Title
Gene regulation and the genomic basis of speciation and adaptation in house mice (Mus musculus)

Author
Mack, Katya L

Publication Date
2018

Peer reviewed|Thesis/dissertation

Gene regulation and the genomic basis of speciation and adaptation in house mice (*Mus musculus*)

By

Katya L. Mack

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael W. Nachman, Chair
Professor Rasmus Nielsen
Professor Craig T. Miller

Fall 2018

Abstract

Gene regulation and the genomic basis of speciation and adaptation in house mice (*Mus musculus*)

by

Katya Mack

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Michael W. Nachman, Chair

Gene expression is a molecular phenotype that is essential to organismal form and fitness. However, how gene regulation evolves over evolutionary time and contributes to phenotypic differences within and between species is still not well understood. In my dissertation, I examined the role of gene regulation in adaptation and speciation in house mice (*Mus musculus*).

In chapter 1, I reviewed theoretical models and empirical data on the role of gene regulation in the origin of new species. I discuss how regulatory divergence between species can result in hybrid dysfunction and point to areas that could benefit from future research.

In chapter 2, I characterized regulatory divergence between *M. m. domesticus* and *M. m. musculus* associated with male hybrid sterility. The major model for the evolution of post-zygotic isolation proposes that hybrid sterility or inviability will evolve as a product of deleterious interactions (i.e., negative epistasis) between alleles at different loci when joined together in hybrids. As the regulation of gene expression is inherently based on interactions between loci, disruption of gene regulation in hybrids may be a common mechanism for post-zygotic isolation. To test this question, I compared expression differences between house mouse subspecies with expression patterns in sterile and fertile male F1 hybrids. I identified extensive regulatory divergence between the subspecies in the testis and found that compensatory *cis*-and *trans*- changes were non-randomly associated with genes that were misexpressed in sterile hybrids, but not in fertile hybrids. These results support the idea that such regulatory interactions may contribute to hybrid incompatibilities and may be major drivers of speciation.

In my third chapter, I used expression quantitative trait locus (eQTL) mapping in tandem with a genome scan for selection to identify adaptive regulatory variation in house mice (*M. m. domesticus*) on the east coast of North America. Mice on the east coast of North America show adaptive differences in body mass in response to latitudinal variation in temperature. I identified genes with clinally varying *cis*-eQTL

for which expression level is correlated with latitude. Among these clinal outliers, I identified two genes (*Adam17* and *Bcat2*) with *cis*-eQTL of large effect that are associated with adaptive body mass variation and for which expression is correlated with body mass both within and between populations. *Adam17* and *Bcat2* deletions affect body mass in mice and these genes have also been linked to obesity in humans. These findings provide strong evidence for *cis*- regulatory elements as essential loci of environmental adaptation in natural populations.

In chapter 4, I used low-coverage whole genome sequencing data from the same individuals to identify and characterize copy number variation in natural populations of house mice on the east coast of North America. Consistent with a role for copy number variation in local adaptation, I identified two regions where copy number is significantly correlated with latitude. One of these regions contains the gene *Trpm8*, which has previously been shown to affect physiological responses to environmental cold in other species. These results suggest that copy number variation significantly contributes to genetic variation in North American house mice and that copy number variation may play an important role in local adaptation.

Finally, in my fifth chapter, I examined the relationship between gene co-expression networks and molecular evolution in natural populations of *M. m. domesticus.* I found that genes that are more central to their co-expression networks (i.e., have more and greater associations with other genes) and genes whose co-expression relationships are conserved across tissues show lower rates of protein evolution, have fewer polymorphisms, and are less likely show regulatory variation. Together, these results are consistent with gene co-expression network structure being a source of evolutionary constraint.

Acknowledgements

The work was generously funded by several grants, including a Doctorate Dissertation Improvement Grant from the NSF; Jerry O. Wolff Fellowship, Wilhelm L. F. Martens Fund, and Louise Kellogg Fund from the Museum of Vertebrate Zoology (MVZ) at UC Berkeley; Graduate Dean's Summer Research Grant and Graduate Division Summer Grant from the UC Berkeley Graduate Division. The work was also supported a NIH grant to Michael Nachman.

First, I would like to thank my advisor and committee chair, Dr. Michael Nachman, for all his guidance, support, and wisdom over the years. I would also like to thank my committee members, Drs. Craig Miller and Rasmus Nielsen, for useful discussions and comments.

I would like to thank members of the Nachman lab and the MVZ/IB community for creating a supportive research environment: Dr. P. Campbell, Dr. M. Phifer-Rixey, Dr. M. Sheehan, Dr. C. Emerling, Dr. T. Suzuki, Dr. K. Ferris, Dr. G. Bradburd, Dr. A. Moeller, M. Ballinger, and S. Banker, E. Voss, D. Manahan, L. Smith, Dr. K. Bi, Dr. J. McGuire, Dr. N. Whiteman, and Dr. E. Rosenblum. A special thanks to N. Bittner whose friendship and kindness helped keep me afloat.

I would also like to express my gratitude to those who – very generously – provided me with opportunities and guidance as an undergraduate that allowed me to pursue a graduate degree: Dr. Patricia Wittkopp, Dr. Milford Wolpoff, and Dr. Jonathan Gruber.

Finally, I would like to thank my family for their enduring support over the last several years (and always): to my father, Julian Mack, who has always supported the nerdiest versions of his daughters; to my sister, Mahalia Mack, whose generosity of spirit is unparalleled; and to all the other Mack-Widman-Hagood-Greys whose energy and love is a constant source of inspiration. Finally, to my husband, Jeremy Richardson, for who words cannot express my gratitude, but I hope he knows what is within my heart.

Chapter 1

Gene Regulation and Speciation

This chapter has been previously published and is reproduced here in accordance with the journal's article sharing policy:

Mack KL, Nachman MW. 2017. Gene regulation and speciation. *Trends in Genet 33*: 68-80.
DOI: 10.1016/j.tig.2016.11.003

Abstract

Understanding the genetic architecture of speciation is a major goal in evolutionary biology. Hybrid dysfunction is thought to arise most commonly through negative interactions between alleles at two or more loci. Divergence between interacting regulatory elements that affect gene expression (i.e., regulatory divergence) may be a common route for these negative interactions to arise. Here we review how regulatory divergence between species can result in hybrid dysfunction, including recent theoretical support for this model. We then discuss the empirical evidence for regulatory divergence between species and evaluate evidence for mis-regulation as a source of hybrid dysfunction. Finally, we review unresolved questions in gene regulation as it pertains to speciation and point to areas that could benefit from future research.

1.1. A role for gene regulation in hybrid sterility and inviability

Understanding the genetic basis of speciation is a longstanding problem in evolutionary biology. The major model for the evolution of intrinsic post-zygotic isolation postulates that hybrid sterility or inviability arises from negative interactions between alleles at different loci when joined together in hybrids. The regulation of gene expression is inherently based on interactions between loci, raising the possibility that disruption of gene regulation in hybrids is a common mechanism for post-zygotic isolation. Although there is accumulating evidence that changes in gene regulation play a prominent role in adaptation (e.g., Chan *et al.* 2010; Jones *et al.* 2012), the role of regulatory evolution in speciation has received less attention. Here, we evaluate the role of regulatory evolution in speciation, and we suggest, both from recent theoretical and empirical studies, that changes in gene regulation play a major role in intrinsic post-zygotic isolation. While our focus is on post-zygotic isolation, regulatory divergence may also play an important role in establishing other reproductive barriers as a by-product of adaptive divergence (i.e., ecological speciation).

1.2. Conceptual framework

Single-locus models of hybrid dysfunction all suffer from the problem that mutations that lower the fitness of heterozygotes (and thus cause reproductive isolation) are unlikely to become established in a new population (e.g., Lande 1979; Hedrick 1981; Walsh 1982). This problem was recognized by Bateson (Bateson 1909), Dobzhansky (Dobzhansky 1937), and Muller (Muller 1940; Muller 1942), who suggested instead that hybrid dysfunction could arise from negative interactions between alleles at two or more loci. In the Bateson-Dobzhansky-Muller (BDM) model, alleles that are adaptive or neutral in their own genetic background are incompatible with alleles at one or more loci on the alternative genetic background (Figure 1). Thus, diverging lineages can accumulate substitutions without any loss of fitness. There is now strong empirical support for this model of intrinsic post-zygotic isolation (Coyne and Orr 2004).

Gene regulation is the process by which cells control the specific amount of gene product (i.e., RNA or protein) produced. Gene regulation is a complex process involving the interaction of DNA sequences, RNA molecules, and proteins, as well as epigenetic modifications. As the interaction of regulatory elements is required for organismal function, interacting regulatory elements are assumed to be co-adapted (e.g., Dover and Flavel 1984). When co-adapted interactions between regulatory elements are disrupted, downstream targets of these elements may be mis-regulated. While disrupted interactions between any of pair of regulatory elements or sequences could result in hybrid incompatibilities, the process of transcription initiation has received the most attention. While we focus mainly on transcriptional control, divergence between regulatory elements affecting other levels of gene regulation (e.g., translation) may also play a role in speciation.

