The University of Maine
DigitalCommons@UMaine

Honors College

Spring 5-2021

Forward Genomics of a Complex Trait: Mammalian Basal Metabolic Rate

Caleigh Charlebois

Follow this and additional works at: https://digitalcommons.library.umaine.edu/honors

 Part of the Genetics and Genomics Commons

This Honors Thesis is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Honors College by an authorized administrator of DigitalCommons@UMaine. For more information, please contact um.library.technical.services@maine.edu.

FORWARD GENOMICS OF A COMPLEX TRAIT:

MAMMALIAN BASAL METABOLIC RATE

by

Caleigh Charlebois

A Thesis Submitted in Partial Fulfillment
of the Requirements for a Degree in Honors
(Zoology)

The Honors College

University of Maine

May 2021

Advisory Committee:

Danielle Levesque, Assistant Professor of Mammalogy, Advisor
Diane Genereux, Postdoctoral Fellow, Broad Institute
Jacquelyn Gill, Associate Professor of Paleoecology and Plant Ecology
Melissa Ladenheim, Associate Dean of the Honors College
Ana Breit, Ph.D. Candidate in Ecology and Environmental Sciences

ABSTRACT

The significance and nature of basal metabolic rate, a metabolic parameter recorded under specific laboratory conditions, are contested among biologists. Although it was most likely important in the evolution of endothermy in mammals and is associated with many other traits inter- and intra-specifically, the specifics of its heritability and its genetic determinants are largely unknown. Two bioinformatics pipelines are available which can associate traits with their genetic correlates given only whole genomes and phenotypes for each animal. However, extant pipelines were created with binary traits in mind. This leaves a void in our ability to associate continuous traits such as basal metabolic rate with genetic regions that influence them. To fill this gap, I developed a technique to augment the existing forward genomics pipeline developed by Hiller *et al.* (2012) by repeatedly analyzing a continuous trait converted to a binary trait via increasing thresholds. The results of my analysis identified a list of genes that have changed more from a reconstructed ancestral state in high BMR than in low BMR mammals. However, the list of genes did not appear to be enriched for genes associated with any biological process, function, or component clearly related to metabolism. Applying these analyses to other continuous traits could provide context for whether this result is unique to BMR, which could make a statement on its lack of straightforward genetic underpinnings, or is a result of the limitations of the forward genomics pipeline.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Dr. Danielle Levesque for working with me to develop my project and revise my writing as well as ensuring I had the resources and contacts to facilitate a bioinformatics thesis despite the subject being relatively new to the lab. She provided the enhanced BMR residual dataset that I used in my analyses. I would also like to thank Dr. Diane Genereux, a researcher from the Broad Institute serving on my committee, for welcoming me to meetings about the Zoonomia project. It has been very valuable for me to witness collaboration between so many researchers so early in my research career. To the rest of my committee members, I am grateful for your flexibility and interest in helping to shape my project and organize meeting times despite the additional stress we have all experienced due to the COVID-19 pandemic and political turmoil during the past academic year. Each of you brings a unique perspective that I have kept in mind as I have written my thesis. Finally, I would like to acknowledge the weekly conversations with Dr. Joy-El Talbot of Iris Data Solutions, LLC which also helped develop my project, especially the methods and statistical analyses I used to control for noise.

TABLE OF CONTENTS

LIST OF FIGURES V
LIST OF TABLES VI
INTRODUCTION 1
BMR AND THE EVOLUTION OF ENDOTHERMY 1
TRAITS ASSOCIATED WITH BMR 3
BMR IN ENERGETICS 7
UNRAVELLING THE GENETIC DETERMINANTS OF BMR 7
FORWARD GENOMICS APPROACHES 10
METHODS 15
BASAL METABOLIC RATE DATASET 15
MODIFYING FORWARD GENOMICS FOR CONTINUOUS TRAITS 15
FUNCTIONAL ENRICHMENT OF CANDIDATE GENES 24
RESULTS 27
QUANTITY OF RESULTS MATCHING ASSUMPTIONS 27
FUNCTIONAL ENRICHMENT OF RESULTING GENETIC REGIONS 32
DISCUSSION 37
NUMBER OF GENES MATCHING ASSUMPTIONS 1-3 37
FUNCTIONAL ENRICHMENT 44
CONCLUSIONS 48
REFERENCES 50
APPENDICES 54
APPENDIX A: R SCRIPTS 55
APPENDIX B: FULL CANDIDATE GENE TABLES 75
ANALYSIS 1 75
ANALYSIS 2 77
AUTHOR'S BIOGRAPHY 83

LIST OF FIGURES

Figure 1. Comparison of basal and standard metabolic rates of example mammals (endotherms) and lizards (ectotherms) of similar masses. 2

Figure 2. Basal metabolic rates of mammalian species plotted over body mass. 5

Figure 3. Mammalian BMR per gram of body mass plotted against body mass. 5

Figure 4. Visualization of the process of using the Forward Genomics pipeline to uncover genes associated with the ability to synthesize vitamin C. 12

Figure 5: Visualization of augmentation to forward genomics pipeline..... 19

Figure 6. The number of genes returned in each step of the experimental for the preliminary trial. The experimental permutation is shown in the larger chart on top and the smaller one outlined blue and the control permutations are the other smaller charts. 28

Figure 7. Total number of genes returned and number of genes remaining after filtering according to assumptions in the experimental permutation of analysis 1. 29

Figure 8. Total genes matching assumptions per threshold in analysis 1. 29

Figure 9. Distribution of test statistic for analysis 1 (number of genes found in the second to highest threshold step) in control vs experimental permutations. 30

Figure 10. Total number of genes and number of genes fitting assumptions in the experimental permutation of analysis 2. 31

Figure 11. Total genes matching expectations per threshold in Analysis 2..... 31

Figure 12. Distribution of test statistic for analysis 2 (number of genes found in the second to highest threshold step) in control vs experimental permutations. 32

LIST OF TABLES

Table 1. Species included both the online Hiller et al. forward genomics tool and the Genoud et al. (2018) BMR dataset. 15

Table 2. Trait loss and retention values according to each threshold in the preliminary analysis. 21

Table 3. Trait loss and retention values according to each threshold in analysis 1. 21

Table 4. Trait loss and retention values according to each threshold in analysis 2. 22

Table 5. GOrilla enrichment terms for analysis 1 33

Table 6. GOrilla enrichment terms for analysis 2 33

Table 7. DAVID enrichment terms for analysis 1 34

Table 8. DAVID enrichment terms for analysis 2 36

Table 9. Candidate genetic regions from analysis 1 75

Table 10. Candidate genetic regions from analysis 2 77

INTRODUCTION

BMR and the Evolution of Endothermy

Basal metabolic rate (BMR) is the rate of metabolism of an endothermic animal which is fully grown, post-absorbtive (not digesting food), non-reproducing, resting at a normal body temperature, and in an inactive phase of its circadian rhythm (Genoud et al. 2018). These stringent conditions maximize the comparability of BMR between species and lead researchers to use the measurement as a proxy to compare metabolic intensity and minimal energy expenditure among birds and mammals (Lovegrove 2000; Genoud et al. 2018). In mammals, BMR is an important parameter because of its relationship to the production of body heat. The basal metabolic rate of endotherms is five to ten times that of the equivalent parameter in ectotherms, standard metabolic rate (Bennett and Ruben 1979; Garland and Albuquerque 2017) and was likely under selection in early mammals as they developed endothermy as far back as the Permian and through to the Cenozoic period (Lovegrove 2012, 2017). Multiple competing theories have been proposed describing the selective pressures that may have led to the increase of BMR and corresponding increase in body temperature in mammals despite the apparent energetic cost of those traits (Lovegrove 2012).

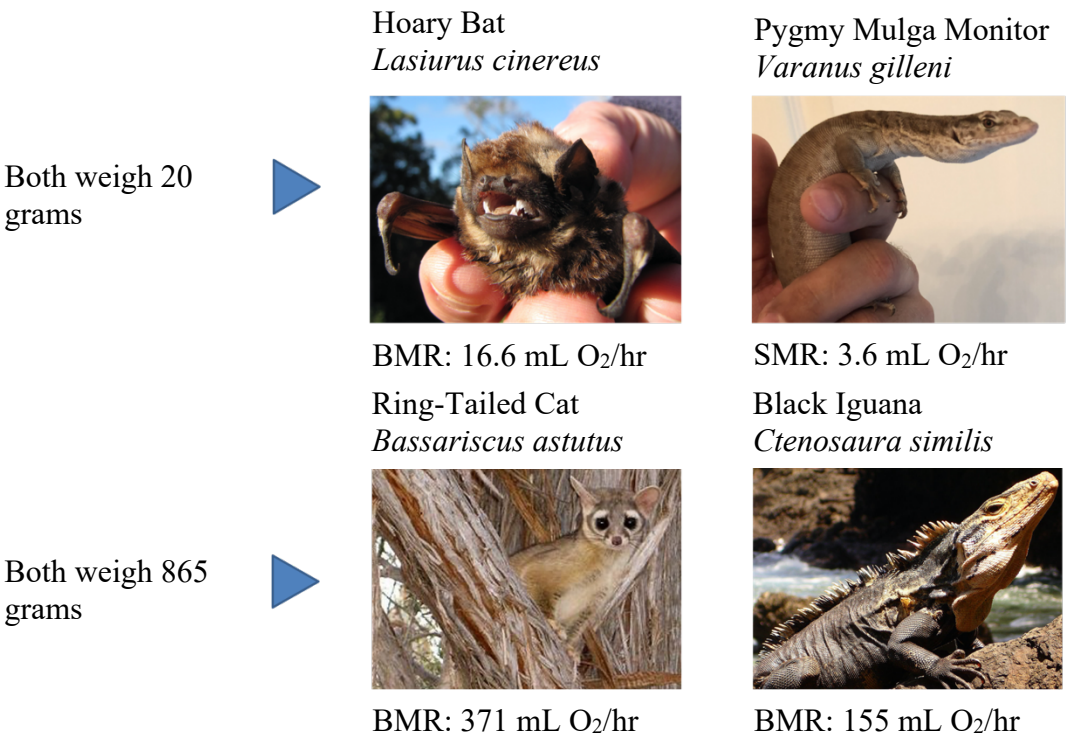


Figure 1. Comparison of basal and standard metabolic rates of example mammals (endotherms) and lizards (ectotherms) of similar masses. Photo credits: Hoary bat, Forest and Kim Starr; pygmy mulga monitor, Sergio of Reptile Highway Inc., ring-tailed cat, National Parks Service; black iguana, Parks Service of Costa Rica. Data from Genoud et al. (2018) and Garland et al. (2017).

Researchers disagree as to whether BMR developed under selection for higher body temperature or if high body temperature was simply a byproduct of increased BMR (Seebacher 2020). The Aerobic Capacity Model argues that BMR first increased as a consequence of selection for increased maximum metabolic rate to support vigorous, sustained activity and locomotion (Bennett and Ruben 1979; Lovegrove 2012). This model makes the assumption that basal metabolic rate initially increased as a side effect of the selection on maximum metabolic rate. If this were the case, it might be expected that the two would also be linked in mammals today. Results of studies evaluating correlation between minimum metabolic rate (which BMR is often used as a proxy for) and maximum metabolic rate or aerobic capacity have been mixed (Sadowska et al. 2005; Auer et al. 2017). For example, studies have shown that basal metabolic rate and peak

metabolic rate are correlated in wild armadillos (Boily 2002) but also that it is possible for selection for mass-independent maximal metabolic rate to happen independently of basal metabolic rate in laboratory mice (Wone et al. 2015). Other models describing the origin of endothermy include the Parental Care Model, the Assimilation Capacity Model, the Correlated Progression Model, and the Plesiomorphic-Apomorphic Endothermy Model. The Plesiomorphic-Apomorphic Endothermy Model proposes that endothermy developed under pressures described by all of the above models from a plesiomorphic or ancestral state where mammals exhibited limited periods of adaptive endothermy to an apomorphic homeothermic state exhibited by some modern mammals (Lovegrove 2012).

Both intra- and inter-specific analyses of the level of BMR are worthwhile in investigating the role of BMR in the evolution of endothermy. Intra-specific analyses may be able to more accurately inform inference regarding natural selection because natural selection occurs on the population and species levels, inter-specific comparisons can work with the greater degree of inter-specific genetic and phenotypic variation, improving the power of statistical analysis (Konarzewski et al. 2005; Konarzewski and Książek 2013).

Traits Associated With BMR

The trait most strongly associated with basal metabolic rate in mammals is body mass, which is responsible for up to 96.8% of the inter-species variation in BMR (McNab 2008). However, there is also a large amount of variation in BMR even among mammalian species of similar mass (Figure 2) (Lovegrove 2003; Genoud et al. 2018). Remaining variation in mammalian BMR can be determined in part by use of torpor,

habitat, type of reproduction, restricted range (McNab 2008), environmental temperature, environmental productivity, rainfall, likely mitochondrial function, and, more controversially, diet and brain size (Lovegrove 2003; White and Kearney 2013). Because these factors are not evenly distributed between phylogenetic groups and zoogeographic regions, it is difficult to determine which would be associated with BMR independent of evolutionary history and which are the result of phylogenetic biases and constraints (Lovegrove 2000). One way that researchers interested in comparative studies can control for the correlation between BMR and mass is by calculating a mass-independent BMR residual. They do this by subtracting the prediction of a linear regression for a species with a certain mass from the species' actual mass-independent BMR to quantify the difference between a species' BMR and the BMR which would be predicted based on the regression for a species of its mass (Figure 3). The mass-independent BMR residual allows researchers to make statements about how slow or fast an organism's metabolism is for its size and investigate other factors influencing the diversity in BMR (Lovegrove 2003).

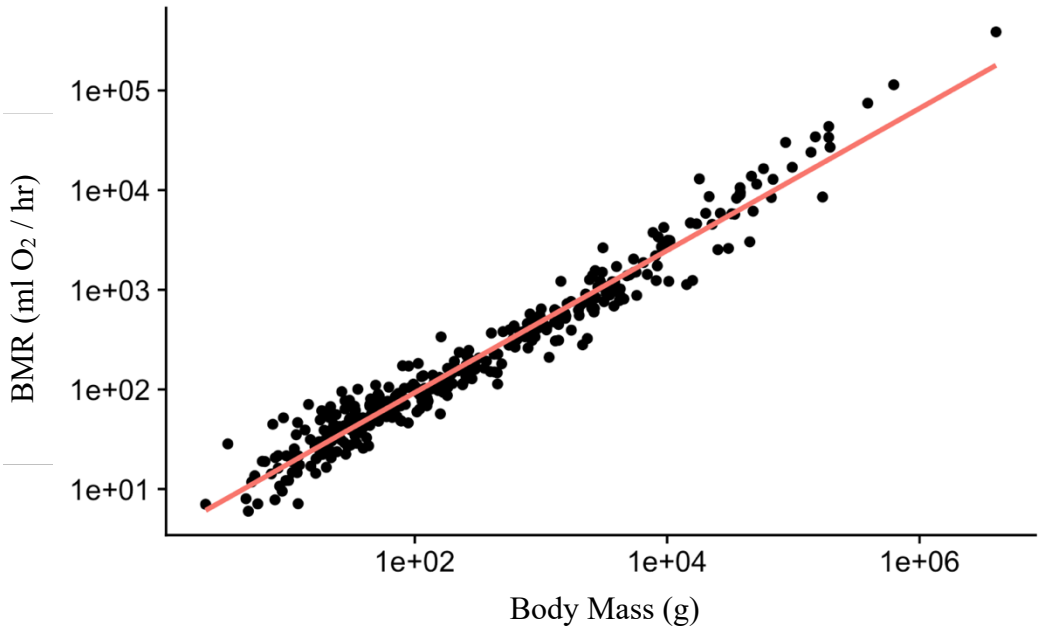


Figure 2. Basal metabolic rates of mammalian species plotted over body mass. Mammalian BMR varies allometrically with mass but is diverse even among mammals of the same mass. A linear regression of BMR over body mass, not corrected for phylogenetic relatedness, is shown as a red line. Plotted with data from Genoud et al. (2018).

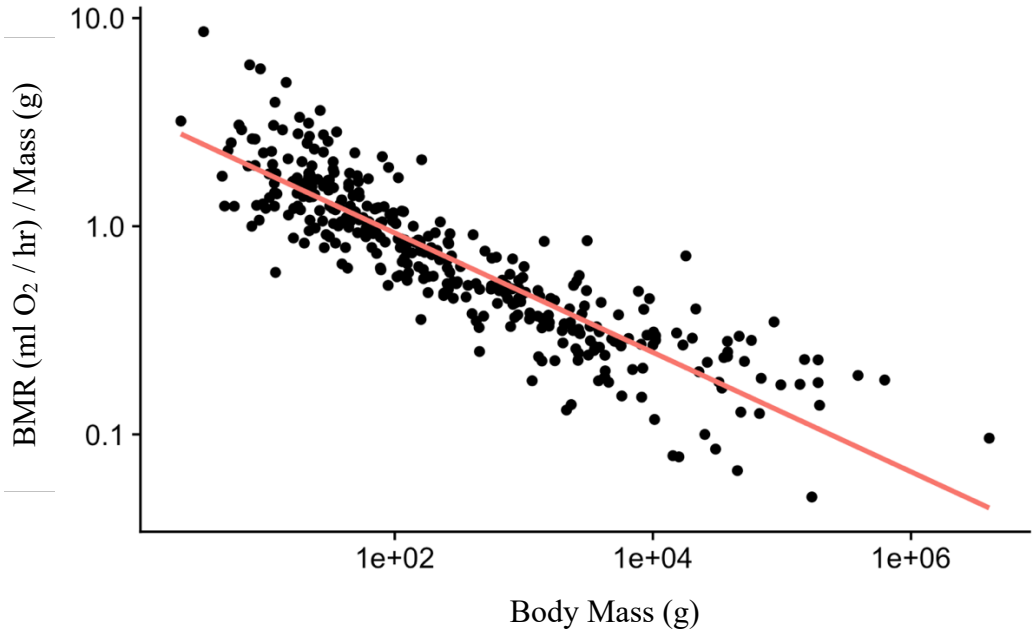


Figure 3. Mammalian BMR per gram of body mass plotted against body mass. Lighter mammals tend to have higher BMRs per gram of body mass. Mass-independent log10 BMR residuals are used to position mammals on a slow-fast metabolic continuum (Lovegrove 2003) and investigate correlation of BMR with other factors. Plotted with data from Genoud et al. (2018).

Complicating analyses of the inter-specific determinants of BMR further is the less studied incidence of intra-specific variation in basal metabolic rate (reviewed by Konarzewski and Książek 2013). Inter-specific analyses rely on one mean value accurately representing BMR in each species, despite variation among individuals and the differences between BMRs of species included in studies being large enough to make intra-specific variation irrelevant (Konarzewski and Książek 2013). Intra-specifically, BMR is correlated with organ masses (Konarzewski and Książek 2013; White and Kearney 2013) and degree of proton leak within cells (Konarzewski and Książek 2013), as well as varying between wild populations of the same species living in different environments (McNab 2008). Artificial selection targeting higher BMR has been found to effect food consumption, voluntary activity levels, immune responses, erythrocyte sizes, oxidative enzyme capacity, and fat mass (Konarzewski and Książek 2013). Notably, artificial selection for higher BMR did not appear to affect oxidative capacity on a treadmill and also did not correlate significantly with body temperature (Konarzewski and Książek 2013).

