (Mortazavi et al., 2008). Resulting fragments, termed reads, are aligned to a reference transcriptome or genome. After reference alignment, a transcriptome-wide profile of gene activity is quantified by mapping reads to annotated coding regions of the genome. Unlike DNA microarray experiments, where genes must be known *a priori*, the unbiased nature of RNA-Seq facilitates discovery of new gene activity, uniquely spliced transcript isoforms, and measurable products of non-protein coding genes. One particular benefit of RNA-Seq studies is capturing the dynamic processes of transcription. Levels of detail range from single-cell specificity, to entire organs treated as nearly homogeneous sets. Unlike genomic experiments, RNA-Seq captures the dynamic transcriptome, facilitating studying how physiology changes in response to environmental and repetitive cues or cycles. Transcriptomic datasets can be further explored by creating

multiple RNA libraries through time, enabling one to study how changes in day/night environment or endogenous neuroendocrinological cues affect genes associated with neurological disorders (Hastings and Goedert, 2013).

## 1.5 Methods in computational behavioral genetics and phenomics

The methods used in this work include inferential, causative, and predictive statistical and machine learning models. I have also had extensive collaboration with colleagues at the MRC Harwell Institute who have performed experimental work, including RNA-sequencing and mouse phenotyping, in Chapter 2. Work in Chapter 3 is partly the result of a collaboration with the University of Jean (ES) during the development of the Aging Neurobehavior Ontology.

### 1.5.1 Statistical inference for gene association studies

Statistical inference for gene association studies relies on generalized linear models, regressing a trait or outcome of interest against a single locus on the genome. Analyses typically include covariates aimed at minimizing confounding variables, including those generated by sequencing or microarray platform, by ancestry or admixture, sex, and by phenotypic covariates which may moderate findings such as body mass index or age. This procedure is repeated hundreds of thousands or millions of times for each allele in the genome, resulting in many independent models. This necessitates a strict control of false positive rates, typically with a p-value of 1e-8 set as a genome-wide standard for significance.

### 1.5.2 Causal inference based on genetic inheritance

Both genome wide association studies and differential gene expression studies use statistical inference to model associations using omics data. Differential transcriptomics studies model how individual transcripts change in bulk expression or differential exon usage between conditions. Genetic epidemiology can model trait - gene and trait - gene x environment effects. Both of these model association, and cannot infer causality. Using traditional statistical approaches, one cannot ascertain if a particular SNP causes an observed trait, or if down-regulation of one gene causes the up-regulation of another, as both of these are experimental designs are observational. Instrumental variable analysis, originally developed in economics, has been used in genetic epidemiology to model the relationship between traits mediated by gene exposure. It was originally used to model the relationship between modifiable exposures and disease (Davey Smith and Ebrahim, 2003a). Instrumental variables are those which robustly predict an exposure but have no independent association with an outcome; any association must be mediated solely by the exposure (Labrecque and Swanson, 2018). Building such causal models rely on genetic data, especially those made available via GWAS studies. Mendelian randomization (MR) assumes its name due to the Mendel's law of independent assortment, which posits that alleles should be randomly distributed. While this is true in Mendelian genetics, alleles in linkage disequilibrium will break this assumption, but a MR experimental design will account for this by selecting loci which are independent. Because assortment is independent, the presence or absence of a particular minor or major allele within a population is randomly distributed. Thus, MR can be thought of as a natural randomized controlled trial, in which participants are randomized at birth, and exposed to different levels due to their likelihood of developing a trait due to genetic exposure (Davey Smith and Ebrahim, 2003b). MR can mimic environmental, physiological, or psychological exposures and have been used to characterise systems biology, model pharmacogenomics, and other areas (Davey Smith and Ebrahim, 2003a). An early application

of MR was performed in 1986, when Katan modeled differential alleles in the APOE gene to investigate the observational relationship between cholesterol levels and cancer (Katan, 2004). Due to one's exposure to an allele occuring prior to an exposure to any outcomes, reverse causation is negated when proper instrumental variables are chosen. Similar to methods in GWAS or transcriptomics, most models used in MR analyses are interpretable linear models, and associations between variables are causal only due to the correct experimental design. MR models have been used in recent years to characterise the relationships between many neurobehavioral traits or psychiatric conditions, including linking smoking behavior (exposure) to bipolar disorder and schizophrenia (outcomes) (Yuan, Yao, and Larsson, 2020) and NR3C1 expression levels (exposure) to psychosis (Iftimovici et al., 2020). In this work, MR will be used to model the relationships between self-reported chronotype (an exposure) and several psychological outcomes. By using causative modeling to investigate the association between chronotype and mental health, this work will provide direction for further longitudinal or interventional experiments using chronotype as a intervention when studying behavioral outcomes. A detailed discussion of these methods used are included in Chapter 3.

### 1.5.3 Machine learning based on genomic and phenotypic data

Modern machine learning has its roots in early attempts to understand the learning process, when a simple model of neuronal processing were developed - the perceptron (Rosenblatt, 1958). As computational resources have increased, machine learning has become useful to the biomedical community. Machine learning is the application of a class of induction algorithms which seek to learn outcomes through iterative training without being explicitly programmed (Kohavi and Provost, 1998). As eloquently proposed in Breiman's two cultures, data modeling and algorithmic modeling approach inference in different ways (Breiman,

2001). Classical statistical modeling seeks to first create a model of the world ($X$ independent variables) which explains $Y$ dependent responses. Based on assumptions about $X$, models are built to attempt to extract information about how the $X$ and $Y$ are related. In contrast, machine learning methods are often viewed as a black box. Instead of explicitly attempting to model how $X$ generates $Y$, algorithms are built that operate on $X$ to predict the $Y$ response variables. This is done independently of assumptions about *how* $X$ leads to $Y$. To achieve good measures of predictive accuracy, machine learning methods depend on large amounts of data and can require vast computational resources, both of which are now available to computational biologists. Two divisions of machine learning processes are distinguished by the presence of labels.

### Unsupervised learning

Given data without labels, a problem is formulated to detect latent patterns present in data. Given a set of data, patterns can be constructed with many different approaches. The most familiar may be method of moments estimators, where the first moment of a sample of data is the mean, and the second moment about the mean is sample variance. These calculations depend on no outside labels of the data. To aid in analysis of multidimensional data from large genomic experiments, dimensionality reduction and clustering algorithms are often used. This thesis uses variations on hierarchical clustering, which models the Euclidean distance between data (such as genes or phenotypes) as a metric of gene-gene similarity. It also uses principal component analysis (PCA) as a dimensionality reduction technique; this is commonly performed in GWAS analyses to account for population structures. In Chapter 3, this work extensively uses clustering based on the similarity of subject's phenotypic traits to one another in a measure called semantic similarity.

### Supervised learning

When given labels associated with data, those labels can serve as targets for supervised learning algorithms. The challenge of predicting phenotype annotations (and punitively assigning functions to genes based on these phenotypes) is a classification problem using phenotypes or phenotypic profiles as targets (dependent variables). Supervised learning approaches used in this work include a variation on random forests, in which initially random decision trees are made out of biological features, and an ensemble of them (a forest) are used to assign the presence or absence of a target trait. The variation adapts trees to better fit data as it is learning over training iterations. Other approaches used in this work involve supervised feature selection with a penalized generalized linear model. To measure the performance of our modeling attempts, the following general strategy is adopted:

- data are split into training and testing sets
- during hyperparameter optimization, training sets are again split into testing and validation as a cross-validation procedure to select a well performing model
- model selection is judged by the area under the receiver operating curve and summary statistics including accuracy, specificity and sensitivity
- after model selection, models learned are evaluated by performance on the yet-unseen testing data
- as an alternative to initial data splitting, newly generated data or additional study cohorts can be used to measure the accuracy of models

This procedure is then repeated several times, and average metrics are used to evaluate the generalizability of the classifier. This is done to protect against train/test splits when features may be unevenly distributed among subjects and when there are a minority of subjects in

the analysis who have a particular target class. When predicting ontology annotations, especially rare phenotypes, class imbalances will occur (Wei and Jr, 2013). Having an equal number of data observations (genes) with and without a given binary phenotype annotation will increase training efficiency. We use both under- and over-sampling, when appropriate, to balance training datasets. When validating predictions on unseen data, class imbalances native to the data are retained, to make predictions more accurate to current levels of gene function annotation. Our modeling thus far has been predicting phenotype annotations from the results of gene expression analysis experiments (below). To mitigate potential model inaccuracy from withholding a percentage of phenotype/gene annotations, a multiple cross-validation approach may be used instead of an initial holdout validation set.

### 1.5.4 Method evaluation

To evaluate supervised machine learning models, I employ diagnostics which can be derived from a confusion matrix:

| | | Truth | |
|---|---|---|---|
| | Positive | Negative | Total |
| Positive Prediction | $a$ | $b$ | $a+b$ |
| Negative Prediction | $c$ | $d$ | $c+d$ |
| Total | $a+c$ | $b+d$ | $N$ |

Table 1.1: *An example of a confusion matrix, or a 2x2 contingency table. The set {a,b,c,d} are all integers, and indicate the number of true positives, false positives, false negatives, and true negatives in a binary classification problem. N represents the total number of entities classified.*

Accuracy is calculated as the sum of true positive and true negatives over the popu-

lation:

$$ACC = \frac{a+d}{a+b+c+d} \tag{1.1}$$

Sensitivity, recall, or the true positive rate, is calculated as the sum of true positive cases over the sum of true and false positives:

$$Sens = \frac{a}{a+b} \tag{1.2}$$

Specificity, or the true negative rate is calculated as the sum of true negative cases over the sum of true negative and false positive cases:

$$Spec = \frac{d}{c+d} \tag{1.3}$$

While accuracy gives a measure of the percent of classifications which are correct, this can be misleading when there is a large imbalance between two cases: for instance, a classifier who classified a set of 100 SNPs as pathogenic, when in reality 99 were, would have an accuracy of 99%. Such a skewed dataset, however, benefits from other measures. I use the area under the receiver operating characteristic (ROC) curve, denoted in this work by AUC. It provides a balanced measure of the trade-off between specificity and sensitivity. An AUC of 0.5 indicates a completely random binary classifier, while an AUC of 1.0 denotes perfect classifier performance (Fawcett, 2006). Confidence intervals around the ROC curve provide measures of the uncertainty of the model's performance.

To evaluate unsupervised learning models in this thesis, two approaches are used. In the first, data are permuted thousands of times, and measures of community or cluster modality (density, node degree) are taken to produce an null, or random, distribution of measures of class centrality. By comparing the modularity of members of a community to the null and counting the frequency of observed values more extreme than those within a community, p-values are generated, producing a measure of how unexpectedly connected a community is with itself compared to data not in that community. I also evaluate unsupervised learning clusters by biological observation: if genes, for instance, participate in more

shared biological processes than would be expected by chance, this leads to the conclusion that the community of genes as a whole may be involved in similar functions or interact with biochemical pathways up or downstream of one another.

Inferential models from GWAS analysis (Chapter 5), and causative models from MR analysis (Chapter 3), are evaluated primarily by the strength and variance of their associations as measured by a log-odds or the transformed $\beta$ coefficient. GWAS are additionally verified by using their features in supervised learning models, and MR analyses have extensive post-hoc tests to ensure statistical robustness and appropriate experimental design.

## 1.6 Thesis aims and objectives

Neurobehavioral disorders are phenotypically and genetically complex, and often diagnosed through observational study or subjective assessment alone. In order to further the understanding of the genetic basis of neurobehavioral disorders, it is necessary to devise ways of both predicting gene/trait associations and to uncover new relationships between the observational traits which neurobehavioral disorders manifest. To make inroads into this domain, several statistical methods and biological principles can be leveraged; from the homologous relationship between mouse and human neurogenetics to modeling causal relationships between traits, though to the semantic modeling of relations between traits and then interrogating their genetic basis. The ultimate aim of this thesis is to predict neurobehavioral gene/trait relationships and provide a basis for prioritizing experimental investigations into genetic causes of neurobehavioral dysfunction. To address this, this thesis aims to:

- 1: Predict genes which are involved in producing abnormal circadian rhythm behavior in mice using supervised machine learning methods

- 2: Investigate the causal relationship between chronotype, a measure of circadian rhythm, and behavioral traits linked to mental illness or psychological disorders in humans

- 3: Mine psychological assessment instruments to model subtle behaviors which, in aggregate, are indicative of complex psychological disorders, such as autism or schizophrenia, and use that to reduce the phenotypic heterogeneity among populations with complex behavioral diagnoses

- 4: Investigate genes which are associated with individuals who share both a common behavioral diagnosis (autism) and share a common phenotypic profile derived from aim 3.

Each of these aims involves modeling of associations between neurobehavioral traits and genetic material which, when expressed or mutated, will influence these traits. Figure 1.1 encapsulates the structure and interconnectedness of this project. The figure may be read from top to bottom, showing the sequence of chapters, relevant data, general methods used, and outcomes. Each chapter's data flows into its methods and then outcomes. In Chapter 2, I am to predict genes involved in abnormal circadian biology. This is accomplished by harnessing data from mouse in predictive models using supervised machine learning to produce genes which may produce circadian phenotypes. Knowing that circadian biology impacts many neurobehavioral and psychosocial traits in humans, Chapter 3 uses genome wide association studies as a basis for modeling the causal relationship between chronotype (a manifestation of circadian rhythms) and various self-reported traits relatd to mental illness and social wellbeing. After finding such associations between chronotype and behavioral traits in a wider population, the thesis turns to focusing on how such traits relate to one another. To address this, the traits which make up complex psychological disorders, including autism and schizophrenia, are mined from gold-standard diagnostic instruments in Chapter

4. These traits are used to engineer biomedical ontologies, on which autism probands are modeled to investigate the underlying (endo)phenotypic structure of a complex neurodevelopmental disorder. Lastly, in Chapter 5, I use phenotypic profiles from Chapter 4 and the genome-wide association study approach from Chapter 3 to investigate the genetic architecture of probands diagnosed with autism yet presenting with distinct phenotypic profiles. Clusters of probands who have a distinct phenotypic presentation are investigated for associations between genetic mutations which separate them from the rest of the studied autism cohort, revealing links to circadian biology studied in Chapter 2. The fundamental biology studied in a model organism in Chapter 2 permiates the thesis through Chapter 5, leading to future work combining complex behavioral phenotypes in population-based human studies with collaborative mechanistic studies detailing individual endophenotypes in mouse.

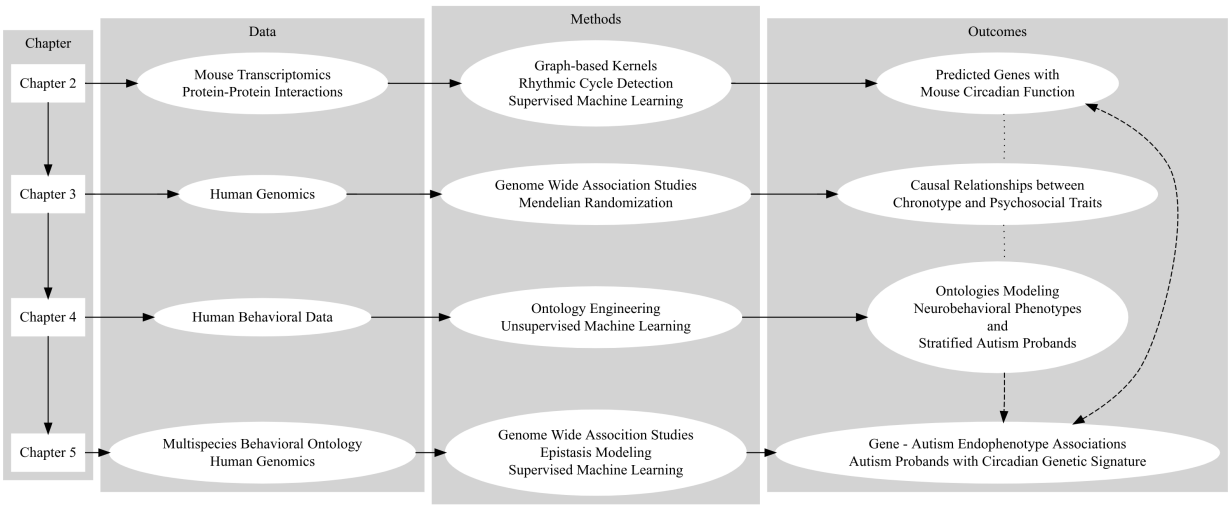



Figure 1.1: *This work is divided into four substantive chapters, each of which attempt to address interrelated but independent aims. Each chapter is connected to each other (left panel, "Chapter"). Solid edges represent an explicit flow of work: in Chapter 2, mouse transcriptomics and protein interaction data feed into kernel, cycling, and machine learning methods. These methods produce results, specifically genes with predicted circadian function in mouse. Outcomes (both biological relationshps and methodological insight) from each chapter indirectly inform the next, as indicated by dashed edges in the rightmost panel (Outcomes). Outcomes from Chapter 4 directly inform those in Chapter 5 (solid dashed edges), while the outcomes of Chapters 2 and 5 inform each other and suggest future research integrating mammalian and human studies.*

# Chapter Two

# Predicting genes involved in abnormal circadian rhythm traits in mouse

## 2.1 Background and Chapter Overview

As described in Chapter 1, most organisms will anticipate regular daily changes to their environment which replay over a 24 hour period, chiefly involving light, food, and temperature. These daily circadian influences, or zeitgebers, can advance or delay the phase of the central circadian clock in the uprachiasmatic nucleus (SCN). The transcriptional-translational feedback loop which regulates circadian rhythms is largely conserved between different cell types, discounting built-in redundancy of a Clock protein paralog in the SCN (Patke, Young, and Axelrod, 2020). The SCN exhibits intercellular coupling, with an interplay between VIP- and GRP- producing neurons in the core with AVP-expressing cells in the shell together regulating the core circadian clock (Hamada, Antle, and Silver, 2004). The SCN acts as the central pacemaker for tissue-specific clocks. Projections from the SCN are largely sent to the diencephalon, and then to the rest of the body (Mieda, 2019).

Every tissue in both mouse and human contains circadian clocks, by which the rhyth-

mic expression of circadian-influenced genes will oscillate in a tissue-specific manner. The SCN acts as the central pacemaker, coordinatingTissue specific transcriptome profiling has revealed differing patters of expression, many of which are able to have sustained rhythm even in isolation of the central pacemaker (Yoo et al., 2004; Yamamoto et al., 2004). Peripheral clocks have been impicated in wound healing, heart rate and respiration, and immune function (Finger, Dibner, and Kramer, 2020). One of the best-studied peripheral circadian clock is in the liver, where metabolism is affected by clock genes including the Per and Cry family. While these tissue-specific clocks maintain their own rhythm, the SCN is still need to coordinate the phases of these tissue-specific clocks (Tahara et al., 2012). They are considered dependent ('slave') oscillators that while dependent on the SCN are strongly influenced by external zeitgebers such as feeding cycles. The circadian role of the liver has received attention: *in-vivo* mouse models have shown a hepatic cytokine to activate clock genes in the liver (Chen et al., 2019), and in humans beneficial effects of dietary fasting have been suggested to be influenced by peripheral metabolic clocks (Lessan and Ali, 2019).

Chronobiological deficits can have system-wide influences ranging from the central nervous sytem, suggesting possible core clock malfunction, to influencing non-neuronal tissues not explained by the SCN's central pacemaking function alone. Genome-wide association and mutational studies have linked variants in core clock genes to sleep/wake disorders, disruptions due to jet-lag, and even sleep-related bone loss (Swanson et al., 2017). Recent GWAS of self-reported chronotype (lark or owl, corresponding to being a 'morning person' or 'night person') reveal several candidate genes associations with previously unexpected circadian function (Hu et al., 2016). The translational importance of understanding circadian biology is not limited to neurobehavioral function; the interplay between metabolism and circadian biology has been highlighted heavily in recent years. The gut microbiome, adipose cytokines, and metabolic hormones from grhelin to leptin are all strongly regulated by circadian biology (Li et al., 2020; Socaciu et al., 2020; Pan, Mota, and Zhang, 2020).

Recent Mendelian randomization studies have also suggested a strong causal link between chronotype (a gross circadian phenotype) and body composition, free fatty acid circulation, and adiposity (Adams and Neuhausen, 2019; Jones et al., 2019).

As such, identification of genes, which contribute to circadian biology, either as part of core clock machinery or up- or down-stream actors, is of great importance. There have been recent attempts, resulting from high throughput proteomic and interaction screens, to prioritize clock genes that bind to CLOCK (Zhao et al., 2007), as well as to identify protein-gene interactions involving the ZFHX3 protein (Parsons et al., 2015). Such high throughput experimental assays are invaluable, and computational predictions, combining several of these features, have been used to screen for potential clock genes (Anafi et al., 2014b). Statistical analysis of high throughput calorimetry data from the International Mouse Phenotyping Consortium have been harnessed to identify novel circadian functions for well-studied genes, further highlighting the utility of integrating data from core clocks, peripheral or exogenous zeitgebers (Zhang et al., 2020).

Creating an *in-silico* method of prioritizing genes of interest, resulting from mammalian circadian phenotyping studies, will dramatically reduce the search space for potentially novel circadian genes, leading to both cost and time savings. My method screened several features, indicative of circadian function, to classify novel circadian genes. I combined measurements of RNA-Sequencing levels, at four different timepoints in the supcrachiasmatic nucleus (SCN) and the liver in mouse, to indicate enrichment of genes in the both the central and peripheral pacemaker. I also assessed genes, expressed in the SCN and liver, in terms of their likelihood of being expressed over time in a oscillating periodic fashion taking 24 hours. Next, potential interactions of proteins, expressed in the SCN, with both known circadian proteins, as well as with any other proteins whose gene progenitors were expressed in the SCN, were studied. The level of expression across several tissues in the mouse body was also considered, and the degree of specific expression in the SCN was cal-

culated. Lastly, my candidate novel genes were further compared to several previous studies characterizing circadian genes in the SCN. I present an analysis of three candidates with previously unknown circadian function, and two candidates, with known circadian function, but not known abnormal circadian phenotype associations in mouse. Of these, I present evidence of circadian phenotypes in Gastrin Releasing Peptide (Grp) mice, thanks to the help of experimentalist colleagues.

The suprachiasmaic nucleus (SCN) is the central pacemaker of the body, maintaining entrainment of peripheral biological clocks in nearly every body tissue (Li et al., 2008). By combining attributes of several known clock genes, I used the RUSBoost machine learning algorithm to rank genes by their likelihood to contribute to the 'abnormal circadian rhythm' phenotype in the mouse. Several known clock genes, with no predicted phenotype in mouse, were suggested to actually contribute to circadian biology on the phenotype level. I also predicted classes of genes not currently known to contribute to circadian phenotypes in mouse or human. These genes are worthy of further study in model organisms to elucidate the molecular mechanisms contributing to clock disruption.

## 2.2 Methods

### 2.2.1 Circadian phenotype selection

Mammalian phenotype to gene annotations were downloaded from the Mouse Genome Database on 30 Nov 2016. Genes annotated to MP:0001393 in the Mammalian Phenotype Ontology (MP), labeled "Abnormal Circadian Rhythm," were extracted, with positive annotations used as targets in my analysis. Genes annotated to any children of 'Abnormal Circadian Rhythm' MP were likewise annotated with MP:0001393 and included as positive

targets.

### RNA-Sequencing and Experimental Validation

In this work, all wet-lab experiments, including phenotyping, collection and RNA extraction, were performed by colleagues in the Nolan group at MRC Harwell Institute. All computational and statistical analysis is my own.

All animal studies were performed under the guidance issued by the Medical Research Council in Responsibility in the Use of Animals for Medical Research (July 1993) and Home Office Project Licenses (30/3384 and 30/3206), with local ethical approval. All animals used in this study were bred and maintained at MRC Harwell. When not being tested, mice were housed in individually ventilated cages under 12/12 h light/dark (LD) conditions with food and water available ad libitum.

## 2.2.2 Circadian Phenotyping of *Grp knockout mice*

Gastrin-releasing peptide (*Grp*) knockout mice were obtained from the International Mouse Phenotyping Consortium (IMPC; https://www.mousephenotype.org/). Six 34 week old homozygous female mice were used for circadian analysis. Nine isogenic 34 week old C57BL/6N females were used as controls. Circadian analysis was performed using the COMPASS system (Brown et al., 2017a). Briefly mice were individually housed and activity data captured by passive infrared sensors for 5 days in a 12:12 LD cycle, followed by 9 days in constant darkness. Data were rebinned using custom python scripts converted to AWD files for analysis on Clocklab (Actimetrics) (Brown et al., 2020).

### 2.2.3 Tissue collection and RNA extraction

C57BL/6J animals were singly housed under a 12:12 hour LD cycle for at least two weeks prior to dissection. Mouse activity was monitored by wheel running (Banks and Nolan 2011) in order to confirm entrainment to LD cycles prior to dissection. Mice were sacrificed by cervical dislocation at either Zt 3, 9, 15 or 21. SCN punches were collected as in Jagannath et al., 2013 (Jagannath et al., 2013) and liver taken from the same animals. Samples were flash frozen on dry ice and stored at -80 degrees C until RNA extraction. Total RNA was extracted using RNAeasy column (QIAGEN). Quality and quantity of RNA were measured using an Agilent Bioanalyzer (Thermo Fisher Scientific). The SCN quality of the tissue dissection was confirmed by qPCR of Six6 (an SCN enriched gene) as described in by Jagannath (Jagannath et al., 2013).

### 2.2.4 RNA sequencing

RNA sequencing was performed by the Oxford Genomics Centre (Wellcome Trust Centre for Human Genetics, University of Oxford). 500 $\mu$g of RNA samples were sent for sequencing. The samples underwent poly-A selection, after which two multiplexed DSN Library Preparations (6 samples/multiplex) were generated. Each multiplex was run on one 50bp PE lane of a HiSeq2000.

### RNA-Sequencing Analysis

To predict novel genes contributing to neurobehavioral phenotypes of circadian rhythm, time-series gene expression studies can be used to exploit observed transcript oscillations across a day. To study circadian rhythms in mice, RNA-Seq libraries were created from

16 male C57BL/6J strain mice. Following procedures previously described (Parsons et al., 2015), animals were sacrificed at four time points in a twenty-four hour range denoted in zeitgeber time (Binder, Hirokawa, and Windhorst, 2009). By convention, zt0 corresponds to lights on, zt12 corresponds to lights off in a twelve-hour light/dark cycle. Mice were sacrificed at zt3, zt9, zt15, and zt21. Samples were taken from the suprachiasmatic nucleus (SCN) of the brain and the liver. After purification and cDNA library creation, samples were sequenced as referenced above.

Raw paired-end reads were analyzed for quality control with FASTQC (Andews, 2015). Reads were aligned to the mm10 genome with the Bowtie2 aligner in the TopHat application suite (Trapnell et al., 2010; Langmead and Salzberg, 2012). HTSeq was used to extract read counts using Ensembl 83 annotation (Flicek et al., 2014; Anders, Pyl, and Huber, 2015). Processed reads were analyzed for statistical distributional assumptions and potential sample mis-annotation. Three samples from each time point and tissue were kept for further analysis, resulting in 24 libraries.

Differential expression between tissues and time points was analyzed with the DESeq2 package (Love, Huber, and Anders, 2014). Assuming a negative-binomial distribution to count data, standard settings were used to fit generalized linear models and perform Wald tests for differential expression. Transcripts expressed with a log-fold change of greater or less than 2 between tissues were considered differentially expressed, subjected to a Benjamini-Hochberg adjusted p-value of $< 0.05$ (Benjamini and Hochberg, 1995). Transcript per Million (TPM) measurements were additionally calculated for each gene, and are reported as the mean TPM in all SCN samples +/- the standard error of the mean.

To test for rhythmic signatures related to circadian rhythms in each tissue, several algorithms were used. The RAIN algorithm was used to detect cycling at four time points (Thaben and Westermark, 2014). Using DESeq2-normalized read count estimates, period

time was explicitly set to twenty-four hours and default RAIN settings and FDR corrections were used. To enable detection of rhythms across the entire 24 hour period with only four time points in each tissue, a second 24 hour cycle was imputed by duplicating the first cycle's values. Visual inspection of core clock-related genes' transcriptomic patterns was performed to verify accuracy of results. Additionally, Lomb-Scargle (Lomb, 1976; Scargle, 1982) and JTKCycle (Hughes, Hogenesch, and Kornacker, 2010a) were implemented. All features from cycling/rhythmic analyses were combined and used as feature input, and correlated features were permuted during training splits to avoid multicollinearity of correlated feature detection methods.

To compare the distribution of expression from known circadian genes across various tissues, transcript-per-million (TPM) expression data from 25 tissues was obtained from the mouse ENCODE consortium (Mouse ENCODE Consortium et al., 2012). Each gene was binary coded if 1 TPM was present in a given tissue and totaled. The resulting vector of tissue totals per gene was mean centered and scaled over the standard deviation to create a tissue ubiquity vector. Since the particular genes represented in my tissue ubiquity metric were chosen by the ENCODE consortium, any genes discovered in my analysis but not included in ENCODE were missing. The tissue distribution of these genes was imputed simply with the median of all gene tissue-wise expression, in order to limit missing values in my downstream classification attempts.

## Protein-protein interactions

To derive a quantitative feature representing interactions of known circadian phenotype proteins with other proteins expressed in the SCN, protein-protein interaction graphs were constructed, including annotation with phenotypes of interest. Protein-protein interaction graphs were constructed from genes expressed in the SCN, at each time point separately, as

follows. Using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database version 10, all default data sources were used with an interaction threshold of medium confidence (0.400) except for textmining which was set to a higher threshold (0.950) (Szklarczyk et al., 2015a). Interactions between proteins are represented as an unweighted undirected graph, and combined into one graph containing all threshold-meeting interactions at each time point. This combined text-mining, fusion event history (proteins fused in some genomes may have similar functions in other species), similar pathway occurrence, and observed co-expression in tissue expression databases into one graphical representation for potential protein-protein interactions. Graphs from each of 4 timepoints were combined, taking the union of each graph to capture genes which were expressed in the SCN at any point across Zt3, Zt9, Zt15 or Zt21. The biological network of interactions is represented as a graph, $G$. In $G$, vertices represent proteins and edges represent undirected theoretical or experimental interactions between them. Vertex sets are represented as $V$, while edge sets are included in $E$. A convenient representation of graph $G$ is the adjacency matrix, $A$. $A_{i,j}$ describes similarity between $V_i$ and $V_j$.

## Diffusion kernel creation

Diffusion kernels are based on the heat equation and can be thought of as discrete versions of Gaussian kernels (Kondor and Lafferty, 2002). To compute the diffusion kernel, the Laplacian of the combined protein-protein interaction graph above $G$ was initially calculated as:

$$L = D - A \tag{2.1}$$

where $D$ is the diagonal matrix and $A$ is the adjacency matrix of the graph. The Laplacian $L$ of $G$ was used to compute the kernel $K$:

$$K = e^{\beta L} \tag{2.2}$$

where $\beta$ is the bandwidth parameter. The effect of diffusing the flow of information across proteins according to the bandwidth parameter $\beta$ can be seen by:

$$e^{\beta L} = I + \beta L + \frac{\beta^2}{2}L^2 + \frac{\beta^3}{6}L^3 + ... \tag{2.3}$$

where $I$ is the identity matrix (Schölkopf, Tsuda, and Vert, 2004). To calculate information diffusion from a random walk around proteins with known circadian phenotypes, the kernel $K$ was multiplied by $P$ to generate a score $S$:

$$S = PK \tag{2.4}$$

where

$$P = \begin{cases} 1, & \text{if an Abnormal Circadian Phenotype is annotated to a a protein} \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

The resulting feature, the diffusion score $S$, was obtained with a bandwidth $B$ set to 0.2. I assembled two features using this approach, each with a score (0,1) for every protein coding gene in my SCN dataset. The first feature was a diffusion score for each gene expressed in the SCN in my dataset and participating in predicted protein-protein interactions as outlined above. The second limited interacting proteins to those which were over-expressed in the SCN when compared to corresponding liver experiments (log fold change $> 2$, FDR adjusted p-value $< 0.05$). Only genes expressed in the SCN and involved in predicted interactions were annotated with features for circadian phenotype predictions.

## Phenotype modeling and classifier evaluation

To predict classification into presence or absence of an abnormal circadian rhythm phenotype, the RUSBoost algorithm was used as implemented in the R environment (Seiffert et al., 2010), (Carnagua, 2015). Fourty-five features (see Supplementary Table 2) were used, comprising

six categories: protein-protein interaction based diffusion scores, SCN expression, evidence of liver and SCN cycling and wave characteristics, evidence of paralogous circadian function in *D. Melanogaster*, and pan-tissue expression. My target was the presence or absence of the 'abnormal circadian rhythm phenotype.' To avoid overfitting, I limited each forest of decision trees to 30 in each iteration of the algorithm. A 70/30 training/test split was used, maintaining the proportion of each class. Normalization parameters included centering both training and test splits, and balancing class in the test split only via SMOTE (Chawla et al., 2002; Torgo, 2010). Feature importance was calculated via the gain of the Gini index given by each variable in a weighted tree, implemented in the Adabag R package as proposed in Hastie et. al. (Alfaro, Gámez, and Garcia, 2013; Hastie, Tibshirani, and Friedman, 2009). To calculate feature importance per gene of interest, local linear approximations of the model's behavior were deduced using the LASSO (least absolute shrinkage and selection operator) method, selecting the ten most explanatory features for each gene using the lime method (Pedersen and Benesty, 2019; Ribeiro, Singh, and Guestrin, 2016).

To test the generalizability of my approach, a repeated resampling study was performed, each generating a new train/test split. Two hundred train/test splits were created, and the algorithm was optimized on training data and validated on the test split. To measure the influence of the graph-based measure on training, during each resampling the diffusion kernel experiment was re-run with additional masking of known circadian genes by setting all scores in the vector $P$ above (eq.4) to zero. Outcomes with and without masking the kernel score were retained. Subjects were considered potentially novel circadian genes if they appeared in the top 90% of re-sampling splits (n = 246).

To characterize the gene function of predicted novel genes across species, gene enrichment analyses were conducted in the Biological Process domain of the Gene Ontology, the Mammalian Phenotype, and the Reactome pathway database using the XGR R package v.1.0.1(Fabregat et al., 2016; Smith and Eppig, 2009; Robinson and Mundlos, 2010). In each

analysis, a hypergeometric test and Benjamini-Hochberg false discovery rate correction were applied, with other settings used as default (Fang et al., 2016; Benjamini and Hochberg, 1995).

## Novel candidate prioritization

The 246 ranked novel candidates, with an emitted probability of $> 0.5$, were kept for further investigation. Of those with higher expression in the SCN than the liver, candidates with no known rhythmic process (GO:0048511) annotations were compared with experimental data from the Allen Mouse Brain Atlas (Jones, Overly, and Sunkin, 2009; Liu et al., 2007) and a pan-tissue circadian gene expression database (Pizarro et al., 2013). A recent meta-analysis of both RNA-Seq and RNA microarray studies investigating gene expression in the SCN and whole brain was also used to bolster evidence from my initial analysis (Brown et al., 2017c). A recent study of RNA-Seq in the SCN using six timepoints was also interrogated (Pembroke et al., 2015), to add annotations of evidence of cycling behavior in genes in an experiment with more timepoints than this current work.

Genes were additionally filtered by their tissue specificity among 27 tissues. RNA-Sequencing for all available tissues were downloaded from the mouse ENCODE repository (Mouse ENCODE Consortium et al., 2012) in Reads per Killobase per Million reads mapped (RPKM) format and averaged for each tissue. RNA-Sequencing reads for each time period of SCN data were converted to RPKM and then a mean was taken.

A recent benchmark of tissue specificity metrics found the *tau* statistic from to be a reliable indicator of the degree to which a transcript is deferentially expressed in one tissue compared to several others (Kryuchkova-Mostacci and Robinson-Rechavi, 2017; Yanai et al., 2005). To calculate tissue specificity, I first took the base 2 log then quantile normalized

the expression of each gene within each tissue. Then within each gene, the expression was democratized into 10 bins of equal density. Next, the *tau* controling factor was calculated, as:

$$\tau = \frac{\sum_{i=1}^{n}(1 - \widehat{x}_i)}{n - 1} \tag{2.6}$$

and:

$$\widehat{x}_i = \frac{xi}{\frac{max}{1 \leq i \leq n} x_i} \tag{2.7}$$

where $x_i$ is the binned gene count (0 - 9) for each tissue divided by the max, and $N$ is the number of tissues which the gene is expressed in. The $\tau$ metric was computed to compare the degree of which each gene was expressed uniquely in each tissue. Genes with a $\tau$ of 0.85 - 1 in the SCN were deemed to be specifically expressed in that tissue. To calculate the specificity of each gene, a final calculation is made:

$$\tau = \tau(\frac{qn}{max}) \tag{2.8}$$

where $qn$ is the quantile-norm gene count of a gene in a given tissue tissue, and $max$ is the maximum quantile-normed expression of that gene in any tissue studied.

## 2.3 Results

### 2.3.1 Machine Learning predicts 246 potential novel circadian genes

I used an ensemble classifier, RUSBoost with decision trees, to predict if genes were considered to be circadian or not, a binary classificaiton problem. To access the performance of my classifier, the first variable importance was measured. As seen in Figure 2.1 A, the feature which explained most of my model performance was a continuous score $S$ from projecting genes expressed in the SCN onto a protein-protein interaction graph, netting us information about how connected each gene is to neighbors who are known to produce circadian traits in

mouse. Measures of circadian amplitude, cycling potential (normalized adjusted p-values), expression in the SCN, expression distribution throughout the mammalian body, and known circadian function in fly homologues contribute to my predictive performance. In order to predict previously unannotated circadian genes, I trained my model on my entire data set, resulting in 246 potential genes which may contribute to the circadian phenotype in mouse, reducing the search space for new circadian genes expressed in the central pacemaker by over 98% Figure 2.1 B, and achieved an area under the receiver operating characteristic curve (AUROC) of 0.93. To test the generalizablity of my method on such imbalanced data, I performed multiple permutations of training/test splits, and I measured performance while masking the graph diffusion features by setting known circadian nodes in the training and testing sets to the mean of each feature. These metrics allow us to compare my ability to recover known circadian genes with other machine learning methods, but the biological relevance of predicted novel circadian genes is best accessed by investigating known gene functions.

### 2.3.2 Feature Characteristics

Several features were used in extracting information relevant to circadian classification. These stemmed primarily from two sources, namely an RNA-Sequencing study of genes expressed in the SCN and the liver, measured at four timepoints during a 24 hour period, and predicted protein-protein interactions among those genes. Differential expression analysis between liver and the SCN at four time points in mouse shows that more genes were over-expressed in the SCN compared to the those over-expressed in the liver. Interestingly, over ten thousand genes exhibited 24-hour cycling in the liver (a peripheral clock) compared to the SCN (Table 2.1). After excluding genes not expressed in the SCN and those not involved in protein-protein interactions in the STRING database, only 12,502 genes were

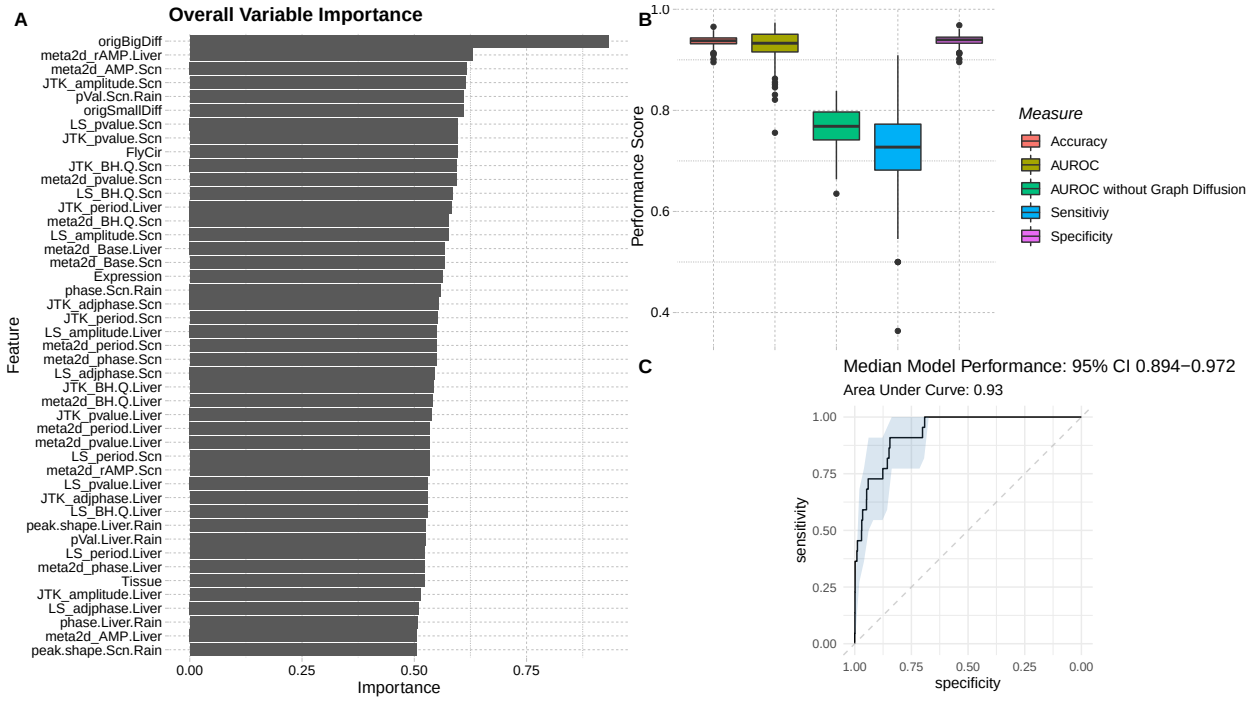



Figure 2.1: *Classification results rely on SCN expression and protein interactions to narrow search space for novel circadian phenotypes. Variable importance of the final model (a) is dominated by the measure of likelihood of predicted protein interaction with known circadian proteins, seconded by measures of rhythmicity in the liver and SCN. When serially trained, I predicted 246 genes previously unannotated with circadian phenotypes to have annotations. Accuracy, area under the ROC curve with and without graph metrics (AUROC, AUROC without Graph Diffusion), Sensitivity, and Specificity are depicted for all 200 models on their respective test holdouts (b). The median performing (by AUC) model's ROC curve is depicted in panel c, with an AUC of 0.93 +/-0.36.*

carried forward for subsequent analysis. Of those, 3058 were significantly over-expressed in the SCN (FDR-adjusted p-value $< 0.05$, log fold change $> 2$).

| | **SCN** | | | **Liver** | | **Both Tissues** | |
|---|---|---|---|---|---|---|---|
| | All | Cycling | PPI | All | Cycling | All | Cycling |
| Expressed | 21865 | 6335 | 12052 | 20138 | 10097 | 18205 | 2877 |
| Over-Expressed | 6533 | 1947 | 3058 | 3995 | 2008 | - | - |

Table 2.1: *SCN and Liver differential expression statistics. A larger number of genes were significantly over-expressed (log fold change $> 2$) in the SCN versus the liver, while the more significantly cycling genes were detected by RAIN in the liver than in the SCN. A final protein-protein interaction network was constructed from 12492 genes expressed in the SCN and participating in interactions in STRING. Of those, 3058 were over-expressed in the SCN. The analysis was limited to the 12492 genes in the large interaction network going forward.*

Gene expression is shown in Figure 2.2. While the differential expression study depicted here indicates several modules of genes overexpressed in the liver (left) or SCN (right), differential expression analysis alone does not reveal known clock genes. Circadian genes with a known phenotypic manifestation in mouse that were expressed in the SCN are shown in the insert, failing to hierarchically cluster into biologically meaningful communities Figure (2.2, insert).

As noted above, previous studies have investigated the specificity vs ubiquity of tissue expression among the community of clock genes. I measured the expression of genes in the SCN and 25 mouse encode tissues (Mouse ENCODE Consortium et al., 2012), showing a uniform distribution of tissues expression among the community of mammalian clock genes, whereas non-clock genes were found to exhibit a somewhat bi-modal distribution (Figure 2.3 A). I also characterized the degree of cycling of all genes within both the SCN and the liver across time, visually shown in Figure 2.3 B. Sixteen exemplar clock genes are shown, and

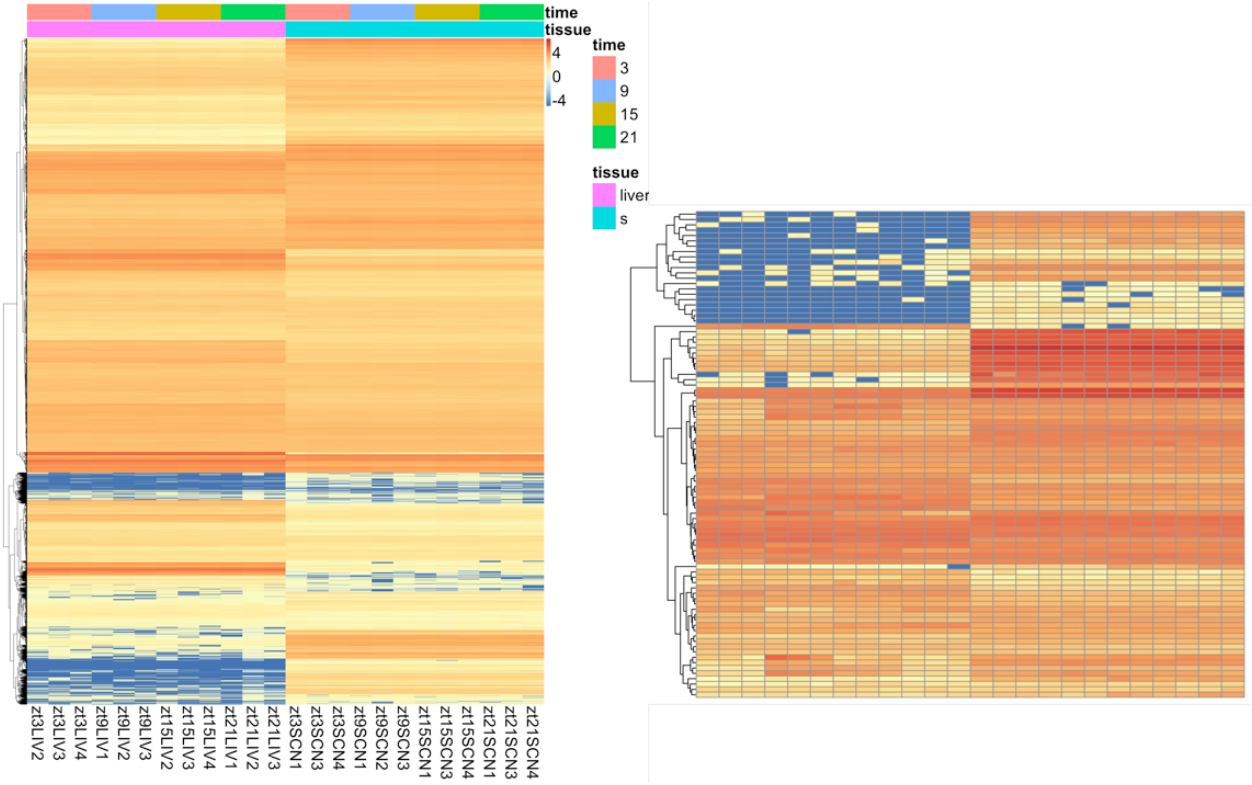



Figure 2.2: *Circadian genes cannot be identified by expression patterns alone. Hierarchical clustering was performed on the rows of a time series of RNA-Sequencing on liver (left, large heatmap) and suprachiasmatic nucleus (right, large heatmap) over four time points (Zt3, Zt9, Zt15, Zt21) with variance stabilizing transformed read counts. Genes contributing to circadian phenotypes (insert) do not cluster together in the larger expression matrix. Circadian contributing genes are more likely to be overexpressed in the SCN compared to the liver (insert).*

it should be noted that while some cycle in the SCN (red) with a parabolic or sinusoidal waveform, others do not indicate expression patterns unique to circadian genes. Several are more highly expressed in the liver than in the SCN, again indicating the complexity of categorizing clock genes.

The classification of each gene as circadian or non-circadian rely on varying features contributing in a non-linear manner towards the decision. Table 2.2 shows the estimated relative feature importance for each gene in a LASSO generalized linear model, highlighting the diversity of cycling-detection and protein-protein interaction metrics which influenced the algorithm's classification probability.

| | phase.Liver.Rain | peak.shape.Scn.Rain | JTK_adjphase.Liver | |
|---|---|---|---|---|
| Sytl4 | 0 | 0.004782532 | -0.001507748 | |
| Grp | -0.001173762 | -0.004860947 | 0.001170079 | |
| Calcr | 0 | 0.004137092 | -0.002211773 | |
| Trh | 0 | 3.860104e-03 | -1.922031e-03 | |
| Npas2 | 0 | 0.003597015 | -0.001559419 | |
| | LS_pvalue.Liver | meta2d_rAMP.Liver | LS_pvalue.Scn | |
| Sytl4 | 0.030499157 | 0.031079645 | 0.024323524 | |
| Grp | -0.021136905 | -0.030154203 | -0.015068329 | |
| Calcr | 0.021488926 | 0.032092530 | 0.026492679 | |
| Trh | 1.807519e-02 | 2.211730e-02 | 2.447960e-02 | |
| Npas2 | 0.017161504 | 0.038264636 | 0.019437223 | |
| | LS_period.Scn | LS_adjphase.Scn | meta2d_period.Scn | |
| Sytl4 | -0.002756362 | 0.001363876 | -0.005244190 | |
| Grp | 0.004348889 | -0.001168412 | 0.003868879 | |
| Calcr | -0.003853733 | 0.001234456 | -0.005247058 | |
| Trh | -2.262622e-03 | 9.459379e-04 | -4.649955e-03 | |
| Npas2 | 0 | 0.001194213 | -0.003534250 | |
| | origBigDiff | meta2d_phase.Scn | meta2d_period.Liver | JTK_amplitude.Scn |
| Sytl4 | 0.303266144 | 0.001337962 | 0 | 0 |
| Grp | -0.304644869 | 0 | 0 | 0 |
| Calcr | 0.314479838 | 0.001138545 | 0 | 0 |
| Trh | 2.905109e-01 | 0 | 0 | 2.861497e-05 |
| Npas2 | 0.270126899 | 0.001325216 | -0.002642194 | 0 |

Table 2.2: *LASSO modeled feature importance weights for selected genes highlight different features important for predicting to have circadian influence in the mouse.*

Numerical quantification of each gene in the liver and SCN, including those not expressed in the SCN and not carried further in my study, is provided in Supplemental Table 1.

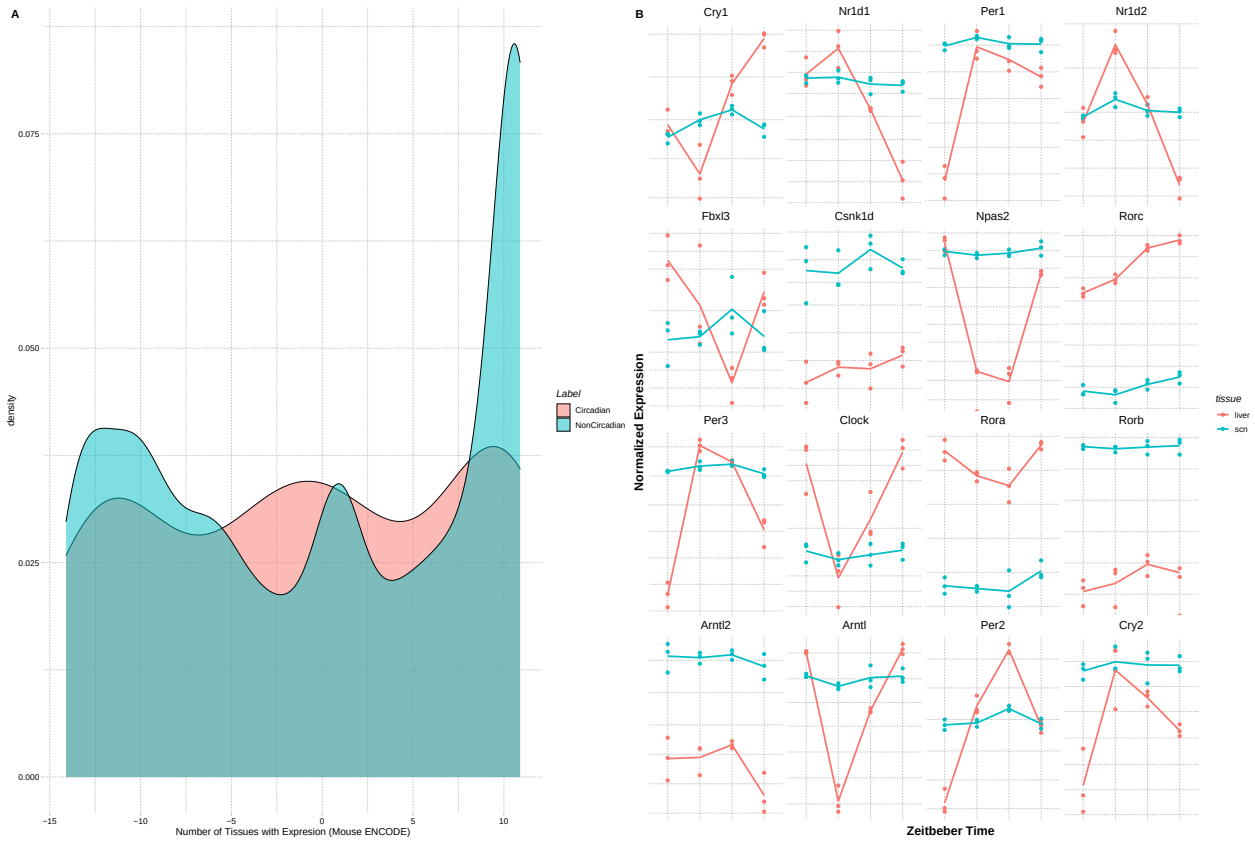



Figure 2.3: *Expression between and within tissues identify circadian characteristics. Evidence of expression was measured (> 1 TPM) in 25 mouse ENCODE tissues and the SCN (in house), depicted in a density plot 3a. Non-clock genes (blue) exhibit a bi-modal expression, most genes expressed specifically or ubiquitously. Clock genes annotated in mouse (pink) exhibit a near uniform distribution. Within liver and SCN tissues, core clock genes are depicted showing cyclic waves across four time points measured (3b). Among the core clock genes, cycling appears more sinusoidal in the SCN (blue) and u-shaped in the liver (red). 8852 genes were significantly cyclic in the SCN, 6596 in the liver (FDR p-value < 0.05, RAIN algorithm).*

Each potentially novel gene is given not only a class, but a ranking based on the emitted probabilities from the RUSBoost algorithm. In Supplemental Table 2, quantified measurements of cycling potential based on three algorithms, non-normalized computed diffusion scores from predicted protein-protein interaction kernels, as well as measurements from other features and external resources are listed for both known circadian actors and all 246 suggested novel circadian genes.

### 2.3.3 Predicted novel circadian genes are enriched for shared biological functions

I accessed both known and unknown circadian genes by performing gene set enrichment analyses using three sets of ontology/gene annotations. I enriched target genes for their annotations in the Biological Process domain of the Gene Ontology to survey shared cascades of molecular functions which contribute to biological effects (Gene Ontology Consortium, 2015). Molecular pathways of those effects were investigated by enriching among pathways in the Reactome database (Fabregat et al., 2016). Lastly, my target for classification was genes categorized as contributing to abnormal circadian rhythms in MP. To survey biological traits observable to human observers, I enriched known and predicted genes in that ontology.

Known circadian genes are highly enriched for not only circadian processes and phenotypes (Figure 2.4 red) but also share with genes exhibiting no circadian mouse phenotype (blue) enrichment for GPCR ligand binding, signaling receptor events, and several behavior related phenotypes and biological processes. While no known circadian mouse phenotypes have been associated to the novel 246 genes enriched in the MGI database [please check that I did not alter the intended meaning], circadian processes associations are recorded in the GO biological process domain, as well as circadian pathways relations are identified in the Reactome database among supposedly non-circadian genes. Both set of genes are enriched

for abnormal sleep patterns, traits which are intricately related to circadian function.

The 246 putative novel circadian genes were segregated into two groups. The circadian-related ignorome can be classes as a category of totally novel circadian genes, i.e. there may be genes with biologically relevant circadian roles which have not yet been envisioned by chronobiologists. Previously studied genes in my case include genes with identified contributions to circardian biology, but who have no such recorded associations in the MGI gene/phenotype annotation database. GO, MP, and Reactome enrichments as in Figure 2.4 above can be seen in Figure 2.5 below, with previously annotated genes in panels A,B,C and the circadian ignorome in D,E,F. Interestingly, both gene sets share signaling pathways and GPCR signaling events which are also shared with well prototypically characterized circadian genes. Among the ignorome, which include genes Trh, Sytl4, and Calcr surveyed above, abnormal synaptic transmission and nervous system physiology are highly enriched indicating strong CNS function (E). Additionally, feeding behavior is highly enriched in the GO.

### 2.3.4 ML highlights the role of Trh expressing neurons in the SCN

Thyrotropin releasing hormone (Trh) is a proprotein regulating biosynthesis of TSH, thyroid stimulating hormone. As such, Trh deficiencies have been associated with hypothyroidism in humans (Gary et al., 2003a), but there is no evidence that plays a circadian role in humans or mammals. My Rusboost based approach predicted potential Trh associations with circadian phenotypes at over 90% probability. Trh is highly expressed in the SCN (233.39 +/- 7.99 TPM) and is predicted to interact with several known clock proteins, including known circadian genes Avp, Cck, Oxt, and Grp. In humans, it has been suggested that Trh is regulated by the circadian entrainment in the SCN (Gary et al., 2003b). When comparing the expression of all my genes across 27 tissues, I found Trh to be highly enriched

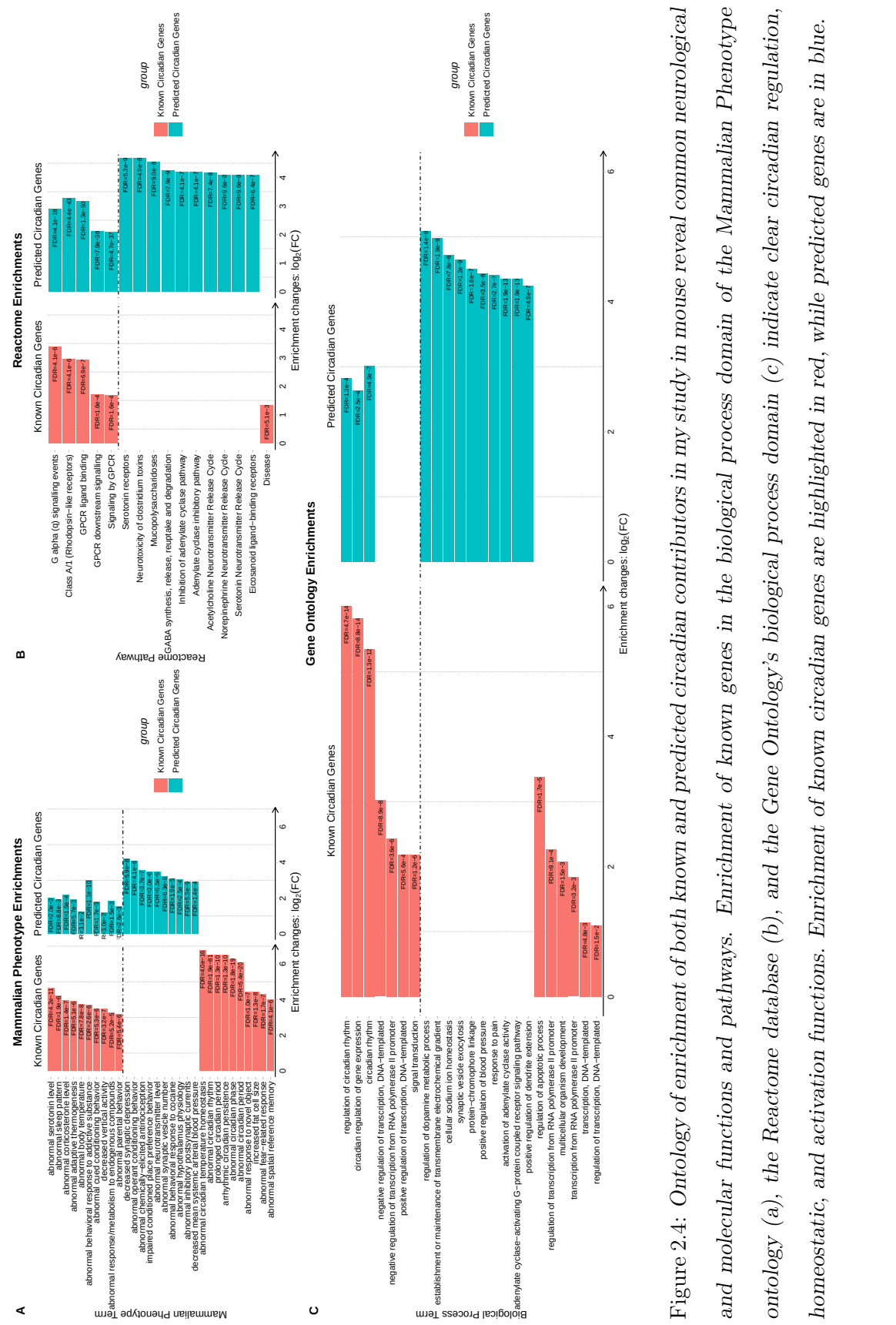



Figure 2.4: *Ontology of enrichment of both known and predicted circadian contributors in my study in mouse reveal common neurological and molecular functions and pathways. Enrichment of known genes in the biological process domain of the Mammalian Phenotype ontology (a), the Reactome database (b), and the Gene Ontology's biological process domain (c) indicate clear circadian regulation, homeostatic, and activation functions. Enrichment of known circadian genes are highlighted in red, while predicted genes are in blue.*

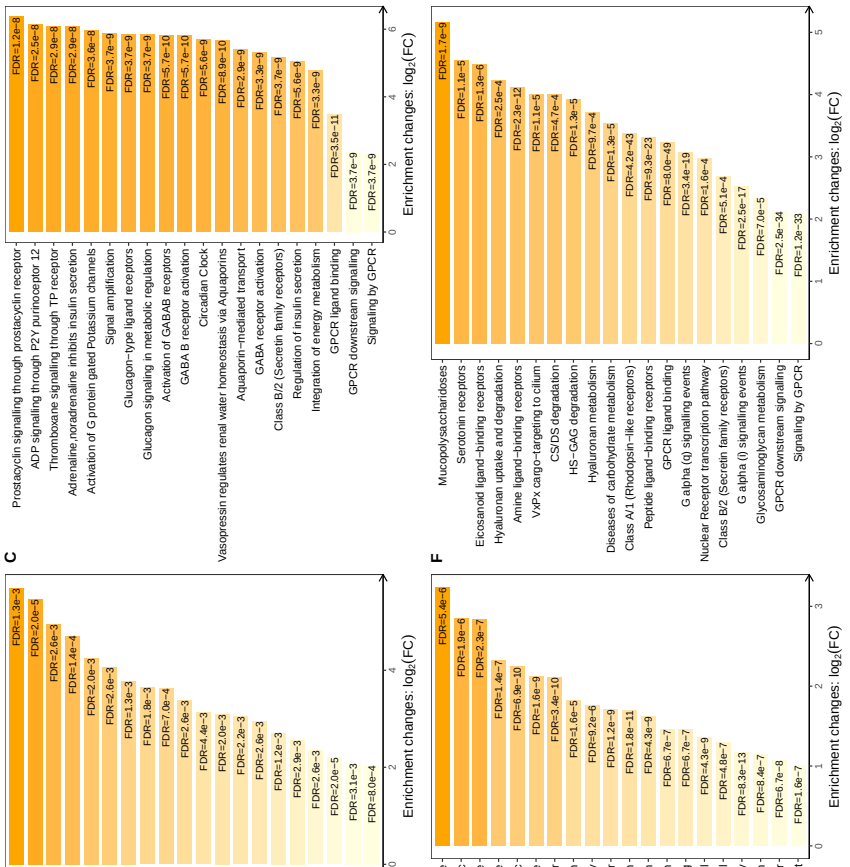



Figure 2.5: *Ontology of enrichment of predicted circadian genes previously annotated with circadian functions (top: a,b,c) outside of mouse compared to predicted genes with no circadian GO annotations reveals shared signalling enrichment (bottom: d,e,f). Annotated ontologies are (left to right) Gene Ontology – Biological Process, Mammalian Phenotype, Reactome. Both known and unknown circadian genes are involved in GPCR signaling pathways (c,f) and synaptic transmission processes (a,d). Abnormal synaptic transmission and neurotransmitter levels are evident in phenotypes of uncharacterized circadian genes (e), while abnormal hormone levels are readily evident in known genes (b).*

in the SCN, with a specificity value $\tau$ value of 0.96 (see methods). In a recent transcriptomics meta-analysis, Trh is highly enriched in the SCN compared to the whole brain (Brown et al., 2017c), and has been shown to cycle rhythmically (Pembroke et al., 2015) in the SCN. In a previous study of a Zfhx3 mutant, Trh was shown to interact with several known clock proteins derived from ChIP-Seq data, including Prokr2, Vip, and Vipr2 (Parsons et al., 2015).

Feature importance, based on RUSBoost algorithm ranking of several known and predicted circadian genes, can be visualized in Figure 2.6. Of note, Trh interacts with several genes and is shown with a high ranking (deep red) in circadian activity while being highly enriched in the SCN compared to the liver.

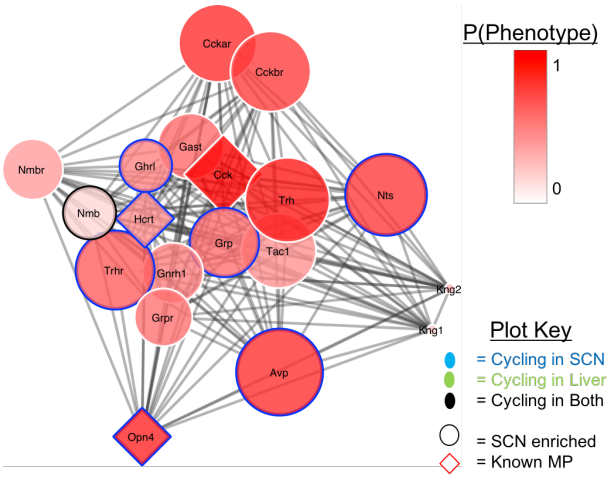



Figure 2.6: *Trh is predicted to contribute to abnormal circadian rhythms. Integration of clock features helps identify known and unknown circadian genes. Genes are situated in a predicted protein-protein interaction graph using an organic layout. Genes with dark red colors have a higher probability of contributing to an abnormal circadian phenotype. Larger nodes are more highly expressed in the SCN compared to the liver (log-fold change). All circular nodes depict genes with no previous mouse abnormal circadian phenotype, while diamonds represent known ground truth for classification. Border circle colors indicate cycling in SCN (blue), liver (green), or both (black) tissues.*

### 2.3.5 ML predicts Synaptotagmin-like protein 4 contributions to circadian rhythms

Several involved in protein exocytosis have recently been shown to be involved in circadian rhythms (Brown et al., 2017c; Yi et al., 2002). Sylt4 was highly expressed in the SCN (mean 29.75, +/- 0.61 TPM) and interacts with genes in a SNARE complex previously shown to be involved in circadian regulation, Vamp2 and Snap25 (Oliver et al., 2012b). Sytl4 is not only enriched in the SCN when compared expression in the liver, but when compared to whole brain tissue as well (Brown et al., 2017c). Brown and colleagues experimentally validated the expressio nof Sytl4 revealed in their meta-analysis. Some experimental studies found the Circadb database report that Sytl4 cycles in the SCN, but my four time-point study failed to detect cycling (Pizarro et al., 2013).

### 2.3.6 Calcitonin receptors may contribute to circadian phenotypes

Calcitonin receptor (Calcr) is a G protein-coupled receptor, activating Gs and Gq alpha subunits when bound to calcitonin (Goda et al., 2018a). While highly expressed in osteoclasts to maintain local calcium homeostasis, it is also expressed in several brain regions. Recent research noted Calcr expression in the mammalian suprachiasmatic nucleus, and the expression Drosophila Calcr ortholog (Goda et al., 2018b). In my study, features relevant to Calcr include high expression in the SCN (36.5 +/- 0.51 TPM), over expression in the SCN compared to the liver, and predicted protein-protein interactions with several well characterized clock proteins including Vip, Vipr2, and Avp. My algorithm predicts Calcr to influence abnormal circadian phenotypes in mouse with $> 90\%$ probability. In addition to my algorithm's high ranking results, indicating cause for further investigation, calculations of tissue specificity reveal a $\tau$ of 1 linked to the SCN, indicating that across 27 mouse tissues measured

its activity is restrained to the SCN. Calcr was demonstrated to be cycling in the SCN in several studies (Pizarro et al., 2013; Pembroke et al., 2015), indicative of possible circadian function.

### 2.3.7 Recovery of core clock genes lacking mammalian phenotypes

Although circadian processes are highly evolutionarily conserved, several discrepancies exist among gene annotations in human and mousse when measuring circadian activity at the phenotype level. I have predicted several known circadian genes to contribute to mouse abnormal circadian phenotypes. Among them are Npas2, Nocturnin, Rorb, Rorc, and Grp.

### 2.3.8 Npas2: A recovered Clock gene paralog

Neuronal PAS domain protein 2 (Npas2) is a transcription factor in humans and mice which acts as a Clock paralog (DeBruyne, Weaver, and Reppert, 2007), rescuing circadian clock activity in Clock gene knockout mice. However, at least withing the MGI database, no known circadian phenotype associations have been reported. My algorithm suggests Npas2 contributes to an abnormal circadian phenotype with a high probability (0.98), being moderately expressed in the SCN (6.57 +/- 0.079 TPM), interacting with the core clock machinery, and cycling in liver tissue. Npas2 was not cycling in my dataset, nor in 3/4 SCN datasets in CircaDB. Npas2 interacts with as a Clock paralog with core clock genes in the transcriptional-translational feedback loop, including Bmal1, Ror proteins, and the Per/Cry complex.

### 2.3.9 Grp knockouts produce a circadian phenotype

Gastrin releasing peptide is well known to stimulate the release of Gastrin in the stomach, and is also well characterized in circadian pathways. McArthur and colleagues showed that Grp shifts the phase of SCN nuclei as part of the circadian entrainment pathway (McArthur et al., 2000). My results depicted Grp cycling in the liver (FDR = 0.0048) and being deferentially overexpressed in the SCN. Grp was expressed with a mean TPM of 6.01 +/- 0.26 in the SCN, and had enriched expression in the SCN compared to the whole brain in both Brown et al's meta-analysis and Pembroke et al's time series study, who additionally found evidence for cycling in the SCN. Grp is known to interact with many clock genes, including known circadian actors, including Oct, Prokr2, Cck, Nmu, and App. To compare the degree to which Grp's role in circadian biology is evident on the phenotypic level, experimental validation was performed. Circadian phenotyping comparing female C57Bl/6N wildtype with $Grp^{-/-}$ mice indicate a reduced circadian amplitude, period, and period onset in constant darkness (DD), and also a decreased period in in a 24 hour light-dark cycle (LD), Figure 2.7. Representative actograms are shown in Figure 2.8, with five days of wheel running in an LD cycle, and the remainding in DD suggesting a loss of entrainment between WT and $Grp^{-/-}$ models.

## 2.4 Discussion

Features for circadian phenotype prediction were generated from two approaches: gene expression studies and mining potential protein-protein interactions. It is worth noting that several canonical clock genes do not share all features measured. For instance, the Clock transcript does not cycle in the SCN, and several clock genes exhibit decreased expression in the SCN compared to their expression levels in the liver. Using the RUSBoost algorithm, which is particularly suited to classifying skewed data, allowed us to capture non-linear

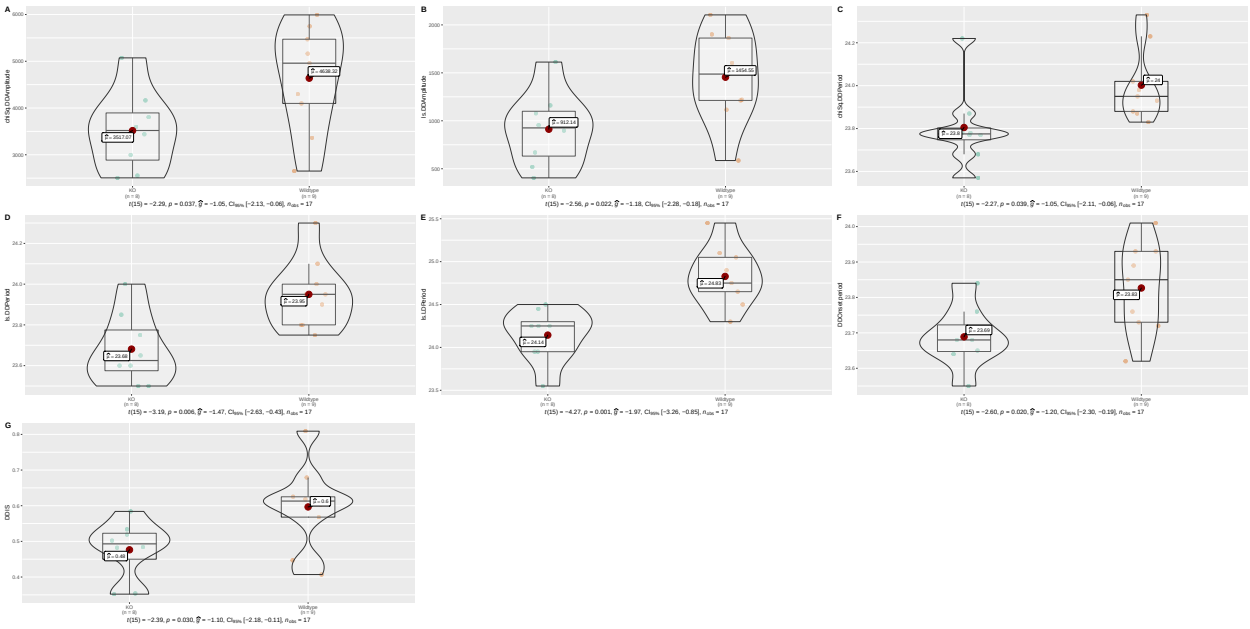
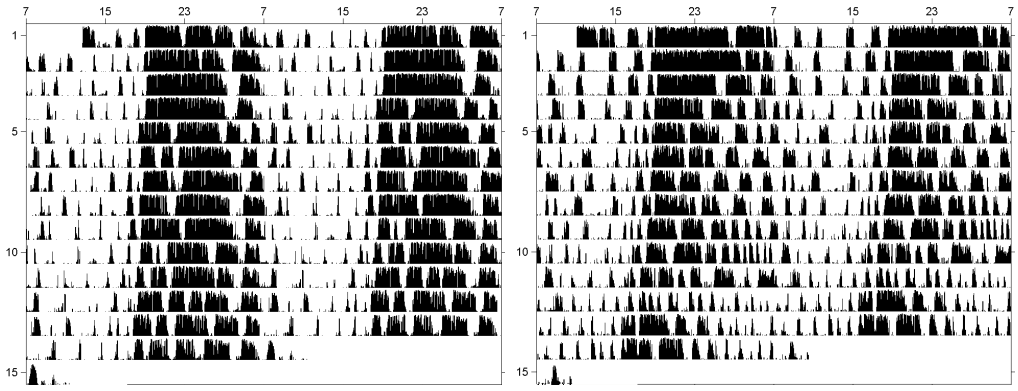


Figure 2.7: *Grp -/- mutants exhibit a statistically significantly set of decreased circadian parameters as defined by periodogram measurements when compared to wild type. This includes amplitude in constant darkness (DD) as measured by a chi-squared (A) and Lomb-Scargle periodogram (B), period in as similarly measured (C,D), decreased period in a 24 hour light-dark cycle measured by Lomb Scargle (E), DD period onset (F), and DD interdaily variability.*




(a) (b)

Figure 2.8: *Under an initial 5 day reference period, wild type (A) and Grp knockout (B) were kept in a 24 hour light-dark (LD) cycle. The following days are under constant darkness (DD). Lights are switched on at 7am and off at 7pm as indicated in the top x-axis label. Y-axis labels indicate day. Lights on is zt0, off zt12. Black histograms measure activity. Wild type activity appears more stable under DD (A) compared with the Grp mutant in identical conditions (B).*

relationships between features generated.

### 2.4.1 Exploiting expression knowledge within and among tissues

The cycling metric was generated by using the RAIN algorithm, and I also employed the widely used cycling detection algorithm JTKCycle (Hughes, Hogenesch, and Kornacker, 2010b). Both JTKCycle and RAIN use non-parametric slope tests to account for statistically unexpected forms. Unlike RAIN, asymmetric expression shapes are assumed to be aperiodic, enabling capture of non-sinusoid rhythms by JTKCycle. My experimental setup was limited to four time points in 24 hours. To compensate for the lack of two periods worth of coverage, I imputed time points for a second day. This was done to increase the power of RAIN to detect rhythms which are evident when observed by eye. Additional cycling experiments were run with JKCycle and the Lomb-Scargle periodogram method (Scargle, 1982; Lomb,

1976). Although imputing another 24 hours of expression data carriers much extrapolation, it was deemed reasonable for detecting 24-hour periodicity. Cycling algorithms perform best with repeated time points within a series as evidenced by comparing cycle detection in public databases (Pizarro et al., 2013), and doubling my time points increased my ability to detect cycling patterns. This does have a major limitation, however. By doubling the time points, I may have also increased the false positive rate of detecting significantly cyclic patterns in the data. This would be particularly likely when the algorithms attempt to compute changes within windows involving both zt21 and the "next day" zt3, where no actual measurement occured. If a high expression of a gene in zt21 is followed by a low expression in the imputed zt3, a false positive oscillation will be observed. This increased false positive rate must be considered when reporting these results. For this reason, I focus on the ability of real or imputed cyclical measurements to predict circadian activity repeated train/test splits, and not in reporting any cycling measurements as scientific findings themselves. To produce more valid results, data with more than four time-points and over 24 hours would need to be used. An observation easily seen when querying public cycling databases is that measures of cycling are not always in agreement, as methods of detecting period, phase, and amplitude all vary. This is evident in my features table, Supplementary Table 2, providing further evidence of the care with which these data should be interpreted. Tissue ubiquity was used previously as a feature in previous circadian gene predictions (Anafi et al., 2014b). To accomplish this, I re-normalized my DESeq2-normalized read counts to Transcripts per Million, giving each gene a percentage of one experiment's total expression. This facilitated comparison of transcripts between labs and tissues with different library sizes. Some discrepancy arose between my read calling and that of mouse ENCODE due to Ensembl transcript versions used. Therefore, some genes included in my dataset had missing tissue ubiquity measurements. While I did impute missing values with median values, this may not reflect reality. This feature did not heavily influence my classifier results, as tissue specificity was used as a post-algorithm augmentation of my results rather than as feature input. The calculated $\tau$ value of SCN-

specific genes was a useful measure of tissue specificity across a wide range of organs and brain regions.

This chapter used, as primary input, data from the liver and the SCN. As mentioned in the introduction, there is an interplay between the regulation of the central and peripheral clocks. A widely studied endogenous zeitgeber is melatonin, which increases expresio nof Cry2, Blam1, and Per1 proteins and maintains oscillations independent of those proteins (Chen, Zhang, and Lee, 2020). Besides melatonin, oscillations within the liver may be the most well-studied. Liver clocks may be entrained non-photically, and feeding cycles are the dominant zeitgeber for these clocks (Finger, Dibner, and Kramer, 2020). Gastrointestinal hormones, including VIP, gastrin, ghrelin, and cholecystokinin are produced with a circadian or rhythmic cycle, and feedback mechanisms to the SCN are currently under heavy investigation especially as this affects adiose homeostasis (Landgraf, Neumann, and Oster, 2017; Landgraf et al., 2015; Crosby et al., 2019; Guo et al., 2005; Buijs et al., 2001). While not confined to the liver, including a peripheral circadian tissue in this work allows for circadian signals which would be lost by looking only at the SCN to be gained. Circadian biology depends on the coupling of oscillators between cells to coordinate the response of external zeitgebers, including those independent of light (Finger, Dibner, and Kramer, 2020). The non-linear interactions produced in this chapter's model by decision trees may indicate multiple levels of cross-talk between the gene expression and protein-protein interaction data, which reflects the interaction of biological clocks within brain organs (SCN, hypothalamus), within distant tissues (liver), and between organs on a system wide level.

### 2.4.2 Walking across connected proteins

As expression alone could not segregate data based on circadian potential, it was hypothesized that combining the expression study with protein-protein interaction data would yield

a better predictor of circadian phenotype influence in mouse. Previous work included measurements of interactions with the CLOCK protein (Zhao et al., 2007). I extended this analysis by incorporating predicted protein-protein interactions of all genes expressed in the SCN from my RNA-Seq analysis above. The kernel method chosen to produce a circadian-related score $S$, with an importance graphically depicted in Figure 2.4 A. The method is essentially a short random walk around proteins previously annotated with circadian MP terms. The score reflected two elements of graph architecture: the degree of nodes (how many proteins a protein was interacting with) and path distance (how may steps it would take to walk between any two given proteins). A similar feature was produced using only deferentially over-expressed genes in the SCN, chosen to reflect the importance of the central pacemaker clock in producing circadian phenotypes in mouse. Proteins near multiple circadian nodes receive a higher score than more isolated nodes, quantifying their relative likelihood of contributing to circadian phenotypes from guilt by association. When performing cross-validation, I chose to mask the evidence of circadian nodes, thus predicting how much my score $S$ relies on previous MP annotations. While the $S$ score was a highly important feature in my analysis, my feature importance metrics highlight the role of traditional metrics of circadian analysis, such as phase and peak amplitude. Using a robust random forest approach, I were able to segregate my data with near 100% recall/sensitivity. However, I aimed to predict potential novel circadian genes, and the performance of the RUSBoost machine provided a balance between a high-positive rate achieved by a random forest, and a low recall achieved by a naive-Bayes approach. My annotations from the Mammalian Phenotype ontology were obtained in 2016, and I inspected records from May 2020 for any of my novel predictions. Of the genes in my analysis (those highly expressed in the SCN), seven which I labeled as 'non circadian' have now been annotated as influencing abnormal circadian rhythms in mice. Of these seven, Pth2r was predicted by my approach to be a circadian gene.