Transcription is regulated by the interaction of *cis*- regulatory elements and *trans*- acting factors. *Cis*-regulatory elements are stretches of non-coding DNA (i.e., promoters, enhancers) that act as binding sites for *trans*- acting factors to regulate mRNA abundance. In the simplest case, the *trans*- acting factors are transcription factor proteins, though other proteins have also been known to act in *trans* to regulate gene expression (Yvert *et al.* 2003). Mutations in *cis*- regulatory regions or in transcription factors can affect mRNA abundance. Transcription factors frequently interact with multiple downstream target sequences and thus may be pleiotropic. In contrast, a single gene may have multiple *cis*- regulatory regions that regulate it in a tissue- and context- specific manner. As a consequence, changes in *cis*- regulatory regions are thought to be less pleiotropic than changes to the transcription factors they bind. The modularity of *cis*- regulatory regions has given rise to the idea that changes to these regions may play a large role in phenotypic evolution, an idea that is now well supported by empirical research (Wray 2007; Wittkopp and Kalay 2012). However, while transcription factors are assumed to evolve more slowly than *cis*- regulatory regions, they can evolve quickly compared to other gene classes (Castillo-Davis *et al.* 2012). Changes to transcription factor proteins have also been implicated in the evolution of novel phenotypes (e.g., Lynch *et al.* 2008).

Despite the role of transcriptional variation in phenotypic evolution, mRNA levels are often constrained on long time scales (Bedford and Hartl 2009). Genome-wide comparisons of mRNA levels between species show widespread reductions in

divergence compared to neutral expectations (Rifkin *et al.* 2008; Lemos *et al.* 2005; Gilad 2006), suggesting that changes in transcript levels are frequently deleterious. Despite the existing constraint on transcript levels, gene regulatory networks themselves are not necessarily well-conserved between species (True and Haag 2001). Interestingly, data on mRNA abundance from yeast, worms, and flies suggest that expression evolution best fits a House-of-Cards model of stabilizing selection (Hodgins-Davis *et al.* 2015) in which mutations generally have large effects that exceed the standing genetic variation (Kingman 1978; Turelli 1984). As a consequence, mutations that affect mRNA abundance can bring down the evolutionary "house of cards" and cause a cascade of changes between co-evolved *cis* and *trans* factors within a gene regulatory network.

Given these theoretical and empirical considerations, the epistatic interactions that underlie gene regulatory networks may lead to dysfunction in hybrids. In the simplest case, regulatory incompatibilities may arise either as a result of (1) the independent divergence of interacting elements between lineages (Figure 2a) or (2) lineage specific co-evolution between elements (Figure 2b). In the first model, populations respond differently to drift or parallel or opposing directional selection. One population fixes a *cis*- regulatory change, the other fixes a *trans* change. In the second model, a *cis* change that affects expression is compensated for by changes to an interacting *trans*- acting factor, or vise versa. In either model, negative interactions between divergent regulatory elements in hybrids may result in the mis-regulation of downstream targets. More complicated models are possible, including *cis* and *trans* changes in both lineages or interactions between more than two loci.

Recent simulations and mathematical models indicate that these kinds of regulatory incompatibilities can evolve quickly if selection is acting (Johnson and Porter 2000; Johnson and Porter 2007; Palmer and Feldman 2009; Tulchinsky *et al.* 2014; Khatri and Goldstein 2015). In particular, regulatory incompatibilities will evolve most quickly as a byproduct of adaptation when *cis* and *trans* regulatory elements diverge under positive selection (Johnson and Porter 2000; Tulchinsky *et al.* 2014). Incompatibilities will evolve more slowly under a model of stabilizing selection, where compensatory changes follow genetic drift (Tulchinsky *et al.* 2014). As transcription factors often regulate the expression of many genes, opposing selective pressures may constrain functional divergence and slow the evolution of regulatory incompatibilities. However, it was recently shown that it is possible for substantial hybrid mis-regulation to arise even when transcription factors are under moderate pleiotropic constraint (Tulchinsky *et al.* 2014b).

1.3. Regulatory divergence between species is widespread

Recent genomic surveys have found abundant evidence for transcriptional regulatory divergence between species. Divergence in putative *cis*- regulatory regions can be inferred through comparisons of transcription factor binding sites between species. While the loss and gain of transcription factor binding sites has generally been rapid over evolutionary time (Villar *et al.* 2014), examination of individual *cis*- regulatory elements has demonstrated that regulatory function can be maintained despite significant sequence divergence (Ludwig *et al.* 2000; Fisher *et*

al. 2006; Hare *et al.* 2008). This observation may be explained by the fixation of functionally compensatory mutations.

Regulatory divergence affecting the expression of individual genes can also be inferred through interspecific crosses. In F1 hybrids, differences in transcript abundance between two alleles indicates that differences between the parents at this locus are due to changes in *cis*, since the two alleles in the F1 are in a common *trans*-acting environment (Cowles *et al.* 2002) (Figure 3a). In contrast, if the two alleles in the F1 show the same level of transcript abundance, this indicates that differences between the parents are due to changes in *trans* (Wittkopp *et al.* 2004) (Figure 3b), although interpretation can be complicated by dominance in regulatory pathways (Porter *et al.* 2016). This approach has now been used to study genome-wide regulatory divergence between species of mice, birds, flies, yeast and plants (e.g., Goncalves *et al.* 2012; Davidson JH and Balakrishnan 2016; McManus *et al.* 2010; Tirosh *et al.* 2009, Shi 2012). Interspecific divergence in *cis* and *trans* is common, with *cis* regulatory variants generally contributing more to divergence between species than variation within species (Tirosh *et al.* 2009; Emerson *et al.* 2010; Coolon *et al.* 2014). However, a significant proportion of regulatory divergence can be attributed to a combination of *cis* and *trans*- acting variants.

When *cis* and *trans* changes are found together, interactions between them can increase or decrease gene expression divergence between species. When *cis* and *trans* variants act in opposition, their effects may buffer one another in a compensatory fashion. Consistent with stabilizing selection, such *cis*-*trans* compensation appears to play a prominent role in regulatory evolution (Goncalves *et al.* 2012; Tirosh *et al.* 2009; Shi *et al.* 2012; Takahasi *et al.* 2011; Mack *et al.* 2016).

The proportion of genes with *cis*-*trans* divergence has also been shown to accumulate with phylogenetic distance. Transgenic assays called “enhancer swaps”, where orthologous regulatory regions are tested in the same *trans* acting-environment, have found that lineage-specific *cis*-*trans* evolution is more common in comparisons between distant than closely related taxa (Gordon and Ruvinsky 2012). Similarly, pairwise comparisons between species of *Drosophila* found that while the number of genes with *cis*- regulatory divergence increased linearly with divergence time, the number of genes with total expression divergence does not (Coolon *et al.* 2014). This suggests that *cis* changes are often compensated for by changes in *trans* variants, or by other *trans* regulatory feedback mechanisms (Denby *et al.* 2012; Bader *et al.* 2015; Fear *et al.* 2016).

A few clear cases of such *cis*-*trans* compensatory evolution have now been reported (Kuo *et al.* 2010; Barrière *et al.* 2012). In the nematodes *Caenorhabditis elegans* and *C. briggsa*e, the expression of the gene *unc-47* is conserved between species even as its regulation has changed. Reciprocal swaps of *C. briggsae* and *C. elegans* regulatory elements identified lineage-specific changes consistent with compensatory *cis*-*tran*s evolution. Regions in the *C. briggsae unc-47* promoter have co-evolved with lineage-specific changes in the *C. briggsae trans*-regulatory environment. Compensatory modifications in regulatory elements associated with *unc-47* represent an example of how gene expression can be maintained despite underlying regulatory divergence (Barrière *et al.* 2012).

1.4. Mis-regulation as a mechanism for hybrid dysfunction

Mis-regulation of genes in hybrids can lead to misexpression, defined as gene expression that falls outside of the range of the parental species. Novel interactions between divergent *cis*- and *trans*- variants are one way misexpression can arise in hybrids. Consistent with this prediction, a number of studies have associated misexpression with *cis-trans* compensatory evolution (McManus *et al.* 2010; Tirosh *et al.* 2009; Mack *et al.* 2016; Landry *et al.* 2005; Schaefke *et al.* 2013, but see also Coolon *et al.* 2014; Bell *et al.* 2013). Misexpression is commonly seen in sterile interspecific hybrids (Michalak and Noor 2003; Ranz *et al.* 2004; Haerty and Singh 2006; Moehring *et al.* 2007; Malone *et al.* 2007; Good *et al.* 2010) and has been shown to accumulate with phylogenetic distance in *Drosophila* (Coolon *et al.* 2014).