Another factor confounding both inter- and intra-species studies of basal metabolic rate is the effect of experimental conditions. Inclusion of species BMR values measured under imprecise or inconsistent conditions can have strong effects on interspecific comparative studies of small clades and smaller effects even on studies of large clades, such as all mammals (Genoud et al. 2018). Limited studies on temperate and polar species also indicate that the season when the BMR measurement is collected may

have an effect, with winter-acclimatized individuals having higher basal metabolic rates than summer-acclimatized ones (Lovegrove 2005).

BMR in Energetics

Even when consistent measurements are collected, the utility and significance of BMR are unclear. Some researchers now acknowledge that BMR represents a benchmark of metabolism under standardized conditions, rather than the lowest metabolic rate that an animal can produce. Metabolic rate falls below BMR under some conditions (White and Kearney 2013). For example, in a sample of 69 adult human subjects, metabolic rate during sleep fluctuated throughout the night and dropped to 90% of BMR for a period (Seale and Conway 1999). The fact that metabolic rate may regularly drop below BMR even in mammals which are not capable of torpor calls into question the validity of past and current research that uses BMR as "a measure of the minimal intensity of the metabolic machinery of a normothermic endotherm, or as a proxy for energy expenditure or requirements of endotherms" (Genoud et al. 2018). Both basal metabolic rate and resting metabolic rate have been accepted at times as measurements of "minimum" endotherm metabolism (Auer et al. 2017). Though researchers should use caution in discerning which applications of BMR are valid, it remains impressive as a highly studied, standardized metabolic parameter that can be compared among at least 817 species of mammals (Genoud et al. 2018).

Unravelling the Genetic Determinants of BMR

Understanding the genetics and genomics of basal metabolic rate has the potential to lend insight into the evolution of higher basal metabolic rate in mammals, causal relationship with associated phenotypes, and degree of genetic determination and

heritability. However, its entanglement with many other phenotypes, some of which have their own genetic determinants and some of which may be the result of acclimatization, makes identifying genes that influence BMR challenging.

Different hypotheses about the evolutionary origins of BMR result in different predictions for what its genetic signature should look like today. While some believe that BMR evolved to support endothermy to allow organisms to sustain stable temperatures aiding in the development of offspring, others argue that BMR evolved under pressures independent from its impact on endothermy, such as the capacity for higher maximum metabolic rate or aerobic capacity, and that endothermy may just be one possible result of this adaptation (Seebacher 2020). Whether body temperature was the target of selection driving evolution of BMR or a side effect of selection for another trait, the two traits eventually became decoupled as mammals were placed under strong pressure to adapt to colder climates (Avaria-Llautureo et al. 2019). A model relying heavily on the assumption that basal metabolic rate is linked to aerobic capacity predicts that the same positive genetic correlation between BMR and maximum metabolic rate present in early mammals should still exist in species under selection today, as it is an inherent part of how metabolism works (Hayes 2010; Nespolo et al. 2011; Konarzewski and Książek 2013). A more moderate form of the model predicts that the genes responsible for the large heritable variations in BMR and aerobic capacity in early mammals have since become fixed in mammal lineages because their adaptive advantages were so great. This would mean the genes that cause intraspecific variation in BMR today may not be the same as the genes that caused variation in early mammals (Nespolo et al. 2011; Konarzewski and Książek 2013).

Although studies focusing on the genomic determinants of variability in BMR were nonexistent as of 2013 (Konarzewski and Książek 2013), since then, a handful of genome-wide association studies on BMR and the adjacent metabolic phenotype RMR in humans have suggested genes that could be associated with the trait. Single nucleotide polymorphisms in the genes NRG3, OR8U8, BCL2L2-PABN1, PABN1, and SLC22A17 were associated with both BMR and body mass index while single nucleotide polymorphisms in FGGY, PTPRD, NPAS3, PKD1L2, and SETBP1 were associated with BMR alone in Korean women (Lee et al. 2016). These genes are thought to regulate metabolic pathways related to obesity (Kim et al. 2019). Mutations in the gene GPR158 have been associated with lower RMR and energy expenditure in individuals from the Pima Nation of American Indians (Piaggi et al. 2017). Another group of genes which some expect to be related to BMR are mitochondrial carrier proteins including UCP-1, which is essential for non-shivering thermogenesis in brown adipose tissue (Dulloo and Samec 2001; Ricquier 2011), although this expectation is tempered by the fact that transgenic mice with heightened UCP-1 activity in skeletal muscle did not display increased weight-specific BMR (Klaus et al. 2005; Konarzewski and Książek 2013) and non-shivering thermogenesis occurs below the thermoneutral zone, at lower temperatures than BMR is measured by definition. If high BMR in endotherms is due to different regulation of genes also present in ectotherms, a wide range of regulatory proteins may also be implicated (Konarzewski and Książek 2013).

Forward Genomics Approaches

Two publications, one by Hiller et al. (2012) and one by Marcovitz et al. (2019), outline different inter-species genomic analyses that associate shared phenotypes with conserved genomic regions given inputs of phenotypes and annotated genome assemblies (Hiller et al. 2012; Marcovitz et al. 2019).

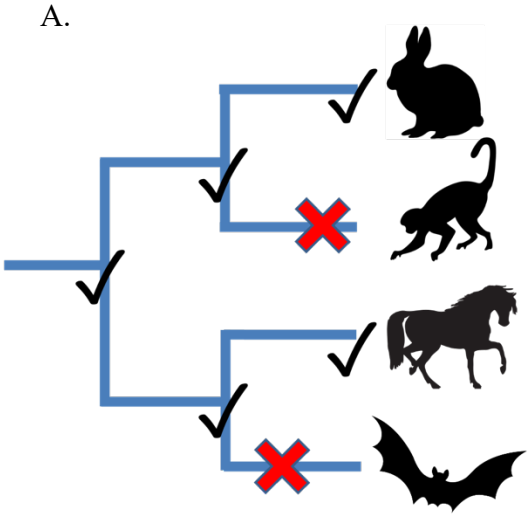
One of these methods for uncovering genetic correlates of a phenotype is the functional enrichment test for convergent evolution published by Marcovitz et al. (2019). This approach relies on the assumption that when species from different branches of the mammal phylogeny develop a convergent trait, trait-related genes should demonstrate amino acid convergence as well (Marcovitz et al. 2019). The Marcovitz *et al.* (2019) analysis reconstructs the likely ancestral states of genes conserved in mammalian lineages which share convergent traits and their outgroups. Next, it searches those genes to flag those which contain more convergent amino-acid substitutions in target lineages with a similar phenotype than in their outgroups. The analysis filters out genes which converge due to relaxation of selection rather than convergent selection by removing genes with an increase in divergent amino acid substitutions as well as convergent ones from the results. Finally, the resulting list of genes is examined for gene ontology terms, records of what molecular functions, cellular components, biological processes, and anatomical features genes are associated with based on prior lab studies of knockout and RNA sequencing studies. Examining which gene ontology terms are more associated with the experimental dataset than in a control group allows researchers to determine whether a higher proportion of phenotype-associated genes are returned by their analysis than would be expected to appear by random chance. The authors primarily framed this

method as a way of determining whether a convergent genetic element is present in the evolution of a convergent trait rather than as a way of identifying individual trait-associated genes (Marcovitz et al. 2019).

The forward genomics pipeline developed by Hiller *et al.* (2012) relies on the opposite assumption: When a phenotype is lost, genomic regions that were associated with that phenotype will begin to accumulate random mutations and diverge. Like the Marcovitz *et al.* (2019) approach, the forward genomics approach uses ancestral state reconstruction. By searching for regions which deviate from the reconstructed ancestral sequence more in species that lack a phenotype than species which retain the phenotype, the forward genomics pipeline identifies genes likely to be associated with the phenotype (Hiller et al. 2012). Researchers used the forward genomics pipeline to correctly identify the inactivated Gulo gene associated with loss of the vitamin C synthesis-capable phenotype by inputting only the full genomes of a set of mammals and information about which species in the set had lost the phenotype (Hiller et al. 2012).

The Hiller *et al.* (2012) forward genomics pipeline requires a multiple genome alignment of mammals and some outgroup species as well as a list of regions highly conserved in vertebrates. A tool called Prequel is then used to reconstruct ancestral sequences for those regions conserved in at least one outgroup species (Siepel et al. 2005; Hiller et al. 2012). For each genomic region, the sequence in each in-group species is assigned a percent identity from 0 (complete loss) to 100 (complete identity) based on its nucleotide identity with the reconstructed ancestral sequence. It outputs regions which have at least 1% less identity with the ancestral species in all trait-loss species than in all trait-retaining species with a percent identity value for that trait (Figure 4).

A.



B.

Species	Trait Loss	Trait Preserving
rabbit		✓
monkey	✓	
horse		✓
bat	✓	

C.

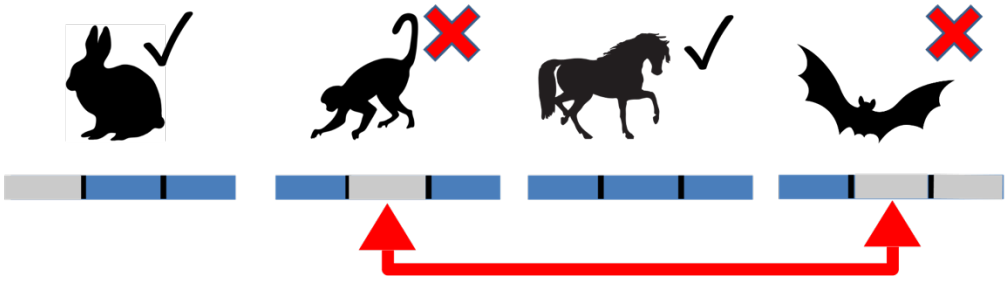


Figure 4. Visualization of the process of using the Forward Genomics pipeline to uncover genes associated with the ability to synthesize vitamin C. A) A small sample of the mammalian tree. Branches where a species has lost the ability to synthesize vitamin C are marked with a red X, while black checks represent the ability to synthesize it. B) A table recording whether a species retains the phenotype of interest. In this example, species which are trait-preserving retain the ability to synthesize vitamin C. C) Bars represent an orthologous section of the genome in the four species, with segments representing individual genes. Blue segments share a high percent genetic identity with the predicted ancestral gene sequence, while grey segments have low identity. The segments marked with red arrows are likely related to the vitamin C phenotype because they have deviated more from the ancestral state in all species which have lost that phenotype than in all species which have retained the phenotype. Modified from Hiller et al. (2012).

The major obstacle to applying either the Marcovitz *et al.* molecular convergence method or the Hiller *et al.* (2012) forward genomics method to BMR is the fact that BMR is a continuous value, not a binary trait. I chose to focus on adapting the Hiller *et al.* (2012) forward genomics pipeline in my thesis rather than the Marcovitz *et al.* pipeline (2019) for three reasons: First, the forward genomics pipeline analyzes a larger selection of conserved genetic regions because it includes non-gene region. The Marcovitz *et al.*

(2019) method includes only coding genes because they can be analyzed for amino acid convergence. Second, the forward genomics pipeline is more accessible to me because it is available through a public user interface at http://phenotree.stanford.edu/public/html/ and the Levesque lab is in contact with the Hiller group as part of existing projects. Finally, there is precedent for applying the forward genomics pipeline to continuous traits.

As well as using it to correctly identify L-gulono-lactone oxidase (Gulo), a gene necessary for the ability to synthesize vitamin C in mammals, Hiller *et al.* (2012) used the pipeline to associate the ATP binding cassette subfamily B member 4 gene (Abcb4) with low levels of biliary phospholipids in guinea pigs and horses. The researchers pointed out a large difference between the guinea pig and horse biliary phospholipid levels (0.11 and 0.38 mM respectively) and the levels of other mammals, which are usually well above 1 mM. They tested the guinea pig in the trait-loss group alone, which resulted in a list of genes too long to analyze. They then grouped the guinea pig and horse, which resulted in only 8 potential phenotype-associated genes, one of which was known phospholipid transporter Abcb4. The researchers did not attempt to determine which genes might be associated with intermediate phospholipid levels, treating the low and high end of biliary phospholipid levels essentially as a binary condition and only testing thresholds until they found the correct threshold to reveal a likely trait-associated gene (Hiller et al. 2012). Commenting on their findings, Hiller *et al.* suggested that the method may have "potential applicability to continuous traits by testing different thresholds" (Hiller et al. 2012). Using the forward genomics pipeline to analyze BMR offers an opportunity to test this possibility more extensively.

My thesis project investigates whether the Hiller *et al.* (2012) forward genomics analysis can be applied to BMR through an extension of the simple binning approach used in their analysis of Abcb4. By running the analysis multiple times on the same data set with low and high groups divided according to increasing thresholds and filtering the results according to additional assumptions, it may be possible to reveal genes which have been lost independently by mammals that convergently evolved increased metabolic rate. If the analysis yields useful results, it could also suggest the usefulness of the approach in analyzing other continuous traits.

METHODS

Basal Metabolic Rate Dataset

BMR data were taken from a subset of the highest quality measurements from the most recent mammalian BMR dataset (Genoud et al. 2018). Mass-independent log10 BMR residuals were then calculated using BMR and body mass (Lovegrove 2003). Eleven species were present both in the online forward genomics tool and the high quality BMR dataset (Table 1).

Table 1. Species included both the online Hiller et al. forward genomics tool and the Genoud et al. (2018) BMR dataset. BMR residuals are calculated from a regression of mass-independent BMR and body mass with data from the Genoud et al. (2018) dataset.

Common Name	**Binomial Name**	**BMR Residual**
Chimp	*Pan troglodytes*	0.375
Marmoset	*Callithrix jacchus*	-0.336
Bushbaby	*Otolemur garnettii*	-0.112
Tree Shrew	*Tupaia belangeri*	-0.122
Mouse	*Mus musculus*	0.108
Kangaroo Rat	*Dipodomys ordii*	0.048
Dolphin	*Tursiops truncatus*	0.743
Dog	*Canis lupus familiaris*	0.293
Megabat	*Pteropus vampyrus*	0.503
Hedgehog	*Erinaceus europaeus*	-0.142
Shrew	*Sorex araneus*	1.142

Modifying Forward Genomics for Continuous Traits

Those affiliated with Hiller *et al.* released two implementations of the forward genomics pipeline. The first, released in 2012, works as described in the introduction and

returns genes which show less identity to the reconstructed ancestral sequence in all trait-loss species than all trait-preserving species (Hiller et al. 2012). The second, published in 2016, has the additional capability to correct for phylogenetic relatedness and rate of evolution, increasing the sensitivity of the results (Prudent et al. 2016). The newer implementation allows violations as long as a certain level of certainty can be maintained (Prudent et al. 2016).

It would have been ideal to run the forward genomics pipeline on an entirely new alignment, taking advantage of the over 200 mammals now sequenced (Genereux et al. 2020). However, running the forward genomics pipeline from the beginning would have involved converting a multi-genome alignment of these species into the correct format, annotating it with highly conserved regions, and implementing the rest of the forward genomics pipeline on the University of Maine's ACG computer cluster. I was not able to obtain these resources in time to generate data for my thesis. Instead, I used the online interface made available for interacting with the Hiller *et al.* (2012) forward genomics pipeline. This web tool employs the 2012 implementation of the forward genomics pipeline and runs analyses against pre-generated percent identity files for genetic regions. Because the 2016 revisions function to increase sensitivity, continued use of the old tool may miss candidate genes but is unlikely to introduce false positives.

The web tool uses a 33-way vertebrate alignment including mammalian species and outgroups that include the chicken, zebra finch, and lizard. The conserved regions were generated with PhastCons (Siepel et al. 2005) and drawn from annotations of known genes downloaded from the UCSC genome browser. Elements within 30 base pairs of one another were merged, and regions of at least 70 base pairs were retained after

merging. Potentially problematically for analysis of the BMR phenotype, regions in the mitochondrial chromosome were excluded. The mitochondria are an important site of energy production in the cell and their DNA encode proteins with metabolism-related functions (Lv et al. 2017).

I ran the Hiller *et al.* forward genomics analysis multiple times, dividing species into high and low BMR groups based on a threshold that increased with each analysis. (Figure 5). I then filtered the results according to a set of assumptions discussed under the subheading "Development of Assumptions."

Development of Assumptions

I developed a set of assumptions building on those inherent to the Forward Genomics pipeline to filter the results for continuous trait-associated genes.

Assumption 1. The same low-BMR associated genes should be lost or changed independently in more than one organism with high BMR. This assumption builds on those that form the foundation of the Forward Genomics pipeline. Importantly, it challenges the assumption that genomic regions which differ more from the ancestral state in all trait-loss organisms than all trait-preserving organisms differ due to relaxation of selection. The assumption may make sense when the lost trait is a binary trait. In the case of BMR, however, the trait "loss" group is simply the group which has changed from the ancestral condition. If a gene deviates more from the ancestral state in all high-BMR "trait-loss" species than all low-BMR "trait-preserving" species, it is also possible that the gene has changed due to adaptation, either divergent or convergent. This assumption also acknowledges that the Forward Genomics pipeline will, by default, return a result even when a region is conserved in only one trait loss species out of the

trait loss group. Such a result would not provide information about a trait being independently lost in more than one species, so for this analysis, I set the Forward Genomics pipeline to return a result only if the genetic region was present in at least two trait-loss species.

Assumption 2. Species with higher BMR should show more loss or change in low-BMR associated genes than species with lower BMR. The number of BMR-associated genes in the output table was expected to increase as the BMR threshold increased.

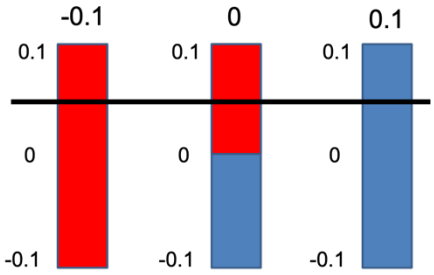
Assumption 3. Genes marked "loss" at a medium BMR should remain "loss" at high BMR. I removed genes which appeared in the results and then disappeared again as the threshold increased because they were not consistently associated with BMR above a given level. This filter for the loss signature of continuous-trait-associated genes narrows the search to return only genes that are consistently, and therefore potentially intrinsically, associated with independent increases in BMR across mammalian orders.

A.

Kangaroo Rat
Dipodomys ordii

BMR vs. Mass Residual:

0.048



B.

Species	BMR Residual	Threshold -0.1	Threshold 0	Threshold 0.1
mouse	0.108	High	High	High
kangaroo rat	0.048	High	High	Low
megabat	-0.084	High	Low	Low
hedgehog	-0.142	Low	Low	Low

C.