### 2.4.3 Prevously studied genes reveal annotation disparities

The protein Neuronal PAS Domain 2, NPAS2, has been shown to have paralogous functions to those of the CLOCK protein. Based on homologous structure, the STRING database predicted NPAS2 to interact with partners of CLOCK. Previous Npas2-knockout experiments have failed to produce circadian phenotypes, though one experiment did observe the "abnormal sleep pattern" phenotype on a mixed-strain background (Garcia et al., 2000). While they observed disrupted sleep bouts, no abnormal circadian phenotypes were reported. A following study reported GO biological process descriptions of circadian rhythm processes, but reported no abnormal circadian mammalian phenotypes (Dudley et al., 2003). Circadian-related activity of Npas2 has been observed in the mouse forebrain and in peripheral tissues, thus a tissue-specific role of Npas2 may explain the lack of abnormal circadian phenotype observed in mouse.

Grp, a well characterized clock protein, was likewise predicted to contribute to circadian phenotypes on perturbation. While Grp has been characterized as a contributing to circadian rhythms for well over a decade (McArthur et al., 2000), there is no evidence of this in the mouse genome database. Indeed, investigation of Grp reveals reported phenotypes of "abnormal grooming behavior," and three 'normal' behavioral/neurological phenotypes. While it is understandable that several circadian genes may be under reported in phenotype databases, or may not produce a circadian disruption observable as a measured trait on the organismal level, using a machine learning approach revealed disparities between literature and organism databases. This validation of my predictions both provides phenotypic evidence of Grp in the MGD database for other researchers to use, but creates a further link on the phenotypic level between circadian biology along the gut-brain axis. As many circadian genes are well characterized but lack mammalian phenotype annotation, finding closes an annotation gap in the widely used Mouse Genome Database (Bult et al., 2019a).

### 2.4.4 Illuminating the circadian ignorome

Several likely candidates among my 246 potential novel genes have external supporting evidence. While tissue specificity, whether measured by a standardized $Z$-score in a meta-analysis, differential expression between tissues, or a $\tau$ measurement is indicative of tissue-related function. In the SCN Calcr, Trh, and Sytl4 are all enriched for SCN-specific action indicating a possible circadian function. A meta-analysis recently experimentally validated the presence of Sytl4 in the SCN, and functional validation is needed to characterize it's potential role. Calcr, the receptor for calcitonin, was previously investigated for circadian function by researchers guided by enrichment in the SCN among orphaned G-protein coupled receptors (Doi et al., 2016). Doi and colleagues' analysis involved generation of a mouse knockout model on the 129P2/OlaHsd $\times$ C57BL/6J strain background, and compared its circadian behavior to a line backcrossed to C57BL/6J for 10 generations. As interstrain variation even among C57BL/6 mice has been shown to affect phenotype, it is reasonable to hypothesize that no abnormal circadian effect was observed due to strain differences (Simon et al., 2013). Additionally, it has been demonstrated that while Calcr is not required for coordinating locomotion-based circadian traits, it has recently been shown to regulate circadian variation in temperature when mice are active (Goda et al., 2018b). Lastly, the role of Trh in circadian rhythm processes has known in essential homeostatic processes in chronobiology (Gary et al., 2003b). While it is expressed in multiple brain regions and body tissues, Trh and Trhr both localize to the SCN in rat (Manaker et al., 1985). As it is also located in retinal ganglia, the hormone may play a role in circadian entrainment via light (Lexow, 1996). In hampsters, injection of TRH into the SCN shifted wheel-running phase (Gary et al., 1996). This again provides evidence of a circadian function, but does not guarantee that perturbation or knockout of a gene, and therefore disruption of endogenous TRH protein, would produce observable circadian phenotypic traits. Just as exogenous Leptin did not produce the expected effect in humans that was observed in mice, further experi-

mentation with Trh will be needed to verify an effect on circadian phenotypes in mammals (El-Haschimi and Lehnert, 2003).

## 2.5 Conclusions

Ultimately, this study demonstrated both the utility of using machine learning to predict abnormal circadian phenotypes, and the potential pitfalls of using observable single traits (phenotypes) as ground truth in biological exploration. Machine learning has been applied in various forms to a range of biological problems, including in recent large scale competitions to access prediction of both biological processes (GO) and potential human phenotypic traits (HPO) in unannotated proteins (Radivojac et al., 2013). Rather than relying simply on amino or nucleic acid sequences, I have demonstrated the utility of combing multiple forms of biological information (protein interactions and multiple measures derived from gene expression) to make biologically relevant predictors. As several of my predicted genes have known circadian roles, I expect that high throughput phenotyping screens such as the International Mouse Phenotyping Consortium may rectify the uncovered annotation disparity between GO and MP databases (Brown and Moore, 2012b). Ultimately I expect several genes ranked high on my list of potential novel genes to provide increased insight into chronobiology. Features included in my predictive analysis cannot discriminate between genes which affect core clock machinery, and thus play a role in circadian periodicity, and genes which are effected downstream of core clock genes and produce peripheral circadian phenotypes, such as abnormal circadian temperature homeostasis. Regardless, as circadian biology impacts every tissue in both human and murine anatomy, I expect my predictions to be useful in prioritizing candidate genes for investigation by experimental chronobiologists.

## 2.6 Chapter Summary

In this chapter, I applied several statistical and machine learning methods to characterize and predict genes which are involved in producing abornomal circadian rhythm phenotypes in mouse. The results demonstrate that leveraging information from orthologs, predicted and verified protein-protein interaction networks, and gene expression across tissues in mouse can improve the ability of classifying circadian genes based on study of the transcriptome. It highlights the importance of looking outside the SCN for rhythmicity, as both liver and SCN were useful features. Thanks to experimental colleagues, one well studied circadian gene, Grp, has been validated to produce abnormal circadian phenotypes in mouse. Others are being used in grant applications and will be analyzed in additional studies, aiming to further validate findings observed in this chapter. Experimentalist colleagues will be able to use both the final gene candidates and the detailed feature annotation, from tissue specific expression to cycling analysis, to guide broad gene characterization experiments. Finally, the central role that protein interaction networks played in this chapter should not be understated. Future applications of this integrative approach will leverage tissue-specific predicted protein interaction networks along with the extensive amount of tissue-specific circadian cycling gene studies available, addressing links between peripheral clocks other than liver and the SCN (Pizarro et al., 2013).

As discussed in this chapter, circadian biology impacts several areas of health from metabolism to neurology. This chapter has used tissue-specific transcriptomic data from mouse to model the likelihood of genetic disruptions influencing circadian biology as revealed by some phenotypic manifestation. A commonly measured circadian phenotype in humans is chronotype, a measure of circadian period which can be dichotomized into "eveningness" and "morningness". In the following chapter, I investigate the relationship between a genetic exposure to chronotype and mental health and social outcomes in humans, using genome-

wide association studies and a method of modeling causation, Mendelian Randomization.

# Chapter Three

# Evaluating the causal relationship between circadian chronotype and psychosocial behavioral traits

## 3.1 Background and Chapter Overview

As discussed in Chapter 1, this thesis addresses questions in neurobehavioral genetics using a combination of approaches, ranging from predictive modeling with transcriptomics to statistical inference using genomics. In this chapter, I use Mendelian Randomization (MR) to the causative relationship between chronotype (a measure of circadian influence) and several neurobehavioral traits. First, I give a brief overview of Genome Wide Association Studies (GWAS), which form the basis of MR analyses. Then I describe the UK Biobank cohort used in this chapter. I then describe the statistical assumptions of MR and the methods used, before presenting results. For a review of MR itself, see (Davey Smith and Ebrahim, 2003b). This chapter contains three analyses relating chronotype to traits. First, I access the influence of chronotype on mental health, hypothesizing that 'morning' and 'evening' chronotyped

individuals will experience different mental health conditions partly due to their chronotype alone. Secondly, in a related study, I investigated measures of social support in the UKBB. Circadian entrainment can be influenced by social cues and patterns, thus chronobiology may influence social support related phenotypes. Lastly, I study the influence of chronotype on keratometry (ophthalmetry) index measurements. These measurements attempt to capture the degree of corneal power and the aim of this work was to employ keratometry measures to assess whether circadian rhythm traits can influence eye morphology, given the fact that light that reaches the eye forms the primary zeitgeber in humans.

### 3.1.1 Genome Wide Association Studies

Prior to the completion of the human genome project (Lander et al., 2001), associations between genetic variants and phenotypic outcomes were obtained through twin or family based linkage disequilibrium studies (Merikangas et al., 1998). While successful in finding causes of high powered monogenic disorders within pedigrees, there was little power to detect associations between genetic variants and complex diseases. Early literature suggested that GWAS would bring increased power to detect small effects common in complex disorders (Risch and Merikangas, 1996). Recently, an omnigenic model of complex disease manifestation has been proposed, in which small variations in SNPs together act to explain the heritability of a complex trait, a notion which would be lost when investigating Mendelian disease (Boyle, Li, and Pritchard, 2017; Wray et al., 2018a). To investigate this omnigenic model, GWAS are retrospective observational studies which test associations between several hundred thousand or often over a million SNPs, chosen in an unbiased manner across the genome, to a given trait (Bush and Moore, 2012). Within the context of a disease, studies often have a case/control design wherein the alleic frequency, or dose, is measured in patients with a condition and matched control subjects. In a naive approach assuming a biallelic SNP and

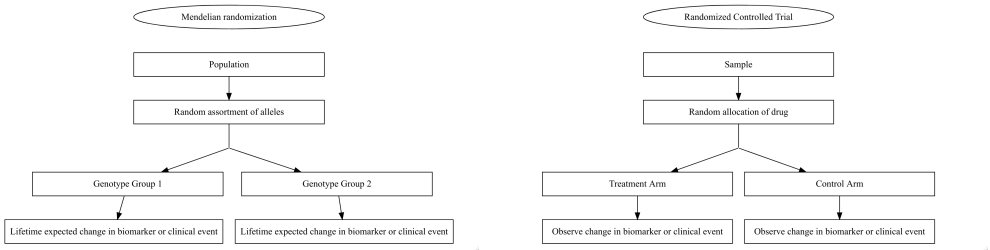
two populations, a $\chi^2$ test can be conducted to obtain a test statistic, and then a p-value of the association based on the extremeness of the statistic under a $\chi$ distribution. From this "2x2" table analysis, odds ratios can be constructed to measure the effect of a SNP on a patient, or case, population characterised by a particular disease. If the prevalence of the disease is known, then a useful relative risk statistic can be obtained (Clarke et al., 2011). If a trait of interest is continuous and not a disease state itself, then linear regression is used. Likewise, instead of a $\chi^2$ test, logistic regression can be used to test for associations between alleleic dose and conditions. In practice covariates are used in the GWAS which is modeled. Typical covariates include any factors which may denote difference between populations not related to the disease itself, whether mediating or moderating variables. These typically include the loadings from a principle component analysis (PCA) which are used to account for the genetic heterogeneity between large population groups, often accounting for variations in ethnicity. Other common covariates include the technology or centre used when performing sequencing or assays, age, body mass index, and sex. GWAS can be facilitated by the availability of genotype and phenotype data from large consortia. In recent years, the use of genomic biobanks (Bycroft et al., 2018; Bourgeois et al., 2017; Gaziano et al., 2016), which combine deep phenotyping with genomic assays and long-term followup in a prospective study design, have facilitated both population level and personalized epidemiological studies.

### 3.1.2 Mendelian Randomization

Epidemiology is the population-level study of the origin (etiology), spread (distribution), and prevention of disease. Traditionally, most levels of inference in epidemiological studies are associative - they can correlate hazardous exposures such as disease or lifestyle factors with health outcomes. Correlation analyses are improved when potential moderating variables

are accounted for. In a hypothetical example, lung cancer rates may be correlated with the use of matches for light cigarettes, but it would be a mistake to say that matches themselves are a primary driver of the disease and not smoking. To model the directionality of an effect, causal models based on instrumental variables are used. Originally derived from economics, instrumental variables (IVs) are variables which strongly predict an exposure but have no association with the outcome studied that in turn is not mediated by the exposure (Kang et al., 2014). Since the advent of widely available GWAS, Mendelian randomization has used genetic variants as IVs. The fundamental assumption is that, because of independent assortment (Mendel's second law), genetic variants should be randomly distributed within a population. If we have deep phenotyping or outcome data, and those data co-occur or co-vary with the presence or absence of an allelic IV, then this IV can be used as a population level randomizer similar to what is seen in a randomized controlled trial (Davey Smith and Ebrahim, 2003b). This mimics the experimental design of a randomized controlled trial, as seen in 3.1a. Here, a study population is randomized into treatment and control groups, which are then given an intervention or treatment (treatment arm) or a placebo or standard of care (control arm). The patients in each group are followed up to investigate the efficacy of an intervention. It is assumed that, over a population matched for age, sex, and other potential confounding variables that other mediators or moderators will be evenly distributed between groups and thus not impact the trial design. In an MR experiment, the design is similar 3.1b. A sample of a population is split, by the random assortment of alleles at birth, into two cohorts with the minor and major allele (assuming a biallelic SNP). Observed effects are treated as lifetime exposures to the outcome of interest, assuming the resiliency of the genetic code to mutations over a normal sample of a population.

By using genetic variants as the basis for causative modeling, MR can allow for stronger causal inference than genetic correlation experiments and model counterfactuals during experimental design. On a level of inferential trust between an observational study




(a) (b)

Figure 3.1: *Study designs of Randomised Controlled Trials and Mendelian randomization. Randomized controlled trails (a) begin with a sample population, randomly split into treatment and control study arms. In each arm, an intervention or therapeutic agent (treatment arm) or a placebo (control arm) is given. Then patients in each arm are tested for changes induced by the drug or placebo. In a Mendelian randomization experiment (b), a population-based sample is included, and the random assortment of alleles in instrumental variables is used to stratify individuals into different genotype groups for each SNP. Observed effects of exposure (SNP) on the outcome measured are considered lifetime exposure effects since the genetic code is resilient.*

in model organisms, an observational study in humans, controlled experimental studies in model organisms, and randomized controlled trials in humans MR can be thought of as stronger than observational evidence and suggestive of experimental effect. In an MR "trial", nature has randomized the population into cases and controls based on the presence or absence of a specific allele, and a treatment affect can be thought of as a proxy for lifetime exposure, since the randomization into treatment and control groups occurred at conception. Early applications of MR modelled exposure to disease and well-known mechanism, such as the influence of C-reactive protein and inflammation, and have been expanded to cover both treatable and non-treatable exposures, modeling variants from either a mechanistic perspective or those found through GWAS (Burgess and Thompson, 2017).

**Mendelian Triangulation and Assumptions**

MR employs IV analysis for causative inference. Any causative modeling study can be graphically displayed using directed acyclic graphs (DAGs) (Pearl, 1995). DAGs are directed, meaning information between two nodes of a graph must flow in one direction from node A to B, but not both. They are acyclic, meaning that on a path form A to B to C, there cannot be a path from C to A. In modeling ontologies, this assures that the transitive properties of relations are maintained and that the logic underlying deductive reasoning will not break. In causal inference and experimental design, this helps ensure that an outcome will not influence an exposure, or that a confounding variable not provide an additional path between an IV and an outcome. There are three fundamental assumptions of causative inference in the MR context which must be satisfied for an IV analysis to be valid (Burgess et al., 2020):

1. The IV (allele) must not associate directly with an exposure
2. The IV must not associate with any potential confiding variables, which then associate with the outcome
3. The IV must not associate directly with the outcome, except as mediated by the exposure.

This is graphically depicted in Figure 3.2. When drawing DAGs for causal inference, a solid line indicates a path between two variables. Dashed lines in this figure indicate paths which, if followed, violate the assumptions above. The genetic variant is the allele used as an IV to segregate the population. The IV is selected by its association with the exposure, either through biochemical knowledge (such as the IL-6 locus for exposure to inflammation) or through a data driven approach (GWAS).

Instrumental variable analysis relies on strong assumptions, and occasionally unver-

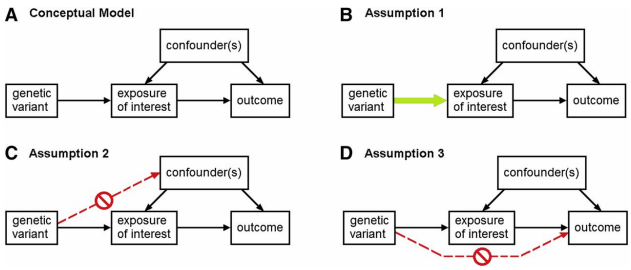



Figure 3.2: *Mendelian randomization experimental design. We are primarily interested in the effect of an exposure on an outcome. Observational studies, which report the associations between an outcome and exposure, are subject to confounding and are by nature correlative only. By using a correctly designed genetic instrumental variable that associates with the exposure but not with outcome, confounding can be avoided via 'Mendelian triangulation.' Panel A depicts the conceptional model. Each solid black line represents directional association, and each dashed curved line represents directional associations which will violate the model if present. The following assumptions must be satisfied:*

*i. The genotype must associate directly with the exposure, panel B. ii. The IV must not associate with any potential confounding variables which then directly associate with the outcome, panel C. iii. The IV must not associate directly with the outcome except where mediated by the exposure, panel D.*

*Originally published in (Sekula et al., 2016)*

ifiable conditions (Labrecque and Swanson, 2018). Subject knowledge is the most reliable method for concluding assumptions are valid, especially when choosing exposures and outcomes correctly. A naieve, phenome-wide search for causation between traits is likely to include among many tests non-plausible experimental designs, leading to an artificial increase in multiple testing burden and design flaws. Adjusting for covariates, including principal component analysis to take into account population differences and combining different methods of MR analysis will assist in adding robustness to MR analysis.

To calculate the causal effect for one SNP of interest between two independent samples, the $\beta$ estimate from linear or logistic regression of the exposure on the allele and outcome on the allele of interest are obtained from GWAS. Then the Wald ratio is obtained, giving a measure of the effect of the exposure on the outcome:

$$\widehat{\theta_j} = \frac{\beta_{Yj}}{\beta_{Xj}} \tag{3.1}$$

where $\beta_{Yj}$ is the effect of the IV on the outcome, and $\beta_{Xj}$ is the effect of the IV on the exposure is obtained for SNP j, and $\widehat{\theta}$ is the effect size. This new effect size forms the basis of all analyses presented in this chapter. In each analysis, the summary-level statistics from GWAS are obtained, namely $\beta$ and the standard error, $se$, of the $\beta$. Rather than being considered individual level data, these are summary statistics reflect population-level estimates in each non-overlapping population used. During the GWAS, as mentioned earlier, these $\beta$s are both adjusted for the same covariates to minimize study bias (Davies, Holmes, and Smith, 2018). These are considered two-sample MR studies, and one-sample MR studies can also be performed. This involves two-stage least-squares regression. First, the exposure is regressed on the IVs and covariates. Predicted values of the exposure from the first regression are obtained, and the outcome is regressed on those predicted values using the same covariates. For two non-overlapping samples, two sample MR using a meta-analytic

method (inverse variance weighted, IVW) is asymptotically equivalent to the 2SLS method (Burgess, Butterworth, and Thompson, 2013). The IVW method is the most powerful estimate of causal effect when all genetic variants are valid IVs (Burgess, Dudbridge, and Thompson, 2016).

### 3.1.3 Previous Mendelian Randomization Studies of Chronotype

With the advent of large biobank cohorts to provide GWAS, there have been several recent MR studies looking at the causal influence between chronotype, sleep and several physiological and behavioral traits. Lind and colleagues found significant genetic correlations between oversleeping, insomnia, and undersleeping exposures with an outcome of post traumatic stress disorder (Lind et al., 2020), but when testing for causality via did not find evidence for causal effects of sleep phenotypes on post traumatic stress. Adams and Neuhausen were interested in the interplay between chronotype and free fatty acid circulation, and also between free fatty acids and type two diabetes (Adams and Neuhausen, 2019). So as to evaluate this, they conducted a two Mendelian randomization studies using two sample data, and found that morning chronotype is associated with lower total fatty acid levels (IVW $\beta$ -0.21, p = 0.02) and that elevated fatty acid levels are associated with a decrease in diabetes, granting a protective effect (IVW $\beta$ -0.23, p = 0.01). They then extended their analysis to include subtypes of free fatty acids and their conclusions held, indicating that a morning chronotype is associated with lower mono-unsaturated fatty acid intake. Richmond and colleagues sought to model sleep traits and risk of breast cancer using Mendelian randomization methods, using chronotype, sleep duration, and insomnia GWAS for instrumental variable selection (Richmond et al., 2019). They modeled the UK Biobank in a one-sample fashion, using a two-stage least squares regression approach instead of splitting the cohort into two non-overlapping samples, and showed a morning chronotype to be protective against breast

cancer (Odds Ratio 0.85). Two sample modeling with an independent cohort supported these findings, showing morning chronotype (IVW OR 0.88) was protective against breast cancer, while increased sleep duration has a detrimental effect (IVW OR 1.19). Gibson investigated bi-directional causal effects between smoking and sleep duration and chronotype (Gibson et al., 2019). They found no clear evidence that smoking initiation influenced sleep behaviors directly, nor evidence for causal effects between chronotype on smoking behavior. However, they did find evidence that insomnia could lead to an increase in smoking behavior (IVW $\beta$ 1.21, p = 0.02) in an underpowered analysis. Treur modelled caffeine consumption and sleep traits, including chronotype, sleep duration, and history of insomnia (Treur et al., 2018). While the association between caffeine consumption and disturbed sleep is well known, and their analysis did show strong genetic correlations between those traits, an extensive two sample MR using IVW and MR-Egger meta-analyses failed to produce significant causal associations. On a wider scale, Lane and colleagues used MR analysis as a follow-up to their first GWAS of chronotype using the UKBB (Lane et al., 2016). They found significant associations between evening phenotype and years of education increasing and self reported schizophrenia diagnosis, and associations between a morning chronotype and a decreased body mass index (BMI).

## 3.2 Methods

A general workflow for the several MR studies in this chapter are depicted in Figure 3.3. Data can be acquired from previously performed GWAS studies with data deposited in the GWAS Catalog, MR Base, or another database. Additionally, researchers may conduct their own GWAS using data they have available. I have done this in order to ensure that I include appropriate covariates during the GWAS itself. Next, to ensure sample independence, the UKBB population is split into two groups randomly, each with half the population. Then

GWAS are conducted and held for further analysis. A discovery set of SNPs are obtained from an outside data source. Next, effect sizes and their variance from each GWAS are extracted for the discovery SNP loci. Data are then combined, and the effect sizes meta-analyzed to produce an MR study. Lastly, post-hoc analyses are performed to test for pleiotropy, directionality of effect between exposure and outcome, and the influence of specific SNPs in the model.

### 3.2.1 Data Acquisition

To obtain an unbiased set of SNPs which associate to chronotype, I downloaded all significant (<= 9.000e-6) variants from a 2016 GWAS conducted by 23andMe (Hu et al., 2016) via GWAS Catalog (MacArthur et al., 2017), accession GCST00342, a total of 77 SNPs on 20 Jan, 2020.

#### Experimental Data

The UK Biobank (UKBB) (Bycroft et al., 2018) was initially proposed by the Wellcome Trust and the Medical Research Council, with the goal of identifying risk factors for human disease (Collins, 2012). The resource focuses on middle age adults, largely of European descent. A plethora of biological samples and exams were taken from each subject, including genetic material, magnetic resonance imaging (MRI) of several body regions, health and lifestyle questionnaires, and medical history data. The UKBB aims to recruit 500,000 individuals, a large sample size reflecting the likelihood of any one individual developing a given disorder or disease (Collins, 2012). Participants signed electronic concent declarations, and the UKBB received ethics approval (Bycroft et al., 2018). For a full description of the resource and every measure available for use upon approved application, see https://biobank.ctsu.ox.ac.uk/crystal/, with further details available as reported previously

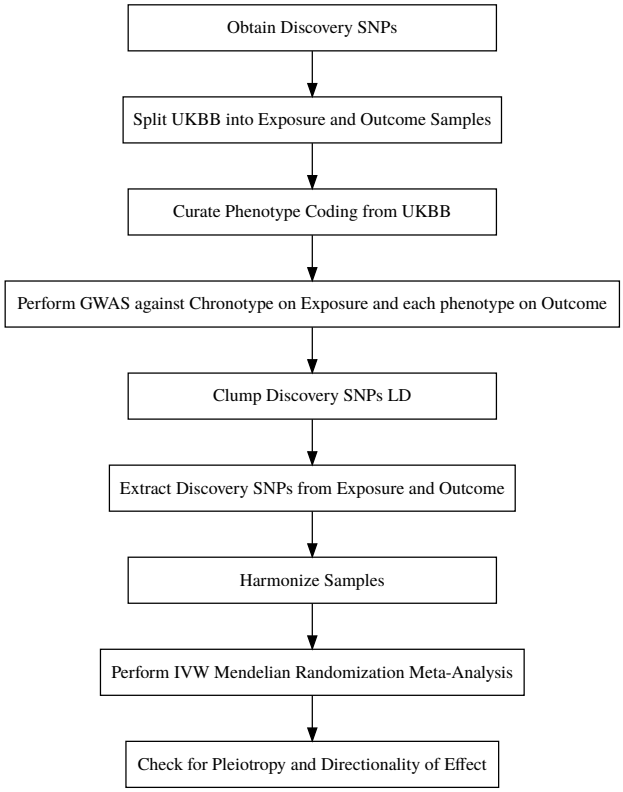



Figure 3.3: *Mendelian Randomization workflow for testing the causal influence of Evening Chronotype on psychosocial and vision traits from the UK Biobank (UKBB). Discovery SNPs associating with chronotype are obtained from an outside population. The UKBB is split into exposure, which has a GWAS performed against Chronotype, and Outcome, which has a GWAS performed against another trait. Discovery SNPs or proxies are extracted from each GWAS, using clumping to ensure SNP independence. Data are combined, and an IVW meta-analysis of effect sizes from GWAS is performed, followed by tests for pleiotropy, directionality, and sensitivity. SNPs = single nucleotide polymorphisms, LD = linkage disequilibrium, GWAS = genome wide association study, IVW = inverse-variance weighted.*

(Bycroft et al., 2018; Collins, 2012). The research in this chapter was performed under project number 31224.

UK resident participants registered with the National Health Service (NHS) were recruited between 40 and 69 years of age between 2006 and 2010. At baseline, participants donated blood and urine, and completed self assessment questionnaires relating to lifestyle, medical and family history (Allen et al., 2014b; Collins, 2007). As of July 2020, 488,264 participants have been recruited and genotyped. Participants completed self-guided questionnaires on centre computers with the aid of in-built help systems when prompted. Genetic data were extracted from blood as previously described (Bycroft et al., 2018; Collins, 2007). Deoxyribonucleic acid (DNA) was extracted at baseline, and initial genetic data from Affymetrix arrays were released in May 2015 and July 2017 (Bycroft et al., 2018). The first release included 150,000 participants, with 50,000 genotype using the Applied Biosystems UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) Axiom Array, and all other participants genotyped via the Applied Biosystems UK biobank Axiom Array (Bycroft et al., 2018). Both arrays share 95% of markers. Genetic SNP arrays can use imputation to increase coverage, which was performed as described (Bycroft et al., 2018). Multiallelic SNPs or those with a minor allele frequency (MAF) of $< 0.01$ (1%) were removed prior to imputation, The UK10K and Haplotype Reference Consortium (HRC) reference pannels were the basis for imputation, performed by the MRC-IEU unit (Howie, Marchini, and Stephens, 2011; Huang et al., 2015). Individuals self identifying as 'White British' and having similar ancestry by clustering with others of the same ancestry in PCA were included, while those with a high degree of kinship to others in the biobank were excluded. The MRC-IEU also preformed quality control during imputation, excluding individuals with a sex mismatch between self reported and genetics, or any individuals with sex-chromosome aneuploidy. For a full description of imputation, see (Bycroft et al., 2018), and for a full description of in-house MRC-IEU quality control see (Ruth Mitchell, 2019).

Phenotypic data were obtained for the following fields for use in three studies in this chapter (Table 3.1). Three studies were conducted. First, a study of the influence of chronotype on self-reported mental health related traits. Then, a study of the effect of chronotype on social support and lifestyle factors. Lastly, a study encompassing the influence of chronotype on measures of eye health, taken by the best keratometry index measurements in the UKBB. When a categorical response has multiple values, they are separated into binary 1/0 outcomes for individual analysis.

| | Trait | Outcome Sample Size | Type | UKBB Showcase | Study |
|---|---|---|---|---|---|
| 1 | Ever highly irritable/argumentative for 2 days | 74750.000 | Binary | 4653 | Mental Health |
| 2 | Leisure/social activities: Religious group | 230684.000 | Binary | 6160 | Social Support |
| 3 | Leisure/social activities: None of the above | 230684.000 | Binary | 6160 | Social Support |
| 4 | Ever manic/hyper for 2 days | 74572.000 | Binary | 4642 | Mental Health |
| 5 | Guilty feelings | 225352.000 | Binary | 2030 | Mental Health |
| 6 | Seen doctor (GP) for nerves, anxiety, tension or depression | 229780.000 | Binary | 2090 | Mental Health |
| 7 | Illness, injury, bereavement, stress in last 2 years: Death of a close relative | 229871.000 | Binary | 6145 | Mental Health |
| 8 | Frequency of tenseness / restlessness in last 2 weeks | 222597.000 | Continuous | 2070 | Mental Health |
| 9 | Illness, injury, bereavement, stress in last 2 years: None of the above | 229871.000 | Binary | 6145 | Mental Health |
| 10 | Fed-up feelings | 226536.000 | Binary | 1960 | Mental Health |
| 11 | 3mm index of best keratometry results (right) | 48004.000 | Continuous | 5237 | Keratometry |
| 12 | Worry too long after embarrassment | 221959.000 | Binary | 2000 | Mental Health |
| 13 | Ever depressed for a whole week | 74950.000 | Binary | 4598 | Mental Health |
| 14 | Illness, injury, bereavement, stress in last 2 years: Serious illness, injury or assault to yourself | 229871.000 | Binary | 6145 | Mental Health |
| 15 | 6mm index of best keratometry results (left) | 41705.000 | Continuous | 5306 | Keratometry |
| 16 | Neuroticism score | 187162.000 | Continuous | 20127 | Mental Health |
| 17 | Manic/hyper symptoms: I was more creative or had more ideas than usual | 15658.000 | Binary | 6156 | Mental Health |
| 18 | Illness, injury, bereavement, stress in last 2 years: Death of a spouse or partner | 229871.000 | Binary | 6145 | Mental Health |
| 19 | Leisure/social activities: Sports club or gym | 230684.000 | Binary | 6160 | Social Support |
| 20 | Frequency of tiredness / lethargy in last 2 weeks | 224510.000 | Continuous | 2080 | Mental Health |
| 21 | 6mm index of best keratometry results (right) | 42152.000 | Continuous | 5251 | Keratometry |
| 22 | Number of depression episodes | 29145.000 | Continuous | 4620 | Mental Health |
| 23 | Suffer from 'nerves' | 222904.000 | Binary | 2010 | Mental Health |
| 24 | Sensitivity / hurt feelings | 224710.000 | Binary | 1950 | Mental Health |
| 25 | Manic/hyper symptoms: All of the above | 15658.000 | Binary | 6156 | Mental Health |
| 26 | Worrier / anxious feelings | 225382.000 | Binary | 1980 | Mental Health |
| 27 | Leisure/social activities: Other group activity | 230684.000 | Binary | 6160 | Social Support |
| 28 | Mood swings | 225810.000 | Binary | 1920 | Mental Health |
| 29 | Number of unenthusiastic/disinterested episodes | 19186.000 | Continuous | 5386 | Mental Health |
| 30 | Tense / 'highly strung' | 223980.000 | Binary | 1990 | Mental Health |
| 31 | Seen a psychiatrist for nerves, anxiety, tension or depression | 230351.000 | Binary | 2100 | Mental Health |
| 32 | Frequency of unenthusiasm / disinterest in last 2 weeks | 223702.000 | Continuous | 2060 | Mental Health |
| 33 | Illness, injury, bereavement, stress in last 2 years: Financial difficulties | 229871.000 | Binary | 6145 | Mental Health |
| 34 | Ever unenthusiastic/disinterested for a whole week | 73271.000 | Binary | 4631 | Mental Health |
| 35 | Manic/hyper symptoms: I was more active than usual | 15658.000 | Binary | 6156 | Mental Health |
| 36 | Irritability | 221084.000 | Binary | 1940 | Mental Health |
| 37 | Miserableness | 227491.000 | Binary | 1930 | Mental Health |
| 38 | 3mm index of best keratometry results (left) | 47999.000 | Continuous | 5292 | Keratometry |

| | | | | | |
|---|---|---|---|---|---|
| 39 | Frequency of friend/family visits | 229915.000 | Continuous | 1031 | Social Support |
| 40 | Longest period of depression | 32350.000 | Continuous | 4609 | Mental Health |
| 41 | Longest period of unenthusiasm / disinterest | 20591.000 | Continuous | 5375 | Mental Health |
| 42 | Frequency of depressed mood in last 2 weeks | 221420.000 | Continuous | 2050 | Mental Health |
| 43 | Able to confide | 224429.000 | Continuous | 2110 | Social Support |
| 44 | Illness, injury, bereavement, stress in last 2 years: Marital separation/divorce | 229871.000 | Binary | 6145 | Mental Health |
| 45 | Risk taking | 223140.000 | Binary | 2040 | Mental Health |
| 46 | Manic/hyper symptoms: None of the above | 15658.000 | Binary | 6156 | Mental Health |
| 47 | Loneliness, isolation | 227682.000 | Binary | 2020 | Mental Health |
| 48 | Illness, injury, bereavement, stress in last 2 years: Serious illness, injury or assault of a close relative | 229871.000 | Binary | 6145 | Mental Health |
| 49 | Nervous feelings | 225350.000 | Binary | 1970 | Mental Health |
| 50 | Length of longest manic/irritable episode | 12624.000 | Binary | 5663 | Mental Health |
| 51 | Manic/hyper symptoms: I was more talkative than usual | 15658.000 | Binary | 6156 | Mental Health |
| 52 | Leisure/social activities: Adult education class | 230684.000 | Binary | 6160 | Social Support |
| 53 | Manic/hyper symptoms: I needed less sleep than usual | 15658.000 | Binary | 6156 | Mental Health |
| 54 | Leisure/social activities: Pub or social club | 230684.000 | Binary | 6160 | Social Support |

Table 3.1: *UK Biobank (UKBB) statistics for each Mendelian Randomization study performed in this chapter.*

### 3.2.2 Performing GWAS

After harmonizing outcomes, PLINKv1.9 was used to perform GWAS against each trait studied, including chronotype (Chang et al., 2015a). During each GWAS, the following procedure was followed.

Quality control was carried out using PLINK R v3.5.0 (R Core Team, 2013). Variants which had a particularly high missing call rate (>0.1) were removed. Individuals with a missing rate of $> 0.05$ were also excluded. A 50kb window with a step size of 5 variants at a time and a 0.5 r2 threshold was used to remove variants in high LD.

Variants with a MAF <0.01 were also excluded. For each analysis, covariates included gender and the first ten principle components. PLINK was used to calculate the principal components using default settings. Prior to any testing, final QC steps were taken per trait, removing variants based on a case control missing rate likelihood of $< 0.001$ when a trait was binary, missing rate 0.05 or a Hardy Weinberg Equilibrium p $<$ 1E-8.

Linear models were run on all continuous traits, and logistic regression models were used for all binary traits. To make the comparison of effect sizes between binary and continuous outcomes efficient, the $\beta$ values from the generalized linear model during logistic regression were retained using PLINK's "–beta" flag. $\beta$, standard error ($se$), and p-values from each GWAS were kept for the data harmonization process.

### 3.2.3 Data Harmonization

When mining each GWAS for the discovery SNPs from 23andMe, a harmonization process was performed using the R TwoSampleMR package (Hemani et al., 2018). First, SNPs from the discovery set were located in the exposure GWAS summary statistics. The strandedness

of each GWAS was checked to make sure that at each allele, the minor and major alleles were equal. If these were reversed, effect sizes were modified to correct for this. Pallendromic SNPs, which contain alleles represented by the same base pairs on both strands of DNA, were discarded. If SNPs were not present, proxies were found using PLINK with an $R^2$ of at least 0.8, and strand was checked again (Chang et al., 2015b). Next, SNPs in the exposure GWAS set were clumped by LD to ensure statistical independence. In a window of 10000 base pairs, an $R^2$ cutoff of $< 0.001$ was set to obtain haplotype blocks, and the European reference panel of the 100,000 Genomes Project (Caulfield et al., 2017). This left 18 SNPs for use as valid, independent IVs. The effect sizes and standard errors from each GWAS were extracted for these 18 SNPs used in this analysis.

### 3.2.4 Causal Inference Modelling

In practice, meta-analyses often use an inverse-variance weighted average to account for the sample size (reflected in variance) of studies included in the meta-analysis. The IVW methods uses the Wald ratios of each SNPs as the "study" the meta-analysis, with a pooled estimate seen in the forest plots below (see results) driving home the meta-analytic nature of multi-SNP MR. Rather than calculate Wald ratios individually, the outcome GWAS $\beta$s or Odds Ratios are regressed on the exposure. The slope of the regression line indicates the strength of the effect, as an increase in the unit of outcome per unit of the exposure(Burgess and Thompson, 2017). In a IVW metaanalysis, IVW estimate is calculated by:

$$\widehat{\beta}_{Yj} = \theta_{IVW}\widehat{\beta}_{Xj} + \epsilon_{Ij}; \epsilon_{Ij} \sim N(0, \sigma^2 se(\widehat{\beta}_{Yj})^2) \tag{3.2}$$

where $\widehat{\theta}$ is the inverse variance weighted average, $se$ is the standard error, and other terms are as above, and I is an error term.

Similarly to other meta-analysis frameworks, by weighting effect sizes by their in-

verse variance, stronger SNPs make a larger contribution to the overall effect size obtained. Compared to other methods used in this analysis, the IVW method has the strongest power simply by not discarding any SNP instruments (Mode) or shrinking their variance (MR Egger). It strongly assumes all instruments are valid, requiring the slope of the meta-regression line to be constrained to zero. Any potential pleiotropy or extreme heterogeneity among SNPs would draw the gradient of the regression line away from the true slope. Heterogeneity can occur when individual SNPs do not converge on an estimate, and can be estimate by Cochran's $Q$ (Higgins and Green, 2011). In this context, heterogeneity may be a sign of horizontal pleiotropy, wherein SNPs effect the outcome by their influence on other confounding traits (Burgess, Small, and Thompson, 2017). In any IVW-based meta-analysis, either fixed or random effects maybe modeled (Higgins and Green, 2011). A fixed effect model applied to MR assumes all instruments are valid, whereas modeling random effects allows for balanced horizontal pleiotropy to be present if it is independent of SNPs effects on the exposure. This is termed the Instrument Strength Independent of Direct Effect (InSIDE) assumption, which is not currently testable. The IVW method also relies on the No Measurement Error (NOME) assumption, assuming SNP-exposure associations are accurate. This can be accessed by calculating F-statistics to test the degree of association between discovery SNPs and the exposure of interest.

A figure comparing the IVW method with the pleiotropy- and assumption-mitigating factors below is shown in Figure 3.4. Each scatter plot depicts an outcome regressed on an exposure with multiple SNPs. The slope is the estimate of the causal effect. If there is no horizontal pleiotorpy, or the pleiotropy is balanced between outcome and effect, inverse-variance weighted regression is used, where the contribution of each SNP is weighted by its inverse variance, so variable SNPs contribute less to the overall study effect size. If directional pleiotropy is suspected, then constraining the intercept to be at zero will allow bias into the model (gray arrow). However if the intercept is not constrained via Egger

regression, an unbiased estimate can be obtained if the instrument-exposure and pleiotropy are uncorrelated (InSIDE assumption). If most instruments are valid (black), and some are invalid (red), a median-based approach will provide an unbiased estimate (black), whereas IVW linear regression would provide a biased estimate (grey). If SNPs are horizontally pleiotropic, they will return biased estimates. Using a mode-based estimator, it is possible to clustering SNPs based on their estimates (grey lines). If the majority of the SNPs are in a cluster which satisfies IV assumptions, then then mode-based estimator is unbiased, if not under powered.

### 3.2.5 Testing for Evidence of Pleiotropy

When an MR study, with multiple valid instruments, is considered as a meta-analysis, meta-analytic tools used to detect bias in studies can be used to detect bias in SNPs. Egger proposed a method for detecting small study bias in meta-analyses, and this has been adopted into an MR context (Bowden et al., 2017; Bowden, Davey Smith, and Burgess, 2015). The Wald ratios of each SNP are used in meta-regression by taking the inverse IVW weights used in IVW analysis, without modeling the intercept. It provides a causal estimate similar to IVW but adjusted for horizontal pleiotropy which would otherwise invalidate IVW (Bowden, Davey Smith, and Burgess, 2015). As the intercept is unconstrained, it estimates the average pleiotropic effect across the SNPs, and the slope provides an estimate of the pleiotropic effect. If the intercept, while not constrained, is not statistically different than zero, then this suggest lack of horizontal pleiotropy and the instruments are assumed valid. To accomplish this, the MR-Egger method relaxes assumption 3 in Figure 3.2. MR-Egger relies on not violating the InSIDE assumption (see above).

MR-Egger regression is an extension of IVW regression. Instead of assuming no

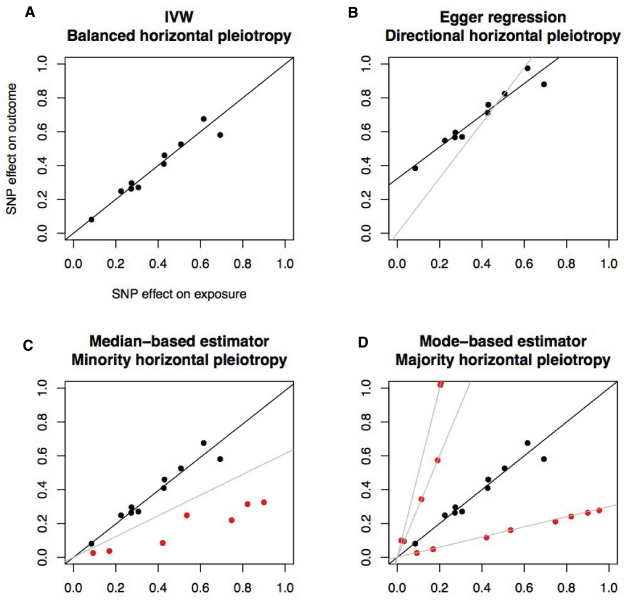



Figure 3.4: *Mendelian Randomization methods have different assumptions. Each scatter plot depicts an outcome regressed on an exposure with multiple SNPs. The slope is the estimate of the causal effect. A) If there is no horizontal pleiotorpy, or the pleiotropy is balanced between outcome and effect, inverse-variance weighted regression is used, where the contribution of each SNP is weighted by its inverse variance, so variable SNPs contribute less to the overall study effect size. B) If directional pleiotropy is suspected, then constraining the intercept to be at zero will allow bias into the model (gray arrow). However if the intercept is not constrained via Egger regression, an unbiased estimate can be obtained if the instrument-exposure and pleiotropy are uncorrelated (InSIDE assumption). C) If most instruments are valid (black), and some are invalid (red), a median-based approach will provide an unbiased estimate (black), whereas IVW linear regression would provide a biased estimate (grey). D) If SNPs are horizontally pleiotropic, they will return biased estimates. Using a mode-based estimator, it is possible to clustering SNPs based on their estimates (grey lines). If the majority of the SNPs are in a cluster which satisfies IV assumptions, then then mode-based estimator is unbiased, if not under powered. Image credit: (Hemani, Bowden, and Davey Smith, 2018).*

intercept term, an intercept is estimated:

$$\widehat{\beta}_{Yj} = \theta_{0E} + \theta_{1E}\widehat{\beta}_{Xj} + \epsilon_{Ej}; \epsilon_{Ej} \sim N(0, \sigma^2 se(\widehat{\beta}_{Yj})^2) \tag{3.3}$$

where $\theta_{0E}$ is the intercept and $\theta_{1E}$ the MR Egger estimate. If the intercept is equal to zero, then the IVW method and MR-Egger will be equivalent (Burgess and Thompson, 2017). During the IVW process in MR-Egger, the effect sizes of each SNP must have the same sign, and this decreases the variation between them (Bowden, Davey Smith, and Burgess, 2015). This renders the MR-Egger method the lowest powered method employed in this chapter, though its robustness to horizontal pleiotropy and ability to test for this in the intercept make it a valuable contribution to methods.

### 3.2.6 Assuming Weak Instrumental Variables: Median and Mode

The weighted median estimate again relies on Wald ratios. First, the Wald ratio is calculated for each SNP. Then, as in the IVW method, a measure of central tendancy is used to produce an overall effect size - but instead of the mean, the median is used. In an unweighted analysis it would be assumed that over half of the instruments are valid, while in an weighted median analysis the assumption is that the at least 50% of the weight of the instruments are valid themselves (Bowden et al., 2016). This approach is robust to directional pleiotropy when compared to a simple IVW meta-analysis. It is also more robust to outliers than either IVW or Mr Egger.

Comparative studies have demonstrated that the power of the weighted median method is similar to the IVW. Moreover, it is not constrained by the InSIDE assumption, contrary to MR Egger, and for it to be valid only half of instruments need to be unbiased (Hemani, Bowden, and Davey Smith, 2018). See Figure 3.4 C for a graphical explanation.

The mode-based estimator (MBE) clusters Wald ratios before calculating random

effects in an IVW meta-analysis (Hartwig, Davey Smith, and Bowden, 2017). The simple MBE uses unweighted analysis, while the weighted MBE uses inverse variance weighting. First, a smooth empirical density function is calculated for each Wald ratio and are then clustered. The ZEMPA assumption (Zero Modal Pleiotropy Assumption) states that the biggest cluster with the same ratio estimates will be valid instruments. Thus, the MBE can provide a valid causal estimate when the largest number of estimates come from valid IVs, even if a majority (those in smaller clusters) are invalid. While not all snps have to be valid IVs, fewer valid IVs is equivalent to fewer samples actually used, and results in lost power compared to IVW methods. I calculated both weighted (by IVW) and unweighted (Wald ratio) methods. See Figure 3.4 D for a graphical explanation.

### 3.2.7 Sensitivity, Bias, and Directionality

To check if possible pleiotrophic or invalid SNPs are dominating models run, I performed a leave-one-out analysis for each experiment. In a leave-one-out senario, the inverse variance weighted average is run using all SNPs except one. This is repeated for every SNP, resulting in 10 IVW average effect sizes in the case of 10 SNPs in the analysis. One outlier effect size would indicate that the SNP left out skews the analysis and may be removed for being an outlier (Hemani, Tilling, and Davey Smith, 2017).

I composed funnel plots to access pleiotopic effects. Originally designed to identify potential biases arising from small studies in meta-analysis, funnel plots are a graphical method of accessing bias, by plotting the effect size against sample size, noting that smaller effects should belong to smaller effect sizes and that there should be a degree of balance in the signs of the effect sizes (Egger et al., 1997). Egger proposed a regression method to test the asymmetry of this relationship, which is the basis of the MR Egger method (Bowden, Davey Smith, and Burgess, 2015). Cochran's $Q$ can be used in combination with funnel plots

to access the degree of pleiotropy. $Q$ is a measure of heterogeneity among calculates the sum of squared differences between SNPs and the pooled effect, and creates a $\chi^2$ test statistic for comparison. The $Q$ statistic is well powered as the number of SNPs increases, indicating that evidence of overdispersion itself does not indicate pleiotropy (Lawlor et al., 2008; Burgess, Small, and Thompson, 2017). Even for a low p-value, indicating high heterogeneity, I have investigated funnel plots, as severe asymmetry would suggest directional pleiotropy and violate MR-IVW assumptions. If symmetry is observed, both fixed- and random-effects IVW would produce similar results (Bowden et al., 2017), and if asymmetry is observed then estimates would differ, and Median or Mode methods should be considered. In this chapter, the IVW analysis is considred the gold standard; if it is significant, other methods are used to confirm results in the presence of pleiotropy or heterogeneity. If all methods of analysis for an experiment concur, the potential causal effect of chronotype is considered more plausible.

I used the Steiger test to access the directional of all causative analyses post-hoc (Hemani, Tilling, and Davey Smith, 2017). This method tests whether the selected IVs are more strongly associated with the exposure than the outcome. The Steiger test first assesses which variables (exposure or outcome) are influenced by the SNPs used, by testing if the SNPs explain more variance in the exposure than in the outcome with a modified $Z$ statistic. If the p-value of the IVW estimate and the Steiger estimate are both significant, the sign of the $Z$ statistic is used to assign the correct causal direction between exposure and outcome.

## 3.3 Results

Results in this chapter are presented as follows. An example GWAS is reported, IVW meta-analyses of all traits tested are shown in graphical form, and then individual traits are

reported. Reports from all traits are included in the Appendix. MR analyses were performed for the influence of chronotype on each trait listed in Table 3.1.

### 3.3.1 Phenome-wide overview of Circadian effects of psychosocial and ophthalmic traits

A linear GWAS of morning - to - evening preference in self-reported chronotype is depicted in Figure 3.5. This GWAS reveals several independent loci which associate highly with chronotype. In green, 18 clumped discovery SNPs from the 23andMe GWAS are depicted. Note that not all of them achieve genome-wide significance levels ($p <= 10^{-}8$). The y-axis shows $-log_10$ p-values, while the x-axis shows the location of each SNP by chromosome.

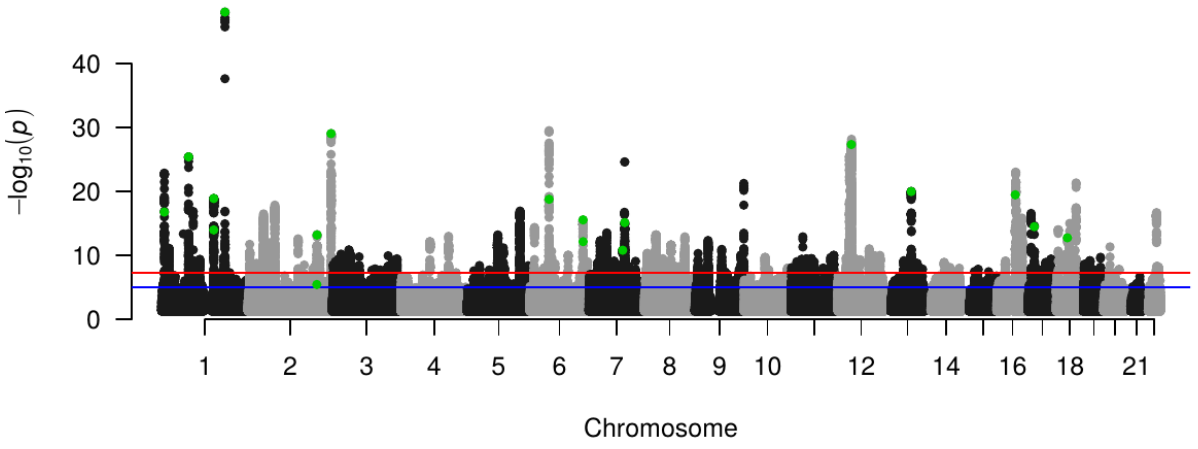



Figure 3.5: *Manhattan plot of morning/evening chronotype GWAS using the UK Biobank population. Along the x-axis, SNPs are represented by chromosome location, and the y-axis depicts $-log_10$ p-values. SNPs from the discovery GWAS are highlighted in green.*

Table 3.2 lists the SNPs used in this study by location, including effect size and p-

values of Chronotype (exposure) and Leisure Activity: Religious Group (outcome). This exposure produced the most extreme IVW p-value ($4.20e-5$). All p-values in exposure (Chronotype) GWAS are highly significant except for one, at the level of $3.5e-6$. No p-values from the outcome GWAS reach genome-wide significance, with a range of (0.0003,0.9). $\beta$ values are likewise smaller from the outcome than the exposure.

| | SNP | Pvalue Chronotype | Pvalue Religous Group | Beta Chronotype | Beta Religous Group |
|---|---|---|---|---|---|
| 1 | rs1064213 | 6.20012e-14 | 0.0850002 | -0.0152572 | -0.00125918 |
| 2 | rs11121022 | 1.59993e-17 | 0.6 | -0.0174991 | 0.00038933 |
| 3 | rs112613078 | 7.19946e-16 | 0.25 | -0.0208278 | 0.00106677 |
| 4 | rs11545787 | 3.10027e-15 | 0.14 | 0.0187265 | -0.00125028 |
| 5 | rs11587758 | 1.20005e-19 | 0.00109999 | -0.0187958 | 0.00244028 |
| 6 | rs12133238 | 1e-14 | 0.0729995 | 0.0174009 | -0.00144885 |
| 7 | rs12617426 | 3.50002e-06 | 0.760001 | 0.00974035 | -0.000234953 |
| 8 | rs12927162 | 3.19963e-20 | 0.01 | 0.0209482 | -0.00210022 |
| 9 | rs4239386 | 1.69981e-13 | 0.26 | 0.0159216 | -0.000872088 |
| 10 | rs4729303 | 1.50003e-11 | 0.75 | -0.0175607 | 0.000294364 |
| 11 | rs4882354 | 4.30031e-28 | 0.012 | -0.0224835 | 0.00185288 |
| 12 | rs509476 | 8.10028e-49 | 0.0219999 | 0.0878165 | -0.00491105 |
| 13 | rs57435966 | 8.49963e-30 | 0.000350002 | 0.0410747 | -0.00465797 |
| 14 | rs62436127 | 6.79986e-13 | 0.9 | -0.019918 | -0.000119614 |
| 15 | rs7547493 | 3.69999e-26 | 0.38 | -0.0281137 | 0.000845037 |
| 16 | rs76223855 | 2.70023e-16 | 0.22 | -0.0798575 | -0.00429174 |
| 17 | rs9475185 | 1.59993e-19 | 0.34 | -0.021635 | 0.000818634 |
| 18 | rs9565309 | 1e-20 | 0.81 | 0.0526882 | -0.000495558 |

Table 3.2: *Discovery SNPs in chronotype and religous group GWAS*

The IVW analysis suggested that eight self-reported mental health traits were indicative of possible causation ($p < 0.05$).

These include associations between a chronotype and manic symptoms, depression, reactions to feelings (sensitivity and worry), and irritability, Figure 3.6. Wide confidence intervals indicate a large confidence interval around the IVW $\beta$. Those confidence intervals crossing the null line (zero) show a null effect. The results of each analysis, including non-significant results, are reported in the online Appendix.

The IVW analysis revealed that two self-reported social traits were indicative of pos-

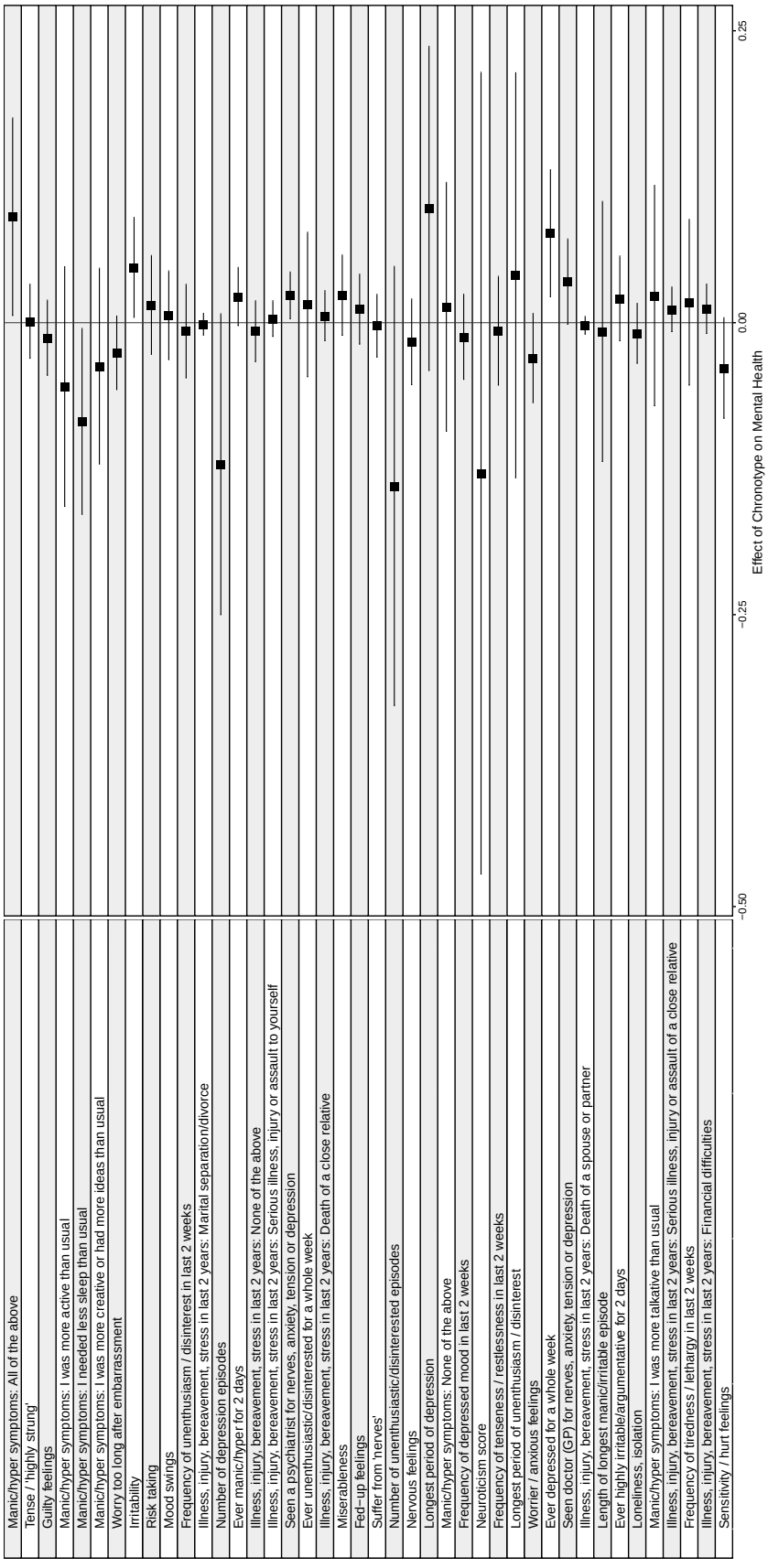



Figure 3.6: *Forest plots of change in standard deviation of each Mental Health trait response as chronotype differs.*

sible causation ($p < 0.05$).

These were related to the attendance at group religious meetings weekly, and the frequency of friend or family visits (Figure 3.6). Wide confidence intervals indicate a large confidence interval around the IVW $\beta$. Those confidence intervals crossing the null line (zero) show a null effect. The results of each analysis, including non-significant results, are reported in the online Appendix.

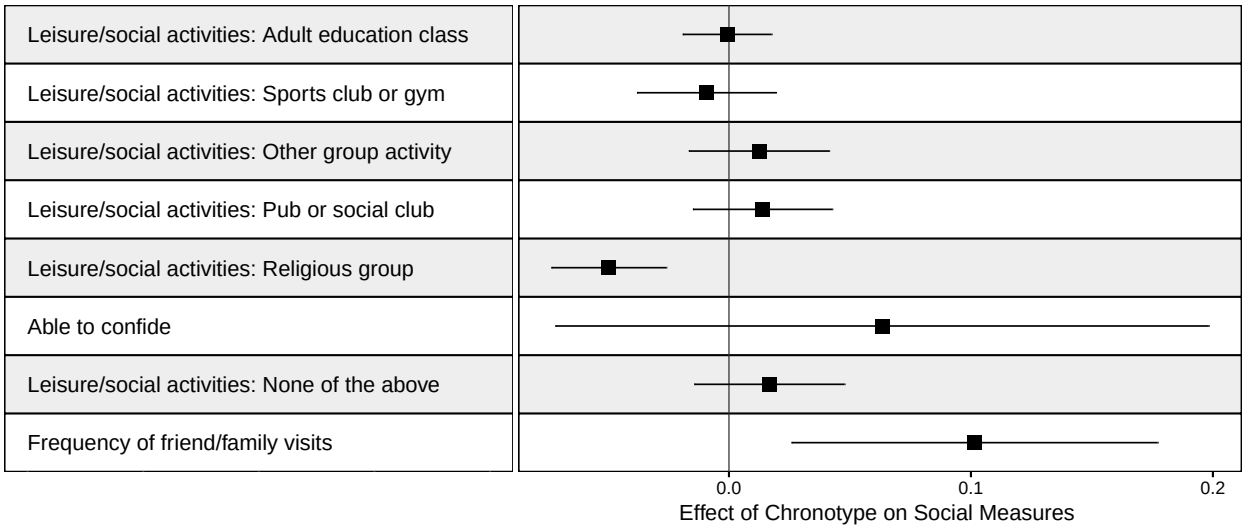



Figure 3.7: *Forest plots of change in standard deviation of each Social Support trait response as chronotype differs.*

Under IVW analysis, one keratometry result was indicative of possible causation.

A 3mm index of the best keratometry result, related to right eye measurements, was significant, and all others, while not significant, were trending in the same direction, Figure 3.6. Wide confidence intervals indicate a large confidence interval around the IVW $\beta$. Those confidence intervals crossing the null line (zero) show a null effect. The results of each analysis, including non-significant results, are reported in the online Appendix.

The IVW results thus far presented provide a phenome-wide level overview of the

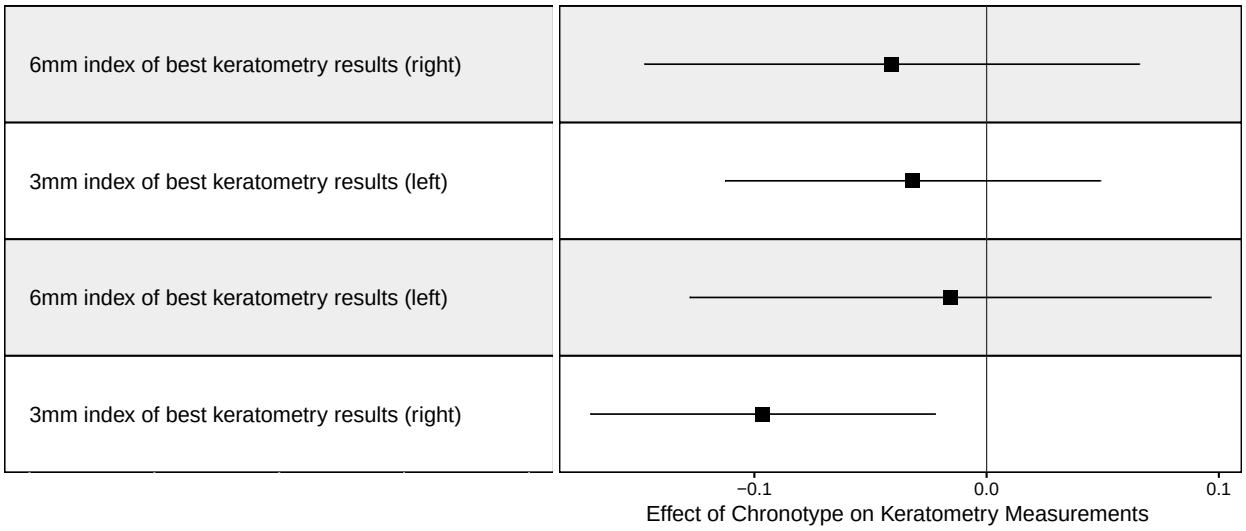



Figure 3.8: *Forest plots of change in standard deviation of each kratometry trait response as chronotype differs.*

effect of chronotype on the phenotypes measured in their respective UKBB domains. As each MR analysis is a self-contained experiment, significant results are reported in the following sections.

### 3.3.2 Morningness and eveningness influence mental health

Regressing behavioral traits on GWAS data revealed chronotype to have a likely causal influence on several tested traits. I observed an increase of 0.090 (IVW pval 0.036, Egger intercept pval 0.970, Q pval 0.213) standard deviation (SD) units of self-reported manic/hyper symptoms [all of the above] for evening compared to morning chronotype (Figure 3.9, Table 3.3). A decrease in the need for sleep while in a high or irritable state of 0.085 SD units (IVW pval 0.037, Egger intercept pval 0.925, Q pval 0.848) for evening compared to morning (Figure 3.10, Table 3.4).

Three self-reported measures of depression produced robust causative models. I ob-

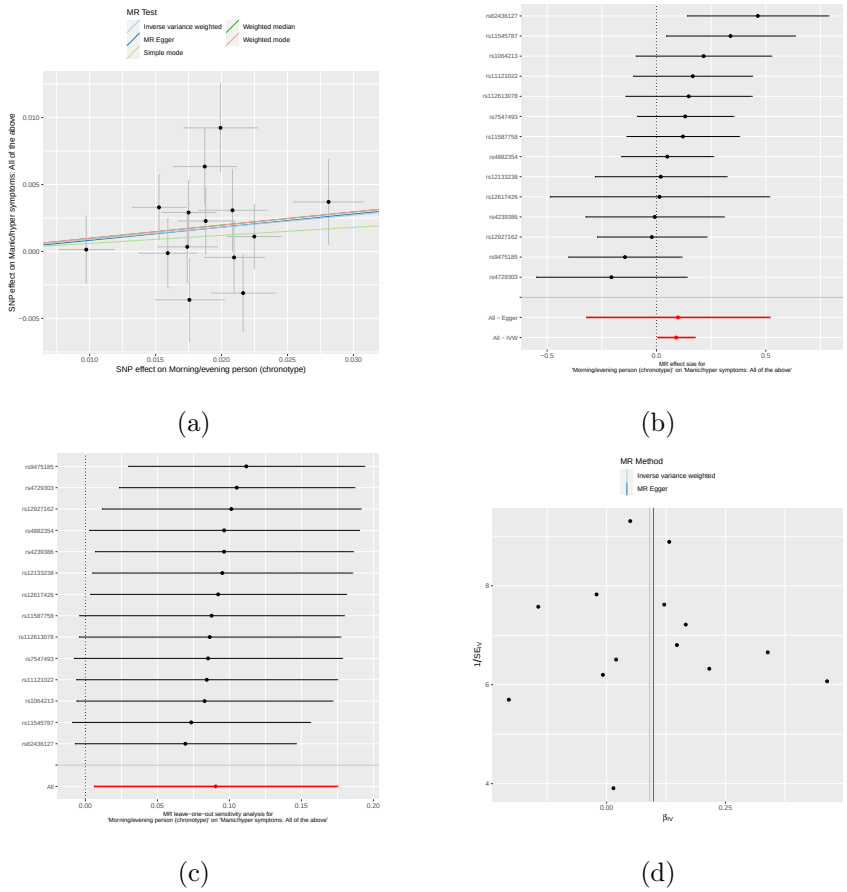



Figure 3.9: *An evening chronotype suggests a small increase in manic symptoms. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

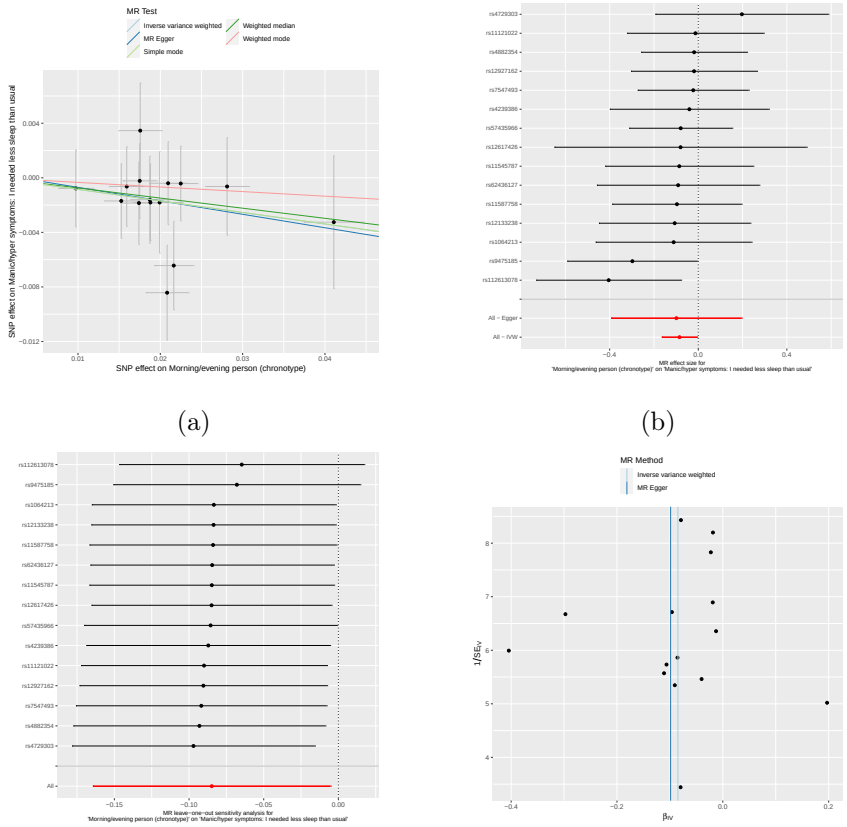



Figure 3.10: *An evening chronotype suggests a decrease in reported need for less sleep, therefore increased duration of sleep. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | 0.090 | 0.043 | 0.036 | 16.700 | 0.213 | | |
| 2 | MR Egger | 0.098 | 0.214 | 0.654 | 16.698 | 0.161 | -0.000 | 0.970 |
| 3 | Simple mode | 0.060 | 0.091 | 0.522 | | | | |
| 4 | Weighted median | 0.098 | 0.053 | 0.064 | | | | |
| 5 | Weighted mode | 0.098 | 0.077 | 0.225 | | | | |

Table 3.3: *Causal effect of evening chronotype on Manic/hyper symptoms: All of the above.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | -0.085 | 0.041 | 0.037 | 8.737 | 0.848 | | |
| 2 | MR Egger | -0.099 | 0.150 | 0.522 | 8.727 | 0.793 | 0.000 | 0.925 |
| 3 | Simple mode | -0.084 | 0.085 | 0.336 | | | | |
| 4 | Weighted median | -0.074 | 0.052 | 0.155 | | | | |
| 5 | Weighted mode | -0.034 | 0.080 | 0.679 | | | | |

Table 3.4: *Causal effect of evening chronotype on Manic/hyper symptoms: I needed less sleep than usual.*

served an increase of 0.023 (IVW pval 0.024, Egger intercept pval 0.907, Q pval 0.112) (SD) incidence of seeing a psychiatrist for nerves, anxiety, tension or repression with an evening chronotype (Figure 3.11, Table 3.5). Likewise, an increase in the likelihood of having had a depressive episode lasting at least a week of 0.076 SD units (IVW pval 0.0.006, Egger intercept pval 0.860, Q pval 0.110) for evening compared to morning, Figure 3.12, Table 3.6.

Lastly, there is a marked decrease in odds of reporting "feelings being easily hurt" when a 'evening person'. This association was not significant under the IVW model, but was with more robust analyses: a weighted median and mode (-0.053 and -0.049 SD units, respectively).

Egger regression suggests lack of horizontal pleiotropy (intercept $p = 0.704$) but strong heterogeneity of SNPs (Q statistic 47, $p < 0.000$) - Figure 3.13, Table 3.7. A similar decrease

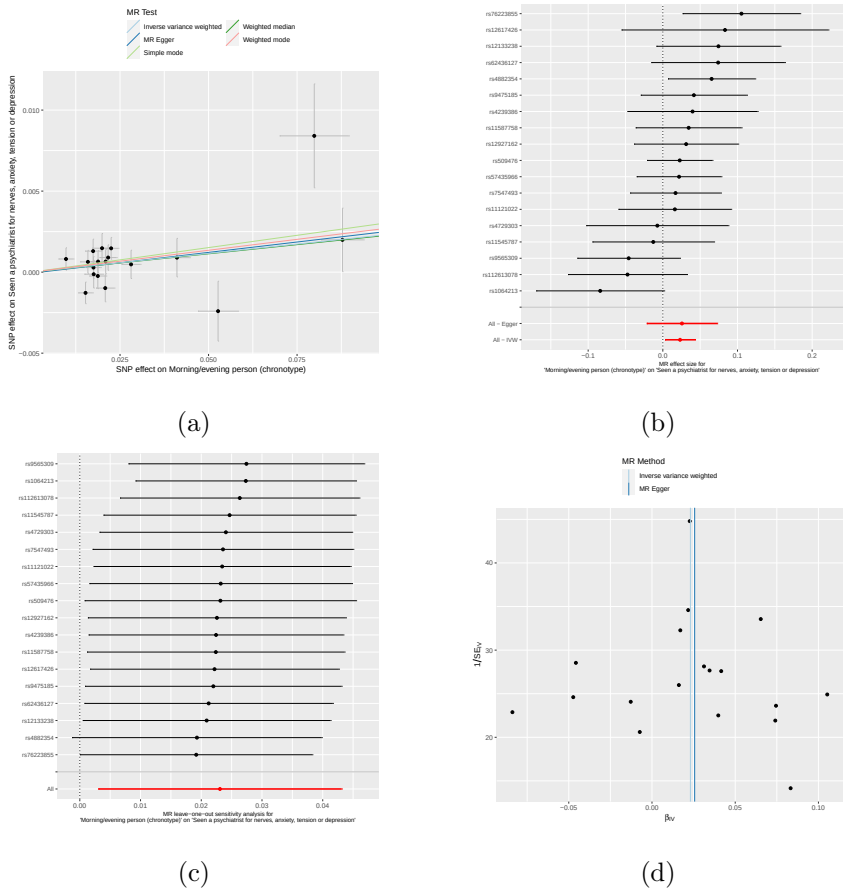



Figure 3.11: *An evening chronotype suggests an increase in self reported anxiety and depression. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

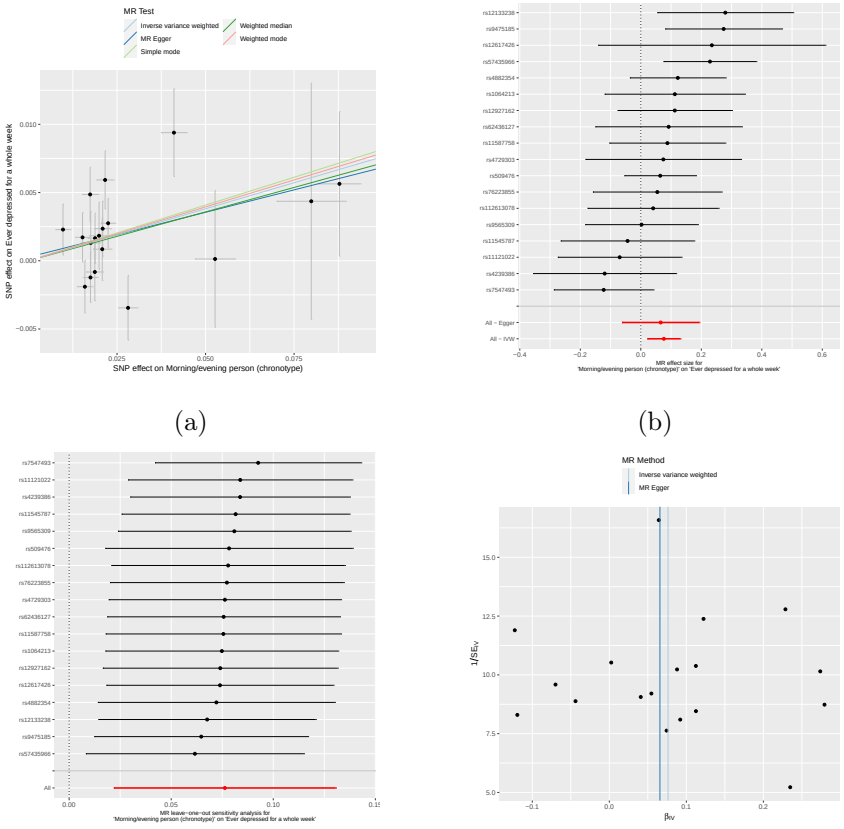



Figure 3.12: *An evening chronotype suggests an increase in self reported depression episodes lasting longer than one week. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | 0.023 | 0.010 | 0.024 | 24.269 | 0.112 | | |
| 2 | MR Egger | 0.026 | 0.024 | 0.302 | 24.248 | 0.084 | -0.000 | 0.907 |
| 3 | Simple mode | 0.030 | 0.021 | 0.169 | | | | |
| 4 | Weighted median | 0.023 | 0.012 | 0.067 | | | | |
| 5 | Weighted mode | 0.027 | 0.018 | 0.150 | | | | |

Table 3.5: *Causal effect of evening chronotype on Seen a psychiatrist for nerves, anxiety, tension or depression.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | 0.076 | 0.028 | 0.006 | 24.365 | 0.110 | | |
| 2 | MR Egger | 0.066 | 0.065 | 0.328 | 24.316 | 0.083 | 0.000 | 0.860 |
| 3 | Simple mode | 0.082 | 0.067 | 0.240 | | | | |
| 4 | Weighted median | 0.072 | 0.034 | 0.033 | | | | |
| 5 | Weighted mode | 0.079 | 0.057 | 0.181 | | | | |

Table 3.6: *Causal effect of evening chronotype on Ever depressed for a whole week.*

was seen in SD unit self reporting of worrying too long after an embarrassing episode, though not significant with IVW (p = 0.104), reached an effect size of $\beta$ -0.052 with a Weighted Median regression (p = 0.06, Egger intercept p = 0.268, Q pval = 0.113) - Figure 3.14, Table 3.8. Finally, a study on irritability returned mixed results. IVW analysis suggests that there is a 0.047 SD unit increase in irritability per SD increase of self-reported evening chronotype (p = 0.031). Non-significant values were returned for robust analyses, indicating lack of power, which coincided with high heterogeneity among SNPs (Q 54, Q pval = $<$ 0.001), while MR Egger intercept p-value suggests unlikely pleiotropy (p = 0.08), Figure and Table 3.15 and 3.9.

Other analyses did not yield suggestive results using the IVW method, see the online appendix, https://github.com/jaw-bioinf/PhdThesis.

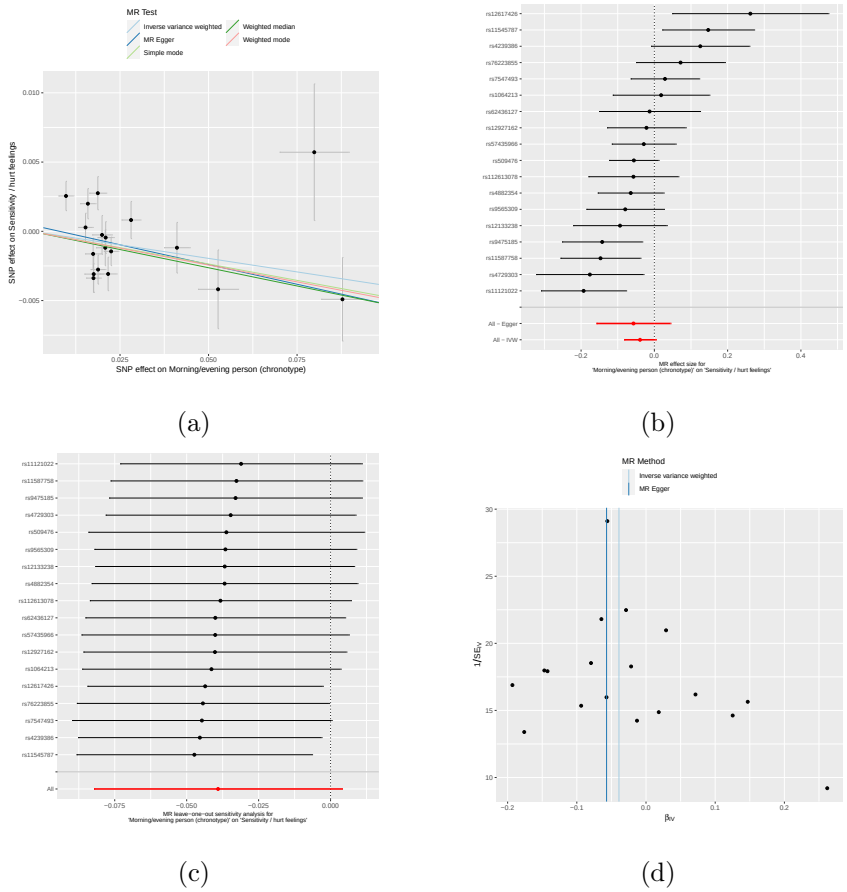



Figure 3.13: *An evening chronotype suggests an decrease in self reported sensitivity to hurt feelings. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

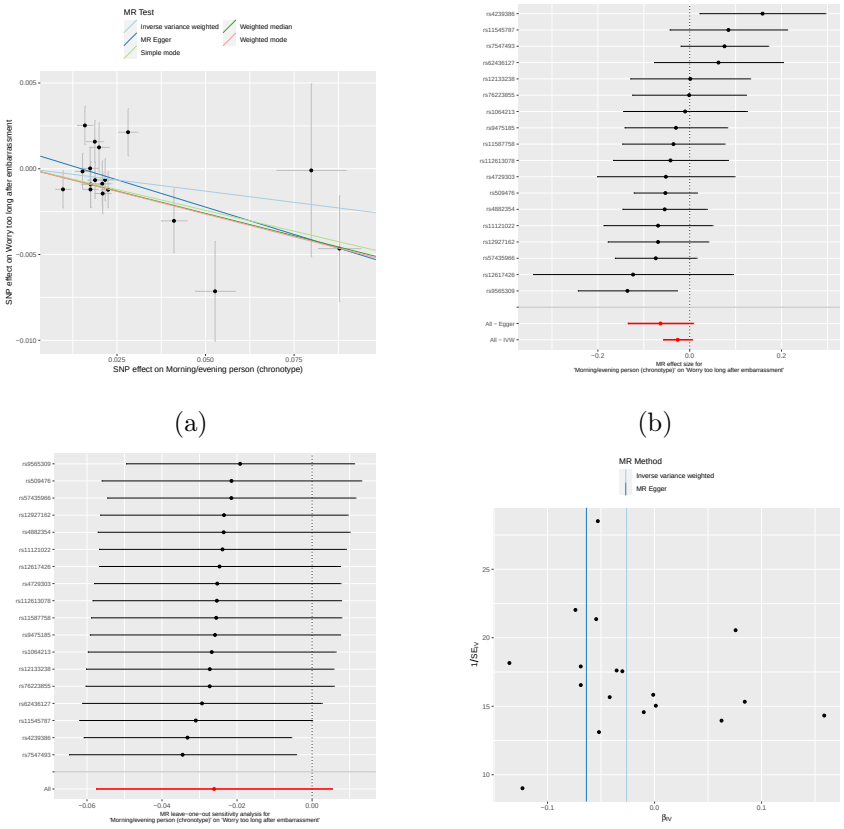
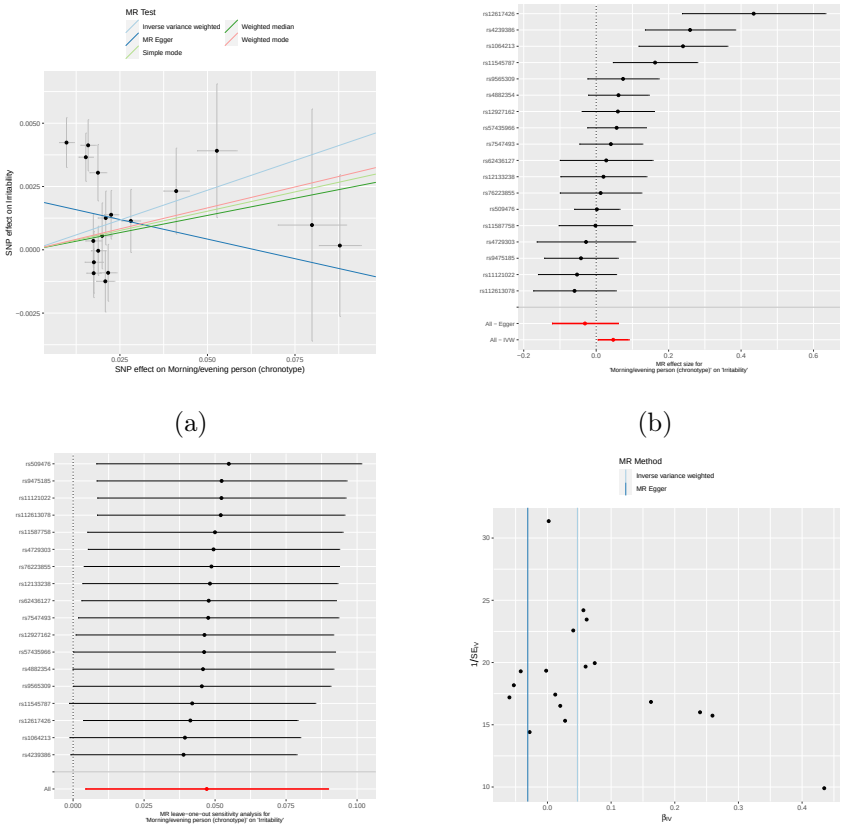



Figure 3.14: *An evening chronotype suggests an decrease in self reported worrying after embarrassing episodes. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*




(a) (b)

(c) (d)

Figure 3.15: *An evening chronotype suggests an increase in irritability. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | -0.039 | 0.022 | 0.075 | 47.398 | 0.000 | | |
| 2 | MR Egger | -0.057 | 0.051 | 0.285 | 46.959 | 0.000 | 0.000 | 0.704 |
| 3 | Simple mode | -0.047 | 0.035 | 0.195 | | | | |
| 4 | Weighted median | -0.053 | 0.020 | 0.010 | | | | |
| 5 | Weighted mode | -0.049 | 0.027 | 0.086 | | | | |

Table 3.7: *Causal effect of evening chronotype on Sensitivity / hurt feelings.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | -0.026 | 0.016 | 0.104 | 24.243 | 0.113 | | |
| 2 | MR Egger | -0.064 | 0.036 | 0.098 | 22.389 | 0.131 | 0.001 | 0.267 |
| 3 | Simple mode | -0.049 | 0.029 | 0.117 | | | | |
| 4 | Weighted median | -0.052 | 0.019 | 0.006 | | | | |
| 5 | Weighted mode | -0.053 | 0.025 | 0.047 | | | | |

Table 3.8: *Causal effect of evening chronotype on Worry too long after embarrassment.*

### 3.3.3 CR influences measures of social support

Four categories of social support measures were studied in relation to self-reported chronotype. Of these, religious behavior and the frequency of visiting family over the last year produced strong associations. Modeling attending group religious activities at least once a week produced a decrease of -0.050 SD units per likelihood $\beta$ of being an evening person ($p < 0.001$, Q pval = 0.041, Egger intercept pval = .959). Self-reported visits to friends or family was coded as an ordinal variable, and increases in this axis indicates less frequent socialization with friends or family (or their absence outside the home). Evening chronotype was indicative of fewer visits (IVW $\beta$ 0.102, p = 0.009), though suggestions of strong heterogeneity (Q pval 0.037) and possible pleiotropy (Egger intercept .004, p = 0.069) were present, and robust methods yielded no signal. See Figure 3.17, Table 3.11.