In some interspecific hybrids, abnormal expression is disproportionately observed in male-biased genes (Michalak and Noor 2003; Ranz 2004) and genes involved in spermatogenesis (Good *et al.* 2010; Sundararajan and Civetta 2011), suggesting that regulatory divergence might underlie some cases of hybrid male sterility. Comparisons between sterile and fertile hybrids of *Drosophila* species (Gomes and Civetta 2015) and house mouse subspecies (Mack *et al.* 2016; Good *et al.* 2010) have found that a greater number of genes are misexpressed in sterile hybrids than in fertile hybrids. Moreover, in house mice, some expression quantitative trait loci (QTL) co-localize with sterility QTL in hybrids, suggesting a causal role for regulatory changes in hybrid male sterility (Turner *et al.* 2014). Also in mice, misexpression in sterile hybrids is associated with compensatory *cis-trans* changes, consistent with a model where disrupted interactions between these types of loci contribute to hybrid sterility (Mack *et al.* 2016).

The X chromosome often plays a central role in post-zygotic isolation (Coyne and Orr 2004; Coyne and Orr 1989). If regulatory divergence underlies hybrid dysfunction, evolutionarily diverged regulation of sex-linked genes may be expected (Johnson and Lachance 2012). Several recent studies have found that expression diverges faster for some genes on the X (in XY taxa) and Z (in ZW taxa) chromosomes than on the autosomes between species (Brawand *et al.* 2011; Llopart 2012; Meisel *et al.* 2012; Dean 2015; Kayserili *et al.* 2012; Coolon *et al.* 2015). Faster divergence of sex-linked gene expression is especially strong for genes with sex-biased effects (male-biased effects in XY taxa and female-biased effects in ZW taxa) (Llopart 2012; Meisel *et al.* 2012; Dean *et al.* 2015; Oka and Shiroishi 2014). However, comparisons of expression patterns in whole tissues may obscure differences in individual cell types. For example, it was recently shown that expression evolution for X-linked genes depends on the developmental stage of spermatogenesis, with genes that are expressed late in spermatogenesis showing slower divergence on the X (Larson *et al.* 2016). Disproportionate misexpression of X-linked genes has also been reported for sterile hybrids (Good *et al.* 2010; Turner *et al.* 2014; Oka and Shiroishi 2014; Bhattacharyya *et al.* 2013).

There are several caveats to bear in mind when considering whether misexpression is causing hybrid sterilty or inviability. First, the widespread misexpression seen in many interspecific crosses can be the result of one or a few upstream changes that cause a cascading effect on genes downstream in a regulatory network (Ortíz-Barrientos *et al.* 2007). This has been seen in hybrids

between *Saccharomyces cerevisiae* and *S. paradoxu*s, where misexpression is primarily due to a shift in the timing of meiosis (Lenz *et al.* 2014). Second, while misexpression in interspecific hybrids has been the subject of intense scrutiny, misexpression has also been observed in intraspecific hybrids where dysfunction is absent (Coolon *et al.* 2014; Gibson *et al.* 2004). Third, changes in cellular composition can also conflate associations between hybrid dysfunction and misexpression. Sterile and inviable animals often have gonads of differing cellular composition or suffer from atrophied tissue relative to their fertile counterparts. As many studies isolate mRNA from whole animals or whole tissues, differences in tissue or cellular composition between sterile or inviabile hybrids and parental species can produce misexpression. As a result, hybrid misexpression that is a direct result of regulatory divergence is likely to be overestimated (Wei *et al.* 2014). In the future, studies that make use of sorted cell populations may mitigate this problem somewhat by comparing gene expression only in equivalent cell types (Larson *et al.* 2016; Campbell *et al.* 2016; Bhattacharyya *et al.* 2014).

1.5. Evidence from speciation genes

Misexpression identified in sterile hybrids provides only indirect evidence of the role of mis-regulation in hybrid dysfunction. "Speciation genes" – defined here as genes that contribute to reproductive isolation – provide the best direct evidence for the role of regulatory divergence in reproductive isolation. Unfortunately, relatively few speciation genes have been identified and molecularly characterized (Presgraves 2010; Maheshwari and Barbash 2011). Despite this limitation, some broad scale patterns have started to emerge. Of the speciation genes identified so far, many have either a putative role in transcriptional or translational regulation, or are themselves misexpressed in hybrids (Table 1). While this pattern is intriguing, it is necessary to characterize the molecular and physiological basis of hybrid dysfunction in each case to determine whether regulatory divergence is causal. Below we discuss a few speciation genes that have been particularly well characterized in *Drosophila* and house mice, highlighting some of the challenges in linking specific mutations to mis-regulation.

1.5.1. *Hmr* and *Lhr.* Hybrid male lethality in crosses between *D. melanogaster* and *D. simulans* can be explained in part by the genes *Hybrid male rescue (Hmr)* and *Lethal hybrid rescue* (*Lhr*). The protein products of *Hmr* and *Lhr* form a complex that localizes to heterochromatic regions of the genome (Brideau *et al.* 2006; Thomae *et al.* 2013) where they transcriptionally repress transposable elements and repetitive sequences (Thomae *et al.* 2013; Satyaki *et al.* 2014) and play a critical role in mitotic chromosome segregation (Thomae *et al.* 2013).

Loss-of-function mutations at *Lhr* in *D. simulans* or at *Hmr* in D*. melanogaster* restore hybrid male viability (Brideau *et al.* 2006, Watanabe 1979; Hutter and Ashburner 1987). The *D. simulans* and *D. melanogaster* orthologs of both genes have diverged extensively under positive selection (Brideau *et al.* 2006). These observations led to the prediction that adaptive functional divergence between *Hmr* and *Lhr* and species-specific heterochromatin sequences causes hybrid dysfunction. However, orthologs of *Lhr* appear to be functionally equivalent: sequence

divergence between *Lhr* orthologs does not affect the localization of the *Lhr* protein, and overexpression of either the *D. simulans* or *D. melanogaster* ortholog has hybrid lethal effects (Brideau and Barbash 2011).

Hybrid lethality is instead a consequence of species-specific changes in the abundance of *Hmr* and *Lhr* protein product. HMR expression is higher in *D. melanogaster*, and LHR expression is higher in *D. simulans.* Increased expression of HMR in *D. melanogaster* and LHR in *D. simulans* results in an elevated amount of the HMR-LHR complex in hybrids. The activity of the HMR-LHR complex is dosage dependent, and overexpression leads to mislocation of the complex (Thomae *et al.* 2013).

As hybrid lethality is a consequence of HMR-LHR overexpression, the observed asymmetrical lethal effects of *D. melanogaster-Hmr* and *D. simulans-Lhr* are likely the result of divergence in regulatory pathways between *D. melanogaster* and *D. simulans* rather than functional divergence between orthologs (Thomae *et al.* 2013; Maheshwari and Barbash 2012). Supporting this hypothesis, transcriptional differences between *Lhr* orthologs in hybrids has been linked to compensatory *cis*-by-*trans* divergence between species in allele-specific expression (Maheshwari and Barbash 2012; Shirata *et al.* 2014).

1.5.2. *Prdm9.* Crosses between *Mus musculus domesticus* and *M. m. musculus* produce sterile hybrid males (Forejt and Iványi 1974). A series of laboratory mapping experiments by Forejt and colleagues (Gregorova *et al.* 1996; Trachtulec *et al.* 1997; Trachtulec *et al.* 2005; Trachtulec *et al.* 2008) led to the positional cloning and identification of *Prdm9* (Mihola *et al.* 2009), the only known hybrid sterility gene in vertebrates. *Prdm9* is believed to interact with yet uncharacterized loci on the X chromosome and autosomes to cause spermatogenic failure in hybrids (Storchová *et al.* 2004, Dzur-Gejdosova *et al.* 2012). Sterile hybrid males show sex-specific failure to pair chromosomes during meiosis as well as misexpression of genes on the X and Y chromosomes (Bhattacharyya *et al.* 2013). While *Prdm9* contains conserved domains associated with transcriptional regulation (Lim 1998; Margolin *et al.* 1994), the effect of *Prdm9* on misexpression may be a secondary consequence of the role of *Prdm9* in meiotic recombination.

Prdm9 has been implicated in recombination rate variation in both humans and mice (Baudat *et al.* 2010; Myers *et al.* 2010; Parvanov *et al.* 2010). During meiosis in mammals, double-stranded breaks are created throughout the genome and then repaired, leading to homologous recombination. These breaks are concentrated in regions called recombination hotspots. In mice, PRDM9 appears to mediate the process of recombination at hotspots by binding to DNA-sequences (Baudat *et al.* 2010). Intriguingly, another QTL implicated in recombination rate variation was recently found to overlap with a hybrid male sterility QTL on the X chromosome (Balcova *et al.* 2016). Altogether, these results suggest a genetic connection between recombination and hybrid sterility (Payseur 2016).

Variation in the number of PRDM9 zinc-finger tandem repeats has been implicated in house mouse sterility (Mihola *et al.* 2009). The PRDM9 zinc-finger array co-evolves with species-specific binding sites. Meiotic drive against recombination hotspots is thought to result in the rapid turnover of these binding

sites. Species-specific erosion of PRDM9 binding sites may explain asymmetric binding of PRDM9 in F1 hybrids that is associated with hybrid sterility. Supporting this prediction, hybrid fertility can be rescued by replacing the sterility associated zinc-finger array with an orthologous region from humans (Davies *et al.* 2016). While it is clear that sterile hybrid males show misexpression of genes on the X and Y chromosomes, the direct role, if any, of *Prdm9* in this misexpression remains unclear.