Gene Symbol	Gene Name	Threshold -0.1	Threshold 0	Threshold 0.1
Gm12185	hypothetical protein LOC620913	Loss	Loss	-
LOC100502936	ubiquitin-conjugating enzyme E2 Q2-like	Loss	-	-
Spinkl	serine protease inhibitor kazal-like protein,	-	Loss	Loss
V1rh13	vomeronasal 1 receptor, H13	Loss	-	Loss

Figure 5: Visualization of augmentation to forward genomics pipeline. A) A continuous phenotype converted to a binary phenotype using bins. The "high" bin is shown in red, while the "low" bin is shown in blue. The threshold between "low" and "high" for each bar is written above them. Depending on the threshold used to create bins, the same organism could have either "low" or "high" BMR. B) Example input data adapting the Forward Genomics pipeline to a continuous trait. The data has been binned using progressively higher thresholds. BMR data from Genoud et al. 2018. C) Example output data from the modified Hiller et al. pipeline. Genes are listed along with their Mouse Genome Informatics gene symbols and which thresholds resulted in their detection. Genes listed as "loss" had lower percent identity with the common ancestor in all species with loss phenotype (high BMR) than all species with preserved phenotype (low BMR) when a particular threshold determined the species with the loss phenotype. Photography credit: US Fish and Wildlife Service.

List of Analyses

First, I conducted a preliminary analysis with thresholds -0.1, 0, 0.1, 0.2, and 0.3 and created five control permutations (created as described under the subheading "Testing Significance" below) in addition to the experimental permutation. Lack of confidence in my results due to a low number of controls (because I needed to enter the phenotype status of each species for each threshold in each permutation by hand) as well as difficulty matching species names in the online forward genomics tool and the Genoud *et al.* (2018) dataset motivated me to create R scripts to streamline these tasks. Compared to the correct data (Table 1), I erroneously included two extra species in these preliminary trials for which there is actually no data in the Genoud *et al.* (2018) dataset, misidentified the species of Megabat referred to by the online tool as *Pteropus giganteus*, and did not include the dolphin (*Tursiops truncatus*) in my analysis (Table 2).

I then conducted two analyses with correct data from Table 1. The first analysis, analysis 1, used the thresholds -0.3, -0.1, 0.1, 0.3, and 0.5. The second, analysis 2, used 10 thresholds set at the same values as species' BMRs so that species were added to the ancestral group one by one, except for the final two. The final two species were added together because otherwise there would be no results in the last step; the forward genomics tool was set to return a genetic region only when two species in the final step both shared it. I ran 200 control permutations (instances of the R scripts run with the BMRs of the species shuffled randomly) to accompany analysis 1 and 100 alongside analysis 2.

Table 2. Trait loss and retention values according to each threshold in the preliminary analysis. Species are labeled “trait-loss” and “trait-retaining” in the Hiller et al. forward genomics tool, but have been labeled “derived” and “ancestral” here for clarity.

Common Name	**Binomial Name**	BMR residual	Threshold -0.1	Threshold 0	Threshold .1	Threshold .2	Threshold .3
Microbat	Hipposideros galeritus	-0.513	ancestral	ancestral	ancestral	ancestral	ancestral
Marmoset	Callithrix jacchus	-0.336	ancestral	ancestral	ancestral	ancestral	ancestral
Hedgehog	Erinaceus europaeus	-0.142	ancestral	ancestral	ancestral	ancestral	ancestral
Tree_Shrew	Tupaia belangeri	-0.122	ancestral	ancestral	ancestral	ancestral	ancestral
Bushbaby	Otolemur garnettii	-0.112	ancestral	ancestral	ancestral	ancestral	ancestral
Megabat	Pteropus giganteus	-0.084	derived	ancestral	ancestral	ancestral	ancestral
Kangaroo_Rat	Dipodomys ordii	0.048	derived	derived	ancestral	ancestral	ancestral
Mouse	Mus musculus	0.108	derived	derived	derived	ancestral	ancestral
Dog	Canis lupus familiaris	0.293	derived	derived	derived	derived	ancestral
Rat	Rattus rattus	0.294	derived	derived	derived	derived	ancestral
Chimp	Pan troglodytes	0.375	derived	derived	derived	derived	derived
Shrew	Sorex araneus	1.142	derived	derived	derived	derived	derived

Table 3. Trait loss and retention values according to each threshold in analysis 1.

Common Name	**Binomial Name**	BMR residual	Threshold -0.3	Threshold -0.1	Threshold 0.1	Threshold 0.3	Threshold 0.5
Marmoset	Callithrix jacchus	-0.336	ancestral	ancestral	ancestral	ancestral	ancestral
Hedgehog	Erinaceus europaeus	-0.142	derived	ancestral	ancestral	ancestral	ancestral
Tree Shrew	Tupaia belangeri	-0.122	derived	ancestral	ancestral	ancestral	ancestral
Bushbaby	Otolemur garnettii	-0.112	derived	ancestral	ancestral	ancestral	ancestral
Kangaroo Rat	Dipodomys ordii	0.048	derived	derived	ancestral	ancestral	ancestral
Mouse	Mus musculus	0.108	derived	derived	derived	ancestral	ancestral
Dog	Canis lupus familiaris	0.293	derived	derived	derived	ancestral	ancestral
Chimp	Pan troglodytes	0.375	derived	derived	derived	derived	ancestral
Megabat	Pteropus vampyrus	0.503	derived	derived	derived	derived	derived
Dolphin	Tursiops truncatus	0.743	derived	derived	derived	derived	derived
Shrew	Sorex araneus	1.142	derived	derived	derived	derived	derived

Table 4. Trait loss and retention values according to each threshold in analysis 2. D stands for "derived" and A stands for "ancestral."

Common Name	**Binomial Name**	BMR residual	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9
Marmoset	Callithrix jacchus	-0.336	A	A	A	A	A	A	A	A	A
Hedgehog	Erinaceus europaeus	-0.142	D	A	A	A	A	A	A	A	A
Tree Shrew	Tupaia belangeri	-0.122	D	D	A	A	A	A	A	A	A
Bushbaby	Otolemur garnettii	-0.112	D	D	D	A	A	A	A	A	A
Kangaroo Rat	Dipodomys ordii	0.048	D	D	D	D	A	A	A	A	A
Mouse	Mus musculus	0.108	D	D	D	D	D	A	A	A	A
Dog	Canis lupus familiaris	0.293	D	D	D	D	D	D	A	A	A
Chimp	Pan troglodytes	0.375	D	D	D	D	D	D	D	A	A
Megabat	Pteropus vampyrus	0.503	D	D	D	D	D	D	D	D	A
Dolphin	Tursiops truncatus	0.743	D	D	D	D	D	D	D	D	D
Shrew	Sorex araneus	1.142	D	D	D	D	D	D	D	D	D

Modified Forward Genomics Pipeline

A major obstacle to the execution of the proposed analysis was the user interface of the online Phenotree Forward Genomics tool made available by Hiller *et al.* on the Stanford webserver. It was intended to allow a casual user to interact with the with the forward genomics approach by analyzing a binary trait of their choice and selecting mammals as "trait-loss," "trait-preserving," or "ignore" by hand. This design made running multiple analyses with changing thresholds slow. To facilitate this project, I wrote a set of scripts in the programming language R (R Core Team 2013) to interact with the website. The included scripts match species found on the online tool with phenotypes, split them into trait-loss and trait-preserving groups depending on user-

supplied thresholds, generate controls, submit the tests to the webpage, download the results, and perform a statistical analysis to generate a p-value from permutation testing with the controls. The scripts in their entirety and a more detailed explanation are appended (Appendix A).

Testing Significance

I used permutation testing to control for noise in my analysis and ascertain whether more genes meeting the assumptions outlined above were found in the experimental group than would be expected by random chance. I borrowed my approach from the Marcovitz *et al.* (2019) molecular convergence test. To establish control permutations, they recalculated their results with different combinations of species labeled as having phenotypes of interest and being in the outgroups (Marcovitz et al. 2019). This established a baseline for results that should be expected due to random noise alone, when the species were not organized according to any trait. For this project, I recorded how many species were moved to the low BMR group at each threshold in my real input data, which I subsequently call the experimental permutation. I created many control permutations of the data by shuffling which species out of those which were included in the BMR dataset were added to the ancestral and derived trait groups during each increasing threshold step but keeping the number of species added to the low group at each step the same as in the experimental permutation. I then plotted a distribution of the test statistic calculated from each of these controls (Figure 8). In this kind of permutation test, the position of the experimental trial on this distribution corresponds to a p-value indicating how likely the same results are to have occurred by chance alone (Wilber 2019).

To determine whether significantly more genes matched my assumptions in the experimental permutation compared to the control permutations, I needed to choose a test statistic which would quantify the number of genes matching my assumptions in a permutation. I first speculated that the most straightforward statistic would be the number of genes returned in the final, most inclusive threshold. Due to the nature of the assumptions used to filter genes, the count of genes present in the final group would include all of the genes found in lower threshold groups as well. However, genes in the final group were not filtered by an additional higher group. This would limit the effect of assumption 2 on the quantity of genes present in the group created according to the highest threshold and mean they would not be filtered to remove genes not consistently associated with increased levels of BMR. Due to this concern, I chose to use the number of genes found in the second to last threshold group, the largest group that still benefitted from the filtering effect of another group with a higher threshold.

Functional Enrichment of Candidate Genes

Without intraspecific lab testing with knockouts, it not possible to know for sure whether a gene is associated with a certain phenotype. However, I can borrow another technique from Marcovitz et al. (2019) and search for enrichment in genes known to be associated with metabolism in the results to infer whether other genes found alongside them may be worthwhile candidates for further testing.

After receiving the lists of total genes lost from the forward genomics tool and filtering them to keep only genes matching assumptions 1 through 3 above, I generated line break separated lists of both target and control genes in each analysis to submit to

web tools that detect enrichment in gene ontology terms. Gene ontology terms are tags added to genes in databases that label them as being related to specific biological processes, anatomical features, or functions. Tools searching for enrichment in gene ontology terms compare the terms associated with a target, or experimental, list to those associated with a background, or control list. Terms that appear significantly more in the target list than the background list are said to be enriched in that list, meaning that they appear more than one would expect due to random chance. This can help researchers determine what functions are over-represented in a target list of genes compared to the background. In this case, I might expect a list of BMR-related genes to be enriched for gene ontology terms relating to the mitochondria or metabolism.

The target list was generated from all genes present in each threshold of the experimental permutation, while the single control list contained all genes from all control permutations. I submitted these lists of gene names to the GOrilla gene ontology enrichment tool (Eden et al. 2007, 2009) and the DAVID (Huang et al. 2009a; b) gene functional classification tool to determine if the gene lists generated by my analyses had more genes with any particular function than would be expected by chance. In the GOrilla tool, the experimental permutation gene list was used as a target list while the control permutation list was used as a background or control list. In the DAVID tool, the experimental and control lists were both used as targets against an available *Mus musculus* background list. I subtracted the terms enriched in the control list from those found in the experimental list, keeping only the terms which were enriched in the results of the experimental but not the control permutations. I selected the *Mus musculus* (mm7) reference assembly when prompted in both tools because the Forward Genomics pipeline

used an extended *Mus musculus* assembly as a starting point to define its conserved genetic regions (Hiller et al. 2012).

I also manually scanned the results for the few genes expected to be associated with mammalian BMR: UCP1 (Ricquier 2011), KLF5 (Choi et al. 2013), NRG3, OR8U8, BCL2L2-PABN1, PABN1, SLC22A17, FGGY, PTPRD, NPAS3, PKD1L2, SETBP1 (Lee et al. 2016), and GPR158 (Piaggi et al. 2017).

RESULTS

Quantity of Results Matching Assumptions

The preliminary analysis with only 5 controls returned 3, 4, 5, 31, and 198 genes respectively as each threshold was applied in the experimental permutation. A p-value was not calculated due to the low number of controls, however the numbers of genes in each step of the experimental trial appeared intermediate among the numbers of genes in each step in the controls (Figure 6).

Analysis 1 indicated two genes lost with threshold -0.3, 4 lost with -0.1, 6 lost with 0.1, 30 lost with 0.3, and 64 lost with 0.5 (Figure 6). According to the test statistic chosen prior to the analysis, quantity of genes lost when species were divided according to the second to highest threshold, this was significantly more candidate genes than were found in the control permutations, with a p-value of 0.02 from a distribution of 200 sample permutations and 1 experimental trial (Figure 8). The mean of the test statistic across all trials was 6.74, the median was 5, and the mode was 2. The experimental permutation returned roughly 23 more candidate genes than the average control permutation in the second to highest threshold. All genes matching assumptions are available in the first table of Appendix B.

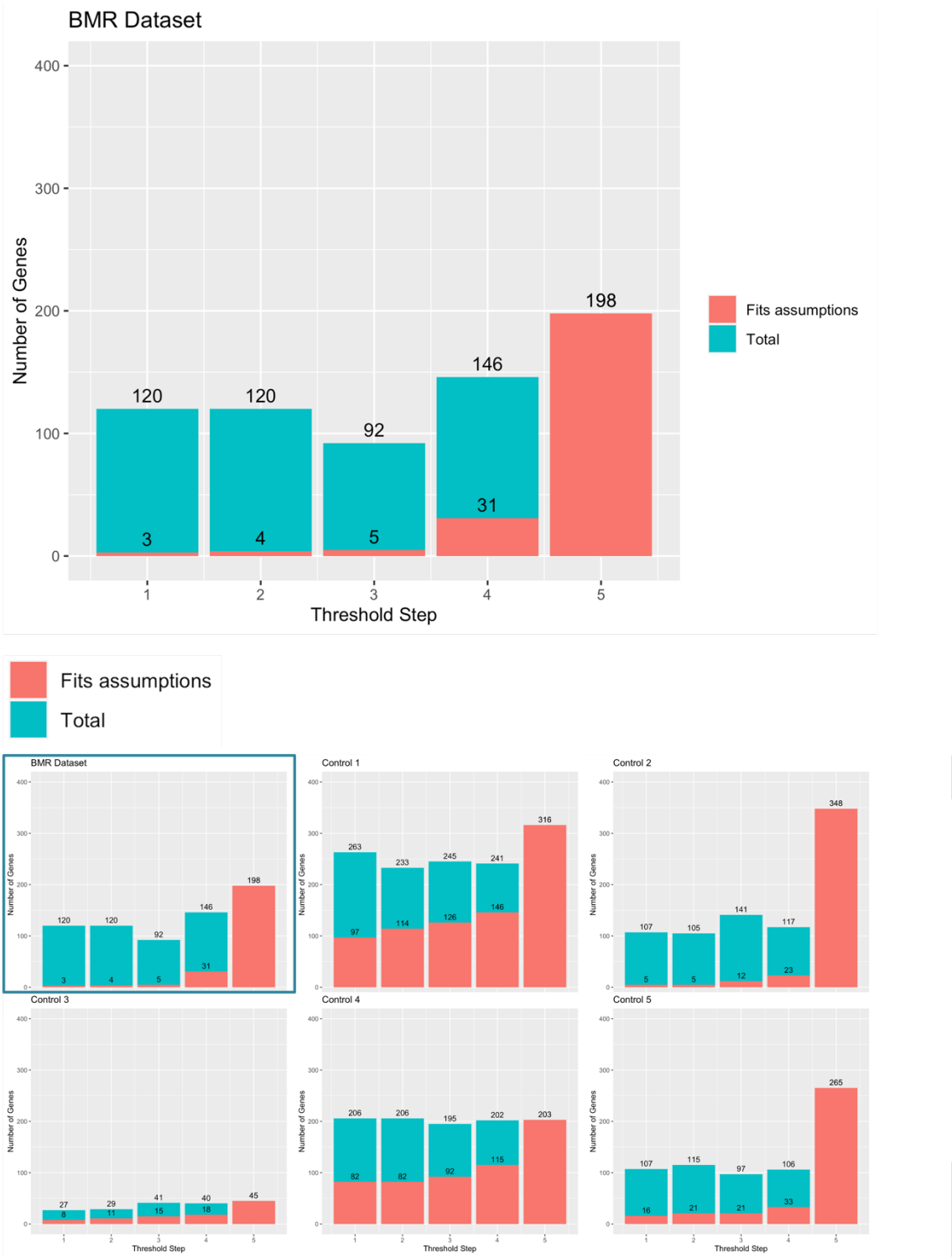


Figure 6. The number of genes returned in each step of the experimental for the preliminary trial. The experimental permutation is shown in the larger chart on top as well as repeated in the smaller one outlined in blue. The control permutations are the five other smaller charts.

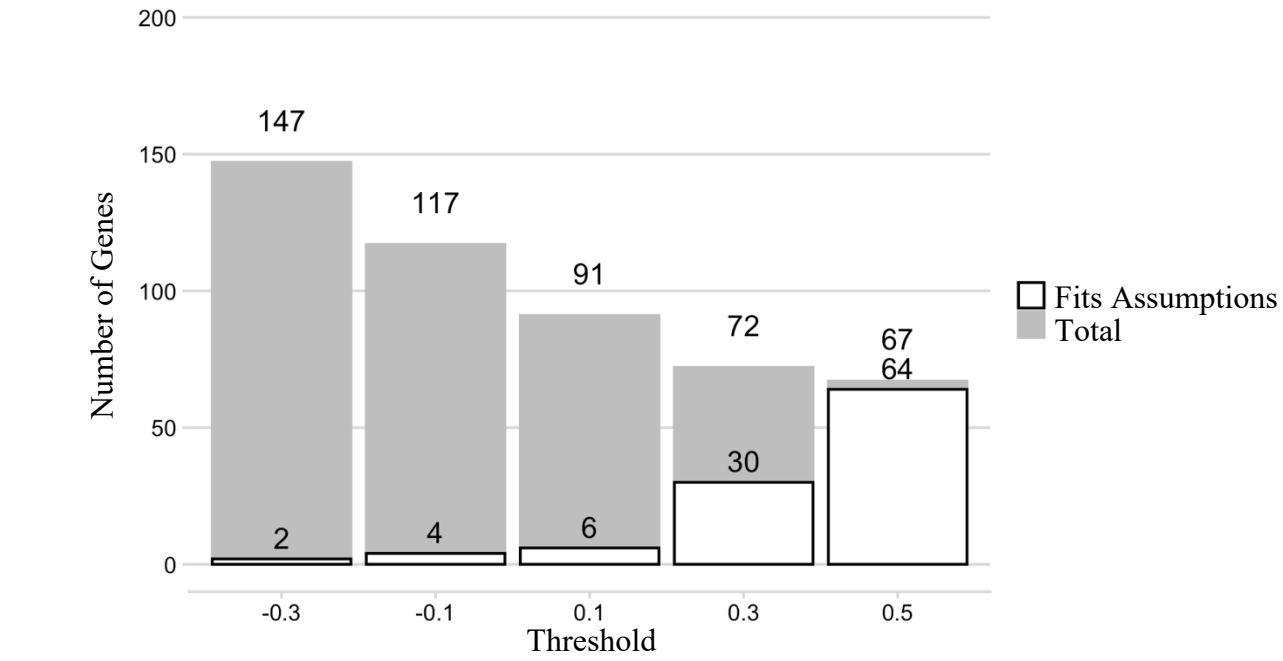


Figure 7. Total number of genes returned and number of genes remaining after filtering according to Assumptions 1-3 in the experimental permutation of analysis 1.