Other analyses did not yield suggestive results using the IVW method, see the online

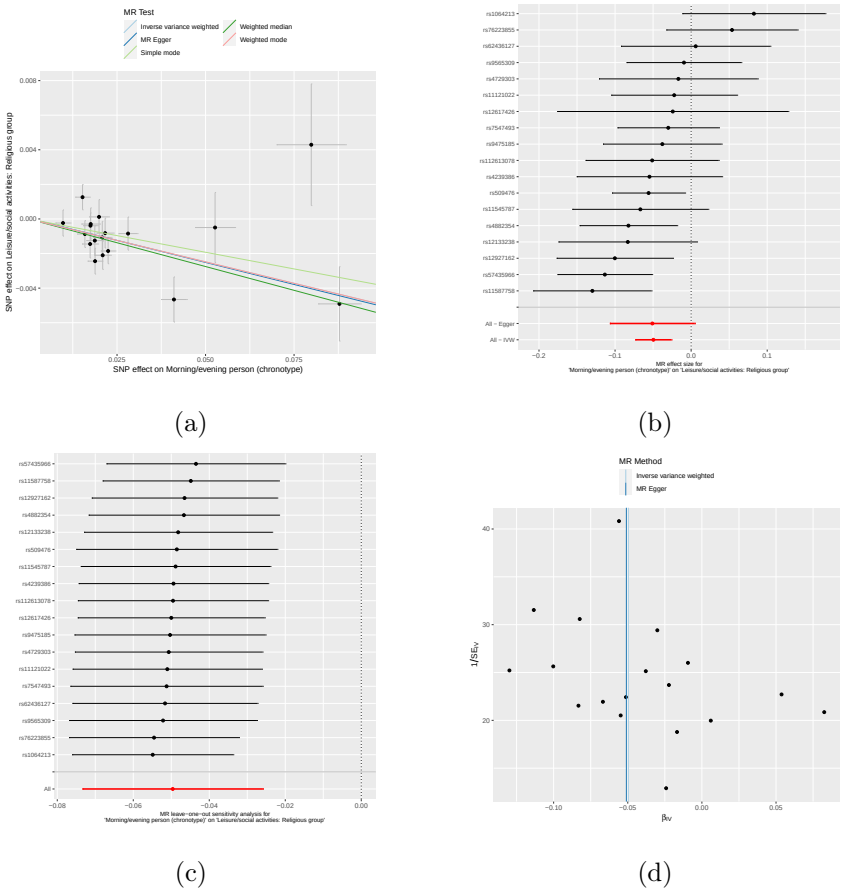



Figure 3.16: *An evening chronotype suggests an decrease in group religious activity. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

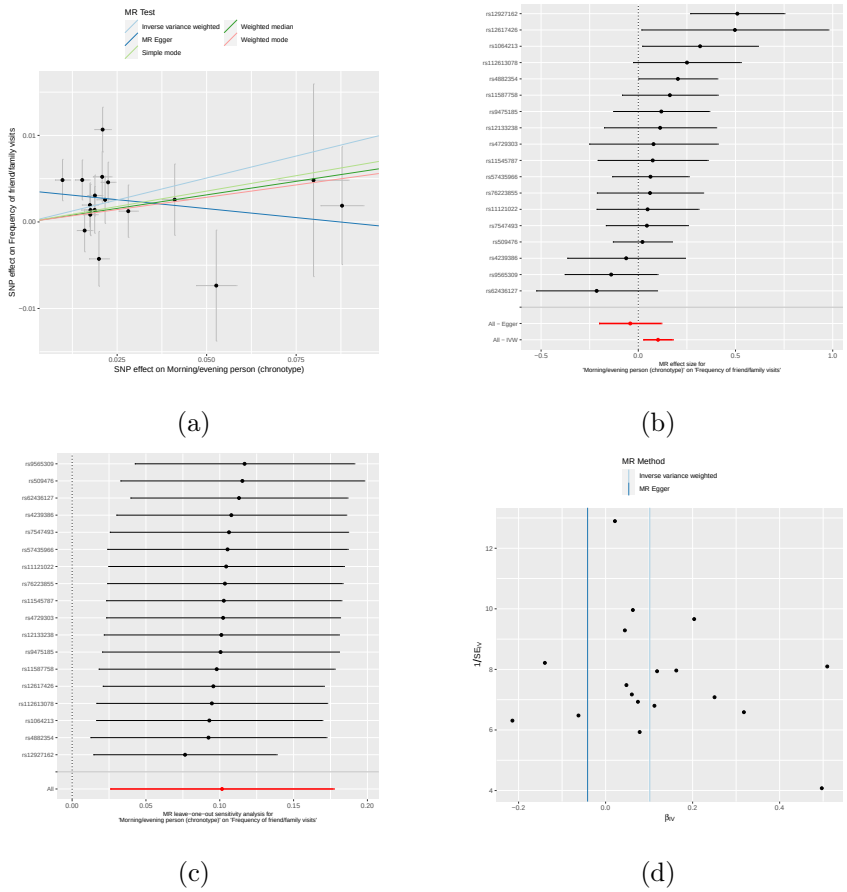



Figure 3.17: *An evening chronotype suggests an increase in the frequency of family visits. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | 0.047 | 0.022 | 0.031 | 54.070 | 0.000 | | |
| 2 | MR Egger | -0.031 | 0.046 | 0.515 | 44.402 | 0.000 | 0.002 | 0.080 |
| 3 | Simple mode | 0.030 | 0.025 | 0.236 | | | | |
| 4 | Weighted median | 0.027 | 0.018 | 0.124 | | | | |
| 5 | Weighted mode | 0.033 | 0.022 | 0.144 | | | | |

Table 3.9: *Causal effect of evening chronotype on Irritability.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | -0.050 | 0.012 | 0.000 | 28.338 | 0.041 | | |
| 2 | MR Egger | -0.051 | 0.029 | 0.093 | 28.334 | 0.029 | 0.000 | 0.959 |
| 3 | Simple mode | -0.039 | 0.025 | 0.141 | | | | |
| 4 | Weighted median | -0.055 | 0.014 | 0.000 | | | | |
| 5 | Weighted mode | -0.050 | 0.021 | 0.028 | | | | |

Table 3.10: *Causal effect of evening chronotype on Leisure/social activities: Religious group.*

appendix, https://github.com/jaw-bioinf/PhdThesis.

### 3.3.4 Chronotype affects eye morphology

Finally, I modelled the causitive influence of evening chronotype on keratometry measurements to study potential interplay between eye physiology and circadian biology. I found a decrease in SD units of 3mm keratometry index in the right eye only per unit of likelihood of having an evening phenotype (IVW $\beta$ -0.096, p = 0.011, Q pval 0.607, Egger intercept pval 0.273. This trend was echoed non-significantly by results from the weighted median method ($\beta$ -0.093, p = 0.078).

Other analyses did not yield suggestive results using the IVW method, see the online appendix, https://github.com/jaw-bioinf/PhdThesis.

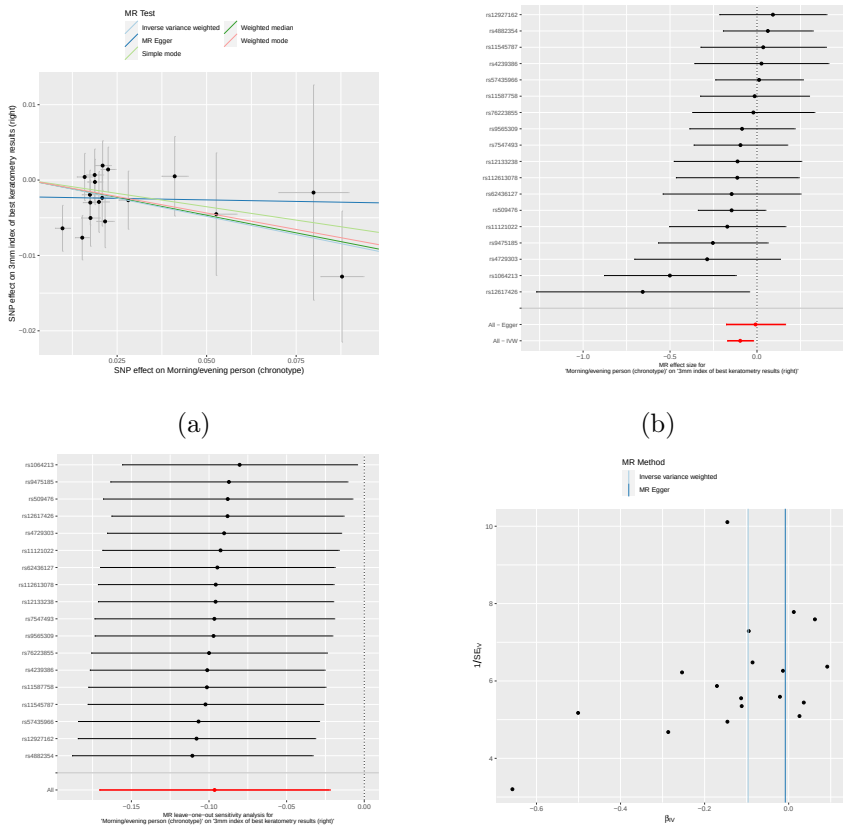



Figure 3.18: *An evening chronotype suggests an decease in corneal strength measured by keratometry index in the right eye. a) scatter plot of MR results. b) Forest plot of SNP contributions to IVW analysis. c) Leave one out sensitivity analysis using IVW method. d) Funnel plot accessing directional pleiotropy.*

| | Method | b | se | pval | Q | Q_pval | intercept | intercept_pval |
|---|---|---|---|---|---|---|---|---|
| 1 | Inverse variance weighted | 0.102 | 0.039 | 0.009 | 28.702 | 0.037 | | |
| 2 | MR Egger | -0.041 | 0.082 | 0.619 | 23.205 | 0.108 | 0.004 | 0.069 |
| 3 | Simple mode | 0.071 | 0.065 | 0.290 | | | | |
| 4 | Weighted median | 0.063 | 0.042 | 0.141 | | | | |
| 5 | Weighted mode | 0.057 | 0.060 | 0.354 | | | | |

Table 3.11: *Causal effect of evening chronotype on Frequency of friend/family visits.*

## 3.4 Discussion

This chapter presents a phenome-wide investigation of the role in circadian rhythms, as reported by chronotype, on the areas of mental health and social support. The results suggest that being a "night owl," or having an evening chronotype, may influence mental health outcomes. The combination of self-reported traits in "Manic Symptoms: all of the above" include feeling more active than usual, getting less sleep, being more talkative, and having more creative ideas than usual. When modeling these traits during hyper or manic episodes alone, sleeping less than usual was also influenced by an evening chronotype. While it has been argued that chronotype and sleep duration are generally independent in a healthy population (Roenneberg et al., 2007), the 43,531 individuals who answered these questions in the UKBB were directed to because they have either been highly irritable and argumentative or manic for a period of two days or more in the past. These may be seen as signs of a diagnosed or undiagnosed bipolar disorder, characterized by periods of hyper and hypo activity. Melo and colleagues recently reported that an evening chronotype is more prevalent in adults with biopolar disorder (Melo et al., 2017), and this chapter adds weight to the hypothesis that circadian biology may contribute to this disorder.

Both measures of depression increased with an evening chronotype, though a self-reported episode of depression had a much larger slope, Figure 3.12 A. Sensitivity analyses

suggest little dominance of any one SNP in the model (B), and a funnel plot (D) suggest any pleiotropy may be balanced across conditions. Though not all SNPs have significant Wald ratios (C) individually, collectively when meta-analysed the effect is significant. While these findings echo previous reporting (Jones et al., 2019), there is a distinction between general measures of anxiety and depression which are associated with an evening chronotype, and the observation than an 'evening person' is more likely to experience a significantly sustained episode of depression. Measures of self-appraisal were tested for in the UKBB. These included having easily hurt feelings and reflecting on embarrassing moments after the fact for a self-described too long period. Both of these studies suggested, via measures other than IVW, that an evening chronotype leads to a decrease in self-reporting these feelings. Self-reported feelings of irritability increased when an evening chronotype was reported, and while most measures were in agreement over the direction of effect, the MR Egger method changed sign, Figure 3.15. Investigating returned plots, it is clear that four SNPs are strong IVs and influence the model to be positive, seen graphically in panel B, though one single SNP does not dominate the model, as seen in the leave-one-out sensitivity study in panel C. The funnel plot (D) suggests a fairly balanced pleiotropy. Overall, while irritability may be caused in part by circadian biology, these findings suggest that the effect is small and additional investigations are warranted before truly judging that chronotype causes the trait.

Of the social support measures studied, there was an increase in attendance at group religious activities at least once a week associated with being a morning person. This effect was quite pronounced, yielding the smallest p-value of any study in this chapter, 4.20e-5. This very significant effect is tempered by the high degree of heterogeneity among the SNPs studied, though not one SNP truly changed the outcome of the model during sensitivity tests (see Figure 3.16 C). There was little evidence of pleiotropy, as the Egger intercept was not significantly different than zero. Nevertheless, there are strong biases to this study design, and it may be that the questions I would ask about the role of chronobiology in

religious participation cannot be answered by the UK Biobank data. For instance, what about non-group religious activity - are morning people still likely to participate? There is an extreme lack of literature on the influence of circadian biology and religion. One recent study investigates the effect of religious participation on diurnal salivary cortisol patterns (Merritt and McCallum, 2013), finding an increase in religious coping mechanisms correlate with increased stress among caregivers for dementia patients. Other literature focus on the interplay between metabolism, circadian biology, and fasting during Ramadan (Chaouachi et al., 2009). This highlights the largest potential source of bias with this study - the mostly white, middle age UK population in this study, who participate in group religious activity, are not as likely to be Muslim as the are to be Christian. Viewing these results as tantalizing but suspect, I must ask if these associations between morning chronotype and religion would hold if the study population practice group activities Friday afternoons instead of Sunday mornings. The decrease in frequency of self-reported visits with friends and family was associated with an evening chronotype, though a look at the scatter plot of results (Figure 3.17 A) do not indicate a strong agreement among methods (again, the Egger method is reversed), nor do they suggest a large effect, as the slopes are nearly flat, and robust methods loose any significant signal. This study highlights several pitfalls when performing a (restricted) phenome-wide study using MR methods. Each study includes multiple analyses, and simply looking at p-values does not tell a whole story. The multiple testing burden on these analyses is immense, but for my purposes p-values are more of a guide than a definitive answer. The low power of MR studies, especially when multiple instruments are used, suggests that p-values should assist in interpretation of studies but a Bonferroni correction may be to strict (McShane et al., 2019; Burgess et al., 2015). Additionally, the coding in the UK Biobank for this trait is conter-intuitive, as a higher 'number' on the ordinal scale does not indicate increased frequency of family/friend visits. Thus, a scientist must pay attention to the coding of traits when using such a public resource.

Finally, I studied the causative role of chronotype on keratometry measurements. Keratometry, or opthalmometry, was accessed in the UK Biobank on over 100,000 participants (Chua et al., 2019). To date, there has been one genome-wide association study performed using any measurements from this field, investigating loci involved in corneal and refractive astigmatism, common to myopia in general (Shah, Guggenheim, and UK Biobank Eye and Vision Consortium, 2018). As the keratometry data are currently used by specialists to address specific questions after deriving new measurements relating to astigmatism from these data, I did not set out to address the role of chronotype in astigmatism itself. The fact that an association was found in the limited number of parameters from this field I tested is in itself interesting and worthy of follow-up by; as light is the primary zeitgeber responsible for circadian entrainment, eye physiology plays a crucial role in chronobiology. This study highlights that there may be bi-directional causation which should be investigated, possibly in mouse and other model organisms.

Each analysis in this study was tested via several methods, from those with the most power (IVW) to others with fewer assumptions about the strength of instrumental variables (Weighted Mode). A strength of this study was that each meta-analytic model of the causal influence of chronotype on a trait was tested with each method and investigated for possible evidence of pleiotropy with the MR Egger method. Vertical pleiotropy, in which SNPs influence multiple traits because of their influence on circadian rhythm as reported by chronotype, is an assumption of Mendelian randomization. Horizontal pleiotropy, where SNPs independently influence multiple traits, leads to the assumptions of causality breaking down (Davey Smith et al., 2020). In this study, although the Egger method can provide an estimate of causal effect robust to such pleiotropy, the presence of pleiotropy was accessed by the intercept. If significantly different than zero ($p < 0.05$), then evidence suggests a lack of unbalanced pleiotropy which would bias the model. In a two-sample MR study, weak instrument bias attenuates results towards the null as SNP-exposure and SNP-outcome

estimates are derived in non-overlapping samples (Burgess and Thompson, 2017). If I had performed a one-sample analysis with two-stage least-squares regression, this would not have been the case, and bias would have artificially inflated the effects seen away from the null hypothesis. Not only do multiple methods add strength to this chapter, but including multiple SNPs is a strength as well, since complex traits such as depression and participation in leisure activities are influenced by many weak effects in many SNPs, part of the omnigenic- and polygenic hypotheses (Boyle, Li, and Pritchard, 2017; Wray et al., 2018a). Though this model is under debate, the multiple loci identified by the chronotype GWAS, Figure 3.5, indicate that no single locus is responsible for chronotype. To reduce the heterogeneity of this study, I pruned SNPs in strong LD with each other. While this only left 18 instruments for this study, they were proxies for the influence of SNPs in each locus acting in tandem. Including all SNPs may have lead to stronger conclusions and larger effect sizes, but would have increased the variance of SNPs and therefore the heterogeneity's of the models. Lastly, the direction of effect for each study was correct, as verified by the Steiger tests; all of which were significant in the correct 'direction.'

## 3.5 Conclusions

I found some evidence that chronotype has causal adverse effects on mental health (increased depression, variety of manic symptoms, sensitivity to hurt feelings, irritability), and is associated with fewer group religious activities but more frequent association with family members. Analyses were largely robust to pleiotropy, with several methods applied to each analysis agreeing in direction and magnitude. Nevertheless, results should be treated with caution, and follow-up studies with large GWAS, such as the Million Veterans Program, will allow for confirmation of these findings before any experimental follow-up may be warranted (Gaziano et al., 2016). This study highlights how genetic epidemiology and social questionnaires can

be combined to answer questions about the causal relationship between circadian rhythm, a potentially modifiable exposure, and outcomes relevant to personal and social well-being.

## 3.6 Chapter Summary

In this chapter, I applied MR methods to investigate the role of circadian chronotype in a series of mental health, social support, and eye physiology traits with the UK Biobank. The results of these three studies suggest that evening chronotype affects several self-reported mental health traits relating to depression, manic symptoms, and irritability. Additionally, significant findings included a link between chronotype and two social measures - group religious participation and interactions with friends and family outside the home. Lastly, a relationship between chronobiology and the refractive index of the cornea indicate a two-way relationship worth of further experimental study. For each of the mental and social studies, verification outside the UK Biobank will provide additional evidence, as self-assessments of mental and social health may be culturally biased. As chronotype is a potentially modifiable trait, these studies suggest possible routes of therapy for depression and and other mental health outcomes. Overall, this study demonstrates how MR and causal analysis can be applied to test the effect of circadian biology on neurobehavioral traits. Several of these associations can be investigated in the mouse to test a mechanism of vertical pleiotropy between circadian biology and other behavioral outcomes.

Here I investigated the relationship between a genetic exposure to evening chronotype and several complex neurobehavioral traits, from episodes of depression to self-reported visits to friends and family. Many of the outcome traits studied were interrelated, including multiple measures of depression or surrounding manic episodes. To study the genetics of traits with multiple presentations, it is useful to break these traits down into observable con-

stituant parts. In the next chapter I create a biomedical ontology to model the relationships between pschological behavioral traits in populations who are aging, have schizophrenia, or are on the autism spectrum. The behavioral traits modeled here move beyond chrontype to other manifestations of aberrant behavior.

# Chapter Four

# Semantic Modeling of Neurobehavioral Phenotypes

This chapter is based in part on the following publications:

Martínez-Santiago, F., García-Viedma, M. R., **Williams, J. A.**, Slater, L. T. & Gkoutos, G. V. Aging Neuro-Behavior Ontology. Applied Ontology 15, 219–239 (2020).

## 4.1 Background and Chapter Overview

From findings in Chapter 3, it is clear that circadian biology has the potential to affect many measurable psychosocial traits, from religious participation to depression. In particular, links between circadian biology and developmental disorders, such as autism (Hu et al., 2009; Carmassi et al., 2019; Ballester et al., 2019; Tordjman et al., 2015; Manning, O'Roak, and Babur, 2019; Manning, O'Roak, and Babur, 2019), and psychiatric disorders, such as schizophrenia (Oliver et al., 2012a; Mansour et al., 2009; Kishi et al., 2011; Jones and Benca, 2015; Karatsoreos, 2014b), are abundant. Far from being homogeneous single traits, which

are easily studied, each disorder is on a spectrum, namely the autism spectrum (ASD) and schizophrenia spectrum (SSD). To study genetic relationships between autism and circadian biology, for instance, it may be helpful to segregate patients by the various phenotypic traits they manifest - identifying them where they are on the spectrum as individuals. As a first step towards this goal, this chapter proposes to model traits related to ASD and SSD using biomedical ontologies. Then, I will demonstrate the utility of ontological modeling using a cohort of autism patients, while continuing to model diverse sets of behavioral processes with ontologies relating to SSD and cognitive decline in elder patients.

### 4.1.1 Biomedical Ontologies

In 2009, Levitis (Levitis, Lidicker, and Freund, 2009) and colleagues conducted surveys and proposed that behavior is the 'internally coordinated responses (actions or in-actions) of whole living organisms (individuals or groups) to internal and/or external stimuli, excluding responses more easily understood as developmental stimuli.' Such a working definition points out how broad the domain of behavior is. In psychiatric genetics, observed behaviors may be classified differently depending on particular settings. During a typical clinical encounter, the behavior observed may be a presenting sign as well as a phenotypic presentation that could account for a larger part of a particular syndrome or form particular disease's symptoms. Endophenotypes may be considered a subclass of phenotypes in general, as they have links to a genetic locus and are not dependent on a patient's transitory state (Gottesman and Gould, 2003). This chapter uses biomedical ontologies to model neurobehavioral traits and endophenotypes relating to three domains of interest, namely autism, schizophrenia, and cognitive decline. To enable this translation, biological processes and observable phenotypes are represented as logical constructs known as ontologies. While an exact definition of ontology in information science is beyond the scope of this work, a working definition is

that an ontology is a descriptive logic knowledge base (Baader, Horrocks, and Sattler, 2008). Description logic is a subset of first-order logic that is decidable, that is implicit knowledge is able to be correctly inferred from axioms, in this use case via subsumption. Ontologies are composed of an ABox (axioms stating facts about the world), a TBox (axioms pertaining to the terminology of the domain of interest). The TBox may sate that every mouse is a mammal, and an ABox may state that an individual is a mouse. In the ontologies used in this chapter and following, the TBox, which pertains to classes of an ontology and not instances, is used. Ontologies are often thought of as knowledge graphs, and can be depicted as directed, acyclic graphs (DAGs), with entities (phenotypes, for example) connected via relations (inter alia part-of, is-a). Ontologies typically exhibit the following four features (Hoehndorf, Schofield, and Gkoutos, 2015):

- Classes and relations
- Domain vocabulary
- Metadata and descriptions
- Axioms and formal definitions

A class is an entity which refers to a set of entities, for example 'mouse' which includes all mice of all strains extant in the world. Classes are intentionally defined, as opposed to an arbitrary set. Relations facilitate describing two different entities. An example relation, 'mitochondrial membrane' in the Gene Ontology (GO)(Ashburner et al., 2000) is *part of* some 'mitochondrial envelope' and *is a* 'organelle membrane'. Both *part of* and *is a* are relations. In addition, standard identifiers facilitate computational integration of ontologies across databases. An example of standard identifiers in use: in the Neuro Behavioral Ontology (Gkoutos, Schofield, and Hoehndorf, 2012), ataxia is represented by the identifier NBO:0000590. Included in this definition is HP:0001251, the identifier of ataxia in the Human Phenotype Ontology

(Robinson et al., 2008), and an the analogous mouse phenotype MP:0001393 (Smith and Eppig, 2009).

Domain vocabulary concerns the list of terms associated with an ontology's classes and relations (above). A class identifier may be identical to a domain vocabulary item. In anatomical ontologies, there may be concurrent descriptions of an anatomical entity. The cerebellum has been termed the 'corpus cerebelli' in human and 'parencephalon' in rat (Swanson, 2003; Truex and Carpenter, 1996). Having a controlled domain vocabulary is particularly useful in an international scientific world, where multiple vocabulary terms can identify the same class.

Metadata and descriptions enable domain experts using an ontology to understand the precise meaning intended when a class was created. Curators of the GO added the following definition to the 'circadian rhythm' term: "Any biological process in an organism that recurs with a regularity of approximately 24 hours." This enables a developmental biologist without neuroscience expertise to understand what the class means in context.

In addition to textual descriptions which are human readable, ontologies include axioms and formal definitions. These machine-readable definitions enable computation over ontologies, such as automated reasoning and integration. Ontologies are commonly represented as directed acyclic graphs, which facilitates probabilistic methods over graphs and storing ontologies as graph data structures.

Behavior is currently modelled in several ontologies, the most widely used of which is the Neuro Behavior Ontology, or NBO (Gkoutos, Schofield, and Hoehndorf, 2012).

### 4.1.2 NBO Model of behavior

The NBO is composed of two branches, each an ontology interacting with the other, Figure 4.1. The first is the Behavioral process (NBO:0000313) branch, which classifies processes in which an organism or group of organisms is involved, and is an extension of the Gene Ontology's Behavior class (Gene Ontology Consortium, 2015). Behaviors are conceptualized on three main axes of classification:

- Response: behavior processes are considered responses to external stimuli, and the "in-response-to" relation axiomatizes this response. For example, visual behavior is in response to some visual perception.
- Intentionality: Intentionality concerns what a behavior is directed towards. An aggressive child's behavior can be directed towards peers or parents; this can be differentiated in NBO with the "is-about" relation.
- Means: Behavior, as a response to an external stimulus, must be mediated by physical attributes. Visual behavior occurs "by-means-of" some visual system.

Having axiomatized behavior processes, the second branch of the NBO is the behavior phenotype domain. Most phenotypes encoded in NBO are manifestations of behavioral processes described in the Behavior Process domain. Phenotypes, or phenotypic traits, are observable characteristics of an organism. Phenotypes encoded in other ontologies, such as Disinhibition (HP:0000734), are translated into NBO equivalents. Taking an example of interest to this chapter, forgetfulness is a phenotype defined by the loss of information already encoded in long-term memory. This is classified in NBO as NBO:0000606, and relates to memory loss behavior in the process arm of the NBO structure. Additionally, forgetfulness is a subclass of cognitive behavior phenotype, indicating that it is also related to cognitive processes.

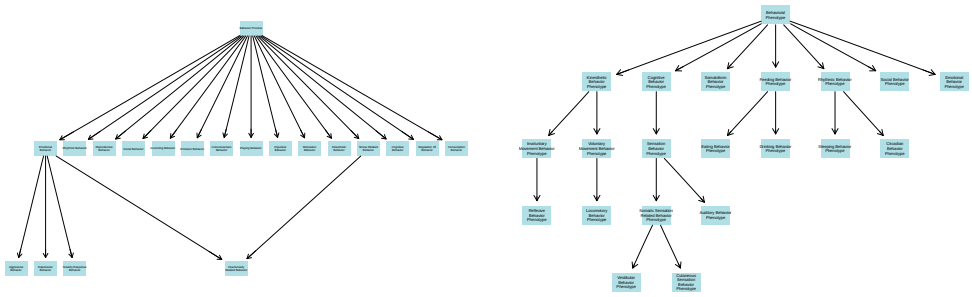
To facilitate interoperability between domains, various strategies have been implemented. Ontology aligners operate in a variety of ways to find terms in one ontology which are equivalent to another (Faria et al., 2013). Alignment may be atomic matches, where terms are logically equivalent, or associated with a degree of probability. To aid in translating phenotypes from one species to another, it is useful to reduce complex phenotypes into their constitute components (Malsen et al., 2011). The Phenotypic Quality Ontology (PATO) was implemented to decompose phenotypes into constituent parts based on entity-quality (E-Q) relations (Gkoutos et al., 2009). For example, 'amyotraphy' in the HPO is an intersection of the term 'atrophied' from PATO and *inheres in* muscle, muscle being a class of the Functional Model Anatomy ontology (Rosse and Jr, 2008). The NBO uses the EQ relations imported from PATO to define the behavioral phenotype branch of NBO. The term forgetfulness, then, is equivalent to:

```
has_quality some (
increased tendency and
towards some memory loss behavior)
```

where "increased tendency" indicates an increased likelihood of participating in (towards relation) memory loss behavior.

### 4.1.3 Phenotypic Representations of Autism and Schizophrenia

Autism spectrum disorder is a neurodevelopmental disease defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) as including impairment in social interaction and communication, and restrictive repetitive behaviors (Association, 2013). Defects in social communication or interaction may include verbal or non-verbal behaviors, from abnormal eye contact to a lack of facial expression. Restricted and repetitive behaviors may



(a) (b)

Figure 4.1: *The Neuro Behavior Ontology (NBO) is structured with two top-level domains, a "Behavior Process" and a "Behavioral Phenotype" domain. The behavior process domain models phenotypic processes, such as a learning behavior. Behavioral phenotypes are pre-composed from behavioral processes. All learning behavior phenotypes, for example, 'participate in' some learning behavior process.*

be stereotyped movements, adherence to ritualized patterns of routines, or rigid thinking patterns. In addition to these example characteristics, the DSM-5 lists hyperreactivity to sensory input as a type of restricted or repetitive behavior pattern.

Diagnostic criteria for schizophrenic spectrum and other psychotic disorders, as indicated by the DSM-5, include delusions, hallucinations, disorganized speech, disorganized or catatonic behavior, and the presence of negative symptoms (avolition, diminished emotional expression). To be diagnosed, patients must also have disrupted areas of life, including work, self-care, or interpersonal relations. Related syndromes, including depressive or bipolar disorder, must be ruled out. Importantly, there is large phenotypic overlap between ASD and SSD - and if there is a history of ASD, then a diagnosis of SSD is only warranted of prominent delusions or hallucinations occur for a period of time (Association, 2013).

### 4.1.4 Current Ontological Representation of Psychological Disorders

Psychological and behavioral disorders are described in several ontologies.

#### Neuro Behavior Ontology

The NBO currently describes psychotic disorder as a top-level behavioral phenotype, directly under the "Behavioral Phenotype" class. Psychotic disorders are further broken down into five subtypes via "is-a" relations to psychotic disorder, including catatonic, disorganized, paranoid, residual, and undifferentiated schizophrenia. Other disorders are present on the same level of the ontology, including dissociative disorders, mood disorder (including bipolar and depressive disorder, and substance induced mood disorders). No representation of autism itself, as an entity, exists in the NBO.

#### Human Phenotype Ontology

The Human Phenotype Ontology (Robinson et al., 2008) (HPO) contains an "Behavioral Abnormality" class, among whose children are "Autistic Behavior." The HPO uses this class to refer to autism spectrum disorder, which can be part of a disease as a phenotypic feature, see Figure 4.2. Immediate child classes, connected by "is a" relations, include hallmarks of autism related traits, such as restrictive behavior and impaired social interactions, as well as additional classes denoting "Autism with high cognitive ability" and "Autism." Additional phenotypic traits are part of the HPO's depiction of autism. Additionally, schizophrenia is defined as a mental disorder characterized by disintegration of thought processes and emotional responsiveness. No child terms of schizophrenia are provided. Biologically closely related terms to SSD include psychosis and its child terms psychotic mentation, and psychotic

episodes with its child transient psychotic episodes. SSD frequently manifests delusions, hallucinations, and paranoia; all of these are in the HPO but do not have any relations to the class "schizophrenia".

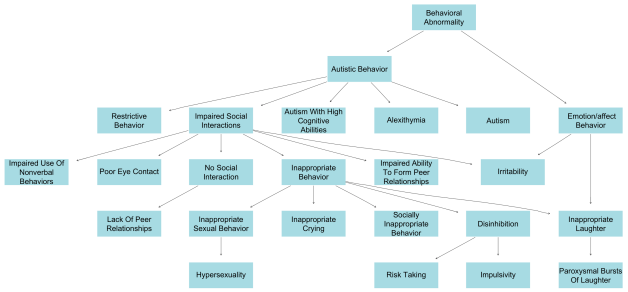



Figure 4.2: *Autistic behavior is modelled in the Human Phenotype Ontology as a type of behavioral abnormality, with many subclasses connected by "is a" relationships.*

### Mental Disease Ontology

The Mental Disease Ontology (Ceusters and Smith, 2010) is an extension of the Mental Functioning Ontology which describes various mental disorders. They designate autism spectrum disorder (and children autistic disorder, atypical autism, and Asperger syndrome) as subclasses of "pervasive developmental disorder," but describe no ASD-related traits. SSD is described with child classes as in the NBO, as a subclass of psychotic disorder.

### Autism and Schizophrenia Specific Spectrum Ontologies

Three autism-specific ontologies have been developed in recent years. Tu et al (Tu et al., 2008) created an autism phenotype ontology following OBO Foundry principles, and comprises 34 classes which represent high-level phenotypes derived from autism diagnostic instruments, and four classes representing the instruments used. The ontology was originally developed for use with the National Database for Autism Research to enable the consor-

tium to organize diagnostic data, and allow researchers to compare diagnostic instruments. The ontology contains SWRL rules to facilitate annotating hypothetical patient traits to the classes in the ontology (Horrocks et al., 2004). More recently, McCray and colleagues created the ASD Phenotype ontology (McCray, Trevvett, and Frost, 2014). As part of the Autism Consortium, their ontology encodes 283 terms across three classes: Personal Traits, Social Competence, and Medical History, including 97 personal traits and 72 social competence entities. Their goal was to integrate the ontology into the Boston-based Autism Consortium database, which no longer exists. From the Autism Diagnostic Inventory, Revised (Lord, Rutter, and Le Couteur, 1994a), they mapped 62 concepts relating to personal traits, and 25 relating to social competence. The ontology has a maximum depth of five and a an average number of siblings of 4. A strength of this ontology is that the compilers were able to map diagnostic criteria from 24 instruments, including a reported 1,883 questions mapping to 97 personal traits, and 931 questions mapping to 72 social competence traits. The bulk of their ontology mapped to medical history questions, not behaviors relating to autism. Lastly, Mugzach et al (Mugzach et al., 2015) combined elements of the ontologies of Tu and McCray into the Autism DSM-ADI-R ontology, or ADAR. ADAR adds SWRL rules to Tu's ontology to infer traits from ADI-R items, and incorporates the phenotype class hierarchy of McCray. Unlike Tu and McCray, their goal was to test the ability of the ADI-R to diagnose individuals with autism using both the DSM-IV and DSM-V criteria. They map coding from the ADI-R directly onto the ontology - for instance, a basic class they encode is "ImaginativePlay-NotAvailable", indicating that an individual assigned to their ontology using SWRL did not have an answer to the question, denoted by a score of 8 for the related question in the ADI-R.

At the present time, no schizophrenia specific trait ontology exists, to the author's knowledge.

### 4.1.5 Chapter contributions

In this chapter, I present ontological modeling of behaviors relating to three complex behavioral domains: schizophrenia, cognitive decline, and autism, each extensions of the Neuro Behavior Ontology. In the cognitive decline domain, I show the integration of pre-existing ontologies for use in modeling behavior, currently being undertaken by colleagues. I explore the phenotypic overlap between schizophrenia and autism, before exploring the diversity of phenotypic presentations of children on the autism spectrum. I then use data from deeply phenotyped people on the autism spectrum to demonstrate the utility of the Psychological Neuro Behavior Ontology (PNBO), the primary output from this chapter, in segregating subjects based on their presenting phenotype into data-driven subgroups based on semantic similarity.

## 4.2 Methods

The ontology build in this chapter, and semantic profiles derived from it, draws heavily from data provided by two diagnostic questionnaires, the Autism Diagnostic Interview - Revised (ADI-R) (Lord, Rutter, and Le Couteur, 1994b), and the Positive and Negative Symptom Scale (PANSS) (Kay, Fiszbein, and Opler, 1987a). Additionally, work with collaborators was essential in building the Aging Neuro Behavior Ontology (ANBO) (Martínez-Santiago et al., 2020). The data used to derive traits from the ADI-R was provided by the Simons Foundation for Autism Research, using data collected for the Simons Simplex Collection. This study was approved by the University of Birmingham's ethical review committee, ERN-17-0879, and assigned SFARI project number 2720.1.

### 4.2.1 Phenotype Extraction from the Simons Simplex Collection

The SSC version 15 dataset cohort contains 2644 families, each with two biological parents, one proband diagnosed with autism spectrum disorder, and at least one unaffected sibling for a total of 10474 individuals, 2292 male probands and 352 female probands (Fischbach and Lord, 2010b). Probands were deeply phenotyped with a variety of medical and behavior related inventories, including the aberrant behavior checklist (Aman et al., 1985), the autism diagnostic observation schedule (ADOS) (Lord et al., 1989), and the ADI-R. The ADI-R was built to diagnose children with ASD, and incorporates questions appropriate for both verbal and nonverbal children. The ADI-R contains 93 questions mapping to several domains, each of which is used to quantitatively diagnose a child wiht ASD. The social domain includes questions relating to peer relationships, nonverbal communication, sharing one's enjoyment with others, and socioemotional measures. The communication domain includes aggregate scores for verbal and non-verbal communication, delayed language learning, ability to make conversation, idiosyncratic speech, and a lack of make believe or imaginative play. The repetitive/restrictive behavior domain contains scores from repeated patterned behavior, compulsive rituals, repetitive mannerisms, and preoccupation with material objects. Children with high scores derived from each domain can be reliably diagnosed with ASD, using an algorithm provided with the instrument (Lord, Rutter, and Le Couteur, 1994a). Using data provided with the SSC, I first identified behaviors relating to the ADI-R to the "behavior process" arm of the NBO ontology, and I then manually added those which did not exist in the current NBO hierarchy. Next, I derived traits from the ADI-R, and mapped those to the "behavioral phenotype" arm of the NBO. Again, where traits did not exist, they were added.

### 4.2.2 Categorical Data Extraction from PANSS

The positive and negative symptom scale (Kay, Fiszbein, and Opler, 1987b) was incorporated into the PNBO in a similar manner as in with the ADI-R. Items in the instrument were first mapped to the 'behavior process' domain, then composed into traits in the 'behavior phenotype' domain of the ontology. Where traits overlapped existing NBO/PNBO classes, they were imported directly into the ontology. As no patient data was available, no individual subject annotations were created.

### 4.2.3 Curating the NBO and creating the PNBO

Once data from each diagnostic instrument was obtained and incorporated into the NBO, several modifications to the NBO were made. Diagnoses of psychological disorders were removed from the "behavior process" arm of the NBO, and either remodeled as phenotypic traits, or discarded. Next, the "object" class from the Relation Ontology (Smith et al., 2005) was added to the NBO, to allow classes to be "about" using an external object. Next, the OntoFox tool (Xiang et al., 2010) was used to extract those portions of the NBO, including classes, relations, and annotations, of the NBO "behavioral phenotype" from the parent class down to each class added from the ADI-R and PANSS. The "behavioral process" domain was left in its entirety to facilitate future expansion of the ontology. Lastly, the extracted classes and relations from each domain were entered into a new ontology, the Psychological Neuro Behavior Ontology, or PNBO.

### 4.2.4 Modeling the relationship between ASD and SSD phenotype

In order to model the relationship between endophenotypes of ASD and SSD as encoded in the PNBO, a graph was created to quantify the degree of overlap between traits. Traits derived from the PANSS and ADI-R were separated, including their ancestors through to the root term "Behavioral Phenotype". The intersection of the terms was then obtained, denoting which terms were exclusive to either survey. To hypothesize about the causal relationship between SSD and ASD conditions, regardless of endophenotype presentation, a Mendelian Randomization analysis was performed, suing the MR Base platform (Hemani et al., 2018) to obtain GWAS SNPs from non-overlapping populations. 73 SNPs were obtained from MR-Base curated summary statistics from a study of 35,476 SSD patients by the Psychiatric Genomics Consortium (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). These SNPs were harmonized with a meta-analysis of 18,381 individuals on the autism spectrum (Grove et al., 2019). Using the TwoSampleMR package, inverse-variance weighted, MR Egger, weighted median, and weighed median Mendelian randomization analyses were performed as in Chapter 3.

### 4.2.5 Modeling Behavior in the ANBO

Working with colleagues, I also created another derivation of the NBO, the Aging Neuro Behavior Ontology (ANBO). This ontology was built to provide a model of the aging process as it relates to activities of daily living (ADLs) and cognitive decline. To this end, the ANBO was built to provide a formal description of cognitive processes relevant to studying an aging population. NBO processes relating to executive function, perception, attention, memory behavior, long-term memory, and proxies were created. Unlike the PNBO, the ANBO is focused on the behavior process domain of the NBO only, and relies on the EQ components of PATO to post-compose phenotypes.

The ANBO is presented for use in conjunction with the OSLE, or Ontology SmartLab Elderly (Martínez-Santiago et al., 2020), an ontology built by my collaborators to model ADLs in a SmartLab environment, in which sensors are attached to devices such as medicine cabinets or televisions to track their use, and their use is compared to pre-planned daily routines modeled with the OSLE. Recorded daily activities in the OSLE are then mapped to the ANBO, to monitor fluctuations in cognitive processes related to cognitive decline. To build ANBO, I first mapped a minimal set of traits from the "Behavioral Process" arm of NBO using OntoFox (Xiang et al., 2010). Additional processes as proposed by collaborators were added. Cognitive processes were modelled by integrating the ANBO with functions from the OSLE using SWRL rules, which allow integrating descriptive logic with constraints to create new classes in the ontology. For example, to denote that a brother of one's parent is one's uncle, a rule may be created:

$$hasParent(?x1, ?x2) \wedge hasBrother(?x2, ?x3) \Rightarrow hasUncle(?x1, ?x3) \tag{4.1}$$

which denotes that if x2 is the parent of x1, and x3 is the brother of x2, then x3 must be the uncle of x1. SWRL rule creation, and all ontology development, were performed in Protege (Musen, 2015).

### 4.2.6 Reasoning and Completeness

After each ontology was built, it was subjected to reasoning to ensure its logical consistency. PNBO, ANBO, including the SWRL rules and collaborators' OSLE, were evaluated with the Hermit reasoner (Shearer, Motik, and Horrocks, 2008). Each ontology was constructed using the OBO Foundry design standards (Smith et al., 2007a).

### 4.2.7 Comparing Individuals via Semantic Similarity

Each proband in the SSC was assayed with the ADI-R, and a numeric score was provided for each question. Data are predominantly coded the following format:

- 0: Behavior not present
- 1: Behavior present in abnormal form, but not not severe enough to meet criteria
- 2: Definite abnormal behavior
- 3: Extreme abnormal behavior
- 7: Abnormality in the area of coding but unspecific
- 8: Not applicable
- 9: Not asked or unknown

Following the algorithm presented in (Lord, Rutter, and Le Couteur, 1994a), I assigned, for each question in the ADI-R, a proband a trait if their score 2 or 3, and all others were considered to be absent. Several of the 93 questions were repeated, asking if the proband *currently* or *ever* presented with the possible trait. If either of these traits were scored positively, they were combined into one trait. This resulted in a corpus of probands, each annotated to several traits in the PNBO ontology.

### 4.2.8 Calculating Semantic Similarity

Once annotated, probands were compared to each other by calculating the semantic similarity shared by their annotated traits. Semantic similarity measures, originally derived for text,

seek to quantify the specificity and commonality of information shared between two corpora. In population genetics, they are widely used alongside ontology-based gene set enrichment to characterise the functions of genes or the similarity of individual diagnoses. I used two similarity measures in this chapter, both based on the measure of the specificity of a class in the ontology. The graph-based structure of an ontology, and the (largely) *is a* relations between them, ensure that the class "delayed echolalia" is more specific than its parent class, "echolalia". This specificity can be calculated using Resnik's method (Resnik, 1995a):

$$IC_Resnik(x) = -logp(x) \tag{4.2}$$

where p is the probability of finding a term in a corpus (or in the ontology):

$$p(x) = \frac{|I(x)|}{|I|} \tag{4.3}$$

and can be calculated directly from the ontology, or from the corpus of terms annotated to the ontology. I calculated the IC of each term in terms of the frequency of a proband being annotated to a term. All annotations take advantage of the transitive property of the *is a* relation, so a proband annotated to the term "delayed echolalia" is also annotated to "echolalia" and all intermediate terms through to "behavioral phenotype", which will have an IC of 0.

To calculate the similarity between classes in the PNBO, I used two methods. Resnik's similarity measure (Resnik, 1995b):

$$Sim_{Resnik}(x, y) = IC(MICA(x, y)) \tag{4.4}$$

where IC is the information content and MICA is the most informative common

ancestor, or the parent of the two terms with the highest IC. Lin (Lin, 1998) modifies this measure:

$$Sim_{Lin}(x, y) = \frac{2 * IC(MICA(x, y))}{IC(x) + IC(y)} \tag{4.5}$$

to standardize the measure over the total information content of each term, x and y. This constrains the measure of similarity to be between 0 and 1, and enables classes that are more distantly located from their MICA to have a lower similarity. To compare sets of classes that were used when comparing sets of probands, I used the Best Match Average approach, which first calculates the maximum average similarity between sets of probands:

$$sim_{MA}(X, Y) = \frac{\sum_{x \in X} max_{y \in Y} sim(x, y)}{|X|} \tag{4.6}$$

where the maximum similarity of any term in proband set Y against each term for proband set X is calculated, and standardized.

The best match average takes the average of the maximum similarity between sets of probands, as they are not guaranteed to be symmetric:

$$sim_{BMA}(X, Y) = \frac{sim_{MA}(X, Y) + sim_{MA}(Y, X)}{2} \tag{4.7}$$

All semantic similarity calculations and ontology plotting were performed in R v3.5, using the OntologyPlot and Ontology Similarity packages (Greene, Richardson, and Turro, 2017).

### 4.2.9 Semantic Clustering of SSC Proband Annotated Traits

To compare probands to each other based on their shared or differential PNBO traits, a pairwise semantic similarity matrix was created for each proband using a Resnik-based BMA approach. Next, hierarchical clustering was performed using the WGCNA package in R (Langfelder and Horvath, 2008) in the following manner. First, the normalized similarity matrix was transformed into a weighted adjacency matrix. Adjacency matrices represent graphs in the form of:

$$a_{i,j} = \begin{cases} 1 \text{ if } sim_{i,j \geq \tau} \\ 0 \text{ if } sim_{i,j < \tau} \end{cases} \tag{4.8}$$

where *sim* is an similarity matrix, usually from correlation between entities but in this case the semantic similarity matrix calculated. Rather than having a thresholding function $\tau$ to constrain the adjacency matrix entries to only 0 or 1, a soft-thresholding was performed:

$$a_{i,j} = |simi, j|^{\beta}, for \beta > 1 \tag{4.9}$$

were $\beta$ was set at 6, which forces less-similar pairs of probands to have much weaker connections in the adjacency matrix. The adjacency matrix was used to calculate the weighted connectivity, or degree, of each node (proband) in the graph.

Next, a topological overlap matrix was created from the adjacency matrix to capture the similarity of nodes: (Zhang and Horvath, 2005):

$$\omega_{i,j} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{min(k_i, k_j) = 1 - a_{i,j}} \tag{4.10}$$

and a dissimilarity matrix crated by $1 - \omega$. Hierarchical clustering was performed on the dissimilarity matrix, using average linkage. Hierarchical clustering results in a tree, or dendogram, representing the relationship between probands. To transform the dendogram into discrete clusters, a dynamic tree cutting algorithm was performed using the WGCNA package (Langfelder, Zhang, and Horvath, 2008). Rather than "cut" a dendogram once, an initial cut is made with a static tree cut, then each resulting cluster is analyzed for fluctuations in the its own subtree. Any clusters exhibiting such a pattern are recursively split, using a limit of 30 probands to indicate a floor for the recursive algorithm. See (Langfelder, Zhang, and Horvath, 2008) Supplementary Data for a detailed explanation.

### 4.2.10 Bootstrap Cluster Validation

To test the resulting clusters' stability, I performed a bootstrapping procedure, using the package NetRep (Ritchie et al., 2016). Two test statistics were investigated: the magnitude of edge weights in each module, and the magnitude of the semantic similarity of each module. To create test statistics, the dissimilarity matrix was permuted 10,000 times, and summary statistics created during each permutation, resulting in a null distribution. The statistics for each module were compared to the null distribution using a one-sided test, and the proportion of null tests more extreme than the observed test statistic provided empirical p-values. Those values were then adjusted for multiple testing by the False Discovery Rate, as in (Benjamini and Hochberg, 1995).

### 4.2.11 Bayesian Semantic Profile Regression

To investigate what underling phenotypic profile the module represents, an adaptation of Similarity Regression using the SimReg package (Greene, Richardson, and Turro, 2016) was

performed. Originally developed to find single SNPs causing rare disease based on the ability of an average semantic similarity profile from the human phenotype ontology to predict disease, this was adapted to predict cluster membership and view the underlying marginal probabilities of an ontology term being associated with cluster membership. Using Baysian regression, two models are compared:

$$y_i \; Bernoulli(p_i); \begin{cases} \gamma = 0\text{: } log(\frac{p_i}{1-p_i} = \alpha) \\ \gamma = 1\text{: } log(\frac{p_i}{1-p_i} = \alpha) + \beta S(\phi, x_i) \end{cases} \tag{4.11}$$

$y_i..y_n$ is a vector of cluster membership: 1 if a patient belongs to a cluster, and 0 if not. $x_i..x_n$ is the vector of phenotypes for each subject, where $xi$ is the minimal set of PNBO phenotypes requried to describe the proband. The set is minimal if and only if only contains directly annotated terms and not others implied by transitive relationships in the PNBO. For instance, if a proband is annotated with "delayed echolalia," their minimal set of traits will not include the parent term "echolalia."

Under the null model assuming $\gamma = 0$ (no association between a phenotype and cluster), *alpha* is the intercept, the proportion of cases in the cluster. Under the alternative, $\beta$ is the coefficient representing a unit increase in phenotypic similarity, using a modified Lin BMA measure, of the patient to a characteristic phenotype $\phi$, on the log-odds of being a member of the cluster. Therefore, the probability that $\gamma = 1$ is greater when the similarity of a proband's phenotype to a cluster profile is larger when if they are indicated as a member of that cluster.

To estimate the characteristic profile $\phi$, a Markov chain Monte Carlo procedure is undertaken under a uniform sampling of up to 2,000 sets of PNBO terms of size k = 5. A mapping function $\nu$ is applied to the vector of phenotypes to account for the fact that not all sampled PNBO sets will be minimal. The prior on the estimated $\widetilde{\phi}$, counting the number of times a trait appears in a sampled distribution under the $gamma = 1$ assumption

(the individual is in cluster membership). For a full description of the models used, and a detailed discussion of MCMC sampling procedures, see (Greene, Richardson, and Turro, 2016) supplementary material section 7. The conditional posterior distribution of each model (null and alternative) can be estimated from MCMC samples where $\gamma = 0$ and 1 and the posterior probability that the alternative model is true is estimated from the number of iterations in which $\gamma = 1$. The marginal posterior probability of a term's inclusion into *phi* for each model was retained, to derive the characteristic minimal phenotype of each cluster.

## 4.3 Results

### 4.3.1 PNBO and ANBO Structure

The PNBO includes 131 entities in the "Behavioral Phenotype" domain specifically to model ASD and SSD traits, including 103 novel classes specific to the PNBO. Including all NBO terms potentially relevant to psychological disorders on heterogeneous spectra and thus kept from the original NBO, statistics for the ontology are given in Table 4.1.

The ANBO is much smaller than the PNBO, can be visualized in Figure 4.3.

Both the PNBO and ANBO take advantage of several types of relations, among them the following:

- by means: A process x occurs by means of a material structure y if and only if x occurs by means of y.
- has participant: A relation between a process and a continuant, in which the continuant is somehow involved in the process

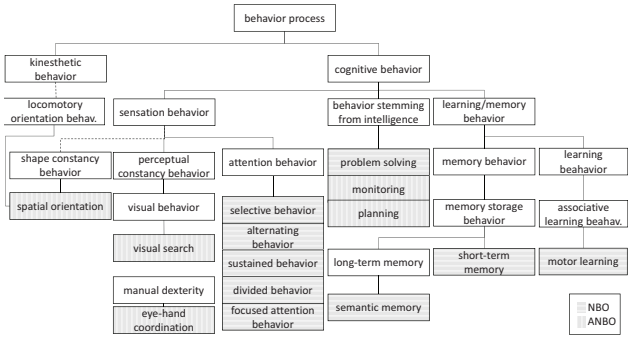



Figure 4.3: *ANBO class diagram. Gray processes are relevant to activities of daily living and are modeled in the OSLE. Horizontal shading indicates behavioral processes native to the NBO, while vertical shading represents processes added for the ANBO. Graphic originally published in (Martínez-Santiago et al., 2020)*

- in response to: Between a process x and a process y if and only if x occurs in response to y
- intersection of: With cardinality of zero or more, this indicates a term is the equivalent to the intersection of several other terms
- is a: a subsumption, indicating that x is a subclass of y, wherein x's specification implies y's specification
- is about: A process x is about some entity y if and only if x is about or directed toward y.
- has input: A phenotype has input a collection of entities with a given property regarding frequency, amount and so on
- part of: a core relation that holds between a part and its whole

### 4.3.2 Integrating the OSLE and ANBO Ontologies

As the majority of the work contained in this chapter involved the PNBO, results from the ANBO in which I played a major role will now be discussed. The ANBO combines behavioral processes in the NBO, entities and qualities from PATO, and anatomy from UBERON (Mungall et al., 2012). It aims to capture the conditions under which a process can be triggered, which results may be expected after a process, and any qualities which define normal or abnormal instances of that process. For example, "visual search" process, ANBO:0000004, is depicted in Figure 4.4. Visual search "is a" visual behavior, which "is about" a physical object quality, and is "in resonse to" visual perception, which is facilitated "by means" of an in-tact visual system. This visual search process (an entity) may have a quality of increased duration, in which case a slow visual scanning phenotype is manifested. Each behavioral phenotype was encoded to depend on high-level UBERON anatomy, and participate in behavioral processes while being denoted by certain qualities.

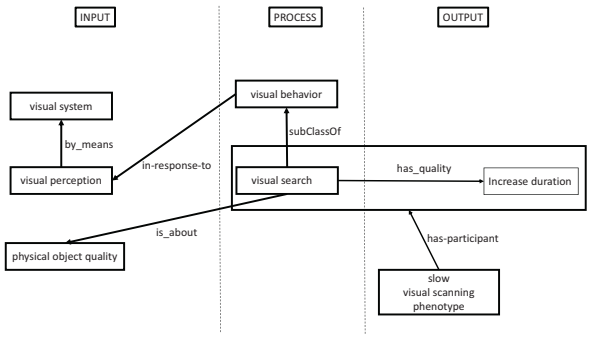



Figure 4.4: *In ANBO, individual classes are composed of input from an anatomical entity, a behavioral process in response to such input, and an expected output. The trait of "Visual Search", ANBO:0000004, is depicted here. Graphic originally published in (Martínez-Santiago et al., 2020)*

Collaborators used the ANBO ontology with the Telehealth Smart Home system to setup a hypothetical experiment, monitoring behaviors modeled with the ANBO by use of

sensors attached to objects involved in activities of daily living. SWRL rules were created for several senarios, as exemplified by by a subject getting a bottle of water, which is an ADL which, in part, has input from a visual search process modeled in ANBO Figure 4.5.

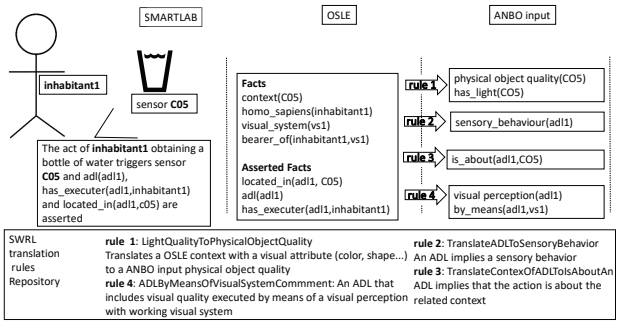



Figure 4.5: *The visual search process is modeled in SWRL rules when a hypothetical sensor (C05) is triggered by a subject obtaining a bottle of water on which a sensor rests. SWRL rules then take in facts that the sensor, when activated, must be operated by a subject with a visual system and who has participated in an assigned activity of daily living, visual perception, by means of ANBO behavior. Graphic originally published in (Martínez-Santiago et al., 2020)*

While the ANBO models activities of daily living in a population which may experience cognitive decline (the elderly), the PNBO models traits found among individuals among all walks of life, some of whom may be on the autism or schizophrenia spectra.

### Overlap between ASD and SSD Phenotypes in the PNBO

A high degree of overlap is apparent between traits in the PNBO derived from ASD and SSD diagnostic interviews, Figure 4.6. To the left, red children from the "Cognitive Behavior Phenotype" annotated from PANSS segregate onto one level. Further yellow children are Learning/Memory behavior phenotypes annotated from ADI-R. The bulk (middle) of the ontology is made up of yellow nodes from the ADI-R, while orange nodes are common to both ontologies, though they may not appear on the diagnostic instruments themselves and

are a product of "is a" relations in the ontology. Distinct traits from each instrument can be closely related: motor retardation (PANSS) is four classes away from abnormal gait (ADRI). Traits with similar names can also be only distantly related: stereotyped thinking (PANSS) is on the third level of the ontology DAG, while Stereotypic motor behavior is on the fifth, six classes away on the DAG, traversing "Behavioral Phenotype."

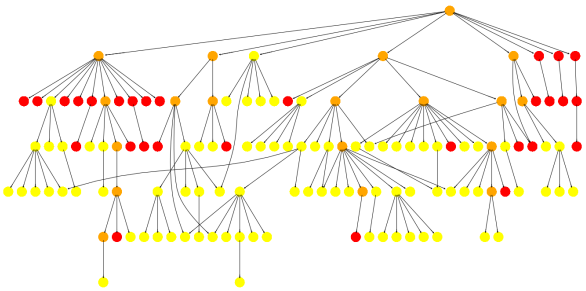


Figure 4.6: *The PNBO includes entities drawn from both the PANSS and the ADI-R, gold-standard psychological instruments designed to diagnose schizophrenia and autism respectively. The structure of the PNBO shows the high degree of overlap between traits manifested by individuals on each spectrum. Autism-specific traits are shown in yellow, schizophrenia specific traits in red, and traits in common in orange. Only the behavioral phenotype domain of the PNBO is shown.*

To query the potential causal relationship between exposure to SSD and autism, a strong association was identified between a genetic predisposition to SSD and a manifestation of ASD. Although there was high heterogeneity (Inverse variance weighted [IVW] Q statistic 174, p = 4e-07, MR Egger intercept p-value 0.06), all methods agreed in the magnitude of effect: IVW $\beta = 0.107$, standard error (se) 0.03, odds ratio (OR) = 1.1; MR Egger $\beta = 0.306$, se 0.11, OR = 1.35, . MR Steiger tests suggest a strong directionality from SSD exposure to ASD outcome (p = 1.e-132). See Figure 4.7.

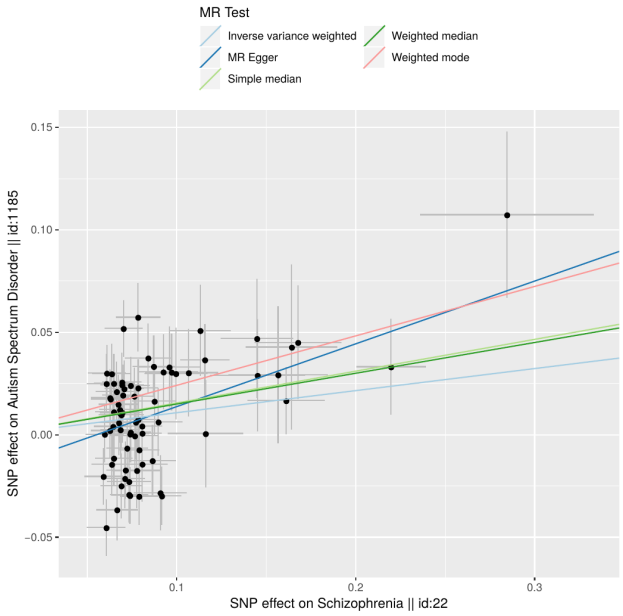



Figure 4.7: *To test the assumption that genetic exposure to schizophrenia may influence the likelihood of autism diagnosis, a Mendelian randomization analysis was performed. Results show the possible causal effect of schizophrenia associated alleles (x-axis) on the development of autism (y-axis). MR = Mendelian randomization. Scales represent the log-odds of phenotype per allelic dose.*

**Interoperablility between PNBO and ANBO with other ontologies**

Each class of the ANBO and PNBO uses entity-quality notation from the PATO ontology (Gkoutos et al., 2009) to enable translation between different ontologies. As an example, the PNBO trait "repetitive use of objects", PNBO:072, can be mapped nearly directly to the Mammalian Phenotype ontology term MP:0001409, "increased stereotypic behavior" as shown below.

```
PNBO_072:  "repetitive use of objects"

'has part' some (
    'increased rate' and
    'inheres in' some 'stereotypic behavior' and
    'has modifier' some abnormal and 'is about' some object)

MP_0001409: "increased stereotypic behavior"

'has part' some (
    'increased rate' and
    'inheres in' some 'stereotypic behavior' and
    'has modifier' some abnormal)
```

while not a perfect match, the EQ notation indicates that both traits are composed of an increased rate of some abnormal quality (when compared to a reference population) that 'inheres in', or belongs to or is a property of, stereotypic behavior. The EQ notation is extended for PNBO:072, by indicating that the behavior 'is about' some object. meaning the behavior is directed towards an external material object - often a toy or other object, in the case of SSC probands annotated with PNBO.

### 4.3.3 Distribution of SSC Phenotypes among Probands and Clustering of Traits

The frequency of trait annotations among probands can be seen in Table 4.2. The trait observed with the highest frequency was "group play with peers," indicating that all but five probands who were phenotyped and whose genetics passed QC did not play well with peers! Indeed, social response abnormalities are seen in the vast majority of probands. 265 probands were annotated with a musical ability phenotype, indicating that they have an increased ability to play music compared to their peers of a similar age. Clustering of traits reduced the dimensionality of the ASD-related traits studied, form 73 annotated in the PNBO to 14 clusters, Figure 4.8. Two of the largest clusters, turquoise and brown, appear nearly intertwined; while smaller clusters with lower-hanging features in the dendogram are more self-contained. Table 4.3 shows the permutation-based, empirical p-values for two

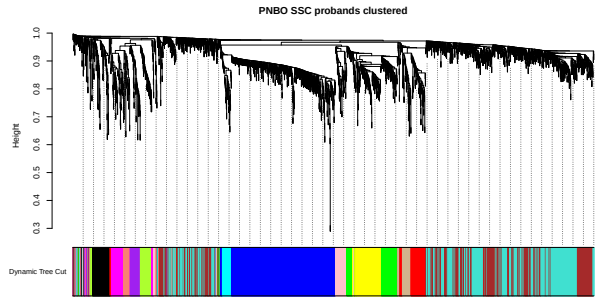



Figure 4.8: *Parents answers on the Autism Diagnostic Interview-Revised were used to model probands' traits in the Psychological Neuro Behavior Ontology (PNBO). A pair-wise semantic similarity score was calculated for each pair of probands using Resnik Best-Match-Average. The resulting matrix was clustered using hierarchical clustering and a dynamic tree cutting algorithm. Clusters are depicted by color.*

test statistics of module stability: the weighted degree of the nodes in the cluster, taken from the topological overlap dissimilarity matrix used as clustering input, and the semantic similarity-based degree, taken from the raw semantic similarity input. The constraints on the

dissimilarity matrix produce more stable modules, as indicated by lower p-values (column 1 vs column 2). The FDR corrected p-values for each test are given in columns 3 and 4. Note that the smallest possible p-value, based on 10,000 permutations, is 0.0001 - a frequently observed output, while the highest p-value, a null result of 1, is also observed for some clusters.

**Semantic Makeup of Phenotype Derived Clusters**

Semantic profiles of each cluster were generated from similarity regression, by combining the marginal probabilities of each PNBO trait's ability to model the likelihood of a proband in a given cluster having membership in only that cluster. Here, I outline the semantic profiles that typify each cluster, from largest (turquoise) to smallest (cyan) in membership.

Each cluster is made up of at least 48 probands, each annotated with a mean of 39 PNBO terms, see Table 4.4. The turquoise cluster, the largest, is associated with playing behavior phenotypes, as well as memory and skill phenotypes, Figure 4.9.

The blue cluster revealed only one minimal trait which could reliably characterise the cluster, abnormal reciprocal conversation. The brown cluster is broad, and encompasses strong associations with deficits in engaging in reciprocal conversation and learning behavior phenotypes, Figure 4.10.

The yellow cluster is very homogeneous, and indicates a combination of self injury, lack of interest in and of playing with other children, and few instances of using one's body to communicate, Figure 4.11.

The green cluster's signature is a loss of language ability, and abnormal social verbalization and response rates, Figure 4.12.

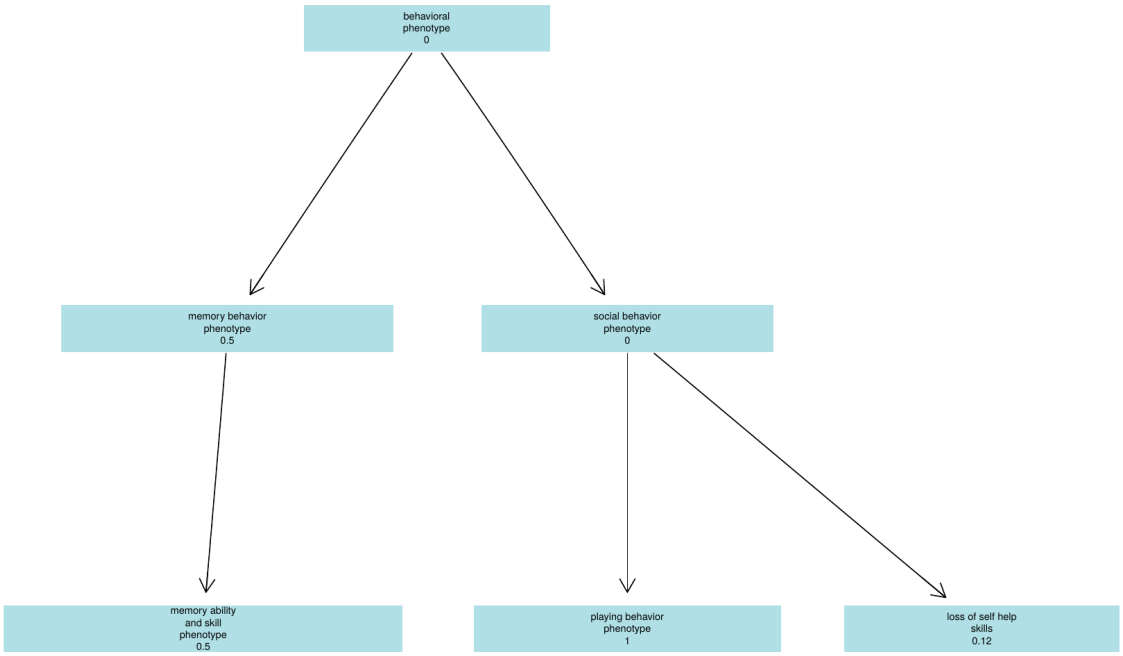



Figure 4.9: *Minimal sub-graph of the PNBO traits in cluster turquoise.*

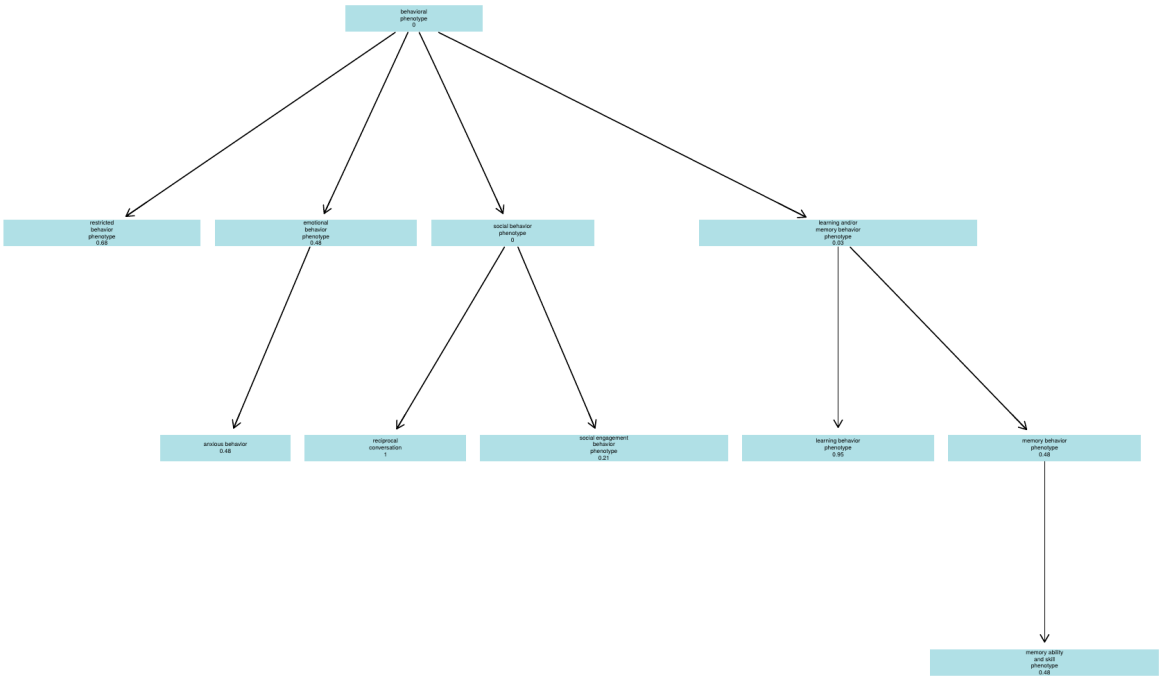



Figure 4.10: *Minimal sub-graph of the PNBO traits in cluster brown.*

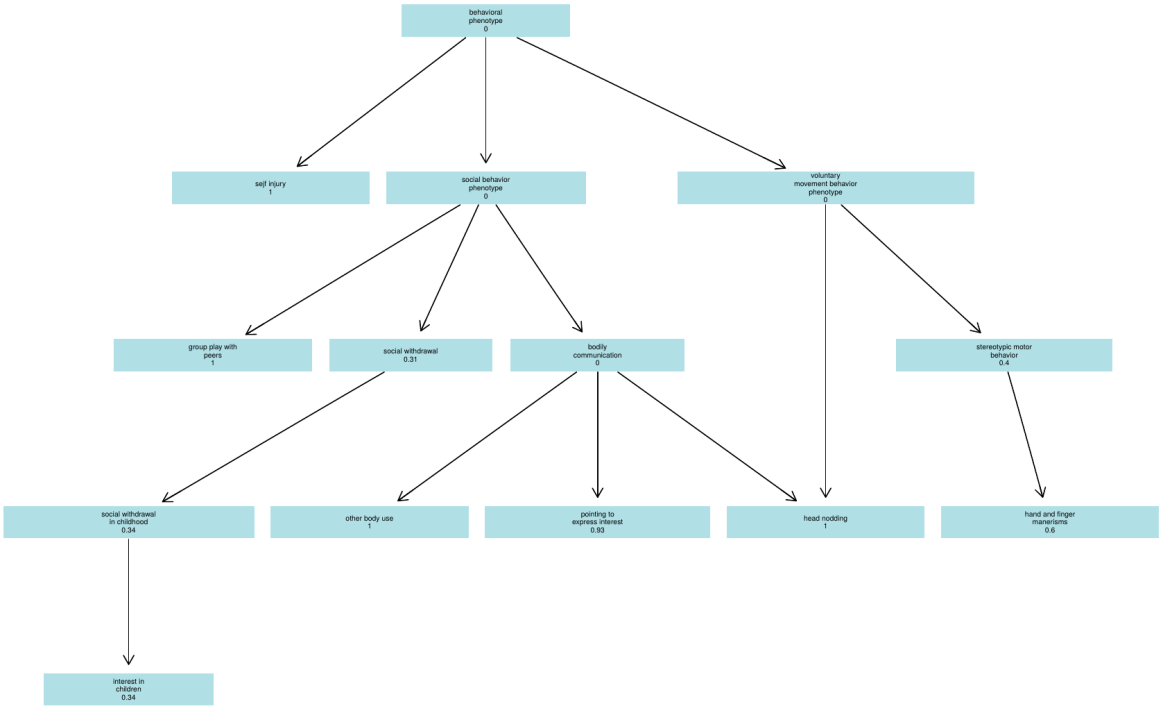



Figure 4.11: *Minimal sub-graph of the PNBO traits in cluster yellow.*

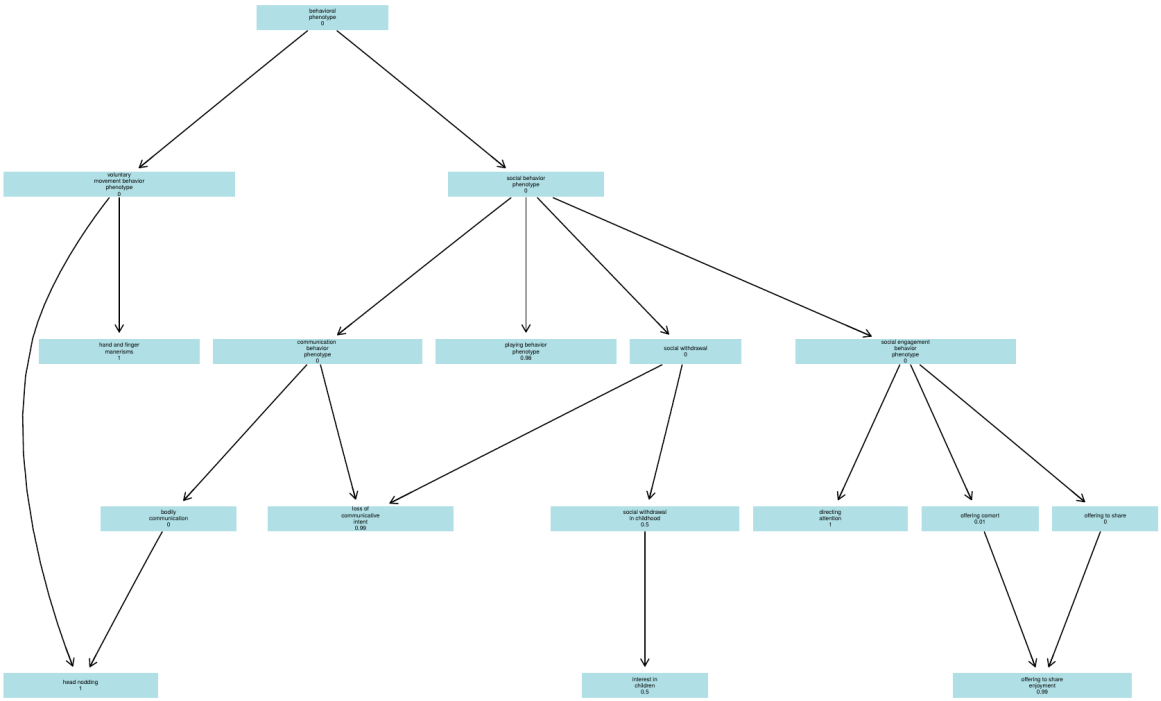



Figure 4.12: *Minimal sub-graph of the PNBO traits in cluster green.*

The red cluster is significantly enriched for defects in coping with changes in the proband's environment, sensitivity to noise, and increased visiospatial ability, which unlike most PNBO traits indicates a gain of function, where an individual is assessed to have increased visiospatial cognition, see Figure 4.13.

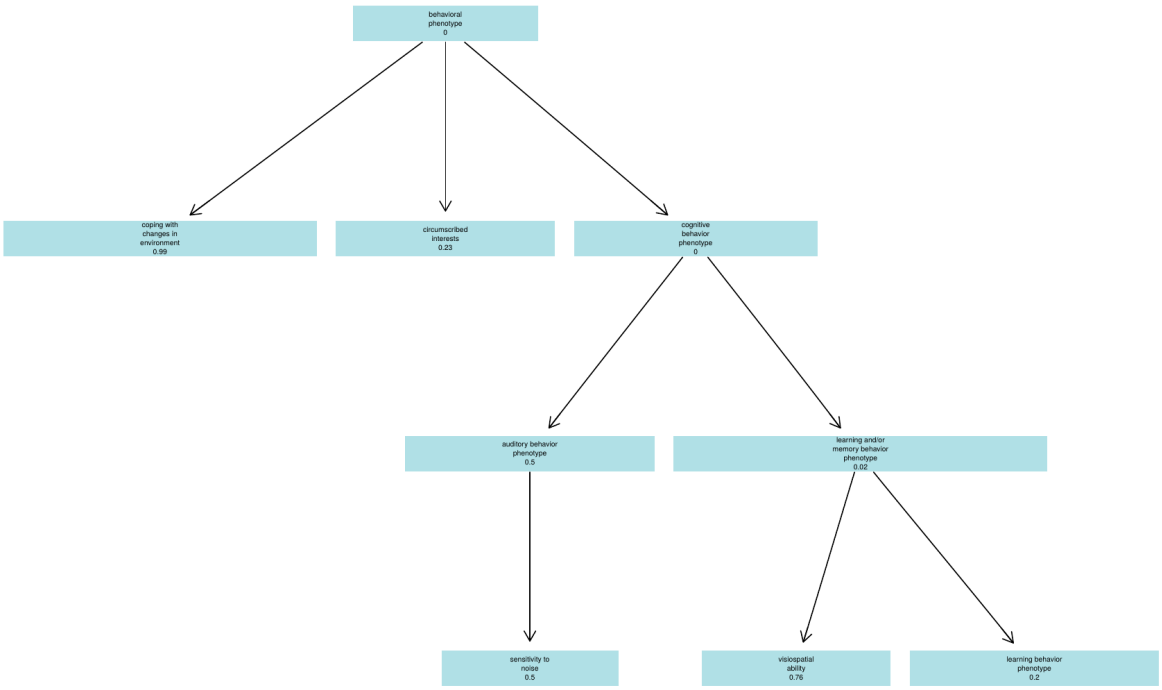



Figure 4.13: *Minimal sub-graph of the PNBO traits in cluster red.*

The black cluster and all following have under 100 probands assigned, and probands who are members of the black cluster have the fewest average traits annotated to them, 28.5 (Table 4.4). Probands typically exhibit defects in voluntary movement, reciprocal conversation, and appropriate social responses (Figure 4.14).