1.6. Open questions and future directions

While the evidence so far suggests that changes in gene regulation may contribute to the origin of new species, there are also cases where hybrid incompatibility appears to be independent of regulatory changes. For example, the speciation genes *Nup160* and *Nup96* cause hybrid inviability in crosses between *Drosophila simulans* and *D. melanogaster*. The protein products of both genes form architectural components of the nuclear pore complex and show evidence of adaptive protein evolution (Presgraves 2003; Tang and Presgraves 2009). We do not wish to provoke a debate on the relative importance of coding versus regulatory mutations to speciation; both surely occur and both are likely to be important in some instances. Instead, we offer several research directions that are likely to be particularly useful in understanding the connection between regulatory divergence and speciation.

First, the study of speciation has benefited from studies of natural populations and from studies that utilize laboratory crosses. However, most of what is known about the role of regulatory divergence in speciation comes from laboratory studies. These studies represent a small sliver of phylogenetic diversity and they rely mainly on model systems (Table 1). If we are interested in understanding generalities of the speciation process, greater taxonomic sampling is necessary. It would also be useful to compare patterns of gene expression in naturally occurring hybrid individuals that contain mixed genetic backgrounds to those seen in laboratory crosses.

Second, there are two aspects of many natural populations that merit further study: the presence of later generation hybrids and the fact that alleles contributing to reproductive isolation may be polymorphic rather than fixed (Cutter 2012). Studying both of these issues in the context of the role of regulatory divergence and reproductive isolation is important. For example, while great progress has been made studying F1 hybrids, using F2 or later generation hybrids makes it possible to identify disrupted gene expression caused by recessive alleles (Turner *et al.* 2014).

Third, most of the focus has been on the role of regulatory divergence in intrinsic post-zygotic isolation. The role of regulatory divergence in other forms of reproductive isolation (i.e., ecological, mating, and gametic) is still largely unexplored. Regulatory divergence may commonly lead to phenotypic differences between populations that result in different kinds of reproductive barriers. In particular, to the extent that changes in gene regulation underlie adaptive evolution, such changes may be quite common in ecological speciation, but this remains to be shown.

Fourth, there is a need to better integrate speciation theory with empirical evidence from gene expression studies. For example, the exposure of recessive mutations on the X (or Z) chromosome in heterogametic hybrids (i.e. XY males or ZW females) has been invoked to explain observations such as Haldane's Rule and the large X effect (Coyne and Orr 2004; Haldane 1922; Coyne 1992). According to this hypothesis, many of the alleles that decrease hybrid fitness are at least partially recessive. It is possible to test the dominance of expression inheritance using crosses or chromosome substitution lines (Gibson *et al.* 2004; Bhattacharyya *et al.* 2014; Lemos *et al.* 2008), and this would help link theoretical predictions with empirical observations of gene expression. Similarly, BDM incompatibilities are predicted to accumulate at a non-linear rate over evolutionary time resulting in a "snowball" effect (Orr 1995). Controlled gene expression studies may be able to determine whether regulatory incompatibilities conform to this prediction and increase nonlinearly with phylogenetic distance.

Fifth, the evolutionary forces that drive regulatory divergence and contribute to hybrid incompatibilities remain largely unknown. Many of the known speciation genes show a signature of positive selection (Presgraves 2010). While this observation is consistent with a model of adaptive divergence driving the evolution of hybrid incompatibilities, a model of compensatory evolution is equally possible. Compensatory evolution requires positive selection to fix compensatory changes to mask the deleterious effects of an earlier mutation.

Finally, while there is significant interest in the role of regulatory divergence in speciation, transcriptional control has received nearly all the attention. The regulation of gene expression is a complex process that may be modulated at many stages, including transcription, translation, and post-translation (Battle *et al.* 2015). The yeast speciation genes *AEP2* and *OLI1* provide one example of how translational mis-regulation can result in hybrid sterility. *AEP2* encodes a mitochondrial protein that translationally regulates *OLI1*. In interspecific hybrids of *S. cerevisiae* and *S. bayanus*, the Aep2 protein is unable to bind to *OLI1* transcripts. The inability of Aep2 to mediate the translation of *OLI1* is thought to result in hybrid sterility (Lee *et al.* 2008). Methodological advances have made the study of post-transcriptional regulation more feasible (Ingolia *et al.* 2009). Allele-specific analyses of translational efficiency can now be used to infer *cis* and *trans* regulatory divergence acting on translation rate (Artieri and Fraser 2014; McManus *et al.* 2014; Hou *et al.* 2015). QTL mapping techniques have been employed to study intraspecific variation in translation and protein abundance (Battle *et al.* 2015, Ghazalpour *et al.* 2011; Skelly *et al.* 124; Wu *et al.* 2013). Studies that combine each of these levels will provide a more complete picture of the role of regulatory divergence in speciation.

1.7. Chapter 1 Table

Table 1. Hybrid incompatibility genes

Locus	Gene name	Species	Phenotype	Molecular function	Evidence of gene regulation[1]	Citation
AEP2	ATPase expression 2	*Saccharomyces bayanus* x *S. cerevisiae*	Sterility	Mitochondrial protein	Regulates translation of *OLI1* transcripts	[118]
OLI1	Oligomycin resistance 1	*S. bayanus* x *S. cerevisiae*	Sterility	F0-ATP synthase subunit	Impaired translation in hybrids	[118]
Ods	Odysseus	*Drosophila mauritiana* x *D. simulans*	Sterility	Regulation of heterochromatic sequences	Encodes a DNA-binding protein that localizes to heterochromatin and regulates their decondensation	[126,127]
agt	O-6-alkylguanine-DNA alkyltransferase	*D. mauritiana* x *D. simulans*	Sterility	DNA binding protein	Encodes a DNA-binding protein for an alkyl-cysteine-S-alkyltransferase	[128]
Taf1	TBP-associated factor 1	*D. mauritiana* x *D. simulans*	Sterility	Transcription factor component	Encodes a DNA-binding protein for a subunit of transcription factor TFIID	[128]
Hmr	Hybrid male rescue	*D. melanogaster* x *D. simulans*	Inviability	Regulation of heterochromatic sequences	Overexpression of HMR/LHR complex in hybrids	[129]
Lhr	Lethal hybrid rescue	*D. melanogaster* x *D. simulans*	Inviability	Regulation of heterochromatic sequences	Overexpression of HMR/LHR complex in hybrids	[85]
gfzf	Suppressor Of Killer-of-prune [Su(Kpn)]	*D. melanogaster* x *D. simulans*	Lethality	Cell cycle regulation	Transcriptional regulator of the RAS/MAPK pathway	[130]
Nup160	Nucleoporin 160	*D. simulans* x *D. melanogaster*	Inviability	Nuclear pore protein	None	[111]
Nup96	Nucleoporin 96	*D. simulans* x *D. melanogaster*	Inviability	Nuclear pore protein	None	[110]
Ovd	Overdrive	*D. pseudoobscura bogatana* x *D. p. pseudoobscura*	Sterility	DNA binding	Encodes a MADF DNA-binding domain	[131]
Hhl	heterochromatin hybrid lethal	*D. melanogaster* x *D. simulans, D. mauritiana, D. sechellia*	Lethality	Unknown	Unclear	[132]
Zhr	Zygotic hybrid rescue	*D. melanogaster* x *D. simulans*	Inviability	Unknown, repetitive DNA	Unclear	[133]
Prdm9	PR domain-containing 9	*Mus musculus musculus* x *M. m. domesticus*	Sterility	Mediates meiotic homologous recombination	Encodes DNA-binding domains associated with transcriptional regulation	[99]
DM1/ DM2	DANGEROUS MIX 1/ 2	*Arabidopsis thaliana*	Lethality	Disease resistance	None	[134]

[1] Putative evidence for regulatory function or of misexpression in hybrids for hybrid incompatibility genes