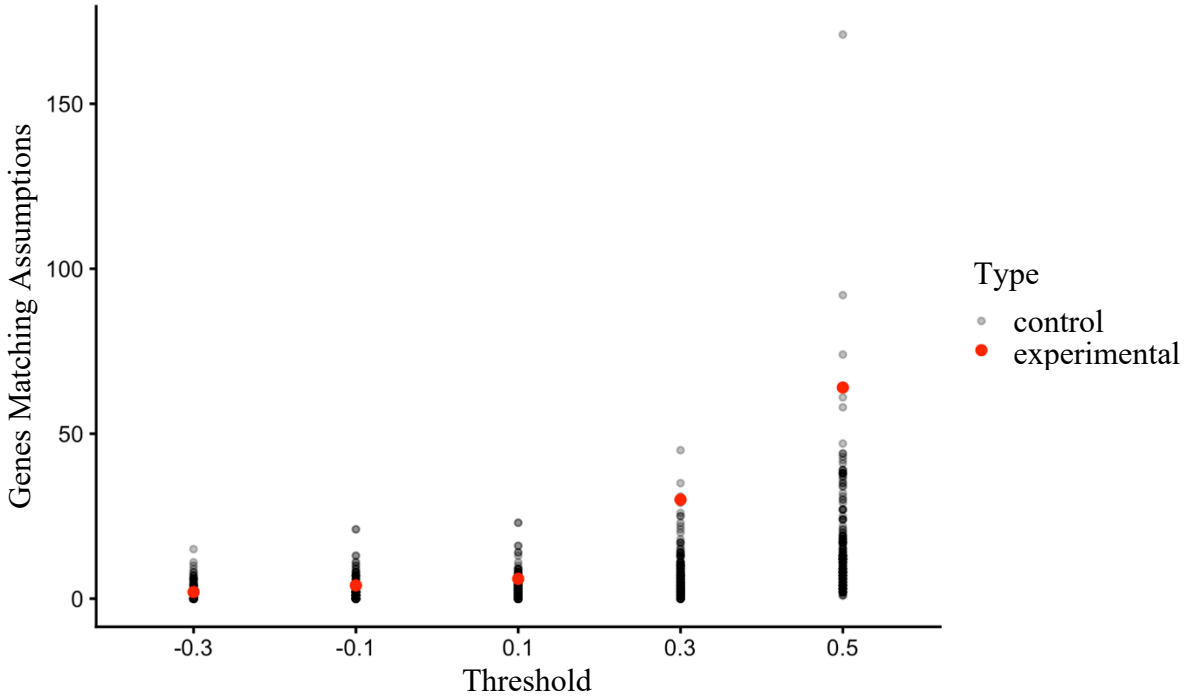


Figure 8. Total genes matching assumptions per threshold in analysis 1. Each point represents the number of candidate genes found in each trial of each threshold group in each control (black) and the experimental (red) permutation.

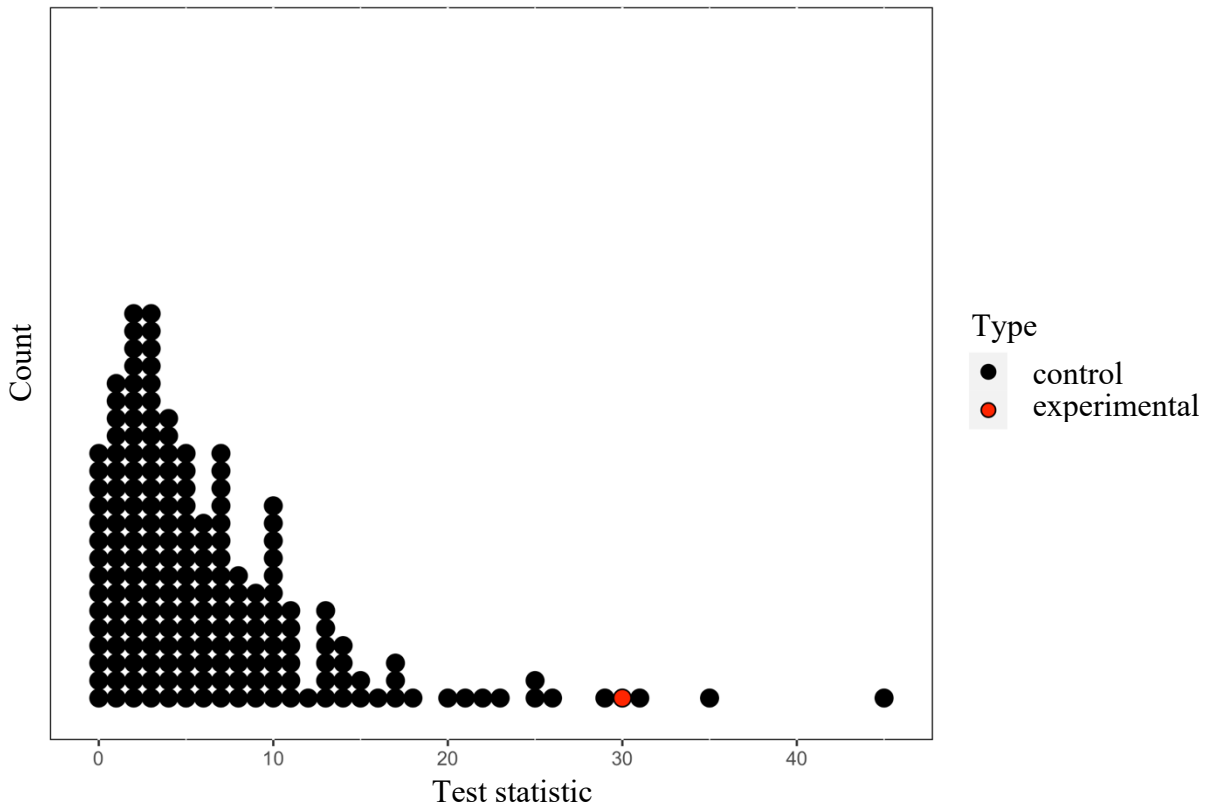


Figure 9. Distribution of test statistic for analysis 1 (number of genes found in the second to highest threshold step) in control vs experimental permutations. Stacked points represent the test statistic, the number of candidate genes found in threshold 0.3, in all control (black) and experimental (red) permutations.

Analysis 2 returned no genes matching Assumptions 1-3 until the sixth threshold (residual>0.293), when the mouse (*Mus musculus*) was added to the trait loss group alongside the marmoset (*Callitrhix jacchus*), hedgehog (*Erinaceus europaeus*), tree shrew (*Tupaia belangeri*), bushbaby (*Otolemur garnettii*), and kangaroo rat (*Dipodomys ordii*) (Figure 9). Only 5 genes matched Assumptions 1-3 when the species were divided by threshold 6, 7 in threshold 7, 16 in threshold 8, and 85 in threshold 9. All genes matching Assumptions 1-3 are available in the second table of Appendix B.

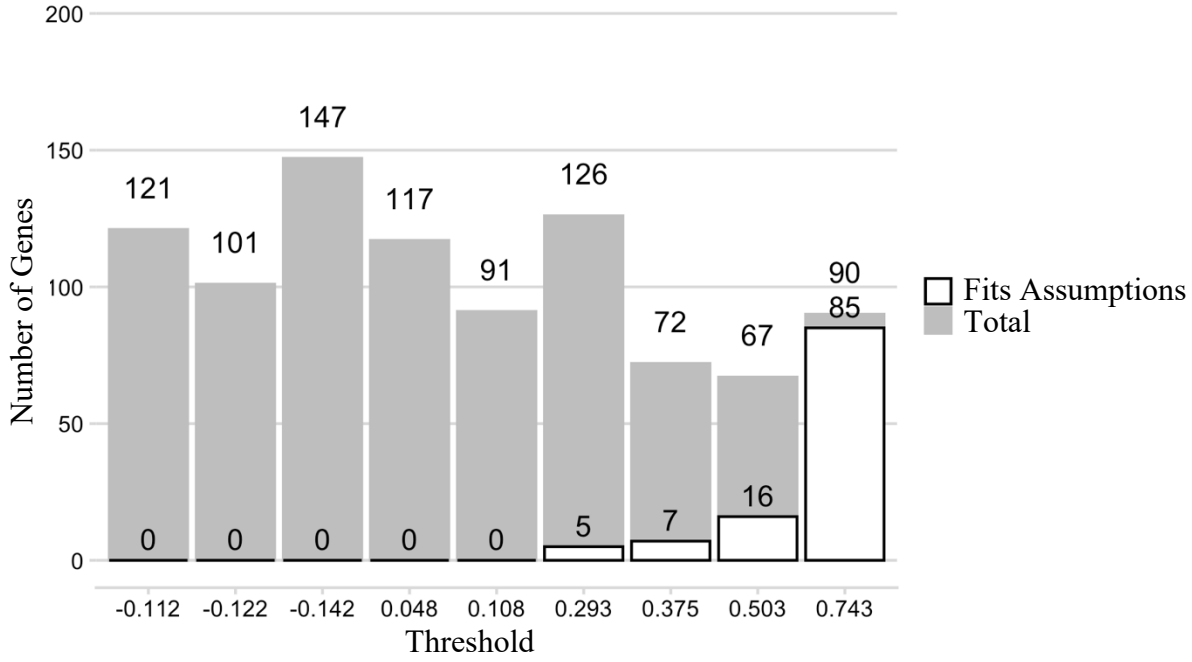


Figure 10. Total number of genes and number of genes fitting Assumptions 1-3 in the experimental permutation of analysis 2.

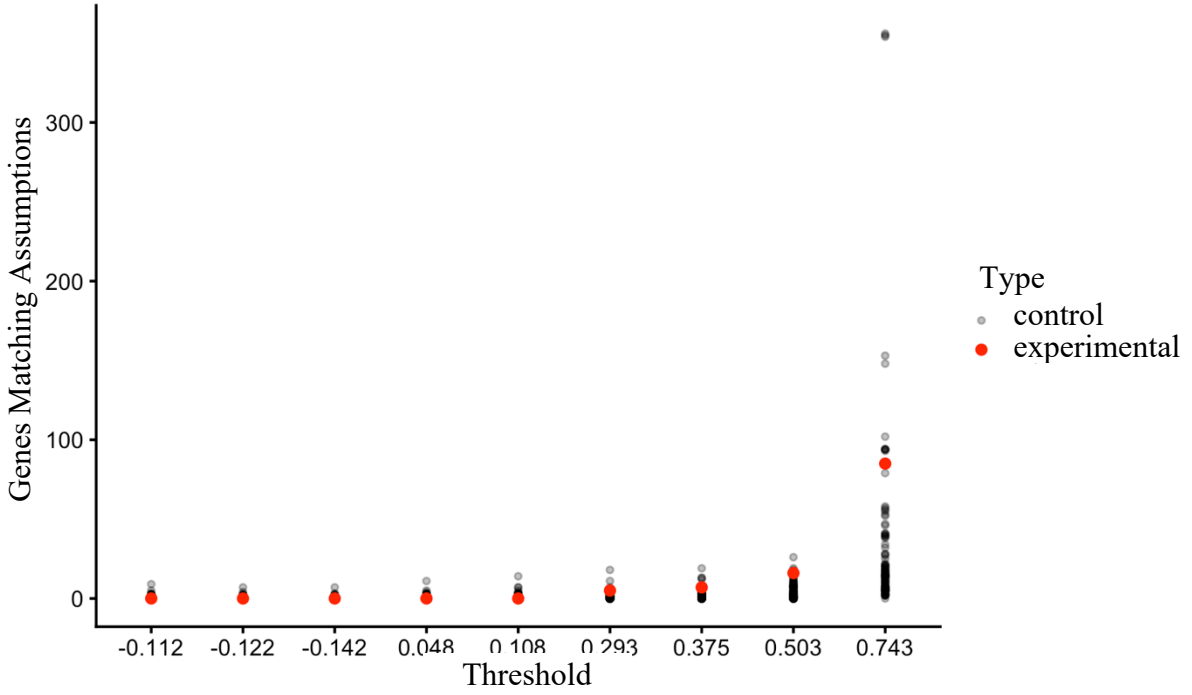


Figure 11. Total genes matching expectations per threshold in Analysis 2. Each point represents the number of candidate genes found in each trial of each threshold group in each control (black) and the experimental (red) permutation.

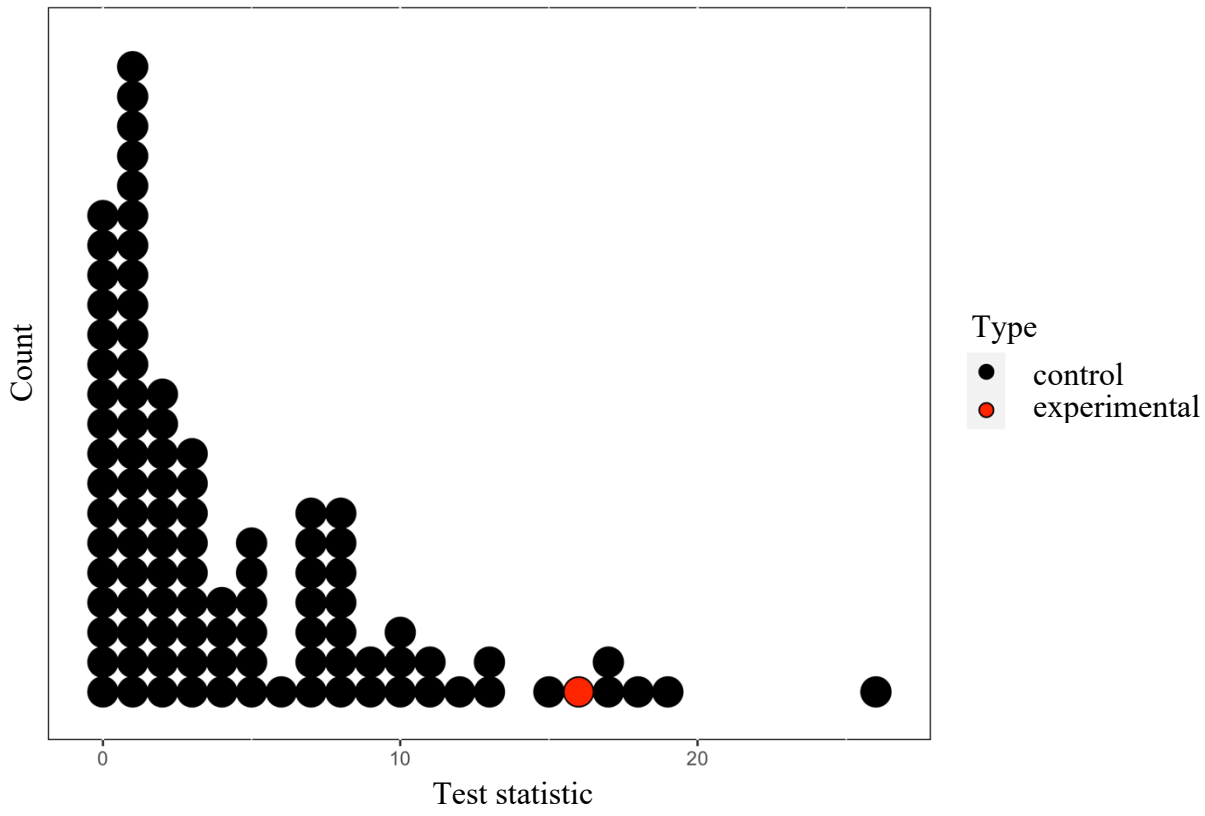


Figure 12. Distribution of test statistic for analysis 2 (number of genes found in the second to highest threshold step) in control vs experimental permutations. Stacked points represent the test statistic, the number of candidate genes found in threshold 0.3, in all control (black) and experimental (red) permutations.

Full tables of genes matching Assumptions 1-3 in analyses 1 and 2 are available in Appendix B. "NA" means that the gene was not more different from the ancestral sequence in more members of the high group than in the low group at that threshold, while "loss" means that it was more different from the ancestral sequence in all members of the high group than all members of the low group at that threshold.

Functional Enrichment of Resulting Genetic Regions

The genes identified as lost in analyses 1 and 2 yielded no statistically significant results when submitted to GOrilla. While there was one term in the function category associated more with the experimental results of analysis 1 than its controls and three

terms in the process category more associated with the results of analysis 2 than its controls, none of these results were statistically significant. The FDR q-value in the fifth column (Table 5, 6) is the false discovery rate, a measure of how likely the result is to appear due to random chance. Smaller values indicate a lower chance that a result is due to chance – a value of 0.05 for example would mean there is a 5% chance of getting that result due to chance alone. Despite the low uncorrected P-values, in the GOrilla results for analysis 1 the FDR q-value was 0.918 and in analysis 2 it was 1 for each term. This means it is highly likely that these results would appear as a result of random chance, and therefore, they should be discarded.

Table 5. GOrilla enrichment terms for analysis 1

Category	GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
Function	GO:0033293	monocarboxylic acid binding	0.00056	0.918	16.27 (2278,7,60,3)	Gstm7, Ptgds, Akr1c6

Table 6. GOrilla enrichment terms for analysis 2

Category	GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
Process	GO:0043303	mast cell degranulation	0.00079	1	35.34 (2827,2,80,2)	Cplx2, Ptgds
Process	GO:0002279	mast cell activation involved in immune response	0.00079	1	35.34 (2827,2,80,2)	Cplx2, Ptgds
Process	GO:0043299	leukocyte degranulation	0.00079	1	35.34 (2827,2,80,2)	Cplx2, Ptgds

The DAVID results for analyses 1 and 2 but were only slightly more significant (Table 7, 8). The false discovery rate (FDR) is a statistic showing the likelihood that a hit in the database is due to random chance (column 5, Tables 7, 8). A value of 1 is equivalent to a 100% chance that the result is due to random chance, and 0.05 is a widely

accepted false discovery rate (Huang et al. 2009a). False discovery rates in my results ranged from 0.11 to 1 in the results of both analyses. Regardless of the likely insignificance of the results, I researched the DAVID results for analysis 1 to understand in more detail what each term meant and which genes were associated with them. As noted in the methods, the results reported here are the gene ontology terms enriched in DAVID for the experimental permutation gene list but not the control permutation gene list. The analysis of the experimental permutation of analysis 1 alone was enriched for 33 terms and the longer control list with all genes found in every control was enriched for 155 terms. Only 7 terms (Table 7) were present in the experimental but not control results.

Table 7. DAVID enrichment terms for analysis 1

Category	Term	Count	Genes	FDR
INTERPRO	IPR018647:Domain of unknown function DUF2075	2	SLFN5, SLFN8	0.0945067
INTERPRO	IPR000215:Serpin family	3	SERPINA3B, SERPINA3G, SERPINA3H	0.11233917
INTERPRO	IPR023796:Serpin domain	3	SERPINA3B, SERPINA3G, SERPINA3H	0.11233917
UP_SEQ_FEATURE	region of interest:RCL	2	SERPINA3B, SERPINA3G	1
INTERPRO	IPR007421:ATPase, AAA-4	2	SLFN5, SLFN8	0.18945803
UP_SEQ_FEATURE	site:Reactive bond	2	SERPINA3B, SERPINA3G	1
INTERPRO	IPR027417:P-loop containing nucleoside triphosphate hydrolase	6	SLFN5, SLFN8, SULT1D1, IQCF4, ABCA14, GM12250	0.58170517

The “SM00093:SERPIN," "IPR000215:Serpin family," and “Serpin domain” terms were shared between the same three genes: SERPINA3B, SERPINA3G, and SERPINA3H. SERPINA3G and SERPINA3B were also responsible for the

“site:Reactive bond” and “region of interest:RCL” terms. "IPR018647:Domain of unknown function DUF2075” and “IPR007421:ATPase, AAA-4” were found in genes SLFN5 and SLFN8, genes which may have roles in hematopoietic cell differentiation and regulation of inflammation respectively (The UniProt Consortium 2021). “IPR027417:P-loop containing nucleoside triphosphate hydrolase” was referred to by genes SLFN5, SLFN8, SULT1D1, IQCF4, ABCA14, and GM12250. The term is assigned to genes that contain the P-loop NTPase fold, a nucleotide-binding protein fold (Leipe et al. 2002; Hunter et al. 2009). SULT1D1 is a sulfotransferase with many substrates for regulatory activity, IQCF4 is a pseudogene, ABCA14 is an ATP-binding cassette transporter, and GM12250 is a GTPase (The UniProt Consortium 2021).