The pink cluster is uniquely made up of individuals who have the restricted behaviors often exhibited by autism patients, including circumscribed interests, unusual sensory interests, and unusual preoccupations, Figure 4.15.

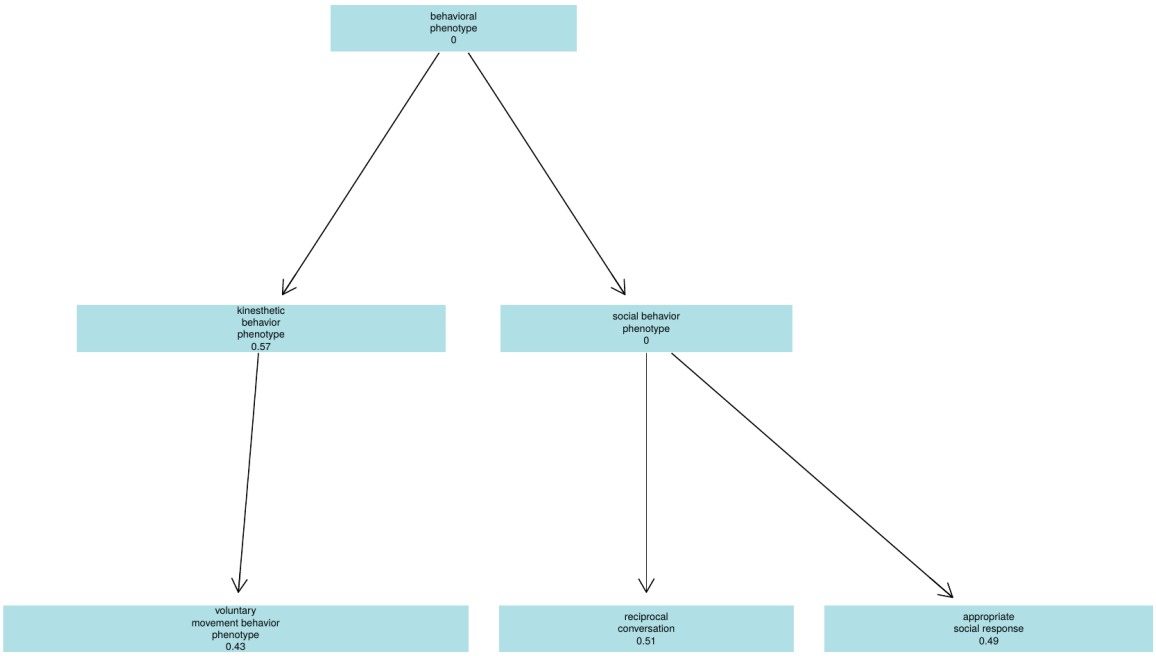



Figure 4.14: *Minimal sub-graph of the PNBO traits in cluster black.*

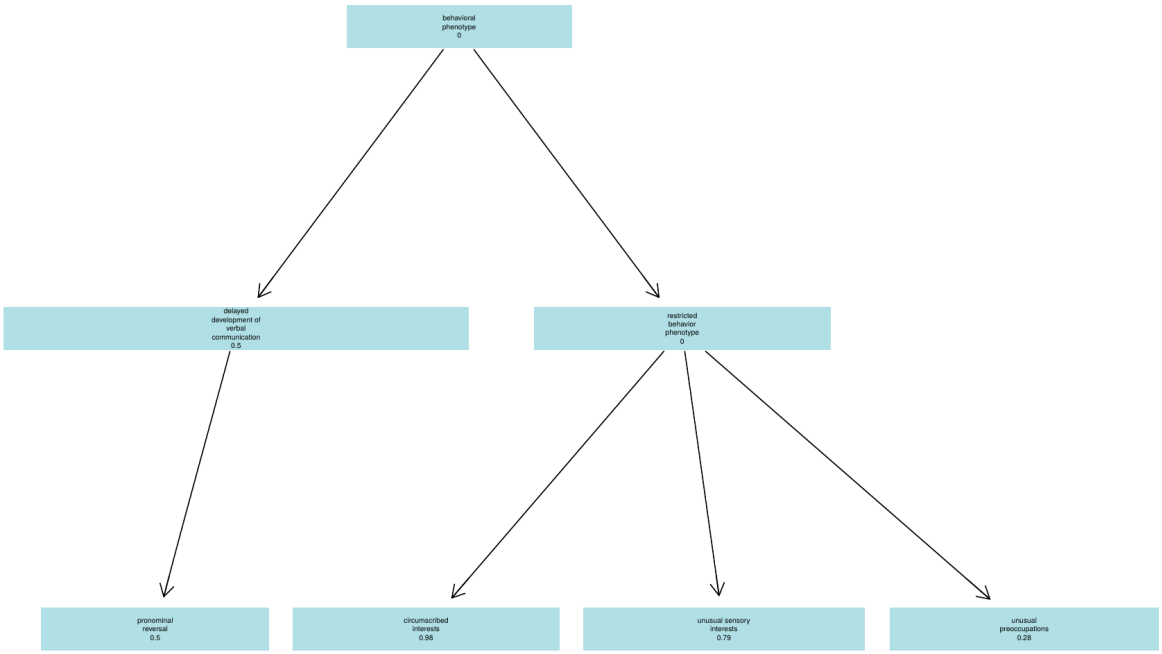



Figure 4.15: *Minimal sub-graph of the PNBO traits in cluster pink.*

The magenta cluster does not have a well defined signature, but self injury (with a 13% probability, and difficulties in coping with changes in the external environment (with a 22% probability) make the cluster of probands distinct.

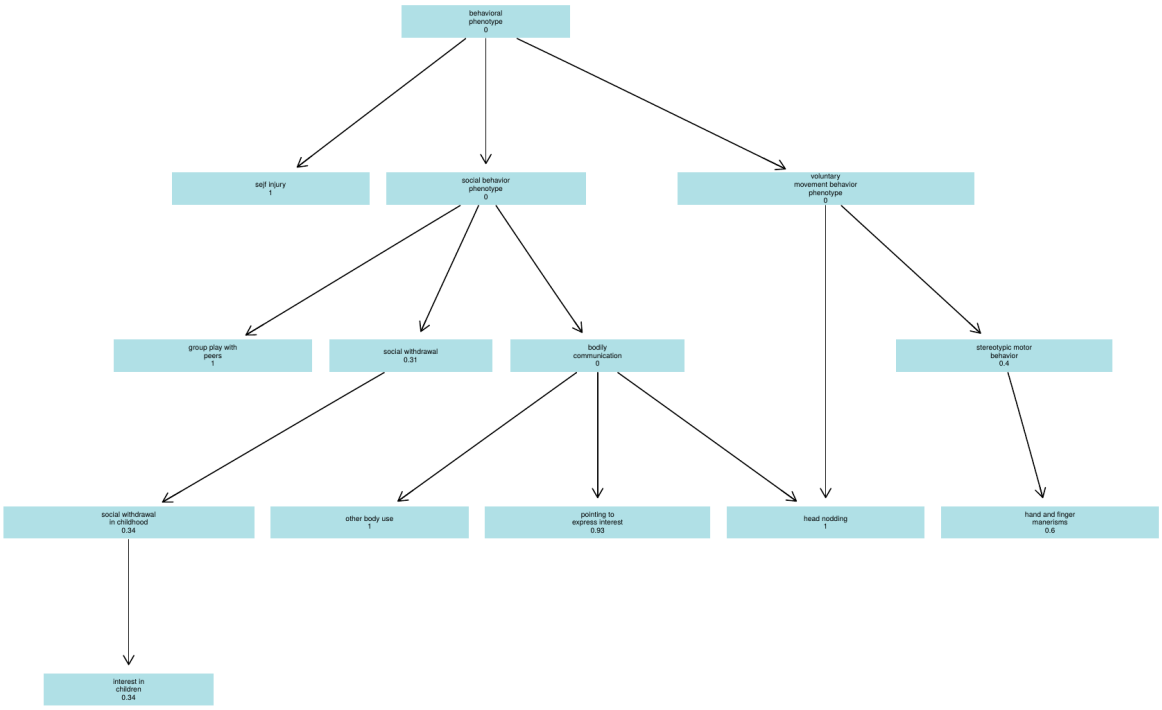



Figure 4.16: *Minimal sub-graph of the PNBO traits in cluster magenta.*

Uniquely among clusters, the purple cohort of 73 probands have a strong likelihood of self-injurious behaviour, with few instances in directing others' attention towards interests, poor eye contact related behavior, and few reported instances of smiling to express happiness, Figure 4.17.

The greenyellow cohort (Figure 4.18) collectively have a strong loss of language ability which was previously acquired, and abnormal social verbalization and response to parental cues.

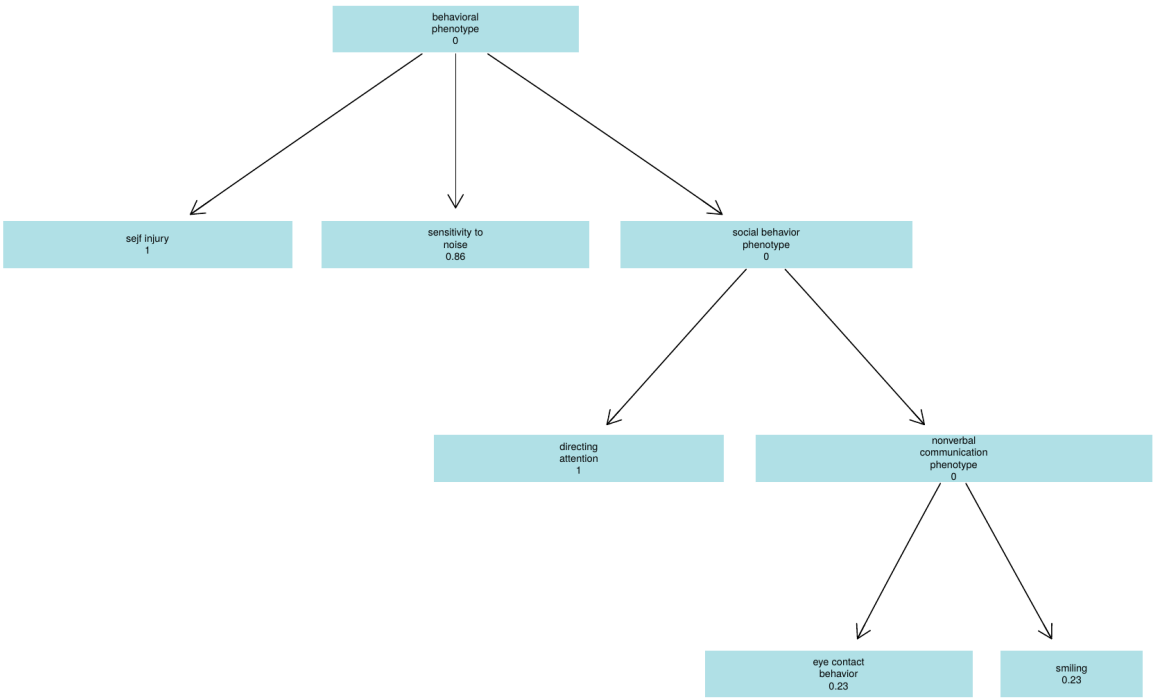



Figure 4.17: *Minimal sub-graph of the PNBO traits in cluster purple.*

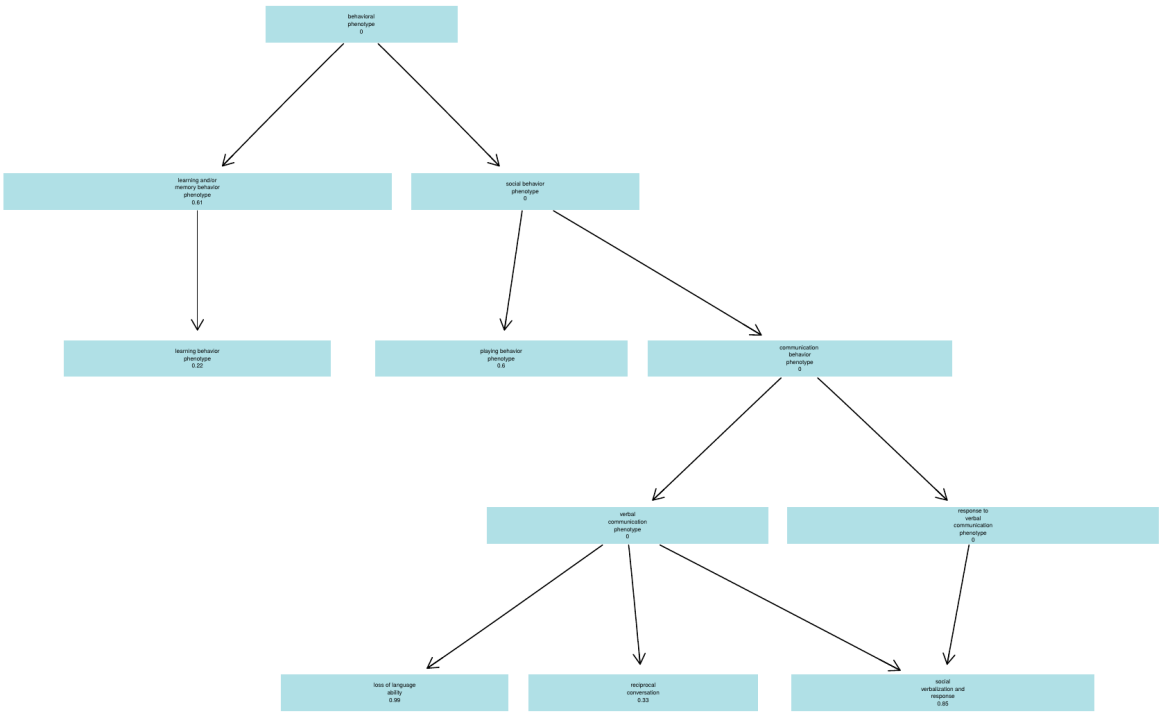



Figure 4.18: *Minimal sub-graph of the PNBO traits in cluster greenyellow.*

The tan subgroup (Figure 4.19) are likely to exhibit abnormal gait, and unlikely to use their heads to nod "yes" to communicate.

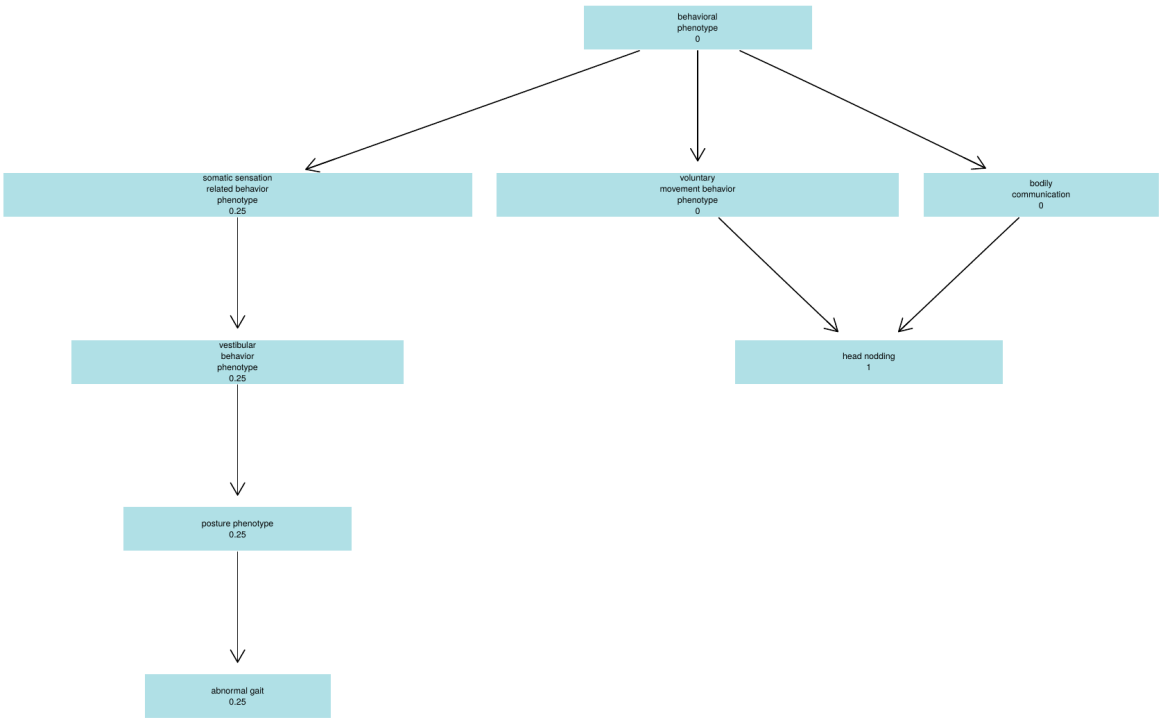



Figure 4.19: *Minimal sub-graph of the PNBO traits in cluster tan.*

The salmon cluster (Figure 4.20) has a strong association (marginal probabilities of 1 each) of having unusual sensory interests, having difficulty in coping with personal environmental changes, and asking inappropriate questions.

Lastly, the cyan cluster, like the blue cluster, contained hetergogeneous behavioral traits with one discovered unifying element, decreased frequency of head nodding. Head nodding, in this case, indicates nodding "yes" to non-verbally communicate agreement.

Taken as a whole, these clusters are often made up of individuals with common phenotypic traits across the spectrum of communication, repetition/restriction, and social

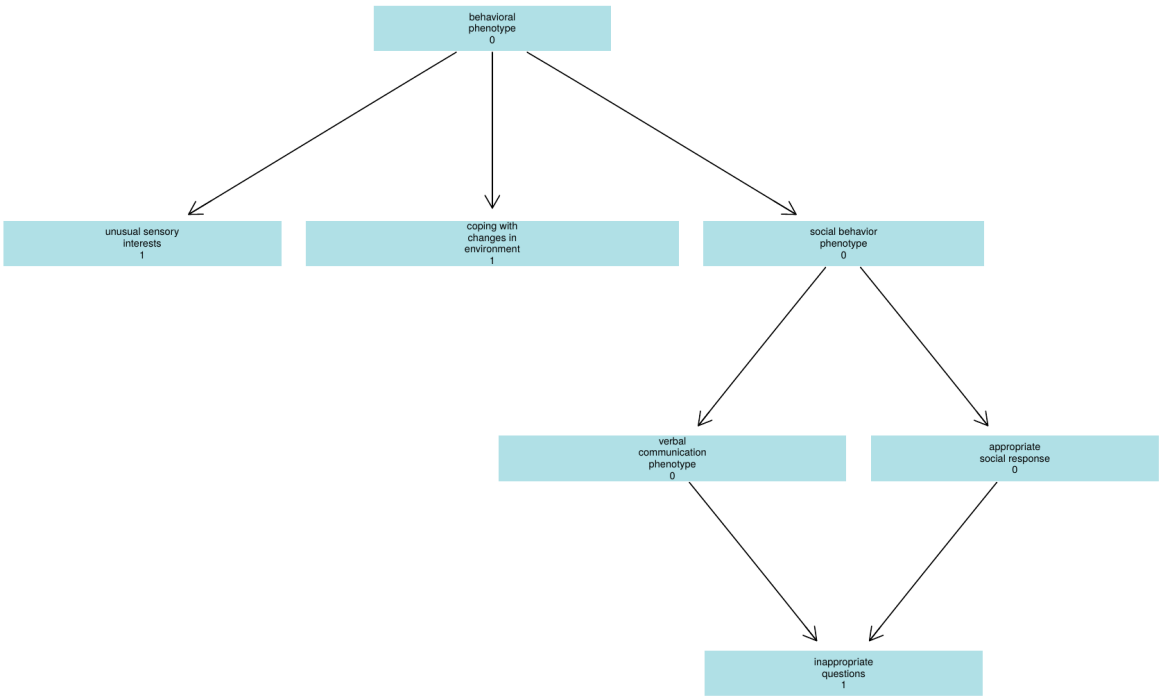



Figure 4.20: *Minimal sub-graph of the PNBO traits in cluster salmon.*

behavior traits. Each cohort can be distinguished by a core set of shared traits, revealing possible sub-types of autism characterised by a spectrum of endophenotypes. The PNBO ontology and the ANBO ontology, as implemented in (Martínez-Santiago et al., 2020), is available in the online appendix, https://github.com/jaw-bioinf/PhdThesis

## 4.4 Discussion

The process of creating the ANBO and the PNBO necessitates first ontology evaluation, and then discussion of the value of the ontologies themselves.

### 4.4.1 Ontology evaluation

The ANBO and PNBO ontologies can be validated several different ways. Hoehndorf and colleagues (Hoehndorf, Dumontier, and Gkoutos, 2013) suggest three distinct modes of ontology validation:

- direct evaluation: accessing intrinsic properties of an ontology, such as consistency and expressivity
- application based evaluation: evaluating an application which itself uses an ontology
- analysis based evaluation: evaluating a scientifically based data analysis which uses an ontology for data integration or analysis itself

In this chapter, I evaluate the ANBO and PNBO by the first criterion. I also perform unsupervised machine learning, in the form of hierarchical clustering, using data informed by the PNBO; thus evaluating the PNBO's ability according to the third criterion as well.

Both the ANBO and PNBO use the DL subset of the OWL query language, which ensures decidability. At different stages, both the Hermit and ELK reasoners were used; as the ANBO or PNBO grow it may be necessary to restrict reasoning to the ELK reasoner as it is guaranteed to classify an ontology in polynomial time (Kazakov, Krötzsch, and Simančík, 2014). Both ontologies follow the principles of the OBO Foundry (Smith et al., 2007b), and are working towards fulfilling as many criteria of the minimum information for reporting ontologies (MIRO) standards as possible (Matentzoglu et al., 2018). Guidelines include basics (ontology name, URL), motivation (need, competition), scope (issue tracking, community), knowledge acquisition (data source acknowledgement), ontology content (metrics, axiom patterns), managing change (sustainability plan), and quality assurance (testing and examples of use). While some criteria for best practices are outstanding for either ontology (issue tracking for ANBO, for instance), this Chapter aims to fulfill the majority of these suggested best practices. Following the OBO Foundry best practices include creating non-overlapping ontologies with a strictly focused content, and importing data from other ontologies using the MIREOT principles (Courtot et al., 2011). Gomez-Perez (Gómez-Pérez, 2001) provides five criteria for validating ontologies using validation approach 1 of (Hoehndorf, Dumontier, and Gkoutos, 2013), which are fulfilled as demonstrated here:

- Consistency: Both ANBO (including SWRL rules) and the PNBO are logically consistant, and evaluation with ELK found no cases of class unsatisfiabiliity, or classes which could contain no instances.

- Completion: That everything which should be in the ontology can be inferred from a reasoner or is explicitly built in. With colleagues, we ensured that at least one ADL was defined for each cognitive process in ANBO; and I ensured that all traits from the PANSS and ADI-R were incorporated into the PNBO.

- Concision: I have pruned behavioral phenotypes from the PNBO which do not currently

relate to either ASD or SSD, while keeping the behavior process domain intact.

- By keeping the behavior domain intact, I facilitate the expandability of the ontology - the effort needed to add new definitions or terms. As knowledge of behavioral endophenotypes grow, this strategy will help ensure the integration of PNBO with other ontologies, including the Mammalian Phenotype ontology.

- Sensitiveness: the ability of the ontology to withstand small changes. By porting the PNBO away from the NBO, I was able to make fundamental changes to the description of psychologically relevant neurobehaviors without breaking the backwards compatability of the NBO, which forms the basis of studying behavior in several ontologies.

### 4.4.2 Why model autism with the PNBO?

Fundamentally, there are two main reasons to model autism with the PNBO instead of previously discussed ontologies, these being interoperability, and the second being its ability to correctly capture phenotypic traits relating to ASD. The PNBO relies on EQ notation from PATO to define behavioral phenotypes by the behavioral processes in which they participate. As illustrated in the EQ example with the Mammalian Phenotype (MP) ontology, this allows endophenotypes derived from children with autism or adults with psychosis to more readily be translated into observable characteristics in mouse. This is essential, because ontology labels (the readable names) can be misleading in certain contexts. The term MP:0002730, head shaking, is a stereotypic behavior and an abnormal head movement. The NBO term NBO:0000023, head shaking, is a behavior process relating to head movement on the horizontal plane, and not a phenotypic trait, nor is it related to stereotypic behavior necessarily. The PNBO term PNBO:051, head shaking, is a lack of shaking the head back and forth to indicate a negative response to questions by others. This, unlike previous uses of 'head shaking,' is not just a physical response to an external stimulus (a behavior), but

also inheres in a specific cultural context in which shaking the head horizontally indicates "no" or displeasure, and has specific communicative intent. The existing autism ontologies, including the ADAR, do not take advantage of EQ notation and are thus unusable by the wider community not focused specifically on autism research in humans. The ability of the PNBO to align with MP and HPO ontologies is a major strength, particularly with the ability of home cage monitoring in mouse and rat experiments to capture unexpected behaviors (Bains et al., 2016; Brown et al., 2017b). Importantly, there can be no diagnosed "autistic" or "schizophrenic" mouse, but endophenotypes of these disorders can map to approximate equivalents. The operational subset of the PNBO needed to characterise ASD is remarkably slim, with only 102 classes needed to incorporate the 93 questions from the ADI-R. When comparing this to the work of Mugzach and colleagues, this may at first seem inadequate. Their ontology, however, includes entities which are deemed phenotypes but are not observable properties of an organism. The class "Imaginative Play," for example, contains two sublasses "ImaginativePlayNotAvailable" and "ImaginativePlayUnknownOrNotAsked". While no doubt useful for predicting if a subject will be classified as autistic, the ontology does not allow interoperability with other ontologies which are used in behavioral genetics, limiting its potential usefulness to the neurobehavioral genetics community in annotating genes with autism-specific phenotypes; if association analyses are performed on the encoded "phenotypes" in the ADAR, interpretation of "ImaginativePlayWithPeersNotAvailable", for example, would be impossible. Where the ADAR is extensive but confounding, the current state of psychological disorder description in the HPO and in the original NBO have flaws which the PNBO seeks to correct. From the NBO, we removed all instances of explicity diagnoses, or re-defined them; while there is depressive behavior in the process arm of the PNBO, the phenotypic manifestation of that behavior has changed from "depressive disorder" to "depression," and children terms such as "major depressive disorder" have been eliminated. Though useful, "major depressive disorder" is not itself a phenotype but a diagnosis, muddling the concept of disease and phenotypic trait. Likewise, schizophrenia is not

directly described by the PNBO, which is not designed to aid diagnosis of the condition, and thus the original NBO term has been removed. While the presence of these terms do not directly impact the ability of the PNBO to model endophenotypes of ASD or SSD, they are a potential source for confusion. When appraising the HPO's depiction of autism and autistic behavior, many problems are evident that give one pause when using it to describe behavioral disorders in general. Figure 4.2 shows the many subclasses of autistic behavior which may be annotated with patients via diagnostic instruments, or with genes from experimental or computational studies. The true path rule, as proposed by the OBO Foundry (Smith et al., 2007b), follows the principle of subsumption: when a gene is annotated to a child term, it must follow transitive relations up the ontology structure. Thus, any gene which is involved in "Lack of peer relationships," a definitive ASD trait, must also be annotated to "Autistic Behavior." Any patient exhibiting hypersexuality or risk taking is likewise annotated with autistic behavior. Given the fact that autism is generally diagnosed at a young age, hypersexuality is unlikely to be an autistic trait, and appears to be so because it is a child of "Inappropriate behaviors." This highlights the social context in which biomedical ontologies must be used if they are to be interpreted correctly. By removing the diagnosis-suggesting "Autistic Behavior" term, and modeling all behaviors on physiology-based behavioral processes, the PNBO works around the confusion which would be evident if persons on the autism spectrum were annotated phenotypes following the HPO. The PNBO also highlights the co-ocurrence of ASD and SSD phenotype.

### 4.4.3 On which spectrum: autism or schizophrenia?

Much previous work has been done on the overlap between ASD and SSD conditions; indeed they were once considered the same disorder. Autism was viewed, in the 1970s, as an early manifestation of SSD (Rutter, 1972 Oct-Dec; Kolvin, 1971). Each condition is di-

agnosed purely based on behavioral assessment, and diagnostic instruments still carry the legacy of early conflation of SSD and ASD. Both the DSM-5 criteria and the PANSS interview both mention autism directly; the DSM as a potential confounding factor and the PANSS as a description of the trait "Preoccupation": "Absorption with internally generated thoughts and feelings and with autistic experiences to the detriment of reality orientation and adaptive behavior." Recent work has highlighted the difficulty in diagnosing ASD in SSD patients because of phenotypic overlap, even with specially made instruments (Deste et al., 2018; Kästner et al., 2015). Figure 4.6 shows the degree of overlap between the two conditions in the PNBO; encoded overlapping traits include both low-level phenotypic traits with high specificity (anxious behavior, reciprocal conversation) but also intermediate level terms (involuntary movement behavior phenotype). By encoding the PANSS and ADR-I into the same ontology, a higher degree of overlap between conditions is seen than would otherwise be expected. This is explained by the transitive nature of subsumption reasoning, as described by the true path rule. This overlap is heterogeneous and does not speak to a directionality of phenotype - however, Mendelian randomization analysis seems to echo a base assumption from the 1970s about the relationship between ASD and SSD. In my causative modeling of SSD exposure to ASD, I used SNPs which were not associated with SSD. Both the effect size of all methods used, and the post-hoc directionality tests suggest a strong possibility that a lifetime exposure to SSD associated genetic loci may cause ASD. There was strong heterogeneity and among SNPs, though the intercept test under MR Egger regression suggests a lack of direction pleiotropy which would invalidate the robust methods used. While no single SNP will likely be the cause of this association, the strong evidence brought forward by the causal model confirms the interconnected presentation, and possible interconnected etiology, of these heterogeneous conditions.

### 4.4.4 Clustering the spectrum

Methods originally developed for weighted gene co-expression analysis were adapted for clustering probands based on the semantic similarity of their phenotypes. Probands were divided into 14 clusters, with no probands failing to cluster (possible using the WGCNA framework if unstable clusters are found during recursive tree cutting). While each cluster was accessed for stability with permutation testing, each cluster was kept for further analysis. The most informative measure tested under permutation was the weighted degree, which calculates the mean weighted degree of nodes in each cluster, the idea being that nodes in a cluster should have higher degree ( a measure of inter-connectivity) than expected by chance. When testing each cluster's stability, the mean semantic similarity of each trait was also calculated. While both measures were robust, the mean semantic similarity of traits alone was not enough to indicate cluster stability. This highlights the usefulness of exponentiating the adjacency matrix before using it to create a dissimilarity matrix for clustering. By applying that transformation, semantically dissimilar pairs of probands's connections are further weakened. The range of raw similarity, measured by Resnik's method, was (2.81, 5.22), indicating the high degree of inherent similarity of probands. Semantic similarity-based regression was used to both independently validate clusters and indicate which combinations of traits were most likely to segregate a cluster from the combination of all others. Two samples, cyan and blue clusters, while varying greatly in size, were themselves heterogeneous enough to not produce robust phenotypic signatures under similarity regression. Some clusters, for example green, produced a signature phenotype profile of highly related traits (loss of language ability and abnormal social verbalization and response rates). Others, such as the red cluster, included gain of function traits - so called savant abilities. While the red cluster was marked by increased visiospatial ability, other abilities possible include computation ability and artistic ability. In the PNBO, these are disjoint with cognitive traits indicating loss of communicative ability. The semantic profiles revealed include several upper-level terms - such as in

the black cluster where probands exhibit abnormal voluntary movement. There are several subclasses of voluntary movement, indicating that the members of the black cluster share several varied movement abnormalities. If a more specific movement related trait helped define the cluster vs all others, it would have been revealed in the similarity regression analysis. Interestingly, the pink cluster is uniquely distinguished by probands who have high levels of several restricted behaviors. Here, sensory interests indicate repetitive preoccupations such as spinning the wheel of a toy car repeatedly, or running in circles. The type of sensory interests are not always well defined by the ADI-R; if they were anatomical components from UBERON could be assigned to them to further segregate interests by body parts or associated behaviors. For example, running repeatedly in circles may be indicated on assessments as an unusual sensory interest, but may also be a voluntary movement trait. When viewing each proband as a constellation of 100 traits, cluster-based similarity regression has been useful for reducing the dimensionality of these traits in subgroups. What remains to be seen, however, are if probands in different clusters have different genetic loci associated to them from GWAS analysis.

### 4.4.5 Extending PNBO and ANBO beyond this chapter

A main goal of the research in this project is to expand both the ANBO and PNBO beyond what is currently presented. To benefit the model organism research community, I aim to align the PNBO with the Mammalian Phenotype ontology. Working with colleagues, I aim to annotate the PNBO with data from patients on the schizophrenia spectrum, leveraging large deeply phenotyped cohorts as part of the Psychosis Immune Mechanism Stratified Medicine Study; see https://www.birmingham.ac.uk/research/mental-health/Psychosis-Immune-Mechanism-Stratified-Medicine-Study.aspx. By applying the PNBO to the various cohorts in this study, I will be able to validate the ability of PNBO to discriminate clusters of

non-autism patients. The PNBO modelled the phenotypic profile of one well-studied cohort of autism patients, the Simons Simplex Collection. By expanding the cohort to others diagnosed with the ADI-R, more precise and better powered characterization of autism subtypes will be possible. Lastly, as the ANBO expands and opportunities to use it in practice with SmartLab begin, it may be possible to extend this approach of studying cognitive processes to the model organism community, by semantically labeling sensors with domains of activity in a home cage. To accomplish this, ANBO would have to be modified to be mouse focused, and extensive consultation with model organism focused neurobehavioral geneticists would be required.

## 4.5 Conclusions and Chapter Summary

In this chapter, I have presented the development of the Psychological Neuro Behavior Ontology, or PNBO, and the Aging Neuro Behavior Ontology, or ANBO. Both ontologies are logically consistent and concise descriptions of their respective domains. The PNBO incorporates phenotypic manifestations of both autism (ASD) and schizophrenia (SSD) spectrum disorders from two diagnostic interviews, the Positive and Negative Syndrome Scale and the Autism Diagnostic Interview - Revised. The degree of overlap between ASD and SSD traits is reflected a possible genetic etiology suggested by a Mendelian randomization analysis using independent SSD and ASD populations. To my knowledge, this chapter includes both the first ontology to represent the SSD phenotype, and to do so in combination with the ASD phenotype. It also includes the first Mendelian randomization study of exposure to SSD on ASD, and results are intriguing even if possible balanced pleiotropy is involved. Lastly, a major work of this chapter was modeling the Simons Simplex Collection cohort with the PNBO, and segregating clusters of probands who manifested similar traits. Robust bootstrapping suggests most clusters of probands exhibit more stability than would be expected

by chance. Additionally, linear modeling of phenotypic profiles which can explain the cluster segregation suggests cluster Independence while explaining which traits are informative in segregating probands in the SSC cohort.

Modeling the relationships between traits exhibited by the probands of the SSC cohort produced several clusters of individuals with a shared phenotypic profile. Autism manifests itself in these various phenotypic profiles, which may have differing genetic etiologies. This leads to the natural complement of detailed phenotyping when genetic data is available: genomics. To test the genetic variability of autism sub-profiles, in the next chapter I employ genome-wide association studies (as in Chapter 3) and complementary analyses to correlate genetic mutations in the SSC cohort with both individual traits encoded in the PNBO and with the clusters derived in this chapter.

| | Metric | Count |
|---|---|---|
| 1 | Axiom | 52872 |
| 2 | Logical axiom count | 9064 |
| 3 | Declaration axioms count | 3649 |
| 4 | Class count | 3255 |
| 5 | Object property count | 226 |
| 6 | Data property count | 0 |
| 7 | Individual count | 4 |
| 8 | Annotation Property count | 166 |
| 9 | SubClassOf | 7058 |
| 10 | EquivalentClasses | 1334 |
| 11 | DisjointClasses | 152 |
| 12 | GCI count | 26 |
| 13 | Hidden GCI Count | 1317 |
| 14 | SubObjectPropertyOf | 222 |
| 15 | EquivalentObjectProperties | 0 |
| 16 | InverseObjectProperties | 41 |
| 17 | DisjointObjectProperties | 0 |
| 18 | FunctionalObjectProperty | 1 |
| 19 | InverseFunctionalObjectProperty | 0 |
| 20 | TransitiveObjectProperty | 34 |
| 21 | SymmetricObjectProperty | 6 |
| 22 | AsymmetricObjectProperty | 1 |
| 23 | ReflexiveObjectProperty | 0 |
| 24 | IrrefexiveObjectProperty | 2 |
| 25 | ObjectPropertyDomain | 60 |
| 26 | ObjectPropertyRange | 55 |
| 27 | SubPropertyChainOf | 78 |
| 28 | AnnotationAssertion | 36959 |

Table 4.1: *PNBO Ontology Full Metrics*

| | Trait | TraitName | Count |
|---|---|---|---|
| 1 | PNBO_050 | group play with peers | 2580 |
| 2 | PNBO_037 | appropriate social response | 2453 |
| 3 | PNBO_080 | social disinhibition | 2446 |
| 4 | PNBO_054 | imaginative play | 2422 |
| 5 | PNBO_086 | spontaneous imitation of actions | 2409 |
| 6 | PNBO_028 | offering to share | 2387 |
| 7 | PNBO_047 | eye contact behavior | 2347 |
| 8 | PNBO_074 | response to children initiating behavior | 2277 |
| 9 | PNBO_023 | reciprocal conversation | 2255 |
| 10 | PNBO_039 | response to voice | 2255 |
| 11 | PNBO_055 | imitative social play | 2218 |
| 12 | PNBO_058 | initiating social activity | 2215 |
| 13 | PNBO_027 | offering comfort | 2208 |
| 14 | PNBO_022 | social verbalization and response | 2204 |
| 15 | PNBO_072 | repetitive use of objects | 2191 |
| 16 | PNBO_059 | interest in children | 2190 |
| 17 | PNBO_065 | pointing to express interest | 2184 |
| 18 | PNBO_067 | abnormal quality of social overtures | 2162 |
| 19 | PNBO_044 | instrumental gesture behavior | 2147 |
| 20 | PNBO_077 | smiling | 2138 |
| 21 | PNBO_068 | facial expressions | 2105 |
| 22 | PNBO_078 | directing attention | 2068 |
| 23 | PNBO_014 | delayed echolalia | 2042 |
| 24 | PNBO_075 | offering to share enjoyment | 2030 |
| 25 | PNBO_061 | intonation phenotype | 2024 |
| 26 | PNBO_048 | friendship maintenance behavior | 2016 |
| 27 | PNBO_084 | unusual sensory interests | 2010 |
| 28 | PNBO_081 | sensitivity to noise | 1983 |
| 29 | PNBO_041 | circumscribed interests | 1962 |
| 30 | PNBO_019 | loss of simple language comprehension | 1961 |
| 31 | PNBO_056 | inappropriate facial expressions | 1931 |
| 32 | PNBO_004 | verbal communication phenotype | 1877 |
| 33 | PNBO_032 | aggressive behavior towards caregiver | 1863 |
| 34 | PNBO_046 | coping with changes in environment | 1847 |
| 35 | PNBO_052 | hand and finger mannerisms | 1632 |
| 36 | PNBO_035 | abnormal response to sensory stimuli | 1604 |
| 37 | PNBO_026 | head nodding | 1577 |
| 38 | PNBO_029 | complex movement behavior | 1487 |
| 39 | PNBO_034 | delayed development of verbal communication | 1438 |
| 40 | PNBO_057 | inappropriate questions | 1393 |
| 41 | PNBO_066 | pronominal reversal | 1391 |
| 42 | PNBO_051 | head shaking | 1357 |
| 43 | PNBO_033 | aggressive behavior towards non-caregiver | 1324 |
| 44 | PNBO_021 | other body use | 1304 |
| 45 | PNBO_043 | compulsive behavior | 1293 |
| 46 | PNBO_076 | self-injury | 1253 |
| 47 | PNBO_024 | abnormal gait | 1211 |
| 48 | PNBO_010 | loss of articulation skills | 1122 |
| 49 | PNBO_085 | verbal rituals | 1018 |
| 50 | PNBO_070 | memory ability and skill phenotype | 929 |
| 51 | PNBO_083 | unusual preoccupations | 855 |
| 52 | PNBO_082 | attachment to objects | 817 |
| 53 | PNBO_025 | neologism phenotype | 753 |
| 54 | PNBO_012 | loss of learned skills | 669 |
| 55 | PNBO_073 | coping with changes in environment not affecting self | 638 |
| 56 | NBO:0000565 | social withdrawal | 610 |
| 57 | PNBO_003 | loss of language ability | 448 |
| 58 | PNBO_071 | visiospatial ability | 406 |
| 59 | PNBO_069 | reading ability phenotype | 368 |
| 60 | PNBO_007 | loss of meaningful communication | 363 |
| 61 | PNBO_008 | loss of communicative intent | 329 |
| 62 | PNBO_064 | musical ability phenotype | 265 |
| 63 | PNBO_036 | computational ability phenotype | 245 |
| 64 | PNBO_013 | midline hand movements | 240 |
| 65 | PNBO_049 | fainting behavior | 215 |
| 66 | PNBO_053 | hyperventilation | 208 |
| 67 | PNBO_011 | loss of language from physical illness | 192 |
| 68 | PNBO_030 | skilled drawing behavior | 173 |
| 69 | PNBO_016 | constructive playing behavior | 164 |
| 70 | NBO:0000591 | motor coordination phenotype | 111 |
| 71 | PNBO_006 | decreased level of communication ability | 106 |
| 72 | PNBO_018 | loss of self-help skills | 105 |
| 73 | PNBO_009 | loss of syntactical skills | 29 |

Table 4.2: *The frequency of traits the Psychological Neuro Behavior Ontology mapped to the Simons Simplex Collection probands, as provided by the Autism Diagnostic Interview - Revised (ADI-R).*

| | Weighted Degree Pval | Semantic Similarity Pval | Degree FDR | Similarity FDR | Module | Module Color |
|---|---|---|---|---|---|---|
| 1 | 0.0001000 | 0.0001000 | 0.0001273 | 0.0002800 | 1 | turquoise |
| 2 | 0.0001000 | 0.0001000 | 0.0001273 | 0.0002800 | 2 | blue |
| 3 | 0.0217978 | 0.0001000 | 0.0234746 | 0.0002800 | 3 | brown |
| 4 | 0.0001000 | 0.0001000 | 0.0001273 | 0.0002800 | 4 | yellow |
| 5 | 0.0001000 | 0.0001000 | 0.0001273 | 0.0002800 | 5 | green |
| 6 | 0.0001000 | 0.3841616 | 0.0001273 | 0.7683232 | 6 | red |
| 7 | 0.0938906 | 1.0000000 | 0.0938906 | 1.0000000 | 7 | black |
| 8 | 0.0001000 | 0.9748025 | 0.0001273 | 1.0000000 | 8 | pink |
| 9 | 0.0001000 | 0.9600040 | 0.0001273 | 1.0000000 | 9 | magenta |
| 10 | 0.0001000 | 1.0000000 | 0.0001273 | 1.0000000 | 10 | purple |
| 11 | 0.0001000 | 0.9505049 | 0.0001273 | 1.0000000 | 11 | greenyellow |
| 12 | 0.0001000 | 0.5857414 | 0.0001273 | 1.0000000 | 12 | tan |
| 13 | 0.0009999 | 1.0000000 | 0.0011666 | 1.0000000 | 13 | salmon |
| 14 | 0.0001000 | 0.0465953 | 0.0001273 | 0.1087225 | 14 | cyan |

Table 4.3: *Modules from hierarichical clustering were subjected to 10,000 label permutations. The pairwise semantic similarity matrix and average distance of the topological overlap matrix of each community was calculated, and test statistics created against these empirical p-values. PNBO = Psychological - Neuro Behavior Ontology*

| Color | Count | TraitMin | TraitMean | TraitMax | TraitSD |
|---|---|---|---|---|---|
| turquoise | 627 | 25 | 39.49 | 56 | 5.2 |
| blue | 524 | 17 | 41.06 | 57 | 6.51 |
| brown | 518 | 21 | 43.08 | 61 | 6.46 |
| yellow | 138 | 24 | 39.67 | 52 | 5.16 |
| green | 123 | 28 | 40.02 | 53 | 4.9 |
| red | 100 | 20 | 36.17 | 57 | 6.21 |
| black | 95 | 18 | 28.55 | 46 | 5.23 |
| pink | 79 | 25 | 37.65 | 51 | 6.05 |
| magenta | 74 | 22 | 34.96 | 49 | 4.97 |
| purple | 73 | 17 | 31.59 | 46 | 5.36 |
| greenyellow | 71 | 24 | 37.58 | 49 | 6.53 |
| tan | 58 | 24 | 33.67 | 42 | 4.9 |
| salmon | 57 | 18 | 31.82 | 45 | 5.81 |
| cyan | 48 | 27 | 36.9 | 51 | 5.91 |

Table 4.4: *Trait distribution within modules. Count is the number of probands assigned to each module. TraitMin, Mean, Max, and SD are the minimum, mean, maximum, and standard deviation of the number of traits assigned to each proband in each cluster.*

# Chapter Five

# Uncovering genetic correlates of autism endophenotypes

## 5.1 Background and Chapter Overview

While Chapter 4 provided a road map for depicting endophenotypic traits among cohorts of individuals on various spectra of disorders, it did not provide a method for associating genetic loci with these traits. To address questions of which loci or combinations of loci may cause manifested phenotype profiles, this chapter employs a variety of methods.

First, I briefly re-introduce the theory behind genome wide association studies (GWAS), and explain the concept of epistasis, and how genetic interactions can be mined to provide insight into networks of genes which contribute to a phenotype. I then review poly-genic models of complex disease as it relates to psychiatric/behavioral disorders. The experimental basis for this chapter is 64 genome wide association studies (GWAS): 50 of endophenotypes in Simons Simplex Collection probands modeled in the PNBO, and 14 of clusters of probands identified by creating a network out of those traits (see Chapter 4). I demonstrate that GWAS of these trait profiles can produce more robust associations than single-trait GWAS,

and that SNPs from cluster-based GWAS can be used to successfully classify probands as belonging to their identified cluster. Lastly, I investigate findings from the best-validated GWAS, providing insight into a possible link between chronobiology and a subgroup of autism patients.

### 5.1.1 Genome Wide Association Studies

As previously described in Chapter 3, genome wide association studies, or GWAS, form the basis of modern genetic epidemiology. GWAS have successfully revealed risk alelles associated with traits from body composition (Fox et al., 2012) to opioid sensitivity (Nishizawa et al., 2014). At heart, they are simple association analyses between a genetic locus (often a SNP) and an measurable or observable phenotype. However, such associations are more complicated when subjects in a GWAS study are related. Family based GWAS offers methods not available to case/control analyses of unrelated individuals (Benyamin, Visscher, and McRae, 2009). These include transmission disequlibrium tests (Spielman, McGinnis, and Ewens, 1993) and family-based designs (Laird and Lange, 2006) which take into account paternal or maternal associations or create artificial controls from a trio of parents and a proband. In the analyses performed in this chapter, I kept a case/control framework while using generalized mixed linear models to account for relationships between cases and controls. In combination with traditional GWAS, I explore gene x gene interactions, investigating epistasis.

### 5.1.2 Epistasis

In terms of quantitative genetics, epistasis can be understood as the non-additive contribution of two variants on a phenotypic trait, as proposed by Fisher (Fisher, 1919). We know

that proteins act together to manifest molecular functions, as evidenced by the utility of the STRING protein-protein interaction database (Szklarczyk et al., 2015b). Because of such interactions, there is a dependency between the loci encoding interacting proteins. This non-independence is what Fisher measured, and can be quantified through various tests at the level of the individual allele. Challenges in identifying epistatic interactions range from the number of combinatorial tests to be performed (a test of 10 loci results in 100 tests, and this scales quadraticly). Recent reviews have concentrated on the role of epistatic interactions may play in mental illness and psychiatric disorders (Webber, 2017), leading away from a monogenic influence on disease to a polygenic approach.

### 5.1.3 Polygeninicity in Psychological Disorders

Polygenicity is the contribution of many alleles, whether SNPs or larger mutations, to phenotype manifestation (Wendt et al., 2020). Such contributions to behavioral disorders are often small, and it is only when combing alleles into a polygenic model does heritability or disease status get explained. Beyond polygenic models, the omnigenic hypothesis propose that highly interconnected networks of gene regulatory networks contribute to a phenotype (Boyle, Li, and Pritchard, 2017). A survey of the GWAS catalog (MacArthur et al., 2017) in March 2020 identified 29 associations between 25 independent loci and autism from 6 published studies, only one of which reached strict genome-wide significance. Compounding this, it has been recently demonstrated that polygenic risk for autism is associated with elevated DNA methylation patterns, furthering the complexity of gene regulation's role in ASD phenotype manifestation (Hannon et al., 2018). Knowing the complexity of the genomics of autism spectrum disorder, this chapter attempts to reduce this complexity by organizing ASD patients phenotypically following the approach depicted in Chapter 4, and then gene association testing on clusters of phenotypicly similar individuals is performed.

## 5.2 Methods

This study was approved by the University of Birmingham's ethical review committee, ERN-17-0879, and assigned the Simons Foundation Autism Research Initiative (SFARI) project number 2720.1. Genetic data was downloaded from the Simons Simplex Collection, comprising 2591 families. Of these, the majority are quads including both parents and an unaffected sibling or parental trios, with some having two probands or more than one sibling. The families were genotyped on one of three Illumina platforms: 1Mv1 (333 families), 1Mv3 (1189 families), Omni2.5 (1069 families).

### 5.2.1 SSC GWAS Quality Control

Quality control was carried out using PLINK v1.9 and R v3.5.0. For each of the three datasets the same procedure was followed. Firstly, variants which had a particularly high missing call rate (>0.1) were removed. Parents were excluded from the analysis, followed by heterogeneity outliers, with a rate greater than 3 standard deviations (SD) from the mean and individuals with a missing rate of $> 0.05$. In order to find related individuals outside of families, identity-by-descent calculations were performed. Regions of known high LD were removed, then remaining regions presenting with high linkage disequilibrium (LD) retained using the PLINK function 'indep-pairwise' moving a 50kb window with a step size of 5 variants at a time and a 0.5 $R^2$ threshold. Identity by descent (IBD) scores calculated between individuals, selecting the individual with the lowest missing rate to retain. Inside family comparisons were made for IBD estimates $> 0.9$ to remove twins or duplicates and outside of family $> 0.1875$ to find those too closely related. Variants not positioned on autosomes or with a MAF <0.01 were excluded before the datasets were merged.

Variants were identified that were shared across the three datasets and the merge was

performed. Any variants which presented with inconsistent base pair position or alleles were removed. For each analysis, a covariates file was built which included gender, genotyping platform, and the first five Principal Components (PC's). Principal Component Analysis (PCA) was carried out by removing known high LD regions using the PLINK function 'indep-pairwise' moving a 50kb window with a step size of 5 variants at a time and a 0.5 $R^2$ threshold.

### 5.2.2 GWAS with Related Subjects

To perform a GWAS analysis with related individuals, a mixed model using the GEMMA software suite was used (Zhou and Stephens, 2012). A generalized linear mixed model is created:

$$y = W\alpha + x\beta + Zu + \epsilon \tag{5.1}$$

where $y$ is the vector of outcomes (presence or absence of a trait), $W$ is a matrix of fixed effect covariates and a column vector of 1s, $\alpha$ is the corresponding vector of coefficients including the intercept, $x$ is the vector of genotypes, $\beta$ is the effect size of each genotypic marker, $Z$ is the identity matrix, and $u$ is the vector of random effects, and $\epsilon$ is the vector of error terms. The mixed effect term $u$ is calculated by:

$$u \sim MVN_m(0, \lambda\tau^{-1}K) \tag{5.2}$$

where $u$ is distributed in a multivariate normal distribution with a mean of 0, $m = n$ below and denotes the number of individuals, $\tau$ is the variance of residual errors, $K$ is a $m$ x $m$ relatedness or kinship matrix (derived from the covariance of the individuals, populated by the proportion of shared alleles), and $\lambda$ is the ratio between the variance components. The error term, $\epsilon$, is modelled as:

$$\epsilon \sim MVN_n(0, \tau^{-1}I_n) \tag{5.3}$$

which is multivariate normally distributed about a mean of 0, where $I$ is the identity matrix. In non-human studies, $m$ could denote the number of strains and $n$ the number of animals with $Z$ representing the strain each animal belongs to, but in this case $m = n$ as stated. By incorporating the mixed-effect $u$, which captures the relatedness of individuals, the assumption that each subject in the GWAS analysis is independent does not have to hold. A kinship matrix was built for the entire cohort of probands and siblings and used for each analysis. Individual fam files were created for each of the 50 different behavioural phenotypes modelled in PNBO which had $>$ 1,000 cases. Prior to any testing, final QC steps were taken per trait, removing variants based on a case control missing rate likelihood of $<$ 0.001, missing rate 0.05 or a Hardy Weinberg Equilibrium p $<$ 1e-8. Finally, case-control association was calculated for the remaining variants.

For each GWAS in this chapter, QQ normality and manhattan plots were generated with ggplot2 (Wickham, 2016). Results of p <= 1e-5, were kept for further investigation.

### 5.2.3 GWAS of Clustered Traits

To test the assumption that individuals who shared common traits would also share a common genotypic architecture underlying those traits, GWAS were performed using the presence or absence of a proband in each cluster identified in Chapter 3. For cluster 1, probands in clusters 2:14 were treated as controls, and probands in cluster 1 as cases. The procedure for performing GWAS on individual PNBO traits was followed, including using a kinship matrix as a mixed effect parameter, and including sex, genotyping platform, and genetic principal components as fixed effects.

### 5.2.4 Epistasis Detection

To test for non-additive interactions between traits, logistic regression epistasis tests for SNP-SNP interactions were performed using CASSI ("CASSI"). After each GWAS performed, any SNPs associated to a trait at a moderate p-value of 1e-3 were retained for analysis with CASSI. A likelihood ratio test statistic to test two models was applied:

$$ln(P(y = case)/P(y = control)) = W\alpha + \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B \tag{5.4}$$

versus

$$ln(P(y = case)/P(y = control)) = W\alpha + \beta_0 + \beta_1 g_A + \beta_2 g_B \tag{5.5}$$

where the $\beta$s are the intercept, allele 1, allele 2, and the interaction term respectively. As previously used, $W\alpha$ is the fixed effect covariate matrix including sex, genotyping chip, and 5 PCs. After calculating the likelihood of each model for a pair of SNPs, the ratio of the reduced model over the full $\lambda = L_0/L_1$ can be transformed into a $\chi^2$ statistic:

$$\chi^2 = -2ln\lambda \tag{5.6}$$

to produce a p-value compared to a $chi^2$ distribution with 1 degree of freedom. Test with p <- 1e-5 were considered statistically significant.

### 5.2.5 GWAS validation via LASSO

To validate the results of clustered GWAS, SNPs identified as significantly associating with each cluster $p <= 1e-5$ or whose pairwise combination was associated at the same threshold were kept. Alleles were extracted and recoded to allelic dose format in a dominant model using PLINK. These SNPs were used as features in a LASSO model, mimicking the framework in Chapter 2 and as published by myself and colleagues (Bravo-Merodio et al., 2019). Using the presence or absence of an individual in a cluster as target and SNPs as features,

data were split split into training (75%) and test (25%), stratified to maintain the proportion of members in the target cluster. Cluster membership was regressed on SNPs and all their pair-wise interactions using the least absolute shrinkage and selection operator, or lasso (Tibshirani, 1994; Zou, 2006), as modified for logistic regression. In this regression analysis, the normal objective function is:

$$Residual sum of squares\,(RSS) = \sum_{i=1}^{N} \left(y_i - x_i^T \beta\right)^2 \tag{5.7}$$

which minimizes the residual sum of squares error. The $\beta$ value, the regression coefficients, minimize this value. In the lasso, a penalizing term is added:

$$\underset{\beta \in R^p}{minimize} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{5.8}$$

which adds a $\lambda$ penalizing term. Here, the left side of the equation is the RSS loss function rewritten, and the right side shows the parameter $\lambda$; this is the penalized sum of the absolute values of the $\beta$ coefficient. To optimize the $\lambda$, hyperparameter,each training set was split 10-folds for cross-validation, and a grid search was used to test values of $\lambda$ in each fold, optimized on the best area under the ROC curve. The best performing model was then tested on the hold-out test set. This procedure was repeated in an outer loop 10 times to account for stratification effects. The caret (Wing et al., 2018) and glmnet (Friedman, Hastie, and Tibshirani, 2010) packages were used for modeling, and all plots were created in the ggplot2 (Wickham, 2016) package in the R computing environment (R Core Team, 2013).

### 5.2.6 Phenome Wide Network Creation

To characterize genetic variants which influence the autism phenome, a graph was created. The Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) was used to annotate significant SNPs from GWAS and epistasis detection for each of the 14 phenome clusters

identified in Chapter 3. For each phenotypic profile identified by Similarity Regression, traits with a marginal probability of association greater than 0.5 were kept. Within each gene/cluster association, the $-log_10$ p-value of the association was multiplied by the marginal probability of each trait/cluster association, resulting in an association score. This resulted in gene/trait associations which formed the nodes of a weighted, undirected graph. Edge weights were the association scores. The graph was visualized using a weighted force-directed layout in Cytoscape v3.9.0 (Shannon et al., 2003).

### 5.2.7 Gene Set Investigations

SNPs identified as significant in the SSC GWAS were annotated with the Ensembl Variant Effect Predictor, using version 97 (McLaren et al., 2016). Variants within protein-coding genes were annotated for coding consequences, and potential deleteriousness with SIFT (Vaser et al., 2016) and PolyPhen (Adzhubei et al., 2010). Gene sets were investigated by performing gene set enrichment analysis using the gProfiler2 R package (Kolberg and Raudvere, 2020; Reimand, Arak, and Vilo, 2011). Databases and ontologies used include the Gene Ontology (Gene Ontology Consortium, 2015), the Kyoto Encyclopedia of Genes and Genomes Database (Kanehisa and Goto, 2000), the Human Phenotype Ontology (Köhler et al., 2019), the CORUM database (Ruepp et al., 2010), REACTOME (Fabregat et al., 2016), and Wikipathways (Slenter et al., 2018). To compare my findings against a known gold standard for autism-related genes, the gene scoring SFARI Gene database was downloaded from the Q2 2019 release.

## 5.3 Results

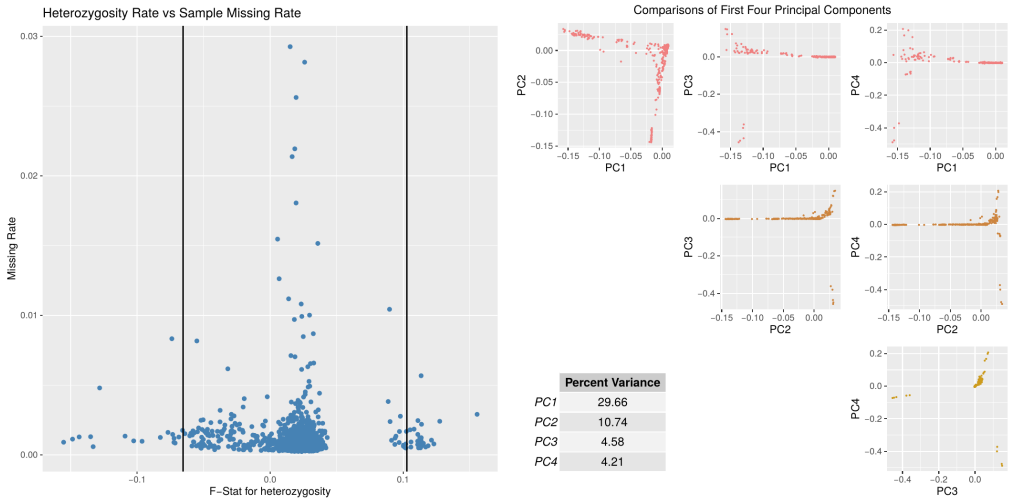
Strict quality control and preprocessing allowed for the assessment of individuals and SNPs, but also facilitated for the mitigation of ethnic background artifacts by including loadings from the principal component analysis of each cohort. The 1Mv1 chip cohort (Figure 5.1) contains individuals with the highest missingness rate of their SNPs, while also being the cohort genotyped - see Table 5.1. The cohort genotyped on 1Mv3, the second generation to be processed in the SSC, is also the largest, and the large degree of variance explained by the first principal component reflects the homogeneity of the cohort (Figure 5.2. The near horizontal lines in the right of panel 5.2b indicate the low explanatory power of dimension 4 and following principal components. The percent of variance explained by the second component is the largest of each cohort. The Omni cohort reflects a wide degree of heterogeneity (Figure 5.3a as indicated by the F-Statistic range (black vertical bars). While not the largest cohort (1016 probands vs 1136 in the 1Mv3 passing QC), it is by far the largest chipset, including over twice the loci passing QC as any other genotyped cohort in the SSC, Table 5.2.

| Individual Filter | 1Mv1 | 1Mv3 | Omni2.5 |
|---|---|---|---|
| Original | 335/1019 | 1193/3433 | 1077/3163 |
| Remove Parents | 335/353 | 1191/2378 | 1077/1025 |
| Check Sex | 335/350 | 1191/1046 | 1077/1024 |
| Heterozygosity | 326/343 | 1140/1012 | 1023/979 |
| Relatedness | 324/343 | 1136/1007 | 1016/978 |

Table 5.1: *Filtering individuals in three arrays. Individuals are segregated into proband/other, where other includes parents and siblings.*

### 5.3.1 Single Trait GWAS fail to reveal significant associations

Fifty GWAS analyses were performed on PNBO-encoded traits from the SSC cohort. No single-trait analysis produced SNPs above a strict genome-wide threshold for significance

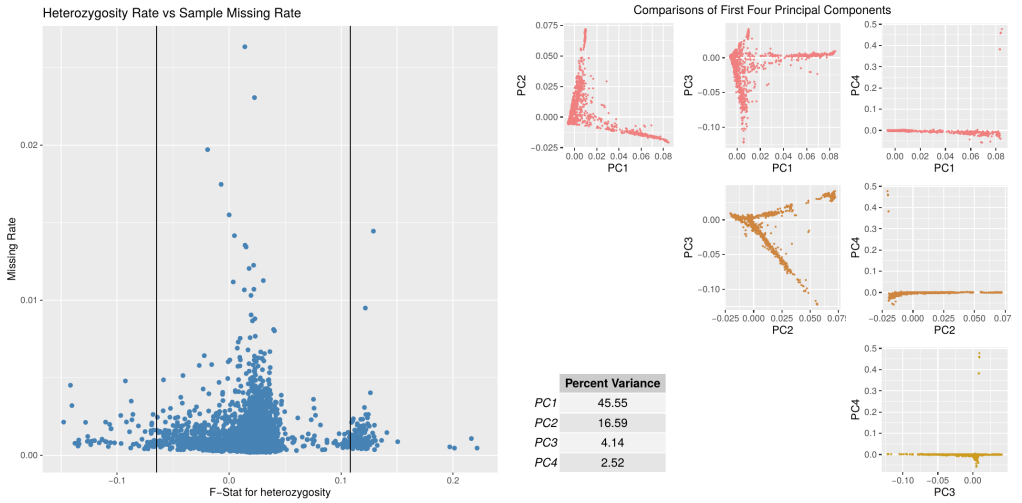



(a) (b)

Figure 5.1: *Diagnostic plots for the Simons Simplex Collection (SSC) cohort genotyped with the Illumina 1Mv1 array. For each individual, panel A indicates the Fisher's test statistic for heterozygosity (x axis), black horizontal lines representing three standard deviations from the mean. The missingness rate for each individual is indicated by the y axis. Panel B indicates the first four principal components of the SNPs belonging to each individual after PCA dimensionality reduction, and the percent of variance explained by each component shown in the key.*

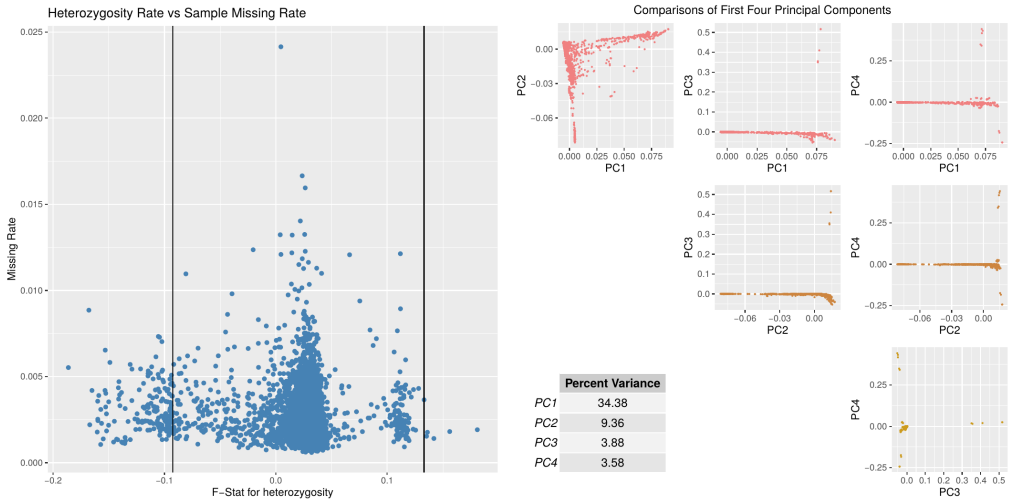
| Individual Filter | 1Mv1 | 1Mv3 | Omni2.5 |
|---|---|---|---|
| Original | 1,072,814 | 1,199,033 | 2,440,283 |
| Autosomes | 1,029,591 | 1,147,689 | 2,383,385 |
| Missingness >0.05 | 1,006,127 | 1,105,550 | 2,380,115 |
| MAF <0.01 | 962,109 | 1,062,277 | 2,236,183 |

Table 5.2: *Filtering variants in three arrays. Within each array, variants were filtered to exclude 10% missing, a minor allele frequency (MAF) of less than 1 %, and exclude sex chromosomes.*




(a) (b)

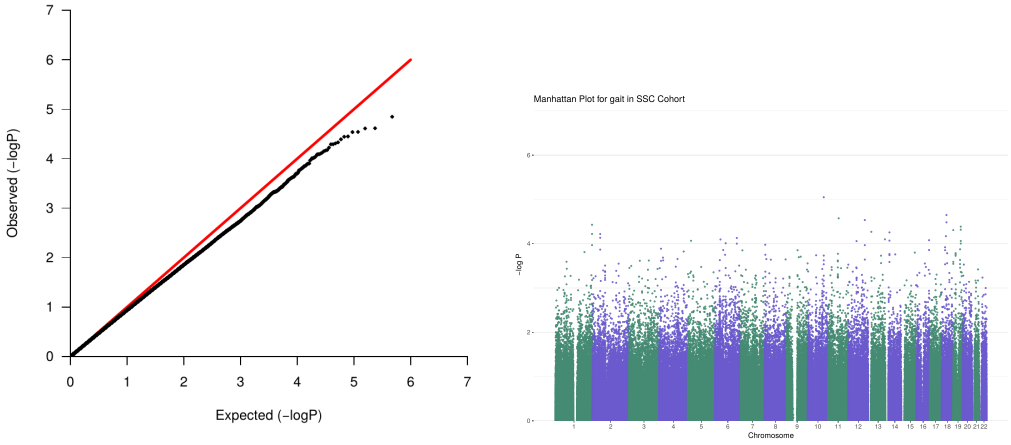
Figure 5.2: *Diagnostic plots for the Simons Simplex Collection (SSC) cohort genotyped with the Illumina 1Mv3 array. For each individual, panel A indicates the Fisher's test statistic for heterozygosity (x axis), black horizontal lines representing three standard deviations from the mean. The missingness rate for each individual is indicated by the y axis. Panel B indicates the first four principal components of the SNPs belonging to each individual after PCA dimensionality reduction, and the percent of variance explained by each component shown in the key.*




(a) (b)

Figure 5.3: *Diagnostic plots for the Simons Simplex Collection (SSC) cohort genotyped with the Illumina Omni2.5 array. For each individual, panel A indicates the Fisher's test statistic for heterozygosity (x axis), black horizontal lines representing three standard deviations from the mean. The missingness rate for each individual is indicated by the y axis. Panel B indicates the first four principal components of the SNPs belonging to each individual after PCA dimensionality reduction, and the percent of variance explained by each component shown in the key.*

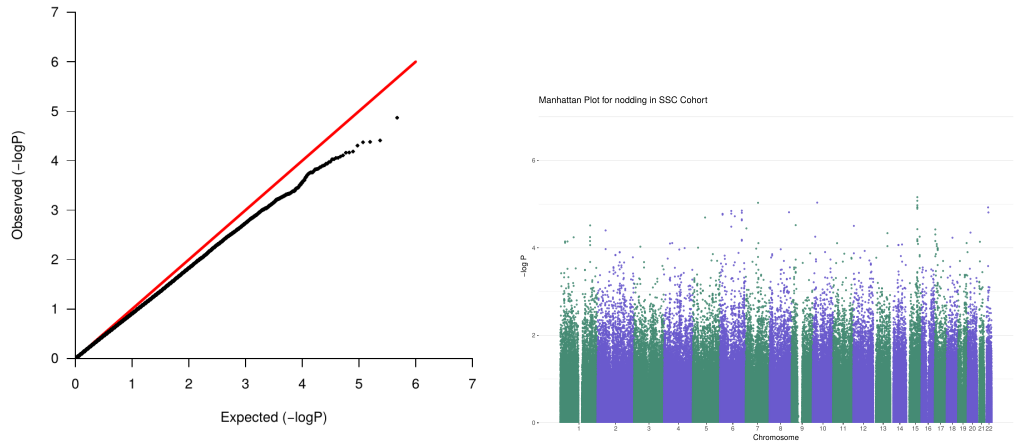
(p < 1e-8), and few produced any hits above a nominal threshold used for inclusion into the GWAS catalog (MacArthur et al., 2017) (p < 1e-5). Two traits are presented, namely an abnormal gait and a lack of head nodding. As seen in Figure 5.4, the GWAS is heavily underpowered ( Figure 5.4a), with a QQ plot whose observed p-values fall far below the expected line The Manhattan plot likewise shows no single locus, and largely no signal.




(a) (b)

Figure 5.4: *GWAS of the SSC cohort whose endophenotype included an abnormal gait as reported by a proband's parent. Panel A shows an under powered QQ plot, with very little signal. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association.*

A GWAS analysis on head nodding produced similar results, see Figure 5.5, with nearly identical QQ statististics; while there appears to be a locus in chromosome 15 which presents several SNPs in high LD, none are statistically significant on a genome-wide scale. GWAS statistics for each individual analysis are available in the online appendix, see https://github.com/jaw-bioinf/PhdThesis.



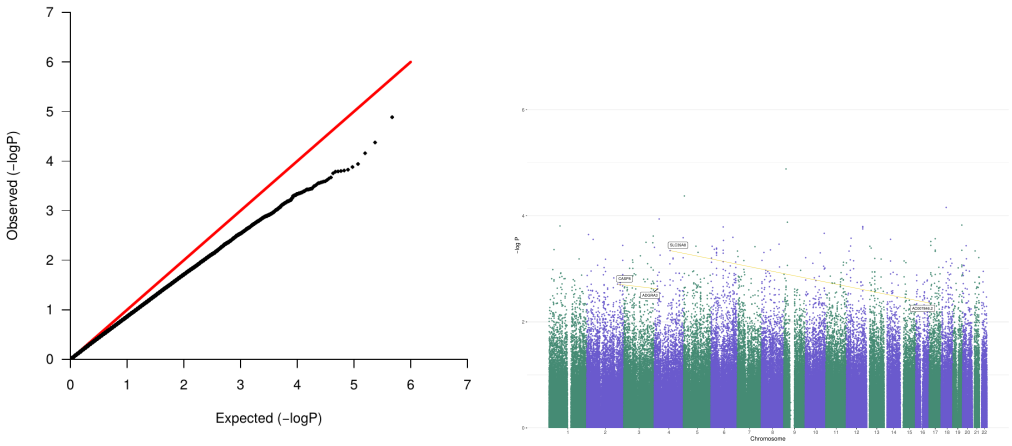

(a) (b)

Figure 5.5: *GWAS of the SSC cohort whose endophenotype included a lack of nodding the head as reported by a proband's parent. Panel A shows an under powered QQ plot, with very little signal. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association.*

### Epistasis detection highlights otherwise non-significant loci

While in most cases there were no genome-wide significant loci revealed when testing each PNBO trait, tests for epistatic interactions did reveal loci which, when combined, had a statistically significant association with PNBO traits ($p < 1e\text{-}5$). The PNBO trait "Delayed Echolalia" can be seen in Figure 5.6. While the GWAS itself was underpowered Figure (5.6a), there were significant epistatic interactions, as can be seen in the Manhattan plot. Intragenic SNPs are labelled with their coding gene, and yellow diagrams indicating interactions between CASP8 and ADGRA3, and between SCL39A8 and AC007846.2. Each interaction is *in trans*, highlighting the multigenic nature of behavioral GWAS.

While the GWAS of abnormal gait in the SSC cohort did not produce any significant results, Table 5.3 indicates 17 significant interactions. Six of them are between SNPs located




(a) (b)

Figure 5.6: *GWAS of the SSC cohort whose endophenotype included delayed echolalia, or the delayed repetition of words and sounds heard by the proband. Panel A shows a QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association. Gold lines indicate epistatic interactions, both intragenic (gene names labeled).*

on chromosomes 11 and 17, each with a $\beta$ of at least 0.43, or an odds ratio (OR) of 1.5. Not all associations are positive, indeed one pairwise epistatic interaction between SNPs on Chr 8 and 14 may be protective against abnormal gait in an autistic population, with a $\beta$ of -17.60, OR of 2e-8.

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p |
|---|---|---|---|---|---|---|---|---|---|
| 1 | rs10903108 | 1 | 25061527 | rs6836731 | 4 | 126776318 | -0.3352 | 0.0739 | 5.69396e-06 |
| 2 | rs1975365 | 1 | 85318295 | rs12120901 | 1 | 241637153 | 1.0540 | 0.2376 | 9.17224e-06 |
| 3 | rs934012 | 2 | 19593206 | rs11707637 | 3 | 25554919 | 0.3354 | 0.0750 | 7.63249e-06 |
| 4 | rs934012 | 2 | 19593206 | rs12432424 | 14 | 35549578 | -0.7752 | 0.1711 | 5.86169e-06 |
| 5 | rs1427538 | 2 | 19627241 | rs11707637 | 3 | 25554919 | 0.3825 | 0.0863 | 9.37178e-06 |
| 6 | rs1530940 | 2 | 236570047 | rs9401672 | 6 | 123820711 | -0.4470 | 0.0847 | 1.3026e-07 |
| 7 | rs13024834 | 2 | 236611291 | rs9401672 | 6 | 123820711 | -0.4034 | 0.0863 | 2.92018e-06 |
| 8 | rs6920980 | 6 | 71083045 | rs9540226 | 13 | 64281942 | 0.7508 | 0.1691 | 8.97587e-06 |
| 9 | rs12375209 | 7 | 10863750 | rs7789727 | 7 | 102576731 | -0.8807 | 0.1963 | 7.23405e-06 |
| 10 | rs4595060 | 7 | 108548069 | rs4782578 | 16 | 82620921 | 0.5280 | 0.1184 | 8.13866e-06 |
| 11 | rs7830660 | 8 | 16460188 | rs10483465 | 14 | 35978788 | -17.6035 | 3.9463 | 8.16772e-06 |
| 12 | rs1152620 | 11 | 65013905 | rs16972147 | 19 | 52544819 | 0.4816 | 0.1026 | 2.65643e-06 |
| 13 | rs1152620 | 11 | 65013905 | rs10419759 | 19 | 52547139 | 0.4844 | 0.1018 | 1.97434e-06 |
| 14 | rs600231 | 11 | 65017222 | rs16972147 | 19 | 52544819 | 0.4396 | 0.0995 | 9.99016e-06 |
| 15 | rs10896016 | 11 | 65092281 | rs16972147 | 19 | 52544819 | 0.4411 | 0.0989 | 8.11992e-06 |
| 16 | rs1784220 | 11 | 65094290 | rs16972147 | 19 | 52544819 | 0.4859 | 0.1020 | 1.922e-06 |
| 17 | rs1784220 | 11 | 65094290 | rs10419759 | 19 | 52547139 | 0.4739 | 0.1011 | 2.76482e-06 |

Table 5.3: *Significant gait Epistasis Results*

An epistatic association analysis of the "head nodding" phenotype revealed 7 significant interactions, including one protective against a lack of communicating yes by a head nod (the explanatory meaning of this trait), between Chrs 1 and 3, $\beta$ -0.92, or an OR 0.39.

While individual traits may not produce many (or any) statistically significant genetic or epistatic associations in the SSC cohort, the PNBO allows us to investigate very closely related traits to see if they produce similar GWAS results.

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p |
|---|---|---|---|---|---|---|---|---|---|
| 1 | rs6673329 | 1 | 161061124 | rs9837045 | 3 | 3044773 | -0.9282 | 0.2088 | 8.81891e-06 |
| 2 | rs2279014 | 2 | 218969420 | rs6481442 | 10 | 52041194 | 0.3257 | 0.0732 | 8.65598e-06 |
| 3 | rs12633333 | 3 | 21027741 | rs4741721 | 9 | 2498296 | 0.3525 | 0.0784 | 6.85237e-06 |
| 4 | rs12633333 | 3 | 21027741 | rs1970074 | 9 | 2499947 | 0.3534 | 0.0781 | 5.98102e-06 |
| 5 | rs6846158 | 4 | 14373558 | rs7725025 | 5 | 83343330 | 0.5811 | 0.1305 | 8.47224e-06 |
| 6 | rs6840522 | 4 | 54725132 | rs17402321 | 7 | 8206434 | 0.6121 | 0.1256 | 1.08975e-06 |
| 7 | rs4446676 | 7 | 103665690 | rs8009761 | 14 | 73958307 | 0.3363 | 0.0739 | 5.42927e-06 |

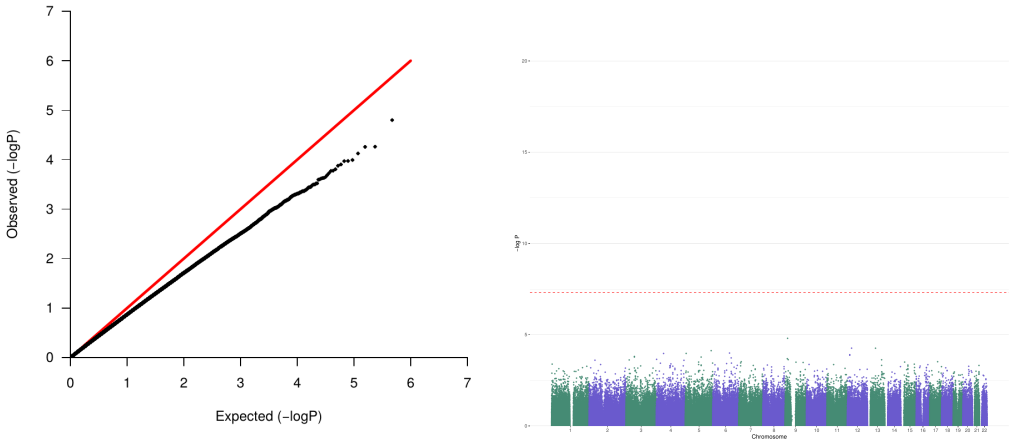
Table 5.4: *Significant nodding Epistasis Results*

### Case study: aggression towards peers (non-caregivers)

The ADI-R asks parents two questions about aggressive behavior, one relating to aggression towards peers (coded as aggression towards non-caregivers), and the other relating to aggression towards caregivers. Each trait is a sibling in the PNBO ontology, with a common parent term (aggressive behavior). When investigating GWAS against a caregiver, no significant results are returned, evident in Figure 5.7. Not only did GWAS not reveal any associations, but epistasis testing did not reveal any interacting SNPs associated with the trait.

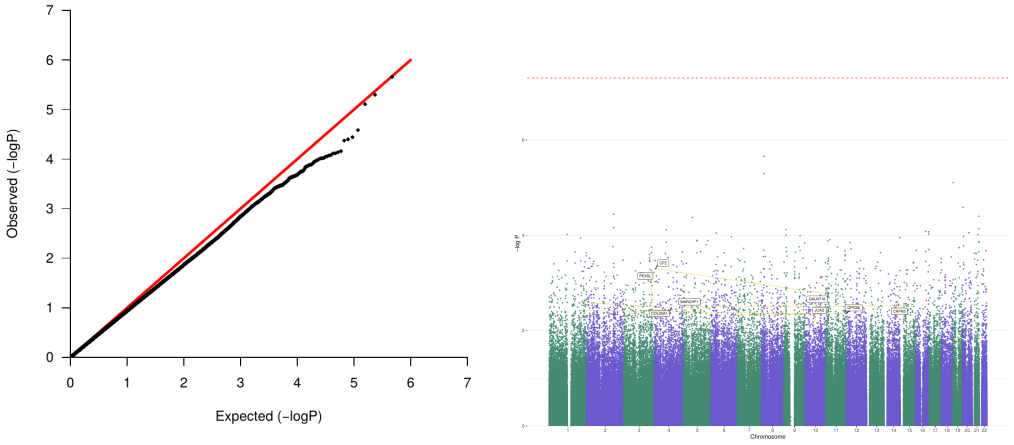
Aggression towards a non-caregiver, however, did result in statistically relevant associations from epistasis testing. While still underpowered, the QQ plot indicates a less-severe observed to expected p-value ratio, and several epistatic interactions are visible in the Manhattan plot in Figure 5.8. Eight epistatic interactions were uncovered, including two *cis* interactions sharing a chromosome, on Chr 3 and 14, see Tables 5.5 and 5.6. Three GWAS SNPs were significant, including several intragenic. These include SNPs within PEX5L, COL25A1, CPZ, MARCHF1, GALNT18, JCAD, OPCML, and CAPN3.