1.8. Chapter 1 Figures

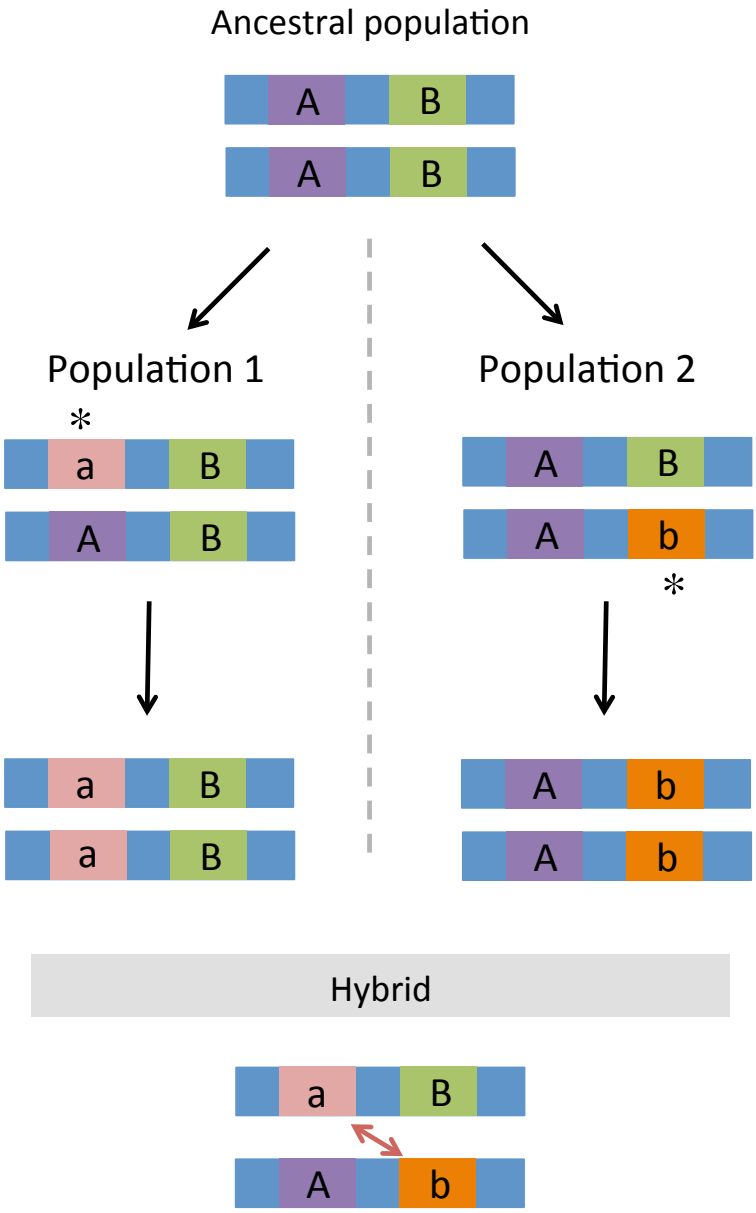


Figure 1. The Bateson-Dobzhansky-Muller model of hybrid incompatibility. In the ancestral population, the genotype is AABB. After the two populations are isolated, new mutations arise independently on each lineage as indicated by the asterisks. In one population, A evolves into a, in the other population B evolves into b. In hybrids, negative interactions between the a and b allele can result in sterility or inviability. The a and b alleles are found together for the first time in hybrids, explaining how this incompatibility could evolve without either lineage experiencing an intermediate state of reduced fitness.

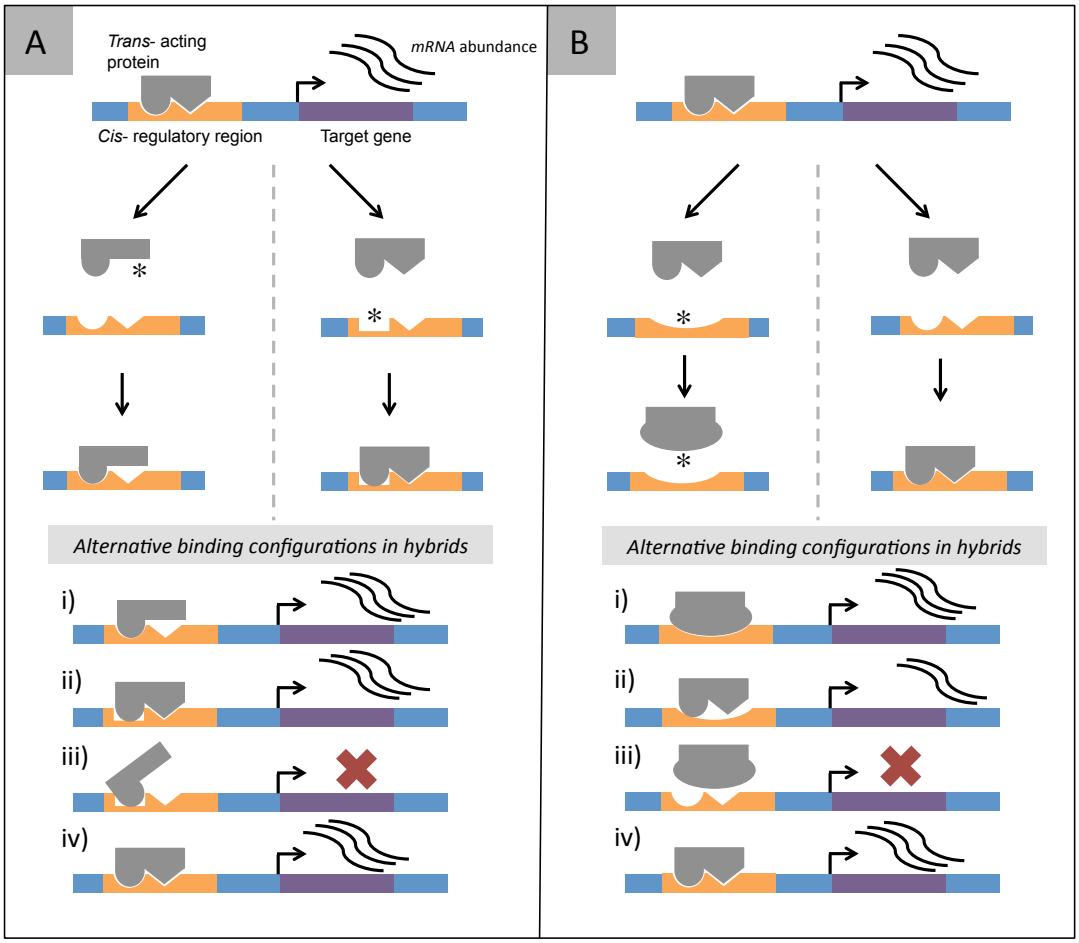


Figure 2. Regulatory divergence as a source of hybrid incompatibilities. Figures A and B are schematics of a 2-locus model for hybrid incompatibilities. Each hybrid incompatibility arises as a consequence of the molecular interactions between a *cis*-regulatory region and a *trans*- acting factor. Changes in binding between interacting regulatory elements affect the expression of a downstream gene. Asterisks represent mutations that become fixed along a lineage. A) A change to a *cis*-regulatory region in one species and the interacting *trans*- acting factor in the other result in hybrid dysfunction. Divergence in this example may be the result of drift or selection. In hybrids, the binding configuration represented by (*iii*) results in mis-regulation, while (*i*), (*ii*), and (*iv*) produce normal transcriptional output. B) Lineage specific co-evolution between *cis* and *trans* regulatory elements result in a hybrid dysfunction. In this example, a change in *cis* is followed by a compensatory change in *trans* to mask the deleterious effect of the first mutation. In hybrids, the binding configuration represented by (*iii*) results in mis-regulation. The binding configuration represented by (*ii*) results in reduced expression compared to the parents, while the binding configurations represented by (*i*) and (*iv*) result in the same expression as in the parents.

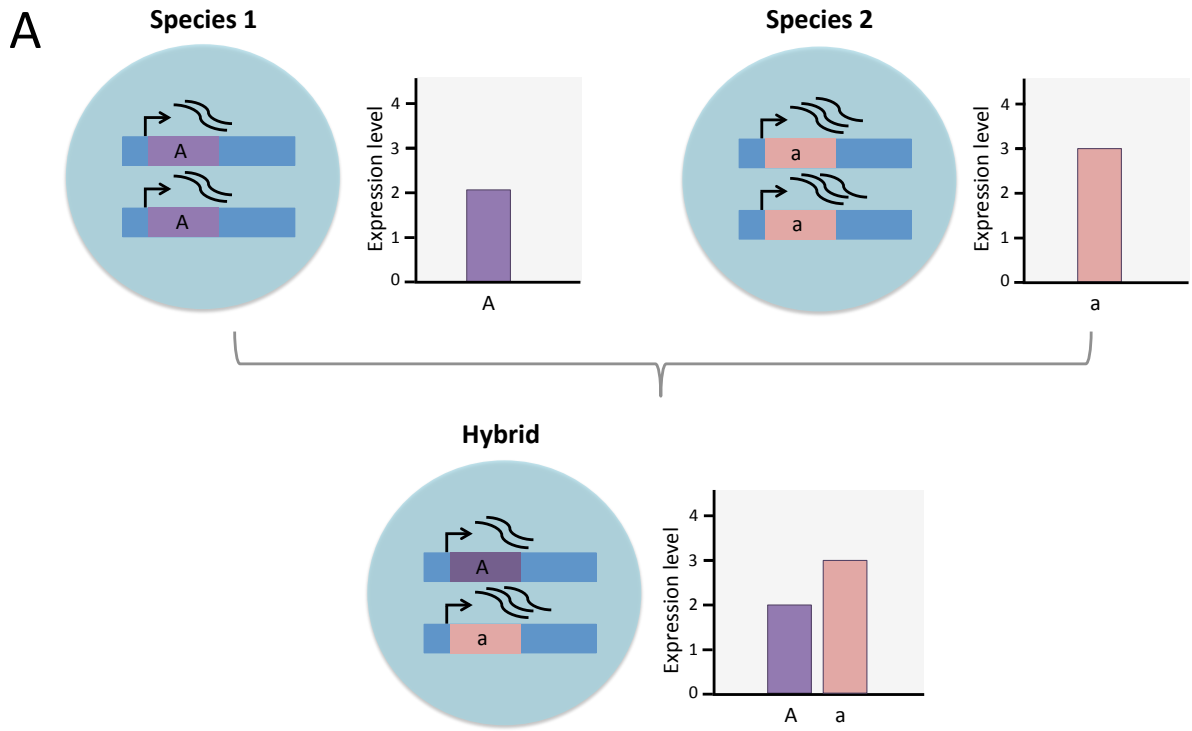


Figure 3. Using allele-specific expression to infer regulatory divergence between species. Differences in the expression of alleles in an F1 can be used to determine whether expression divergence between the parents is due to changes in *cis* or to changes in *trans*. A) Species 1 carries the *A* allele while Species 2 carries the *a* allele. In the parental species, the transcript abundance of *A* is 2 and the transcript abundance of *a* is 3. Differences in the expression of the *A* and *a* alleles in the F1 hybrid suggests *cis*- regulatory divergence between Species 1 and 2, since these two alleles are in the same *trans*-acting environment in the F1. B) *A* and *a* have equal transcript abundances in the F1 hybrid despite the difference in expression seen between the parents. This suggests that differences between the parents are due to changes in *trans*.