None of genes expected to be associated with BMR based on prior speculation and human genome-wide association studies, were found in the results. It is possible not all of them had mouse orthologs to include in the analysis, as most were uncovered in humans. The only gene confirmed to be in the dataset was UCP1, which had appeared in the results during prior analysis not listed here where the shrew was placed in a category alone. Genes searched for were: UCP1 (Ricquier 2011), KLF5 (Choi et al. 2013), NRG3, OR8U8, BCL2L2-PABN1, PABN1, SLC22A17, FGGY, PTPRD, NPAS3, PKD1L2, SETBP1 (Lee et al. 2016), and GPR158 (Piaggi et al. 2017).

Table 8. DAVID enrichment terms for analysis 2

Category	Term	Count	Genes	FDR
UP_SEQ_FEATURE	compositionally biased region:Cys-rich	4	KRT33A, PLSCR4, ADAM1B, ADAM21	1
INTERPRO	IPR023795:Protease inhibitor I4, serpin, conserved site	3	SERPINA12, SERPINA3G, SERPINA3H	0.48873073
INTERPRO	IPR013733:Protein-arginine deiminase (PAD), central domain	2	PADI3, PADI4	0.48873073
INTERPRO	IPR013530:Protein-arginine deiminase, C-terminal	2	PADI3, PADI4	0.48873073
INTERPRO	IPR004303:Protein-arginine deiminase	2	PADI3, PADI4	0.48873073
INTERPRO	IPR013732:Protein-arginine deiminase (PAD) N-terminal	2	PADI3, PADI4	0.48873073
GOTERM_BP_DIRECT	GO:0018101~protein citrullination	2	PADI3, PADI4	1
SMART	SM00093:SERPIN	3	SERPINA12, SERPINA3G, SERPINA3H	0.55378325
GOTERM_MF_DIRECT	GO:0004668~protein-arginine deiminase activity	2	PADI3, PADI4	1
KEGG_PATHWAY	mmu00982:Drug metabolism - cytochrome P450	3	ALDH3B2, UGT2A1, FMO6	1
PIR_SUPERFAMILY	PIRSF001247:protein-arginine deiminase	2	PADI3, PADI4	0.13468544
INTERPRO	IPR023796:Serpin domain	3	SERPINA12, SERPINA3G, SERPINA3H	0.48873073
INTERPRO	IPR000215:Serpin family	3	SERPINA12, SERPINA3G, SERPINA3H	0.48873073
GOTERM_MF_DIRECT	GO:0005198~structural molecule activity	4	KRT33A, LAD1, KRT20, SPRR1A	1
INTERPRO	IPR008972:Cupredoxin	2	PADI3, PADI4	1
UP_KEYWORDS	Calmodulin-binding	3	MYH1, RIT2, SMTNL1	1
UP_KEYWORDS	Thiol protease inhibitor	2	CST8, SERPINA3G	1
INTERPRO	IPR027417:P-loop containing nucleoside triphosphate hydrolase	7	MYH1, SLFN5, PFKFB1, RIT2, GNAT3, SULT6B1, GM12250	1
INTERPRO	IPR002957:Keratin, type I	2	KRT33A, KRT20	1
GOTERM_BP_DIRECT	GO:0012501~programmed cell death	2	GSDMA, PDCD5	1

DISCUSSION

I adapted the Hiller forward genomics pipeline was to analyze continuous traits using progressively higher thresholds to convert them to binary traits and examined the results for significance using permutation testing. I found that the number of genes matching Assumptions 1-3 from my methods was significantly higher in the experimental the control group. However, when I searched for genes I found no increase in genes known to have any particular function compared to controls and found no genes known to be associated with BMR in my gene lists.

Number of Genes Matching Assumptions 1-3

The preliminary analysis showed that the number of genes lost in high BMR species in the experimental permutation was not different from the numbers lost in the same thresholds in the control permutations. Despite the low number of control permutations, some returned numbers of lost genes higher than the experimental permutation and some returned numbers lower (Figure 6). In analysis 1, which was the first trial I ran using the correct BMR data and with 200 controls, the number of lost genes in high BMR species at thresholds -0.3, -0.1, and 0.1 in the experimental permutation were still in the middle of distributions of numbers of genes lost in high BMR species at those thresholds in the control permutations. However, the number of genes lost in high BMR species at thresholds of 0.3 and 0.5 in the experimental permutation was significantly higher than the amount of genes lost in high BMR species at those thresholds in all but three out of 200 controls (Figure 8), indicating that more genes were lost in the experimental data compared to the controls at high BMR thresholds than at low ones.

There are two factors which likely explain this discrepancy between results in the preliminary test and analysis 1. First, I misidentified some species in the online tool during the preliminary test. The online forward genomics tool uses common names to refer to the species being analyzed. Some of the common names are ambiguous because they could refer to multiple species, such as "dolphin" and "megabat." While I was selecting phenotypes from the Genoud et al. (2018) dataset by hand to correspond with the animals present in the tool, I chose the first or only species matching the common name which also had phenotype data available. This method resulted in an incorrect selection of species (Table 2).

Later, during development of the pipeline, I identified the genome assemblies that actually corresponded to the common names in the tool with the help of the genome codes present in the URL of the page. When these correct species names were chosen, 11 species remained that both had BMR phenotypes in the Genoud et al. (2018) dataset and were present in the Forward Genomics tool (Table 1, 3). One species was identified only as "megabat" in the online tool. There were multiple different megabats with BMR residuals available in the Genoud et al. (2018) dataset with widely ranging values. *Pteropus giganteus*, the bat I initially assumed was the megabat in the tool, had a BMR residual of -0.084. The bat actually corresponding to the alignment in the tool, *Pteropus vampyrus*, had a BMR residual of 0.503. Incorrectly identifying one of the high BMR species as a low BMR species, along with including three species which did not end up having BMR data available at all, likely impacted the results. This is especially likely because the Forward Genomics pipeline requires strictly that a gene diverge from the ancestral state more in all of the trait loss species than all of the trait retaining species.

The second factor was that when I conducted the preliminary test, I set the thresholds slightly differently, at -0.1, 0, 0.1, 0.2, and 0.3. I changed the thresholds to -0.3, -0.1, 0.1, 0.3, and 0.5 in the experimental trial to prevent the final two thresholds splitting the species into the same two groups, as they would otherwise have done with the phenotype data (Table 1).

At a standard alpha level of .05, I can reject the null hypothesis that the number of genes meeting Assumptions 1-3 at the fourth threshold level in the experimental permutation of analysis 1 is the result of random chance. If Assumptions 1-3 set out in the methods for what a continuous trait should look like are correct, and if the test statistic correctly represents those assumptions, then this result indicates significantly more genes in the experimental trial met the assumptions expected of a gene associated with a continuous trait than in any control trial, and therefore, some of the genes found are most likely related to BMR.

When I selected the number of genes matching Assumptions 1-3 in the second to last threshold step as the test statistic, I hoped that it would be sufficient to quantify the number of genes matching Assumptions across the thresholds in the analysis. This hope relied on my unfounded preconception that the relationship between the numbers of genes in the different thresholds would be about the same across the experimental and control trials. This did not turn out to be the case. Instead, it is visually clear that the number of genes returned for each threshold in the experimental permutation starts out below the distribution of control permutation values in the first threshold used and ends near the top of the distribution by the final threshold in both analyses 1 and 2 (Figure 7, 10). This means that the chosen test statistic does not seem to account for the difference

between experimental and control permutations. The goal is to view genes associated with different levels of BMR, so it is more useful to look separately at where the number of results for the test permutation falls on the control distribution for each threshold.

Before discussing the specific relationships between the number of genes at each threshold in the distribution of control permutations and the experimental permutations, I would like to reiterate the purpose of the control distributions. In all analyses, I would expect that higher numbers of potentially trait-associated genes should appear at higher thresholds because assumption 2 outlined in the methods removes more results from lower thresholds than higher ones. The control distributions for each threshold (Figure 8, 11) show the increase in lost genes with increasing threshold that is expected as an artifact of Assumptions 1-3. The difference between the pattern of genetic regions returned by the experimental data and the pattern established by the control distributions is what I consider when analyzing the results. In analysis 1, it appears that the numbers of genetic regions found when species were divided by thresholds -0.3, -0.1, and 0.1 were in the middle of control distributions and increased to be near the top of the control distributions at thresholds 0.3 and 0.5 (Figure 7). This means that elevated levels of independent gene loss compared to the levels in the control permutations became visible as chimp, megabat, dolphin, and shrew were the only species remaining in the high BMR (derived) group at threshold and continued to occupy a similar position relative to the control distribution when only megabat, dolphin, and shrew remained in the high BMR group.

From analysis 1 alone, it would appear that the number of genes matching assumptions in the experimental permutation abruptly jumped to significant levels at

thresholds 0.3 and 0.5. However, analysis 2 shows a gradual increase of the number of genes in the experimental permutation from the bottom of the distribution of control values at threshold 0.108, intermediate at 0.293 and 0.375, and to near the top of the distribution of control values at thresholds 0.503 and 0.742 (Figure 11). This means the number of genes fitting assumptions in the experimental permutation increased in relation to the control distribution as the mouse, dog, chimp, and megabat, were moved from the high BMR to the low BMR group, leaving only dolphin and shrew together in the high BMR group at threshold 0.742.

In both analyses, as the threshold to be considered high BMR got higher and more species were moved to the low BMR (ancestral), the remaining high BMR species had an increasingly significant amount of independently lost genes. In both analyses, this occurred around the 0.3 BMR residual cutoff when chimp, megabat, dolphin and shrew were considered high BMR. Analysis 2, in which species were moved from the high to the low BMR group one by one, showed that the increase in the amount of lost genes in experimental and control permutations was gradual rather than abrupt. Together, this appears to indicate my analysis returned genes associated with BMR that were lost or changed from the ancestral state more in all in high BMR species than all low BMR species when high thresholds (>0.3) were used, but not when low thresholds were used. I am cautious in speculating about the fact that no pattern was present at low thresholds because the sample consisted of only 11 mammalian species and there was no comparison to ectothermic outgroups with much lower BMR than all mammals. Very low levels of mammalian BMR could be considered a derived trait if they are lower than

those in the last common ancestor they share with other mammals even though high BMR is a derived trait for mammals as a whole (Avaria-Llautureo et al. 2019).

A high level of BMR appears to be associated with changes to a consistent set of genes across taxa that have developed high BMR independently. If the genetic causes and effects of BMR were not shared between high BMR species, the associated genetic regions would not show up as results of the forward genomics analysis, as they would not be more different from the ancestral state in all high BMR species than all low BMR species. Furthermore, the results of analysis 2 seem to imply that high BMR was associated with changes to additional genetic regions as BMR increased instead of only more changes to the same genetic regions.

Regions which did show up in results could either be independently lost due to relaxation of selective pressures or independently changed convergently or divergently due to selective pressures. While the original paper publishing the forward genomics pipeline assumed that loss of percent identity with the ancestral state would be due to relaxed selection, it does not compare sequences among species in the derived group or filter convergent changes from appearing in the results (Hiller et al. 2012). In the case of BMR, it seems unlikely that genes used at low BMR became unused and experienced a relaxation of selection at high levels of BMR rather than simply changing as a cause or effect of high BMR. This is important, because it means the genes found are not necessarily limited to "low BMR" genes which have no function in high BMR organisms. Instead, they could be genes related to the trait of BMR in general.

In analysis 2, the same genetic regions changed independently as mammals developed high levels of BMR and an increasingly high amount of regions was

associated with high BMR levels as the threshold was raised. The results of analysis 2 appear to support the idea that variation on BMR in extant species acts on new genetic targets that were not associated with the initial development of high BMR in early mammals (Nespolo et al. 2011; Konarzewski and Książek 2013). The results do not, however, preclude the idea suggested by some researchers that some genetic regions related to BMR in early mammals still influence it in extant mammals (Hayes 2010; Nespolo et al. 2011; Konarzewski and Książek 2013). If those genetic regions are responsible for BMR variation, they did not behave in a way consistent with the assumptions of the forward genomics pipeline and my filtering assumptions. No genes at all were associated with differences between the high and low BMR groups when relatively low BMR mammals were considered part of the derived group in the early thresholds of analysis 2.

It is unclear whether genes first associated with BMR in early mammals should have shown up if present and meeting the assumptions of my methods. The genomic regions used in the forward genomics pipeline were conserved in vertebrates, not just mammals. The list of outgroup species the forward genomics pipeline used to produce its ancestral sequence reconstructions included mammal species which were outgroups to the species in the tool but also the chicken, zebra finch, and a species of lizard. I quickly checked for orthologs for genetic regions Slfn8 through Serpina3g using OrthoDB, and found that all of them have orthologs in the chicken genome. I do not know, however, whether they were considered conserved in non-mammalian species in the multi-alignment used by the forward genomics pipeline. It is possible that some were, but more

control over the parameters of the analysis would be necessary to constrain the analysis to genetic regions that were conserved in both mammalian and non-mammalian species.

Functional Enrichment

To determine if an particular functions were overrepresented in the lists of genes matching Assumptions 1-3 in analyses 1 and 2, I submitted the gene lists to GOrilla and DAVID gene ontology search tools. Neither GOrilla (Eden et al. 2007, 2009) nor DAVID (Huang et al. 2009a; b) gene ontology enrichment search tools returned statistically significant results. When I investigated the results of the GOrilla search on analysis 1 despite their lack of statistical significance, no obvious relationship to metabolism was shared among the ontology terms that were investigated or the genes that were assigned the terms. I cannot take this to mean that no genes are related to BMR. That would be very unlikely, because BMR has been found to be heritable and correlated with other traits in artificial selection studies on wild and laboratory organisms (Konarzewski et al. 2005; Gębczyński and Konarzewski 2009; Konarzewski and Książek 2013; Sadowska et al. 2015; Wone et al. 2015). In fact, this result does not necessarily contradict my other findings that BMR-related genes were likely present in the results of the analysis, elevating the number of genes found in the higher thresholds in the experimental permutations over the numbers found in those thresholds in the control permutations. All that it means is that the list of genes which changed in high BMR mammals were not significantly more related to any tissue expressions, functions, cellular components, or processes compared to the list of genes from the control permutations.

The lack of statistically significant results from DAVID and GOrilla could be due to the size of the gene lists that resulted from the analyses. Due to the low number of species present and low sensitivity required to get relevant results with the original 2012 forward genomics pipeline, the gene lists I got from analyses 1 and 2 were small. The target lists for analysis 1 and analysis 2 contained only 64 and 86 genes respectively. The documentation on DAVID suggests that gene lists between 100 and 2,000 are a good size for analysis with the tool, and analysis of smaller lists will be limited in its statistical power (Huang et al. 2009a). In the publication announcing GOrilla, it was tested on a set of 14,565 genes, giving an indication of its intended use case (Eden et al. 2009). Repeating this analysis using the more sensitive 2016 forward genomics pipeline (Prudent et al. 2016) on a larger set of species with more recent lists of conserved regions might yield a gene list with a length more suitable for DAVID or GOrilla gene ontology enrichment analysis.

The lack of enrichment in gene ontology terms could also accurately represent a lack of enrichment in ontology terms in BMR-associated genes. This could be because BMR-associated genes are likely to have many biological roles (Konarzewski and Książek 2013) and studies have not been conducted to identify and label many genes with ontology terms relating to BMR yet. In the molecular convergence pipeline created by Marcovitz *et al.* (2019), gene lists generated by the pipeline were examined for enrichment in tissue expression in certain tissues. In species which echolocated, genes which were associated with the cochlear ganglion in the brain, a region associated with sound processing, had more conserved amino acids than would be expected by chance. The authors used this as proof that the pipeline was working. However, it was only

possible to connect the tissue expression in the cochlear ganglion with hearing because prior laboratory research had shown that the region was related to echolocation. In the case of BMR, there is limited research to show which regions and processes are connected to it, so there may not be extensive data that could appear in a gene ontology search to confirm my methods. Results could also show no enrichment if the results returned by my method are the other traits which also covary with BMR rather than genes related to BMR itself (White and Kearney 2013).

I did not find any of the specific genes thought to be related to BMR based on previous studies in the results of analyses 1 or 2. Other than UCP-1, I was not sure whether any of these genes were present in the dataset at all, and therefore cannot comment on their lack of presence in the results of my analyses. In my initial trials that placed the shrew, the mammal with the highest mass-independent BMR residual, in the high BMR category alone, results showed that UCP-1 had changed more in the shrew compared to the reconstructed ancestral sequence than in all other mammals. The fact that UCP-1 showed up as a gene loss in the shrew confirmed that the gene was present in the dataset. It does not appear in the analyses included in this thesis because prior to the analyses included here, the settings on the forward genomics pipeline were changed to require two high-BMR organisms to have a gene in order to include it in the results. In analysis 2, the dolphin, which was the sole mammal alongside the shrew in the high BMR group when the highest threshold was used, either did not have UCP-1 included in its alignment or caused the sequence for the gene in the high BMR group not to vary more from the ancestral sequence in all high organisms than it did for all low organisms. The absence of UCP-1 in these results despite its presence in the dataset indicates that

changes in UCP-1 are not consistently associated with BMR over a certain level in mammals. This is consistent with findings that show UCP-1 is not related to BMR despite its role facilitating the conversion of energy into heat. For example, while transgenic mice that expressed UCP-1 in skeletal muscle displayed increased activity and heat loss, their BMR was not significantly different from wild type mice (Klaus et al. 2005).

CONCLUSIONS

My extension of the forward genomics pipeline uncovered a significantly higher number of independently changed genes using BMR data than in randomly generated control data but no enrichment for any known functions in the genes found. This is the opposite of results Marcovitz *et al.* (2019) found using their molecular convergence pipeline to investigate echolocation, aquatic lifestyle, and high-altitude habitat phenotypes. In their analysis, they confirmed prior research stating that there was not an overall higher amount of amino acid convergence in organisms sharing those phenotypes (Thomas and Hahn 2015; Zou and Zhang 2015) but found functional enrichment in convergent phenotype-related genes (Marcovitz et al. 2019). The differences between the results of the two methods could indicate that BMR is a trait more central to the biology of organisms and correlates with more molecular changes throughout the genome than echolocation, aquatic lifestyle, and high-altitude habitats. But given that a small number of genes returned by my analysis, it seems more likely that functional enrichment results were limited by the low sensitivity of the Hiller *et al.* (2012) pipeline or by the lack of genes labeled with BMR-related functions in prior studies.