After GWAS and tests for significant epistatic interactions were performed on each PNBO trait, intragenic SNPs were annotated with their corresponding genes, and inter-




(a) (b)

Figure 5.7: *GWAS of the SSC cohort whose endophenotype included aggression directed towards caregivers. Panel A shows a QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association.*




(a) (b)

Figure 5.8: *GWAS of the SSC cohort whose endophenotype included aggression directed towards non-caregivers, usually other children. Panel A shows a QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association.*

| | rs | chr | ps | beta | se | p |
|---|---|---|---|---|---|---|
| 1 | rs6530854 | 8 | 15280833 | 0.0432 | 0.0091 | 2.193156e-06 |
| 2 | rs6530863 | 8 | 15291861 | 0.0403 | 0.0088 | 5.054047e-06 |
| 3 | rs35684719 | 19 | 1852649 | -0.0853 | 0.0191 | 7.806984e-06 |

Table 5.5: *Significant aggression towards a non-caregiver GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p |
|---|---|---|---|---|---|---|---|---|---|
| 1 | rs12141375 | 1 | 201447985 | rs2505127 | 10 | 30409866 | 0.5838 | 0.1261 | 3.62889e-06 |
| 2 | rs4078437 | 1 | 227325540 | rs1526131 | 4 | 110053688 | 0.4728 | 0.0822 | 8.8008e-09 |
| 3 | rs11801153 | 1 | 227327711 | rs1526131 | 4 | 110053688 | 0.4372 | 0.0817 | 8.74462e-08 |
| 4 | rs4855129 | 3 | 181161256 | rs2315505 | 3 | 183835768 | 0.3466 | 0.0755 | 4.3705e-06 |
| 5 | rs2302580 | 4 | 8659534 | rs7120582 | 11 | 11508648 | -0.3424 | 0.0769 | 8.57002e-06 |
| 6 | rs1877314 | 4 | 164709468 | rs17764849 | 15 | 40483167 | -17.5639 | 3.9234 | 7.57961e-06 |
| 7 | rs11154637 | 6 | 131780207 | rs871437 | 11 | 132309082 | 0.4900 | 0.1000 | 9.67749e-07 |
| 8 | rs11159233 | 14 | 76295162 | rs12436912 | 14 | 85891516 | 0.3185 | 0.0702 | 5.76176e-06 |

Table 5.6: *Significant aggression towards a non-caregiver Epistasis Results*

sected with SFARI Gene. Gene set enrichment of those intersections included genes listed as syndromic, and with scores 1 and 2 of potential deleteriousness, but none with score 3 (the category with least support). Of intragenic SNPs, which overlap with SFARI Gene, significant associations include post-synapse, dendritic, and neuron projection processes, several synaptic cellular component locations, and glutamate binding functions. Genes were enriched in the "Glutamatergic synapse" KEGG pathway (all adjusted $p < 0.05$) - see Table 5.7. Likewise, genes not found in the SFARI Gene database were also highly enriched for synapse organization, and enriched in several synapse-related cellular components including the glutamatergic synapse specifically, as well the postsynaptic membrane and a variety of plasma membrane and intracellular vesicle locations. The genes were also enriched for participation in the KEGG calcium signaling pathway. While individual trait-level GWAS did fail to produce significant GWAS-level results, performing GWAS on the clusters of traits

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0179 | GO:0099173 | GO:BP | postsynapse organization |
| 2 | 0.0201 | GO:0097061 | GO:BP | dendritic spine organization |
| 3 | 0.0291 | GO:0106027 | GO:BP | neuron projection organization |
| 4 | 0.0353 | GO:0007399 | GO:BP | nervous system development |
| 5 | 0.0043 | GO:0032279 | GO:CC | asymmetric synapse |
| 6 | 0.0062 | GO:0098984 | GO:CC | neuron to neuron synapse |
| 7 | 0.0225 | GO:0098590 | GO:CC | plasma membrane region |
| 8 | 0.0499 | GO:0014069 | GO:CC | postsynaptic density |
| 9 | 0.0497 | GO:0016595 | GO:MF | glutamate binding |
| 10 | 0.0043 | KEGG:04724 | KEGG | Glutamatergic synapse |

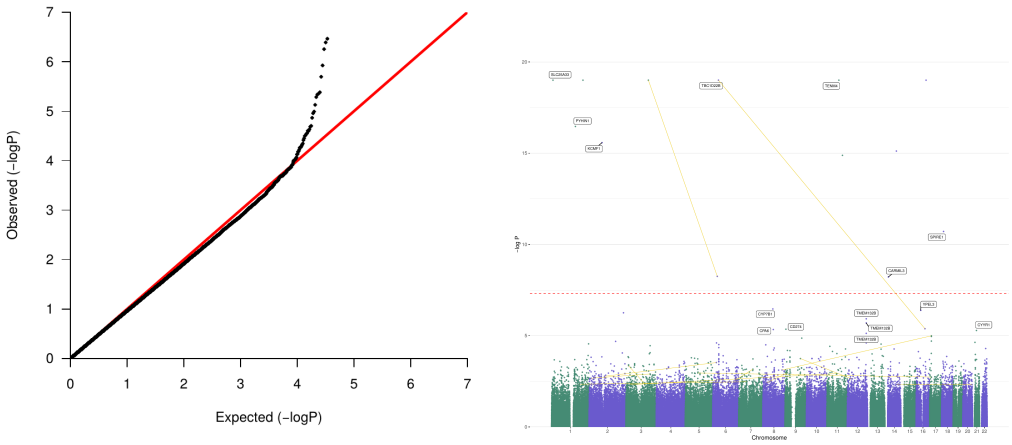
Table 5.7: *Gene ontology and KEGG gene set enrichment results of genes in SFARI Gene with significant SNPs from individual PNBO analysed traits. P-values are analytically corrected (permutation based) in GProfiler software.*

discovered in Chapter 4 indicate significant associations between phenotypic profiles and SNPs in the SSC cohort.

### 5.3.2 Phenome profile GWAS and epistasis studies reveal significant genetic associations for diverse autism phenotypes

GWAS and epistatic interaction tests were performed for each of the 14 clusters identified in Chapter 4. Results are all presented in order by cluster, from largest to smallest proband membership.

The turquoise cluster's traits include high level traits (memory ability and skill, playing behavior, and loss of self-help skills). It is associated with significantly interacting (Table 5.10) and single SNPs (Table 5.9, Figure 5.9). Enrichment produced few results (Table 5.11).



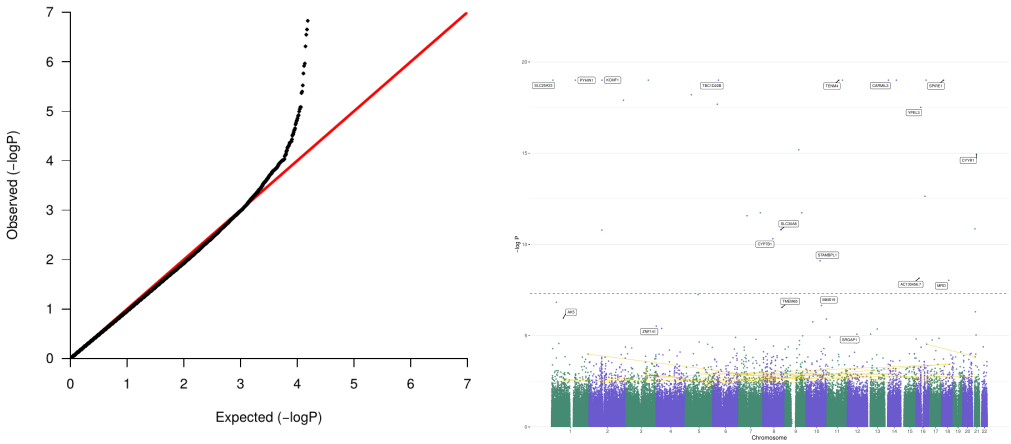

(a) (b)

Figure 5.9: *GWAS of the SSC cohort who belong to cluster 1, turquoise. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

The blue cluster singularly included one significantly associated trait, lack of reciprocal conversation. It is associated with significantly interacting (Table 5.13) and single SNPs (Table 5.12, Figure 5.10). The cluster is enriched for two protein interaction complexes (Table 5.14).

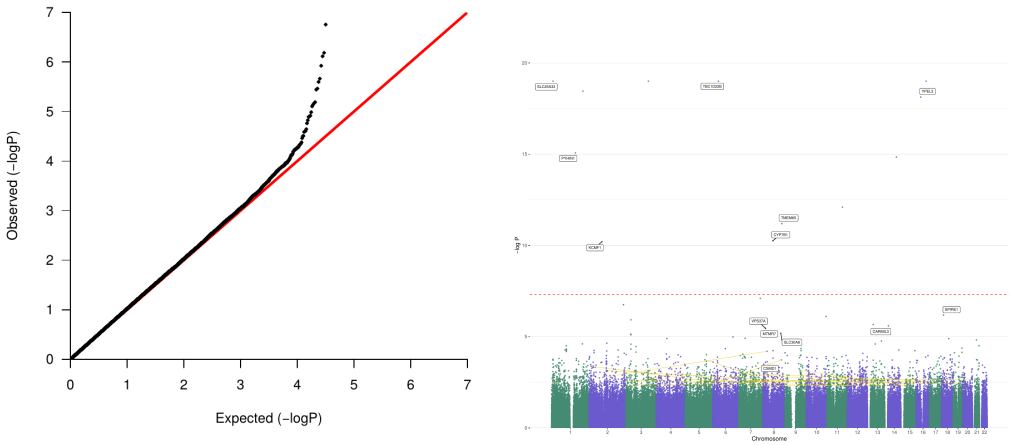
The brown cluster is phenotypically diverse, including high level (restricted behavior) and more specific (anxious behavior) traits. It is associated with significantly interacting (Table 5.16) and single SNPs (Table 5.15, Figure 5.11). The cluster is enriched for bile acid synthesis, and zinc homeostasis, among others (Table 5.17).

The yellow cluster is phenotypically diverse, but specific to communication and social traits. It is associated with significantly interacting (Table 5.19) and single SNPs (Table 5.18, Figure 5.12). The cluster is enriched for receptor localization to synapse processes (Table 5.20).



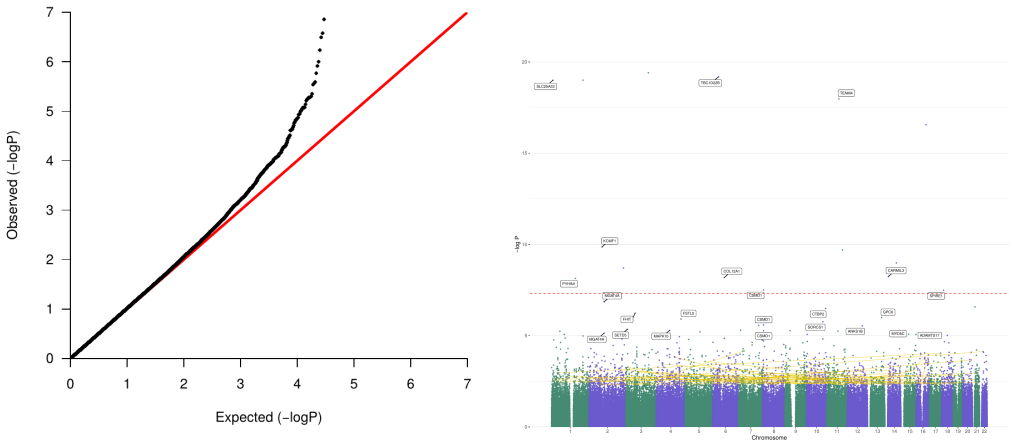

(a) (b)

Figure 5.10: *GWAS of the SSC cohort who belong to cluster 2, blue. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*




(a) (b)

Figure 5.11: *GWAS of the SSC cohort who belong to cluster 3, brown. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*



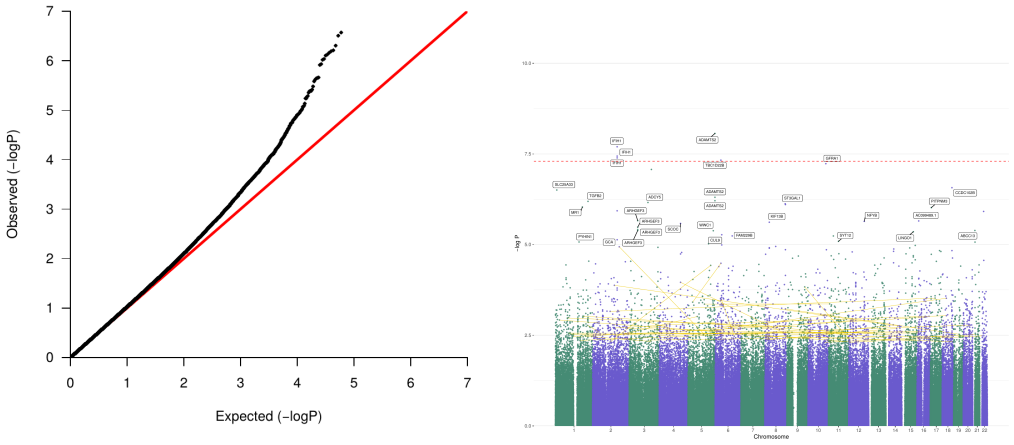

(a) (b)

Figure 5.12: *GWAS of the SSC cohort who belong to cluster 4, yellow. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

The green cluster's phenotypic signature is broad, including lack of communication (head nodding and loss of intent to communicate), lack of interest in children, and lack of offering to share one's enjoyment with others. It is associated with significantly interacting (Table 5.22) and single SNPs (Table 5.21, Figure 5.13). The cluster is enriched for T-cell leukemia virus 1 infection ( Table 5.23).

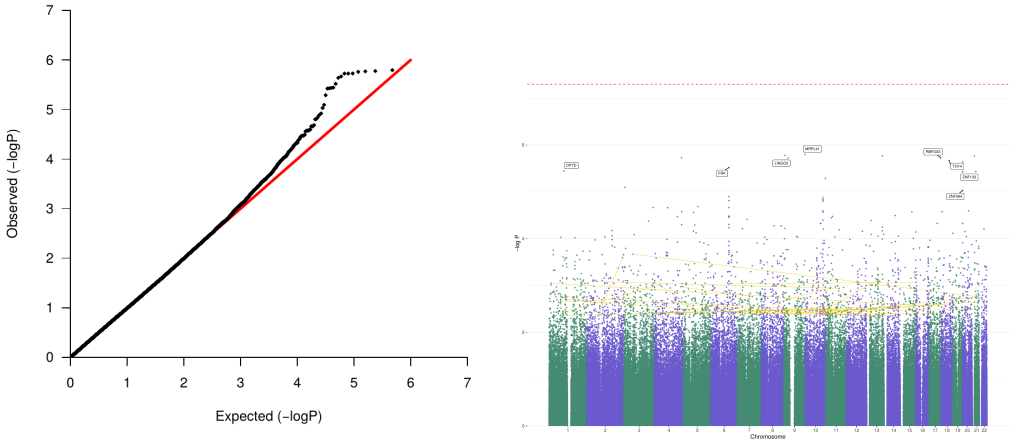
The red cluster's phenotypic signature represents above normal visiospatial ability and sensitivity to noise. It is associated with significantly interacting (Table 5.25) and single SNPs (Table 5.24, Figure 5.14). The cluster is enriched for one protein interaction complex (Table 5.26).

The black cluster's phenotypic signature is vague, including a lack of appropriate social response, abnormal reciprocal conversation, and abnormalities i voluntary movement behavior. It is associated with significantly interacting (Table 5.28) and single SNPs (Table




(a) (b)

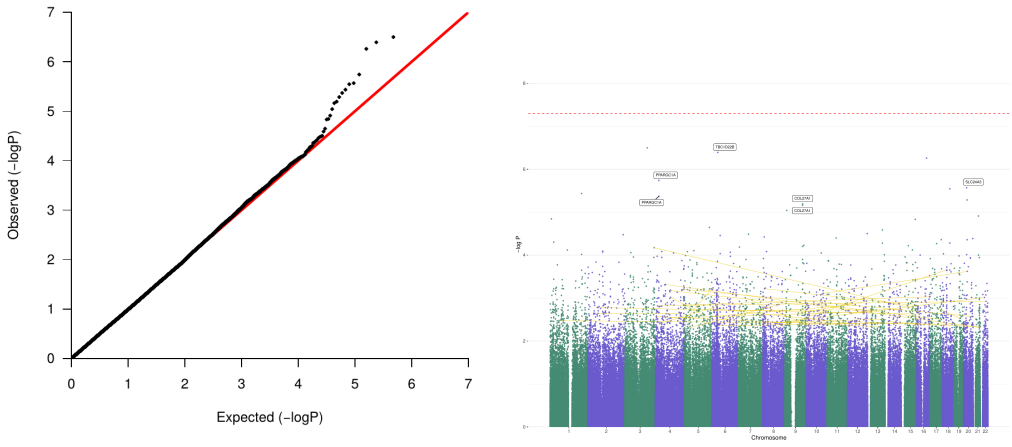
Figure 5.13: *GWAS of the SSC cohort who belong to cluster 5, green. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*




(a) (b)

Figure 5.14: *GWAS of the SSC cohort who belong to cluster 6, red. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

5.27, Figure 5.15). The cluster is associated with metabolism and energy use (Table 5.29).



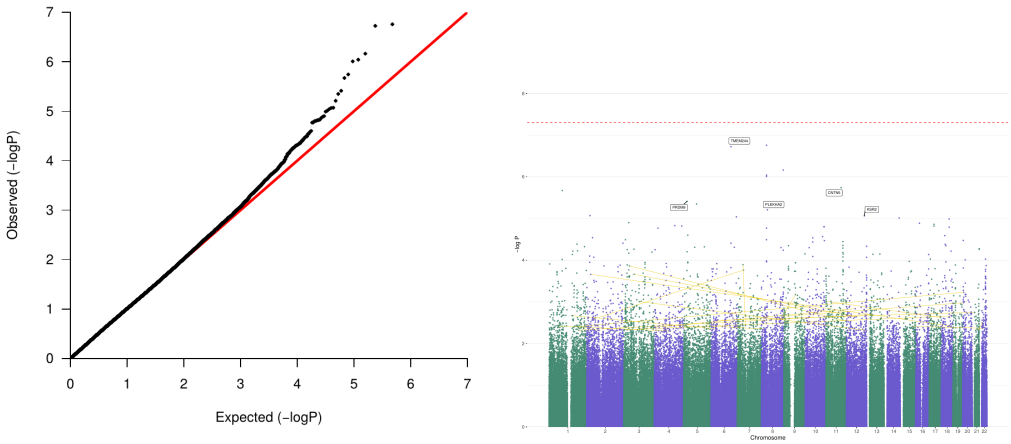

(a) (b)

Figure 5.15: *GWAS of the SSC cohort who belong to cluster 7, black. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

The pink cluster's phenotypic signature includes several types of restricted behaviors. It is associated with significantly interacting (Table 5.31) and single SNPs (Table 5.30, Figure 5.16). The cluster is associated with recombination hotspot binding (Table 5.32).

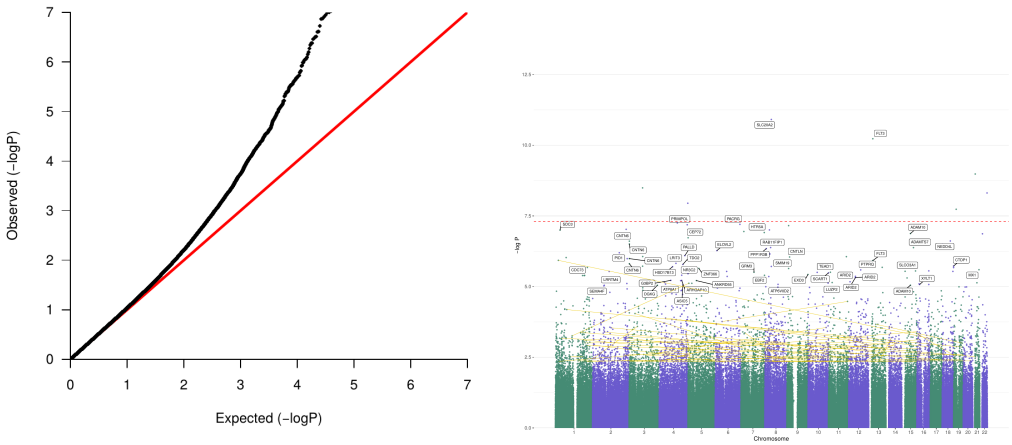
The magenta cluster's phenotypic signature is made up of poor coping mechanisms. It is associated with significantly interacting (Table 5.34) and single SNPs (Table 5.33, Figure 5.17). The cluster is associated with localization in the postsynaptic membrane (Table 5.35).

The purple cluster's phenotypic makeup involves self injury, sensitivity to noise, and social communication traits. It is associated with significantly interacting (Table 5.37) and single SNPs (Table 5.36, Figure 5.18). The cluster is associated with three interacting protein complexes involving DZIP1 (Table 5.38).



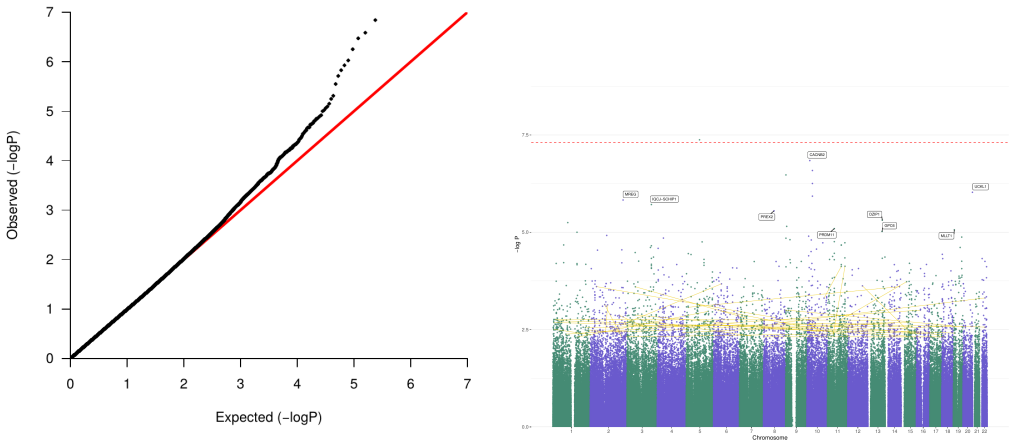

(a) (b)

Figure 5.16: *GWAS of the SSC cohort who belong to cluster 8, pink. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*




(a) (b)

Figure 5.17: *GWAS of the SSC cohort who belong to cluster 9, magenta. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*



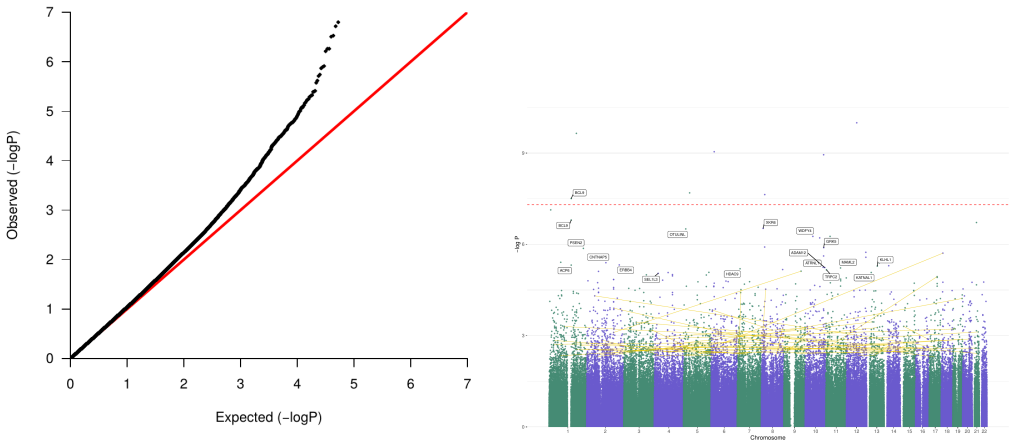

(a) (b)

Figure 5.18: *GWAS of the SSC cohort who belong to cluster 10, purple. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

The green-yellow cluster's phenotypic makeup is broad, including learning, playing, and communication traits. It is associated with significantly interacting (Table 5.40) and single SNPs (Table 5.39, Figure 5.19). The cluster is associated with Notch signaling pathways (Table 5.41).

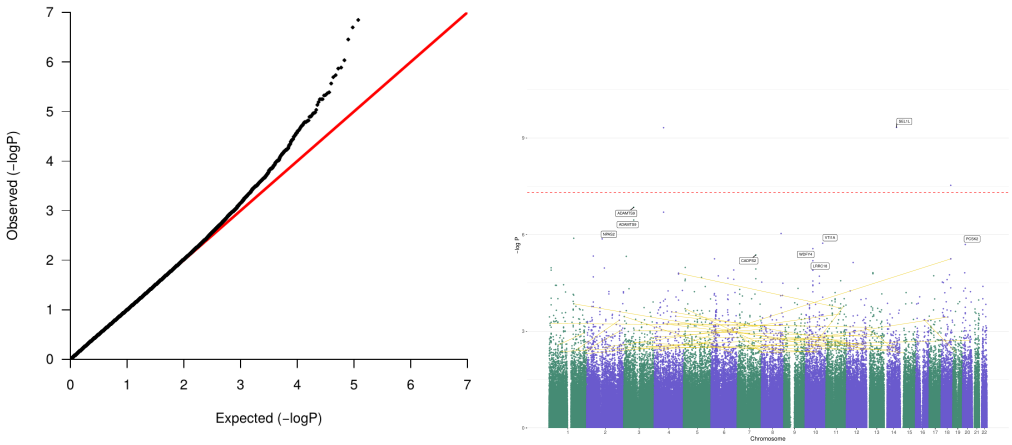
The tan cluster's phenotypic makeup is specific, involving head nodding and gait abnormalities. It is associated with significantly interacting (Table 5.43) and single SNPs (Table 5.42, Figure 5.20). The cluster is associated with circadian biology and the SNARE complex ( Table 5.44).

The salmon cluster's phenotypic makeup includes unusual sensory interests, coping with environmental changes, and asking socially inappropriate questions. It is associated with significantly interacting (Table 5.46) and single SNPs (Table 5.45, Figure 5.21). The cluster is associated with several protein complexes (Table 5.47).



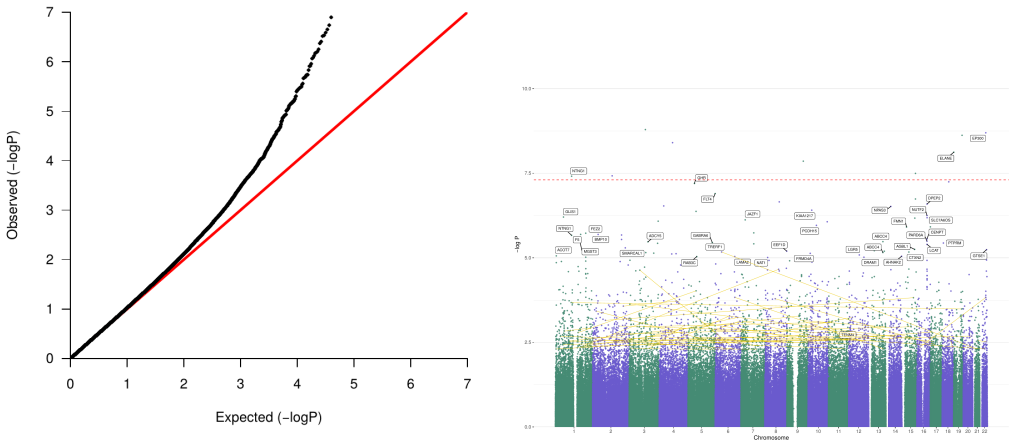

(a) (b)

Figure 5.19: *GWAS of the SSC cohort who belong to cluster 11, green-yellow. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*




(a) (b)

Figure 5.20: *GWAS of the SSC cohort who belong to cluster 12, tan. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*



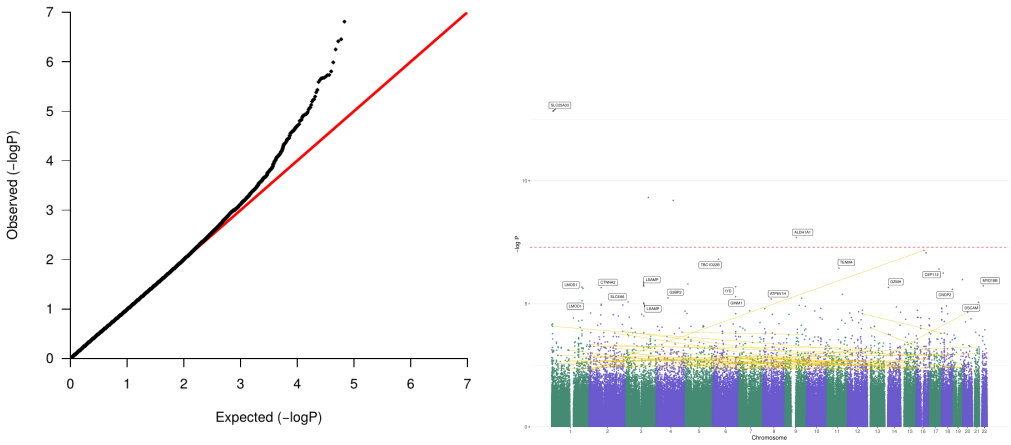

(a) (b)

Figure 5.21: *GWAS of the SSC cohort who belong to cluster 13, salmon. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

Lastly, the cyan cluster's phenotypic makeup includes head nodding. It is associated with significantly interacting (Table 5.49) and single SNPs (Table 5.48, Figure 5.22). The cluster is associated with the transcription factor PRDM1 and one protein complex (Table 5.50).

### 5.3.3 SNPs predict autism endophenotype-derived cluster membership

Without access to external datasets of autism patients phenotyped with the PNBO and genetic data available, I designed cluster classification experiments to test the biological validity of phenotype-derived clusters. My working hypothesis was that if SNPs associated to a cluster via GWAS can classify probands correctly into this cluster, then there may be




(a) (b)

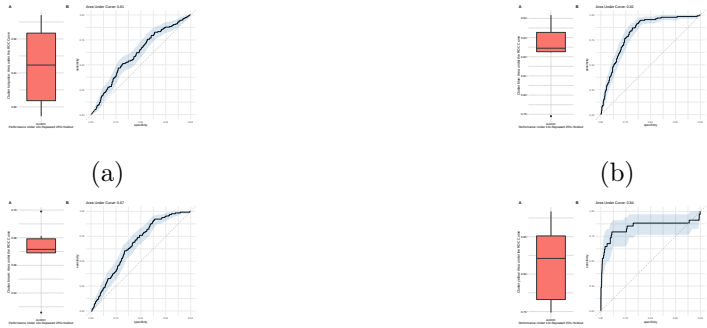
Figure 5.22: *GWAS of the SSC cohort who belong to cluster 14, cyan. Panel A shows an under powered QQ plot. Panel B shows the corresponding Manhattan plot. X-axis represents chromosome position, Y-axis represents the -log10 p-value of association, and yellow bands indicate epistatic interaction.*

a genetic etiology to the distinct phenotype profile of that cluster. To validate this, 1/4 of probands in a cluster were excluded from classifier training and used to test the performance of the classifier.

The coded allele at each SNP (major, minor, missing) made up features used for classifying SNPs into their cluster. The experiment was repeated 10 times, and the median Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) values retained. The worst performing was an AUC of 0.61 (Figure 5.23a in the turquoise module (the largest). The highest was in the tan cluster, the 12th smallest, with a median AUC of 0.94, seen in Figure 5.25d.

While each cluster was evaluated with 10 different train/test splits (drawn at random), few showed large AUC score variance, as indicated by a lack of outliers (outside of 1.5 times the interquartile range). Predictive models, which accurately captured cluster membership,

include those used to classify turquoise, yellow, green, red, black, purple, green-yellow, tan, salmon, and cyan; see Figures 5.23d, 5.24a, 5.24b, 5.24c, 5.25b, 5.25c, 5.26a, and 5.26b. Others were more variable including the blue, brown, pink, and magenta models; see Figures 5.23b, 5.23c, 5.24d, and 5.25a.



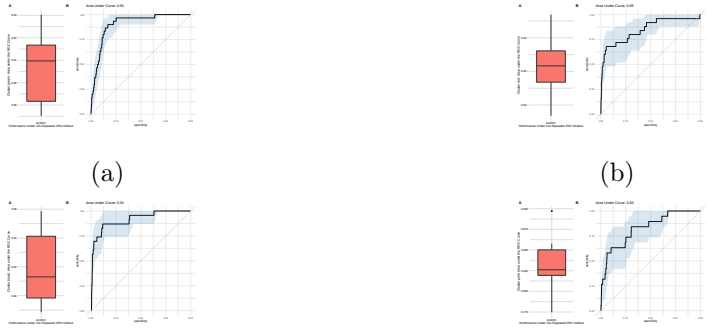
(a) (b)

(c) (d)

Figure 5.23: *Predictive ability of SNPs identified from GWAS in the turquoise, blue, brown, and yellow cluster to predict cluster membership of 25% holdout validation set, repeated 10 times. Panels A presents areas under the ROC curve for 10 repeated 75/25 train/test splits. The median-performing ROC curve is plotted in Panels B.*

### 5.3.4 Phenome-wide network reveals gene-driven relationships between autism related traits

Thus far, each analysis focused on on a single trait or cluster. Each cluster of probands is made up of multiple traits, many intersecting. To view a phenome-wide map of genes influencing their several traits in clusters, a graph, whose edges represent the -log P-value of association between a gene and a trait, with wider edges indicating a stronger association, was generated . Only SNPs associating with known genes are included, and interactions are collapsed to the level of the gene. The only edges in the graph are links between genes and traits; no links between PNBO traits from the ontology, or links between genes sharing an epistatic interaction, are shown, Figure 5.27.



(a) (b) (c) (d)

Figure 5.24: *Predictive ability of SNPs identified from GWAS in the green, red, black, and pink cluster to predict cluster membership of 25% holdout validation set, repeated 10 times. Panels A presents areas under the ROC curve for 10 repeated 75/25 train/test splits. The median-performing ROC curve is plotted in Panels B.*

(a) (b) (c) (d)

Figure 5.25: *Predictive ability of SNPs identified from GWAS in the magenta, purple, greenyellow, and tan cluster to predict cluster membership of 25% holdout validation set, repeated 10 times. Panels A presents areas under the ROC curve for 10 repeated 75/25 train/test splits. The median-performing ROC curve is plotted in Panels B.*

(a) (b)

Figure 5.26: *Predictive ability of SNPs identified from GWAS in the salmon and cyan cluster to predict cluster membership of 25% holdout validation set, repeated 10 times. Panels A presents areas under the ROC curve for 10 repeated 75/25 train/test splits. The median-performing ROC curve is plotted in Panels B.*

Traits are in green, genes are in white, with a data-driven edge-weighted force directed layout show. From this bird's eye view of gene/trait associations made in this chapter, it is apparent that loss of language ability, social verbalization and response, and learning and/or memory behavioral phenotypes are with an interconnected cluster of genes, which reflects results observed in the green-yellow cluster. Another cluster of genes connects traits related to the visiospatial ability and sensitivity to noise from cluster red.

## 5.4 Discussion

In this chapter, 64 GWAS and epistatic analysis experiments were performed, and 14 were validated by holdout. At the endophenotype level, GWAS failed to find significant associations between alleles and traits. In the SSC cohort, nearly 40 out of 100 individual traits will map to an individual, giving each proband a mix of seemingly uncorrelated traits which may be expressed more or less strongly. Because traits are assigned from the ADI-R instrument, subjectivity in interpreting answers may bias how accurate assignments of individual traits are. Moreover, these findings are not unique to this work. Chaste and colleagues performed GWAS analyses on the SSC cohort (Chaste et al., 2015). They manually split the cohort into 10 overlapping groups for separate analyses based on autism diagnosis, IQ,

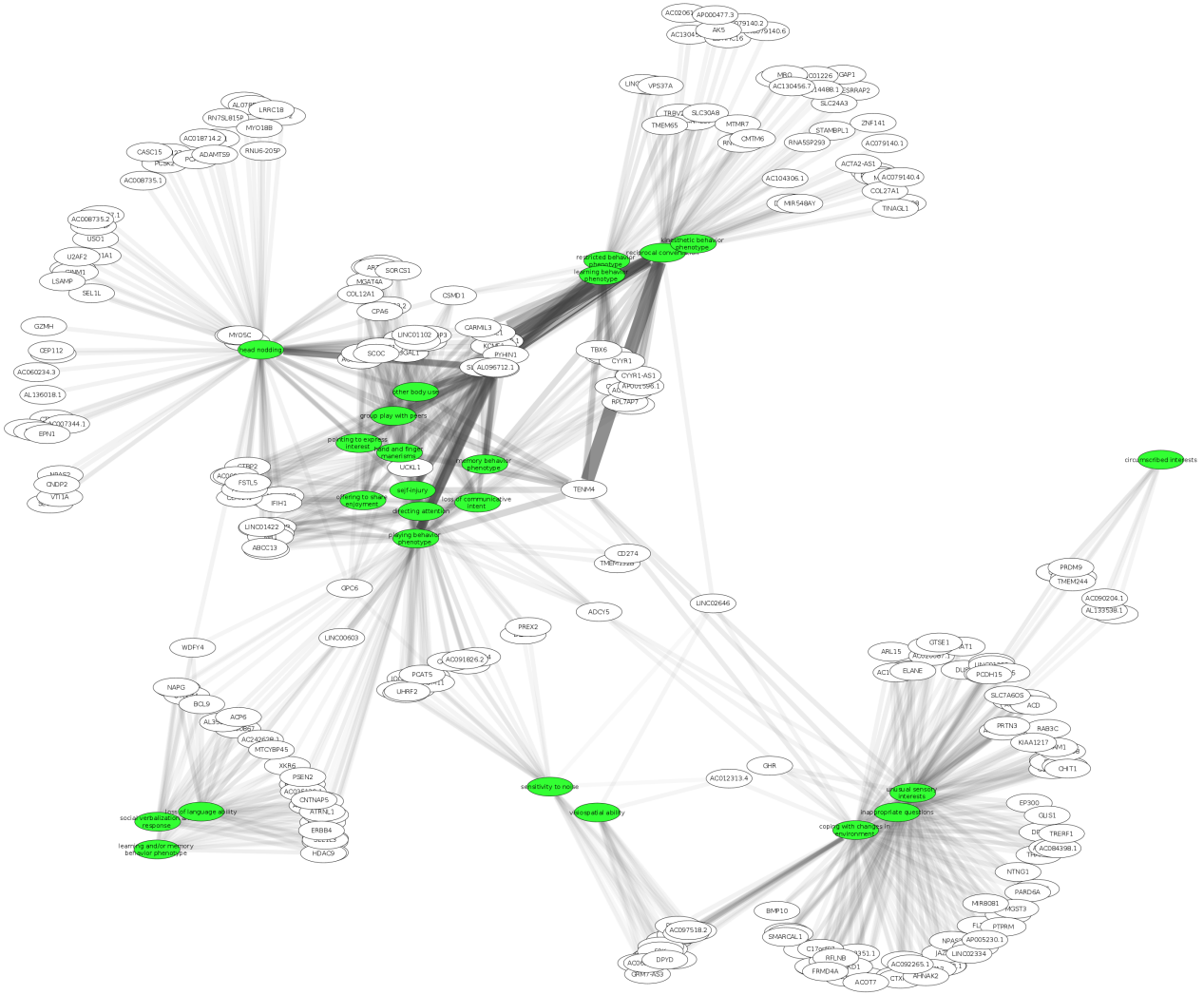



Figure 5.27: *Significant gene/trait associations from SSC cohort cluster GWAS and gene/gene interactions from epistasis studies are connected as a network. Wider and darker edges represent stronger associations and lower p-values. PNBO traits are in green, genes are in white. The layout is an edge-weighted force directed layout, simulated in Cytoscape.*

having circumscribed interests in the ADI-R, difficulty with change in the ADI-R, sensitivity to noise, and from measurements derived from the Autism Diagnostic Observation Schedule (Lord et al., 1989). They reported no genome-wide significant signal, hypothesizing that among the phenotypically similar SSC cohort divisions based on phenotype are still to homogeneous. In each case of trait selection, rather than split the cohort into groups based of phenotypic profile, they analysed single traits following the approach described in the beginning of this chapter. Epistatic interaction analysis did not improve the signal, although evidence of epistatic interactions associated with individual phenotypes ($p < 1e\text{-}6$), when no GWAS association was identified, was found. This suggests that the non-additive effect of SNPs making a small contribution toward a phenotype may explain more of the aberrant traits than single SNP effects.

### 5.4.1 Cluster-based GWAS find novel genomic associations for autism traits

Contrary to single trait-based GWAS, cluster GWAS and epistatic analysis produced genome-wide significant results which were not seen at the individual trait level. These associations included 27 SNPs associated with their cluster's traits at $p < 1e\text{-}20$, far exceeding the genome-wide standard of 1e-8. To validate clusters GWAS associations, supervised learning experiments aiming to classify probands into their cluster were performed. After training a classifier to place probands in their cluster or the rest of the SSC population, holdout probands were then tested. Previous approaches have used GWAS summary statistics and expected AUCs of around 0.75 for a well-performing classifier (Patron et al., 2019). Ours performed much better, with several GWAS validating with AUCs above 0.9, up to 0.94 in the tan module. Their model, however, did not include SNPs from epistatic interactions, which were not found in GWAS, and did not perform feature selection via the LASSO. Ad-

ditionally, I made artificial epistatic interactions between all SNPs in my GWAS validation classifier by including interaction terms for each SNP used to attempt to classify probands. This may have explained the increased performance among some clusters, as complex disease is often underpinned by complex genetic models. The two largest clusters had the lowest AUC scores assigned, likely indicative of the very high level phenotypic signature which defined the large number of probands in the two clusters. Given that most probands were assigned play-related behavior traits in PNBO, which make up the majority of the turquoise cluster's phenotypic signature, this is not surprising.

**Core circadian genes associate with communication and gait phenotypes**

Interestingly, the best preforming validation ROC score was identified for the second smallest "tan" cluster, see Figure 5.25d. Of nine SNPs within protein coding regions, one SNP rs4851384 is embedded in the intron of NPAS2. NPAS2 was discussed in Chapter 2 as an under-annotated circadian gene; it clearly has a biological function in the core circadian pacemaker but, due to being a paralog of the CLOCK gene, does not yet have associations with abnormal circadian rhythms in mouse. A survey of the other nine genes in this cluster reveal CADPS2, whose mouse homolog is annotated in the Disease Ontology (Schriml et al., 2012) with ASD. None of these genes are currently annotated to their cluster's phenotypic profile of head nodding and abnormal gait. Cadps2 is associated with abnormal circadian behavior (further strengthening the module's circadian connection) and abnormal cerebellum morphology, affecting Purkinje and granule cells (Bult et al., 2019b). There are several PNBO-enoded ASD traits which involve fine motor skills, controlled by the cerebellum. Mouse models of Fragile-X syndrome often exhibit neurodegeneration, abnormal gait, and autism characteristics while also being implicated in Purkinje cell dysfunction (Tsai et al., 2012; Sundberg and Sahin, 2015). Given this evidence, behavioral mouse phenotyping may be needed on Sel1l2 and Lrrc18, which have not been phenotyped by the International Mouse

Phenotyping Consortium. The IMPC SHIRPA exam, which includes observations of gait in all screens, may be sufficient to prompt further behavioral assays (Hampson and Blatt, 2015; Brown and Moore, 2012a; Hatcher et al., 2001). This cross-species analysis demonstrates the utility of combining model organism with human (and potentially translational) research, especially in the field of neurobehavioural genetics.

### 5.4.2 Gene-phenotype network highlights novel gene/trait relationships

Ultimately, this chapter was able to associate multiple genetic loci with multiple phenotypic traits. No single locus was strongly associated with any of the 14 clusters, suggesting that networks of genes acting together in *trans* with small effect sizes are together responsible for complex traits.

However, whereas most GWAS studies would associate multiple loci with a single trait, by performing GWAS on clusters of traits, effectively a single dependant variable as a substitute for a dependent multivariate model used. This produces a many - to - many annotation problem, where many SNPs are each associated with a set of traits in combination. How does one annotate a gene to a profile of traits, if each trait has a different probability of representing the cluster?

The gene/trait network, Figure 5.27, is one approach for addressing this question. Underling that network is the probability of a trait associating with a cluster, multiplied by the strength of the association. Thus, each SNP may be more or less well annotated with multiple PNBO-derived traits. A threshold of 0.5 was set to inform the annotation graph, meaning that if a trait has a marginal probability of characterising a phenotype cluster of at least 0.5 (from Chapter 4), a SNP/trait association was made. For example,

strong associations exist between "coping with a changing environment" and genes linked to sensitivity to noise. SNPs linked to many behavioral traits associated with social phenotypes such as group pay with peers and pointing to express interest. They also associate with self-injury and repetitive hand and finger mannerisms in a different part of the graph. Of more interest, the periphery of the graph shows loss of language ability, social verbalization and response, and a high level learning or memory phenotype associated with a cluster of genes. These traits are isolated from the others parts of the graph. The genes associated to them are enriched for Notch signaling, indicating roles in developmental biology. These traits also uniquely defined a cluster phenotypically; the same is true of sensitivity to noise and visiospatial ability. These results suggest that while phenotypic heterogeneity among the SSC cohort may be difficult to untangle, there is a likely shared genetic basis for characteristic behaviors of some phenotypically similar probands in the SSC cohort.

## 5.5 Conclusions and Chapter Summary

In this chapter, genetic associations between autism spectrum disorder endophenotypes mapped in the Psychological Neuro Behavior Ontology (PNBO) among families with probands and unaffected siblings were investigated. Genetic associations between SNPs and individual traits were statistically underpowered, and produced no genome-wide significant results. When using mixed models and surveying clustered traits, several genome-wide significant associations were found. GWAS of PNBO profile clusters were well powered, indicating a strong correlation between genotypic and phenotypic homogeneity. Epistasis models performed on clustered traits unveiled gene/gene interactions passing strict genome-wide significance levels. In lieu of available cohorts for GWAS replication studies, predictive models tested the performance of SNPs identified through GWAS and epistatic interactions to classify probands by phenotypic profile. This novel approach showed that diverse autism phenotype profiles

can be classified by genetics. Lastly, specific trait/trait relations were found when surveying significantly enriched genetic activity amongst phenotype clusters. The best-performing predictive model was annotated with genes involved in circadian rhythm biology, revealing a subset of probands who may have circadian abnormalities though have not been tested for it. Thus, this chapter links autism spectrum disorder and chronology indirectly, suggesting an underlying polygenic link between circadian abnormalities and communicative and movement related traits.

Following are cluster-specific tabular results, including SNPs and associations to each cluster's trait profile both directly and via epistatic interactions with other loci. Additional tables incldude gene enrichment results for significant intragenic SNPs. After the last result tables are presented for each SSC cluster, the next chapter dicusses the project as a comprehensive whole.

## 5.6 Cluster Result Tables

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | 0.2252 | 0.0094 | 4.379245e-119 | SLC25A33 |
| 2 | rs3008409 | 1 | 31814425 | 0.0565 | 0.0107 | 1.487914e-07 | |
| 3 | rs2602948 | 1 | 77530012 | 0.0350 | 0.0072 | 1.08932e-06 | AK5 |
| 4 | rs16841336 | 1 | 157167751 | 0.0851 | 0.0058 | 3.525406e-48 | PYHIN1 |
| 5 | rs10779486 | 1 | 206805932 | 0.1379 | 0.0074 | 1.619351e-75 | |
| 6 | rs4832336 | 2 | 83911474 | -0.0479 | 0.0071 | 1.645032e-11 | |
| 7 | rs17025837 | 2 | 85136390 | 0.0896 | 0.0063 | 2.179066e-44 | KCMF1 |
| 8 | rs2396261 | 2 | 226253774 | 0.0658 | 0.0074 | 1.259342e-18 | |
| 9 | rs11925421 | 3 | 147370852 | 0.1927 | 0.0086 | 6.195197e-107 | |
| 10 | rs11729526 | 4 | 323310 | 0.0383 | 0.0082 | 3.009371e-06 | ZNF141 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11 | rs4833120 | 4 | 35516217 | -0.0281 | 0.0061 | 4.012208e-06 | |
| 12 | rs10053502 | 5 | 40014929 | -0.0748 | 0.0084 | 6.297018e-19 | |
| 13 | rs16901158 | 5 | 83891214 | -0.0544 | 0.0100 | 5.553335e-08 | |
| 14 | rs2394173 | 6 | 29881137 | 0.0616 | 0.0070 | 2.082146e-18 | |
| 15 | rs6914506 | 6 | 37374934 | -0.1152 | 0.0049 | 1.470992e-115 | TBC1D22B |
| 16 | rs331833 | 7 | 54629782 | -0.0530 | 0.0076 | 2.667352e-12 | |
| 17 | rs17837474 | 7 | 142056577 | -0.0568 | 0.0080 | 1.8312e-12 | |
| 18 | rs16931350 | 8 | 65701037 | -0.0581 | 0.0088 | 4.863796e-11 | CYP7B1 |
| 19 | rs3019885 | 8 | 118094826 | 0.0398 | 0.0059 | 1.58287e-11 | SLC30A8 |
| 20 | rs12114111 | 8 | 125398550 | -0.0292 | 0.0057 | 2.843439e-07 | TMEM65 |
| 21 | rs1887387 | 9 | 90069741 | 0.0464 | 0.0057 | 6.535233e-16 | |
| 22 | rs4978649 | 9 | 109716918 | 0.0433 | 0.0061 | 1.832544e-12 | |
| 23 | rs11238776 | 10 | 42098785 | 0.1046 | 0.0218 | 1.723645e-06 | |
| 24 | rs7908852 | 10 | 90684773 | 0.0855 | 0.0139 | 7.865417e-10 | STAMBPL1 |
| 25 | rs11818837 | 10 | 99209983 | 0.1540 | 0.0297 | 2.230238e-07 | MMS19 |
| 26 | rs7913323 | 10 | 132050446 | 0.0340 | 0.0070 | 1.217173e-06 | |
| 27 | rs7946005 | 11 | 78464992 | 0.1297 | 0.0072 | 1.442226e-69 | TENM4 |
| 28 | rs11225401 | 11 | 102113169 | 0.0968 | 0.0061 | 6.03376e-56 | |
| 29 | rs17100060 | 12 | 62739380 | -0.0335 | 0.0075 | 8.276557e-06 | SRGAP1 |
| 30 | rs9316212 | 13 | 20727309 | -0.0348 | 0.0078 | 8.25745e-06 | |
| 31 | rs2137512 | 13 | 64399588 | -0.0314 | 0.0068 | 4.303749e-06 | |
| 32 | rs1003761 | 14 | 23602251 | 0.0914 | 0.0094 | 5.07282e-22 | CARMIL3 |
| 33 | rs4903419 | 14 | 76054408 | 0.0964 | 0.0061 | 4.322848e-55 | |
| 34 | rs4780805 | 16 | 19312146 | 0.0413 | 0.0071 | 7.015834e-09 | AC130456.7 |
| 35 | rs13330491 | 16 | 30012701 | -0.0684 | 0.0078 | 3.076694e-18 | YPEL3 |
| 36 | rs13329856 | 16 | 57491588 | -0.0598 | 0.0081 | 2.302452e-13 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 37 | rs235083 | 16 | 64760511 | -0.1138 | 0.0053 | 1.744564e-99 | |
| 38 | rs486743 | 18 | 12495089 | -0.0742 | 0.0052 | 1.134169e-44 | SPIRE1 |
| 39 | rs2849233 | 18 | 46585551 | 0.0334 | 0.0058 | 9.244312e-09 | MRO |
| 40 | rs11910489 | 21 | 16968775 | 0.0825 | 0.0122 | 1.403848e-11 | |
| 41 | rs4596064 | 21 | 19786452 | 0.0394 | 0.0078 | 4.893777e-07 | |
| 42 | rs2828789 | 21 | 24363246 | 0.0617 | 0.0139 | 9.103979e-06 | |
| 43 | rs17002187 | 21 | 26762664 | -0.0460 | 0.0057 | 1.1432e-15 | CYYR1 |

Table 5.12: *Significant blue GWAS Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0235 | CORUM:6807 | CORUM | ADRA1A-CXCR4 complex |
| 2 | 0.0205 | GO:0050808 | GO:BP | synapse organization |
| 3 | 0.0000 | GO:0098978 | GO:CC | glutamatergic synapse |
| 4 | 0.0002 | GO:0030054 | GO:CC | cell junction |
| 5 | 0.0006 | GO:0045202 | GO:CC | synapse |
| 6 | 0.0007 | GO:0071944 | GO:CC | cell periphery |
| 7 | 0.0007 | GO:0005886 | GO:CC | plasma membrane |
| 8 | 0.0008 | GO:0097060 | GO:CC | synaptic membrane |
| 9 | 0.0013 | GO:0098794 | GO:CC | postsynapse |
| 10 | 0.0020 | GO:0099699 | GO:CC | integral component of synaptic membrane |
| 11 | 0.0020 | GO:0045211 | GO:CC | postsynaptic membrane |
| 12 | 0.0038 | GO:0099240 | GO:CC | intrinsic component of synaptic membrane |
| 13 | 0.0070 | GO:0099056 | GO:CC | integral component of presynaptic membrane |
| 14 | 0.0112 | GO:0005901 | GO:CC | caveola |
| 15 | 0.0120 | GO:0098889 | GO:CC | intrinsic component of presynaptic membrane |
| 16 | 0.0156 | GO:0016020 | GO:CC | membrane |
| 17 | 0.0185 | GO:0031012 | GO:CC | extracellular matrix |
| 18 | 0.0185 | GO:0031226 | GO:CC | intrinsic component of plasma membrane |
| 19 | 0.0185 | GO:0099055 | GO:CC | integral component of postsynaptic membrane |
| 20 | 0.0217 | GO:0005887 | GO:CC | integral component of plasma membrane |
| 21 | 0.0217 | GO:0098590 | GO:CC | plasma membrane region |
| 22 | 0.0217 | GO:0098936 | GO:CC | intrinsic component of postsynaptic membrane |
| 23 | 0.0250 | GO:0042383 | GO:CC | sarcolemma |
| 24 | 0.0321 | GO:0098793 | GO:CC | presynapse |
| 25 | 0.0321 | GO:0062023 | GO:CC | collagen-containing extracellular matrix |
| 26 | 0.0321 | GO:0045121 | GO:CC | membrane raft |
| 27 | 0.0321 | GO:0098857 | GO:CC | membrane microdomain |
| 28 | 0.0321 | GO:0044853 | GO:CC | plasma membrane raft |
| 29 | 0.0435 | GO:0098589 | GO:CC | membrane region |
| 30 | 0.0492 | GO:0097708 | GO:CC | intracellular vesicle |
| 31 | 0.0492 | GO:0031224 | GO:CC | intrinsic component of membrane |
| 32 | 0.0492 | GO:0031410 | GO:CC | cytoplasmic vesicle |
| 33 | 0.0045 | GO:0005509 | GO:MF | calcium ion binding |
| 34 | 0.0145 | KEGG:04020 | KEGG | Calcium signaling pathway |

Table 5.8: *Gene ontology and KEGG gene set enrichment results of genes not annotated in SFARI Gene, which have significant SNPs from individual PNBO analysed traits. P-values are analytically corrected (permutation based) in GProfiler software.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | -0.1252 | 0.0106 | 1.355673e-31 | SLC25A33 |
| 2 | rs16841336 | 1 | 157167751 | -0.0536 | 0.0063 | 3.442908e-17 | PYHIN1 |
| 3 | rs10779486 | 1 | 206805932 | -0.0820 | 0.0081 | 8.339206e-24 | |
| 4 | rs17025837 | 2 | 85136390 | -0.0571 | 0.0069 | 2.64675e-16 | KCMF1 |
| 5 | rs2396261 | 2 | 226253774 | -0.0405 | 0.0081 | 5.565648e-07 | |
| 6 | rs11925421 | 3 | 147370852 | -0.1013 | 0.0096 | 1.063148e-25 | |
| 7 | rs2394173 | 6 | 29881137 | -0.0444 | 0.0076 | 5.607052e-09 | |
| 8 | rs6914506 | 6 | 37374934 | 0.0718 | 0.0055 | 2.449977e-38 | TBC1D22B |
| 9 | rs16931350 | 8 | 65701037 | 0.0487 | 0.0095 | 3.433225e-07 | CYP7B1 |
| 10 | rs7341614 | 8 | 68723993 | 0.1358 | 0.0296 | 4.600234e-06 | CPA6 |
| 11 | rs10125854 | 9 | 5454065 | 0.0630 | 0.0137 | 4.417458e-06 | CD274 |
| 12 | rs7946005 | 11 | 78464992 | -0.0799 | 0.0079 | 1.167039e-23 | TENM4 |
| 13 | rs11225401 | 11 | 102113169 | -0.0536 | 0.0067 | 1.316471e-15 | |
| 14 | rs4765231 | 12 | 124270220 | -0.0423 | 0.0087 | 1.187936e-06 | TMEM132B |
| 15 | rs4765028 | 12 | 124278681 | -0.0368 | 0.0082 | 7.45518e-06 | TMEM132B |
| 16 | rs11058097 | 12 | 124287014 | -0.0400 | 0.0084 | 2.012821e-06 | TMEM132B |
| 17 | rs1003761 | 14 | 23602251 | -0.0598 | 0.0103 | 6.047103e-09 | CARMIL3 |
| 18 | rs4903419 | 14 | 76054408 | -0.0542 | 0.0067 | 7.658302e-16 | |
| 19 | rs13330491 | 16 | 30012701 | 0.0431 | 0.0085 | 4.075843e-07 | YPEL3 |
| 20 | rs13329856 | 16 | 57491588 | 0.0407 | 0.0088 | 4.146549e-06 | |
| 21 | rs235083 | 16 | 64760511 | 0.0623 | 0.0059 | 5.425392e-26 | |
| 22 | rs486743 | 18 | 12495089 | 0.0387 | 0.0057 | 1.949031e-11 | SPIRE1 |
| 23 | rs17002187 | 21 | 26762664 | 0.0284 | 0.0062 | 5.17815e-06 | CYYR1 |

Table 5.9: *Significant turquoise GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs12089482 | 1 | 183864117 | rs1689310 | 7 | 52983956 | 1.0042 | 0.2159 | 3.3078e-06 | | |
| 2 | rs10779486 | 1 | 206805932 | rs11225401 | 11 | 102113169 | -0.9935 | 0.1759 | 1.6097e-08 | | |
| 3 | rs10779486 | 1 | 206805932 | rs4903419 | 14 | 76054408 | -0.8547 | 0.1757 | 1.14832e-06 | | |
| 4 | rs748615 | 2 | 13250777 | rs16827846 | 3 | 158771579 | -16.9912 | 3.8248 | 8.89616e-06 | | |
| 5 | rs12712101 | 2 | 101426229 | rs9393387 | 6 | 23396822 | 0.3867 | 0.0845 | 4.6701e-06 | | |
| 6 | rs2396261 | 2 | 226253774 | rs6914506 | 6 | 37374934 | 0.4563 | 0.0921 | 7.21786e-07 | | TBC1D22B |
| 7 | rs11549705 | 3 | 17107009 | rs3749213 | 3 | 123616520 | -0.8504 | 0.1891 | 6.92634e-06 | | |
| 8 | rs11549705 | 3 | 17107009 | rs10934592 | 3 | 123617932 | -0.8506 | 0.1892 | 6.92298e-06 | | |
| 9 | rs17018468 | 3 | 26662581 | rs2508191 | 11 | 63509558 | -16.2046 | 3.6469 | 8.85541e-06 | | |
| 10 | rs11925421 | 3 | 147370852 | rs2394173 | 6 | 29881137 | -1.4784 | 0.3183 | 3.4118e-06 | | |
| 11 | rs11925421 | 3 | 147370852 | rs4903419 | 14 | 76054408 | -1.0787 | 0.2247 | 1.58633e-06 | | |
| 12 | rs11925421 | 3 | 147370852 | rs486743 | 18 | 12495089 | 0.9562 | 0.1937 | 7.996e-07 | | SPIRE1 |
| 13 | rs11925421 | 3 | 147370852 | rs7278383 | 21 | 27500390 | 1.7747 | 0.3768 | 2.47226e-06 | | |
| 14 | rs10053502 | 5 | 40014929 | rs486743 | 18 | 12495089 | -0.4105 | 0.0892 | 4.20932e-06 | | SPIRE1 |
| 15 | rs6452463 | 5 | 81868125 | rs11651038 | 17 | 11167226 | -0.5360 | 0.1203 | 8.29945e-06 | | |
| 16 | rs6452463 | 5 | 81868125 | rs1546560 | 17 | 11169414 | -0.5363 | 0.1203 | 8.24343e-06 | | |
| 17 | rs6914506 | 6 | 37374934 | rs13329856 | 16 | 57491588 | -0.4824 | 0.0972 | 6.99962e-07 | TBC1D22B | |
| 18 | rs6914506 | 6 | 37374934 | rs235083 | 16 | 64760511 | -0.4957 | 0.0713 | 3.69418e-12 | TBC1D22B | |
| 19 | rs480456 | 6 | 147889216 | rs4257948 | 7 | 93578688 | -0.4184 | 0.0936 | 7.73479e-06 | | |
| 20 | rs4263777 | 8 | 25783203 | rs551112 | 9 | 14433468 | -0.8769 | 0.1952 | 7.07747e-06 | | |
| 21 | rs10429356 | 8 | 126150662 | rs4792121 | 17 | 11141667 | -0.5742 | 0.1291 | 8.67985e-06 | | |
| 22 | rs12686783 | 9 | 100901588 | rs3210635 | 12 | 26955499 | -0.6130 | 0.1353 | 5.86506e-06 | | |
| 23 | rs7946005 | 11 | 78464992 | rs11225401 | 11 | 102113169 | -0.7171 | 0.1576 | 5.35793e-06 | TENM4 | |
| 24 | rs7946005 | 11 | 78464992 | rs4903419 | 14 | 76054408 | -0.7035 | 0.1585 | 9.02146e-06 | TENM4 | |
| 25 | rs7946005 | 11 | 78464992 | rs235083 | 16 | 64760511 | 0.6037 | 0.1325 | 5.24508e-06 | TENM4 | |
| 26 | rs7946005 | 11 | 78464992 | rs486743 | 18 | 12495089 | 0.7445 | 0.1258 | 3.28291e-09 | TENM4 | SPIRE1 |
| 27 | rs11225401 | 11 | 102113169 | rs235083 | 16 | 64760511 | 0.4424 | 0.0915 | 1.3385e-06 | | |
| 28 | rs10162458 | 14 | 77174941 | rs2295000 | 20 | 61007383 | -0.6148 | 0.1284 | 1.6795e-06 | | |
| 29 | rs235083 | 16 | 64760511 | rs486743 | 18 | 12495089 | -0.2986 | 0.0662 | 6.56885e-06 | | SPIRE1 |
| 30 | rs7186168 | 16 | 86543637 | rs4792121 | 17 | 11141667 | 0.4059 | 0.0918 | 9.81487e-06 | | |

Table 5.10: *Significant turquoise Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0499 | REAC:R-HSA-5579013 | REAC | Defective CYP7B1 causes Spastic paraplegia 5A, autosomal recessive (SPG5A) and Congenital bile acid synthesis defect 3 (CBAS3) |

Table 5.11: *Significant gene set enrichment results of genes with SNPs found by turquoise cluster analysis. P-values are analytically adjusted.*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | rs17025837 | 2 | 85136390 | -1.0634 | 0.1314 | 5.69145e-16 | SLC25A33 | KCMF1 |
| 2 | rs6664362 | 1 | 9536458 | rs10053502 | 5 | 40014929 | 1.5329 | 0.3299 | 3.37052e-06 | SLC25A33 | |
| 3 | rs6664362 | 1 | 9536458 | rs16931350 | 8 | 65701037 | 1.3496 | 0.2883 | 2.85701e-06 | SLC25A33 | CYP7B1 |
| 4 | rs6664362 | 1 | 9536458 | rs1887387 | 9 | 90069741 | -0.6083 | 0.1324 | 4.30855e-06 | SLC25A33 | |
| 5 | rs6664362 | 1 | 9536458 | rs11238776 | 10 | 42098785 | -1.8342 | 0.3687 | 6.54073e-07 | SLC25A33 | |
| 6 | rs6664362 | 1 | 9536458 | rs5030416 | 11 | 36489064 | -0.7433 | 0.1465 | 3.87038e-07 | SLC25A33 | |
| 7 | rs6664362 | 1 | 9536458 | rs1003761 | 14 | 23602251 | -0.8616 | 0.1729 | 6.20901e-07 | SLC25A33 | CARMIL3 |
| 8 | rs6664362 | 1 | 9536458 | rs486743 | 18 | 12495089 | 1.3808 | 0.2775 | 6.49246e-07 | SLC25A33 | SPIRE1 |
| 9 | rs6664362 | 1 | 9536458 | rs11910489 | 21 | 16968775 | -1.3202 | 0.2190 | 1.65162e-09 | SLC25A33 | |
| 10 | rs6664362 | 1 | 9536458 | rs17002187 | 21 | 26762664 | 0.9024 | 0.1982 | 5.30623e-06 | SLC25A33 | CYYR1 |
| 11 | rs6687006 | 1 | 38445830 | rs2924338 | 18 | 51326785 | -0.9532 | 0.2072 | 4.19675e-06 | | |
| 12 | rs6690381 | 1 | 83928327 | rs13415553 | 2 | 221126448 | -3.1959 | 0.6105 | 1.64724e-07 | | |
| 13 | rs16841336 | 1 | 157167751 | rs17025837 | 2 | 85136390 | -0.4429 | 0.0793 | 2.33708e-08 | PYHIN1 | KCMF1 |
| 14 | rs16841336 | 1 | 157167751 | rs5030416 | 11 | 36489064 | -0.4897 | 0.0903 | 5.89178e-08 | PYHIN1 | |
| 15 | rs11803913 | 1 | 203139465 | rs11925421 | 3 | 147370852 | -0.7989 | 0.1515 | 1.33751e-07 | | |
| 16 | rs10779486 | 1 | 206805932 | rs17025837 | 2 | 85136390 | -0.4889 | 0.0909 | 7.39718e-08 | | KCMF1 |
| 17 | rs10779486 | 1 | 206805932 | rs11238776 | 10 | 42098785 | -1.4005 | 0.2926 | 1.70295e-06 | | |
| 18 | rs2574672 | 1 | 240482668 | rs663366 | 8 | 98528011 | -0.8496 | 0.1838 | 3.79235e-06 | | |
| 19 | rs11898209 | 2 | 45872037 | rs12549933 | 8 | 80992623 | -0.5751 | 0.1270 | 5.9624e-06 | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | rs17025837 | 2 | 85136390 | rs11925421 | 3 | 147370852 | -0.7958 | 0.1126 | 1.55212e-12 | KCMF1 | |
| 21 | rs17025837 | 2 | 85136390 | rs9322193 | 6 | 149960836 | -0.3705 | 0.0762 | 1.17391e-06 | KCMF1 | |
| 22 | rs17025837 | 2 | 85136390 | rs1887387 | 9 | 90069741 | -0.3553 | 0.0744 | 1.77149e-06 | KCMF1 | |
| 23 | rs17025837 | 2 | 85136390 | rs11823088 | 11 | 19146328 | 0.6435 | 0.1393 | 3.86113e-06 | KCMF1 | |
| 24 | rs17025837 | 2 | 85136390 | rs11225401 | 11 | 102113169 | -0.5004 | 0.0811 | 6.96136e-10 | KCMF1 | |
| 25 | rs17025837 | 2 | 85136390 | rs4903419 | 14 | 76054408 | -0.4412 | 0.0802 | 3.80492e-08 | KCMF1 | |
| 26 | rs6809953 | 3 | 117493781 | rs9383431 | 6 | 18800063 | -0.5738 | 0.1289 | 8.56706e-06 | | |
| 27 | rs782437 | 3 | 128899892 | rs6788460 | 3 | 138324139 | 0.4671 | 0.0908 | 2.65296e-07 | | |
| 28 | rs782437 | 3 | 128899892 | rs9831130 | 3 | 138332662 | 0.4453 | 0.0902 | 7.96323e-07 | | |
| 29 | rs11925421 | 3 | 147370852 | rs16931350 | 8 | 65701037 | 1.3044 | 0.2620 | 6.40999e-07 | | CYP7B1 |
| 30 | rs11925421 | 3 | 147370852 | rs1887387 | 9 | 90069741 | -0.5915 | 0.1140 | 2.13373e-07 | | |
| 31 | rs11925421 | 3 | 147370852 | rs1003761 | 14 | 23602251 | -0.8264 | 0.1556 | 1.08115e-07 | | CARMIL3 |
| 32 | rs11925421 | 3 | 147370852 | rs11910489 | 21 | 16968775 | -1.0415 | 0.1880 | 3.04159e-08 | | |
| 33 | rs6443856 | 3 | 184362805 | rs387946 | 12 | 69162165 | -1.2677 | 0.2831 | 7.53719e-06 | | |
| 34 | rs6443856 | 3 | 184362805 | rs632547 | 12 | 69162296 | -1.2696 | 0.2830 | 7.27077e-06 | | |
| 35 | rs3902731 | 5 | 87314833 | rs11152086 | 18 | 54391010 | -1.1719 | 0.2488 | 2.46512e-06 | | |
| 36 | rs3844554 | 5 | 87317210 | rs11152086 | 18 | 54391010 | -1.1536 | 0.2492 | 3.65516e-06 | | |
| 37 | rs13172873 | 5 | 118665442 | rs3789950 | 10 | 98152781 | 1.1497 | 0.2585 | 8.68879e-06 | | |
| 38 | rs582677 | 6 | 5953061 | rs7251000 | 19 | 4665468 | -0.4884 | 0.1098 | 8.58508e-06 | | |
| 39 | rs7754397 | 6 | 91847214 | rs8127479 | 21 | 26012227 | -1.5186 | 0.3305 | 4.3388e-06 | | |
| 40 | rs11758609 | 6 | 105948645 | rs11626056 | 14 | 51303026 | -0.6441 | 0.1454 | 9.43036e-06 | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 41 | rs4458717 | 6 | 132950572 | rs17837474 | 7 | 142056577 | -0.8846 | 0.1792 | 7.96785e-07 | | |
| 42 | rs4458717 | 6 | 132950572 | rs16931350 | 8 | 65701037 | -1.2024 | 0.2109 | 1.19041e-08 | | CYP7B1 |
| 43 | rs6922844 | 6 | 150360447 | rs7022051 | 9 | 95042804 | 1.1557 | 0.2616 | 9.92936e-06 | | |
| 44 | rs11761505 | 7 | 47069736 | rs10876864 | 12 | 54687352 | 0.4749 | 0.0988 | 1.51414e-06 | | |
| 45 | rs17837474 | 7 | 142056577 | rs16931350 | 8 | 65701037 | -1.5678 | 0.2736 | 1.00398e-08 | | CYP7B1 |
| 46 | rs17837474 | 7 | 142056577 | rs12114111 | 8 | 125398550 | -0.8455 | 0.1400 | 1.52957e-09 | | TMEM65 |
| 47 | rs10755971 | 8 | 52577973 | rs7923523 | 10 | 95192322 | 0.5616 | 0.1170 | 1.58222e-06 | | |
| 48 | rs12541541 | 8 | 52579465 | rs7923523 | 10 | 95192322 | 0.5391 | 0.1198 | 6.81784e-06 | | |
| 49 | rs16931350 | 8 | 65701037 | rs12114111 | 8 | 125398550 | -0.6981 | 0.1577 | 9.61961e-06 | CYP7B1 | TMEM65 |
| 50 | rs4307325 | 8 | 124490578 | rs17375044 | 11 | 35173476 | -1.0275 | 0.2148 | 1.72692e-06 | | |
| 51 | rs1887387 | 9 | 90069741 | rs11225401 | 11 | 102113169 | -0.4008 | 0.0829 | 1.3395e-06 | | |
| 52 | rs11238776 | 10 | 42098785 | rs11225401 | 11 | 102113169 | -1.0909 | 0.2392 | 5.1104e-06 | | |
| 53 | rs548102 | 10 | 105854556 | rs8063675 | 16 | 68841080 | 2.4075 | 0.5193 | 3.55187e-06 | | |
| 54 | rs5030416 | 11 | 36489064 | rs4903419 | 14 | 76054408 | -0.4518 | 0.0923 | 9.8978e-07 | | |
| 55 | rs7946005 | 11 | 78464992 | rs11910489 | 21 | 16968775 | -0.6999 | 0.1580 | 9.42932e-06 | TENM4 | |
| 56 | rs11225401 | 11 | 102113169 | rs1003761 | 14 | 23602251 | -0.5900 | 0.1018 | 6.88186e-09 | | CARMIL3 |
| 57 | rs11225401 | 11 | 102113169 | rs486743 | 18 | 12495089 | 0.9056 | 0.1849 | 9.71786e-07 | | SPIRE1 |
| 58 | rs380835 | 12 | 69136889 | rs2149313 | 14 | 23313477 | -0.7588 | 0.1707 | 8.78126e-06 | | |
| 59 | rs7156661 | 14 | 70131302 | rs4636985 | 18 | 12068224 | 0.5602 | 0.1262 | 9.10903e-06 | | |
| 60 | rs4903419 | 14 | 76054408 | rs11910489 | 21 | 16968775 | -0.6136 | 0.1361 | 6.48894e-06 | | |
| 61 | rs2162549 | 15 | 78303041 | rs2221434 | 16 | 77508627 | -0.8880 | 0.1918 | 3.67171e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 62 | rs8063675 | 16 | 68841080 | rs2824699 | 21 | 18483226 | 2.8982 | 0.6053 | 1.68158e-06 |

Table 5.13: *Significant blue Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0497 | CORUM:7222 | CORUM | MMS19-XPD complex |
| 2 | 0.0497 | CORUM:7223 | CORUM | MMS19-FAM96B complex |

Table 5.14: *Significant gene set enrichment results of genes with SNPs found by blue cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | -0.1140 | 0.0098 | 6.045372e-31 | SLC25A33 |
| 2 | rs16841336 | 1 | 157167751 | -0.0471 | 0.0058 | 8.530469e-16 | PYHIN1 |
| 3 | rs10779486 | 1 | 206805932 | -0.0672 | 0.0075 | 3.520136e-19 | |
| 4 | rs17025837 | 2 | 85136390 | -0.0420 | 0.0064 | 6.236332e-11 | KCMF1 |
| 5 | rs2396261 | 2 | 226253774 | -0.0389 | 0.0074 | 1.760571e-07 | |
| 6 | rs164208 | 3 | 32519945 | 0.0273 | 0.0061 | 7.229022e-06 | |
| 7 | rs164219 | 3 | 32539488 | 0.0305 | 0.0063 | 1.192366e-06 | |
| 8 | rs3853720 | 3 | 32590673 | 0.0272 | 0.0061 | 7.851331e-06 | |
| 9 | rs11925421 | 3 | 147370852 | -0.0968 | 0.0088 | 1.32862e-27 | |
| 10 | rs6914506 | 6 | 37374934 | 0.0593 | 0.0051 | 5.092402e-31 | TBC1D22B |
| 11 | rs17837474 | 7 | 142056577 | 0.0433 | 0.0080 | 7.93706e-08 | |
| 12 | rs11555737 | 8 | 17198414 | 0.0403 | 0.0087 | 3.449944e-06 | VPS37A |
| 13 | rs7015243 | 8 | 17220931 | 0.0413 | 0.0089 | 3.610716e-06 | MTMR7 |
| 14 | rs16931350 | 8 | 65701037 | 0.0577 | 0.0088 | 5.578833e-11 | CYP7B1 |
| 15 | rs3019885 | 8 | 118094826 | -0.0265 | 0.0059 | 6.479774e-06 | SLC30A8 |
| 16 | rs12114111 | 8 | 125398550 | 0.0389 | 0.0056 | 6.399247e-12 | TMEM65 |
| 17 | rs10829351 | 10 | 129865332 | -0.0353 | 0.0071 | 7.688361e-07 | |
| 18 | rs7946005 | 11 | 78464992 | -0.0693 | 0.0073 | 3.799564e-21 | TENM4 |
| 19 | rs11225401 | 11 | 102113169 | -0.0442 | 0.0062 | 7.932704e-13 | |
| 20 | rs9548183 | 13 | 37559893 | -0.0293 | 0.0065 | 6.773426e-06 | |
| 21 | rs1924198 | 13 | 37628805 | -0.0303 | 0.0064 | 2.171432e-06 | |
| 22 | rs1003761 | 14 | 23602251 | -0.0445 | 0.0095 | 2.52818e-06 | CARMIL3 |
| 23 | rs4903419 | 14 | 76054408 | -0.0494 | 0.0062 | 1.44213e-15 | |
| 24 | rs13330491 | 16 | 30012701 | 0.0692 | 0.0078 | 7.493377e-19 | YPEL3 |
| 25 | rs235083 | 16 | 64760511 | 0.0539 | 0.0054 | 4.174809e-23 | |
| 26 | rs486743 | 18 | 12495089 | 0.0265 | 0.0053 | 6.556181e-07 | SPIRE1 |

Table 5.15: *Significant brown GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs16861446 | 1 | 18221371 | rs2798353 | 1 | 47413628 | 1.5482 | 0.3489 | 9.10384e-06 | | |
| 2 | rs1149064 | 1 | 30315613 | rs511773 | 2 | 13543565 | 0.5905 | 0.1332 | 9.33032e-06 | | |
| 3 | rs16841336 | 1 | 157167751 | rs10779486 | 1 | 206805932 | -1.1393 | 0.1833 | 5.08374e-10 | PYHIN1 | |
| 4 | rs16841336 | 1 | 157167751 | rs11925421 | 3 | 147370852 | -1.4548 | 0.2774 | 1.57319e-07 | PYHIN1 | |
| 5 | rs16841336 | 1 | 157167751 | rs7946005 | 11 | 78464992 | -0.8703 | 0.1698 | 2.98342e-07 | PYHIN1 | TENM4 |
| 6 | rs16841336 | 1 | 157167751 | rs11225401 | 11 | 102113169 | -0.5718 | 0.1211 | 2.35013e-06 | PYHIN1 | |
| 7 | rs16841336 | 1 | 157167751 | rs4903419 | 14 | 76054408 | -0.6564 | 0.1271 | 2.40881e-07 | PYHIN1 | |
| 8 | rs10779486 | 1 | 206805932 | rs2396261 | 2 | 226253774 | -1.0879 | 0.2392 | 5.41111e-06 | | |
| 9 | rs10779486 | 1 | 206805932 | rs6914506 | 6 | 37374934 | 0.6451 | 0.1428 | 6.2931e-06 | | TBC1D22B |
| 10 | rs10779486 | 1 | 206805932 | rs11225401 | 11 | 102113169 | -1.0480 | 0.1915 | 4.44974e-08 | | |
| 11 | rs10779486 | 1 | 206805932 | rs4903419 | 14 | 76054408 | -0.8963 | 0.1925 | 3.23644e-06 | | |
| 12 | rs10779486 | 1 | 206805932 | rs235083 | 16 | 64760511 | 0.8737 | 0.1594 | 4.19024e-08 | | |
| 13 | rs10779486 | 1 | 206805932 | rs486743 | 18 | 12495089 | 0.6892 | 0.1536 | 7.2125e-06 | | SPIRE1 |
| 14 | rs511773 | 2 | 13543565 | rs17097224 | 5 | 140708718 | -2.8161 | 0.6119 | 4.18622e-06 | | |
| 15 | rs13402137 | 2 | 91334683 | rs7832778 | 8 | 27181804 | -2.0163 | 0.4550 | 9.36948e-06 | | |
| 16 | rs10206874 | 2 | 120876615 | rs9988693 | 10 | 4369270 | 0.4632 | 0.0962 | 1.45726e-06 | | |
| 17 | rs2396261 | 2 | 226253774 | rs11925421 | 3 | 147370852 | -2.2440 | 0.4296 | 1.76062e-07 | | |
| 18 | rs2396261 | 2 | 226253774 | rs2394173 | 6 | 29881137 | -0.6360 | 0.1373 | 3.59189e-06 | | |
| 19 | rs2396261 | 2 | 226253774 | rs11225401 | 11 | 102113169 | -0.9079 | 0.1478 | 8.05579e-10 | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | rs2396261 | 2 | 226253774 | rs235083 | 16 | 64760511 | 0.4829 | 0.0991 | 1.1096e-06 | | |
| 21 | rs2396261 | 2 | 226253774 | rs486743 | 18 | 12495089 | 0.4315 | 0.0942 | 4.69904e-06 | | SPIRE1 |
| 22 | rs4381740 | 2 | 235698824 | rs16879540 | 4 | 27200415 | -17.2377 | 3.7457 | 4.18497e-06 | | |
| 23 | rs10180205 | 2 | 240165374 | rs4618447 | 5 | 177799781 | -1.3536 | 0.2975 | 5.37353e-06 | | |
| 24 | rs9861677 | 3 | 29472993 | rs4374629 | 4 | 99630806 | 0.4309 | 0.0932 | 3.79037e-06 | | |
| 25 | rs13325694 | 3 | 150740094 | rs6924196 | 6 | 131631199 | -0.5837 | 0.1108 | 1.38346e-07 | | |
| 26 | rs11735537 | 4 | 144370151 | rs2188148 | 7 | 91046536 | -1.1761 | 0.2605 | 6.32338e-06 | | |
| 27 | rs2355617 | 4 | 146429885 | rs3848701 | 20 | 57536928 | -0.4536 | 0.0957 | 2.16634e-06 | | |
| 28 | rs30708 | 5 | 128987981 | rs4922069 | 8 | 19518995 | -0.8433 | 0.1798 | 2.71299e-06 | | |
| 29 | rs17338894 | 5 | 164848265 | rs1396208 | 12 | 24449587 | 0.4803 | 0.1063 | 6.17479e-06 | | |
| 30 | rs2267639 | 6 | 30688616 | rs7843326 | 8 | 120942523 | -0.9458 | 0.2096 | 6.38434e-06 | | |
| 31 | rs6914506 | 6 | 37374934 | rs235083 | 16 | 64760511 | -0.4966 | 0.0769 | 1.06689e-10 | TBC1D22B | |
| 32 | rs6914506 | 6 | 37374934 | rs486743 | 18 | 12495089 | -0.3528 | 0.0733 | 1.48384e-06 | TBC1D22B | SPIRE1 |
| 33 | rs10239402 | 7 | 36754882 | rs17691755 | 19 | 34485216 | -0.6212 | 0.1152 | 6.94713e-08 | | |
| 34 | rs7797417 | 7 | 46188335 | rs10859030 | 12 | 89614699 | -1.7266 | 0.3894 | 9.23292e-06 | | |
| 35 | rs4947572 | 7 | 51149992 | rs8071990 | 17 | 4526387 | -0.4730 | 0.1046 | 6.18877e-06 | | |
| 36 | rs10247961 | 7 | 75147682 | rs2725627 | 15 | 51979784 | -0.6196 | 0.1352 | 4.56701e-06 | | |
| 37 | rs9644372 | 8 | 4645818 | rs2076421 | 16 | 1071394 | 1.4867 | 0.3339 | 8.47268e-06 | CSMD1 | |
| 38 | rs1947717 | 9 | 99145556 | rs28364553 | 15 | 40492634 | 1.5533 | 0.3324 | 2.97645e-06 | | |
| 39 | rs2242187 | 9 | 99177215 | rs28364553 | 15 | 40492634 | 1.5653 | 0.3290 | 1.95973e-06 | | |
| 40 | rs7023854 | 9 | 103331633 | rs3824951 | 11 | 5611809 | -0.4790 | 0.0974 | 8.70492e-07 | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 41 | rs1983812 | 9 | 103372864 | rs3824951 | 11 | 5611809 | 0.4332 | 0.0961 | 6.54686e-06 | | |
| 42 | rs12292693 | 11 | 64693295 | rs17073814 | 13 | 80638674 | 1.0129 | 0.2243 | 6.2858e-06 | | |
| 43 | rs7946005 | 11 | 78464992 | rs11225401 | 11 | 102113169 | -0.9678 | 0.1802 | 7.888e-08 | TENM4 | |
| 44 | rs7946005 | 11 | 78464992 | rs235083 | 16 | 64760511 | 0.7355 | 0.1439 | 3.18345e-07 | TENM4 | |
| 45 | rs11225401 | 11 | 102113169 | rs4903419 | 14 | 76054408 | -0.6355 | 0.1329 | 1.74557e-06 | | |
| 46 | rs235083 | 16 | 64760511 | rs486743 | 18 | 12495089 | -0.3868 | 0.0722 | 8.28396e-08 | | SPIRE1 |