Chapter 2

Gene regulation and speciation in house mice

This chapter has been previously published and is reproduced here in accordance with the journal's article sharing policy:

Mack KL, Campbell P, Nachman MW. 2016. Gene regulation and speciation in house mice. *Genome res* 26: 451-461.
DOI: 10.1101/gr.195743.115

Abstract

One approach to understanding the process of speciation is to characterize the genetic architecture of postzygotic isolation. While the majority of work in this area has focused on identifying incompatibilities between protein coding genes, negative epistatic interactions between divergent regulatory elements might also contribute to reproductive isolation. Here we take advantage of a cross between house mouse subspecies, where hybrid dysfunction is largely unidirectional, to test several key predictions about regulatory divergence and reproductive isolation. Regulatory divergence between *M. m. musculus* and *M. m. domesticus* was characterized by studying allele-specific expression in fertile hybrid males using mRNA-sequencing of whole testes. We found extensive regulatory divergence between *M. m. musculus* and *M. m. domesticus*, largely attributable to *cis*- regulatory changes. When both *cis*- and *trans*- changes occurred, they were observed in opposition much more often than expected under a neutral model, providing strong evidence of widespread compensatory evolution. We also found evidence for lineage-specific positive selection on a subset of genes related to transcriptional regulation. Comparisons of fertile and sterile hybrid males identified a set of genes that were uniquely mis-expressed in sterile individuals. Lastly, we discovered a nonrandom association between these genes and genes showing evidence of compensatory evolution, consistent with the idea that regulatory interactions might contribute to Dobzhansky-Muller incompatibilities and be important in speciation.

2.1. Introduction

Forty years ago, King and Wilson argued that differences between chimpanzees and humans could not be explained by changes in protein sequences alone (King and Wilson 1975). Since then, there has been a lively debate about the relative importance of changes in gene regulation versus changes in gene structure in adaptive evolution (e.g. Hoekstra and Coyne 2007; Carroll 2008), and some recent studies have revealed a major role for regulatory changes in adaptation (e.g. Jones *et al.* 2012).

The role of gene regulation in speciation has received less attention. This is somewhat surprising since gene regulation requires interactions between loci, and disrupted interactions between loci in hybrids (Dobzhansky-Muller incompatibilities) are thought to underlie many examples of post-zygotic

reproductive isolation. At the transcriptional level, gene expression is a consequence of the interaction of *cis*- regulatory elements and *trans*- acting factors. *Cis*- regulatory regions are stretches of noncoding DNA that bind *trans* acting factors to regulate mRNA abundance. Thus, negative epistatic interactions between *cis*- and *trans*- regulatory elements in hybrids might be important in reproductive isolation.

One powerful way to identify *cis*- and *trans*- changes is to compare expression differences between species with expression differences between alleles in interspecific hybrids (Fig. 1) (Cowles *et al.* 2002; Wittkopp *et al.* 2004). This approach has now been used in a number of crosses in flies, yeast, mice**,** and plants (Table 1). These studies have led to an emerging understanding of regulatory divergence within and between species as well as some understanding of the causes of mis-expression in hybrids.

Lacking in these studies is a direct association with reproductive isolation through a hybrid sterility or inviability phenotype. House mice (*Mus musculus*) provide a good opportunity for making links between hybrid sterility phenotypes, mis-expression in hybrids, and regulatory divergence between lineages. House mice consist of three main subspecies that diverged recently and are isolated to varying degrees by hybrid male sterility. Over the past four decades, house mice have been developed as a model system for the study of mammalian hybrid sterility (e.g. Forejt and Iványi 1974; Forejt 1985,1996; Oka *et al.* 2004, 2007, 2010, 2014; Britton-Davidian *et al.* 2005; Good *et al.* 2008a, 2010; Mihola *et al.* 2009; Bhattacharyya *et al.* 2013, 2014). Sterility is highly polymorphic between different laboratory strains and in natural populations (Forejt and Iványi 1974; Good *et al.* 2008a; Vyskocilova *et al.* 2009; Bhattacharyya *et al.* 2014). Importantly, crosses between a wild-derived inbred line of *M. m. musculus* (PWK/PhJ) and a wild-derived inbred line of *M. m. domesticus* (LEWES/EiJ) result in infertile hybrid males in one direction and fertile hybrid males in the reciprocal direction. Infertile hybrid males in this cross have significantly reduced testis weight and sperm count compared to pure subspecies (Good *et al.* 2008a). For simplicity, hereafter we refer to these hybrid males with lowered fertility as "sterile" though sterility is not complete in all individuals. By comparing sterile and fertile hybrid males, it is possible to disentangle mis-expression that is associated with sterility from mis-expression that is simply a consequence of hybridization.

In a previous study using genome-wide microarray data, hybrid male sterility in this cross was associated with widespread over-expression of the *M. m. musculus* X Chromosome during spermatogenesis and mis-expression at a number of autosomal genes (Good *et al.* 2010). This work suggested that differences in gene regulation might be important in reproductive isolation. More recently, Turner *et al.* (2014) mapped sterility quantitative trait loci (QTL) and expression QTL (eQTL) in an F2 cross using different strains of *M. m. musculus* and *M. m. domesticus.* They identified a large role for *trans*-eQTL as well as a number of complex regulatory network interactions related to sterility (Turner *et al.* 2014). However, the mapping approach was not designed to identify allele-specific expression patterns in F1's and did not address the relative importance of *cis*- and *trans*- changes to regulatory divergence between these subspecies.

Here we compare expression differences between house mouse subspecies with expression patterns in sterile and fertile F1 hybrids. This allows us to address a number of related issues. First, we describe the proportion of changes between subspecies that are due to changes in *cis*, *trans*, or both. Second, when both kinds of changes occur, they may occur in the same direction or in the opposite direction. If gene expression is largely under stabilizing selection, as experimental work suggests (Denver *et al.* 2005; Lemos *et al.* 2005; Gilad *et al.* 2006), *cis*- and *trans*-variants that act in opposite directions may be more common than expected by chance. We test this prediction. Third, the identification of *cis*-eQTL allows us to ask whether differences in expression are driven by positive selection (Bullard *et al.* 2010; Fraser *et al.* 2010, 2011) and, if so, to identify classes of genes that are under selection. Fourth, we identify mis-expression (i.e., changes greater than 1.25-fold on a $\log_2$ scale between the hybrid and both parents) in sterile and fertile hybrids. Comparing sterile and fertile hybrids allows us to identify those genes that are mis-expressed only in sterile mice and thereby associate mis-expression with hybrid sterility. While this approach does not distinguish between the specific genes causing sterility from those that are mis-expressed as a downstream consequence of causative genes, it does identify a set of candidate genes for reproductive isolation and it makes specific testable predictions. In particular, we test the hypothesis that these candidate genes are disproportionately governed by compensatory evolution, as expected if regulatory interactions contribute to Dobzhansky-Muller incompatibilities.

2.2. Results

2.2.1. Extensive *cis* regulatory divergence between *M. m. musculus* and *M. m. domesticus*

To characterize the contribution of *cis*- and *trans*-acting variants to divergence between *M. m. musculus* and *M. m. domesticus*, we compared expression differences in whole testis between subspecies with allele-specific expression in their fertile hybrid using three replicates per genotype (Fig. 1a). Since hybrids inherit alleles from both parents that meet in the same *trans*- acting environment, differences in expression between parents that are also seen between alleles in hybrids can be inferred to be the result of one or more *cis*- regulatory variants (Cowles *et al.* 2004). Alternatively, when a gene is differentially expressed between subspecies but not between alleles in the hybrid, we can infer divergence in one or more *trans* variants (Wittkopp *et al.* 2004).