My project provides two short lists of genes which may be lost or independently in association with a high position on the slow-fast BMR continuum in mammals (Appendix B) as well as an R script that can be readily applied to analyze data from any phenotype with my augmented version of the Hiller (2012) forward genomics pipeline. Applying this pipeline to additional phenotypes would provide additional context for the BMR results. If the augmented pipeline were to return a statistically significant amount of genes for other continuous traits, it could be worthwhile to build a new set of percent

identity values the many mammalian species recently sequenced (Genereux et al. 2020) and apply a similar binning approach with data from the more sensitive 2016 forward genomics pipeline (Prudent et al. 2016). If applying this pipeline to BMR, it would also be worthwhile to include a more robust comparison of mammalian conserved regions with those of their closely related outgroups. My results suggest that developing this approach further could help us understand the genetic underpinnings of BMR and its related traits.

REFERENCES

AUER, S. K., S. S. KILLEN, AND E. L. REZENDE. 2017. Resting vs. active: a meta-analysis of the intra- and inter-specific associations between minimum, sustained, and maximum metabolic rates in vertebrates. Functional Ecology 31:1728–1738.

AVARIA-LLAUTUREO, J., C. E. HERNÁNDEZ, E. RODRÍGUEZ-SERRANO, AND C. VENDITTI. 2019. The decoupled nature of basal metabolic rate and body temperature in endotherm evolution. Nature 572:651–654.

BENNETT, A. F., AND J. A. RUBEN. 1979. Endothermy and Activity in Vertebrates. Science 206:649–654.

BOILY, P. 2002. Individual variation in metabolic traits of wild nine-banded armadillos (Dasypus novemcinctus), and the aerobic capacity model for the evolution of endothermy. Journal of Experimental Biology 205:3207–3214.

CHOI, J. R., I.-S. KWON, D. Y. KWON, M.-S. KIM, AND M. LEE. 2013. TT Mutant Homozygote of Kruppel-like Factor 5 Is a Key Factor for Increasing Basal Metabolic Rate and Resting Metabolic Rate in Korean Elementary School Children. Genomics & Informatics 11:263–271.

CRAN TEAM, D. TEMPLE LANG, AND T. KALIBERA. 2013. XML: Tools for Parsing and Generating XML Within R and S-Plus.

DULLOO, A. G., AND S. SAMEC. 2001. Uncoupling proteins: their roles in adaptive thermogenesis and substrate metabolism reconsidered. The British Journal of Nutrition 86:123–139.

EDEN, E., D. LIPSON, S. YOGEV, AND Z. YAKHINI. 2007. Discovering Motifs in Ranked Lists of DNA Sequences. PLOS Computational Biology 3:e39.

EDEN, E., R. NAVON, I. STEINFELD, D. LIPSON, AND Z. YAKHINI. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:48.

GARLAND, T., JR., AND R. L. ALBUQUERQUE. 2017. Locomotion, Energetics, Performance, and Behavior: A Mammalian Perspective on Lizards, and Vice Versa. Integrative and Comparative Biology 57:252–266.

GĘBCZYŃSKI, A. K., AND M. KONARZEWSKI. 2009. Locomotor activity of mice divergently selected for basal metabolic rate: a test of hypotheses on the evolution of endothermy. Journal of Evolutionary Biology 22:1212–1220.

GENEREUX, D. P. ET AL. 2020. A comparative genomics multitool for scientific discovery and conservation. Nature 587:240–245.

GENOUD, M., K. ISLER, AND R. D. MARTIN. 2018. Comparative analyses of basal rate of metabolism in mammals: data selection does matter. Biological Reviews 93:404–438.

HAYES, J. P. 2010. Metabolic rates, genetic constraints, and the evolution of endothermy. Journal of Evolutionary Biology 23:1868–1877.

HILLER, M., B. T. SCHAAR, V. B. INDJEIAN, D. M. KINGSLEY, L. R. HAGEY, AND G. BEJERANO. 2012. A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Reports 2:817–823.

HUANG, D. W., B. T. SHERMAN, AND R. A. LEMPICKI. 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4:44–57.

HUANG, D. W., B. T. SHERMAN, AND R. A. LEMPICKI. 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37:1–13.

HUNTER, S. ET AL. 2009. InterPro: the integrative protein signature database. Nucleic Acids Research 37:D211–D215.

KIM, M. J., J. H. KIM, M.-S. KIM, H. J. YANG, M. LEE, AND D. Y. KWON. 2019. Metabolomics Associated with Genome-Wide Association Study Related to the Basal Metabolic Rate in Overweight/Obese Korean Women. Journal of Medicinal Food 22:499–507.

KLAUS, S., B. RUDOLPH, C. DOHRMANN, AND R. WEHR. 2005. Expression of uncoupling protein 1 in skeletal muscle decreases muscle energy efficiency and affects thermoregulation and substrate oxidation. Physiological Genomics 21:193–200.

KONARZEWSKI, M., AND A. KSIĄŻEK. 2013. Determinants of intra-specific variation in basal metabolic rate. Journal of Comparative Physiology B 183:27–41.

KONARZEWSKI, M., A. KSIAZEK, AND I. B. LAPO. 2005. Artificial selection on metabolic rates and related traits in rodents. Integrative and Comparative Biology 45:416–425.

LEE, M., D. Y. KWON, M.-S. KIM, C. R. CHOI, M.-Y. PARK, AND A. KIM. 2016. Genome-wide association study for the interaction between BMR and BMI in obese Korean women including overweight. Nutrition Research and Practice 10:115–124.

LEIPE, D. D., Y. I. WOLF, E. V. KOONIN, AND L. ARAVIND. 2002. Classification and evolution of P-loop GTPases and related ATPases11Edited by J. Thornton. Journal of Molecular Biology 317:41–72.

LOVEGROVE, B. G. 2000. The Zoogeography of Mammalian Basal Metabolic Rate. The American Naturalist 156:201–219.

LOVEGROVE, B. G. 2003. The influence of climate on the basal metabolic rate of small mammals: a slow-fast metabolic continuum. Journal of Comparative Physiology B 173:87–112.

LOVEGROVE, B. G. 2005. Seasonal thermoregulatory responses in mammals. Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology 175:231–247.

LOVEGROVE, B. G. 2012. The evolution of endothermy in Cenozoic mammals: a plesiomorphic-apomorphic continuum. Biological Reviews 87:128–162.

LOVEGROVE, B. G. 2017. A phenology of the evolution of endothermy in birds and mammals. Biological Reviews 92:1213–1240.

LV, J., M. BHATIA, AND X. WANG. 2017. Roles of Mitochondrial DNA in Energy Metabolism. Mitochondrial DNA and Diseases:71–83.

MARCOVITZ, A. ET AL. 2019. A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales. Proceedings of the National Academy of Sciences 116:21094–21103.

MCNAB, B. K. 2008. An analysis of the factors that influence the level and scaling of mammalian BMR. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 151:5–28.

NESPOLO, R. F., L. D. BACIGALUPE, C. C. FIGUEROA, P. KOTEJA, AND J. C. OPAZO. 2011. Using new tools to solve an old problem: the evolution of endothermy in vertebrates. Trends in Ecology & Evolution 26:414–423.

PIAGGI, P. ET AL. 2017. A Genome-Wide Association Study Using a Custom Genotyping Array Identifies Variants in GPR158 Associated With Reduced Energy Expenditure in American Indians. Diabetes 66:2284–2295.

PRUDENT, X., G. PARRA, P. SCHWEDE, J. G. ROSCITO, AND M. HILLER. 2016. Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations Between Species' Phenotypic and Genomic Differences. Molecular Biology and Evolution 33:2135–2150.

R CORE TEAM. 2013. R: A language and program for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RICQUIER, D. 2011. Uncoupling Protein 1 of Brown Adipocytes, the Only Uncoupler: A Historical Perspective. Frontiers in Endocrinology 2.

SADOWSKA, E. T. ET AL. 2005. Genetic Correlations Between Basal and Maximum Metabolic Rates in a Wild Rodent: Consequences for Evolution of Endothermy. Evolution 59:672–681.

SADOWSKA, E. T. ET AL. 2015. Evolution of basal metabolic rate in bank voles from a multidirectional selection experiment. Proceedings of the Royal Society B: Biological Sciences 282.

SEALE, J. L., AND J. M. CONWAY. 1999. Relationship between overnight energy expenditure and BMR measured in a room-sized calorimeter. European Journal of Clinical Nutrition 53:107.

SEEBACHER, F. 2020. Is Endothermy an Evolutionary By-Product? Trends in Ecology & Evolution 35:503–511.

SIEPEL, A. ET AL. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15:1034–1050.

THE UNIPROT CONSORTIUM. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49:D480–D489.

THOMAS, G. W. C., AND M. W. HAHN. 2015. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. Molecular Biology and Evolution 32:1232–1236.

WHITE, C. R., AND M. R. KEARNEY. 2013. Determinants of inter-specific variation in basal metabolic rate. Journal of Comparative Physiology B 183:1–26.

WICKHAM, H. ET AL. 2019. Welcome to the Tidyverse. Journal of Open Source Software 4:1686.

WILBER, J. 2019. Permutation Test: Visual Explanation. <https://www.jwilber.me/permutationtest/> (8 April 2021).

WONE, B. W. M. ET AL. 2015. A strong response to selection on mass-independent maximal metabolic rate without a correlated response in basal metabolic rate. Heredity 114:419–427.

ZOU, Z., AND J. ZHANG. 2015. No genome-wide protein sequence convergence for echolocation. Molecular Biology and Evolution 32:1237–1241.

APPENDICES

APPENDIX A: R SCRIPTS

These R scripts were written to allow the user to make changes to the dataset being used (with the user specifying the file path and names of the columns containing species names and synonyms and the phenotype to be investigated), number and value of thresholds, number of controls, and whether a "high" phenotype or a "low" phenotype is considered trait-loss by defining variables. It uses the packages Tidyverse (Wickham et al. 2019) for data manipulation and XML (CRAN Team et al. 2013) for HTML parsing. It can easily be applied to any dataset with species names and phenotypes exported as a tab-separated-values file from a spreadsheet program. Running the "00_phenotree_pipeline.R" script (Appendix 1) after assigning variables performs all of the following tasks in sequence:

Loading Phenotype. The first script in the R pipeline reads a tab-separated-values file containing a phenotype dataset and uses a user-supplied vector of column names containing the binomial species names and synonyms and a separate variable storing the name of the column of desired phenotypes to extract the named phenotype for each species present in both the online tool and the provided dataset. If a species is not present in the first column of column names provided, the script continues to try additional columns until they are exhausted.

Creating URLs. When the user inputs which mammals are trait-loss, trait-preserving, and "ignore," as well as selecting the number of allowed violations, the Forward Genomics tool appears to generate a URL that, when entered, instructs the script on the website to calculate the results. The URLs look like this:

http://phenotree.stanford.edu/public/html/main.py?numspecies=27&species_select_hg18=ignore&species_select_panTro2=loss&species_select_gorGor1=ignore&species_select_ponAbe2=ignore&species_select_rheMac2=ignore&species_select_calJac1=preserving&species_select_tarSyr1=ignore&species_select_micMur1=ignore&species_select_otoGar1=loss&species_select_tupBel1=ignore&species_select_mm9=loss&species_select_rn5=ignore&species_select_dipOrd1=loss&species_select_cavPor3=ignore&species_select_speTri1=ignore&species_select_oryCun1=ignore&species_select_ochPri2=ignore&species_select_vicPac1=ignore&species_select_turTru1=loss&species_select_bosTau4=ignore&species_select_equCab2=ignore&species_select_felCat3=ignore&species_select_canFam2=ignore&species_select_myoLuc1=ignore&species_select_pteVam1=loss&species_select_eriEur1=loss&species_select_sorAra1=loss&min_loss=1&min_preserving=1&num_violations=0

It is possible to generate a URL matching a desired set of loss, preserving, and ignore values for species without using the user interface on the website, saving considerable time. Instead, the script uses the *sprintf* function to insert "loss," "preserving," and "ignore" into the URL as a string according to values in a table generated from thresholds. These tables have species names in the first column and then additional columns for each user-supplied threshold and calculated values for whether a species is "loss" or "preserving" according to that threshold and whether a high or low phenotype should be loss. The URLs generated from these tables are then added to a new data frame, with one URL per threshold.

Downloading Results. The next script iterates over each data frame containing URLs in turn, downloading the source code of the pages the URLs point to. The source code for these pages does not contain the results, which are hosted separately in a link which changes with each submission. The link is present in the source code downloaded using the URL. The script obtains all links in the document using the getHTMLLinks

function from the XML package and the correct link can be searched for using grep because it is the only link to contain “tmp.” The “export.tsv” file corresponding to the request is downloaded from this link.

To avoid being mistaken for a DDoS attack or other malicious activity, the script waits 5 seconds between requests to the Stanford webpage hosting the Forward Genomics tool. Along with the time it takes the website to calculate results, this wait time means that it may take hours to take a sample of 100 controls with 5 thresholds each, which totals to 500 requests. Despite the time required, running this script is much faster and less tedious than interacting with the site manually.

Matching Assumptions. After saving the results from the web tool to the correct output folders for the experimental file and each control, the script loads the results and filters them to find the genes present in each trial that meet the assumptions described under the earlier “Development of Assumptions” heading. It proved difficult to solve the problem of keeping only genes which had the desired pattern of loss, remaining lost at all higher thresholds once they became lost at a lower one. An initial solution worked for the specific case that there were five thresholds but was not easy to generalize for a user-supplied number of thresholds.

Ultimately, the solution chosen was to iterate over the columns from right to left and test whether the contents of the row in that column were equal to “loss” or “NA.” If they were equal to loss, the statement dataframe[row, column]==”loss” would evaluate to the Boolean TRUE. If not, it would evaluate to FALSE. TRUE is equivalent to 1 in R while FALSE is equivalent to 0, so summing the row would return the number of lost genes present in that row up to the column being currently tested. If all cells to the right

of the column currently being tested have "loss," the sum of the logical statement on the cells in that row for those columns should be equal to the distance from the right of the column being tested. If the sum is any less, it means there is a gap, and that gene should only be included if the value of the cell in the column currently being tested is the one that is NA. (Because that means the program can't prove that the gene hasn't stopped being a loss, and the iteration for the next column over will remove it if not.)

Once filtered from the whole set of genes returned by the Forward Genomics tool, the potential candidate genes were assigned to their own dataframe, one dataframe per experimental or control trial. This script also calculated the value of the test statistic for each dataframe as it was created and added that to another frame.

Statistics. The final R script generates a p value based on the table of test statistics generated in the previous step as well as creating a visualization of the test statistic distribution.

00_phenotree_pipeline.R

```
# Setup

  # set project folder working directory
  setwd("/Users/levesquelab/Desktop/CWC_Phenotree_Pipeline")
  # load packages
  require(tidyverse)

  # set subdirectory for data and graphics
    # when making a new subdirectory, copy and rename the existing
directory before running script
    # instead of creating a directory from scratch
  subdirectory_name <- "BMR_fix_thresholds"
  setwd(subdirectory_name)

#### INPUT ####
# Load phenotype data for species represented in Phenotool

  # file path to phenotype dataset relative to project folder
  pheno_dataset_fp <- "data/phenotypes_to_investigate.tsv"
```

```
  # vector of column titles of columns containing species bionomials
and synonyms
  species_col_name <- c("species binomial", "species_synonyms")
  # column title of continuous phenotype of interest in the phenotype
dataset table
  phenotype_name <- "BMR.resids"

  source("../scripts/load_pheno_data.R")


# Create URLs to run Phenotree website with chosen thresholds
  # URL CSV saves at "data/url_table.csv"

  # desired thresholds as double vector
    # if pheno equal to threshold, species will be considered "high"
    # Thresholds should go from lower to higher number for the analysis
part to work
    # Don't pick a threshold with no species on one side of it for your
phenotype...
    # It won't return any results.
  bin_thresholds <- c(-0.3, -0.1, 0.1, 0.3, 0.5)

  # Is a high value a loss? (TRUE or FALSE?)
  highloss <- TRUE

  # Pick minimum preserving and loss species to have data for gene in
the Phenotree queries
    # (Currently same number for all steps, so keep in mind # species
in groups in the
    # first and final steps.)
  min_loss <- 2
  min_pres <- 1

  source("../scripts/create_bins.R")


# Create a user-defined number of controls, records of what was in
them, and also URLS for them
  # Running this will delete output data present in the folders so
remember that

  # Desired number of controls
  num_controls <- 200

  source("../scripts/create_controls.R")

#### OUTPUT ####

# Script to automatically input URLs and download results

# If not using this script:
  # As you run the thresholds with the URLs, click "export as tsv" and
save the tsv in
  # the matching folder that was created in phenotree_output.
  # Leave the files with the name "export.tsv"

  source("../scripts/scrape.R")
```

```
# Script to extract results which match assumptions and create test
statistic table

  source("../scripts/match_assumptions.R")

# Statistical analysis
  # Calculate P value

  source("../scripts/statistics.R")
```

load_pheno_data.R

```
# Purpose: Convert generic data TSV to format with names used in the
Phenotree tool
# and column with desired trait for each species that is present in
both the dataset
# and the tool.

# file reading
  # load file that has the list of species from the tool
species_in_tool_df <- read_csv("data/species_in_tool.csv")
  # load file with continuous traits and species name and pick relevant
columns
pheno_dataset_df <- read_tsv(pheno_dataset_fp)
pheno_dataset_df <- pheno_dataset_df[, c(species_col_name,
phenotype_name)]

# rename phenotype_name column so it works better in dplyr
pheno_dataset_df <- rename(pheno_dataset_df, pheno_col =
colnames(pheno_dataset_df[,phenotype_name]))

# pull out rows from data table that match species names from the tool
  # good place to practice pipe syntax in the future?
  # ahhh, I was trying to do it as a recursive function at first but
the for statement
  # works much better.

i <- 1

species_in_tool_df <- species_in_tool_df

species_in_tool_df$pheno_col <- NA

for (i in species_col_name) {

species_in_tool_df <- left_join(species_in_tool_df,
                           pheno_dataset_df[,c(i,"pheno_col")],
                           by= c("species"=i))

species_in_tool_df <- mutate(species_in_tool_df,
                                  pheno_col =
ifelse(is.na(pheno_col.x), pheno_col.y, pheno_col.x))

species_in_tool_df <- species_in_tool_df[,c(1,2,3,6)]
```

```
}

rm(i, pheno_dataset_df, pheno_dataset_fp, species_col_name)
```