Table 5.16: *Significant brown Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0017 | CORUM:6780 | CORUM | RAD6A-KCMF1-UBR4 complex |
| 2 | 0.0248 | GO:0008396 | GO:MF | oxysterol 7-alpha-hydroxylase activity |
| 3 | 0.0248 | GO:0015218 | GO:MF | pyrimidine nucleotide transmembrane transporter activity |
| 4 | 0.0248 | GO:0033783 | GO:MF | 25-hydroxycholesterol 7alpha-hydroxylase activity |
| 5 | 0.0248 | GO:0047092 | GO:MF | 27-hydroxycholesterol 7-alpha-monooxygenase activity |
| 6 | 0.0460 | KEGG:00120 | KEGG | Primary bile acid biosynthesis |
| 7 | 0.0227 | REAC:R-HSA-5579013 | REAC | Defective CYP7B1 causes Spastic paraplegia 5A, autosomal recessive (SPG5A) and (CBAS3) |
| 8 | 0.0427 | REAC:R-HSA-1660517 | REAC | Synthesis of PIPs at the late endosome membrane |
| 9 | 0.0427 | REAC:R-HSA-175474 | REAC | Assembly Of The HIV Virion |
| 10 | 0.0427 | REAC:R-HSA-435354 | REAC | Zinc transporters |
| 11 | 0.0427 | REAC:R-HSA-435368 | REAC | Zinc efflux and compartmentalization by the SLC30 family |
| 12 | 0.0427 | REAC:R-HSA-174490 | REAC | Membrane binding and targetting of GAG proteins |
| 13 | 0.0427 | REAC:R-HSA-193807 | REAC | Synthesis of bile acids and bile salts via 27-hydroxycholesterol |
| 14 | 0.0427 | REAC:R-HSA-174495 | REAC | Synthesis And Processing Of GAG, GAGPOL Polyproteins |
| 15 | 0.0427 | REAC:R-HSA-1855183 | REAC | Synthesis of IP2, IP, and Ins in the cytosol |
| 16 | 0.0451 | REAC:R-HSA-193368 | REAC | Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol |
| 17 | 0.0451 | REAC:R-HSA-192105 | REAC | Synthesis of bile acids and bile salts |
| 18 | 0.0451 | REAC:R-HSA-211976 | REAC | Endogenous sterols |
| 19 | 0.0451 | REAC:R-HSA-425410 | REAC | Metal ion SLC transporters |
| 20 | 0.0451 | REAC:R-HSA-9615710 | REAC | Late endosomal microautophagy |
| 21 | 0.0451 | REAC:R-HSA-264876 | REAC | Insulin processing |
| 22 | 0.0451 | REAC:R-HSA-917729 | REAC | Endosomal Sorting Complex Required For Transport (ESCRT) |
| 23 | 0.0451 | REAC:R-HSA-162588 | REAC | Budding and maturation of HIV virion |
| 24 | 0.0463 | REAC:R-HSA-5579029 | REAC | Metabolic disorders of biological oxidation enzymes |
| 25 | 0.0497 | REAC:R-HSA-194068 | REAC | Bile acid and bile salt metabolism |
| 26 | 0.0420 | WP:WP465 | WP | Tryptophan metabolism |
| 27 | 0.0420 | WP:WP3529 | WP | Zinc homeostasis |
| 28 | 0.0420 | WP:WP4545 | WP | Oxysterols derived from cholesterol |
| 29 | 0.0457 | WP:WP43 | WP | Oxidation by Cytochrome P450 |

Table 5.17: *Significant gene set enrichmet results of genes with SNPs found by brown cluster analysis. P-values are FDR adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | 0.0551 | 0.0054 | 1.846251e-24 | SLC25A33 |
| 2 | rs904218 | 1 | 54733577 | 0.0251 | 0.0055 | 5.74492e-06 | |
| 3 | rs4465227 | 1 | 83805074 | 0.0241 | 0.0054 | 8.421985e-06 | |
| 4 | rs16841336 | 1 | 157167751 | 0.0185 | 0.0032 | 7.265087e-09 | PYHIN1 |
| 5 | rs10779486 | 1 | 206805932 | 0.0400 | 0.0041 | 6.131656e-22 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 6 | rs17025837 | 2 | 85136390 | 0.0224 | 0.0035 | 1.369008e-10 | KCMF1 |
| 7 | rs10173578 | 2 | 98638084 | 0.0711 | 0.0135 | 1.39271e-07 | MGAT4A |
| 8 | rs7586108 | 2 | 98656954 | 0.0545 | 0.0121 | 7.26247e-06 | MGAT4A |
| 9 | rs4467263 | 2 | 216412203 | 0.0284 | 0.0064 | 8.941384e-06 | |
| 10 | rs2396261 | 2 | 226253774 | 0.0244 | 0.0041 | 1.956478e-09 | |
| 11 | rs17050371 | 3 | 9482379 | 0.0539 | 0.0117 | 4.446303e-06 | SETD5 |
| 12 | rs7613687 | 3 | 60124976 | 0.0366 | 0.0073 | 5.830227e-07 | FHIT |
| 13 | rs11925421 | 3 | 147370852 | 0.0451 | 0.0049 | 3.889461e-20 | |
| 14 | rs6855169 | 4 | 87409854 | 0.0174 | 0.0038 | 5.17517e-06 | MAPK10 |
| 15 | rs359502 | 4 | 163061103 | 0.0566 | 0.0116 | 1.21883e-06 | FSTL5 |
| 16 | rs4594876 | 5 | 96200799 | 0.0587 | 0.0130 | 6.150582e-06 | |
| 17 | rs6914506 | 6 | 37374934 | -0.0256 | 0.0028 | 6.577782e-20 | TBC1D22B |
| 18 | rs636444 | 6 | 75886187 | 0.0681 | 0.0117 | 6.537207e-09 | COL12A1 |
| 19 | rs4719303 | 7 | 12212182 | 0.0710 | 0.0155 | 5.019359e-06 | |
| 20 | rs2333890 | 7 | 131063176 | 0.0467 | 0.0099 | 2.742906e-06 | |
| 21 | rs1881690 | 7 | 154556020 | 0.0248 | 0.0056 | 9.72057e-06 | |
| 22 | rs9644372 | 8 | 4645818 | 0.0448 | 0.0081 | 3.103544e-08 | CSMD1 |
| 23 | rs17071510 | 8 | 4658756 | 0.0410 | 0.0087 | 2.570519e-06 | CSMD1 |
| 24 | rs17071569 | 8 | 4663824 | 0.0360 | 0.0079 | 5.303253e-06 | CSMD1 |
| 25 | rs7859572 | 9 | 32127688 | 0.0551 | 0.0121 | 5.305183e-06 | |
| 26 | rs7477268 | 10 | 5360986 | -0.0150 | 0.0034 | 8.443174e-06 | |
| 27 | rs10491052 | 10 | 108384604 | 0.0211 | 0.0044 | 1.705825e-06 | SORCS1 |
| 28 | rs3781436 | 10 | 126685259 | 0.0256 | 0.0050 | 3.207543e-07 | CTBP2 |
| 29 | rs341093 | 11 | 71914693 | 0.0407 | 0.0090 | 5.639215e-06 | |
| 30 | rs7946005 | 11 | 78464992 | 0.0358 | 0.0040 | 1.052576e-18 | TENM4 |
| 31 | rs11225401 | 11 | 102113169 | 0.0214 | 0.0034 | 2.011892e-10 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 32 | rs7970246 | 12 | 98033341 | 0.0503 | 0.0107 | 2.895588e-06 | ANKS1B |
| 33 | rs17234467 | 13 | 92867191 | 0.0758 | 0.0155 | 1.000154e-06 | GPC6 |
| 34 | rs1003761 | 14 | 23602251 | 0.0300 | 0.0051 | 5.651741e-09 | CARMIL3 |
| 35 | rs4903419 | 14 | 76054408 | 0.0207 | 0.0034 | 1.031482e-09 | |
| 36 | rs11070885 | 15 | 50289148 | 0.0209 | 0.0047 | 8.394346e-06 | MYO5C |
| 37 | rs2573590 | 15 | 98340757 | 0.0691 | 0.0155 | 8.49181e-06 | ADAMTS17 |
| 38 | rs235083 | 16 | 64760511 | -0.0252 | 0.0030 | 2.736873e-17 | |
| 39 | rs486743 | 18 | 12495089 | -0.0160 | 0.0029 | 3.271965e-08 | SPIRE1 |
| 40 | rs16975208 | 18 | 37538035 | 0.0421 | 0.0095 | 9.649039e-06 | |
| 41 | rs11910489 | 21 | 16968775 | 0.0340 | 0.0066 | 2.647598e-07 | |

Table 5.18: *Significant yellow GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs12034719 | 1 | 8882740 | rs1873061 | 17 | 3340194 | 0.8962 | 0.1995 | 7.07328e-06 | | |
| 2 | rs2105217 | 1 | 31700974 | rs550353 | 5 | 31421753 | 1.1716 | 0.2641 | 9.15523e-06 | | |
| 3 | rs10493102 | 1 | 41781108 | rs2280183 | 16 | 26987458 | 0.8757 | 0.1928 | 5.54145e-06 | | |
| 4 | rs1109255 | 1 | 41864295 | rs12450609 | 17 | 71249027 | -1.4880 | 0.3338 | 8.28709e-06 | | |
| 5 | rs12757818 | 1 | 104761380 | rs4382869 | 11 | 14393309 | -1.4173 | 0.3018 | 2.65012e-06 | | |
| 6 | rs486753 | 1 | 206395372 | rs4861505 | 4 | 183508343 | -1.6241 | 0.3430 | 2.19098e-06 | | |
| 7 | rs1332777 | 1 | 235820963 | rs3799879 | 6 | 46672927 | 1.3297 | 0.2770 | 1.58329e-06 | | |
| 8 | rs1332777 | 1 | 235820963 | rs16874838 | 6 | 46673023 | 1.2941 | 0.2796 | 3.68625e-06 | | |
| 9 | rs1842088 | 1 | 235824870 | rs3799879 | 6 | 46672927 | 1.2723 | 0.2750 | 3.73154e-06 | | |
| 10 | rs1842088 | 1 | 235824870 | rs16874838 | 6 | 46673023 | 1.2358 | 0.2776 | 8.50978e-06 | | |
| 11 | rs2192982 | 2 | 51369176 | rs12467106 | 2 | 79301209 | 1.6400 | 0.3699 | 9.25034e-06 | | |
| 12 | rs1861108 | 2 | 59274551 | rs8106157 | 19 | 49784782 | 2.1188 | 0.4754 | 8.32022e-06 | | |
| 13 | rs2042091 | 2 | 64622985 | rs1519790 | 2 | 151319638 | -1.2995 | 0.2625 | 7.39318e-07 | | |
| 14 | rs10201616 | 2 | 75529368 | rs2693698 | 14 | 98788972 | -0.8607 | 0.1885 | 4.948e-06 | | |
| 15 | rs580041 | 2 | 166658756 | rs816623 | 10 | 599945 | -2.7486 | 0.6142 | 7.64718e-06 | | |
| 16 | rs7587026 | 2 | 166686996 | rs816623 | 10 | 599945 | -2.7405 | 0.6085 | 6.66797e-06 | | |
| 17 | rs10515949 | 2 | 207446158 | rs35294541 | 14 | 50974619 | -1.6834 | 0.3809 | 9.90623e-06 | | |
| 18 | rs10933164 | 2 | 227568915 | rs1812576 | 8 | 62335588 | -1.3489 | 0.2840 | 2.02992e-06 | | |
| 19 | rs3773341 | 3 | 12591533 | rs6886410 | 5 | 10161500 | 0.9705 | 0.1778 | 4.776e-08 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs1873628 | 3 | 70193317 | rs12111597 | 7 | 34836526 | -2.0540 | 0.4596 | 7.85256e-06 |
| 21 | rs9874533 | 3 | 78363609 | rs7777560 | 7 | 29428441 | 1.6559 | 0.3490 | 2.08683e-06 |
| 22 | rs964243 | 3 | 83649900 | rs2100425 | 15 | 46021021 | 1.2354 | 0.2786 | 9.2362e-06 |
| 23 | rs9290057 | 3 | 101636807 | rs269876 | 5 | 22626814 | -2.3294 | 0.4954 | 2.57288e-06 |
| 24 | rs10516998 | 4 | 40820851 | rs11072326 | 15 | 69851773 | -1.2871 | 0.2620 | 9.00425e-07 |
| 25 | rs2291182 | 4 | 71236237 | rs10486180 | 7 | 7528105 | 1.7258 | 0.3709 | 3.27583e-06 |
| 26 | rs17148759 | 4 | 71256426 | rs10486180 | 7 | 7528105 | 1.6189 | 0.3620 | 7.7579e-06 |
| 27 | rs7660807 | 4 | 71423140 | rs10486180 | 7 | 7528105 | 1.8648 | 0.3704 | 4.77229e-07 |
| 28 | rs7678480 | 4 | 87306344 | rs10215217 | 7 | 25127658 | 1.0351 | 0.2246 | 4.06314e-06 |
| 29 | rs12498329 | 4 | 119530972 | rs1574196 | 8 | 62656389 | 0.9820 | 0.2135 | 4.25482e-06 |
| 30 | rs724043 | 4 | 150120535 | rs11665868 | 19 | 57045646 | 2.9996 | 0.6296 | 1.89802e-06 |
| 31 | rs269876 | 5 | 22626814 | rs6099116 | 20 | 54374675 | -2.1664 | 0.4600 | 2.48396e-06 |
| 32 | rs4435901 | 5 | 117399047 | rs1792005 | 11 | 122947776 | 1.2235 | 0.2687 | 5.27761e-06 |
| 33 | rs7356549 | 5 | 129751415 | rs8073016 | 17 | 66027687 | -1.1027 | 0.2297 | 1.58098e-06 |
| 34 | rs2636112 | 5 | 143140380 | rs9471773 | 6 | 42199623 | -1.2516 | 0.2805 | 8.14772e-06 |
| 35 | rs2636112 | 5 | 143140380 | rs9493282 | 6 | 99611098 | 0.9448 | 0.1988 | 2.00947e-06 |
| 36 | rs3792815 | 5 | 143172404 | rs9493282 | 6 | 99611098 | 0.9417 | 0.2070 | 5.36743e-06 |
| 37 | rs11134531 | 5 | 168144746 | rs10788336 | 10 | 86005175 | 0.9124 | 0.2039 | 7.60982e-06 |
| 38 | rs11134531 | 5 | 168144746 | rs4933317 | 10 | 86006146 | 0.9252 | 0.2041 | 5.78925e-06 |
| 39 | rs28011 | 5 | 179339835 | rs1189133 | 14 | 55951645 | 0.9330 | 0.2089 | 7.94133e-06 |
| 40 | rs9381137 | 6 | 42183439 | rs180623 | 10 | 117774312 | -0.9005 | 0.2030 | 9.13006e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs9688879 | 6 | 117806498 | rs7944394 | 11 | 313649 | 1.1729 | 0.2271 | 2.39325e-07 |
| 42 | rs6569053 | 6 | 119539583 | rs12580674 | 12 | 63204613 | -2.2643 | 0.5124 | 9.905e-06 |
| 43 | rs10215217 | 7 | 25127658 | rs11219115 | 11 | 122704777 | 1.2651 | 0.2735 | 3.7211e-06 |
| 44 | rs2285738 | 7 | 25148437 | rs11219115 | 11 | 122704777 | 1.2167 | 0.2665 | 4.98167e-06 |
| 45 | rs7800248 | 7 | 29434639 | rs860798 | 21 | 38054651 | 0.9782 | 0.2205 | 9.15855e-06 |
| 46 | rs255112 | 7 | 30703658 | rs10149902 | 14 | 55952495 | 1.3518 | 0.3047 | 9.1307e-06 |
| 47 | rs255112 | 7 | 30703658 | rs2269193 | 21 | 37028413 | 0.8909 | 0.1966 | 5.89267e-06 |
| 48 | rs10267710 | 7 | 95578271 | rs9554019 | 13 | 26849711 | 1.4706 | 0.3228 | 5.212e-06 |
| 49 | rs4731523 | 7 | 128317512 | rs2298389 | 22 | 23283349 | -2.8521 | 0.5945 | 1.60363e-06 |
| 50 | rs2272750 | 8 | 28265756 | rs11111796 | 12 | 102730058 | 1.1918 | 0.2606 | 4.81667e-06 |
| 51 | rs1606133 | 8 | 50215053 | rs2849890 | 21 | 17807769 | 1.3259 | 0.2837 | 2.95774e-06 |
| 52 | rs1434927 | 8 | 69702396 | rs976825 | 9 | 78118411 | -0.9969 | 0.2227 | 7.57185e-06 |
| 53 | rs16924243 | 9 | 6247054 | rs2835119 | 21 | 36039929 | -17.1460 | 3.8803 | 9.92747e-06 |
| 54 | rs943452 | 10 | 6655460 | rs17005224 | 12 | 78030993 | -3.0671 | 0.6333 | 1.28012e-06 |
| 55 | rs2765161 | 13 | 23468779 | rs2693698 | 14 | 98788972 | -0.8782 | 0.1842 | 1.8668e-06 |
| 56 | rs9508434 | 13 | 28860368 | rs9931155 | 16 | 9892024 | 0.8193 | 0.1816 | 6.43107e-06 |
| 57 | rs10483305 | 14 | 25626745 | rs731119 | 16 | 54955872 | -0.9807 | 0.2021 | 1.2197e-06 |
| 58 | rs10483305 | 14 | 25626745 | rs9937484 | 16 | 55022018 | -1.0663 | 0.2064 | 2.39145e-07 |
| 59 | rs8061733 | 16 | 26983403 | rs7196183 | 16 | 77228455 | 0.8645 | 0.1953 | 9.57569e-06 |

Table 5.19: *Significant yellow Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0451 | GO:0097120 | GO:BP | receptor localization to synapse |

Table 5.20: *Significant gene set enrichment results of genes with SNPs found by yellow cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | -0.0260 | 0.0051 | 3.124247e-07 | SLC25A33 |
| 2 | rs16841336 | 1 | 157167751 | -0.0133 | 0.0030 | 8.459175e-06 | PYHIN1 |
| 3 | rs3789357 | 1 | 179286940 | 0.0304 | 0.0062 | 9.199761e-07 | MR1 |
| 4 | rs2000220 | 1 | 216654259 | 0.0153 | 0.0031 | 6.426457e-07 | TGFB2 |
| 5 | rs1879113 | 2 | 104450001 | 0.0623 | 0.0141 | 9.985805e-06 | |
| 6 | rs7558405 | 2 | 162831332 | 0.0307 | 0.0063 | 1.168289e-06 | |
| 7 | rs6734769 | 2 | 162847031 | 0.0394 | 0.0070 | 2.003524e-08 | IFIH1 |
| 8 | rs12476567 | 2 | 162856657 | 0.0377 | 0.0069 | 4.064585e-08 | IFIH1 |
| 9 | rs10439256 | 2 | 162860597 | 0.0379 | 0.0069 | 3.620302e-08 | IFIH1 |
| 10 | rs16846646 | 2 | 162923669 | 0.0318 | 0.0071 | 7.33516e-06 | GCA |
| 11 | rs6978 | 3 | 56736755 | 0.0191 | 0.0041 | 3.306814e-06 | ARHGEF3 |
| 12 | rs3821414 | 3 | 56737285 | 0.0193 | 0.0041 | 2.186296e-06 | ARHGEF3 |
| 13 | rs1009119 | 3 | 56738268 | 0.0190 | 0.0041 | 3.836417e-06 | ARHGEF3 |
| 14 | rs6765444 | 3 | 56738659 | 0.0188 | 0.0041 | 4.079054e-06 | ARHGEF3 |
| 15 | rs6795648 | 3 | 124571756 | 0.0448 | 0.0090 | 6.891151e-07 | ADCY5 |
| 16 | rs11925421 | 3 | 147370852 | -0.0246 | 0.0046 | 8.467602e-08 | |
| 17 | rs10016497 | 4 | 141433672 | 0.0181 | 0.0039 | 2.612259e-06 | SCOC |
| 18 | rs10155508 | 5 | 136222522 | 0.0363 | 0.0082 | 9.451178e-06 | |
| 19 | rs11949188 | 5 | 167680766 | 0.0234 | 0.0051 | 4.143152e-06 | WWC1 |
| 20 | rs10074081 | 5 | 178552380 | 0.0353 | 0.0061 | 8.609612e-09 | ADAMTS2 |
| 21 | rs4700788 | 5 | 178553161 | 0.0210 | 0.0042 | 6.157401e-07 | ADAMTS2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 22 | rs3822601 | 5 | 178554182 | 0.0212 | 0.0042 | 4.95439e-07 | ADAMTS2 |
| 23 | rs6914506 | 6 | 37374934 | 0.0143 | 0.0026 | 4.687496e-08 | TBC1D22B |
| 24 | rs16896333 | 6 | 43275633 | 0.0549 | 0.0121 | 5.342441e-06 | CUL9 |
| 25 | rs17073336 | 6 | 112524916 | 0.0661 | 0.0146 | 5.760098e-06 | FAM229B |
| 26 | rs876058 | 8 | 29081350 | 0.0158 | 0.0034 | 2.413463e-06 | KIF13B |
| 27 | rs12543905 | 8 | 134637913 | 0.0215 | 0.0043 | 7.481305e-07 | ST3GAL1 |
| 28 | rs6578122 | 8 | 136155923 | 0.0270 | 0.0054 | 7.858254e-07 | |
| 29 | rs3781509 | 10 | 117847336 | 0.0579 | 0.0107 | 5.862896e-08 | GFRA1 |
| 30 | rs1872706 | 11 | 32533274 | 0.0151 | 0.0033 | 5.776211e-06 | |
| 31 | rs11227655 | 11 | 66555934 | 0.0588 | 0.0132 | 8.110952e-06 | SYT12 |
| 32 | rs12427339 | 12 | 103048947 | 0.0248 | 0.0052 | 2.281135e-06 | NFYB |
| 33 | rs12898861 | 15 | 75770857 | 0.0144 | 0.0031 | 4.42175e-06 | LINGO1 |
| 34 | rs8053650 | 16 | 11450946 | 0.0287 | 0.0061 | 2.225265e-06 | AC099489.1 |
| 35 | rs12601618 | 17 | 6330731 | 0.0289 | 0.0059 | 9.562703e-07 | PITPNM3 |
| 36 | rs9962971 | 18 | 64621573 | 0.0563 | 0.0109 | 2.69358e-07 | CCDC102B |
| 37 | rs2403768 | 21 | 14593706 | 0.0614 | 0.0133 | 4.077156e-06 | ABCC13 |
| 38 | rs9984213 | 21 | 15442521 | 0.0234 | 0.0053 | 8.524214e-06 | |
| 39 | rs9974747 | 21 | 15446136 | 0.0251 | 0.0055 | 5.1839e-06 | |
| 40 | rs9306428 | 22 | 25594282 | 0.0373 | 0.0077 | 1.217336e-06 | |

Table 5.21: *Significant green GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs3010109 | 1 | 27151140 | rs11226850 | 11 | 105164739 | 2.8624 | 0.6105 | 2.74803e-06 | | |
| 2 | rs12025436 | 1 | 80078572 | rs1452333 | 9 | 28110384 | -1.3014 | 0.2933 | 9.0877e-06 | | |
| 3 | rs2503273 | 1 | 88046378 | rs11075030 | 16 | 11883915 | 1.0811 | 0.2421 | 7.95436e-06 | | |
| 4 | rs2503273 | 1 | 88046378 | rs11075032 | 16 | 11911951 | 1.0787 | 0.2432 | 9.2194e-06 | | |
| 5 | rs10881450 | 1 | 107190366 | rs10492555 | 13 | 35607109 | -17.1996 | 3.8527 | 8.03373e-06 | | |
| 6 | rs11583252 | 1 | 117275111 | rs2545394 | 5 | 66441899 | -17.4176 | 3.9373 | 9.69977e-06 | | |
| 7 | rs1891737 | 1 | 163195377 | rs6472078 | 8 | 49907900 | 0.9183 | 0.2060 | 8.24406e-06 | | |
| 8 | rs1409813 | 1 | 182324982 | rs7143462 | 14 | 76092986 | 1.3118 | 0.2926 | 7.34496e-06 | | |
| 9 | rs12755165 | 1 | 201286370 | rs11916000 | 3 | 144846715 | -2.7839 | 0.6256 | 8.5953e-06 | | |
| 10 | rs6691953 | 1 | 214721745 | rs1866670 | 2 | 42163826 | 1.7707 | 0.3844 | 4.09264e-06 | | |
| 11 | rs7574069 | 2 | 116461957 | rs2702959 | 17 | 31301319 | -1.8928 | 0.4146 | 4.98898e-06 | | |
| 12 | rs6735366 | 2 | 117090007 | rs7724036 | 5 | 150265001 | 1.2624 | 0.2681 | 2.49552e-06 | | |
| 13 | rs6719835 | 2 | 144504859 | rs10873325 | 14 | 79531042 | 1.2082 | 0.2596 | 3.25006e-06 | | |
| 14 | rs6759721 | 2 | 147503753 | rs12193414 | 6 | 2238019 | -1.5902 | 0.3475 | 4.74317e-06 | | |
| 15 | rs2706035 | 2 | 147529174 | rs6869001 | 5 | 60943136 | 2.3592 | 0.5311 | 8.90825e-06 | | |
| 16 | rs16863814 | 2 | 177000920 | rs4301259 | 5 | 117288025 | -1.5359 | 0.3448 | 8.43626e-06 | | |
| 17 | rs6722501 | 2 | 202501511 | rs9936526 | 16 | 59162462 | 1.4549 | 0.3192 | 5.16022e-06 | | |
| 18 | rs11675143 | 2 | 230583555 | rs4554985 | 12 | 24684085 | -0.8614 | 0.1924 | 7.56811e-06 | | |
| 19 | rs748465 | 3 | 127690482 | rs9958551 | 18 | 43490826 | -16.8299 | 3.6601 | 4.26018e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs6784120 | 3 | 134844999 | rs1564896 | 4 | 96667710 | 2.4703 | 0.5196 | 1.99179e-06 |
| 21 | rs13084806 | 3 | 193211750 | rs6122673 | 20 | 46443923 | 2.7787 | 0.5452 | 3.46503e-07 |
| 22 | rs987218 | 4 | 141397778 | rs16902124 | 8 | 128426400 | -16.8295 | 3.7496 | 7.17697e-06 |
| 23 | rs1398186 | 4 | 141400596 | rs980712 | 10 | 15680110 | -1.6746 | 0.3650 | 4.46394e-06 |
| 24 | rs1398186 | 4 | 141400596 | rs896431 | 10 | 15696680 | -1.6853 | 0.3530 | 1.81028e-06 |
| 25 | rs1398186 | 4 | 141400596 | rs2039910 | 10 | 15752834 | -1.6896 | 0.3565 | 2.13979e-06 |
| 26 | rs1436969 | 5 | 30725657 | rs2051072 | 6 | 37669843 | 0.8802 | 0.1797 | 9.64798e-07 |
| 27 | rs4389670 | 5 | 72646286 | rs2388807 | 16 | 53294960 | -2.3385 | 0.4972 | 2.55497e-06 |
| 28 | rs17163950 | 5 | 109093607 | rs7939444 | 11 | 7738998 | -1.6382 | 0.3438 | 1.88407e-06 |
| 29 | rs6872842 | 5 | 146443032 | rs4799581 | 18 | 27415509 | -0.8271 | 0.1796 | 4.11984e-06 |
| 30 | rs7723948 | 5 | 146445272 | rs2704058 | 18 | 27411566 | -0.8004 | 0.1791 | 7.8098e-06 |
| 31 | rs7723948 | 5 | 146445272 | rs4799581 | 18 | 27415509 | -0.8783 | 0.1796 | 1.00104e-06 |
| 32 | rs6939085 | 6 | 40739033 | rs931546 | 6 | 125840174 | -16.8805 | 3.6300 | 3.31456e-06 |
| 33 | rs6939230 | 6 | 40739043 | rs931546 | 6 | 125840174 | -16.8818 | 3.6098 | 2.91624e-06 |
| 34 | rs9381214 | 6 | 42801937 | rs10781329 | 9 | 70808065 | -1.2184 | 0.2741 | 8.81273e-06 |
| 35 | rs11980453 | 7 | 31404897 | rs878202 | 18 | 68864017 | 1.8288 | 0.4096 | 8.01959e-06 |
| 36 | rs4282513 | 7 | 135986301 | rs7317997 | 13 | 113781019 | 1.5554 | 0.3314 | 2.69543e-06 |
| 37 | rs13253327 | 8 | 9347868 | rs854380 | 14 | 24340175 | -0.8046 | 0.1795 | 7.39399e-06 |
| 38 | rs17150066 | 8 | 9347921 | rs854380 | 14 | 24340175 | -0.9277 | 0.1832 | 4.14057e-07 |
| 39 | rs2004678 | 8 | 10172645 | rs36400 | 14 | 71429746 | 2.9652 | 0.6567 | 6.31394e-06 |
| 40 | rs1480803 | 8 | 136202180 | rs4511648 | 19 | 35798852 | 1.2950 | 0.2831 | 4.79367e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs10119193 | 9 | 109336507 | rs17132028 | 10 | 4119009 | 3.1801 | 0.6917 | 4.27896e-06 |
| 42 | rs7047863 | 9 | 110590683 | rs12228854 | 12 | 46683187 | 2.5555 | 0.5702 | 7.38978e-06 |
| 43 | rs7047863 | 9 | 110590683 | rs11168370 | 12 | 46721547 | 2.3114 | 0.5207 | 9.0462e-06 |
| 44 | rs7047863 | 9 | 110590683 | rs9971764 | 12 | 130814085 | 3.6700 | 0.7595 | 1.35017e-06 |
| 45 | rs1349323 | 9 | 117781539 | rs4766152 | 12 | 3620389 | -0.9393 | 0.2007 | 2.8774e-06 |
| 46 | rs11192698 | 10 | 107811288 | rs2824700 | 21 | 18484365 | 1.3977 | 0.2974 | 2.6108e-06 |
| 47 | rs2725829 | 11 | 76733434 | rs4511648 | 19 | 35798852 | 1.6364 | 0.3491 | 2.76447e-06 |
| 48 | rs7973859 | 12 | 3855485 | rs5746996 | 22 | 15983801 | 0.8547 | 0.1900 | 6.8435e-06 |
| 49 | rs17662791 | 14 | 51086033 | rs9958551 | 18 | 43490826 | 3.4493 | 0.7545 | 4.84095e-06 |
| 50 | rs10484088 | 14 | 51102810 | rs9958551 | 18 | 43490826 | 17.9096 | 3.8728 | 3.75638e-06 |

Table 5.22: *Significant green Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0179 | KEGG:05166 | KEGG | Human T-cell leukemia virus 1 infection |

Table 5.23: *Significant gene set enrichment results of genes with SNPs found by green cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs12144395 | 1 | 97988897 | 0.0572 | 0.0123 | 3.605114e-06 | DPYD |
| 2 | rs17045707 | 3 | 6700175 | 0.0129 | 0.0029 | 8.053122e-06 | |
| 3 | rs2068197 | 4 | 178461632 | 0.0473 | 0.0099 | 1.871678e-06 | |
| 4 | rs562819 | 6 | 116385418 | 0.0145 | 0.0031 | 3.030728e-06 | FRK |
| 5 | rs12353094 | 9 | 7533336 | 0.0216 | 0.0045 | 1.675326e-06 | |
| 6 | rs17771757 | 9 | 28189845 | 0.0295 | 0.0062 | 1.880324e-06 | LINGO2 |
| 7 | rs1048370 | 9 | 139566627 | 0.0652 | 0.0136 | 1.60149e-06 | MRPL41 |
| 8 | rs11017185 | 10 | 131986157 | 0.0261 | 0.0057 | 5.136763e-06 | |
| 9 | rs7335956 | 13 | 101986271 | 0.0176 | 0.0037 | 1.697828e-06 | |
| 10 | rs8081464 | 17 | 74703613 | 0.0262 | 0.0055 | 1.87884e-06 | RBFOX3 |
| 11 | rs1440475 | 18 | 51130347 | 0.0392 | 0.0083 | 2.150068e-06 | TCF4 |
| 12 | rs3764532 | 19 | 63621034 | 0.0154 | 0.0035 | 9.331308e-06 | ZNF584 |
| 13 | rs11667591 | 19 | 63629449 | 0.0164 | 0.0035 | 2.290426e-06 | |
| 14 | rs1122955 | 19 | 63638015 | 0.0162 | 0.0035 | 3.771492e-06 | ZNF132 |
| 15 | rs12980907 | 19 | 63648138 | 0.0163 | 0.0035 | 3.724861e-06 | |
| 16 | rs2823375 | 21 | 15853950 | 0.0175 | 0.0037 | 1.734587e-06 | |
| 17 | rs9977399 | 21 | 23535457 | 0.0306 | 0.0066 | 3.677517e-06 | |

Table 5.24: *Significant red GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs17385885 | 1 | 15372802 | rs7502538 | 17 | 13299190 | -17.7143 | 3.7003 | 1.69035e-06 | | |
| 2 | rs1321117 | 1 | 30400206 | rs12881810 | 14 | 71993639 | 1.0136 | 0.2150 | 2.41313e-06 | | |
| 3 | rs13426 | 1 | 32028673 | rs3766379 | 1 | 159074339 | 1.4875 | 0.2981 | 6.06304e-07 | | |
| 4 | rs13426 | 1 | 32028673 | rs6671710 | 1 | 159075909 | 1.3881 | 0.3022 | 4.34981e-06 | | |
| 5 | rs13426 | 1 | 32028673 | rs512525 | 1 | 159080367 | 1.3813 | 0.3024 | 4.92128e-06 | | |
| 6 | rs13426 | 1 | 32028673 | rs11582719 | 1 | 159090436 | 1.4119 | 0.3019 | 2.91276e-06 | | |
| 7 | rs4915873 | 1 | 63307674 | rs9822730 | 3 | 114337365 | -1.2383 | 0.2742 | 6.31023e-06 | | |
| 8 | rs284169 | 1 | 91986030 | rs705099 | 4 | 37271341 | -1.1248 | 0.2209 | 3.55266e-07 | | |
| 9 | rs2207701 | 1 | 115111035 | rs1042717 | 5 | 148186839 | -2.7862 | 0.6137 | 5.63112e-06 | | |
| 10 | rs3765821 | 1 | 239860854 | rs6737623 | 2 | 42144989 | -2.2711 | 0.5073 | 7.56743e-06 | | |
| 11 | rs6737623 | 2 | 42144989 | rs12363683 | 11 | 87459489 | 2.6636 | 0.5837 | 5.04272e-06 | | |
| 12 | rs2567599 | 2 | 73847582 | rs11563057 | 2 | 234563208 | 2.4924 | 0.5341 | 3.06772e-06 | | |
| 13 | rs162810 | 3 | 7691889 | rs7324378 | 13 | 50577045 | -17.2315 | 3.6314 | 2.08341e-06 | | |
| 14 | rs780362 | 3 | 59734612 | rs1892669 | 21 | 31213372 | -1.5527 | 0.3275 | 2.12719e-06 | | |
| 15 | rs2971458 | 3 | 184697981 | rs3794370 | 13 | 22654627 | 1.3251 | 0.2906 | 5.10049e-06 | | |
| 16 | rs16861112 | 4 | 47890699 | rs17209451 | 5 | 65952734 | 1.2977 | 0.2860 | 5.6869e-06 | | |
| 17 | rs7668895 | 4 | 53151173 | rs7340852 | 4 | 65192764 | 1.2893 | 0.2765 | 3.11407e-06 | | |
| 18 | rs7668895 | 4 | 53151173 | rs4403082 | 4 | 65193230 | 1.2987 | 0.2803 | 3.60218e-06 | | |
| 19 | rs7668895 | 4 | 53151173 | rs766818 | 4 | 65195214 | 1.2530 | 0.2818 | 8.71394e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs3733506 | 4 | 53154377 | rs7340852 | 4 | 65192764 | 1.2912 | 0.2764 | 2.99961e-06 |
| 21 | rs3733506 | 4 | 53154377 | rs4403082 | 4 | 65193230 | 1.3011 | 0.2803 | 3.44955e-06 |
| 22 | rs3733506 | 4 | 53154377 | rs766818 | 4 | 65195214 | 1.2554 | 0.2817 | 8.34194e-06 |
| 23 | rs6840875 | 4 | 53299344 | rs7340852 | 4 | 65192764 | 1.4748 | 0.2890 | 3.32978e-07 |
| 24 | rs6840875 | 4 | 53299344 | rs4403082 | 4 | 65193230 | 1.5008 | 0.2929 | 3.00065e-07 |
| 25 | rs6840875 | 4 | 53299344 | rs766818 | 4 | 65195214 | 1.4251 | 0.2925 | 1.10087e-06 |
| 26 | rs4865408 | 4 | 53314124 | rs7340852 | 4 | 65192764 | 1.4411 | 0.2837 | 3.78264e-07 |
| 27 | rs4865408 | 4 | 53314124 | rs4403082 | 4 | 65193230 | 1.4678 | 0.2873 | 3.24594e-07 |
| 28 | rs4865408 | 4 | 53314124 | rs766818 | 4 | 65195214 | 1.3969 | 0.2874 | 1.17207e-06 |
| 29 | rs10009096 | 4 | 53318494 | rs7340852 | 4 | 65192764 | 1.4706 | 0.2860 | 2.73289e-07 |
| 30 | rs10009096 | 4 | 53318494 | rs4403082 | 4 | 65193230 | 1.4850 | 0.2899 | 3.02432e-07 |
| 31 | rs10009096 | 4 | 53318494 | rs766818 | 4 | 65195214 | 1.4092 | 0.2894 | 1.12051e-06 |
| 32 | rs6855459 | 4 | 64086972 | rs2806890 | 13 | 45360061 | -1.1212 | 0.2369 | 2.21014e-06 |
| 33 | rs6534554 | 4 | 127502750 | rs3758893 | 11 | 120480600 | -1.7188 | 0.3836 | 7.45224e-06 |
| 34 | rs1993616 | 4 | 174056940 | rs219816 | 7 | 98448531 | -17.0403 | 3.6457 | 2.95293e-06 |
| 35 | rs1993616 | 4 | 174056940 | rs219815 | 7 | 98448620 | -21.6502 | 4.8658 | 8.60983e-06 |
| 36 | rs9361116 | 6 | 77470954 | rs2211848 | 21 | 38289313 | 1.5022 | 0.3397 | 9.75834e-06 |
| 37 | rs11153611 | 6 | 116871256 | rs2516737 | 16 | 2041750 | -2.2439 | 0.5071 | 9.65482e-06 |
| 38 | rs11153611 | 6 | 116871256 | rs1800720 | 16 | 2045401 | -2.2607 | 0.5100 | 9.30948e-06 |
| 39 | rs2051950 | 7 | 86833858 | rs8012614 | 14 | 73024217 | -1.6106 | 0.3637 | 9.48057e-06 |
| 40 | rs11977045 | 7 | 130480376 | rs4875371 | 8 | 4479917 | -1.5120 | 0.3111 | 1.17044e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs1439794 | 7 | 130485199 | rs4875371 | 8 | 4479917 | -1.3963 | 0.3060 | 5.03862e-06 |
| 42 | rs7462672 | 8 | 11935618 | rs933577 | 17 | 50807334 | 1.4449 | 0.2996 | 1.41339e-06 |
| 43 | rs1531848 | 8 | 94306462 | rs10835507 | 11 | 29268995 | 0.9141 | 0.2064 | 9.51145e-06 |
| 44 | rs882259 | 9 | 2589564 | rs3911554 | 11 | 80827181 | 0.9511 | 0.1985 | 1.66173e-06 |
| 45 | rs885578 | 9 | 107906281 | rs9320004 | 18 | 58029028 | 17.8708 | 3.7113 | 1.47085e-06 |
| 46 | rs10509936 | 10 | 113026683 | rs953905 | 13 | 73891001 | 1.0962 | 0.2473 | 9.30599e-06 |
| 47 | rs10736280 | 10 | 121827417 | rs2823357 | 21 | 15836776 | -3.0359 | 0.6601 | 4.2461e-06 |
| 48 | rs10736280 | 10 | 121827417 | rs2179030 | 21 | 15837982 | -2.9384 | 0.6331 | 3.45988e-06 |
| 49 | rs4751072 | 10 | 130899243 | rs7124697 | 11 | 24605938 | 0.9887 | 0.2219 | 8.37049e-06 |
| 50 | rs9804629 | 11 | 82413697 | rs8129780 | 21 | 21916220 | -18.4187 | 4.0891 | 6.65694e-06 |
| 51 | rs9804629 | 11 | 82413697 | rs17003913 | 21 | 21917013 | -23.1629 | 5.0585 | 4.67141e-06 |
| 52 | rs12602753 | 17 | 62332794 | rs9320004 | 18 | 58029028 | 1.1025 | 0.2202 | 5.52669e-07 |
| 53 | rs12602753 | 17 | 62332794 | rs978572 | 18 | 58073170 | 1.0726 | 0.2253 | 1.93415e-06 |
| 54 | rs12602753 | 17 | 62332794 | rs2980981 | 18 | 58098098 | 1.0292 | 0.2206 | 3.07117e-06 |

Table 5.25: *Significant red Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0497 | CORUM:5262 | CORUM | TCF4-CTNNB1 complex |

Table 5.26: *Significant gene set enrichment results of genes with SNPs found by red cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs10779486 | 1 | 206805932 | -0.0156 | 0.0034 | 3.672534e-06 | |
| 2 | rs11925421 | 3 | 147370852 | -0.0206 | 0.0040 | 3.170029e-07 | |
| 3 | rs13122003 | 4 | 23368426 | -0.0128 | 0.0028 | 4.27892e-06 | PPARGC1A |
| 4 | rs4697418 | 4 | 23368811 | -0.0133 | 0.0028 | 1.813714e-06 | PPARGC1A |
| 5 | rs6914506 | 6 | 37374934 | 0.0118 | 0.0023 | 4.047782e-07 | TBC1D22B |
| 6 | rs12235745 | 9 | 12530358 | 0.0167 | 0.0038 | 9.082487e-06 | |
| 7 | rs1766059 | 9 | 116078932 | -0.0131 | 0.0029 | 6.857047e-06 | COL27A1 |
| 8 | rs10982134 | 9 | 116090819 | 0.0130 | 0.0029 | 6.407809e-06 | COL27A1 |
| 9 | rs235083 | 16 | 64760511 | 0.0123 | 0.0025 | 5.498986e-07 | |
| 10 | rs1822215 | 18 | 49501501 | 0.0135 | 0.0029 | 2.852245e-06 | |
| 11 | rs16980635 | 20 | 19394626 | -0.0130 | 0.0028 | 2.706448e-06 | SLC24A3 |
| 12 | rs1538218 | 20 | 22444290 | 0.0613 | 0.0134 | 5.181347e-06 | |

Table 5.27: *Significant black GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs7537470 | 1 | 58774809 | rs10871631 | 18 | 64564484 | 1.4029 | 0.3031 | 3.69222e-06 | | |
| 2 | rs3136247 | 2 | 47866603 | rs13113465 | 4 | 165917658 | 1.1195 | 0.2512 | 8.31952e-06 | | |
| 3 | rs10865358 | 2 | 66900069 | rs9939674 | 16 | 16831521 | -1.3684 | 0.3060 | 7.76812e-06 | | |
| 4 | rs1008800 | 3 | 191677262 | rs8042025 | 15 | 92391548 | 1.1654 | 0.2563 | 5.44436e-06 | | |
| 5 | rs4487342 | 4 | 7568079 | rs17349444 | 13 | 82122720 | -17.3758 | 3.8501 | 6.38988e-06 | | |
| 6 | rs9918079 | 4 | 15154508 | rs2930230 | 16 | 83874666 | -2.3389 | 0.5193 | 6.67054e-06 | | |
| 7 | rs11728037 | 4 | 15203996 | rs2930230 | 16 | 83874666 | -2.4446 | 0.5050 | 1.29232e-06 | | |
| 8 | rs2571494 | 4 | 29743146 | rs2253907 | 6 | 31444849 | 0.9725 | 0.2078 | 2.87074e-06 | | |
| 9 | rs2571494 | 4 | 29743146 | rs2844558 | 6 | 31448412 | -0.9268 | 0.1977 | 2.75928e-06 | | |
| 10 | rs7663168 | 4 | 67020209 | rs414845 | 21 | 39176335 | -1.7270 | 0.3737 | 3.81087e-06 | | |
| 11 | rs4619859 | 4 | 94918032 | rs7896076 | 10 | 15912833 | -17.4349 | 3.9409 | 9.68226e-06 | | |
| 12 | rs732020 | 4 | 94920377 | rs7896076 | 10 | 15912833 | -17.4286 | 3.9450 | 9.96727e-06 | | |
| 13 | rs1453871 | 4 | 100167429 | rs304379 | 5 | 121796570 | 0.9916 | 0.2206 | 6.95238e-06 | | |
| 14 | rs1238741 | 4 | 100202335 | rs7323392 | 13 | 23164521 | -2.3326 | 0.5195 | 7.11327e-06 | | |
| 15 | rs8086 | 4 | 185914415 | rs7806 | 20 | 54412578 | 1.1596 | 0.2601 | 8.2869e-06 | | |
| 16 | rs1803898 | 4 | 185938543 | rs2284060 | 22 | 36873399 | -0.9556 | 0.2036 | 2.67939e-06 | | |
| 17 | rs7698980 | 4 | 186412191 | rs2064108 | 6 | 5609463 | -1.0657 | 0.2392 | 8.37314e-06 | | |
| 18 | rs10473829 | 5 | 27337279 | rs7821554 | 8 | 106867851 | 2.3199 | 0.5088 | 5.12265e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 19 | rs9313564 | 5 | 171274454 | rs12587001 | 14 | 103671930 | -1.3761 | 0.2809 | 9.60461e-07 |
| 20 | rs307805 | 5 | 180010093 | rs1791619 | 12 | 51057341 | -2.0340 | 0.4470 | 5.36448e-06 |
| 21 | rs2844558 | 6 | 31448412 | rs7996365 | 13 | 67200857 | -1.9982 | 0.4399 | 5.57184e-06 |
| 22 | rs2855448 | 6 | 33244553 | rs1796135 | 12 | 77512679 | 0.9289 | 0.2070 | 7.23374e-06 |
| 23 | rs6954595 | 7 | 19905468 | rs702827 | 7 | 29457928 | 0.9967 | 0.2217 | 6.90662e-06 |
| 24 | rs17513961 | 7 | 46359092 | rs12242999 | 10 | 16717982 | 1.4122 | 0.3119 | 5.96359e-06 |
| 25 | rs921773 | 8 | 13275292 | rs2294618 | 16 | 1754392 | -17.2755 | 3.7218 | 3.45483e-06 |
| 26 | rs16883597 | 8 | 35139007 | rs12323080 | 13 | 100841576 | 1.9783 | 0.4473 | 9.7467e-06 |
| 27 | rs1947300 | 8 | 35195174 | rs12323080 | 13 | 100841576 | 2.2536 | 0.5099 | 9.87656e-06 |
| 28 | rs17456596 | 8 | 80946553 | rs7288416 | 22 | 27129436 | 1.0053 | 0.2192 | 4.50629e-06 |
| 29 | rs10956824 | 8 | 93257078 | rs8086318 | 18 | 29517318 | -0.9459 | 0.2039 | 3.49948e-06 |
| 30 | rs1331436 | 9 | 88358520 | rs12831803 | 12 | 124127104 | 1.8995 | 0.4240 | 7.47958e-06 |
| 31 | rs9988765 | 10 | 82705777 | rs4789374 | 17 | 72423955 | -17.3086 | 3.7812 | 4.70532e-06 |
| 32 | rs7073243 | 10 | 112592257 | rs11908625 | 20 | 23378230 | 2.7529 | 0.6086 | 6.09455e-06 |
| 33 | rs3741845 | 12 | 10853382 | rs1791619 | 12 | 51057341 | -1.1086 | 0.2486 | 8.24971e-06 |
| 34 | rs3741845 | 12 | 10853382 | rs10459511 | 14 | 55336569 | -1.7295 | 0.3242 | 9.56024e-08 |
| 35 | rs10881068 | 12 | 45868962 | rs2839010 | 21 | 46067460 | 0.9623 | 0.2147 | 7.41232e-06 |
| 36 | rs2272550 | 14 | 21661828 | rs7187470 | 16 | 20327276 | -1.1308 | 0.2512 | 6.72258e-06 |
| 37 | rs4794255 | 17 | 46907675 | rs3746236 | 19 | 39990801 | -1.0497 | 0.2367 | 9.18601e-06 |

Table 5.28: *Significant black Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0249 | GO:1990843 | GO:CC | subsarcolemmal mitochondrion |
| 2 | 0.0249 | GO:1990844 | GO:CC | interfibrillar mitochondrion |
| 3 | 0.0442 | WP:WP3407 | WP | FTO Obesity Variant Mechanism |
| 4 | 0.0442 | WP:WP4191 | WP | Caloric restriction and aging |

Table 5.29: *Significant gene set enrichment results of genes with SNPs found by black cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs7531445 | 1 | 86439481 | 0.0203 | 0.0043 | 2.132489e-06 | |
| 2 | rs7558478 | 2 | 21041664 | 0.0458 | 0.0103 | 8.519934e-06 | |
| 3 | rs1896701 | 5 | 23504549 | 0.0239 | 0.0052 | 3.875019e-06 | PRDM9 |
| 4 | rs13362182 | 5 | 85575598 | 0.0364 | 0.0079 | 4.480616e-06 | |
| 5 | rs17058082 | 6 | 130214612 | 0.0471 | 0.0090 | 1.885517e-07 | TMEM244 |
| 6 | rs9364788 | 6 | 165378269 | 0.0126 | 0.0028 | 9.090796e-06 | |
| 7 | rs4475440 | 8 | 32846452 | 0.0411 | 0.0084 | 9.092389e-07 | |
| 8 | rs16880105 | 8 | 32849239 | 0.0410 | 0.0084 | 9.825619e-07 | |
| 9 | rs16880178 | 8 | 32939262 | 0.0508 | 0.0097 | 1.748787e-07 | |
| 10 | rs9643901 | 8 | 38898804 | 0.0138 | 0.0030 | 6.163741e-06 | PLEKHA2 |
| 11 | rs875056 | 8 | 142652032 | 0.0191 | 0.0038 | 6.880124e-07 | |
| 12 | rs1453580 | 11 | 99286357 | 0.0344 | 0.0072 | 1.809675e-06 | CNTN5 |
| 13 | rs11068753 | 12 | 116864562 | 0.0162 | 0.0036 | 8.607632e-06 | KSR2 |
| 14 | rs4905757 | 14 | 98252379 | 0.0243 | 0.0055 | 9.735744e-06 | |

Table 5.30: *Significant pink GWAS Results*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs7513083 | 1 | 20010708 | 0.0420 | 0.0086 | 1.180802e-06 | |
| 2 | rs10914230 | 1 | 31142100 | 0.0570 | 0.0107 | 1.004611e-07 | SDC3 |
| 3 | rs356392 | 1 | 71618040 | 0.0490 | 0.0100 | 9.407864e-07 | |
| 4 | rs4262503 | 1 | 180518932 | 0.0227 | 0.0049 | 4.063218e-06 | |
| 5 | rs12085373 | 1 | 191430770 | 0.0450 | 0.0098 | 4.082378e-06 | CDC73 |
| 6 | rs7519081 | 1 | 194324511 | 0.0445 | 0.0096 | 4.026014e-06 | |
| 7 | rs17022092 | 1 | 212418241 | 0.0353 | 0.0074 | 2.111922e-06 | |
| 8 | rs11900406 | 2 | 74748856 | 0.0357 | 0.0080 | 9.439586e-06 | SEMA4F |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | rs17014064 | 2 | 77494416 | 0.0521 | 0.0117 | 8.994745e-06 | LRRTM4 |
| 10 | rs10178717 | 2 | 109809934 | 0.0328 | 0.0070 | 2.922429e-06 | |
| 11 | rs2922309 | 2 | 177188300 | 0.0182 | 0.0036 | 6.345909e-07 | |
| 12 | rs16863360 | 2 | 222673958 | 0.0480 | 0.0090 | 9.439669e-08 | |
| 13 | rs10211582 | 2 | 227299685 | 0.0424 | 0.0095 | 7.775491e-06 | |
| 14 | rs16825543 | 2 | 229556193 | 0.0498 | 0.0102 | 1.027337e-06 | PID1 |
| 15 | rs9878022 | 3 | 1118424 | 0.0184 | 0.0036 | 3.129339e-07 | CNTN6 |
| 16 | rs6442206 | 3 | 1124544 | 0.0177 | 0.0036 | 1.024663e-06 | CNTN6 |
| 17 | rs1504076 | 3 | 1129559 | 0.0208 | 0.0043 | 1.556434e-06 | CNTN6 |
| 18 | rs2872338 | 3 | 1137041 | 0.0220 | 0.0043 | 2.47368e-07 | CNTN6 |
| 19 | rs11925472 | 3 | 88921014 | 0.0578 | 0.0097 | 3.22674e-09 | |
| 20 | rs6788974 | 3 | 88954827 | 0.0435 | 0.0091 | 1.901243e-06 | |
| 21 | rs11921073 | 3 | 89693118 | 0.0550 | 0.0112 | 8.641648e-07 | |
| 22 | rs7649531 | 3 | 187356269 | 0.0361 | 0.0081 | 8.977747e-06 | DGKG |
| 23 | rs11930898 | 4 | 34710972 | 0.0399 | 0.0089 | 7.150223e-06 | |
| 24 | rs6447170 | 4 | 42262372 | 0.0202 | 0.0045 | 8.131334e-06 | ATP8A1 |
| 25 | rs11934086 | 4 | 76867439 | 0.0332 | 0.0073 | 6.030591e-06 | G3BP2 |
| 26 | rs7698406 | 4 | 88453617 | 0.0518 | 0.0109 | 1.984204e-06 | HSD17B13 |
| 27 | rs9994891 | 4 | 111010902 | 0.0313 | 0.0065 | 1.507277e-06 | LRIT3 |
| 28 | rs17050077 | 4 | 120234346 | 0.0500 | 0.0092 | 5.608889e-08 | |
| 29 | rs17007008 | 4 | 142360954 | 0.0476 | 0.0105 | 6.164935e-06 | |
| 30 | rs6824449 | 4 | 149210445 | 0.0114 | 0.0025 | 6.055101e-06 | ARHGAP10 |
| 31 | rs11933380 | 4 | 149216852 | 0.0121 | 0.0026 | 2.391101e-06 | |
| 32 | rs5534 | 4 | 149220535 | 0.0113 | 0.0025 | 3.963045e-06 | NR3C2 |
| 33 | rs4396973 | 4 | 153866528 | 0.0485 | 0.0095 | 3.460692e-07 | |
| 34 | rs6817090 | 4 | 156999620 | 0.0141 | 0.0032 | 8.673273e-06 | ASIC5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 35 | rs17314234 | 4 | 157030205 | 0.0196 | 0.0041 | 1.644914e-06 | TDO2 |
| 36 | rs17054390 | 4 | 169766756 | 0.0500 | 0.0101 | 8.028544e-07 | PALLD |
| 37 | rs10013501 | 4 | 185851469 | 0.0577 | 0.0107 | 6.49959e-08 | PRIMPOL |
| 38 | rs10023035 | 4 | 189707217 | 0.0583 | 0.0102 | 1.126058e-08 | |
| 39 | rs4481275 | 4 | 191066832 | 0.0281 | 0.0060 | 3.510118e-06 | |
| 40 | rs7711748 | 5 | 689589 | 0.0608 | 0.0117 | 1.876282e-07 | CEP72 |
| 41 | rs4540151 | 5 | 5411146 | 0.0470 | 0.0105 | 7.656081e-06 | |
| 42 | rs16902110 | 5 | 12304916 | 0.0464 | 0.0098 | 2.404024e-06 | |
| 43 | rs16894229 | 5 | 25229794 | 0.0476 | 0.0106 | 7.668456e-06 | |
| 44 | rs16884834 | 5 | 55448597 | 0.0345 | 0.0076 | 6.13885e-06 | ANKRD55 |
| 45 | rs6869150 | 5 | 71832199 | 0.0513 | 0.0108 | 2.281972e-06 | ZNF366 |
| 46 | rs1150555 | 6 | 11108503 | 0.0582 | 0.0116 | 5.338111e-07 | ELOVL2 |
| 47 | rs7767058 | 6 | 75640352 | 0.0454 | 0.0097 | 2.91443e-06 | |
| 48 | rs6934826 | 6 | 79101608 | 0.0373 | 0.0081 | 4.38008e-06 | |
| 49 | rs9456853 | 6 | 163639068 | 0.0532 | 0.0098 | 6.301062e-08 | PACRG |
| 50 | rs2526615 | 7 | 19024189 | 0.0558 | 0.0105 | 1.127028e-07 | |
| 51 | rs10226401 | 7 | 86295461 | 0.0496 | 0.0106 | 3.101765e-06 | GRM3 |
| 52 | rs7792679 | 7 | 154504393 | 0.0564 | 0.0106 | 1.225709e-07 | HTR5A |
| 53 | rs6946401 | 7 | 155921602 | 0.0359 | 0.0078 | 3.989848e-06 | |
| 54 | rs9886459 | 8 | 9032820 | 0.0518 | 0.0103 | 4.546778e-07 | PPP1R3B |
| 55 | rs17054447 | 8 | 25759079 | 0.0491 | 0.0107 | 4.409374e-06 | EBF2 |
| 56 | rs16876321 | 8 | 29984758 | 0.0570 | 0.0107 | 9.911058e-08 | |
| 57 | rs6987116 | 8 | 32971639 | 0.0554 | 0.0111 | 7.016618e-07 | |
| 58 | rs7003403 | 8 | 37860475 | 0.0575 | 0.0113 | 4.160963e-07 | RAB11FIP1 |
| 59 | rs16891358 | 8 | 42509836 | 0.0791 | 0.0116 | 1.226689e-11 | SLC20A2 |
| 60 | rs6990159 | 8 | 42523979 | 0.0308 | 0.0065 | 1.940991e-06 | SMIM19 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 61 | rs13439165 | 8 | 87120095 | 0.0389 | 0.0087 | 8.341251e-06 | ATP6V0D2 |
| 62 | rs16926678 | 9 | 10960165 | 0.0188 | 0.0040 | 2.291648e-06 | |
| 63 | rs4097990 | 9 | 11053519 | 0.0218 | 0.0040 | 6.974418e-08 | |
| 64 | rs16935681 | 9 | 17460126 | 0.0460 | 0.0094 | 8.980852e-07 | CNTLN |
| 65 | rs7027646 | 9 | 139341403 | 0.0233 | 0.0050 | 3.7343e-06 | EXD3 |
| 66 | rs7907639 | 10 | 60537173 | 0.0417 | 0.0089 | 3.141399e-06 | |
| 67 | rs12416478 | 10 | 132736140 | 0.0128 | 0.0029 | 7.5594e-06 | |
| 68 | rs1329152 | 10 | 135153022 | 0.0282 | 0.0060 | 3.28626e-06 | SCART1 |
| 69 | rs7924536 | 11 | 12874326 | 0.0283 | 0.0061 | 3.113337e-06 | TEAD1 |
| 70 | rs12292743 | 11 | 24574951 | 0.0341 | 0.0076 | 7.201329e-06 | LUZP2 |
| 71 | rs7110236 | 11 | 118908226 | 0.0368 | 0.0075 | 8.742096e-07 | |
| 72 | rs12314360 | 12 | 8711391 | 0.0396 | 0.0088 | 6.752011e-06 | |
| 73 | rs16913692 | 12 | 18035565 | 0.0452 | 0.0098 | 3.970053e-06 | |
| 74 | rs12307979 | 12 | 44533920 | 0.0348 | 0.0076 | 4.798102e-06 | ARID2 |
| 75 | rs7303119 | 12 | 44541262 | 0.0371 | 0.0081 | 4.787352e-06 | ARID2 |
| 76 | rs10161510 | 12 | 44543753 | 0.0374 | 0.0081 | 4.343722e-06 | ARID2 |
| 77 | rs1997279 | 12 | 79461884 | 0.0201 | 0.0043 | 2.568212e-06 | PTPRQ |
| 78 | rs7317427 | 13 | 27518563 | 0.0487 | 0.0100 | 1.202044e-06 | FLT3 |
| 79 | rs7993593 | 13 | 27541728 | 0.0677 | 0.0103 | 5.891557e-11 | FLT3 |
| 80 | rs16944562 | 13 | 89938215 | 0.0469 | 0.0099 | 2.086672e-06 | |
| 81 | rs4778547 | 15 | 21323644 | 0.0383 | 0.0082 | 2.880616e-06 | |
| 82 | rs1400896 | 15 | 34413201 | 0.0367 | 0.0080 | 4.950426e-06 | |
| 83 | rs8039750 | 15 | 56740661 | 0.0472 | 0.0089 | 1.330753e-07 | ADAM10 |
| 84 | rs720273 | 15 | 56811819 | 0.0469 | 0.0105 | 8.937799e-06 | ADAM10 |
| 85 | rs3759822 | 15 | 66355366 | 0.0582 | 0.0114 | 3.343013e-07 | |
| 86 | rs16970108 | 15 | 76886530 | 0.0474 | 0.0094 | 4.184148e-07 | ADAMTS7 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 87 | rs2302085 | 15 | 90495211 | 0.0473 | 0.0101 | 2.798805e-06 | SLCO3A1 |
| 88 | rs8060156 | 16 | 13673522 | 0.0382 | 0.0086 | 8.015921e-06 | |
| 89 | rs6498703 | 16 | 17410330 | 0.0132 | 0.0030 | 8.419755e-06 | XYLT1 |
| 90 | rs4547344 | 16 | 78341778 | 0.0118 | 0.0026 | 8.434265e-06 | |
| 91 | rs13338235 | 16 | 78355131 | 0.0126 | 0.0027 | 2.779489e-06 | |
| 92 | rs7239517 | 18 | 1760403 | 0.0538 | 0.0105 | 3.069724e-07 | |
| 93 | rs8088506 | 18 | 53915357 | 0.0597 | 0.0115 | 2.436038e-07 | NEDD4L |
| 94 | rs17056358 | 18 | 70950152 | 0.0506 | 0.0108 | 2.906366e-06 | |
| 95 | rs567717 | 18 | 75610863 | 0.0476 | 0.0100 | 1.876543e-06 | CTDP1 |
| 96 | rs524643 | 18 | 75616126 | 0.0472 | 0.0099 | 2.128941e-06 | |
| 97 | rs7250069 | 19 | 16682566 | 0.0620 | 0.0110 | 1.841046e-08 | |
| 98 | rs2823875 | 21 | 16847242 | 0.0434 | 0.0071 | 1.042645e-09 | |
| 99 | rs171479 | 21 | 28078751 | 0.0386 | 0.0086 | 7.339007e-06 | |
| 100 | rs8132871 | 21 | 41732690 | 0.0162 | 0.0034 | 2.532079e-06 | MX1 |
| 101 | rs5748302 | 22 | 17939409 | 0.0539 | 0.0102 | 1.357992e-07 | |
| 102 | rs7290537 | 22 | 48312295 | 0.0605 | 0.0103 | 4.880806e-09 | |

Table 5.33: *Significant magenta GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs7520924 | 1 | 62273143 | rs12640158 | 4 | 104940847 | 1.7675 | 0.3751 | 2.44971e-06 | | |
| 2 | rs12410256 | 1 | 155496893 | rs1056032 | 6 | 28978954 | 18.1830 | 4.0424 | 6.85652e-06 | | |
| 3 | rs12410256 | 1 | 155496893 | rs7762289 | 6 | 29018214 | 18.3579 | 4.0951 | 7.3634e-06 | | |
| 4 | rs12410256 | 1 | 155496893 | rs4947256 | 6 | 29029919 | 18.2365 | 4.0648 | 7.24373e-06 | | |
| 5 | rs16834300 | 1 | 190684303 | rs6054399 | 20 | 6547698 | 16.8988 | 3.7025 | 5.01514e-06 | | |
| 6 | rs1033928 | 1 | 194430520 | rs3736919 | 13 | 35837688 | -1.1001 | 0.2458 | 7.61243e-06 | | |
| 7 | rs7604720 | 2 | 26591404 | rs9789218 | 18 | 64284356 | 1.0979 | 0.2477 | 9.34751e-06 | | |
| 8 | rs17025044 | 2 | 40169403 | rs17690319 | 7 | 41084586 | -17.5766 | 3.5960 | 1.01995e-06 | | |
| 9 | rs3754568 | 2 | 46224718 | rs6054399 | 20 | 6547698 | -1.0248 | 0.2306 | 8.81437e-06 | | |
| 10 | rs9811522 | 3 | 5854366 | rs925319 | 18 | 66623186 | -17.1518 | 3.6727 | 3.0102e-06 | | |
| 11 | rs9834970 | 3 | 36831034 | rs12364918 | 11 | 130246625 | 1.8564 | 0.3823 | 1.20128e-06 | | |
| 12 | rs1553656 | 3 | 36834709 | rs12364918 | 11 | 130246625 | -1.7587 | 0.3894 | 6.29289e-06 | | |
| 13 | rs4789 | 3 | 36844443 | rs12364918 | 11 | 130246625 | -1.7590 | 0.3892 | 6.2122e-06 | | |
| 14 | rs2167180 | 3 | 67763797 | rs2612060 | 12 | 64529769 | -2.0481 | 0.4486 | 4.97941e-06 | | |
| 15 | rs2316263 | 3 | 101382786 | rs4883536 | 12 | 131746819 | -1.0119 | 0.2263 | 7.73409e-06 | | |
| 16 | rs1812310 | 4 | 5272479 | rs12230547 | 12 | 57526292 | -17.1231 | 3.7837 | 6.0271e-06 | | |
| 17 | rs4100758 | 5 | 148626971 | rs17420770 | 8 | 20503992 | 1.7703 | 0.3789 | 2.98131e-06 | | |
| 18 | rs6458116 | 6 | 39507637 | rs2823188 | 21 | 15550904 | 2.3543 | 0.5069 | 3.4081e-06 | | |
| 19 | rs7807193 | 7 | 37901198 | rs11983950 | 7 | 53422386 | -1.4160 | 0.3161 | 7.48468e-06 | | |
| 20 | rs11774078 | 8 | 125017223 | rs10746936 | 9 | 75976117 | 1.2251 | 0.2690 | 5.24179e-06 | | |
| 21 | rs10962915 | 9 | 1721076 | rs2664374 | 20 | 5607444 | 1.2852 | 0.2709 | 2.09985e-06 | | |
| 22 | rs622173 | 11 | 1842723 | rs12449822 | 17 | 74867283 | 17.2612 | 3.7549 | 4.28722e-06 | | |
| 23 | rs1005842 | 15 | 79768620 | rs13058323 | 22 | 20770829 | 2.8986 | 0.5920 | 9.77038e-07 | | |
| 24 | rs12916179 | 15 | 79845673 | rs13058323 | 22 | 20770829 | 2.4541 | 0.5514 | 8.54643e-06 | | |
| 25 | rs12913258 | 15 | 79849884 | rs13058323 | 22 | 20770829 | 2.5603 | 0.5607 | 4.96805e-06 | | |
| 26 | rs9980212 | 21 | 39376918 | rs2110412 | 22 | 19349125 | 1.2227 | 0.2692 | 5.55825e-06 | | |

Table 5.31: *Significant pink Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0499 | GO:0010844 | GO:MF | recombination hotspot binding |

Table 5.32: *Significant gene set enrichment results of genes with SNPs found by pink cluster analysis. P-values are analytically adjusted.*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs7513083 | 1 | 20010708 | rs17825455 | 17 | 5783519 | -2.7033 | 0.5975 | 6.05738e-06 | | |
| 2 | rs2473277 | 1 | 22234432 | rs17135441 | 16 | 1621753 | -17.2259 | 3.5822 | 1.51926e-06 | | |
| 3 | rs12026014 | 1 | 38833082 | rs7196594 | 16 | 24298206 | 3.1005 | 0.6937 | 7.8484e-06 | | |
| 4 | rs1746747 | 1 | 75915283 | rs11672222 | 19 | 56919728 | 3.2168 | 0.6772 | 2.03014e-06 | | |
| 5 | rs17118192 | 1 | 98575051 | rs7548237 | 1 | 183394873 | -1.8104 | 0.3859 | 2.70676e-06 | | |
| 6 | rs1282275 | 1 | 111431094 | rs4540151 | 5 | 5411146 | 18.5410 | 4.0662 | 5.12015e-06 | | |
| 7 | rs1282275 | 1 | 111431094 | rs234369 | 19 | 45168577 | 2.9522 | 0.6469 | 5.02603e-06 | | |
| 8 | rs7515448 | 1 | 150638659 | rs6126303 | 20 | 49740539 | -2.7786 | 0.6228 | 8.14578e-06 | | |
| 9 | rs6427202 | 1 | 167795454 | rs2907465 | 17 | 58738171 | -1.0689 | 0.2364 | 6.12601e-06 | | |
| 10 | rs7546105 | 1 | 187289233 | rs16936162 | 10 | 35928643 | 3.6672 | 0.7687 | 1.83831e-06 | | |
| 11 | rs12065500 | 1 | 197740157 | rs7005571 | 8 | 66914425 | 4.1640 | 0.8917 | 3.01893e-06 | | |
| 12 | rs12065500 | 1 | 197740157 | rs7005641 | 8 | 102193073 | 4.0596 | 0.8083 | 5.09669e-07 | | |
| 13 | rs9430018 | 1 | 208063165 | rs11068512 | 12 | 116413880 | -17.3912 | 3.9364 | 9.96194e-06 | | |
| 14 | rs9430019 | 1 | 208117417 | rs2903728 | 19 | 61100107 | 2.4636 | 0.5239 | 2.56851e-06 | | |
| 15 | rs823198 | 2 | 2869699 | rs6800366 | 3 | 54665399 | 17.8172 | 3.9975 | 8.30775e-06 | | |
| 16 | rs2693822 | 2 | 6083374 | rs2352224 | 6 | 67207301 | 3.7601 | 0.7453 | 4.52819e-07 | | |
| 17 | rs6724922 | 2 | 79824756 | rs7762767 | 6 | 90058184 | 1.7550 | 0.3912 | 7.24621e-06 | | |
| 18 | rs10178940 | 2 | 123311751 | rs9604511 | 13 | 113486684 | 2.8007 | 0.5982 | 2.83994e-06 | | |
| 19 | rs10178940 | 2 | 123311751 | rs7489888 | 13 | 113502415 | 2.5770 | 0.5765 | 7.8304e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs2551649 | 2 | 208064454 | rs10092425 | 8 | 67098659 | 2.8143 | 0.6303 | 7.99944e-06 |
| 21 | rs6755425 | 2 | 208065237 | rs10092425 | 8 | 67098659 | 3.1164 | 0.6604 | 2.3678e-06 |
| 22 | rs16866930 | 2 | 226861513 | rs10499163 | 6 | 130004326 | 4.0666 | 0.8669 | 2.71551e-06 |
| 23 | rs2686702 | 3 | 4165452 | rs12361365 | 11 | 68348448 | -1.9281 | 0.4229 | 5.13788e-06 |
| 24 | rs6599257 | 3 | 38779592 | rs4872439 | 8 | 26382300 | 1.2245 | 0.2673 | 4.63299e-06 |
| 25 | rs2699975 | 3 | 110453409 | rs2893954 | 10 | 65718694 | 17.7018 | 3.9239 | 6.443e-06 |
| 26 | rs893883 | 3 | 144050017 | rs853553 | 6 | 110976298 | -17.5461 | 3.8360 | 4.78366e-06 |
| 27 | rs7624371 | 3 | 183294929 | rs7904164 | 10 | 119613646 | -17.8961 | 3.9399 | 5.56486e-06 |
| 28 | rs4298137 | 4 | 14412892 | rs992592 | 5 | 98095732 | -1.8689 | 0.4093 | 4.96418e-06 |
| 29 | rs9291218 | 4 | 42185437 | rs4899038 | 14 | 60859545 | -1.8708 | 0.4223 | 9.40518e-06 |
| 30 | rs10222924 | 4 | 48711518 | rs7965179 | 12 | 51791594 | -1.7789 | 0.4026 | 9.92621e-06 |
| 31 | rs2768957 | 4 | 48748279 | rs7965179 | 12 | 51791594 | -1.7835 | 0.4035 | 9.8402e-06 |
| 32 | rs1051447 | 4 | 48758629 | rs7965179 | 12 | 51791594 | -1.8574 | 0.4016 | 3.75441e-06 |
| 33 | rs28498138 | 4 | 71515512 | rs7831728 | 8 | 29261722 | 3.9938 | 0.8852 | 6.43056e-06 |
| 34 | rs28690964 | 4 | 71515698 | rs7831728 | 8 | 29261722 | 3.9912 | 0.8852 | 6.51892e-06 |
| 35 | rs1962495 | 4 | 115532764 | rs2059117 | 4 | 181678796 | 1.1316 | 0.2544 | 8.68676e-06 |
| 36 | rs1962495 | 4 | 115532764 | rs1376508 | 8 | 35081747 | -1.8596 | 0.4195 | 9.28618e-06 |
| 37 | rs7674531 | 4 | 165105510 | rs10499163 | 6 | 130004326 | 2.7837 | 0.6092 | 4.88203e-06 |
| 38 | rs4690554 | 4 | 178836765 | rs697550 | 5 | 14125453 | 17.0624 | 3.7910 | 6.76863e-06 |
| 39 | rs7669469 | 4 | 181700302 | rs11150032 | 16 | 76595394 | -17.4714 | 3.3131 | 1.33896e-07 |
| 40 | rs11943826 | 4 | 181711225 | rs11150032 | 16 | 76595394 | -2.8583 | 0.6382 | 7.50889e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs4426911 | 5 | 125806834 | rs4782717 | 16 | 81167218 | -1.7019 | 0.3678 | 3.71386e-06 |
| 42 | rs4426911 | 5 | 125806834 | rs459756 | 21 | 26598272 | 1.3782 | 0.2839 | 1.20289e-06 |
| 43 | rs2161747 | 5 | 146651578 | rs9595770 | 13 | 47219459 | -1.5016 | 0.3045 | 8.17068e-07 |
| 44 | rs7760394 | 6 | 45686882 | rs7928765 | 11 | 126287995 | -2.2074 | 0.4418 | 5.84518e-07 |
| 45 | rs9456696 | 6 | 161989559 | rs9604511 | 13 | 113486684 | 18.5316 | 3.8816 | 1.80375e-06 |
| 46 | rs11770984 | 7 | 26137739 | rs2925877 | 8 | 129318573 | 1.0398 | 0.2341 | 8.9224e-06 |
| 47 | rs7783212 | 7 | 153773484 | rs4918948 | 10 | 97306030 | -1.6880 | 0.3612 | 2.9665e-06 |
| 48 | rs7831728 | 8 | 29261722 | rs2813277 | 10 | 51573668 | 4.0065 | 0.8970 | 7.95521e-06 |
| 49 | rs7831728 | 8 | 29261722 | rs17058178 | 13 | 38014394 | 3.1900 | 0.7110 | 7.22913e-06 |
| 50 | rs7831728 | 8 | 29261722 | rs10149020 | 14 | 52968356 | 3.3903 | 0.7325 | 3.6867e-06 |
| 51 | rs7831728 | 8 | 29261722 | rs8031014 | 15 | 33108637 | 3.0094 | 0.6769 | 8.76352e-06 |
| 52 | rs6558205 | 8 | 48244020 | rs11247118 | 15 | 97989207 | 1.5805 | 0.3539 | 7.96403e-06 |
| 53 | rs4551346 | 8 | 63565715 | rs7974366 | 12 | 12829195 | 1.6016 | 0.3553 | 6.54995e-06 |
| 54 | rs17085042 | 9 | 83578988 | rs7404578 | 16 | 17404445 | 1.8676 | 0.3975 | 2.618e-06 |
| 55 | rs4518734 | 9 | 123708019 | rs11831821 | 12 | 43763574 | 2.4895 | 0.5579 | 8.12116e-06 |
| 56 | rs1904024 | 10 | 53511343 | rs1002244 | 21 | 18477887 | 2.4747 | 0.5526 | 7.51381e-06 |
| 57 | rs11231918 | 11 | 64513102 | rs1336666 | 13 | 101849898 | 3.6445 | 0.7975 | 4.88521e-06 |
| 58 | rs10774812 | 12 | 114349795 | rs4902046 | 14 | 60859258 | -17.0918 | 3.8533 | 9.18007e-06 |
| 59 | rs10774812 | 12 | 114349795 | rs4899038 | 14 | 60859545 | -2.5990 | 0.5518 | 2.47932e-06 |
| 60 | rs17079354 | 13 | 23269225 | rs1629174 | 19 | 44578058 | 1.6854 | 0.3802 | 9.29416e-06 |
| 61 | rs7331256 | 13 | 32281103 | rs4899038 | 14 | 60859545 | -23.3283 | 5.0950 | 4.68074e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 62 | rs9595770 | 13 | 47219459 | rs191207 | 19 | 57493758 | 3.6303 | 0.7548 | 1.51306e-06 |
| 63 | rs338122 | 13 | 70141283 | rs10483479 | 14 | 36515241 | 1.4729 | 0.3260 | 6.24364e-06 |
| 64 | rs7184991 | 16 | 9048552 | rs2972588 | 19 | 8814922 | 1.8008 | 0.3740 | 1.47488e-06 |

Table 5.34: *Significant magenta Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0065 | GO:0097060 | GO:CC | synaptic membrane |
| 2 | 0.0106 | GO:0045211 | GO:CC | postsynaptic membrane |
| 3 | 0.0429 | GO:0099572 | GO:CC | postsynaptic specialization |

Table 5.35: *Significant gene set enrichment results of genes with SNPs found by magenta cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs12729332 | 1 | 98772342 | 0.0536 | 0.0118 | 5.644269e-06 | |
| 2 | rs11692400 | 2 | 216547176 | 0.0200 | 0.0042 | 1.484727e-06 | MREG |
| 3 | rs4680509 | 3 | 160634953 | 0.0155 | 0.0033 | 1.940174e-06 | IQCJ-SCHIP1 |
| 4 | rs17419291 | 5 | 87816188 | 0.0239 | 0.0044 | 4.218069e-08 | |
| 5 | rs6997672 | 8 | 69109036 | 0.0434 | 0.0093 | 2.823701e-06 | PREX2 |
| 6 | rs7855148 | 9 | 1506528 | 0.0340 | 0.0067 | 3.366948e-07 | |
| 7 | rs10975612 | 9 | 6498149 | 0.0211 | 0.0047 | 7.058319e-06 | |
| 8 | rs11813051 | 10 | 18783066 | 0.0551 | 0.0105 | 1.44058e-07 | CACNB2 |
| 9 | rs10827540 | 10 | 36116780 | 0.0255 | 0.0051 | 5.592256e-07 | |
| 10 | rs945494 | 10 | 36141815 | 0.0235 | 0.0046 | 2.593002e-07 | |
| 11 | rs4934770 | 10 | 36155566 | 0.0182 | 0.0037 | 1.181975e-06 | |
| 12 | rs7122974 | 11 | 45097320 | 0.0167 | 0.0037 | 8.066621e-06 | PRDM11 |
| 13 | rs9524352 | 13 | 93630098 | 0.0195 | 0.0044 | 9.549363e-06 | GPC6 |
| 14 | rs17235334 | 13 | 95072901 | 0.0175 | 0.0038 | 4.888065e-06 | DZIP1 |
| 15 | rs1560694 | 19 | 6218827 | -0.0108 | 0.0024 | 8.676698e-06 | MLLT1 |
| 16 | rs751327 | 20 | 62048057 | 0.0480 | 0.0098 | 9.406004e-07 | UCKL1 |