Only reads that could be assigned preferentially to either *M. m. musculus* or *M. m. domesticus* were retained for analysis (see Table S1 for read counts). This allowed us to measure allele-specific expression in hybrids by comparing the relative number of reads mapping to the genome of each subspecies. After excluding genes with low read counts from the analysis, 9,851 autosomal genes could be tested for regulatory divergence (see supplemental methods). Of genes that could be tested, approximately 24% (2,349 genes) showed evidence of divergence due to

one or more variant acting in *cis* alone, 9% (883 genes) showed evidence of divergence due to one more variants acting in *trans* alone, and 44% (4,349 genes) showed evidence of divergence in both *cis* and *trans* (Fig. 1b).

The median regulatory divergence between subspecies in *trans* alone (0.58 $\log_2$ fold change) was significantly lower than the median divergence in *cis* alone (0.65 $\log_2$ fold change) (Wilcoxon rank-sum test, p=0.00019). Genes with an upper-quartile $\log_2$ fold change between subspecies ($|\log_2$ fold change$|>0.96$) were also enriched for variants acting in *cis* alone relative to those in *trans* alone (40% *cis* alone, 9% *trans* alone; Fisher's exact test p=0.0003).

2.2.2.Widespread compensatory evolution

Genes with evidence of divergence in *cis* and *trans* can be further subdivided into categories based on their contribution to expression differences between subspecies and their direction of action. Genes with evidence of divergence in both *cis* and *trans* were divided into three subgroups (Landry *et al.* 2005; McManus *et al.* 2010)(see supplemental methods and Fig. 1): 1) *cis x trans*, where there was significant differential expression between subspecies, significant differential expression between alleles in the hybrid, and where the subspecies with higher expression contributed the lower expressed allele in the hybrid; 2) compensatory, where the subspecies did not show differences in expression, but alleles in hybrids were significantly different; and 3) *cis + trans*, where there was significant differential expression between subspecies, significant differential expression between alleles in the hybrid, and where the subspecies with the higher expression level contributed the higher expressed allele in the hybrid. We further subdivided genes in this last category, *cis + trans,* into cases where *cis* and *trans* variants act in the same direction and cases where these variants act in opposition (Fig. S1). Of genes with evidence of both *cis* and *trans* divergence, the majority were categorized as *cis + trans* (24%, or 2,392 genes); in the majority of these, *cis* and *trans* variants act in opposition (1,626 genes) rather than in the same direction (766 genes) (Fig 1B). Thirteen percent of genes were categorized as compensatory (1,309 genes). A minority of genes showed evidence of *cis* x *trans* divergence (7%, 648 genes) (Table S2).

Under a neutral model, we expect an equal number of genes to show divergence due to *cis* and *trans* variants acting in opposition and *cis* and *trans* variants acting in the same direction. An excess of *cis* and *trans* changes acting to reinforce one another would be consistent with directional selection to alter expression level. Alternatively, an excess of *cis* and *trans* variants acting in opposition would be evidence for compensatory evolution and widespread stabilizing selection to maintain expression level. Genes categorized as *cis* x *trans*, compensatory, and a subset of *cis + trans* (where variants act in opposition), show evidence of *cis* and *trans* changes acting in opposite directions (Fig. 1b). In contrast, a subset of genes categorized as *cis + trans* show evidence of *cis* and *trans* changes that are acting in the same direction. By deriving neutral expectations from the number of independent *cis* and *trans* changes acting in the same and opposite directions, we tested for bias in directionality (see methods). The proportion of *cis*- and *trans*- changes that act in opposition was extremely inflated compared to the

neutral expectation (Table 2, $p<0.0001$) providing evidence for widespread compensatory evolution.

2.2.3. Adaptive evolution of *cis*-regulatory elements

Changes in *cis*- variants are potentially targets for selection on gene expression level as *cis*-regulatory regions act as context-dependent regulators on which selection may act efficiently (reviewed by Wray 2007). To test for lineage-specific selection on genes with divergent *cis*-acting variants between the subspecies, a gene set approach was employed (Bullard *et al.* 2010; Fraser *et al.* 2010, 2011). Under a neutral model, an equal number of genes will be up- and down- regulated by *cis* variants. If a gene set associated with a biological function deviates from the null expectation by presenting a significant directional bias, we can infer lineage-specific selection. We tested this by grouping genes with only *cis*-acting variants by Gene Ontology (GO) terms (see supplemental methods). Three non-independent biological process GO terms were identified with significant enrichment for biased directionality: 1) Transcription, DNA-templated (GO:0006351, $p=0.0004$), 2) Positive regulation of transcription from RNA polymerase II promoter (GO:0045944, $p=0.02$), and 3) Regulation of transcription, DNA-templated (GO:0006355, $p=0.02$). These interrelated gene sets collectively include 410 genes with putative evidence of selection, and show biased directionality towards upregulation in *M. m. musculus* (or down-regulation in *M. m. domesticus*).

2.2.4. Mis-expression in hybrids

Crosses between *M. m. domesticus* (LEWES/EiJ) and *M. m. musculus* (PWK/PhJ) result in fertile hybrid males when the mother is *M. m. domesticus* and sterile hybrid males when the mother is *M. m. musculus.* To identify differences in expression between fertile and sterile hybrids and to identify mis-expression, we summed reads mapping to both the *M. m. domesticus* and *M. m. musculus* allele for each sample and then for each genotype (see supplemental material). Total read counts for fertile and sterile hybrids are strongly correlated with the read counts of both subspecies (Fig. S2).

First, we compared expression patterns on the X Chromosome between sterile and fertile mice. Previous work suggests a large role for the *M. m. musculus* X Chromosome in hybrid male sterility (Good *et al.* 2010; Storchová *et al.* 2004; Good *et al.* 2008a; Oka *et al.* 2004, 2014; Bhattacharyya *et al.* 2013, 2014). Genes remaining in the analysis after filtering for low read counts were distributed across the X Chromosome. In fertile hybrids, the number of genes expressed above and below the level seen in *M. m. domesticus* was nearly equal, while in sterile hybrids the majority of genes were expressed above the level seen in *M. m. musculus* (Fig. 2)(Fisher's exact test, $p<0.0001$; Table S3). We next compared fold changes on the X to fold changes on the autosomes. Fold changes were calculated between both subspecies and between the sterile and fertile hybrids for 10,264 genes. The ratio of genes over-expressed on the X versus the autosomes in the sterile hybrid was significant (Fisher's exact test, $p<0.0001$; Table S4), while there was no significant difference between these ratios in the fertile hybrid (Fisher's exact test, $p=1.0$; Table

S4). Together, these results suggest that the X Chromosome in the sterile hybrid is uniquely overexpressed compared to the fertile hybrid and to the autosomes. Over-expression of genes on the X Chromosome in sterile hybrids is consistent with previous work based on microarrays (Good *et al.* 2010). It is also consistent with expression studies of germ cells that were sorted by developmental stage (Campbell *et al.* 2013), indicating that overexpression of genes on the X is not an artifact of differences in the cellular composition of the testes of sterile and fertile mice (see Discussion).

Next, we focused on patterns of expression of autosomal genes. Comparing the number of reads mapping to a gene in the hybrid and in the pure subspecies allowed us to identify mis-expressed genes and to infer the mode of inheritance for expression for each gene (Fig S2). Genes that showed less than a 1.25-$\log_2$ fold change between the hybrid and both subspecies were considered "similar" regardless of significance (Gibson *et al.* 2004; McManus *et al.* 2010). Since this is a conservative cut-off, we found that most genes showed similar levels of expression in hybrids and in pure subspecies (86%, or 8,834 genes, and 90%, or 9,300 genes, of genes in the sterile and fertile hybrid, respectively; Table S5). While the number of genes categorized as similar in this analysis is higher than in previous studies, this is unsurprising given the short divergence time between *M. m. musculus* and *M. m. domesticus*. Genes that did not demonstrate conserved expression patterns were divided into *dominant*, *additive*, and *mis-expressed* (see supplemental methods and Table S5). Where 28 genes were mis-expressed in the fertile hybrid, 63 genes were mis-expressed in the sterile hybrid (Table S5). In the fertile hybrid an equal number of genes were mis-expressed above and below the level of both subspecies, while in the sterile hybrid significantly more genes were over-expressed (Fisher's exact test, p=0.0006; Table S6). Eleven mis-expressed genes were shared between the sterile and fertile hybrid, all of which were over-expressed.

Genes that are over- or under-expressed in the sterile hybrid to the exclusion of the fertile hybrid are of interest as potential candidates for hybrid incompatibilities. First, we identified genes for which the number of reads mapping to the fertile and sterile hybrid was significantly different. Then we eliminated genes with less than a 1-$\log_2$ fold difference between the sterile hybrid and both subspecies. A 1-$\log_2$ fold change corresponds to an expression difference that is two-fold higher or lower, so differences between the sterile hybrid and each subspecies at this threshold may be biologically meaningful. We identified 202 genes at a 5% FDR with these criteria, hereafter referred to as genes with "aberrant expression" for simplicity. These 202 genes were enriched for 39 non-independent GO terms at a 5% false discovery rate, the most highly significant of which were: 1) positive regulation of gene expression (FDR q-value=0.0115), 2) positive regulation of RNA metabolic process (FDR q-value=0.0139), and, 3) regulation of cell migration (FDR q-value= 0.0236) (Eden *et al.* 2009). Of these aberrantly expressed genes, 17 were associated with only a *cis* regulatory change and thus could be included in the test for positive selection. Remarkably, 12 of these 17 genes were identified as targets of positive selection in the analysis above, representing a highly significant over-enrichment of positively selected genes among those associated with hybrid sterility (Fisher's Exact Test, p<0.0001; Table S7).