create_bins.R

```
# Purpose: Take table from load_pheno_data.R and calculuate
preservation or loss according to
  # any amount of user-specified thresholds
  # output a URL for the phenotree tool for each threshold

# Making new column to tell whether species is a loss, retention, or
ignore for that threshold
# This is different depending on whether a high value or low value is
the loss condition

threshold_df <- species_in_tool_df
thresh_col_rec <- c()

  # Section for if high = loss
if (highloss==TRUE) {
  for(i in bin_thresholds) {
    thresh_col <- paste("threshold", i , sep="_")
    threshold_df <- mutate(threshold_df,
                           "threshold_{i}" := ifelse(is.na(pheno_col),
                                                     "ignore",
                                                     ifelse(pheno_col
>= i,
                                                            "loss",

"preserving")))
    # This is to make the URL creation step later easier
    thresh_col_rec <- append(thresh_col_rec, thresh_col)
  }
}
  # Nearly identical section for if low = loss (highloss=FALSE)
if (highloss==FALSE) {
  for(i in bin_thresholds) {
    thresh_col <- paste("threshold", i , sep="_")
    threshold_df <- mutate(threshold_df,
                           "threshold_{i}" := ifelse(is.na(pheno_col),
                                                     "ignore",
                                                     ifelse(pheno_col <
i,
                                                            "loss",

"preserving")))
    # This is to make the URL creation step later easier
    thresh_col_rec <- append(thresh_col_rec, thresh_col)
  }
}

rm(i, thresh_col)

write_csv(threshold_df,
"data/phenotree_input/experimental/input_table.csv")
```

```
# That part was fun!
# Now create the string to put into the webpage for each threshold..?
# I can do it in the form of a table for human use for now

url_prefix <-
"http://phenotree.stanford.edu/public/html/main.py?numspecies=27"
url_suffix <-
sprintf("&min_loss=%s&min_preserving=%s&num_violations=0",
                      min_loss, min_pres)

url_vec <- c()

for(i in thresh_col_rec) {
# You could also build the string using the repeating parts and codes
which would be fun to try
  url_middle <- do.call(sprintf, c(fmt =
"&species_select_hg18=%s&species_select_panTro2=%s&species_select_gorGo
r1=%s&species_select_ponAbe2=%s&species_select_rheMac2=%s&species_selec
t_calJac1=%s&species_select_tarSyr1=%s&species_select_micMur1=%s&specie
s_select_otoGar1=%s&species_select_tupBel1=%s&species_select_mm9=%s&spe
cies_select_rn5=%s&species_select_dipOrd1=%s&species_select_cavPor3=%s&
species_select_speTri1=%s&species_select_oryCun1=%s&species_select_ochP
ri2=%s&species_select_vicPac1=%s&species_select_turTru1=%s&species_sele
ct_bosTau4=%s&species_select_equCab2=%s&species_select_felCat3=%s&speci
es_select_canFam2=%s&species_select_myoLuc1=%s&species_select_pteVam1=%
s&species_select_eriEur1=%s&species_select_sorAra1=%s",
                                  as.list(threshold_df[[i]])))

  url_full <- paste(url_prefix, url_middle, url_suffix, sep="",
collapse=NULL)
  url_vec <- append(url_vec, url_full)
}
url_df <- as_tibble_col(thresh_col_rec, column_name = "threshold")
url_df$url <- url_vec

rm(i, url_middle, url_prefix, url_suffix, url_vec, url_full,
species_in_tool_df)

write_csv(url_df, "data/phenotree_input/experimental/url_table.csv")

rm(url_df)
```

create_controls.R

```
# Purpose: Create random controls that match the signature of
additional loss/gains
  # found in each threshold stage of the experimental data

# Get rid of ignored species because we want only the species used in
the experimental group
control_gen_df <- filter(threshold_df,
threshold_df[,thresh_col_rec[1]]!="ignore")

# Establish signature of experimental dataset

loss_signature <- c()
```

```
for(i in thresh_col_rec) {
  loss_signature <- append(loss_signature,
sum(control_gen_df[,i]=="loss"))
}

rm(i)

# Remove old threshold values from control_gen_df
control_gen_df <- control_gen_df[,c(1,2,3)]

# Generate data with same signature

num_controls <- 1:num_controls

# Delete old files in the phenotree_input folder with "control" in the
title
unlink (list.files("data/phenotree_input",
                   full.names = TRUE)[grep("control",

list.files("data/phenotree_input",
                                                        full.names =
TRUE))],
        recursive=TRUE)

for(j in num_controls) {

  control_gen_df$pheno_col <- sample(nrow(control_gen_df))

  lsindex <- 1

  for(i in thresh_col_rec) {

    control_gen_df <- mutate(control_gen_df,
                             "{i}" := ifelse(pheno_col <=
loss_signature[lsindex] , "loss", "preserving"))
    lsindex <- lsindex + 1
  }

  # Add back in the "ignore" data

  control_gen_df <- left_join(threshold_df[,c(1,2,3)],
                              control_gen_df)

  control_gen_df_temp <- control_gen_df[,c(1,2,3,4)]
  control_gen_df <- control_gen_df[,5:length(colnames(control_gen_df))]

  control_gen_df[is.na(control_gen_df)] <- "ignore"

  control_gen_df <- bind_cols(control_gen_df_temp,
                              control_gen_df)
  rm(control_gen_df_temp)

  # Create a folder and record of loss / preservation in the control to
go in folder with its URL

  dir.create(sprintf("data/phenotree_input/control_%s", j))
```

```
  write_csv(control_gen_df,
sprintf("data/phenotree_input/control_%s/input_table.csv",
                                      j))

  rm(i, lsindex)

  # Create the URLs themselves and put in the same folder
    # Copy of URL creation in create_bins.R, would be good to make a
function

  url_prefix <-
"http://phenotree.stanford.edu/public/html/main.py?numspecies=27"
  url_suffix <-
sprintf("&min_loss=%s&min_preserving=%s&num_violations=0",
                         min_loss, min_pres)

  url_vec <- c()

  for(i in thresh_col_rec) {
    url_middle <- do.call(sprintf, c(fmt =
"&species_select_hg18=%s&species_select_panTro2=%s&species_select_gorGo
r1=%s&species_select_ponAbe2=%s&species_select_rheMac2=%s&species_selec
t_calJac1=%s&species_select_tarSyr1=%s&species_select_micMur1=%s&specie
s_select_otoGar1=%s&species_select_tupBel1=%s&species_select_mm9=%s&spe
cies_select_rn5=%s&species_select_dipOrd1=%s&species_select_cavPor3=%s&
species_select_speTri1=%s&species_select_oryCun1=%s&species_select_ochP
ri2=%s&species_select_vicPac1=%s&species_select_turTru1=%s&species_sele
ct_bosTau4=%s&species_select_equCab2=%s&species_select_felCat3=%s&speci
es_select_canFam2=%s&species_select_myoLuc1=%s&species_select_pteVam1=%
s&species_select_eriEur1=%s&species_select_sorAra1=%s",
                                     as.list(control_gen_df[[i]])))
    url_full <- paste(url_prefix, url_middle, url_suffix, sep="",
collapse=NULL)
    url_vec <- append(url_vec, url_full)
  }
  url_df <- as_tibble_col(thresh_col_rec, column_name = "threshold")
  url_df$url <- url_vec

  write_csv(url_df,
sprintf("data/phenotree_input/control_%s/url_table.csv", j))

  rm(i, url_middle, url_prefix, url_suffix, url_vec, url_full, url_df)

}

rm(j, loss_signature, control_gen_df)

# Also create matching output directories for use later:

  # Delete old
  unlink (list.files("data/phenotree_output",
                     full.names = TRUE)[grep("control",

list.files("data/phenotree_output",
                                                          full.names =
TRUE))],
```

```
          recursive=TRUE)

  # Make new folders for each control and each threshold
  for (i in num_controls) {
    dir.create(sprintf("data/phenotree_output/control_%s", i))
    for (j in thresh_col_rec) {
    dir.create(sprintf("data/phenotree_output/control_%s/%s", i, j))
    }
  }

  rm(i, j)

  # Make output directory for experimental group since number of
thresholds could change:
  unlink (list.files("data/phenotree_output/experimental", full.names =
TRUE),
          recursive=TRUE)
  for (i in thresh_col_rec) {
    dir.create(sprintf("data/phenotree_output/experimental/%s", i))
  }
```

scrape.R

```
require(XML)

# Function to download outputs from the phenotree site given a certain
input URL
phenotree_downloader <- function (input_url, output_path){
  links <- getHTMLLinks(input_url,
                        externalOnly=TRUE,
                        xpQuery = "//a/@href",
                        relative = TRUE)
  if (sum(grepl("tmp", links))==1) {
    result_link <- grep("tmp", links)

    result_link <- links[result_link]

    download.file(result_link, output_path)

  }
  # Handle if there is no data for that combination on the website
  else {
    write_file(c("no_data"), output_path)
  }
}

# Use function on experimental data URL table

url_df <- read_csv("data/phenotree_input/experimental/url_table.csv")

for (i in seq_along(url_df$url)) {

  input_url <- url_df$url[i]
  output_path <-
sprintf("data/phenotree_output/experimental/%s/export.tsv",
url_df$threshold[i])
```

```
  phenotree_downloader(input_url=input_url, output_path = output_path)

  # Sleep for 10 seconds after each input/download to be nice to the
website
  Sys.sleep(10)
}

# Use function on each URL from each control's URL table
for (j in num_controls) {

  url_df <-
read_csv(sprintf("data/phenotree_input/control_%s/url_table.csv",j))

  for (i in seq_along(url_df$url)) {

    input_url <- url_df$url[i]
    output_path <-
sprintf("data/phenotree_output/control_%s/%s/export.tsv", j,
url_df$threshold[i])

    phenotree_downloader(input_url=input_url, output_path =
output_path)

    # Sleep for a number of seconds after each input/download to be
nice to the website
    Sys.sleep(5)
}

}

rm(i, j, input_url, output_path, phenotree_downloader)
```

match_assumptions.R

```
# read and trim tsv files to important info, also combining them into
one file
  # which marks lost genes present in each threshold step
  # "NA" means they weren't present in that step

# Decide which threshold will be the test statistic
  # second to last is length(thresh_col_rec)-1
test_stat_row <- length(thresh_col_rec)-1

#### EXPERIMENTAL ####

# If no data, make table for threshold with no genes
output_txt <-
read_file(sprintf("data/phenotree_output/experimental/%s/export.tsv",
                  thresh_col_rec[1]))

if (output_txt == "no_data") {
  tbl_colnames <- c("#gene symbol", "gene name", thresh_col_rec[1])
  output_df <- read_csv("\n", col_names = tbl_colnames) # all character
type
  rm(tbl_colnames)
```

```
} else {

  # If there is data, read it and prepare to match assumptions
  output_df <-
read_tsv(sprintf("data/phenotree_output/experimental/%s/export.tsv",
                                 thresh_col_rec[1]),
                         skip=2)
  output_df[,thresh_col_rec[1]] <- "loss"
  output_df <- output_df[,c(1,2,9)]
}

for(i in c(2:length(thresh_col_rec))) {

  # Check output as plain text to see if data is present
  output_txt <-
read_file(sprintf("data/phenotree_output/experimental/%s/export.tsv",
                                   thresh_col_rec[i]))

  # If no data, make blank output_df_temp
  if (output_txt == "no_data") {

    tbl_colnames <- c("#gene symbol",
                      "gene name",
                      thresh_col_rec[i])
    output_df_temp <- read_csv("\n",
                               col_names = tbl_colnames) # all
character type
    rm(tbl_colnames)

  } else {

    # If there is data, put it in output_df_temp
    output_df_temp <-
read_tsv(sprintf("data/phenotree_output/experimental/%s/export.tsv",
                                     thresh_col_rec[i]),
                                 skip=2)
    output_df_temp[,thresh_col_rec[i]] <- "loss"
    output_df_temp <- output_df_temp[,c(1,2,9)]
  }

  # Join output_df_temp to the previously made output_df
    output_df <- full_join(output_df,
                           output_df_temp,
                           by=c("#gene symbol", "gene name"))
}

rm(output_df_temp)

colnames(output_df)[1:2] <- c("gene_symbol", "gene_name")

write_csv(output_df,
"data/phenotree_output/experimental/total_genes.csv")

# make dataframe setup for observations about tests at each threshold

gene_counts <- tibble(colnames(output_df)[3:length(output_df)])
```

```
colnames(gene_counts) <- "threshold"

# count total amount of genes lost in each column of the gene list and
add them to a vector

temp_count <- c(sum(output_df[thresh_col_rec[1]]=="loss", na.rm=TRUE))

for (i in c(2:length(thresh_col_rec))) {
  temp_count <- append(temp_count,
sum(output_df[thresh_col_rec[i]]=="loss", na.rm=TRUE))
}

# Make that vector part of the gene_counts dataframe
gene_counts$"total_genes" <- c(temp_count)
rm(temp_count)

# Now add the number of genes that match assumptions to that in a new
column!
  # First separating out the genes so they can be recorded and counted
  # Oh my god this is such a better way of doing this than I had
before...
  # Starts at the last column

if (highloss==TRUE) {
  rev_output_df_length <- rev(c(3:length(output_df[1,])))
  # Make sure you're only working with things present in final
threshold to begin with
  output_df <- filter(output_df,
output_df[rev_output_df_length[1]]=="loss")
  # Keep rows where the last n columns are equal to
  for(i in seq_along(rev_output_df_length)) {
    output_df <- filter(output_df,

is.na(output_df[,rev_output_df_length[i]])==TRUE |
rowSums(output_df[,length(output_df):rev_output_df_length[i]]=="loss",
                                    na.rm=TRUE)==i)
}
}

#Same counting as earlier

# count total amount of genes lost in each column of the gene list and
add them to a vector

temp_count <- c(sum(output_df[thresh_col_rec[1]]=="loss", na.rm=TRUE))

for (i in c(2:length(thresh_col_rec))) {
  temp_count <- append(temp_count,
sum(output_df[thresh_col_rec[i]]=="loss", na.rm=TRUE))
}

write_csv(output_df,
"data/phenotree_output/experimental/candidate_genes.csv")

experimental_output_df <- output_df

# Make that vector part of the gene_counts dataframe
```

```
gene_counts$"candidate_genes" <- c(temp_count)
rm(temp_count)

write_csv(gene_counts,
"data/phenotree_output/experimental/gene_counts.csv")

# Calculate test statistic and create test statistic table
test_statistic <- gene_counts[[test_stat_row,3]]
test_statistic_df <- tribble(
  ~type, ~trial, ~test_statistic,
  "experimental", "experimental", test_statistic
)


#So now I have my gene list, potential candidate list, and list of the
counts of genes in each... Nice.
#Bar chart time!

# Barplot

# Why is this putting the steps out of order?
ggplot(gene_counts, aes(threshold)) +
  geom_col(aes(y=total_genes, fill="Total"))+
  geom_col(aes(y=candidate_genes, fill="Fits assumptions")) +
  geom_text(aes(y=total_genes, label = total_genes), vjust = -0.5) +
  geom_text(aes(y=candidate_genes, label = candidate_genes), vjust = -
0.5) +
  ggtitle("BMR Dataset") +
  xlab("Threshold") +
  ylim(0,400) +
  ylab("Number of Genes") +
  labs(fill="")

ggsave(
  "data/phenotree_output/experimental/summary_plot.png",
  plot=last_plot()
)

#### CONTROLS ####

for(j in num_controls) {

    # Check output as plain text to see if data is present
  output_txt <-
read_file(sprintf("data/phenotree_output/control_%s/%s/export.tsv",
                                  j,
                                  thresh_col_rec[1]))

  # If no data, make blank output_df with column names
  if (output_txt == "no_data") {

    tbl_colnames <- c("#gene symbol",
                      "gene name",
                      thresh_col_rec[1])
    output_df <- read_csv("\n",
                                col_names = tbl_colnames) # all
character type
```

```
    rm(tbl_colnames)

  # If data present, make output_df with data
  } else {
output_df <-
read_tsv(sprintf("data/phenotree_output/control_%s/%s/export.tsv",
                                j,
                                thresh_col_rec[1]),
                         skip=2)
output_df[,thresh_col_rec[1]] <- "loss"
output_df <- output_df[,c(1,2,9)]
}

for(i in c(2:length(thresh_col_rec))) {

  output_txt <-
read_file(sprintf("data/phenotree_output/control_%s/%s/export.tsv",
                                    j,
                                    thresh_col_rec[i]))

  # If no data, make blank output_df_temp
  if (output_txt == "no_data") {

    tbl_colnames <- c("#gene symbol",
                      "gene name",
                      thresh_col_rec[i])
    output_df_temp <- read_csv("\n",
                               col_names = tbl_colnames) # all
character type
    rm(tbl_colnames)

    # Or make output_df_temp with data
  } else {

  output_df_temp <-
read_tsv(sprintf("data/phenotree_output/control_%s/%s/export.tsv",
                                       j,
                                       thresh_col_rec[i]),
                                skip=2)
  output_df_temp[,thresh_col_rec[i]] <- "loss"
  output_df_temp <- output_df_temp[,c(1,2,9)]
  }

  output_df <- full_join(output_df,
                         output_df_temp,
                         by=c("#gene symbol", "gene name"))
}

rm(output_df_temp)

colnames(output_df)[1:2] <- c("gene_symbol", "gene_name")

write_csv(output_df,
sprintf("data/phenotree_output/control_%s/total_genes.csv", j))

# make dataframe setup for observations about tests at each threshold
```

```
gene_counts <- tibble(colnames(output_df)[3:length(output_df)])
colnames(gene_counts) <- "threshold"

# count total amount of genes lost in each column of the gene list and
add them to a vector

temp_count <- c(sum(output_df[thresh_col_rec[1]]=="loss", na.rm=TRUE))

for (i in c(2:length(thresh_col_rec))) {
  temp_count <- append(temp_count,
sum(output_df[thresh_col_rec[i]]=="loss", na.rm=TRUE))
}

# Make that vector part of the gene_counts dataframe
gene_counts$"total_genes" <- c(temp_count)
rm(temp_count)

# Now add the number of genes that match assumptions to that in a new
column!
# First separating out the genes so they can be recorded and counted
# Oh my god this is such a better way of doing this than I had
before...
# Starts at the last column

if (highloss==TRUE) {
  rev_output_df_length <- rev(c(3:length(output_df[1,])))
  # Make sure you're only working with things present in final
threshold to begin with
  output_df <- filter(output_df,
output_df[rev_output_df_length[1]]=="loss")
  # Keep rows where the last n columns are equal to
  for(i in seq_along(rev_output_df_length)) {
    output_df <- filter(output_df,

is.na(output_df[,rev_output_df_length[i]])==TRUE |
rowSums(output_df[,length(output_df):rev_output_df_length[i]]=="loss",

na.rm=TRUE)==i)
  }
}

#Same counting as earlier

# count total amount of genes lost in each column of the gene list and
add them to a vector

temp_count <- c(sum(output_df[thresh_col_rec[1]]=="loss", na.rm=TRUE))

for (i in c(2:length(thresh_col_rec))) {
  temp_count <- append(temp_count,
sum(output_df[thresh_col_rec[i]]=="loss", na.rm=TRUE))
}

write_csv(output_df,
sprintf("data/phenotree_output/control_%s/candidate_genes.csv", j))

# Make that vector part of the gene_counts dataframe
```

```
gene_counts$"candidate_genes" <- c(temp_count)
rm(temp_count)

write_csv(gene_counts,
sprintf("data/phenotree_output/control_%s/gene_counts.csv", j))

# Calculate test statistic and add to test statistic table
test_statistic <- gene_counts[[test_stat_row,3]]

test_statistic_df <-
  add_row(test_statistic_df,
          type = "control",
          trial = sprintf("control_%s", j),
          test_statistic = test_statistic)

}

rm(gene_counts,
   output_df,
   i,
   j,
   rev_output_df_length,
   test_stat_row
   )
```