Table 5.36: *Significant purple GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs11122109 | 1 | 6610945 | rs9424522 | 1 | 230519655 | -1.4552 | 0.3280 | 9.14473e-06 | | |
| 2 | rs2235564 | 1 | 6635701 | rs8052587 | 16 | 24796094 | 17.1808 | 3.8246 | 7.04889e-06 | | |
| 3 | rs6659873 | 1 | 6646393 | rs8052587 | 16 | 24796094 | 17.1689 | 3.8213 | 7.02494e-06 | | |
| 4 | rs2789740 | 1 | 14542883 | rs1888560 | 14 | 20234085 | 1.6415 | 0.3641 | 6.52689e-06 | | |
| 5 | rs194637 | 1 | 33174360 | rs6413504 | 19 | 11102915 | 1.1736 | 0.2588 | 5.77179e-06 | | |
| 6 | rs194640 | 1 | 33176068 | rs6413504 | 19 | 11102915 | 1.1739 | 0.2588 | 5.73066e-06 | | |
| 7 | rs11161721 | 1 | 86260502 | rs8055371 | 16 | 25635344 | -1.7543 | 0.3802 | 3.93939e-06 | | |
| 8 | rs10799928 | 1 | 161761408 | rs10076979 | 5 | 145027619 | 17.5446 | 3.8895 | 6.45913e-06 | | |
| 9 | rs17026635 | 1 | 214618990 | rs1471744 | 5 | 154798692 | -23.9991 | 5.4202 | 9.52492e-06 | | |
| 10 | rs2618663 | 1 | 235777616 | rs10932037 | 2 | 204533591 | 2.0911 | 0.4698 | 8.55964e-06 | | |
| 11 | rs13405142 | 2 | 41409785 | rs6998543 | 8 | 82063399 | -16.7546 | 3.7633 | 8.50153e-06 | | |
| 12 | rs2874316 | 2 | 95173331 | rs2042484 | 2 | 208071307 | 1.9369 | 0.4156 | 3.16075e-06 | | |
| 13 | rs2946592 | 2 | 100594107 | rs882575 | 2 | 114734725 | -1.1089 | 0.2388 | 3.42507e-06 | | |
| 14 | rs10182570 | 2 | 165817880 | rs1445728 | 5 | 13260997 | 1.3874 | 0.2833 | 9.74896e-07 | | |
| 15 | rs10174400 | 2 | 165833465 | rs1445728 | 5 | 13260997 | 1.4200 | 0.2824 | 4.95218e-07 | | |
| 16 | rs1011584 | 2 | 190606463 | rs4146282 | 3 | 109979738 | 3.0712 | 0.6162 | 6.24003e-07 | | |
| 17 | rs6759008 | 2 | 216677487 | rs4715247 | 6 | 51791833 | 5.7512 | 1.2063 | 1.86457e-06 | | |
| 18 | rs7579078 | 2 | 230935683 | rs754970 | 11 | 45008403 | 1.1193 | 0.2519 | 8.84584e-06 | | |
| 19 | rs6803088 | 3 | 2643594 | rs1837763 | 15 | 69494429 | -2.8238 | 0.6251 | 6.25991e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs807193 | 3 | 53591378 | rs12783710 | 10 | 82396254 | -17.4838 | 3.7611 | 3.34217e-06 |
| 21 | rs1492001 | 3 | 54981222 | rs2347785 | 7 | 47294909 | -1.1395 | 0.2540 | 7.24856e-06 |
| 22 | rs1368515 | 3 | 97031729 | rs2052162 | 7 | 77009553 | 1.4506 | 0.3127 | 3.51203e-06 |
| 23 | rs1368515 | 3 | 97031729 | rs4729594 | 7 | 77136688 | 1.4535 | 0.3131 | 3.43792e-06 |
| 24 | rs10804512 | 3 | 97040061 | rs2052162 | 7 | 77009553 | 1.4693 | 0.3124 | 2.55907e-06 |
| 25 | rs10804512 | 3 | 97040061 | rs4729594 | 7 | 77136688 | 1.4728 | 0.3127 | 2.48231e-06 |
| 26 | rs13101933 | 4 | 163369235 | rs2757117 | 14 | 58764396 | -17.4512 | 3.8621 | 6.226e-06 |
| 27 | rs2067587 | 5 | 13446197 | rs744654 | 14 | 94438034 | 2.1072 | 0.4224 | 6.08319e-07 |
| 28 | rs6885207 | 5 | 76180991 | rs17205951 | 5 | 82884080 | -1.7955 | 0.4023 | 8.06019e-06 |
| 29 | rs163129 | 5 | 78315576 | rs2414049 | 15 | 48453732 | 1.3551 | 0.2997 | 6.14625e-06 |
| 30 | rs7720835 | 5 | 117269389 | rs972894 | 15 | 35751832 | 1.2737 | 0.2862 | 8.60466e-06 |
| 31 | rs13196352 | 6 | 49621726 | rs11064111 | 12 | 6232623 | 2.1282 | 0.4669 | 5.16984e-06 |
| 32 | rs3779837 | 8 | 2036057 | rs2032357 | 18 | 20373239 | -2.5278 | 0.5455 | 3.58927e-06 |
| 33 | rs11780116 | 8 | 73957810 | rs5750310 | 22 | 35538606 | 1.4139 | 0.3195 | 9.59883e-06 |
| 34 | rs7006008 | 8 | 129676131 | rs3830068 | 17 | 77233304 | 1.4650 | 0.3215 | 5.18877e-06 |
| 35 | rs7017613 | 8 | 139309627 | rs2304144 | 19 | 57197045 | -17.0745 | 3.5660 | 1.68287e-06 |
| 36 | rs12251539 | 10 | 12405881 | rs17105725 | 10 | 87359244 | 1.8001 | 0.4016 | 7.36376e-06 |
| 37 | rs11258865 | 10 | 14186438 | rs9612028 | 22 | 41953966 | -24.4211 | 5.4523 | 7.49827e-06 |
| 38 | rs10732804 | 10 | 112785011 | rs4130618 | 11 | 91176411 | -2.2566 | 0.4395 | 2.8355e-07 |
| 39 | rs7112451 | 11 | 37182142 | rs531860 | 11 | 115766886 | 18.5902 | 4.1010 | 5.81367e-06 |
| 40 | rs1483397 | 11 | 83314550 | rs837127 | 15 | 35031478 | -1.5133 | 0.3364 | 6.82307e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs1483397 | 11 | 83314550 | rs844380 | 15 | 35032208 | -1.4482 | 0.3253 | 8.52616e-06 |
| 42 | rs1450998 | 12 | 96550445 | rs12904629 | 15 | 64274276 | 1.2851 | 0.2753 | 3.05149e-06 |
| 43 | rs7154375 | 14 | 22996199 | rs2581641 | 18 | 71119261 | 1.1984 | 0.2604 | 4.17771e-06 |
| 44 | rs8014874 | 14 | 23008418 | rs2581641 | 18 | 71119261 | 1.2142 | 0.2613 | 3.37157e-06 |
| 45 | rs1028589 | 14 | 23010199 | rs2581641 | 18 | 71119261 | 1.2642 | 0.2658 | 1.9749e-06 |

Table 5.37: *Significant purple Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0497 | CORUM:6524 | CORUM | DZIP1-GLI3 complex |
| 2 | 0.0497 | CORUM:6531 | CORUM | CEP164-DZIP1 complex |
| 3 | 0.0497 | CORUM:6533 | CORUM | DZIP1-IFT88 complex |

Table 5.38: *Significant gene set enrichment results of genes with SNPs found by purple cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs2379146 | 1 | 11405504 | 0.0438 | 0.0081 | 7.356094e-08 | |
| 2 | rs4526656 | 1 | 77055155 | 0.0350 | 0.0076 | 3.886543e-06 | |
| 3 | rs7512513 | 1 | 145553128 | 0.0158 | 0.0029 | 3.118e-08 | BCL9 |
| 4 | rs946903 | 1 | 145563890 | 0.0151 | 0.0029 | 1.589736e-07 | BCL9 |
| 5 | rs885239 | 1 | 145594226 | 0.0140 | 0.0031 | 4.808661e-06 | ACP6 |
| 6 | rs7550430 | 1 | 179427776 | 0.0675 | 0.0106 | 2.251652e-10 | |
| 7 | rs16846590 | 1 | 225131343 | 0.0519 | 0.0107 | 1.354884e-06 | PSEN2 |
| 8 | rs971121 | 2 | 76814935 | 0.0111 | 0.0025 | 7.744659e-06 | |
| 9 | rs10202684 | 2 | 125259581 | 0.0229 | 0.0050 | 4.007289e-06 | CNTNAP5 |
| 10 | rs1568519 | 2 | 212714637 | 0.0216 | 0.0047 | 4.739713e-06 | ERBB4 |
| 11 | rs9993404 | 4 | 25440137 | 0.0400 | 0.0090 | 8.982442e-06 | SEL1L3 |
| 12 | rs17015461 | 4 | 90491000 | 0.0503 | 0.0113 | 8.450987e-06 | |
| 13 | rs11562926 | 4 | 118869490 | 0.0115 | 0.0026 | 9.858816e-06 | |
| 14 | rs152615 | 5 | 14645979 | 0.0529 | 0.0103 | 3.11351e-07 | OTULINL |
| 15 | rs10053502 | 5 | 40014929 | 0.0183 | 0.0033 | 2.031656e-08 | |
| 16 | rs10475791 | 5 | 165980465 | 0.0453 | 0.0101 | 8.327628e-06 | |
| 17 | rs11964520 | 6 | 20199553 | 0.0583 | 0.0095 | 9.02915e-10 | |
| 18 | rs17139557 | 7 | 18677723 | 0.0519 | 0.0115 | 6.341957e-06 | HDAC9 |
| 19 | rs4840528 | 8 | 10872062 | 0.0245 | 0.0048 | 2.970369e-07 | XKR6 |
| 20 | rs6980533 | 8 | 20781438 | 0.0463 | 0.0095 | 1.223912e-06 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 | rs10092318 | 8 | 21414363 | 0.0592 | 0.0106 | 2.312417e-08 | |
| 22 | rs16907174 | 8 | 138517940 | 0.0459 | 0.0102 | 6.733312e-06 | |
| 23 | rs472936 | 9 | 112669312 | 0.0267 | 0.0060 | 7.607175e-06 | |
| 24 | rs17011340 | 10 | 49756248 | 0.0544 | 0.0108 | 5.468037e-07 | WDFY4 |
| 25 | rs7100357 | 10 | 94451324 | 0.0537 | 0.0107 | 6.132173e-07 | |
| 26 | rs6585325 | 10 | 117053520 | 0.0517 | 0.0114 | 5.665307e-06 | ATRNL1 |
| 27 | rs10490912 | 10 | 120136330 | 0.0558 | 0.0118 | 2.407318e-06 | |
| 28 | rs11198385 | 10 | 120143942 | 0.0629 | 0.0103 | 1.135182e-09 | |
| 29 | rs11818150 | 10 | 121060687 | 0.0350 | 0.0072 | 1.278625e-06 | GRK5 |
| 30 | rs7069895 | 10 | 127903886 | 0.0516 | 0.0114 | 5.79542e-06 | ADAM12 |
| 31 | rs12422109 | 11 | 3591662 | 0.0358 | 0.0080 | 6.868221e-06 | TRPC2 |
| 32 | rs1036864 | 11 | 27177824 | 0.0544 | 0.0108 | 5.474938e-07 | |
| 33 | rs12788404 | 11 | 95567860 | -0.0123 | 0.0027 | 5.952731e-06 | MAML2 |
| 34 | rs2920027 | 12 | 67241298 | 0.0475 | 0.0073 | 1.011208e-10 | |
| 35 | rs11609868 | 12 | 126602165 | 0.0171 | 0.0036 | 1.825437e-06 | |
| 36 | rs7308784 | 12 | 126814588 | 0.0483 | 0.0103 | 2.656139e-06 | |
| 37 | rs17633078 | 13 | 29728634 | 0.0202 | 0.0045 | 8.412617e-06 | KATNAL1 |
| 38 | rs4884871 | 13 | 69490705 | 0.0125 | 0.0027 | 5.162202e-06 | KLHL1 |
| 39 | rs2019845 | 13 | 81866894 | 0.0212 | 0.0046 | 4.072522e-06 | |
| 40 | rs10130493 | 14 | 30050171 | 0.0466 | 0.0102 | 5.039719e-06 | |
| 41 | rs3865365 | 18 | 10515480 | 0.0499 | 0.0105 | 1.944334e-06 | |
| 42 | rs171479 | 21 | 28078751 | 0.0451 | 0.0086 | 1.91667e-07 | |

Table 5.39: *Significant greenyellow GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs3828051 | 1 | 30970386 | rs7815682 | 8 | 18623918 | 1.9864 | 0.4263 | 3.16241e-06 | | |
| 2 | rs6425939 | 1 | 35401376 | rs4618582 | 7 | 66546427 | 1.2207 | 0.2682 | 5.31044e-06 | | |
| 3 | rs4348764 | 1 | 68610398 | rs6743106 | 2 | 68595464 | 5.6031 | 1.2236 | 4.67165e-06 | | |
| 4 | rs11210357 | 1 | 74106296 | rs4074617 | 6 | 3858629 | -17.7804 | 3.7574 | 2.22276e-06 | | |
| 5 | rs3916208 | 1 | 95033472 | rs10440995 | 7 | 11198252 | -17.0745 | 3.8581 | 9.61453e-06 | | |
| 6 | rs2810418 | 1 | 100416498 | rs871429 | 11 | 36183135 | 1.4439 | 0.3256 | 9.23334e-06 | | |
| 7 | rs946818 | 1 | 176264420 | rs3007061 | 14 | 46308356 | -1.9142 | 0.4306 | 8.77885e-06 | | |
| 8 | rs11799643 | 1 | 207946222 | rs1988083 | 17 | 50146705 | 2.7173 | 0.5796 | 2.75047e-06 | | |
| 9 | rs1366990 | 1 | 234948326 | rs10254035 | 7 | 70039082 | 17.5660 | 3.7695 | 3.16156e-06 | | |
| 10 | rs4390163 | 1 | 237562512 | rs3120890 | 13 | 54609561 | 1.3305 | 0.2957 | 6.81754e-06 | | |
| 11 | rs4390163 | 1 | 237562512 | rs1582147 | 13 | 54628925 | 1.3311 | 0.2957 | 6.73788e-06 | | |
| 12 | rs4390163 | 1 | 237562512 | rs641385 | 13 | 54771026 | 1.3187 | 0.2948 | 7.70887e-06 | | |
| 13 | rs11903474 | 2 | 52100230 | rs7216300 | 17 | 56023823 | 3.8464 | 0.8557 | 6.95987e-06 | | |
| 14 | rs6743106 | 2 | 68595464 | rs327832 | 5 | 125200482 | -16.8887 | 3.7043 | 5.13326e-06 | | |
| 15 | rs6743106 | 2 | 68595464 | rs4149313 | 9 | 106626574 | 17.1767 | 3.5725 | 1.52431e-06 | | |
| 16 | rs7606395 | 2 | 76734718 | rs2830976 | 21 | 27782693 | 2.0499 | 0.4204 | 1.08491e-06 | | |
| 17 | rs17024323 | 2 | 84043438 | rs3003202 | 13 | 69469515 | -2.9748 | 0.6155 | 1.34288e-06 | | |
| 18 | rs17024323 | 2 | 84043438 | rs2472285 | 13 | 69478105 | -17.7830 | 3.8921 | 4.90148e-06 | | |
| 19 | rs6733249 | 2 | 98093335 | rs166353 | 4 | 26222553 | -17.2881 | 3.9066 | 9.62875e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs6743113 | 2 | 175860968 | rs472936 | 9 | 112669312 | -17.5115 | 3.7326 | 2.71132e-06 |
| 21 | rs4894203 | 2 | 177418213 | rs7047329 | 9 | 132229317 | -17.1244 | 3.8295 | 7.76114e-06 |
| 22 | rs10497567 | 2 | 181111709 | rs10814814 | 9 | 4009700 | 1.7146 | 0.3851 | 8.51567e-06 |
| 23 | rs2372757 | 3 | 81305242 | rs2631864 | 8 | 21156364 | 1.2811 | 0.2878 | 8.50915e-06 |
| 24 | rs10470524 | 3 | 185251030 | rs1510973 | 5 | 118117697 | -1.4302 | 0.3166 | 6.25099e-06 |
| 25 | rs12233824 | 4 | 7784743 | rs9380934 | 6 | 12337924 | -1.2382 | 0.2555 | 1.25781e-06 |
| 26 | rs6833151 | 4 | 24343929 | rs8103801 | 19 | 14546934 | 5.0580 | 1.1361 | 8.50016e-06 |
| 27 | rs6833151 | 4 | 24343929 | rs10415055 | 19 | 14549725 | 18.8564 | 4.1985 | 7.08313e-06 |
| 28 | rs4861074 | 4 | 40648464 | rs2836889 | 21 | 39407444 | -1.6516 | 0.3673 | 6.88788e-06 |
| 29 | rs2412494 | 4 | 54327916 | rs17760491 | 13 | 95965153 | 1.8601 | 0.4210 | 9.96685e-06 |
| 30 | rs355679 | 4 | 78697530 | rs28735138 | 9 | 139393985 | 18.3068 | 4.1044 | 8.18514e-06 |
| 31 | rs345328 | 4 | 86982925 | rs4806798 | 19 | 59858533 | -2.2048 | 0.4927 | 7.65065e-06 |
| 32 | rs1553142 | 4 | 129806899 | rs9603226 | 13 | 37041586 | -17.1104 | 3.8619 | 9.3977e-06 |
| 33 | rs11133706 | 5 | 12363717 | rs12198596 | 6 | 103902072 | 1.8297 | 0.3731 | 9.36843e-07 |
| 34 | rs1966938 | 5 | 93608749 | rs497264 | 19 | 60801728 | -1.8487 | 0.3731 | 7.23127e-07 |
| 35 | rs9393069 | 6 | 8800947 | rs1880535 | 8 | 69434121 | 1.5351 | 0.3341 | 4.32425e-06 |
| 36 | rs9262539 | 6 | 31098469 | rs9305467 | 21 | 31971660 | -17.4354 | 3.4843 | 5.61443e-07 |
| 37 | rs9471115 | 6 | 39544026 | rs3865365 | 18 | 10515480 | 19.2794 | 4.3559 | 9.59812e-06 |
| 38 | rs593493 | 6 | 65114770 | rs11630814 | 15 | 88439163 | -1.1026 | 0.2421 | 5.26003e-06 |
| 39 | rs10440995 | 7 | 11198252 | rs6948941 | 7 | 24368645 | -3.4814 | 0.7354 | 2.20195e-06 |
| 40 | rs10440995 | 7 | 11198252 | rs6980533 | 8 | 20781438 | -18.0562 | 2.9506 | 9.38654e-10 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs10440995 | 7 | 11198252 | rs10092318 | 8 | 21414363 | -2.6757 | 0.5829 | 4.42812e-06 |
| 42 | rs10440995 | 7 | 11198252 | rs575764 | 11 | 22441652 | -2.9788 | 0.6436 | 3.68629e-06 |
| 43 | rs10440995 | 7 | 11198252 | rs1367860 | 13 | 46541160 | -3.5470 | 0.7908 | 7.2735e-06 |
| 44 | rs10440995 | 7 | 11198252 | rs16991956 | 20 | 44918454 | -17.6237 | 3.5238 | 5.69165e-07 |
| 45 | rs221134 | 7 | 28943536 | rs13045110 | 20 | 48270025 | -1.8694 | 0.4207 | 8.86395e-06 |
| 46 | rs221151 | 7 | 28958387 | rs13045110 | 20 | 48270025 | -2.3065 | 0.5185 | 8.6449e-06 |
| 47 | rs10112584 | 8 | 10316913 | rs7828503 | 8 | 27385398 | 3.9735 | 0.8914 | 8.28422e-06 |
| 48 | rs2607058 | 8 | 95296944 | rs7831168 | 8 | 138689939 | -1.2499 | 0.2760 | 5.91704e-06 |
| 49 | rs4506215 | 8 | 123591503 | rs763512 | 17 | 32963644 | -1.1925 | 0.2673 | 8.16839e-06 |
| 50 | rs10814814 | 9 | 4009700 | rs8063474 | 16 | 53100356 | -1.3832 | 0.2958 | 2.92979e-06 |
| 51 | rs16938162 | 9 | 20753840 | rs7349 | 10 | 31857911 | 1.2521 | 0.2814 | 8.64043e-06 |
| 52 | rs11144075 | 9 | 76529111 | rs3095821 | 15 | 55363208 | 1.7919 | 0.4026 | 8.53269e-06 |
| 53 | rs11008356 | 10 | 31465689 | rs2669017 | 12 | 75564882 | -1.9340 | 0.4367 | 9.49989e-06 |
| 54 | rs10128422 | 10 | 130837828 | rs2198517 | 17 | 51367182 | -3.2100 | 0.7016 | 4.74958e-06 |
| 55 | rs2669017 | 12 | 75564882 | rs9597698 | 13 | 57738510 | 1.1389 | 0.2490 | 4.78758e-06 |
| 56 | rs3120890 | 13 | 54609561 | rs1951187 | 14 | 31977943 | 1.2457 | 0.2563 | 1.17078e-06 |
| 57 | rs3120890 | 13 | 54609561 | rs941749 | 14 | 31998385 | 1.3059 | 0.2661 | 9.19626e-07 |
| 58 | rs1582147 | 13 | 54628925 | rs1951187 | 14 | 31977943 | 1.2464 | 0.2563 | 1.15095e-06 |
| 59 | rs1582147 | 13 | 54628925 | rs941749 | 14 | 31998385 | 1.3067 | 0.2660 | 9.03505e-07 |
| 60 | rs641385 | 13 | 54771026 | rs1951187 | 14 | 31977943 | 1.1726 | 0.2548 | 4.19486e-06 |
| 61 | rs641385 | 13 | 54771026 | rs941749 | 14 | 31998385 | 1.3008 | 0.2654 | 9.52124e-07 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 62 | rs12886510 | 14 | 40666305 | rs2615258 | 15 | 42291407 | -1.2745 | 0.2796 | 5.16825e-06 |
| 63 | rs12915222 | 15 | 21724598 | rs6564427 | 16 | 75924089 | 1.4652 | 0.3109 | 2.44552e-06 |

Table 5.40: *Significant greenyellow Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0224 | KEGG:04330 | KEGG | Notch signaling pathway |
| 2 | 0.0011 | REAC:R-HSA-2894862 | REAC | Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants |
| 3 | 0.0011 | REAC:R-HSA-2644606 | REAC | Constitutive Signaling by NOTCH1 PEST Domain Mutants |
| 4 | 0.0011 | REAC:R-HSA-2894858 | REAC | Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer |
| 5 | 0.0011 | REAC:R-HSA-2644602 | REAC | Signaling by NOTCH1 PEST Domain Mutants in Cancer |
| 6 | 0.0011 | REAC:R-HSA-2644603 | REAC | Signaling by NOTCH1 in Cancer |
| 7 | 0.0021 | REAC:R-HSA-1980143 | REAC | Signaling by NOTCH1 |
| 8 | 0.0300 | REAC:R-HSA-350054 | REAC | Notch-HLH transcription pathway |
| 9 | 0.0419 | REAC:R-HSA-1980145 | REAC | Signaling by NOTCH2 |
| 10 | 0.0251 | WP:WP3845 | WP | Canonical and Non-canonical Notch signaling |
| 11 | 0.0310 | WP:WP4905 | WP | 1q21.1 copy number variation syndrome |

Table 5.41: *Significant gene set enrichment results of genes with SNPs found by greenyellow cluster analysis. P-values are analytically adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs10917804 | 1 | 161825373 | 0.0183 | 0.0038 | 1.296552e-06 | |
| 2 | rs12986863 | 2 | 42986051 | 0.0105 | 0.0023 | 4.623089e-06 | |
| 3 | rs4851384 | 2 | 100928070 | 0.0180 | 0.0037 | 1.356349e-06 | NPAS2 |
| 4 | rs2657606 | 3 | 16151302 | 0.0108 | 0.0024 | 4.735653e-06 | |
| 5 | rs6766801 | 3 | 64476738 | 0.0530 | 0.0104 | 3.518725e-07 | ADAMTS9 |
| 6 | rs17070965 | 3 | 64502168 | 0.0560 | 0.0106 | 1.421178e-07 | ADAMTS9 |
| 7 | rs12644729 | 4 | 60612082 | 0.0410 | 0.0066 | 4.792338e-10 | |
| 8 | rs10009793 | 4 | 60660113 | 0.0329 | 0.0063 | 2.010529e-07 | |
| 9 | rs2328662 | 6 | 22149112 | 0.0138 | 0.0030 | 5.633088e-06 | |
| 10 | rs12706405 | 7 | 121746982 | 0.0114 | 0.0025 | 4.280018e-06 | CADPS2 |
| 11 | rs4871014 | 8 | 128340162 | 0.0112 | 0.0023 | 9.236507e-07 | |
| 12 | rs11599366 | 10 | 3762249 | 0.0135 | 0.0029 | 4.094138e-06 | |
| 13 | rs210271 | 10 | 42318996 | 0.0095 | 0.0021 | 9.285532e-06 | |
| 14 | rs11101535 | 10 | 49734437 | 0.0126 | 0.0027 | 2.720994e-06 | WDFY4 |
| 15 | rs7094610 | 10 | 49792187 | 0.0102 | 0.0022 | 6.452029e-06 | LRRC18 |
| 16 | rs11195970 | 10 | 114222079 | 0.0179 | 0.0037 | 1.84804e-06 | VTI1A |
| 17 | rs3803174 | 12 | 43743585 | 0.0345 | 0.0077 | 7.346486e-06 | |
| 18 | rs11499034 | 14 | 81042194 | 0.0539 | 0.0086 | 4.685786e-10 | SEL1L |
| 19 | rs2706600 | 18 | 61913993 | 0.0225 | 0.0041 | 2.952041e-08 | |
| 20 | rs7233624 | 18 | 61946597 | 0.0235 | 0.0052 | 5.641393e-06 | |
| 21 | rs1388176 | 18 | 61955870 | 0.0225 | 0.0050 | 5.650304e-06 | |
| 22 | rs6034804 | 20 | 17298685 | 0.0201 | 0.0042 | 2.022442e-06 | PCSK2 |

Table 5.42: *Significant tan GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs12563394 | 1 | 7303334 | rs10429632 | 9 | 134316627 | -1.8560 | 0.4033 | 4.19639e-06 | | |
| 2 | rs919055 | 1 | 60235624 | rs12039989 | 1 | 231564859 | -2.3579 | 0.5138 | 4.46025e-06 | | |
| 3 | rs4287129 | 1 | 74788735 | rs9569913 | 13 | 33104418 | 3.4697 | 0.7781 | 8.2328e-06 | | |
| 4 | rs10917799 | 1 | 161792101 | rs10501429 | 11 | 77957438 | 1.3856 | 0.2775 | 5.92581e-07 | | |
| 5 | rs1934533 | 1 | 161810065 | rs10501429 | 11 | 77957438 | 1.2591 | 0.2699 | 3.0773e-06 | | |
| 6 | rs11803436 | 1 | 180306065 | rs7355673 | 2 | 213112634 | 1.3042 | 0.2932 | 8.67048e-06 | | |
| 7 | rs7520347 | 1 | 239507041 | rs9358493 | 6 | 22269883 | 1.6514 | 0.3242 | 3.51577e-07 | | |
| 8 | rs28477715 | 1 | 240207464 | rs16824173 | 2 | 228468580 | -21.8462 | 4.9293 | 9.33961e-06 | | |
| 9 | rs11681404 | 2 | 106102745 | rs10501745 | 11 | 90281003 | -2.5166 | 0.5234 | 1.52626e-06 | | |
| 10 | rs13315983 | 3 | 2905995 | rs9376393 | 6 | 139367404 | -2.2107 | 0.5005 | 9.9945e-06 | | |
| 11 | rs11708252 | 3 | 63483860 | rs3812689 | 10 | 70725502 | 2.2041 | 0.4911 | 7.17843e-06 | | |
| 12 | rs17045666 | 3 | 67097455 | rs9559900 | 13 | 110325330 | 17.5897 | 3.8789 | 5.76743e-06 | | |
| 13 | rs17048091 | 3 | 69008468 | rs237733 | 20 | 47375387 | 3.7309 | 0.8230 | 5.80689e-06 | | |
| 14 | rs6794608 | 3 | 150665245 | rs17106990 | 14 | 37268283 | 1.6133 | 0.3568 | 6.14495e-06 | | |
| 15 | rs9289788 | 3 | 150690624 | rs17106990 | 14 | 37268283 | 1.9302 | 0.4156 | 3.41595e-06 | | |
| 16 | rs17632548 | 3 | 178765503 | rs2221894 | 8 | 28847078 | -22.0786 | 4.9580 | 8.46258e-06 | | |
| 17 | rs10008679 | 4 | 23643806 | rs1388176 | 18 | 61955870 | 2.0707 | 0.4610 | 7.04983e-06 | | |
| 18 | rs12647295 | 4 | 59176173 | rs9563614 | 13 | 33087333 | 3.2846 | 0.7219 | 5.3626e-06 | | |
| 19 | rs6849677 | 4 | 84421707 | rs500951 | 15 | 23924879 | 2.6582 | 0.5845 | 5.42465e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs9307824 | 4 | 84439636 | rs500951 | 15 | 23924879 | 2.6060 | 0.5897 | 9.91743e-06 |
| 21 | rs313949 | 4 | 114240542 | rs1829127 | 12 | 72715377 | -17.1197 | 3.7726 | 5.68123e-06 |
| 22 | rs4241923 | 4 | 140062058 | rs17799219 | 7 | 51105350 | -1.9204 | 0.4241 | 5.95896e-06 |
| 23 | rs12502016 | 4 | 158569059 | rs1944790 | 11 | 96366083 | -2.0572 | 0.4484 | 4.4871e-06 |
| 24 | rs1865353 | 4 | 158579217 | rs1944790 | 11 | 96366083 | -2.0669 | 0.4481 | 3.9787e-06 |
| 25 | rs6536237 | 4 | 158583138 | rs1944790 | 11 | 96366083 | -1.9879 | 0.4491 | 9.57797e-06 |
| 26 | rs1049344 | 5 | 10733487 | rs10099164 | 8 | 126458440 | -1.9717 | 0.4318 | 4.976e-06 |
| 27 | rs3822414 | 5 | 10755688 | rs10099164 | 8 | 126458440 | -1.9779 | 0.4286 | 3.94038e-06 |
| 28 | rs7725550 | 5 | 11427223 | rs1829127 | 12 | 72715377 | -2.0408 | 0.4458 | 4.70738e-06 |
| 29 | rs7702488 | 5 | 11431282 | rs1829127 | 12 | 72715377 | -2.0201 | 0.4148 | 1.11439e-06 |
| 30 | rs13167730 | 5 | 79406003 | rs7298766 | 12 | 531917 | 1.7179 | 0.3786 | 5.69543e-06 |
| 31 | rs17341747 | 5 | 103598057 | rs7771953 | 6 | 27379322 | -17.1239 | 3.6265 | 2.33622e-06 |
| 32 | rs4720028 | 7 | 31124806 | rs2124154 | 8 | 2631200 | 2.8986 | 0.6538 | 9.28418e-06 |
| 33 | rs10246618 | 7 | 53250628 | rs17119756 | 14 | 83704185 | 1.5370 | 0.3411 | 6.61509e-06 |
| 34 | rs664112 | 7 | 70786039 | rs532849 | 18 | 42526088 | -2.5354 | 0.5253 | 1.38897e-06 |
| 35 | rs9641363 | 7 | 105855399 | rs11779546 | 8 | 26730853 | 1.2893 | 0.2876 | 7.36106e-06 |
| 36 | rs4726342 | 7 | 140501563 | rs5027573 | 10 | 87881252 | 2.1600 | 0.4877 | 9.48129e-06 |
| 37 | rs4726342 | 7 | 140501563 | rs9323455 | 14 | 64397041 | 2.9386 | 0.5834 | 4.7309e-07 |
| 38 | rs10112737 | 8 | 126354508 | rs7961656 | 12 | 103546283 | -18.5535 | 3.4055 | 5.09074e-08 |
| 39 | rs17156937 | 10 | 16847751 | rs7123583 | 11 | 116105231 | 2.3735 | 0.5303 | 7.61678e-06 |
| 40 | rs1585962 | 11 | 38621133 | rs12597 | 12 | 92320556 | -1.3848 | 0.3134 | 9.92898e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs10773558 | 12 | 127643363 | rs11626215 | 14 | 94365382 | -1.2226 | 0.2538 | 1.44995e-06 |
| 42 | rs4784805 | 16 | 55998908 | rs4808983 | 19 | 19743476 | 2.5561 | 0.5589 | 4.80444e-06 |

Table 5.43: *Significant tan Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0069 | CORUM:876 | CORUM | SNARE complex (VAMP3, STX6, VTI1A) |
| 2 | 0.0069 | CORUM:877 | CORUM | SNARE complex (VAMP4, STX6, STX16, VTI1a, VTI1b) |
| 3 | 0.0069 | CORUM:4999 | CORUM | VCP-VIMP-DERL1-DERL2-HRD1-SEL1L complex |
| 4 | 0.0069 | CORUM:6886 | CORUM | AUP1-OS9-SEL1L-UBC6e-UBXD8 complex |
| 5 | 0.0069 | CORUM:7271 | CORUM | ARNTL-NPAS2 complex |
| 6 | 0.0125 | CORUM:6859 | CORUM | HRD1 complex |
| 7 | 0.0255 | KEGG:04130 | KEGG | SNARE interactions in vesicular transport |
| 8 | 0.0255 | KEGG:04710 | KEGG | Circadian rhythm |

Table 5.44: *Significant gene set enrichmet results of genes with SNPs found by tan cluster analysis. P-values are FDR adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6577570 | 1 | 6309863 | 0.0262 | 0.0059 | 8.77777e-06 | ACOT7 |
| 2 | rs12118066 | 1 | 29958054 | 0.0329 | 0.0068 | 1.476096e-06 | |
| 3 | rs941124 | 1 | 53826672 | 0.0281 | 0.0056 | 6.24286e-07 | GLIS1 |
| 4 | rs2218404 | 1 | 107754128 | 0.0316 | 0.0067 | 2.162715e-06 | NTNG1 |
| 5 | rs558009 | 1 | 107758362 | 0.0343 | 0.0062 | 3.876878e-08 | NTNG1 |
| 6 | rs10918226 | 1 | 163884596 | 0.0156 | 0.0033 | 3.451386e-06 | MGST3 |
| 7 | rs9332628 | 1 | 167764355 | 0.0221 | 0.0046 | 2.035565e-06 | F5 |
| 8 | rs2494302 | 1 | 201511837 | 0.0154 | 0.0035 | 9.629035e-06 | |
| 9 | rs2494279 | 1 | 201522441 | 0.0174 | 0.0036 | 1.875029e-06 | |
| 10 | rs6734660 | 2 | 8010477 | 0.0382 | 0.0085 | 6.980981e-06 | |
| 11 | rs2287105 | 2 | 36671811 | 0.0332 | 0.0070 | 2.041979e-06 | FEZ2 |
| 12 | rs3792223 | 2 | 68945648 | 0.0220 | 0.0047 | 3.217716e-06 | BMP10 |
| 13 | rs13428497 | 2 | 129693628 | 0.0457 | 0.0083 | 3.784852e-08 | |
| 14 | rs10432509 | 2 | 192094767 | 0.0149 | 0.0032 | 2.876863e-06 | |
| 15 | rs2203748 | 2 | 192105436 | 0.0165 | 0.0035 | 2.11851e-06 | |
| 16 | rs284552 | 2 | 217021289 | 0.0281 | 0.0061 | 5.084279e-06 | SMARCAL1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 17 | rs9851625 | 3 | 107615782 | 0.0548 | 0.0091 | 1.623243e-09 | |
| 18 | rs2715694 | 3 | 110023833 | 0.0147 | 0.0033 | 7.019351e-06 | |
| 19 | rs6795648 | 3 | 124571756 | 0.0292 | 0.0063 | 3.379739e-06 | ADCY5 |
| 20 | rs6780511 | 3 | 162969359 | 0.0412 | 0.0092 | 7.297414e-06 | |
| 21 | rs16861986 | 3 | 188557190 | 0.0337 | 0.0073 | 3.735657e-06 | |
| 22 | rs2124075 | 3 | 192370979 | 0.0423 | 0.0086 | 8.281619e-07 | |
| 23 | rs1836702 | 4 | 30022467 | 0.0396 | 0.0077 | 2.934216e-07 | |
| 24 | rs7673475 | 4 | 83215207 | 0.0458 | 0.0102 | 7.701697e-06 | |
| 25 | rs11936273 | 4 | 88554532 | 0.0470 | 0.0080 | 3.9524e-09 | |
| 26 | rs16997168 | 4 | 111848488 | 0.0124 | 0.0028 | 7.566087e-06 | |
| 27 | rs2972781 | 5 | 42752083 | 0.0293 | 0.0054 | 6.355426e-08 | GHR |
| 28 | rs37540 | 5 | 53213804 | 0.0252 | 0.0050 | 4.248416e-07 | |
| 29 | rs965795 | 5 | 57980586 | 0.0095 | 0.0021 | 9.218401e-06 | RAB3C |
| 30 | rs11953502 | 5 | 160972260 | 0.0487 | 0.0105 | 3.541367e-06 | GABRA6 |
| 31 | rs307835 | 5 | 179966211 | 0.0384 | 0.0073 | 1.265326e-07 | FLT4 |
| 32 | rs12663510 | 6 | 42436047 | 0.0236 | 0.0052 | 6.737154e-06 | TRERF1 |
| 33 | rs265387 | 6 | 129606637 | 0.0107 | 0.0024 | 9.173991e-06 | LAMA2 |
| 34 | rs6977730 | 7 | 28100505 | 0.0207 | 0.0042 | 7.567479e-07 | JAZF1 |
| 35 | rs2435279 | 7 | 84893553 | 0.0194 | 0.0041 | 1.833612e-06 | |
| 36 | rs2463664 | 7 | 84917896 | 0.0180 | 0.0039 | 3.865175e-06 | |
| 37 | rs7003666 | 8 | 18107049 | 0.0247 | 0.0056 | 9.809486e-06 | NAT1 |
| 38 | rs2513764 | 8 | 95176653 | 0.0201 | 0.0039 | 2.223534e-07 | |
| 39 | rs2382962 | 8 | 144740535 | 0.0158 | 0.0035 | 6.355282e-06 | EEF1D |
| 40 | rs16925649 | 9 | 108403551 | 0.0286 | 0.0050 | 1.393939e-08 | |
| 41 | rs17154797 | 10 | 14267295 | 0.0357 | 0.0080 | 7.40503e-06 | FRMD4A |
| 42 | rs12253961 | 10 | 24307122 | 0.0518 | 0.0102 | 3.915021e-07 | KIAA1217 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 43 | rs4935474 | 10 | 55292241 | 0.0276 | 0.0057 | 1.100928e-06 | PCDH15 |
| 44 | rs2886405 | 10 | 127142934 | 0.0110 | 0.0022 | 8.611043e-07 | |
| 45 | rs6488271 | 12 | 10241957 | 0.0152 | 0.0034 | 6.357687e-06 | |
| 46 | rs1148980 | 12 | 70252605 | 0.0463 | 0.0104 | 8.239239e-06 | LGR5 |
| 47 | rs17032130 | 12 | 100855303 | 0.0354 | 0.0080 | 9.626446e-06 | DRAM1 |
| 48 | rs2257147 | 13 | 37421512 | 0.0308 | 0.0068 | 6.166626e-06 | |
| 49 | rs4148511 | 13 | 94589810 | 0.0335 | 0.0074 | 6.988155e-06 | ABCC4 |
| 50 | rs2619313 | 13 | 94589862 | 0.0354 | 0.0076 | 3.328072e-06 | ABCC4 |
| 51 | rs16964614 | 13 | 103774966 | 0.0412 | 0.0091 | 6.389575e-06 | |
| 52 | rs1012023 | 14 | 33025960 | 0.0231 | 0.0045 | 3.036911e-07 | NPAS3 |
| 53 | rs10133168 | 14 | 104507267 | 0.0158 | 0.0036 | 9.00513e-06 | AHNAK2 |
| 54 | rs12594522 | 15 | 30855341 | 0.0505 | 0.0104 | 1.229724e-06 | FMN1 |
| 55 | rs12591918 | 15 | 32807797 | 0.0313 | 0.0068 | 4.95362e-06 | |
| 56 | rs2413886 | 15 | 46265512 | 0.0262 | 0.0058 | 6.757172e-06 | CTXN2 |
| 57 | rs1566091 | 15 | 84566901 | 0.0106 | 0.0023 | 5.642319e-06 | AGBL1 |
| 58 | rs13379679 | 15 | 85704281 | 0.0495 | 0.0100 | 6.719055e-07 | |
| 59 | rs8042285 | 15 | 90116648 | 0.0346 | 0.0066 | 1.828344e-07 | |
| 60 | rs8030496 | 15 | 90122915 | 0.0353 | 0.0064 | 3.203015e-08 | |
| 61 | rs35356834 | 16 | 66253866 | 0.0267 | 0.0057 | 3.216372e-06 | PARD6A |
| 62 | rs1113232 | 16 | 66437328 | 0.0270 | 0.0058 | 2.75359e-06 | CENPT |
| 63 | rs28679372 | 16 | 66440433 | 0.0252 | 0.0050 | 5.610266e-07 | NUTF2 |
| 64 | rs5923 | 16 | 66531454 | 0.0225 | 0.0049 | 4.012146e-06 | LCAT |
| 65 | rs1133090 | 16 | 66578969 | 0.0263 | 0.0051 | 2.577475e-07 | DPEP2 |
| 66 | rs11860071 | 16 | 66894393 | 0.0258 | 0.0052 | 6.53352e-07 | SLC7A6OS |
| 67 | rs7501522 | 17 | 285907 | 0.0346 | 0.0071 | 1.217473e-06 | |
| 68 | rs7235119 | 18 | 2034076 | 0.0448 | 0.0094 | 1.835765e-06 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 69 | rs9748638 | 18 | 7867052 | 0.0182 | 0.0039 | 3.660023e-06 | PTPRM |
| 70 | rs9958551 | 18 | 43490826 | 0.0469 | 0.0086 | 5.682477e-08 | |
| 71 | rs740021 | 19 | 803104 | 0.0314 | 0.0054 | 7.566308e-09 | ELANE |
| 72 | rs1274688 | 19 | 55692129 | 0.0557 | 0.0093 | 2.397298e-09 | |
| 73 | rs20554 | 22 | 39883205 | 0.0534 | 0.0089 | 2.0182e-09 | EP300 |
| 74 | rs8135305 | 22 | 45097718 | 0.0216 | 0.0048 | 5.759904e-06 | GTSE1 |

Table 5.45: *Significant salmon GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs12047204 | 1 | 4368607 | rs7946005 | 11 | 78464992 | 1.7910 | 0.3901 | 4.39917e-06 | | TENM4 |
| 2 | rs1541318 | 1 | 4534838 | rs643394 | 6 | 126635960 | -1.5134 | 0.2944 | 2.73001e-07 | | |
| 3 | rs1541318 | 1 | 4534838 | rs4724679 | 7 | 5451609 | -17.1773 | 3.7331 | 4.19874e-06 | | |
| 4 | rs705191 | 1 | 33374139 | rs7966642 | 12 | 23782609 | 18.8706 | 3.5243 | 8.58381e-08 | | |
| 5 | rs3902720 | 1 | 50377924 | rs17093638 | 9 | 136259102 | 1.2244 | 0.2679 | 4.87643e-06 | | |
| 6 | rs10127537 | 1 | 58562458 | rs11810097 | 1 | 232357275 | 1.1701 | 0.2637 | 9.1167e-06 | | |
| 7 | rs10801851 | 1 | 91519594 | rs1328531 | 9 | 79579741 | 1.8565 | 0.4126 | 6.7977e-06 | | |
| 8 | rs743113 | 1 | 94215327 | rs7626405 | 3 | 20151852 | -2.2311 | 0.4997 | 7.99887e-06 | | |
| 9 | rs12080747 | 1 | 101858086 | rs820007 | 2 | 173831519 | -17.7399 | 3.7119 | 1.76065e-06 | | |
| 10 | rs2127980 | 1 | 105034488 | rs745831 | 19 | 51715250 | 18.1066 | 3.9116 | 3.67549e-06 | | |
| 11 | rs3134871 | 1 | 150565857 | rs271245 | 5 | 53223060 | -17.8103 | 3.8408 | 3.53286e-06 | | |
| 12 | rs1208373 | 1 | 167441820 | rs913891 | 20 | 55851133 | 17.2376 | 3.7057 | 3.2938e-06 | | |
| 13 | rs1938522 | 1 | 185702203 | rs16901239 | 8 | 90951107 | -2.1930 | 0.4799 | 4.87962e-06 | | |
| 14 | rs4025021 | 1 | 199695134 | rs7661552 | 4 | 166259824 | -17.5400 | 3.9542 | 9.17575e-06 | | |
| 15 | rs2551325 | 2 | 23833233 | rs16882141 | 6 | 52127622 | -17.6612 | 3.7828 | 3.02957e-06 | | |
| 16 | rs7562996 | 2 | 68940864 | rs1328674 | 13 | 46339708 | 2.9105 | 0.6262 | 3.35506e-06 | | |
| 17 | rs3792223 | 2 | 68945648 | rs1328674 | 13 | 46339708 | 3.4105 | 0.6811 | 5.52391e-07 | BMP10 | |
| 18 | rs6743132 | 2 | 85307104 | rs16969846 | 15 | 76516794 | -2.3816 | 0.5290 | 6.72647e-06 | | |
| 19 | rs10208875 | 2 | 134469526 | rs8033596 | 15 | 60914346 | 2.4620 | 0.4724 | 1.87147e-07 | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | rs10200377 | 2 | 146407901 | rs2566830 | 6 | 123006981 | 1.6218 | 0.3605 | 6.85146e-06 | |
| 21 | rs820007 | 2 | 173831519 | rs10838239 | 11 | 44112011 | 17.9575 | 4.0181 | 7.85295e-06 | |
| 22 | rs16839858 | 2 | 204075021 | rs3866838 | 4 | 111973264 | -1.5618 | 0.3433 | 5.37838e-06 | |
| 23 | rs7634161 | 3 | 2541831 | rs7946005 | 11 | 78464992 | 1.7198 | 0.3871 | 8.90318e-06 | TENM4 |
| 24 | rs7634161 | 3 | 2541831 | rs2427537 | 20 | 61846671 | -1.8606 | 0.3824 | 1.14268e-06 | |
| 25 | rs12636309 | 3 | 2548531 | rs2427537 | 20 | 61846671 | -1.7883 | 0.3879 | 4.01153e-06 | |
| 26 | rs11714594 | 3 | 4412484 | rs6414745 | 4 | 11990689 | -1.2889 | 0.2598 | 7.03247e-07 | |
| 27 | rs3911945 | 3 | 5107977 | rs11204411 | 17 | 19564494 | -17.3735 | 3.8840 | 7.70934e-06 | |
| 28 | rs2887976 | 3 | 64290071 | rs13125670 | 4 | 10122170 | 1.3308 | 0.2984 | 8.21186e-06 | |
| 29 | rs11128181 | 3 | 70690895 | rs11754094 | 6 | 106216673 | 1.3513 | 0.3035 | 8.48149e-06 | |
| 30 | rs16838468 | 3 | 98414526 | rs831465 | 11 | 20135461 | 2.4301 | 0.5313 | 4.78625e-06 | |
| 31 | rs4299484 | 3 | 109672317 | rs9377141 | 6 | 148826747 | -24.2549 | 5.3091 | 4.91175e-06 | |
| 32 | rs11917814 | 3 | 119334708 | rs16928055 | 11 | 35870147 | 1.5676 | 0.3416 | 4.45501e-06 | |
| 33 | rs11927890 | 3 | 133484772 | rs17046936 | 4 | 166835457 | 1.9884 | 0.4344 | 4.71527e-06 | |
| 34 | rs970235 | 3 | 152304758 | rs10746798 | 9 | 89157155 | -1.1704 | 0.2511 | 3.13817e-06 | |
| 35 | rs970235 | 3 | 152304758 | rs3793708 | 10 | 5805643 | -2.3536 | 0.5269 | 7.92367e-06 | |
| 36 | rs1346943 | 3 | 152307484 | rs10746798 | 9 | 89157155 | -1.1745 | 0.2514 | 2.97422e-06 | |
| 37 | rs1346943 | 3 | 152307484 | rs3793708 | 10 | 5805643 | -2.3482 | 0.5262 | 8.10085e-06 | |
| 38 | rs11730321 | 4 | 57433226 | rs193858 | 7 | 105363793 | -2.2079 | 0.4714 | 2.81213e-06 | |
| 39 | rs1387964 | 4 | 71296502 | rs9354072 | 6 | 94764903 | 16.9583 | 3.7979 | 8.00114e-06 | |
| 40 | rs2194225 | 5 | 35919561 | rs10483936 | 14 | 79442984 | -1.5988 | 0.3548 | 6.59064e-06 | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 41 | rs302483 | 5 | 88179847 | rs10828393 | 10 | 23329068 | -16.6583 | 3.7335 | 8.1249e-06 | |
| 42 | rs3776996 | 5 | 153777242 | rs9347621 | 6 | 162670416 | -2.8689 | 0.6398 | 7.32203e-06 | |
| 43 | rs12526605 | 6 | 1566624 | rs929026 | 22 | 33676625 | 2.2750 | 0.4602 | 7.69568e-07 | |
| 44 | rs12529159 | 6 | 24055592 | rs989996 | 7 | 117429488 | -17.6914 | 3.8365 | 4.00199e-06 | |
| 45 | rs12663510 | 6 | 42436047 | rs2830412 | 21 | 26994130 | 2.1415 | 0.4750 | 6.51705e-06 | TRERF1 |
| 46 | rs9388726 | 6 | 130077505 | rs9518226 | 13 | 100398863 | -1.8832 | 0.4208 | 7.63203e-06 | |
| 47 | rs6456182 | 6 | 170388033 | rs10774348 | 12 | 5573729 | 1.3579 | 0.3010 | 6.42949e-06 | |
| 48 | rs12539924 | 7 | 10309563 | rs2114076 | 11 | 25272547 | -2.4557 | 0.5303 | 3.64873e-06 | |
| 49 | rs36916 | 7 | 14969354 | rs169851 | 9 | 92767566 | -1.3064 | 0.2829 | 3.87523e-06 | |
| 50 | rs11975368 | 7 | 52562628 | rs12277367 | 11 | 107945626 | -1.7715 | 0.3687 | 1.55476e-06 | |
| 51 | rs1979600 | 7 | 154062483 | rs953334 | 12 | 24363082 | -1.5339 | 0.3367 | 5.20707e-06 | |
| 52 | rs6991079 | 8 | 12777125 | rs929026 | 22 | 33676625 | 1.9771 | 0.4289 | 4.02291e-06 | |
| 53 | rs1809437 | 8 | 68691402 | rs16970072 | 17 | 29924662 | -1.5233 | 0.3241 | 2.60409e-06 | |
| 54 | rs7875771 | 9 | 2923001 | rs929026 | 22 | 33676625 | -17.8825 | 3.4586 | 2.3352e-07 | |
| 55 | rs169851 | 9 | 92767566 | rs2872615 | 17 | 8855418 | -1.1918 | 0.2613 | 5.09678e-06 | |
| 56 | rs734632 | 9 | 97858467 | rs8017553 | 14 | 22807362 | 1.2013 | 0.2687 | 7.78357e-06 | |
| 57 | rs11573709 | 9 | 109125076 | rs9555773 | 13 | 110571109 | -1.5544 | 0.3518 | 9.96228e-06 | |
| 58 | rs12219721 | 10 | 114930505 | rs12279209 | 11 | 82630953 | 20.2670 | 4.3531 | 3.22792e-06 | |
| 59 | rs1080015 | 11 | 2468103 | rs16988469 | 20 | 37838349 | 17.3197 | 3.5842 | 1.34997e-06 | |
| 60 | rs10838382 | 11 | 44877202 | rs2427537 | 20 | 61846671 | -1.7710 | 0.3997 | 9.3838e-06 | |
| 61 | rs7946005 | 11 | 78464992 | rs2427537 | 20 | 61846671 | -2.3607 | 0.5162 | 4.7948e-06 | TENM4 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 62 | rs7946005 | 11 | 78464992 | rs2834428 | 21 | 34561844 | 3.1717 | 0.5047 | 3.30008e-10 | TENM4 |
| 63 | rs7946005 | 11 | 78464992 | rs4817630 | 21 | 34567605 | 2.5893 | 0.4567 | 1.42796e-08 | TENM4 |
| 64 | rs12587046 | 14 | 87249362 | rs2830412 | 21 | 26994130 | 17.8064 | 3.8326 | 3.38457e-06 | |
| 65 | rs12444250 | 16 | 19031915 | rs133076 | 22 | 39411507 | -21.5003 | 4.6143 | 3.17007e-06 | |
| 66 | rs8097281 | 18 | 57193201 | rs3919594 | 18 | 68922380 | -17.2152 | 3.8679 | 8.55575e-06 | |
| 67 | rs16988469 | 20 | 37838349 | rs913891 | 20 | 55851133 | 17.8300 | 3.8765 | 4.23414e-06 | |
| 68 | rs2834428 | 21 | 34561844 | rs4044210 | 22 | 45164979 | -1.7960 | 0.3950 | 5.43591e-06 | |

Table 5.46: *Significant salmon Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0083 | CORUM:1779 | CORUM | TGF-beta-receptor-PAR6 complex |
| 2 | 0.0083 | CORUM:5375 | CORUM | EGR-EP300 complex |
| 3 | 0.0083 | CORUM:5261 | CORUM | TCF4-CTNNB1-EP300 complex |
| 4 | 0.0083 | CORUM:6354 | CORUM | Sox4-beta-catenin-p300 complex |
| 5 | 0.0083 | CORUM:6502 | CORUM | Estrogen receptor complex (ESR1, EP300, NCOA1) |
| 6 | 0.0083 | CORUM:2642 | CORUM | SMAD1-P300 complex |
| 7 | 0.0083 | CORUM:6586 | CORUM | VEcad-VEGFR complex |
| 8 | 0.0083 | CORUM:5535 | CORUM | PLCB3-PARD3-PARD6A complex |
| 9 | 0.0083 | CORUM:1831 | CORUM | PIAS3-SMAD3-P300 complex |
| 10 | 0.0083 | CORUM:6144 | CORUM | CENP-T-W complex |
| 11 | 0.0083 | CORUM:1521 | CORUM | p300-SMAD1-STAT3 complex |
| 12 | 0.0083 | CORUM:5388 | CORUM | SERPINA1-ELA2 complex |
| 13 | 0.0083 | CORUM:6942 | CORUM | FZD5-LGR5-LRP6 complex |
| 14 | 0.0083 | CORUM:1160 | CORUM | ING1-p300-PCNA complex |
| 15 | 0.0083 | CORUM:1158 | CORUM | p33ING1b-p300 complex |
| 16 | 0.0083 | CORUM:7328 | CORUM | AML1-HIPK2-p300 complex |
| 17 | 0.0083 | CORUM:918 | CORUM | PAR3-PAR6-PALS1 complex |
| 18 | 0.0083 | CORUM:5740 | CORUM | NRP2-VEGFR3 complex |
| 19 | 0.0083 | CORUM:831 | CORUM | PAR-6-PAR-3-VE-cadherin complex, endothelial |
| 20 | 0.0083 | CORUM:829 | CORUM | PAR-6-VE-cadherin complex, endothelial |
| 21 | 0.0083 | CORUM:571 | CORUM | p300-CBP-p270 complex |
| 22 | 0.0083 | CORUM:7564 | CORUM | LGR5-RNF43-RSPO1 complex |
| 23 | 0.0083 | CORUM:98 | CORUM | p300-MDM2-p53 protein complex |
| 24 | 0.0083 | CORUM:7300 | CORUM | PARD3B-PARD6A-PRKCI complex |

| | | | | |
|---|---|---|---|---|
| 25 | 0.0083 | CORUM:5534 | CORUM | PLCB1-PARD3-PARD6A complex |
| 26 | 0.0096 | CORUM:4 | CORUM | Multisubunit ACTR coactivator complex |
| 27 | 0.0096 | CORUM:7444 | CORUM | HOOK2-PAR3-Par6alpha-PKCiota complex |
| 28 | 0.0096 | CORUM:6653 | CORUM | EP300-KAT2B-TBX5-WWTR1 complex |
| 29 | 0.0096 | CORUM:6146 | CORUM | CENP-T-W-S-X heterotetramer complex |
| 30 | 0.0108 | CORUM:5260 | CORUM | TCF4-CTNNB1-SUMO1-EP300-HADAC6 complex |
| 31 | 0.0108 | CORUM:1471 | CORUM | (RB2, E2F5, HDAC1, SUV39H1, P300) |
| 32 | 0.0108 | CORUM:905 | CORUM | KIF3A/B-PAR-3-aPKC-PAR-6 complex |
| 33 | 0.0119 | CORUM:5118 | CORUM | RBL2 complex |
| 34 | 0.0119 | CORUM:2638 | CORUM | HES1 promoter corepressor complex |
| 35 | 0.0119 | CORUM:927 | CORUM | CENP-A NAC complex |
| 36 | 0.0135 | CORUM:570 | CORUM | p300-CBP-p270-SWI/SNF complex |
| 37 | 0.0236 | CORUM:2639 | CORUM | HES1 promoter-Notch enhancer complex |
| 38 | 0.0236 | CORUM:1179 | CORUM | CENP-A NAC-CAD complex |
| 39 | 0.0247 | CORUM:7388 | CORUM | CENP-A nucleosomal complex |
| 40 | 0.0301 | CORUM:7581 | CORUM | ESR1-TRAP/Mediator coactivator-complex |

Table 5.47: *Significant gene set enrichment results of genes with SNPs found by salmon cluster analysis. P-values are FDR adjusted.*

| | rs | chr | ps | beta | se | p | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs6664362 | 1 | 9536458 | 0.0240 | 0.0032 | 1.483981e-13 | SLC25A33 |
| 2 | rs2360546 | 1 | 200155592 | 0.0117 | 0.0025 | 2.108225e-06 | LMOD1 |
| 3 | rs16849442 | 1 | 200158880 | 0.0106 | 0.0024 | 7.466283e-06 | LMOD1 |
| 4 | rs10779486 | 1 | 206805932 | 0.0118 | 0.0025 | 2.355672e-06 | |
| 5 | rs12471455 | 2 | 79763857 | 0.0177 | 0.0037 | 2.172188e-06 | CTNNA2 |
| 6 | rs2289129 | 3 | 14483315 | 0.0186 | 0.0042 | 8.383137e-06 | SLC6A6 |
| 7 | rs2944401 | 3 | 117044246 | 0.0152 | 0.0034 | 9.204589e-06 | LSAMP |
| 8 | rs2972479 | 3 | 117048717 | 0.0163 | 0.0034 | 1.872022e-06 | LSAMP |
| 9 | rs11925421 | 3 | 147370852 | 0.0184 | 0.0029 | 4.73261e-10 | |
| 10 | rs11934086 | 4 | 76867439 | 0.0271 | 0.0060 | 5.735106e-06 | G3BP2 |
| 11 | rs2713946 | 4 | 111417254 | 0.0328 | 0.0053 | 6.26633e-10 | |
| 12 | rs6866220 | 5 | 17078067 | 0.0403 | 0.0084 | 1.570396e-06 | |
| 13 | rs6914506 | 6 | 37374934 | -0.0088 | 0.0017 | 1.546578e-07 | TBC1D22B |
| 14 | rs1137086 | 6 | 149945290 | 0.0239 | 0.0052 | 5.054793e-06 | GINM1 |
| 15 | rs9372012 | 6 | 150766274 | 0.0202 | 0.0042 | 2.004581e-06 | IYD |
| 16 | rs16919513 | 8 | 54805952 | 0.0323 | 0.0071 | 6.35616e-06 | ATP6V1H |
| 17 | rs7827405 | 8 | 69927679 | 0.0293 | 0.0063 | 3.71653e-06 | |
| 18 | rs1888203 | 9 | 74713677 | 0.0426 | 0.0076 | 2.028845e-08 | ALDH1A1 |
| 19 | rs7029300 | 9 | 109192354 | 0.0137 | 0.0030 | 5.883305e-06 | |
| 20 | rs7946005 | 11 | 78464992 | 0.0125 | 0.0024 | 3.517999e-07 | TENM4 |
| 21 | rs11225401 | 11 | 102113169 | 0.0093 | 0.0020 | 4.135928e-06 | |
| 22 | rs2236336 | 14 | 24145569 | 0.0287 | 0.0060 | 2.141021e-06 | GZMH |
| 23 | rs1420995 | 16 | 50068938 | 0.0488 | 0.0090 | 6.728779e-08 | |
| 24 | rs235083 | 16 | 64760511 | -0.0096 | 0.0018 | 8.533857e-08 | |
| 25 | rs1420791 | 17 | 61345212 | 0.0220 | 0.0043 | 3.859276e-07 | CEP112 |
| 26 | rs9959968 | 18 | 10427781 | 0.0190 | 0.0038 | 5.643485e-07 | |
| 27 | rs7577 | 18 | 70339327 | 0.0109 | 0.0023 | 2.563995e-06 | CNDP2 |
| 28 | rs12151353 | 19 | 60878398 | 0.0365 | 0.0075 | 1.031181e-06 | |
| 29 | rs16999673 | 21 | 40610439 | 0.0355 | 0.0080 | 8.672033e-06 | DSCAM |
| 30 | rs4822679 | 22 | 24722172 | 0.0240 | 0.0050 | 1.85396e-06 | MYO18B |

Table 5.48: *Significant cyan GWAS Results*

| | ID1 | CHR1 | BP1 | ID2 | CHR2 | BP2 | beta | se | p | Gene1 | Gene2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs12128446 | 1 | 6977194 | rs9542596 | 13 | 70621297 | -1.9700 | 0.4418 | 8.23735e-06 | | |
| 2 | rs2297977 | 1 | 43187875 | rs1535692 | 13 | 93832750 | 1.5306 | 0.3327 | 4.20554e-06 | | |
| 3 | rs2297972 | 1 | 43190613 | rs11203200 | 21 | 42687644 | 1.6399 | 0.3620 | 5.91541e-06 | | |
| 4 | rs11162094 | 1 | 76417668 | rs12446798 | 16 | 28191236 | 2.2932 | 0.5122 | 7.57619e-06 | | |
| 5 | rs12726148 | 1 | 102860260 | rs7086624 | 10 | 73369151 | -17.6720 | 3.5360 | 5.79921e-07 | | |
| 6 | rs1241182 | 1 | 103120688 | rs7086624 | 10 | 73369151 | -17.6075 | 3.4055 | 2.33648e-07 | | |
| 7 | rs13375429 | 1 | 111904845 | rs11954658 | 5 | 140901595 | 3.8089 | 0.8518 | 7.76249e-06 | | |
| 8 | rs11102354 | 1 | 112235189 | rs16966406 | 15 | 36264553 | 17.8819 | 4.0332 | 9.26558e-06 | | |
| 9 | rs6703705 | 1 | 113175228 | rs10203969 | 2 | 217193795 | 19.1931 | 4.2631 | 6.72647e-06 | | |
| 10 | rs6703705 | 1 | 113175228 | rs12880418 | 14 | 78239581 | -17.5491 | 3.8575 | 5.3805e-06 | | |
| 11 | rs12132927 | 1 | 151698185 | rs7585718 | 2 | 11640781 | 17.4523 | 3.8720 | 6.56357e-06 | | |
| 12 | rs9435886 | 1 | 246395636 | rs7626182 | 3 | 195328616 | -17.2831 | 3.5659 | 1.25461e-06 | | |
| 13 | rs6760967 | 2 | 4647199 | rs166213 | 16 | 64456437 | 1.7659 | 0.3957 | 8.09826e-06 | | |
| 14 | rs1558628 | 2 | 10395873 | rs17021982 | 3 | 2911576 | -2.3225 | 0.5234 | 9.1228e-06 | | |
| 15 | rs1558628 | 2 | 10395873 | rs12153515 | 5 | 164564372 | 1.6423 | 0.3635 | 6.253e-06 | | |
| 16 | rs6758633 | 2 | 10397652 | rs12153515 | 5 | 164564372 | 1.6634 | 0.3698 | 6.85873e-06 | | |
| 17 | rs6758633 | 2 | 10397652 | rs1638586 | 11 | 66931173 | 1.6241 | 0.3602 | 6.52401e-06 | | |
| 18 | rs7585718 | 2 | 11640781 | rs12313273 | 12 | 120547393 | 18.2520 | 3.9889 | 4.74574e-06 | | |
| 19 | rs7585718 | 2 | 11640781 | rs3741595 | 12 | 120563572 | 4.1800 | 0.9058 | 3.93458e-06 | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | rs7585718 | 2 | 11640781 | rs6127466 | 20 | 36276297 | 18.2759 | 3.8876 | 2.58847e-06 |
| 21 | rs6733711 | 2 | 12498718 | rs1286738 | 3 | 25589536 | 17.2235 | 3.6827 | 2.91379e-06 |
| 22 | rs1530394 | 2 | 31217236 | rs2834114 | 21 | 33451104 | 1.8019 | 0.3999 | 6.61117e-06 |
| 23 | rs207440 | 2 | 31415916 | rs812315 | 12 | 56279757 | -2.9958 | 0.6683 | 7.38007e-06 |
| 24 | rs2273659 | 2 | 32781367 | rs759159 | 7 | 55146894 | 1.4506 | 0.3205 | 6.0088e-06 |
| 25 | rs7596775 | 2 | 58015024 | rs10934665 | 3 | 125525876 | -18.0649 | 3.9136 | 3.91362e-06 |
| 26 | rs7585161 | 2 | 76723811 | rs1710896 | 3 | 10992136 | 2.4261 | 0.5036 | 1.45564e-06 |
| 27 | rs16844617 | 2 | 141217162 | rs6794294 | 3 | 65618598 | 1.6192 | 0.3635 | 8.4005e-06 |
| 28 | rs7604580 | 2 | 158769504 | rs16970049 | 15 | 38038887 | 2.1940 | 0.4782 | 4.46417e-06 |
| 29 | rs7579360 | 2 | 158769781 | rs16970049 | 15 | 38038887 | 2.2057 | 0.4782 | 3.98312e-06 |
| 30 | rs6740826 | 2 | 168735801 | rs10770407 | 12 | 18978417 | -2.8427 | 0.6125 | 3.46891e-06 |
| 31 | rs540524 | 2 | 169465176 | rs16926560 | 11 | 44536546 | 17.7545 | 3.8100 | 3.1616e-06 |
| 32 | rs4125973 | 2 | 180654320 | rs1420995 | 16 | 50068938 | 2.5140 | 0.5626 | 7.87318e-06 |
| 33 | rs6717445 | 2 | 221643665 | rs17094645 | 14 | 96475737 | 4.3796 | 0.9638 | 5.51924e-06 |
| 34 | rs17050639 | 3 | 9920978 | rs6925338 | 6 | 994017 | -17.8520 | 4.0070 | 8.38129e-06 |
| 35 | rs1710896 | 3 | 10992136 | rs3028 | 9 | 89341182 | -17.4121 | 3.9083 | 8.38177e-06 |
| 36 | rs4340668 | 3 | 24972917 | rs12880418 | 14 | 78239581 | -22.8939 | 5.0397 | 5.5531e-06 |
| 37 | rs10934665 | 3 | 125525876 | rs3741595 | 12 | 120563572 | 18.7760 | 3.8680 | 1.20852e-06 |
| 38 | rs10934665 | 3 | 125525876 | rs12880418 | 14 | 78239581 | -18.1415 | 3.8859 | 3.03276e-06 |
| 39 | rs4599370 | 4 | 169878084 | rs16928463 | 9 | 114903731 | 17.7206 | 3.8768 | 4.85452e-06 |
| 40 | rs17064782 | 4 | 178550268 | rs12153515 | 5 | 164564372 | 1.6372 | 0.3690 | 9.15257e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 41 | rs17739994 | 4 | 178550508 | rs12153515 | 5 | 164564372 | 1.9466 | 0.4095 | 2.00144e-06 |
| 42 | rs11724481 | 4 | 188588781 | rs12363087 | 11 | 45069446 | 20.2031 | 4.4527 | 5.70008e-06 |
| 43 | rs13186768 | 5 | 18084440 | rs12441058 | 15 | 59164913 | -1.2767 | 0.2857 | 7.87279e-06 |
| 44 | rs6451722 | 5 | 43747135 | rs16966406 | 15 | 36264553 | -17.8555 | 3.9949 | 7.8353e-06 |
| 45 | rs809798 | 5 | 53284120 | rs12877019 | 13 | 25377030 | -1.3646 | 0.3071 | 8.83086e-06 |
| 46 | rs809798 | 5 | 53284120 | rs6134722 | 20 | 12856451 | -17.7285 | 3.8899 | 5.1747e-06 |
| 47 | rs10474648 | 5 | 80444622 | rs9927125 | 16 | 86029374 | 17.5885 | 3.8208 | 4.15837e-06 |
| 48 | rs12153515 | 5 | 164564372 | rs885201 | 12 | 2982639 | 1.5626 | 0.3456 | 6.12601e-06 |
| 49 | rs4395771 | 6 | 3249777 | rs7976542 | 12 | 79020516 | 18.3472 | 3.9578 | 3.55738e-06 |
| 50 | rs7761394 | 6 | 21889931 | rs8084524 | 18 | 41168213 | 17.1916 | 3.8868 | 9.73047e-06 |
| 51 | rs6901028 | 6 | 62042822 | rs1011387 | 8 | 128351093 | 2.3961 | 0.5321 | 6.70988e-06 |
| 52 | rs12662237 | 6 | 148150583 | rs3741595 | 12 | 120563572 | -2.8634 | 0.6459 | 9.29141e-06 |
| 53 | rs9384449 | 6 | 150725088 | rs12154752 | 7 | 124837765 | 1.4320 | 0.2848 | 4.96966e-07 |
| 54 | rs9322486 | 6 | 155296741 | rs9901870 | 17 | 72575732 | 17.2411 | 3.7977 | 5.62825e-06 |
| 55 | rs17707505 | 7 | 22124589 | rs4945142 | 11 | 76484170 | 2.3170 | 0.4719 | 9.09473e-07 |
| 56 | rs739627 | 7 | 24820021 | rs16926560 | 11 | 44536546 | 2.2266 | 0.4732 | 2.53641e-06 |
| 57 | rs12546334 | 8 | 15764239 | rs2292910 | 11 | 45860189 | 3.0271 | 0.6820 | 9.06791e-06 |
| 58 | rs10108727 | 8 | 15816860 | rs10441687 | 9 | 38154364 | 1.4948 | 0.3344 | 7.83732e-06 |
| 59 | rs7842248 | 8 | 89697728 | rs17514846 | 15 | 89217554 | -3.3992 | 0.7428 | 4.73109e-06 |
| 60 | rs11784608 | 8 | 96271460 | rs9584922 | 13 | 98182541 | 19.2755 | 4.3413 | 8.99505e-06 |
| 61 | rs11142907 | 9 | 73397341 | rs12589344 | 14 | 85027423 | 16.7630 | 3.6866 | 5.44255e-06 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 62 | rs12220112 | 10 | 8852761 | rs12589344 | 14 | 85027423 | 3.1679 | 0.7081 | 7.69236e-06 |
| 63 | rs1638586 | 11 | 66931173 | rs3803452 | 15 | 55367038 | 2.2581 | 0.4903 | 4.11244e-06 |
| 64 | rs7105776 | 11 | 72401751 | rs12589344 | 14 | 85027423 | 3.1567 | 0.6968 | 5.88177e-06 |
| 65 | rs871639 | 12 | 67725342 | rs7204722 | 16 | 85987740 | 1.6518 | 0.3362 | 8.93642e-07 |
| 66 | rs6581845 | 12 | 67741183 | rs7204722 | 16 | 85987740 | 1.7100 | 0.3449 | 7.10923e-07 |
| 67 | rs1873570 | 12 | 98961863 | rs16975208 | 18 | 37538035 | 3.0999 | 0.6724 | 4.0265e-06 |
| 68 | rs7296162 | 12 | 98969115 | rs16975208 | 18 | 37538035 | 3.1095 | 0.7009 | 9.1386e-06 |
| 69 | rs7296162 | 12 | 98969115 | rs16997918 | 22 | 35925862 | -17.5512 | 3.8066 | 4.01126e-06 |
| 70 | rs7963085 | 12 | 112709174 | rs6087433 | 20 | 31017374 | -17.2793 | 3.8685 | 7.94709e-06 |
| 71 | rs3741595 | 12 | 120563572 | rs1500106 | 12 | 124315518 | -1.8022 | 0.4053 | 8.72556e-06 |
| 72 | rs7335200 | 13 | 22555518 | rs17721321 | 20 | 5244078 | -1.6987 | 0.3714 | 4.79016e-06 |
| 73 | rs17649047 | 15 | 49840521 | rs6044870 | 20 | 1754234 | 19.6224 | 4.0980 | 1.68202e-06 |
| 74 | rs17649255 | 15 | 49908897 | rs6044870 | 20 | 1754234 | 18.8020 | 3.9719 | 2.20363e-06 |
| 75 | rs12592953 | 15 | 94329349 | rs6417104 | 18 | 45253512 | -1.3819 | 0.2851 | 1.25191e-06 |
| 76 | rs3102350 | 16 | 87907437 | rs2836034 | 21 | 38179247 | 2.5746 | 0.5817 | 9.58299e-06 |

Table 5.49: *Significant cyan Epistasis Results*

| | p_value | term_id | source | term_name |
|---|---|---|---|---|
| 1 | 0.0499 | CORUM:7357 | CORUM | TRIM25-G3BP2 complex |
| 2 | 0.0197 | TF:M03916 | TF | Factor: PRDM1; motif: NRAAAGTGAAAGTNN |
| 3 | 0.0197 | TF:M03916_0 | TF | Factor: PRDM1; motif: NRAAAGTGAAAGTNN; match class: 0 |

Table 5.50: *Significant gene set enrichment results of genes with SNPs found by cyan cluster analysis. P-values are analytically adjusted.*

# Chapter Six

# Discussion and Thesis Conclusions

The genesis of this body of work was the near insurmountable task of understanding the genetic underpinnings of behavioral and developmental disorders affecting human health, and a desire to contribute to the ability of basic researchers to prioritize areas of focus in a world ever more drowned in biological data. All projects grow, and this work started small by classifying circadian functions of genes expressed in the suprachiasmatic nucleus, deep in the anterior hypothalamus of the mouse brain. As disrupted circadian rhythmicity and expression is a known hallmark of psychological disorders, I exploited massive biobank data to draw inferences about the relationship between chronotype and several measures of psychosocial wellbeing. I then moved to characterize behavioral traits relating to complex neurobehavioral dysfunction in autism and psychosis, before testing phenotypic models by exploring genetic associations among diverse phenotypes. Before setting out on this work, I aimed to attempt to accomplish for interrelated tasks:

- 1: Predict genes which are involved in producing abnormal circadian rhythm behavior in mice using supervised machine learning methods
- 2: Investigate the causal relationship between chronotype, a measure of circadian rhythm, and behavioral traits linked to mental illness or psychological disorders in

humans

- 3: Mine psychological assessment instruments to model subtle behaviors which, in aggregate, are indicative of complex psychological disorders such as autism or schizophrenia, and use that to reduce the phenotypic heterogeneity among populations with complex behavioral diagnoses

- 4: Investigate genes which are associated with individuals who share both a common behavioral diagnosis (autism) and share a common phenotypic profile derived from aim 3.

## 6.1 Overview of major findings

In Chapter 2, I used a multi-modal machine learning model to classify genes which may contribute to abnormal circadian rhythm function in mouse. I was able to exploit measurements of rhythmicity and gene expression values in predictions. I showed that combining measurements of rhythmicity, bulk expression, tissue specificity, and protein-protein interactions can generate robust predictions of circadian gene function. Including protein data substantially improved model performance, illustrating the value in multimodal data integration for gene function prediction.

Previously studied genes circadian genes, such as NPAS2, revealed a disparity between Mammalian Phenotype and Gene Ontology annotated circadian traits. This finding demonstrates the care that must be used in basing findings off of bioinformatics databases, while suggesting the Mouse Genome Database (Bult et al., 2019b) to be a conservative gold standard of experimentally based gene annotations. Working with colleagues, I predicted Grp to contribute to circadian phenotypes; this was experimentally validated, facilitating correcting this oversight in the MGD database (Bult et al., 2019a). Potentially more useful,

though, is that I predicted 246 novel genes to be involved in circadian traits. While predictions do not necessarily reflect the biological reality, and I do not expect every gene to produce a phenotype, several had strong external evidence to bolster their reliability. Among these, several had tissue-specific profiles for the SCN, indicating a lack of expression in other brain regions as well as throughout the mouse body.

In Chapter 3, I performed Mendelian randomization experiments using the UK Biobank cohort in a mental-health based phenome-wide manner, investigating the potential causal relationships between evening chronotype and mental health. Significant associations were identified between exposure to evening chronotype and experiencing a combination of "manic symptoms," including being more active than usual, getting less sleep, and being more talkative and communicative than usual. This association, while not surprising, is fairly non-specific and includes sleep traits. Though sleep is related to circadian exposure, the two phenomena are distinct; yet the causal association could be bidirectional. Measures of emotional self-appraisal, including having one's feelings hurt easily and reflecting on embarrassing moments, were negatively associated with an evening chronotype. Of the measures tested, several did not produce associations (see online Appendix). These included psychosocial factors, such as friendship satisfaction, financial situation satisfaction, and work/job satisfaction. Although these were modeled, heterogeneity was high and even negative results are likely unreliable, unless additional confounding variables are accounted for in experimental design. Without appropriately designing causal diagrams when modeling behavioral traits, false positive results are common (Burgess, Small, and Thompson, 2017). Social support measures were included in the analysis, the first time chronotype has been suggested to have a causal role in hobbies or gatherings. Unsurprisingly from a western cultural context, there was a highly significant ($p = 4.2e\text{-}5$) causal association between a morning chronotype and resulting religious gatherings. Though effects were highly variable, the overall effect suggests causation and opens up a potential area of study between chronotype and social events.

While questions about the context of religion and chronotype are interesting, modeling gene x environment interactions may illuminate this connection, and this result should be repeated in a multivariate Mendelian randomization experiment to control for other indicators of extroversion, personality type, or other factors which may influence social gatherings. Finally in this chapter, I modelled the relationship between chronotype and keratometry measurements, and found one (out of four modelled) potential association. While light as a zeitgeber travels through the eye to regulate chronotype, these findings suggest that chronotype may have an affect on eye morphology and angles of refraction. To pursue this finding experimentally, phenotype screens via the International Mouse Phenotyping Consortium could be conducted since eye morphology data is routinely collected. A strength of this study was the robustness of statistical methods, which attempted to attenuate bidirectional (balanced) pleiotropy as well as weak genetic instruments. This was balanced by the choice of instruments used, none were in linkage disequilibrium (ensuring independence), and they were selected from an independent cohort to avoid the "winner's curse". Using a two-sample approach also catered for this analysis to be justifiably conservative, will pleiotropy biasing associations toward the null. One additional note about Chapter 3 that needs restating, is that the 18 SNPs acted as proxy for a lifetime exposure to an evening chronotype; and in this analysis the biological roles of the individual SNPs are less relevant than in a *cis* Mendelian study focused on one locus. Chapter 3 completed my focus on chronotype, and the phenotypic complexity of many traits discussed in this chapter motivated a new appraisal of the state of behavioral ontologies.

In Chapter 4, I co-created the Aging Neuro Behavior Ontology and created the Psychological Neuro Behavior Ontology, ANBO and PNBO respectively. Each ontology is designed to meet the best practice standards for ontology creation and maintenance; more importantly both are logically consistent and contain traits needed to model select cognitive and behavioral traits. While this study focused on human behavioral traits relating to

autism and schizophrenia, the ontology itself remains species neutral, and patterning the E-Q notation will facilitate working with murine neurogeneticists. There are several reasons why modeling behaviors with the PNBO is useful; the first being the structure of the ontology - this is shared by a large portion of the Neuro Behavior Ontology, and the primary value is in creating new knowledge that is not implicitly coded into the ontology. New class relations are created when reasoners run, and annotations are expanded when probands, patients, or genes area annotated to the ontology thanks to the true path rule. This facilitates mapping individuals to traits that, on the surface, may not have much in common but share common ancestor terms, linking individuals close in function. The utility of this approach was demonstrated when surveying the concurrence of ASD and SSD traits. The diagnosis of disorders historically overlap (Rutter, 1972 Oct-Dec), but this lack of distinction is echoed today by a recent meta-analysis of ASD and SSD GWAS, where the consortium (Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017) identified a "neurodevelopmental hub" when combining cohorts into one analysis. Their analysis could be augmented by phenotyping participants with the PNBO and performing GWAS for traits which overlap between ASD and SSD affected individuals, providing deeper insight into very heterogeneous syndromes. A useful aspect of the PNBO is that it separates the concepts of disease and trait, something that the NBO and Human Phenotype ontologies do not delinate as successfully. Thankfully this issue does not arise in the Mammalian Phenotype ontology, likely because mouse geneticists are not trying to diagnose mice but rather observe traits. This work assumes behavioral disorders as a constellation of traits that different clinicians or diagnosticians may assign different labels (Asperger's vs ASD in the DSM-IV, for instance), but whose phenotypic manifestations have physiological or environmental causes. An experimental result in Chapter 4, which contributes to the literature, was clustering the spectrum, or segregating probands by their indicative phenotype profiles. These phenotypic profiles of probands in the clusters were often at high levels of the ontology, but some clusters of probands exhibited well known trait profiles, such as having several restricted behaviors as

a hallmark.

In Chapter 5, I extended my analysis of autism endophenotypes in the SSC cohort by attempting genome-wide association studies. As is not surprising given the sample size of some traits tested, no increased power was seen when performing GWAS associations on individual traits. Likewise, epistatic interactions did not reach genome-wide significance to a strict p $<$ 1e-8 standard. Interestingly, epistatic interactions associated to individual traits and to trait profiles had both negative (or less than 1 odds ratio), and positive (more than one OR) $\beta$s, indicating that protective and deleterious pairs of interacting genes were found. Epistatic interaction tests were run under a model of dominant inheritance, and changing these assumptions may enable detection of different interacting alleles. Contrary to the poor performance of single-trait GWAS, cluster based GWAS had good performance, with some SNPs in a range of clusters reaching extremely low p-values. Cluster GWAS were validated with a machine learning approach, wherein the cohort was split into training/test sets the the allele (coded 0,1,2) for each SNP deemed significant to that cluster via association studies were used as features belonging to their respective probands. These genetic features were used in a LASSO penalized logistic regression model to predict cluster membership; the holdout probands were then used to test the classifier's learned performance. In the absence of another cohort to study, this validation strategy is innovative. It cannot replace GWAS replication on an independently phenotyped cohort, but does allow a measure of the ability of the genomic makeup of a cluster's probands to predict that the probands belong to that cluster. Lastly, the highest performing cluster in validation experiments, made up of 95 SNPs, of which nine were intragenic. Of these, two genes were indicative of circadian function - NPAS2, discussed in Chapter 2, and CADPS2. In humans, CADPS2 is not annotated to autism traits, but in mouse it was observed to be related ASD traits and abnormal gait, which is a hallmark of the tan cluster these genes are associated with. Without being informed by model organism studies, analysis of this cluster would have been incomplete.

Innovations in this chapter include not only individual cluster/gene relationships, but the network-based view of interrelated traits and genes associated with them in the SSC cohort. When annotating an ontology term with a gene or an individual, there is no room for doubt, and no probabilities associated with that annotation; this is necessary for logical completeness and to enable reasoners to make inferences. However, biology is seldom definitive, and having a probabilistic way of annotating genes to ontology terms would be an innovation that could be pursued outside of this work.

This project focused on revealing gene/trait relationships within neurobehavioral processes and the broader domain of neurological and mental health disorders. Serveral methods and and applications presented here have implications beyond this domain, especially where complex syndromes or highly comorbid conditions exist. The PNBO was used to encode specific phenotypic traits and serve as a basis for modeling the semantic similarity of patients who presented with multiple variations of said traits. This allowed for the stratification of patients that incorporated the relationship between the observed traits and included otherwise hidden dependencies between patient symptoms which may be shared by portions of the cohort. This methods may be applied to comorbid diseases when rich phenotyping data is available, and may be particularly suited to common disorders such as sarcopenia or other highly prevalent multifactorial syndromes. In order to properly harness such an application, corrections to the Human Phenotype ontology mentioned in this thesis, and the incorporation of parts of the ANBO and PNBO into a broarder ontological framework, will be needed. The improved PNBO and ANBO descriptions of behavior which can be transferred to model organisms through the Phenotype and Trait Ontology to augment and refine the Mammalian Phenotype ontology. Incoproration of analagous PNBO and ANBO classes into the MP ontology may facilitate a more precise translation of observed behavioral observations when systematically studying models of non-behavioral or non-neurological disorders. Thus, improved ontologies may be useful in not only stratifiying patients with neurological

and non-neurological disorders, but in elicidating the behavioral effects of mutant models designed to study non-neurological phenomena.

## 6.2 Contributions the literature

- Several novel circadian-related genes have been proposed; among them are genes implicated in the SNARE complex (Sytl4), a G-protein coupled receptor (Calcr), and a hormone (Trh).

- With the creation of large genomics consortia and biobanks, including 23andMe and the UK Biobank, several chronotype GWAS have been published. This facilitated a comprehensive causal analysis of the influence of chronotype on reported measures of mental health, social support, and eye physiology.

- During this work, two biomedical ontologies were created. The Aging Neuro Behavior Ontology models behavioral processes relating to cognitive decline, and the Psychological Neuro Behavior Ontology models behavior phenotypes exhibited by patients on the autism and schizophrenia spectra. When published, PNBO will be the first biomedical ontology to attempt to model both autism and schizophrenia related behaviors together, and the first to model schizophrenia traits at all.

- I have discovered novel autism trait related loci in GWAS of the SSC cohort, and using a polygenic model have validated those via machine learning methods. In particular, NPAS2 and CADPS2 were implicated in gait and communication related phenotypes in an autistic population.

In summary, these contributions provide new knowledge to the scientific community of the potential role of characterized genes in circadian biology, of the likely causal influence

of chronotype on social and mental health, and provide novel robust ways of characterizing the complex phenotype of autism and schizophrenia patients.

## 6.3 Limitations

### 6.3.1 Limitations of transcriptome characterization

A strength of this thesis was its use of both model and human organisms as sources for creating new inferences, and using mammalian genetics to infer possible function to human associated genes. This is also a limitation, as not all phenotypic traits are guaranteed to phenocopy between mouse and human. Indeed, even the genetic background of inbred C57BL/6 mouse sub-strains (J and N) can dramatically influence the phenotype observed (Simon et al., 2013).

### 6.3.2 Polygenic causal associations

A limitation in Chapter 3 is the data-driven nature of SNP selection in the discovery dataset for Mendelian randomization. If I was estimating the effect of increased inflammation using Mendelian randomization, I would investigate inflammatory cytokines for *cis* based associations, picking cytokines based on the pathology or traits of interest as prevously performed (Collaboration (CCGC), 2011). To perform an analogous experiment, I could have chosen core clock genes as the focus of analysis.

## 6.4 Planed future work

There are several extension to this work which may be done; some are planned for validation and others to continue themes of this project.

### 6.4.1 Replication of gene associations

Both a limitation and future work, I aim to replicate the GWAS and phenotype associations presented herein by applying to the Autism Genetic Resource Exchange, AGRE (Geschwind et al., 2001). The AGRE holds an independent replication cohort of deeply phenotyped (ADOS, ADR-I, and others) autism probands for use in both phenotype cluster analysis and GWAS. As the input into clusters in this work was the semantic similarity of probands; and that relationship to the whole will change when 1) any new terms are added to the PNBO in which probands participate, and 2) new probands are added to the analysis, a perfect replication may not be possible. However, because of the hierarchical nature of the ontology probands with not identical but similar traits should still cluster together.

### 6.4.2 Validation of PNBO in schizophrenia patients

As part of the PIMS study currently underway at the universities of Birmingham and Cambridge, colleagues are preforming a meta-analysis of cohorts of schizophrenia patients from five cohorts, investigating inflammatory biomarkers of the SSD. Several cohorts are deeply phenotyped, and I aim to work with psychiatrists to interpret diagnostic and assessment tools used to enhance the PNBO and annotate study participants with PNBO traits. While current unsupervised learning methods by the group have not segregated patients based on phenotype, they are not using a structured, semantic approach to phenotyping patients.

The approach developed here may help, and would further validate the use of the PNBO for patient stratification.

### 6.4.3 Extending the PNBO ontology to more assessment instruments

To that end, I plan on extending the PNBO with more traits based on relevant psychological or psychiatric assessments. This will enable the PNBO to capture more phenotypic variation among ASD and SSD patients, and deeper phenotyping will allow more human traits to be mapped to mouse via E-Q statements. Having psychosis or autism related traits mapped directly to mouse may help guide researchers in planning phenotyping experiments, saving valuable resources and ensuring translatability of any findings.

### 6.4.4 Working with colleagues to validate circadian genes and gene/trait causality

To validate findings, I plan to work with colleagues from MRC Harwell Institute to plan behavioral phenotyping experiments of candidate circadian and autism-related traits. As the Institute changes, collaborations will be maintained.

### 6.4.5 Large scale studies and future plans

The initial methods used in this work, especially the methods in Chapters 3, 4, and 5, will be applied within biobank-scale data. During the course of this work, I have become an affiliate researcher at the Lawrence Berkely National Laboratory in Berkeley, CA, USA to obtain access to the Million Veteran Program (MVP) which aims to genotype and phenotype

current and former US military personnel (Gaziano et al., 2016). Phenotypic information is broad, as in the UK Biobank, and streams of research include behavioral concerns including substance abuse, mental disorders, as well as exposure-specific heterogeneous syndroms incluidng Persian Gulf Syndrome. To facilitate patient stratification, I will ontologize the information from the MVP biobank relevant to mental health but absent from the PNBO, and combine the existing Human Phenotype ontology with the PNBO to stratify patients based on psychological and pysiological traits. Whereas the UK Biobank does not focus on a population at high risk for mental illness such as post-traumatic stress disorder (PTST) or suicidal idiation, the MVP participants will have increased likelihood of poor mental health outcomes compared to the general population who have not been at war in some fashion since at least 2001. Once patients are stratified, GWAS and epistatic testing will be performed, as well as gene/environment interaction studies to take into account specific exposures unique to this population. Accessing such a large biobank will also serve as an independnt replication set for chronotype GWAS performed in this study and those of outher outcomes, from depression to group religious participation. Applying Mendelian Randomization, as discussed in Chapter 3, to these populations will allow for the modeling of causal relationships between genetic succeptability to PTSD and moderating or mediating exposures and quality of life outcomes. Findings will have potential implications for treatment of these and other populations beyond the biobank setting. The Psychiatric GWAS Consortium meta-analizes several psyciatric GWAS studies (Psychiatric GWAS Consortium Steering Committee, 2009), and will additionally provide fruitful ground for patient stratification using more recent studies from the consoritum which include diagnostic and deep phenotyping information. Performing patient stratification at the biobank or meta-analysis scale will ensure that the approches devised in this thesis are exploited to directly benefit basic research into genetic epidemiology and potentially translate into human health. By maintaining collaborations with experimentalists working with mouse models, findings from these wider studies be investiated mechanistically to put computational insights into

biological context.

## 6.5 Conclusions

This thesis aims to formulate models to predict genes involved in neurobehavioral abnormalities. Journeying from mouse to human to back again, it has demonstrated the inherent dependence that human genetics has on model organism work. With colleagues, I predicted and validated circadian functions brought about by the Grp gene and prioritized several others for experimental validation. Moving to human, this work investigated the causal links between chronotype and a series of mental health, social, and opthalmic traits. This work confirmed previously known associations (depression), while proposing novel social and mental traits influenced by circadian genetics, while finding null associations with several mental traits. Knowing that circadian biology influences more aspects of mental health than were modeled using the UK Biobank self assessment, I then modeled autism and schizophrenia related traits using the new Psychological Neuro Behavior Ontology. This ontology was used to develop a new method of segregating autistic probands by endophenotype similarity. Lastly, this work contributed 64 possible GWAS to the literature. Most were insignificant, as expected, but clustering probands led to highly significant results, including potential new links between genes associated with abnormal gait and head nodding in humans and autistic behaviors and chronotype in mouse.

# Appendix One

# Online Appendix

Several supplementary files are available for chapters 2 - 5 can be accessed at http://github.com/jaw-bioinf/PhdThesis:

- Chapter 2: Appendix 1:
- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter2/AppendixTable1.csv
- Chapter 2: Appendix 2:
- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter2/AppendixTable2.csv
- Chapter 3: All mental health trait Mendelian results:
- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter3/all_mental.tsv
- Chapter 3: All social support trait Mendelian results:
- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter3/all_social.tsv
- Chapter 3: All keratometry trait Mendelian results:
- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter3/all_keratometry.tsv
- Chapter 4: The PNBO ontology

- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter4/pnbo.obo
- Chapter 4: The ANBO ontology
- https://github.com/jaw-bioinf/PhdThesis/blob/master/AppendixChapter4/anbo.owl
- Chapter 5: All nominally significant GWAS results from single PNBO trait analyses
- https://github.com/jaw-bioinf/PhdThesis/tree/master/AppendixChapter5

# References


Adams, Charleen D. and Susan L. Neuhausen (2019). "Evaluating causal associations between chronotype and fatty acids and between fatty acids and type 2 diabetes: A Mendelian randomization study". In: *Nutrition, metabolism, and cardiovascular diseases: NMCD* 29.11, pp. 1176–1184. ISSN: 1590-3729. DOI: 10.1016/j.numecd.2019.06.020.

Adzhubei, Ivan A. et al. (Apr. 2010). "A Method and Server for Predicting Damaging Missense Mutations". eng. In: *Nat. Methods* 7.4, pp. 248–249. ISSN: 1548-7105. DOI: 10.1038/nmeth0410-248.

Albrecht, Urs (Apr. 26, 2012). "Timing to perfection: the biology of central and peripheral circadian clocks". In: *Neuron* 74.2, pp. 246–260. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2012.04.006.

Alfaro, Esteban, Matıas Gámez, and Noelia Garcia (2013). "Adabag: An R package for classification with boosting and bagging". In: *Journal of Statistical Software* 54.2, pp. 1–35. URL: https://www.jstatsoft.org/article/view/v054i02/adabag_An_R_Package_for_Classification_with_Boosting_and_Bagging.pdf (visited on 01/06/2017).

Allen, Naomi E. et al. (Feb. 2014a). "UK Biobank Data: Come and Get It". en. In: *Science Translational Medicine* 6.224, 224ed4–224ed4. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.3008601.

Allen, Naomi E. et al. (Feb. 19, 2014b). "UK Biobank Data: Come and Get It". In: *Science Translational Medicine* 6.224. Publisher: American Association for the Advancement of Science Section: Editorial, 224ed4–224ed4. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.3008601. URL: https://stm.sciencemag.org/content/6/224/224ed4 (visited on 07/28/2020).

Aman, M. G. et al. (Mar. 1985). "The Aberrant Behavior Checklist: A Behavior Rating Scale for the Assessment of Treatment Effects". eng. In: *Am J Ment Defic* 89.5, pp. 485–491. ISSN: 0002-9351.

Anafi, Ron C. et al. (Apr. 2014a). "Machine Learning Helps Identify CHRONO as a Circadian Clock Component". en. In: *PLOS Biology* 12.4, e1001840. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001840.

— (Apr. 15, 2014b). "Machine Learning Helps Identify CHRONO as a Circadian Clock Component". In: *PLOS Biology* 12.4, e1001840. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001840. URL: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001840 (visited on 04/17/2018).

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (Jan. 15, 2015). "HTSeq–a Python framework to work with high-throughput sequencing data". In: *Bioinformatics (Oxford, England)* 31.2, pp. 166–169. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu638.

Andews, Simon (2015). "FastQC: A Quality Control tool for High Throughput Sequence Data". Version 0.11.4. In: URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 05/17/2017).

Ashburner, M. et al. (May 2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nature Genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556.

Association, American Psychiatric (May 31, 2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. 5th edition edition. Washington, D.C: American Psychiatric Publishing. 1000 pp. ISBN: 978-0-89042-555-8.

Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium (2017). "Meta-Analysis of GWAS of over 16,000 Individuals with Autism Spectrum Disorder Highlights a Novel Locus at 10q24.32 and a Significant Overlap with Schizophrenia". eng. In: *Mol Autism* 8, p. 21. ISSN: 2040-2392. DOI: 10.1186/s13229-017-0137-9.

Baader, Franz, Ian Horrocks, and Ulrike Sattler (2008). "Chapter 3 Description Logics". en. In: *Foundations of Artificial Intelligence*. Vol. 3. Elsevier, pp. 135–179. ISBN: 978-0-444-52211-5. DOI: 10.1016/S1574-6526(07)03003-9.

Bains, Rasneer S. et al. (2016). "Analysis of Individual Mouse Activity in Group Housed Animals of Different Inbred Strains Using a Novel Automated Home Cage Analysis System". eng. In: *Front Behav Neurosci* 10, p. 106. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2016.00106.

Ballester, Pura et al. (2019). "Sleep Problems in Adults with Autism Spectrum Disorder and Intellectual Disability". en. In: *Autism Research* 12.1, pp. 66–79. ISSN: 1939-3806. DOI: 10.1002/aur.2000.

Barnard, Alun R. and Patrick M. Nolan (May 2008). "When Clocks Go Bad: Neurobehavioural Consequences of Disrupted Circadian Timing". In: *PLOS Genetics* 4.5, e1000040. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000040.

Benjamini, Yoav and Yosef Hochberg (Jan. 1, 1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. ISSN: 0035-9246. URL: http://www.jstor.org/stable/2346101 (visited on 11/05/2015).

Benyamin, Beben, Peter M. Visscher, and Allan F. McRae (Feb. 2009). "Family-Based Genome-Wide Association Studies". eng. In: *Pharmacogenomics* 10.2, pp. 181–190. ISSN: 1744-8042. DOI: 10.2217/14622416.10.2.181.

"Zeitgeber Time" (2009). In: *Encyclopedia of Neuroscience*. Ed. by MarcD. Binder, Nobutaka Hirokawa, and Uwe Windhorst. Springer Berlin Heidelberg, pp. 4399–4399. ISBN: 978-3-540-23735-8. URL: http://dx.doi.org/10.1007/978-3-540-29678-2_6473.

Boudellioua, Imane et al. (Apr. 2017). "Semantic Prioritization of Novel Causative Genomic Variants". eng. In: *PLoS Comput. Biol.* 13.4, e1005500. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005500.

Bourgeois, Florence T. et al. (Dec. 15, 2017). "Development of the Precision Link Biobank at Boston Children's Hospital: Challenges and Opportunities". In: *Journal of Personalized Medicine* 7.4. ISSN: 2075-4426. DOI: 10.3390/jpm7040021.

Bowden, Jack, George Davey Smith, and Stephen Burgess (Apr. 2015). "Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression". eng. In: *Int J Epidemiol* 44.2, pp. 512–525. ISSN: 1464-3685. DOI: 10.1093/ije/dyv080.

Bowden, Jack et al. (May 2016). "Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator". In: *Genetic Epidemiology* 40.4, pp. 304–314. ISSN: 1098-2272. DOI: 10.1002/gepi.21965.

Bowden, Jack et al. (2017). "A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization". In: *Statistics in Medicine* 36.11, pp. 1783–1802. ISSN: 1097-0258. DOI: 10.1002/sim.7221.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard (June 15, 2017). "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7, pp. 1177–1186. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.05.038. URL: http://www.sciencedirect.com/science/article/pii/S0092867417306293 (visited on 07/16/2019).

Bravo-Merodio, Laura et al. (May 14, 2019). "-Omics biomarker identification pipeline for translational medicine". In: *Journal of Translational Medicine* 17.1, p. 155. ISSN: 1479-5876. DOI: 10.1186/s12967-019-1912-5.

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures". In: *Statistical Science* 16.3, pp. 199–215. ISSN: 0883-4237. URL: http://www.jstor.org/stable/2676681 (visited on 05/20/2017).

Brown, Laurence A. et al. (Apr. 24, 2017a). "COMPASS: Continuous Open Mouse Phenotyping of Activity and Sleep Status". In: *Wellcome Open Research* 1. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.9892.2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5140024/ (visited on 07/29/2020).

— (Apr. 2017b). "COMPASS: Continuous Open Mouse Phenotyping of Activity and Sleep Status". In: *Wellcome Open Res* 1. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.9892.2.

Brown, Laurence A. et al. (Sept. 29, 2017c). "Meta-analysis of transcriptomic datasets identifies genes enriched in the mammalian circadian pacemaker". In: *Nucleic Acids Research* 45.17, pp. 9860–9873. ISSN: 0305-1048. DOI: 10.1093/nar/gkx714. URL: https://academic.oup.com/nar/article/45/17/9860/4084660 (visited on 01/22/2018).

Brown, Laurence A. et al. (2020). "Simultaneous assessment of circadian rhythms and sleep in mice using passive infrared sensors: a user guide". In: *Current Protocols in Mouse Biology* In Press.

Brown, Steve D. M. and Mark W. Moore (Oct. 2012a). "The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping". In: *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 23.9, pp. 632–640. ISSN: 1432-1777. DOI: 10.1007/s00335-012-9427-x.

— (May 2012b). "Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium". In: *Disease Models & Mechanisms* 5.3, pp. 289–292. ISSN: 1754-8403. DOI: 10.1242/dmm.009878. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3339821/ (visited on 09/09/2015).

Buijs, R. M. et al. (Mar. 2001). "Parasympathetic and Sympathetic Control of the Pancreas: A Role for the Suprachiasmatic Nucleus and Other Hypothalamic Centers That Are

Involved in the Regulation of Food Intake". eng. In: *The Journal of Comparative Neurology* 431.4, pp. 405–423. ISSN: 0021-9967. DOI: 10.1002/1096-9861(20010319)431:4<405::aid-cne1079>3.0.co;2-d.

Bult, Carol J. et al. (2019a). "Mouse Genome Database (MGD) 2019". In: *Nucleic Acids Research* 47 (D1), pp. D801–D806. ISSN: 1362-4962. DOI: 10.1093/nar/gky1056.

— (Aug. 2019b). "Mouse Genome Database (MGD) 2019". eng. In: *Nucleic Acids Res.* 47.D1, pp. D801–D806. ISSN: 1362-4962. DOI: 10.1093/nar/gky1056.

Burgess, Stephen, Adam Butterworth, and Simon G. Thompson (Nov. 2013). "Mendelian Randomization Analysis with Multiple Genetic Variants Using Summarized Data". eng. In: *Genet. Epidemiol.* 37.7, pp. 658–665. ISSN: 1098-2272. DOI: 10.1002/gepi.21758.

Burgess, Stephen, Frank Dudbridge, and Simon G. Thompson (May 2016). "Combining Information on Multiple Instrumental Variables in Mendelian Randomization: Comparison of Allele Score and Summarized Data Methods". eng. In: *Stat Med* 35.11, pp. 1880–1906. ISSN: 1097-0258. DOI: 10.1002/sim.6835.

Burgess, Stephen, Dylan S Small, and Simon G Thompson (Oct. 2017). "A review of instrumental variable estimators for Mendelian randomization". In: *Statistical Methods in Medical Research* 26.5, pp. 2333–2355. ISSN: 0962-2802. DOI: 10.1177/0962280215597579. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5642006/ (visited on 05/25/2020).

Burgess, Stephen and Simon G. Thompson (2017). "Interpreting findings from Mendelian randomization using the MR-Egger method". In: *European Journal of Epidemiology* 32.5, pp. 377–389. ISSN: 0393-2990. DOI: 10.1007/s10654-017-0255-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506233/ (visited on 06/05/2020).

Burgess, Stephen et al. (July 2015). "Using Published Data in Mendelian Randomization: A Blueprint for Efficient Identification of Causal Risk Factors". en. In: *Eur J Epidemiol* 30.7, pp. 543–552. ISSN: 1573-7284. DOI: 10.1007/s10654-015-0011-z.

Burgess, Stephen et al. (Apr. 28, 2020). "Guidelines for performing Mendelian randomization investigations". In: *Wellcome Open Research* 4, p. 186. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.15555.2. URL: https://wellcomeopenresearch.org/articles/4-186/v2 (visited on 06/27/2020).

Burton, Christopher et al. (Feb. 2013). "Activity Monitoring in Patients with Depression: A Systematic Review". eng. In: *J Affect Disord* 145.1, pp. 21–28. ISSN: 1573-2517. DOI: 10.1016/j.jad.2012.07.001.

Bush, William S. and Jason H. Moore (Dec. 2012). "Chapter 11: Genome-Wide Association Studies". In: *PLoS Comput Biol* 8.12. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1002822.

Bycroft, Clare et al. (2018). "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726, pp. 203–209. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0579-z. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6786975/ (visited on 07/27/2020).

Carmassi, Claudia et al. (June 2019). "Systematic Review of Sleep Disturbances and Circadian Sleep Desynchronization in Autism Spectrum Disorder: Toward an Integrative Model of a Self-Reinforcing Loop". In: *Front Psychiatry* 10. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2019.00366.

Carnagua, Stephen (2015). "RUSBoost - R package". In: *GitHub Repository*. URL: https://github.com/SteveOhh/RUSBoost (visited on 05/08/2017).

Caulfield, Mark et al. (Dec. 2017). *The 100,000 Genomes Project Protocol*. DOI: 10.6084/m9.figshare.4530893.v4.

Ceusters, Werner and Barry Smith (Dec. 2010). "Foundations for a Realist Ontology of Mental Disease". In: *Journal of Biomedical Semantics* 1.1, p. 10. ISSN: 2041-1480. DOI: 10.1186/2041-1480-1-10.

Chang, Christopher C. et al. (Dec. 1, 2015a). "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1. Publisher: Oxford Academic.

DOI: 10.1186/s13742-015-0047-8. URL: https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533 (visited on 04/13/2020).

Chang, Christopher C. et al. (Dec. 2015b). "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets". en. In: *GigaScience* 4.1. DOI: 10.1186/s13742-015-0047-8.

Chaouachi, Anis et al. (Dec. 2009). "Effects of Ramadan Intermittent Fasting on Sports Performance and Training: A Review". eng. In: *Int J Sports Physiol Perform* 4.4, pp. 419–434. ISSN: 1555-0265. DOI: 10.1123/ijspp.4.4.419.

Chaste, Pauline et al. (May 2015). "A Genomewide Association Study of Autism Using the Simons Simplex Collection: Does Reducing Phenotypic Heterogeneity in Autism Increase Genetic Homogeneity?" In: *Biol Psychiatry* 77.9, pp. 775–784. ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2014.09.017.

Chawla, Nitesh V. et al. (June 2002). "SMOTE: Synthetic Minority over-Sampling Technique". In: *Journal of Artificial Intelligence Research* 16.1, pp. 321–357. ISSN: 1076-9757.

Chen, Dongmei, Tao Zhang, and Tae Ho Lee (Aug. 2020). "Cellular Mechanisms of Melatonin: Insight from Neurodegenerative Diseases". In: *Biomolecules* 10.8. ISSN: 2218-273X. DOI: 10.3390/biom10081158.

Chen, Siyu et al. (June 2019). "Angptl8 Mediates Food-Driven Resetting of Hepatic Circadian Clock in Mice". eng. In: *Nature Communications* 10.1, p. 3518. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11513-1.

Chua, Sharon Yu Lin et al. (Feb. 2019). "Cohort Profile: Design and Methods in the Eye and Vision Consortium of UK Biobank". eng. In: *BMJ Open* 9.2, e025077. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2018-025077.

Clarke, Geraldine M et al. (Feb. 2011). "Basic statistical analysis in genetic case-control studies". In: *Nature protocols* 6.2, pp. 121–133. ISSN: 1754-2189. DOI: 10.1038/nprot.

2010.182. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3154648/ (visited on 07/27/2020).

Collaboration (CCGC), C. Reactive Protein Coronary Heart Disease Genetics (Feb. 2011). "Association between C Reactive Protein and Coronary Heart Disease: Mendelian Randomisation Analysis Based on Individual Participant Data". en. In: *BMJ* 342. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.d548.

Collins, Rory (2007). "UK Biobank Protocol". en. In: p. 112.

— (Mar. 31, 2012). "What makes UK Biobank special?" In: *The Lancet* 379.9822. Publisher: Elsevier, pp. 1173–1174. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(12)60404-8. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(12)60404-8/abstract (visited on 07/28/2020).

Costello, E Jane et al. (2004). "Prevalence of psychiatric". In: *Mental health services: A public health perspective*. Publisher: Oxford University Press, USA, p. 111.

Courtot, Mélanie et al. (Jan. 2011). "MIREOT: The Minimum Information to Reference an External Ontology Term". en. In: *Applied Ontology* 6.1, pp. 23–33. ISSN: 1570-5838. DOI: 10.3233/AO-2011-0087.

Crosby, Priya et al. (Feb. 2019). "Insulin/IGF-1 Drives PERIOD Synthesis to Entrain Circadian Rhythms with Feeding Time". eng. In: *Cell* 177.4, 896–909.e20. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.02.017.

Davey Smith, George and Shah Ebrahim (Feb. 2003a). "'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?" en. In: *Int J Epidemiol* 32.1, pp. 1–22. ISSN: 0300-5771. DOI: 10.1093/ije/dyg070.

— (Feb. 1, 2003b). "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" In: *International Journal of Epidemiology* 32.1. Publisher: Oxford Academic, pp. 1–22. ISSN: 0300-5771. DOI:

10.1093/ije/dyg070. URL: https://academic.oup.com/ije/article/32/1/1/642797 (visited on 09/03/2020).

Davey Smith, George et al. (Feb. 2020). "Mendel's Laws, Mendelian Randomization and Causal Inference in Observational Data: Substantive and Nomenclatural Issues". en. In: *Eur J Epidemiol* 35.2, pp. 99–111. ISSN: 1573-7284. DOI: 10.1007/s10654-020-00622-7.

Davies, Neil M., Michael V. Holmes, and George Davey Smith (July 12, 2018). "Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians". In: *BMJ* 362. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting. ISSN: 0959-8138, 1756-1833. DOI: 10.1136/bmj.k601. URL: https://www.bmj.com/content/362/bmj.k601 (visited on 06/14/2020).

DeBruyne, Jason P., David R. Weaver, and Steven M. Reppert (May 2007). "CLOCK and NPAS2 have overlapping roles in the suprachiasmatic circadian clock". In: *Nature Neuroscience* 10.5, pp. 543–545. ISSN: 1097-6256. DOI: 10.1038/nn1884.

Deste, Giacomo et al. (Dec. 2018). "Looking through Autistic Features in Schizophrenia Using the PANSS Autism Severity Score (PAUSS)". eng. In: *Psychiatry Res* 270, pp. 764–768. ISSN: 1872-7123. DOI: 10.1016/j.psychres.2018.10.074.

Dickinson, Mary E. et al. (Sept. 2016). "High-Throughput Discovery of Novel Developmental Phenotypes". eng. In: *Nature* 537.7621, pp. 508–514. ISSN: 1476-4687. DOI: 10.1038/nature19356.

Doi, Masao et al. (Feb. 17, 2016). "Gpr176 is a Gz-linked orphan G-protein-coupled receptor that sets the pace of circadian behaviour". In: *Nature Communications* 7.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–13. ISSN: 2041-1723. DOI: 10.1038/ncomms10583. URL: https://www.nature.com/articles/ncomms10583 (visited on 03/05/2020).

Dudley, Carol A. et al. (July 18, 2003). "Altered patterns of sleep and behavioral adaptability in NPAS2-deficient mice". In: *Science (New York, N.Y.)* 301.5631, pp. 379–383. ISSN: 1095-9203. DOI: 10.1126/science.1082795.

Egger, M. et al. (Sept. 1997). "Bias in Meta-Analysis Detected by a Simple, Graphical Test". eng. In: *BMJ* 315.7109, pp. 629–634. ISSN: 0959-8138. DOI: 10.1136/bmj.315.7109.629.

El-Haschimi, K. and H. Lehnert (Feb. 2003). "Leptin resistance - or why leptin fails to work in obesity". In: *Experimental and Clinical Endocrinology & Diabetes: Official Journal, German Society of Endocrinology [and] German Diabetes Association* 111.1, pp. 2–7. ISSN: 0947-7349. DOI: 10.1055/s-2003-37492.

Fabregat, Antonio et al. (Jan. 4, 2016). "The Reactome pathway Knowledgebase". In: *Nucleic Acids Research* 44 (D1), pp. D481–487. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1351.

Fang, Hai et al. (2016). "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits". In: *Genome Medicine* 8, p. 129. ISSN: 1756-994X. DOI: 10.1186/s13073-016-0384-y. URL: http://dx.doi.org/10.1186/s13073-016-0384-y (visited on 05/18/2017).

Faria, Daniel et al. (2013). "The AgreementMakerLight Ontology Matching System". In: *On the Move to Meaningful Internet Systems: Otm 2013 Conferences*. Ed. by R. Meersman et al. Vol. 8185. WOS:000329655300035. Berlin: Springer-Verlag Berlin, pp. 527–541. ISBN: 978-3-642-41029-1 978-3-642-41030-7.

Fawcett, Tom (June 2006). "An Introduction to ROC Analysis". en. In: *Pattern Recognition Letters* 27.8, pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

Finger, Anna-Marie, Charna Dibner, and Achim Kramer (2020). "Coupled Network of the Circadian Clocks: A Driving Force of Rhythmic Physiology". en. In: *FEBS Letters* 594.17, pp. 2734–2769. ISSN: 1873-3468. DOI: 10.1002/1873-3468.13898.

Fischbach, Gerald D. and Catherine Lord (Oct. 2010a). "The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors". eng. In: *Neuron* 68.2, pp. 192–195. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2010.10.006.

Fischbach, Gerald D. and Catherine Lord (Oct. 21, 2010b). "The Simons Simplex Collection: a resource for identification of autism genetic risk factors". In: *Neuron* 68.2, pp. 192–195. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2010.10.006.

Fisher, Ronald A (1919). "XV.—The correlation between relatives on the supposition of Mendelian inheritance." In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52.2. Publisher: Royal Society of Edinburgh Scotland Foundation, pp. 399–433.

Flicek, Paul et al. (Jan. 2014). "Ensembl 2014". In: *Nucleic Acids Research* 42 (Database issue), pp. D749–755. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1196.

Fox, Caroline S. et al. (2012). "Genome-Wide Association for Abdominal Subcutaneous and Visceral Adipose Reveals a Novel Locus for Visceral Fat in Women". eng. In: *PLoS Genet.* 8.5, e1002695. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002695.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

Garcia, J. A. et al. (June 23, 2000). "Impaired cued and contextual memory in NPAS2-deficient mice". In: *Science (New York, N.Y.)* 288.5474, pp. 2226–2230. ISSN: 0036-8075.

Gary, K. A. et al. (July 1996). "Thyrotropin-releasing hormone phase shifts circadian rhythms in hamsters." In: *Neuroreport* 7.10, pp. 1631–1634. ISSN: 0959-4965. URL: http://europepmc.org/abstract/med/8904771 (visited on 04/17/2018).

Gary, Keith A. et al. (May 2003a). "The Thyrotropin-Releasing Hormone (TRH) Hypothesis of Homeostatic Regulation: Implications for TRH-Based Therapeutics". eng. In: *J. Pharmacol. Exp. Ther.* 305.2, pp. 410–416. ISSN: 0022-3565. DOI: 10.1124/jpet.102.044040.

— (May 2003b). "The thyrotropin-releasing hormone (TRH) hypothesis of homeostatic regulation: implications for TRH-based therapeutics". In: *The Journal of Pharma-*

*cology and Experimental Therapeutics* 305.2, pp. 410–416. ISSN: 0022-3565. DOI: 10.1124/jpet.102.044040.

Gaziano, John Michael et al. (Feb. 2016). "Million Veteran Program: A mega-biobank to study genetic influences on health and disease". In: *Journal of Clinical Epidemiology* 70, pp. 214–223. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2015.09.016.

Gene Ontology Consortium (Jan. 2015). "Gene Ontology Consortium: going forward". In: *Nucleic Acids Research* 43 (Database issue), pp. D1049–1056. ISSN: 1362-4962. DOI: 10.1093/nar/gku1179.

Geschwind, Daniel H. et al. (Aug. 2001). "The Autism Genetic Resource Exchange: A Resource for the Study of Autism and Related Neuropsychiatric Conditions". In: *Am J Hum Genet* 69.2, pp. 463–466. ISSN: 0002-9297.

Gibson, Mark et al. (2019). "Evidence for Genetic Correlations and Bidirectional, Causal Effects Between Smoking and Sleep Behaviors". In: *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco* 21.6, pp. 731–738. ISSN: 1469-994X. DOI: 10.1093/ntr/nty230.

Gkoutos, Georgios V., Paul N. Schofield, and Robert Hoehndorf (2012). "The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes". In: *International Review of Neurobiology* 103, pp. 69–87. ISSN: 2162-5514. DOI: 10.1016/B978-0-12-388408-4.00004-6.

Gkoutos, Georgios V. et al. (2009). "Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO". In: *Conference Proceedings* 2009, pp. 7069–7072. ISSN: 1557-170X. DOI: 10.1109/IEMBS.2009.5333362. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398700/ (visited on 05/24/2017).

Goda, Tadahiro et al. (Feb. 2018a). "Calcitonin Receptors Are Ancient Modulators for Rhythms of Preferential Temperature in Insects and Body Temperature in Mammals". en. In: *Genes & Development*. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.307884.117.

Goda, Tadahiro et al. (Feb. 12, 2018b). "Calcitonin receptors are ancient modulators for rhythms of preferential temperature in insects and body temperature in mammals". In: *Genes & Development*. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.307884.117. URL: http://genesdev.cshlp.org/content/early/2018/02/12/gad.307884.117 (visited on 04/12/2018).

Gómez-Pérez, Asunción (2001). "Evaluation of Ontologies". en. In: *International Journal of Intelligent Systems* 16.3, pp. 391–409. ISSN: 1098-111X. DOI: 10.1002/1098-111X(200103)16:3<391::AID-INT1014>3.0.CO;2-2.

Gottesman, Irving I. and Todd D. Gould (Apr. 1, 2003). "The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions". In: *American Journal of Psychiatry* 160.4, pp. 636–645. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.160.4.636. URL: http://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.160.4.636 (visited on 05/21/2017).

Greene, Daniel, Sylvia Richardson, and Ernest Turro (Mar. 3, 2016). "Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases". In: *American Journal of Human Genetics* 98.3, pp. 490–499. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2016.01.008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827100/ (visited on 06/20/2019).

— (Jan. 5, 2017). "ontologyX: a suite of R packages for working with ontological data". In: *Bioinformatics (Oxford, England)*. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw763.

Grove, Jakob et al. (Mar. 2019). "Identification of Common Genetic Risk Variants for Autism Spectrum Disorder". en. In: *Nature Genetics* 51.3, pp. 431–444. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0344-8.

Guo, Hongnian et al. (Feb. 2005). "Differential Control of Peripheral Circadian Rhythms by Suprachiasmatic-Dependent Neural Signals". eng. In: *Proceedings of the National*

*Academy of Sciences of the United States of America* 102.8, pp. 3111–3116. ISSN: 0027-8424. DOI: 10.1073/pnas.0409734102.

Hamada, Toshiyuki, Michael C. Antle, and Rae Silver (2004). "Temporal and Spatial Expression Patterns of Canonical Clock Genes and Clock-Controlled Genes in the Suprachiasmatic Nucleus". en. In: *European Journal of Neuroscience* 19.7, pp. 1741–1748. ISSN: 1460-9568. DOI: 10.1111/j.1460-9568.2004.03275.x.

Hampson, David R. and Gene J. Blatt (2015). "Autism Spectrum Disorders and Neuropathology of the Cerebellum". eng. In: *Front Neurosci* 9, p. 420. ISSN: 1662-4548. DOI: 10.3389/fnins.2015.00420.

Hannon, Eilis et al. (Mar. 2018). "Elevated Polygenic Burden for Autism Is Associated with Differential DNA Methylation at Birth". In: *Genome Medicine* 10.1, p. 19. ISSN: 1756-994X. DOI: 10.1186/s13073-018-0527-4.

Hartwig, Fernando Pires, George Davey Smith, and Jack Bowden (Dec. 2017). "Robust Inference in Summary Data Mendelian Randomization via the Zero Modal Pleiotropy Assumption". In: *Int J Epidemiol* 46.6, pp. 1985–1998. ISSN: 0300-5771. DOI: 10.1093/ije/dyx102.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84857-0 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7. URL: http://link.springer.com/10.1007/978-0-387-84858-7 (visited on 05/17/2017).

Hastings, Michael H and Michel Goedert (Oct. 2013). "Circadian clocks and neurodegenerative diseases: time to aggregate?" In: *Current Opinion in Neurobiology* 23.5, pp. 880–887. ISSN: 0959-4388. DOI: 10.1016/j.conb.2013.05.004. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3782660/ (visited on 05/18/2017).

Hatcher, J. P. et al. (Nov. 2001). "Development of SHIRPA to Characterise the Phenotype of Gene-Targeted Mice". eng. In: *Behav. Brain Res.* 125.1-2, pp. 43–47. ISSN: 0166-4328.

Hemani, Gibran, Jack Bowden, and George Davey Smith (Aug. 1, 2018). "Evaluating the potential role of pleiotropy in Mendelian randomization studies". In: *Human Molecular Genetics* 27 (R2), R195–R208. ISSN: 0964-6906. DOI: 10.1093/hmg/ddy163. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6061876/ (visited on 07/29/2020).

Hemani, Gibran, Kate Tilling, and George Davey Smith (Nov. 2017). "Orienting the Causal Relationship between Imprecisely Measured Traits Using GWAS Summary Data". eng. In: *PLoS Genet.* 13.11, e1007081. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007081.

Hemani, Gibran et al. (May 30, 2018). "The MR-Base platform supports systematic causal inference across the human phenome". In: *eLife* 7. Ed. by Ruth Loos. Publisher: eLife Sciences Publications, Ltd, e34408. ISSN: 2050-084X. DOI: 10.7554/eLife.34408. URL: https://doi.org/10.7554/eLife.34408 (visited on 07/20/2020).

"Cochrane Handbook for Systematic Reviews of Interventions" (2011). In: ed. by JPT Higgins and S Green. URL: http://www.cochrane-handbook.org (visited on 12/02/2013).

Hoehndorf, Robert, Michel Dumontier, and Georgios V. Gkoutos (Nov. 2013). "Evaluation of Research in Biomedical Ontologies". In: *Brief Bioinform* 14.6, pp. 696–712. ISSN: 1467-5463. DOI: 10.1093/bib/bbs053.

Hoehndorf, Robert, Paul N. Schofield, and Georgios V. Gkoutos (Nov. 2015). "The role of ontologies in biological and biomedical research: a functional perspective". In: *Briefings in Bioinformatics* 16.6, pp. 1069–1080. ISSN: 1467-5463. DOI: 10.1093/bib/bbv011. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4652617/ (visited on 05/24/2017).

Hoehndorf, Robert et al. (Jan. 28, 2015). "Aber-OWL: a framework for ontology-based data access in biology". In: *BMC Bioinformatics* 16.1. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0456-9. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384359/ (visited on 05/11/2016).

Horrocks, Ian et al. (2004). "SWRL: A semantic web rule language combining OWL and RuleML". In: *W3C Member submission* 21.79, pp. 1–31.

Howey, R. "CASSI". Version 2.51. In: (). URL: http://www.staff.ncl.ac.uk/richard.howey/cassi/.

Howie, Bryan, Jonathan Marchini, and Matthew Stephens (Nov. 2011). "Genotype Imputation with Thousands of Genomes". eng. In: *G3 (Bethesda, Md.)* 1.6, pp. 457–470. ISSN: 2160-1836. DOI: 10.1534/g3.111.001198.

Hu, Valerie W. et al. (Apr. 2009). "Gene Expression Profiling Differentiates Autism Case-Controls and Phenotypic Variants of Autism Spectrum Disorders: Evidence for Circadian Rhythm Dysfunction in Severe Autism". eng. In: *Autism Res* 2.2, pp. 78–97. ISSN: 1939-3806. DOI: 10.1002/aur.73.

Hu, Youna et al. (Feb. 2, 2016). "GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person". In: *Nature Communications* 7.1. Number: 1 Publisher: Nature Publishing Group, p. 10448. ISSN: 2041-1723. DOI: 10.1038/ncomms10448. URL: https://www.nature.com/articles/ncomms10448 (visited on 05/27/2020).

Huang, Jie et al. (Sept. 2015). "Improved Imputation of Low-Frequency and Rare Variants Using the UK10K Haplotype Reference Panel". eng. In: *Nature Communications* 6, p. 8111. ISSN: 2041-1723. DOI: 10.1038/ncomms9111.

Hughes, Michael E., John B. Hogenesch, and Karl Kornacker (Oct. 2010a). "JTK_CYCLE: An Efficient Non-Parametric Algorithm for Detecting Rhythmic Components in Genome-Scale Datasets". In: *J Biol Rhythms* 25.5, pp. 372–380. ISSN: 0748-7304. DOI: 10.1177/0748730410379711.

— (Oct. 2010b). "JTK_CYCLE: an efficient non-parametric algorithm for detecting rhythmic components in genome-scale datasets". In: *Journal of biological rhythms* 25.5, pp. 372–380. ISSN: 0748-7304. DOI: 10.1177/0748730410379711. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3119870/ (visited on 09/22/2015).

Hughes, Michael E. et al. (Oct. 2017). "Guidelines for Genome-Scale Analysis of Biological Rhythms". In: *J Biol Rhythms* 32.5, pp. 380–393. ISSN: 0748-7304. DOI: 10.1177/0748730417728663.

Hughes, Steven et al. (Jan. 2015). "Chapter Six - Photic Regulation of Clock Systems". en. In: *Methods in Enzymology*. Ed. by Amita Sehgal. Vol. 552. Circadian Rhythms and Biological Clocks, Part B. Academic Press, pp. 125–143. DOI: 10.1016/bs.mie.2014.10.018.

Iftimovici, Anton et al. (2020). "Stress, Cortisol and NR3C1 in At-Risk Individuals for Psychosis: A Mendelian Randomization Study". eng. In: *Front Psychiatry* 11, p. 680. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2020.00680.

Jagannath, Aarti et al. (Aug. 29, 2013). "The CRTC1-SIK1 pathway regulates entrainment of the circadian clock". In: *Cell* 154.5, pp. 1100–1111. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.08.004.

Jones, Allan R., Caroline C. Overly, and Susan M. Sunkin (Nov. 2009). "The Allen Brain Atlas: 5 years and beyond". In: *Nature Reviews Neuroscience* 10.11, pp. 821–828. ISSN: 1471-003X. DOI: 10.1038/nrn2722. URL: http://www.nature.com/nrn/journal/v10/n11/full/nrn2722.html?cookies=accepted (visited on 11/30/2016).

Jones, Rachel Maree et al. (Oct. 2013). "Genome-Wide Association Study of Autistic-Like Traits in a General Population Study of Young Adults". In: *Front Hum Neurosci* 7. ISSN: 1662-5161. DOI: 10.3389/fnhum.2013.00658.

Jones, Samuel E. et al. (Jan. 29, 2019). "Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms". In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 343. ISSN: 2041-1723. DOI: 10.1038/s41467-018-08259-7. URL: https://www.nature.com/articles/s41467-018-08259-7 (visited on 06/14/2020).

Jones, Stephanie G. and Ruth M. Benca (Dec. 2015). "Circadian Disruption in Psychiatric Disorders". In: *Sleep Medicine Clinics*. Science of Circadian Rhythms 10.4, pp. 481–493. ISSN: 1556-407X. DOI: 10.1016/j.jsmc.2015.07.004.

Kanehisa, M. and S. Goto (Jan. 1, 2000). "KEGG: kyoto encyclopedia of genes and genomes". In: *Nucleic Acids Research* 28.1, pp. 27–30. ISSN: 0305-1048.

Kang, Hyunseung et al. (Sept. 21, 2014). "Instrumental Variables Estimation with Some Invalid Instruments and its Application to Mendelian Randomization". In: *arXiv:1401.5755 [stat]*. arXiv: 1401.5755. URL: http://arxiv.org/abs/1401.5755 (visited on 07/20/2020).

Karatsoreos, Ilia N. (May 6, 2014a). "Links between Circadian Rhythms and Psychiatric Disease". In: *Frontiers in Behavioral Neuroscience* 8. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2014.00162. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4018537/ (visited on 01/18/2017).

— (May 2014b). "Links between Circadian Rhythms and Psychiatric Disease". In: *Front Behav Neurosci* 8. ISSN: 1662-5153. DOI: 10.3389/fnbeh.2014.00162.

Kästner, Anne et al. (May 2015). "Autism beyond Diagnostic Categories: Characterization of Autistic Phenotypes in Schizophrenia". In: *BMC Psychiatry* 15. ISSN: 1471-244X. DOI: 10.1186/s12888-015-0494-x.

Katan, M. B. (Feb. 2004). "Apolipoprotein E Isoforms, Serum Cholesterol, and Cancer. 1986". eng. In: *Int J Epidemiol* 33.1, p. 9. ISSN: 0300-5771. DOI: 10.1093/ije/dyh312.

Kay, S. R., A. Fiszbein, and L. A. Opler (1987a). "The positive and negative syndrome scale (PANSS) for schizophrenia". In: *Schizophrenia Bulletin* 13.2, pp. 261–276. ISSN: 0586-7614. DOI: 10.1093/schbul/13.2.261.

— (1987b). "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". eng. In: *Schizophr Bull* 13.2, pp. 261–276. ISSN: 0586-7614. DOI: 10.1093/schbul/13.2.261.

Kazakov, Yevgeny, Markus Krötzsch, and František Simančík (June 2014). "The Incredible ELK: From Polynomial Procedures to Efficient Reasoning with E $\beta$ $\mathcal {E}

\mathcal {L}$ Ontologies". In: *Journal of Automated Reasoning* 53.1, pp. 1–61. ISSN: 0168-7433, 1573-0670. DOI: 10.1007/s10817-013-9296-3. URL: http://link.springer.com/10.1007/s10817-013-9296-3 (visited on 04/07/2019).

Kishi, T. et al. (Apr. 2011). "SIRT1 Gene, Schizophrenia and Bipolar Disorder in the Japanese Population: An Association Study". eng. In: *Genes Brain Behav.* 10.3, pp. 257–263. ISSN: 1601-183X. DOI: 10.1111/j.1601-183X.2010.00661.x.

Kohavi, Ron and Foster Provost (Feb. 1998). "Glossary of Terms". In: *Machine Learning* 30.2, pp. 271–274. ISSN: 0885-6125. URL: http://dl.acm.org/citation.cfm?id=288808.288815 (visited on 05/20/2017).

Kolberg, Liis and Uku Raudvere (2020). *Gprofiler2: Interface to the 'g:Profiler' Toolset.* Manual. R package version 0.1.9.

Kolvin, I. (Apr. 1971). "Studies in the Childhood Psychoses. I. Diagnostic Criteria and Classification". eng. In: *Br J Psychiatry* 118.545, pp. 381–384. ISSN: 0007-1250. DOI: 10.1192/bjp.118.545.381.

Kondor, Risi Imre and John Lafferty (2002). "Diffusion kernels on graphs and other discrete input spaces". In: *ICML*. Vol. 2, pp. 315–322. URL: https://pdfs.semanticscholar.org/6320/770fe216ebbba769b9f0a006669b616a03d0.pdf (visited on 02/08/2017).

Kryuchkova-Mostacci, Nadezda and Marc Robinson-Rechavi (Mar. 1, 2017). "A benchmark of gene expression tissue-specificity metrics". In: *Briefings in Bioinformatics* 18.2. Publisher: Oxford Academic, pp. 205–214. ISSN: 1467-5463. DOI: 10.1093/bib/bbw008. URL: https://academic.oup.com/bib/article/18/2/205/2562739 (visited on 07/29/2020).

Köhler, Sebastian et al. (Jan. 8, 2019). "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources". In: *Nucleic Acids Research* 47 (D1), pp. D1018–D1027. ISSN: 0305-1048. DOI: 10.1093/nar/gky1105. URL: https://academic.oup.com/nar/article/47/D1/D1018/5198478 (visited on 12/20/2019).

Labrecque, Jeremy and Sonja A. Swanson (2018). "Understanding the Assumptions Underlying Instrumental Variable Analyses: A Brief Review of Falsification Strategies and Related Tools". eng. In: *Curr Epidemiol Rep* 5.3, pp. 214–220. ISSN: 2196-2995. DOI: 10.1007/s40471-018-0152-1.

Laird, Nan M. and Christoph Lange (May 2006). "Family-Based Designs in the Age of Large-Scale Gene-Association Studies". eng. In: *Nat. Rev. Genet.* 7.5, pp. 385–394. ISSN: 1471-0056. DOI: 10.1038/nrg1839.

Lander, Eric S. et al. (Feb. 15, 2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062. URL: http://www.nature.com/nature/journal/v409/n6822/abs/409860a0.html (visited on 09/09/2015).

Landgraf, Dominic, Anne-Marie Neumann, and Henrik Oster (Dec. 2017). "Circadian Clock-Gastrointestinal Peptide Interaction in Peripheral Tissues and the Brain". eng. In: *Best Practice & Research. Clinical Endocrinology & Metabolism* 31.6, pp. 561–571. ISSN: 1878-1594. DOI: 10.1016/j.beem.2017.10.007.

Landgraf, Dominic et al. (Mar. 2015). "Oxyntomodulin Regulates Resetting of the Liver Circadian Clock by Food". eng. In: *eLife* 4, e06253. ISSN: 2050-084X. DOI: 10.7554/eLife.06253.

Landsteiner, Karl (1901). "Ueber Agglutinationserscheinungen normalen mensclinchen Blutes". In: *Wiener Klinische Wochenschrift* 14, pp. 1132–1134.

— (Jan. 2, 1961). "On Agglutination of Normal Human Blood". In: *Transfusion* 1.1, pp. 5–8. ISSN: 1537-2995. DOI: 10.1111/j.1537-2995.1961.tb00005.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1537-2995.1961.tb00005.x/abstract (visited on 02/07/2018).

Lane, Jacqueline M. et al. (Mar. 9, 2016). "Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank". In: *Nature Communications* 7, p. 10889. ISSN: 2041-1723. DOI: 10.1038/ncomms10889.

Langfelder, Peter and Steve Horvath (Dec. 29, 2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9, p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559. URL: https://doi.org/10.1186/1471-2105-9-559 (visited on 07/11/2018).

Langfelder, Peter, Bin Zhang, and Steve Horvath (Mar. 2008). "Defining Clusters from a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R". en. In: *Bioinformatics* 24.5, pp. 719–720. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm563.

Langmead, Ben and Steven L. Salzberg (Apr. 2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4, pp. 357–359. ISSN: 1548-7091. DOI: 10.1038/nmeth.1923. URL: http://www.nature.com.ezp-prod1.hul.harvard.edu/nmeth/journal/v9/n4/full/nmeth.1923.html (visited on 06/26/2015).

Lawlor, Debbie A. et al. (2008). "Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology". en. In: *Statistics in Medicine* 27.8, pp. 1133–1163. ISSN: 1097-0258. DOI: 10.1002/sim.3034.

Lessan, Nader and Tomader Ali (May 2019). "Energy Metabolism and Intermittent Fasting: The Ramadan Perspective". eng. In: *Nutrients* 11.5. ISSN: 2072-6643. DOI: 10.3390/nu11051192.

Levitis, Daniel A., William Z. Lidicker, and Glenn Freund (July 2009). "Behavioural biologists do not agree on what constitutes behaviour". In: *Animal Behaviour* 78.1, pp. 103–110. ISSN: 00033472. DOI: 10.1016/j.anbehav.2009.03.018. URL: http://linkinghub.elsevier.com/retrieve/pii/S0003347209001730 (visited on 05/21/2017).

Lexow, Nedra (Jan. 1, 1996). "Localization and expression of thyrotropin-releasing hormone in rat and cat retina". In: *Dissertations available from ProQuest*, pp. 1–127. URL: https://repository.upenn.edu/dissertations/AAI9712962.

Li, Marilyn M. et al. (Nov. 1, 2015). "A multicenter, cross-platform clinical validation study of cancer cytogenomic arrays". In: *Cancer Genetics* 208.11, pp. 525–536. ISSN: 2210-7762. DOI: 10.1016/j.cancergen.2015.08.002.

Li, Ran et al. (Oct. 2008). "CLOCK/BMAL1 Regulates Human Nocturnin Transcription through Binding to the E-Box of Nocturnin Promoter". eng. In: *Mol. Cell. Biochem.* 317.1-2, pp. 169–177. ISSN: 0300-8177. DOI: 10.1007/s11010-008-9846-x.

Li, Yuying et al. (July 13, 2020). "Circadian Rhythms and Obesity: Timekeeping Governs Lipid Metabolism". In: *Journal of Pineal Research*, e12682. ISSN: 1600-079X. DOI: 10.1111/jpi.12682.

Lin, Dekang (1998). "An Information-Theoretic Definition of Similarity". In: *In Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, pp. 296–304.

Lind, Mackenzie J. et al. (Apr. 15, 2020). "Molecular genetic overlap between posttraumatic stress disorder and sleep phenotypes". In: *Sleep* 43.4. ISSN: 1550-9109. DOI: 10.1093/sleep/zsz257.

Liu, Zheng et al. (Apr. 16, 2007). "Study of gene function based on spatial co-expression in a high-resolution mouse brain atlas". In: *BMC Systems Biology* 1, p. 19. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-19. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1863433/ (visited on 11/30/2016).

Lomb, N. R. (Feb. 1, 1976). "Least-squares frequency analysis of unequally spaced data". In: *Astrophysics and Space Science* 39.2, pp. 447–462. ISSN: 1572-946X. DOI: 10.1007/BF00648343. URL: https://doi.org/10.1007/BF00648343 (visited on 03/10/2020).

Lord, C., M. Rutter, and A. Le Couteur (Oct. 1994a). "Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders". In: *Journal of Autism and Developmental Disorders* 24.5, pp. 659–685. ISSN: 0162-3257. DOI: 10.1007/BF02172145.

— (Oct. 1994b). "Autism Diagnostic Interview-Revised: A Revised Version of a Diagnostic Interview for Caregivers of Individuals with Possible Pervasive Developmental Disorders". eng. In: *J Autism Dev Disord* 24.5, pp. 659–685. ISSN: 0162-3257. DOI: 10.1007/BF02172145.

Lord, C. et al. (June 1989). "Autism Diagnostic Observation Schedule: A Standardized Observation of Communicative and Social Behavior". eng. In: *J Autism Dev Disord* 19.2, pp. 185–212. ISSN: 0162-3257. DOI: 10.1007/BF02211841.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. ISSN: 1465-6914. DOI: 10.1186/s13059-014-0550-8.

MacArthur, Jacqueline et al. (Jan. 2017). "The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog)". In: *Nucleic Acids Research* 45.D1, pp. D896–D901. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1133.

Malsen, Annetrude J. G. de Mooij-van et al. (Aug. 1, 2011). "Cross-species behavioural genetics: A starting point for unravelling the neurobiology of human psychiatric disorders". In: *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 35.6, pp. 1383–1390. ISSN: 1878-4216. DOI: 10.1016/j.pnpbp.2010.10.003.

Manaker, S. et al. (Jan. 1, 1985). "Autoradiographic localization of thyrotropin-releasing hormone receptors in the rat central nervous system". In: *Journal of Neuroscience* 5.1, pp. 167–174. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.05-01-00167.1985. URL: http://www.jneurosci.org/content/5/1/167 (visited on 04/17/2018).

Mandillo, Silvia et al. (Aug. 2008). "Reliability, Robustness, and Reproducibility in Mouse Behavioral Phenotyping: A Cross-Laboratory Study". eng. In: *Physiol. Genomics* 34.3, pp. 243–255. ISSN: 1531-2267. DOI: 10.1152/physiolgenomics.90207.2008.

Manning, Hannah, Brian J. O'Roak, and Özgün Babur (May 2019). "Mutually Exclusive Autism Mutations Point to the Circadian Clock and PI3K Signaling Pathways". en. In: *bioRxiv*, p. 653527. DOI: 10.1101/653527.

Mansour, Hader A. et al. (Nov. 2009). "Association Study of 21 Circadian Genes with Bipolar I Disorder, Schizoaffective Disorder, and Schizophrenia". eng. In: *Bipolar Disord* 11.7, pp. 701–710. ISSN: 1399-5618. DOI: 10.1111/j.1399-5618.2009.00756.x.

Martínez-Santiago, Fernando et al. (Jan. 2020). "Aging Neuro-Behavior Ontology". en. In: *Applied Ontology* 15.2, pp. 219–239. ISSN: 1570-5838. DOI: 10.3233/AO-200229.

Martínez-Santiago, Fernando et al. (Jan. 1, 2020). "Aging Neuro-Behavior Ontology". In: *Applied Ontology* 15.2. Publisher: IOS Press, pp. 219–239. ISSN: 1570-5838. DOI: 10.3233/AO-200229. URL: https://content.iospress.com/articles/applied-ontology/ao200229 (visited on 08/06/2020).

Matentzoglu, Nicolas et al. (Jan. 18, 2018). "MIRO: guidelines for minimum information for the reporting of an ontology". In: *Journal of Biomedical Semantics* 9.1, p. 6. ISSN: 2041-1480. DOI: 10.1186/s13326-017-0172-7. URL: https://doi.org/10.1186/s13326-017-0172-7 (visited on 08/07/2020).

McArthur, A. J. et al. (July 15, 2000). "Gastrin-releasing peptide phase-shifts suprachiasmatic nuclei neuronal rhythms in vitro". In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 20.14, pp. 5496–5502. ISSN: 0270-6474.

McCray, Alexa T., Philip Trevvett, and H. Robert Frost (Apr. 2014). "Modeling the autism spectrum disorder phenotype". In: *Neuroinformatics* 12.2, pp. 291–305. ISSN: 1559-0089. DOI: 10.1007/s12021-013-9211-4.

McLaren, William et al. (2016). "The Ensembl Variant Effect Predictor". In: *Genome Biology* 17.1, p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

McLaren, William et al. (June 2016). "The Ensembl Variant Effect Predictor". eng. In: *Genome Biol.* 17.1, p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

McShane, Blakeley B. et al. (Mar. 2019). "Abandon Statistical Significance". In: *The American Statistician* 73.sup1, pp. 235–245. ISSN: 0003-1305. DOI: 10.1080/00031305.2018.1527253.

Melo, Matias C. A. et al. (Aug. 2017). "Chronotype and Circadian Rhythm in Bipolar Disorder: A Systematic Review". eng. In: *Sleep Med Rev* 34, pp. 46–58. ISSN: 1532-2955. DOI: 10.1016/j.smrv.2016.06.007.

Merikangas, Kathleen R. and Alison K. Merikangas (Jan. 2016). "Chapter 2 - Contribution of Genetic Epidemiology to Our Understanding of Psychiatric Disorders". en. In: *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*. Ed. by Thomas Lehner, Bruce L. Miller, and Matthew W. State. San Diego: Academic Press, pp. 27–50. ISBN: 978-0-12-800105-9. DOI: 10.1016/B978-0-12-800105-9.00002-0.

Merikangas, Kathleen R. et al. (Nov. 1998). "Familial Transmission of Substance Use Disorders". en. In: *Arch Gen Psychiatry* 55.11, pp. 973–979. ISSN: 0003-990X. DOI: 10.1001/archpsyc.55.11.973.

Merritt, Marcellus M. and T. J. McCallum (Jan. 2013). "Too Much of a Good Thing?: Positive Religious Coping Predicts Worse Diurnal Salivary Cortisol Patterns for Overwhelmed African American Female Dementia Family Caregivers". eng. In: *Am J Geriatr Psychiatry* 21.1, pp. 46–56. ISSN: 1545-7214. DOI: 10.1016/j.jagp.2012.10.006.

Mieda, Michihiro (2019). "The Network Mechanism of the Central Circadian Pacemaker of the SCN: Do AVP Neurons Play a More Critical Role Than Expected?" English. In: *Frontiers in Neuroscience* 13. ISSN: 1662-453X. DOI: 10.3389/fnins.2019.00139.

Moon, Joung-Ho et al. (Sept. 2016). "Advanced Circadian Phase in Mania and Delayed Circadian Phase in Mixed Mania and Depression Returned to Normal after Treatment of Bipolar Disorder". English. In: *EBioMedicine* 11, pp. 285–295. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2016.08.019.

Mortazavi, Ali et al. (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. URL: http://www.nature.com/nmeth/journal/v5/n7/abs/nmeth.1226.html (visited on 09/13/2015).

Mouse ENCODE Consortium et al. (2012). "An encyclopedia of mouse DNA elements (Mouse ENCODE)". In: *Genome Biology* 13.8, p. 418. ISSN: 1465-6914. DOI: 10.1186/gb-2012-13-8-418.

Mugzach, Omri et al. (Aug. 2015). "An ontology for Autism Spectrum Disorder (ASD) to infer ASD phenotypes from Autism Diagnostic Interview–Revised data". In: *Journal of biomedical informatics* 56, pp. 333–347. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2015.06.026. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4532604/ (visited on 08/06/2020).

Mungall, Christopher J. et al. (Jan. 31, 2012). "Uberon, an integrative multi-species anatomy ontology". In: *Genome Biology* 13.1, R5. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-1-r5. URL: https://doi.org/10.1186/gb-2012-13-1-r5 (visited on 04/07/2019).

Murray, C, A Lopez, and others (2002). "World Health Report 2002: reducing risks, promoting healthy life". In: *Geneva: World Health Organization*, p. 186.

Musen, Mark A. (June 2015). "The Protégé Project: A Look Back and a Look Forward". In: *AI matters* 1.4, pp. 4–12. ISSN: 2372-3483. DOI: 10.1145/2757001.2757003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883684/ (visited on 04/07/2019).

Nishizawa, D et al. (Jan. 2014). "Genome-Wide Association Study Identifies a Potent Locus Associated with Human Opioid Sensitivity". In: *Mol Psychiatry* 19.1, pp. 55–62. ISSN: 1359-4184. DOI: 10.1038/mp.2012.164.

Oliver, Peter L. et al. (Feb. 2012a). "Disrupted Circadian Rhythms in a Mouse Model of Schizophrenia". In: *Curr Biol* 22.4, pp. 314–319. ISSN: 0960-9822. DOI: 10.1016/j.cub.2011.12.051.

Oliver, Peter L. et al. (Feb. 21, 2012b). "Disrupted Circadian Rhythms in a Mouse Model of Schizophrenia". In: *Current Biology* 22.4, pp. 314–319. ISSN: 0960-9822. DOI: 10.1016/j.cub.2011.12.051. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3356578/ (visited on 04/16/2018).

Pan, Xiaoyue, Samantha Mota, and Boyang Zhang (2020). "Circadian Clock Regulation on Lipid Metabolism and Metabolic Diseases". In: *Advances in Experimental Medicine and Biology* 1276, pp. 53–66. ISSN: 0065-2598. DOI: 10.1007/978-981-15-6082-8_5.

Parsons, Michael J. et al. (July 30, 2015). "The Regulatory Factor ZFHX3 Modifies Circadian Function in SCN via an AT Motif-Driven Axis". In: *Cell* 162.3, pp. 607–621. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.06.060. URL: http://www.sciencedirect.com/science/article/pii/S0092867415008302 (visited on 01/21/2016).

Patke, Alina, Michael W. Young, and Sofia Axelrod (Feb. 2020). "Molecular Mechanisms and Physiological Importance of Circadian Rhythms". en. In: *Nature Reviews Molecular Cell Biology* 21.2, pp. 67–84. ISSN: 1471-0080. DOI: 10.1038/s41580-019-0179-2.

Patron, Jonas et al. (Dec. 2019). "Assessing the Performance of Genome-Wide Association Studies for Predicting Disease Risk". en. In: *PLOS ONE* 14.12, e0220215. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0220215.

Pearl, Judea (1995). "Causal Diagrams for Empirical Research". In: *Biometrika* 82.4, pp. 669–688. ISSN: 0006-3444. DOI: 10.2307/2337329.

Pedersen, Thomas Lin and Michaël Benesty (2019). *lime: Local interpretable model-agnostic explanations*. URL: https://CRAN.R-project.org/package=lime.

Pembroke, William G. et al. (Nov. 2, 2015). "Temporal transcriptomics suggest that twin-peaking genes reset the clock". In: *eLife*, e10518. ISSN: 2050-084X. DOI: 10.7554/eLife.10518. URL: http://elifesciences.org/content/early/2015/11/02/eLife.10518 (visited on 01/21/2016).

Pizarro, Angel et al. (Jan. 2013). "CircaDB: a database of mammalian circadian gene expression profiles". In: *Nucleic Acids Research* 41 (Database issue), pp. D1009–1013. ISSN: 1362-4962. DOI: 10.1093/nar/gks1161.

Psychiatric GWAS Consortium Steering Committee (Jan. 2009). "A Framework for Interpreting Genome-Wide Association Studies of Psychiatric Disorders". eng. In: *Molecular Psychiatry* 14.1, pp. 10–17. ISSN: 1476-5578. DOI: 10.1038/mp.2008.126.

R Core Team (2013). "R: A language and environment for statistical computing." Version 3.0.2. In: URL: http://www.R-project.org/.

Radivojac, Predrag et al. (Mar. 2013). “A large-scale evaluation of computational protein function prediction”. In: *Nature methods* 10.3, pp. 221–227. ISSN: 1548-7091. DOI: 10.1038/nmeth.2340. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3584181/ (visited on 04/17/2018).

Reimand, Jüri, Tambet Arak, and Jaak Vilo (July 2011). “G:Profiler—a Web Server for Functional Interpretation of Gene Lists (2011 Update)”. In: *Nucleic Acids Res* 39.Web Server issue, W307–W315. ISSN: 0305-1048. DOI: 10.1093/nar/gkr378.

Resnik, Philip (Aug. 1995a). “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 448–453. ISBN: 978-1-55860-363-9.

— (Aug. 20, 1995b). “Using information content to evaluate semantic similarity in a taxonomy”. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 448–453. ISBN: 978-1-55860-363-9. (Visited on 08/31/2020).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “"Why Should I Trust You?": Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, CA, USA, august 13-17, 2016*, pp. 1135–1144.

Richmond, Rebecca C. et al. (June 26, 2019). “Investigating causal relations between sleep traits and risk of breast cancer in women: mendelian randomisation study”. In: *BMJ (Clinical research ed.)* 365, p. l2327. ISSN: 1756-1833. DOI: 10.1136/bmj.l2327.

Risch, N. and K. Merikangas (Sept. 1996). “The Future of Genetic Studies of Complex Human Diseases”. eng. In: *Science* 273.5281, pp. 1516–1517. ISSN: 0036-8075. DOI: 10.1126/science.273.5281.1516.

Ritchie, Scott C. et al. (July 2016). "A Scalable Permutation Approach Reveals Replication and Preservation Patterns of Network Modules in Large Datasets". In: *Cell Systems* 3.1, pp. 71–82. DOI: 10.1016/j.cels.2016.06.012.

Robinson, P. N. and S. Mundlos (June 2010). "The human phenotype ontology". In: *Clinical Genetics* 77.6, pp. 525–534. ISSN: 1399-0004. DOI: 10.1111/j.1399-0004.2010.01436.x.

Robinson, Peter N. et al. (Nov. 17, 2008). "The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease". In: *American Journal of Human Genetics* 83.5, pp. 610–615. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2008.09.017. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668030/ (visited on 08/06/2020).

Roenneberg, Till et al. (Dec. 2007). "Epidemiology of the Human Circadian Clock". eng. In: *Sleep Med Rev* 11.6, pp. 429–438. ISSN: 1087-0792. DOI: 10.1016/j.smrv.2007.07.005.

Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain". In: *Psychological Review*, pp. 65–386.

Rosse, Cornelius and José L. V. Mejino Jr (2008). "The Foundational Model of Anatomy Ontology". In: *Anatomy Ontologies for Bioinformatics*. Ed. by Albert Burger BSc MSc, Duncan Davidson BSc, and Richard Baldock BSc. Computational Biology 6. Springer London, pp. 59–117. ISBN: 978-1-84628-884-5 978-1-84628-885-2. DOI: 10.1007/978-1-84628-885-2_4. URL: http://link.springer.com/chapter/10.1007/978-1-84628-885-2_4 (visited on 05/27/2017).

Ruepp, Andreas et al. (Jan. 2010). "CORUM: the comprehensive resource of mammalian protein complexes–2009". In: *Nucleic Acids Research* 38 (Database issue), pp. D497–501. ISSN: 1362-4962. DOI: 10.1093/nar/gkp914.

Ruth Mitchell, Elsworth (Feb. 20, 2019). "MRC IEU UK Biobank GWAS pipeline version 2". data.bris. In: Library Catalog: data.bris.ac.uk. DOI: 10.5523/bris.pnoat8cxo0u52p6ynfaekeigi. URL: https://data.bris.ac.uk/data/dataset/pnoat8cxo0u52p6ynfaekeigi (visited on 06/15/2020).

Rutter, M. (1972 Oct-Dec). "Childhood Schizophrenia Reconsidered". eng. In: *J Autism Child Schizophr* 2.4, pp. 315–337. ISSN: 0021-9185. DOI: 10.1007/BF01537622.

Scargle, J.D. (Dec. 1982). "Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data". In: *The Astrophysical Journal* 263. tex.added-at: 2010-02-23T20:59:31.000+0100 tex.biburl: https://www.bibsonomy.org/bibtex/21a tex.interhash: 913df1a1ac7e9f01402c08a84cd1a016 tex.intrahash: 1a42a6f1235b5e3fa20ea6eca1075 tex.timestamp: 2010-02-23T20:59:43.000+0100 tex.where: IO-OC2.1-Arquiv:P-S, pp. 835–853.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (July 2014). "Biological Insights from 108 Schizophrenia-Associated Genetic Loci". eng. In: *Nature* 511.7510, pp. 421–427. ISSN: 1476-4687. DOI: 10.1038/nature13595.

Schriml, Lynn Marie et al. (Jan. 2012). "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic Acids Research* 40 (Database issue), pp. D940–946. ISSN: 1362-4962. DOI: 10.1093/nar/gkr972.

Schölkopf, Bernhard, Koji Tsuda, and Jean-Philippe Vert (2004). *Kernel Methods in Computational Biology*. Google-Books-ID: SwAooknaMXgC. MIT Press. 428 pp. ISBN: 978-0-262-19509-6.

Seiffert, Chris et al. (Jan. 2010). "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40.1, pp. 185–197. ISSN: 1083-4427, 1558-2426. DOI: 10.1109/TSMCA.2009.2029559. URL: http://ieeexplore.ieee.org/document/5299216/ (visited on 01/10/2017).

Sekula, Peggy et al. (Nov. 2016). "Mendelian Randomization as an Approach to Assess Causality Using Observational Data". en. In: *JASN* 27.11, pp. 3253–3265. ISSN: 1046-6673, 1533-3450. DOI: 10.1681/ASN.2016010098.

Shah, Rupal L., Jeremy A. Guggenheim, and UK Biobank Eye and Vision Consortium (Dec. 2018). "Genome-Wide Association Studies for Corneal and Refractive Astigmatism

in UK Biobank Demonstrate a Shared Role for Myopia Susceptibility Loci". eng. In: *Hum. Genet.* 137.11-12, pp. 881–896. ISSN: 1432-1203. DOI: 10.1007/s00439-018-1942-8.

Shannon, Paul et al. (Nov. 2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13.11, pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303.

Shearer, Rob, Boris Motik, and Ian Horrocks (Jan. 2008). "HermiT: A highly-efficient OWL reasoner". In: vol. 432. OWLED.

Simon, Michelle M et al. (2013). "A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains". In: *Genome Biology* 14.7, R82. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-7-r82. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053787/ (visited on 01/09/2018).

Slenter, Denise N et al. (Jan. 2018). "WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research". In: *Nucleic Acids Res* 46.Database issue, pp. D661–D667. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1064.

Smedley, Damian and Peter N. Robinson (2015). "Phenotype-Driven Strategies for Exome Prioritization of Human Mendelian Disease Genes". eng. In: *Genome Med* 7.1, p. 81. ISSN: 1756-994X. DOI: 10.1186/s13073-015-0199-2.

Smith, Barry et al. (2005). "Relations in Biomedical Ontologies". In: *Genome Biol* 6.5, R46. ISSN: 1465-6906. DOI: 10.1186/gb-2005-6-5-r46.

Smith, Barry et al. (Nov. 2007a). "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration". eng. In: *Nat. Biotechnol.* 25.11, pp. 1251–1255. ISSN: 1087-0156. DOI: 10.1038/nbt1346.

Smith, Barry et al. (Nov. 2007b). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". In: *Nature Biotechnology* 25.11, pp. 1251–1255. ISSN: 1087-0156. DOI: 10.1038/nbt1346.

Smith, Cynthia L. and Janan T. Eppig (Dec. 2009). "The mammalian phenotype ontology: enabling robust annotation and comparative analysis". In: *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 1.3, pp. 390–399. ISSN: 1939-005X. DOI: 10.1002/wsbm.44.

Socaciu, Andreea Iulia et al. (July 21, 2020). "Melatonin, an ubiquitous metabolic regulator: functions, mechanisms and effects on circadian disruption and degenerative diseases". In: *Reviews in Endocrine & Metabolic Disorders*. ISSN: 1573-2606. DOI: 10.1007/s11154-020-09570-9.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens (Mar. 1993). "Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM)". eng. In: *Am. J. Hum. Genet.* 52.3, pp. 506–516. ISSN: 0002-9297.

Sundberg, Maria and Mustafa Sahin (Dec. 2015). "Cerebellar Development and Autism Spectrum Disorder in Tuberous Sclerosis Complex". eng. In: *J. Child Neurol.* 30.14, pp. 1954–1962. ISSN: 1708-8283. DOI: 10.1177/0883073815600870.

Surrey, Lea F. et al. (2016). "The Genomic Era of Clinical Oncology: Integrated Genomic Analysis for Precision Cancer Care". In: *Cytogenetic and Genome Research* 150.3, pp. 162–175. ISSN: 1424-8581, 1424-859X. DOI: 10.1159/000454655.

Swanson, Christine M. et al. (Oct. 1, 2017). "Bone Turnover Markers After Sleep Restriction and Circadian Disruption: A Mechanism for Sleep-Related Bone Loss in Humans". In: *The Journal of Clinical Endocrinology and Metabolism* 102.10, pp. 3722–3730. ISSN: 1945-7197. DOI: 10.1210/jc.2017-01147.

Swanson, Larry (Dec. 11, 2003). *Brain Maps: Structure of the Rat Brain: Vol 3*. 3 edition. Amsterdam ; Boston: Academic Press. ISBN: 978-0-12-610582-7.

Szklarczyk, Damian et al. (Jan. 2015a). "STRING v10: protein-protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43 (Database issue), pp. D447–452. ISSN: 1362-4962. DOI: 10.1093/nar/gku1003.

Szklarczyk, Damian et al. (Jan. 2015b). "STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life". eng. In: *Nucleic Acids Res.* 43.Database issue, pp. D447–452. ISSN: 1362-4962. DOI: 10.1093/nar/gku1003.

Tahara, Yu et al. (June 2012). "In Vivo Monitoring of Peripheral Circadian Clocks in the Mouse". eng. In: *Current biology: CB* 22.11, pp. 1029–1034. ISSN: 1879-0445. DOI: 10.1016/j.cub.2012.04.009.

Takahashi, Joseph S. (Mar. 2017). "Transcriptional architecture of the mammalian circadian clock". In: *Nature Reviews Genetics* 18.3, pp. 164–179. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.150. URL: https://www.nature.com/nrg/journal/v18/n3/abs/nrg.2016.150.html (visited on 05/26/2017).

Takumi, Toru et al. (Mar. 2020). "Behavioral Neuroscience of Autism". en. In: *Neuroscience & Biobehavioral Reviews*. IBNS 2017 - Contemporary Contributions to Basic and Translational Behavioral Neuroscience Research 110, pp. 60–76. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2019.04.012.

Thaben, Paul F. and Pål O. Westermark (Dec. 2014). "Detecting Rhythms in Time Series with RAIN". In: *Journal of Biological Rhythms* 29.6, pp. 391–400. ISSN: 0748-7304. DOI: 10.1177/0748730414553029. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4266694/ (visited on 05/17/2017).

Tibshirani, Robert (1994). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288.

Tordjman, Sylvie et al. (2015). "Autism as a Disorder of Biological and Behavioral Rhythms: Toward New Therapeutic Perspectives". English. In: *Front. Pediatr.* 3. ISSN: 2296-2360. DOI: 10.3389/fped.2015.00001.

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC. URL: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Trapnell, Cole et al. (May 2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5, pp. 511–515. ISSN: 1546-1696. DOI: 10.1038/nbt.1621.

Treur, Jorien L. et al. (2018). "Investigating genetic correlations and causal effects between caffeine consumption and sleep behaviours". In: *Journal of Sleep Research* 27.5, e12695. ISSN: 1365-2869. DOI: 10.1111/jsr.12695.

Truex, R. C. and Malcolm B. Carpenter (Mar. 1, 1996). *Human Neuroanatomy*. Ed. by Andre Parent. 9th Revised edition edition. Baltimore: Lippincott Williams and Wilkins. 2171 pp. ISBN: 978-0-683-06752-1.

Tsai, Peter T. et al. (Aug. 2012). "Autistic-like Behaviour and Cerebellar Dysfunction in Purkinje Cell Tsc1 Mutant Mice". eng. In: *Nature* 488.7413, pp. 647–651. ISSN: 1476-4687. DOI: 10.1038/nature11310.

Tu, Samson W. et al. (2008). "Using an Integrated Ontology and Information Model for Querying and Reasoning about Phenotypes: The Case of Autism". In: *AMIA Annual Symposium Proceedings* 2008, pp. 727–731. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655950/ (visited on 08/14/2020).

Vaser, Robert et al. (Jan. 2016). "SIFT Missense Predictions for Genomes". en. In: *Nat Protoc* 11.1, pp. 1–9. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2015.123.

Wang, Zhong, Mark Gerstein, and Michael Snyder (Jan. 2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484. URL: http://www.nature.com/nrg/journal/v10/n1/abs/nrg2484.html (visited on 05/18/2017).

Watmuff, Bradley et al. (June 2016). "Disease Signatures for Schizophrenia and Bipolar Disorder Using Patient-Derived Induced Pluripotent Stem Cells". eng. In: *Molecular and Cellular Neurosciences* 73, pp. 96–103. ISSN: 1095-9327. DOI: 10.1016/j.mcn.2016.01.003.

Webber, Caleb (Apr. 2017). "Epistasis in Neuropsychiatric Disorders". en. In: *Trends in Genetics* 33.4, pp. 256–265. ISSN: 0168-9525. DOI: 10.1016/j.tig.2017.01.009.

Wei, Qiong and Roland L. Dunbrack Jr (July 9, 2013). "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics". In: *PLOS ONE* 8.7, e67863. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0067863. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067863 (visited on 05/31/2017).

Wendt, Frank R. et al. (May 2020). "Heterogeneity and Polygenicity in Psychiatric Disorders: A Genome-Wide Perspective". In: *Chronic Stress (Thousand Oaks)* 4. ISSN: 2470-5470. DOI: 10.1177/2470547020924844.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: http://ggplot2.org.

Wing, Max Kuhn Contributions from Jed et al. (2018). *caret: Classification and Regression Training*. URL: https://CRAN.R-project.org/package=caret.

Wray, Naomi R. et al. (June 14, 2018a). "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model". In: *Cell* 173.7, pp. 1573–1580. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.05.051. URL: http://www.sciencedirect.com/science/article/pii/S0092867418307141 (visited on 07/16/2019).

Wray, Naomi R. et al. (May 2018b). "Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression". en. In: *Nature Genetics* 50.5, pp. 668–681. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0090-3.

Xiang, Zuoshuang et al. (June 2010). "OntoFox: Web-Based Support for Ontology Reuse". In: *BMC Research Notes* 3.1, p. 175. ISSN: 1756-0500. DOI: 10.1186/1756-0500-3-175.

Yamamoto, Takuro et al. (Oct. 2004). "Transcriptional Oscillation of Canonical Clock Genes in Mouse Peripheral Tissues". eng. In: *BMC molecular biology* 5, p. 18. ISSN: 1471-2199. DOI: 10.1186/1471-2199-5-18.

Yanai, Itai et al. (Mar. 1, 2005). "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification". In: *Bioinformatics (Oxford, England)* 21.5, pp. 650–659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti042.

Yi, Zhaohong et al. (Mar. 2002). "The Rab27a/granuphilin complex regulates the exocytosis of insulin-containing dense-core granules". In: *Molecular and Cellular Biology* 22.6, pp. 1858–1867. ISSN: 0270-7306.

Yoo, Seung-Hee et al. (Apr. 2004). "PERIOD2::LUCIFERASE Real-Time Reporting of Circadian Dynamics Reveals Persistent Circadian Oscillations in Mouse Peripheral Tissues". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.15, pp. 5339–5346. ISSN: 0027-8424. DOI: 10.1073/pnas.0308709101.

Yuan, Shuai, Honghui Yao, and Susanna C. Larsson (Aug. 2020). "Associations of Cigarette Smoking with Psychiatric Disorders: Evidence from a Two-Sample Mendelian Randomization Study". eng. In: *Sci Rep* 10.1, p. 13807. ISSN: 2045-2322. DOI: 10.1038/s41598-020-70458-4.

Zhang, Bin and Steve Horvath (2005). "A General Framework for Weighted Gene Co-Expression Network Analysis". eng. In: *Stat Appl Genet Mol Biol* 4, Article17. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1128.

Zhang, Tao et al. (2020). "High-throughput discovery of genetic determinants of circadian misalignment". In: *PLoS genetics* 16.1, e1008577. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1008577.

Zhao, Wen-Ning et al. (Mar. 2007). "CIPC is a mammalian circadian clock protein without invertebrate homologues". In: *Nature Cell Biology* 9.3, pp. 268–275. ISSN: 1476-4679. DOI: 10.1038/ncb1539. URL: https://www.nature.com/articles/ncb1539 (visited on 04/17/2018).

Zhou, Xiang and Matthew Stephens (June 2012). "Genome-Wide Efficient Mixed Model Analysis for Association Studies". In: *Nat Genet* 44.7, pp. 821–824. ISSN: 1061-4036. DOI: 10.1038/ng.2310.

Zou, Hui (Dec. 1, 2006). "The adaptive lasso and its oracle properties". In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429. ISSN: 0162-1459. DOI: 10.1198/016214506000000735. URL: https://experts.umn.edu/en/publications/the-adaptive-lasso-and-its-oracle-properties (visited on 03/05/2018).