statistics.R

```
# Count number of control test statistics higher than the experimental
one
test_stat_above_ct <- test_statistic_df$test_statistic
exp_test_stat <- test_stat_above_ct[1]
test_stat_above_ct <- test_stat_above_ct[2:length(test_stat_above_ct)]
test_stat_above_ct <- test_stat_above_ct[test_stat_above_ct >
exp_test_stat]
test_stat_above_ct <- length(test_stat_above_ct)

# Calculate P value with that!
P_value <- (test_stat_above_ct+1) / (length(num_controls) + 1)

p <- test_statistic_df %>%
  ggplot( aes(x=test_statistic, fill=type)) +
  geom_dotplot(binwidth=1) +
  ggtitle("Distribution of Test Statistics") +
  scale_fill_manual(values=c("black", "red")) +
  theme(
    plot.title = element_text(size=15),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    panel.grid.major.y = element_blank(),
    panel.grid.minor.y = element_blank()
  )

p
ggsave(
  "graphics/test_statistic_distribution.png",
  plot=p
```

```
)

# Create summary table and plot of entire trial

# Create initial count data from control group
gene_counts <-
read_csv("data/phenotree_output/experimental/gene_counts.csv")
gene_counts$threshold <- substring(gene_counts$threshold, 11)
#gene_counts$threshold <- as.double(gene_counts$threshold)

gene_counts <- add_column(gene_counts, "type" = c("experimental"),
.before="threshold")

# Add data from experimental groups
for (i in num_controls) {
  gene_counts_temp <-
read_csv(sprintf("data/phenotree_output/control_%s/gene_counts.csv",
i))
  gene_counts_temp$threshold <- substring(gene_counts_temp$threshold,
11)
  #gene_counts_temp$threshold <- as.double(gene_counts_temp$threshold)
  gene_counts_temp <- add_column(gene_counts_temp, "type" =
c("control"), .before="threshold")
  gene_counts <- rbind(gene_counts, gene_counts_temp)
}
rm(gene_counts_temp)
gene_counts$threshold <- factor(gene_counts$threshold,
levels=bin_thresholds)

# Generate plot
gene_counts %>%
  arrange(type) %>%
ggplot(aes(x=threshold, y=candidate_genes)) +
  geom_point(aes(col=type, shape=type, alpha=type)) +
  scale_shape_manual(values=c(20, 19)) +
  scale_alpha_manual(values=c(.3, 1)) +
  scale_color_manual(values=c("black", "red")) +
  ggtitle("Total Candidate Genes Per Threshold")

ggsave(
  "graphics/candidate_gene_summary.png",
  plot = last_plot()
)
```

gene_ontology.R

```
# Create enter delimited file of gene symbols for the target
(experimental) group
  GO_target <-
read_csv("data/phenotree_output/experimental/candidate_genes.csv")
  GO_target <- GO_target$gene_symbol
  write_lines(GO_target, "data/GO_target.txt")
  rm(GO_target)

# Create enter delimited file of gene symbols for background (control)
group
```

```
  GO_background <-
read_csv("data/phenotree_output/control_1/candidate_genes.csv")
  GO_background <- GO_background$gene_symbol

  for (i in 2:length(num_controls)) {
    GO_background_temp <-
read_csv(sprintf("data/phenotree_output/control_%s/candidate_genes.csv"
, i))
    GO_background_temp <- GO_background_temp$gene_symbol
    GO_background <- append(GO_background, GO_background_temp)
  }
  GO_background <- unique(GO_background)
  write_lines(GO_background, "data/GO_background.txt")
  rm(GO_background_temp)
```

APPENDIX B: FULL CANDIDATE GENE TABLES

Analysis 1

Table 9. Candidate genetic regions from analysis 1

MGI Gene Symbol	Gene Name	Threshold -0.3	Threshold -0.1	Threshold 0.1	Threshold 0.3	Threshold 0.5
Slfn8	schlafen 8 isoform 1	loss	loss	loss	loss	loss
Tmem181c-ps	Mus musculus adult male corpora quadrigemina cDNA, RIKEN full-length enriched library, clone:B230309D09 product:unclassifiable, full insert sequence.	loss	loss	loss	loss	loss
4932414N04Rik	hypothetical protein LOC75721	NA	loss	loss	loss	loss
A430033K04Rik	hypothetical protein LOC243308	NA	loss	loss	loss	loss
Gm13154	SubName: Full=Novel protein similar to Rex2; Flags: Fragment;	NA	NA	loss	loss	loss
Gm5631	hypothetical protein LOC434674	NA	NA	loss	loss	loss
Dcpp2	demilune cell and parotid protein 2	NA	NA	NA	loss	loss
Cts6	cathepsin 6	NA	NA	NA	loss	loss
Serpina3g	serine protease inhibitor A3G	NA	NA	NA	loss	loss
Vmn2r85	vomeronasal 2, receptor 85	NA	NA	NA	loss	loss
Vmn2r102	vomeronasal receptor Vmn2r102	NA	NA	NA	loss	loss
Zfp36l3	zinc finger protein 36-like 3	NA	NA	NA	loss	loss
Vmn2r88	vomeronasal 2, receptor 88	NA	NA	NA	loss	loss
Gm12250	predicted gene 12250	NA	NA	NA	loss	loss
Vmn2r18	vomeronasal 2, receptor 18	NA	NA	NA	loss	loss
Serpina3h	SubName: Full=Putative uncharacterized protein Serpina3h;	NA	NA	NA	loss	loss
Zfp607	zinc finger proten 607	NA	NA	NA	loss	loss
Olfr136	olfactory receptor 136	NA	NA	NA	loss	loss
Gm4846	flavin-containing monooxygenase 13	NA	NA	NA	loss	loss
Zfp780b	SubName: Full=Zinc finger protein 780B;	NA	NA	NA	loss	loss
Serpina3b	serine protease inhibitor A3B precursor	NA	NA	NA	loss	loss

Skint7	selection and upkeep of intraepithelial T-cells	NA	NA	NA	loss	loss
5430413K10Rik	hypothetical protein LOC433492	NA	NA	NA	loss	loss
AK050745	Mus musculus 9 days embryo whole body cDNA, RIKEN full-length enriched library, clone:D030014K19 product:SET domain, bifurcated 1, full insert sequence.	NA	NA	NA	loss	loss
Akr1c6	estradiol 17 beta-dehydrogenase 5	NA	NA	NA	loss	loss
Olfr622	olfactory receptor 622	NA	NA	NA	loss	loss
Olfr535	olfactory receptor 535	NA	NA	NA	loss	loss
Olfr1494	olfactory receptor 1494	NA	NA	NA	loss	loss
Olfr652	olfactory receptor 652	NA	NA	NA	loss	loss
Cyp2j7	cytochrome P450, family 2, subfamily j,	NA	NA	NA	loss	loss
B3gnt6	UDP-GlcNAc:betaGal	NA	NA	NA	NA	loss
Olfr1415	olfactory receptor 1415	NA	NA	NA	NA	loss
Iqcf4	IQ motif containing F4	NA	NA	NA	NA	loss
A630073D07Rik	hypothetical protein LOC381819	NA	NA	NA	NA	loss
Olfr415	olfactory receptor 415	NA	NA	NA	NA	loss
Zdhhc11	probable palmitoyltransferase ZDHHC11	NA	NA	NA	NA	loss
Slfn5	schlafen family member 5	NA	NA	NA	NA	loss
Clca6	calcium-activated chloride channel regulator 4	NA	NA	NA	NA	loss
Gstm7	glutathione S-transferase Mu 7	NA	NA	NA	NA	loss
Gas1	growth arrest-specific protein 1	NA	NA	NA	NA	loss
Gm12597	alpha-interferon precursor	NA	NA	NA	NA	loss
Sult1d1	sulfotransferase family 1D, member 1	NA	NA	NA	NA	loss
Rnaset2b	ribonuclease T2B	NA	NA	NA	NA	loss
BC089491	selenoprotein V	NA	NA	NA	NA	loss
Smtnl1	smoothelin-like protein 1	NA	NA	NA	NA	loss
2310033P09Rik	multiple myeloma tumor-associated protein 2	NA	NA	NA	NA	loss
Olfr1505	olfactory receptor 1505	NA	NA	NA	NA	loss
Gm5105	hypothetical protein LOC329763	NA	NA	NA	NA	loss
Abca14	ATP-binding cassette, sub-family A (ABC1),	NA	NA	NA	NA	loss

Olfr920	olfactory receptor 920	NA	NA	NA	NA	loss
Trim31	E3 ubiquitin-protein ligase TRIM31	NA	NA	NA	NA	loss
Adam1b	disintegrin and metalloproteinase	NA	NA	NA	NA	loss
Catsper1	cation channel sperm-associated protein 1	NA	NA	NA	NA	loss
Gpr25	probable G-protein coupled receptor 25	NA	NA	NA	NA	loss
Krt9	keratin, type I cytoskeletal 9	NA	NA	NA	NA	loss
Olfr871	olfactory receptor 871	NA	NA	NA	NA	loss
Taf3	transcription initiation factor TFIID subunit 3	NA	NA	NA	NA	loss
Gm561	hypothetical protein LOC228715	NA	NA	NA	NA	loss
Ptgds	prostaglandin-H2 D-isomerase	NA	NA	NA	NA	loss
2200002J24Rik	hypothetical protein LOC69147	NA	NA	NA	NA	loss
Olfr412	olfactory receptor 412	NA	NA	NA	NA	loss
Man2b2	epididymis-specific alpha-mannosidase precursor	NA	NA	NA	NA	loss
Lor	loricrin	NA	NA	NA	NA	loss
Naca	nascent polypeptide-associated complex subunit	NA	NA	NA	NA	loss

Analysis 2

Table 10. Candidate genetic regions from analysis 2

MGI Gene symbol	Gene name	Threshold -0.142	Threshold -0.122	Threshold -0.112	Threshold 0.048	Threshold 0.108	Threshold 0.293	Threshold 0.375	Threshold 0.503	Threshold 0.743
Serpina3g	serine protease inhibitor A3G	NA	NA	NA	NA	NA	loss	loss	loss	loss
Serpina3h	SubName: Full=Putative uncharacterized protein Serpina3h;	NA	NA	NA	NA	NA	loss	loss	loss	loss
Olfr136	olfactory receptor 136	NA	NA	NA	NA	NA	loss	loss	loss	loss
Zfp780b	SubName: Full=Zinc finger protein 780B;	NA	NA	NA	NA	NA	loss	loss	loss	loss
AK050745	Mus musculus 9 days embryo whole body cDNA, RIKEN full-	NA	NA	NA	NA	NA	loss	loss	loss	loss

	length enriched library, clone:D030014K19 product:SET domain, bifurcated 1, full insert sequence.									
Gm12250	predicted gene 12250	NA	NA	NA	NA	NA	NA	loss	loss	loss
Olfr652	olfactory receptor 652	NA	NA	NA	NA	NA	NA	loss	loss	loss
Olfr415	olfactory receptor 415	NA	NA	NA	NA	NA	NA	NA	loss	loss
Slfn5	schlafen family member 5	NA	NA	NA	NA	NA	NA	NA	loss	loss
Gm12597	alpha-interferon precursor	NA	NA	NA	NA	NA	NA	NA	loss	loss
Smtnl1	smoothelin-like protein 1	NA	NA	NA	NA	NA	NA	NA	loss	loss
Adam1b	disintegrin and metalloproteinase	NA	NA	NA	NA	NA	NA	NA	loss	loss
Taf3	transcription initiation factor TFIID subunit 3	NA	NA	NA	NA	NA	NA	NA	loss	loss
Ptgds	prostaglandin-H2 D-isomerase	NA	NA	NA	NA	NA	NA	NA	loss	loss
2200002J24Rik	hypothetical protein LOC69147	NA	NA	NA	NA	NA	NA	NA	loss	loss
Man2b2	epididymis-specific alpha-mannosidase precursor	NA	NA	NA	NA	NA	NA	NA	loss	loss
2310057J18Rik	hypothetical protein LOC67719 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Padi3	protein-arginine deiminase type-3	NA	NA	NA	NA	NA	NA	NA	NA	loss
Krt33a	keratin, type I cuticular Ha3-I	NA	NA	NA	NA	NA	NA	NA	NA	loss
Aldh3b2	aldehyde dehydrogenase 3 family, member B2	NA	NA	NA	NA	NA	NA	NA	NA	loss
Padi4	protein-arginine	NA	NA	NA	NA	NA	NA	NA	NA	loss

	deiminase type-4									
E430018J23Rik	hypothetical protein LOC101604	NA	NA	NA	NA	NA	NA	NA	NA	loss
Bet3l	trafficking protein particle complex subunit	NA	NA	NA	NA	NA	NA	NA	NA	loss
Gsdma	gasdermin-A	NA	NA	NA	NA	NA	NA	NA	NA	loss
Klrg1	killer cell lectin-like receptor subfamily G	NA	NA	NA	NA	NA	NA	NA	NA	loss
Plscr4	phospholipid scramblase 4	NA	NA	NA	NA	NA	NA	NA	NA	loss
Pfkfb1	6-phosphofructo-2-kinase/fructose-2,	NA	NA	NA	NA	NA	NA	NA	NA	loss
Cplx2	complexin-2	NA	NA	NA	NA	NA	NA	NA	NA	loss
Serpina12	serpin A12 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Rit2	GTP-binding protein Rit2	NA	NA	NA	NA	NA	NA	NA	NA	loss
Trmt2b	tRNA (uracil-5-)-methyltransferase homolog	NA	NA	NA	NA	NA	NA	NA	NA	loss
Pdia4	protein disulfide-isomerase A4	NA	NA	NA	NA	NA	NA	NA	NA	loss
Hisppd1	histidine acid phosphatase domain containing 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Olfr654	olfactory receptor 654	NA	NA	NA	NA	NA	NA	NA	NA	loss
9230105E10Rik	tripartite motif protein TRIM5	NA	NA	NA	NA	NA	NA	NA	NA	loss
Anxa9	annexin A9	NA	NA	NA	NA	NA	NA	NA	NA	loss
Opn3	opsin-3	NA	NA	NA	NA	NA	NA	NA	NA	loss
2010109K11Rik	hypothetical protein LOC72123	NA	NA	NA	NA	NA	NA	NA	NA	loss
Zc3h12d	probable ribonuclease ZC3H12D	NA	NA	NA	NA	NA	NA	NA	NA	loss
Myh1	myosin-1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Mfsd9	major facilitator superfamily	NA	NA	NA	NA	NA	NA	NA	NA	loss

	domain-containing									
Adam21	disintegrin and metalloproteinase	NA	NA	NA	NA	NA	NA	NA	NA	loss
Fmo6	flavin containing monooxygenase 6	NA	NA	NA	NA	NA	NA	NA	NA	loss
Sult6b1	sulfotransferase 6B1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Arrb2	beta-arrestin-2	NA	NA	NA	NA	NA	NA	NA	NA	loss
BC048599	putative trypsin-X3 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Fbrsl1	fibrosin-like 1 isoform 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Prdx2	peroxiredoxin-2	NA	NA	NA	NA	NA	NA	NA	NA	loss
Ddi1	protein DDI1 homolog 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Gnat3	guanine nucleotide-binding protein G(t) subunit	NA	NA	NA	NA	NA	NA	NA	NA	loss
Pdcd5	programmed cell death protein 5	NA	NA	NA	NA	NA	NA	NA	NA	loss
Ttll13	tubulin polyglutamylase TTLL13	NA	NA	NA	NA	NA	NA	NA	NA	loss
Prrt4	proline-rich transmembrane protein 4	NA	NA	NA	NA	NA	NA	NA	NA	loss
Krt20	keratin, type I cytoskeletal 20	NA	NA	NA	NA	NA	NA	NA	NA	loss
BC109180	Mus musculus activated spleen cDNA, RIKEN full-length enriched library, clone:F830003B07 product:hypothetical protein, full insert sequence.	NA	NA	NA	NA	NA	NA	NA	NA	loss
M13677	Mouse T-cell receptor active beta-	NA	NA	NA	NA	NA	NA	NA	NA	loss

	chain V-region V2DJ mRNA.									
Il17rb	interleukin-17 receptor B precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Olfr437	olfactory receptor 437	NA	NA	NA	NA	NA	NA	NA	NA	loss
Trpa1	transient receptor potential cation channel	NA	NA	NA	NA	NA	NA	NA	NA	loss
1700012B07Rik	hypothetical protein LOC69324 isoform 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Odf1	outer dense fiber protein 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Plag1	zinc finger protein PLAG1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Cyp11b2	cytochrome P450 11B2, mitochondrial	NA	NA	NA	NA	NA	NA	NA	NA	loss
Sgk3	serine/threonine-protein kinase Sgk3	NA	NA	NA	NA	NA	NA	NA	NA	loss
1700020C07Rik	Tsg23	NA	NA	NA	NA	NA	NA	NA	NA	loss
Hsd17b11	estradiol 17-beta-dehydrogenase 11	NA	NA	NA	NA	NA	NA	NA	NA	loss
Ugt2a1	UDP-glucuronosyltransferase 2A1 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Cst8	cystatin-8 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Fgf23	fibroblast growth factor 23 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
Zfp174	zinc finger protein 174	NA	NA	NA	NA	NA	NA	NA	NA	loss
Ttc1	tetratricopeptide repeat protein 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
4921507P07Rik	hypothetical protein LOC70821	NA	NA	NA	NA	NA	NA	NA	NA	loss
Fam53a	dorsal neural-tube nuclear protein	NA	NA	NA	NA	NA	NA	NA	NA	loss
Lad1	ladinin-1	NA	NA	NA	NA	NA	NA	NA	NA	loss

AK040202	Mus musculus 0 day neonate thymus cDNA, RIKEN full-length enriched library, clone:A430077D02 product:unclassifiable, full insert sequence.	NA	NA	NA	NA	NA	NA	NA	NA	loss
Elk4	ETS domain-containing protein Elk-4	NA	NA	NA	NA	NA	NA	NA	NA	loss
Rell1	RELT-like protein 1 precursor	NA	NA	NA	NA	NA	NA	NA	NA	loss
4930588N13Rik	hypothetical protein LOC75860	NA	NA	NA	NA	NA	NA	NA	NA	loss
Pram1	PML-RARA-regulated adapter molecule 1	NA	NA	NA	NA	NA	NA	NA	NA	loss
Cdkl4	cyclin-dependent kinase-like 4	NA	NA	NA	NA	NA	NA	NA	NA	loss
Pde7a	high affinity cAMP-specific 3',5'-cyclic	NA	NA	NA	NA	NA	NA	NA	NA	loss
Arhgef6	rho guanine nucleotide exchange factor 6	NA	NA	NA	NA	NA	NA	NA	NA	loss
Adcy6	adenylate cyclase type 6	NA	NA	NA	NA	NA	NA	NA	NA	loss
Sprr1a	cornifin-A	NA	NA	NA	NA	NA	NA	NA	NA	loss
Xpnpep3	probable Xaa-Pro aminopeptidase 3	NA	NA	NA	NA	NA	NA	NA	NA	loss

AUTHOR'S BIOGRAPHY

Caleigh W. Charlebois was born in Southern California on May 18, 1998. Her family moved to China, Maine when she was very young, and she graduated from Erskine Academy in 2016. At University of Maine, she is majoring in zoology with a minor in professional writing. She was a recipient of the Goldwater Award in 2020 and the Wallace C. and Janet S. Dunham Prize in 2021.

After graduation, she plans to pursue a Master's program in Zoology with the Irwin lab at the University of British Columbia in Vancouver, where she has been accepted with an NSERC CGS-M scholarship. She also hopes to gain experience in science writing and communication and continue to pursue her musical and artistic hobbies.