A subset of the genes that are aberrantly expressed uniquely in the sterile hybrid are associated with male reproductive phenotypes or cell cycle control in laboratory mice, or are highly expressed in the testis relative to other tissues, making them potential candidates for reproductive incompatibilities between the subspecies (Table 3)(phenotype and expression data collected from Eppig *et al.* 2015; Su *et al.* 2004; Wu *et al.* 2009). Notably, five genes (*Adgrg1*, *Itpka*, *Mtcl1*, *Myl10*, and *Micall2*) have been identified in regions of overlap between the results of a genome-wide differentiation study between the subspecies (Phifer-Rixey *et al.* 2014), a QTL mapping study on measures of hybrid male sterility (White *et al.* 2011), and in regions of low introgression across the *M. m. musculus* and *M. m. domesticus* hybrid zone (Janoušek *et al.* 2012).

2.2.5. Compensatory evolution is associated with mis-expression in sterile hybrids

If *cis*- and *trans*- changes interact epistatically to result in hybrid incompatibilities, we expect divergence between subspecies that involves both *cis* and *trans* changes to be associated with novel expression patterns in the sterile hybrid. Genes with both *cis* and *trans*- changes in opposing directions should be particularly enriched if the breakdown of co-adapted regulatory machinery contributes to mis-expression in sterile hybrids. To test this hypothesis, we examined the regulatory categories associated with genes that were mis-expressed in sterile hybrids (genes with a greater than 1.25-$\log_2$ fold change between the sterile hybrid and both subspecies) (Table S8). A number of the mis-expressed genes could not be analyzed for regulatory divergence due to low read counts. Of the genes that remained in the analysis, there was a non-random association between *cis* and *trans* variants acting in opposing directions and mis-expression in the sterile hybrid compared to genes where *cis* or *trans* variants acted alone or in the same direction (Fisher's exact test, $p<0.0001$; Table 4). Genes categorized as strictly compensatory, where there was no significant difference in expression between subspecies despite significant differences between alleles in the hybrid, were the most enriched in the mis-expressed gene set (Fisher's exact test, $p=0.0004$; Table S9). Far fewer mis-expressed genes were retained for analysis from the fertile hybrid (17 genes total). No regulatory category was enriched in the mis-expressed gene set of the fertile hybrid, although this may be due to lack of power given the low number of genes tested (Fisher's exact test, $p=1.0$; Table S10).

Next, we repeated this analysis using the previously described "aberrantly expressed" genes (Table S11) (i.e. a more relaxed cut-off in which expression was at least 1-$\log_2$ fold different between the sterile hybrid and both subspecies). As above, genes for which *cis*- and *trans*- variants acted in opposition were enriched compared to genes for which *cis*- and *trans*- variants acted independently or in the same direction (Fisher's exact test, $p<0.0001$; Table S12). Likewise, strictly compensatory changes again were especially enriched in this differentiated gene set (Fisher's exact test, $p<0.0001$; Table S13). Finally, to further investigate the relationship between compensatory evolution and mis-expression in the sterile hybrid, genes were binned based on $\log_2$ fold changes between the sterile hybrid and both subspecies.

As fold change increased, the proportion of genes where *cis* and *trans* variants act in opposition increased (Fig. 3).

2.2.6. Expression comparisons between multiple subspecies lines

The findings described above were based on a small number of wild-derived inbred lines. This limits the extent to which our conclusions speak to regulatory divergence between *M. m. musculus* and *M. m. domesticus* in general as opposed to regulatory divergence between these particular lines. To expand this analysis and look more generally at expression divergence between the subspecies, we took advantage of data from a recent study that analyzed the testis transcriptomes from 7 lines of *M. m. domesticus* and 8 lines of *M. m. musculus* (Phifer-Rixey *et al.* 2014). While Phifer-Rixey *et al.* (2014) included more lines, coverage per line was lower than in our analysis. Still, overlap between the two datasets is high: 77% of the genes in Phifer-Rixey *et al.* (2014) were represented in our data. We re-analyzed the data of Phifer-Rixey *et al.* (2014) for this subset of 9,779 genes that were shared between the two studies.

Importantly, genes that were differentially expressed in the data of Phifer-Rixey *et al.* (2014) overlap significantly with genes that have significant parental ratios in our analysis (hypergeometric test, p=1.749e-16). Genes categorized as *cis* and *cis + trans* where variants act in the same direction were particularly enriched in this overlap, making up 57% of the genes found to be differentially expressed between *M. m. musculus* and *M. m. domesticus* in both analyses (p<0.0001). Conversely, genes where *cis* and *trans* variants act in opposing directions (*cis x trans* and subset of *cis + trans* categories) showed the lowest proportion of overlap.

We also reanalyzed the data from Phifer-Rixey *et al.* (2014) to see if our conclusions about *cis*- changes subject to positive selection were general. Genes with significantly different expression between *M. m. musculus* and *M. m. domesticus* in Phifer-Rixey *et al.* (2014) that overlapped with genes identified in our analysis as divergent in *cis* alone were categorized based on directionally. Genes in the three sets we identified as targets of selection (biological process GO terms GO:006351, GO:0045944, and GO:0006355; see results above) were then subjected to a hypergeometric test as in the previous analysis. Despite the reduction in genes represented in each gene set, all three sets maintained biased directionality at a 10% false discovery rate in this new analysis based on a larger number of inbred lines.

The general concordance between these datasets suggests that many of the conclusions described above do not simply represent line effects but instead characterize regulatory divergence between these two subspecies more generally.

2.3. Discussion

We characterized regulatory divergence in testis between *Mus musculus domesticus* and *Mus musculus musculus* as well as aberrant expression associated with sterility in hybrids. We identified evidence of widespread compensatory evolution consistent with stabilizing selection as well as evidence for lineage-specific positive selection on a subset of genes related to transcriptional regulation.

Lastly, we identified genes with aberrant expression unique to sterile hybrids. These sterility-associated genes were non-randomly associated with *cis*- and *trans*-changes that act in opposition to one another, consistent with the idea that regulatory changes might underlie Dobzhansky-Muller incompatibilities and be important in speciation.

2.3.1. Regulatory divergence between *M. m. domesticus* and *M. m. musculus*

A large number of genes in this study showed evidence of gene expression divergence between *M. m. domesticus* and *M. m. musculus*. To mitigate the potential effects of inbreeding, we crossed two different inbred lines within each subspecies to create heterozygous individuals against which inter-subspecific hybrids could be compared. This approach, which is rarely used in studies of expression evolution, eliminates differences in gene expression that arise between subspecies as a result of differences in inbreeding depression and it eliminates expression differences between the subspecies and hybrids as a result of heterosis. We also compared our results to an independent expression study that included more inbred lines (Phifer-Rixey *et al.* 2014). Without population level sampling, it is impossible to distinguish between line-specific effects and subspecific differences. However, by characterizing the intersection between these two datasets, we identified patterns that are more likely to be representative of subspecific differences. The high correspondence between the two studies despite their differences in depth and breath suggests that we have captured a large proportion of subspecific divergence.

The majority of the regulatory divergence between *M. m. musculus* and *M. m. domesticus* was the consequence of *cis* variants, either alone or together with one or more *trans* variants. Conversely, regulatory divergence due to *trans* variants alone was relatively rare, accounting for only a small proportion of genes tested. Comparisons between the median expression differences associated with variants acting in *cis* or *trans* alone revealed that *cis* variants were of greater magnitude. Consistent with the results presented here, divergence in *cis* has been demonstrated to be more common than divergence in *trans* in insects and nematodes (Gordon and Ruvinsky 2012), and was previously shown to contribute to a larger proportion of differentially expressed genes in the liver between the house mouse subspecies *M. m. castaneus* (CAST/EiJ) and *M. m. domesticus* (C57BL/67) (Goncalves *et al.* 2012). Similarly, Crowley *et al.* (2015) found allelic imbalance consistent with *cis* regulatory effects in 85% of testable genes in comparisons between mouse subspecies. These results stand in contrast to those of McManus *et al.* (2010) and Coolon *et al.* (2014), both of whom found a large proportion of expression divergence to be the result of *trans*- differences in *Drosophila* crosses. Elevated *trans*- divergence in these two studies may be due to demographic or biological differences between species or to differences in the experimental methods (e.g., the use of whole files versus specialized tissue types, number of replicates, etc.).

Studies in yeast and flies suggest that *cis*-regulatory divergence typically contributes more to differences between species than to differences within species (Tirosh *et al.* 2009; Emerson *et al.* 2010) and increases consistently and proportionately with divergence time (Coolon *et al.* 2014). While *cis*-regulatory variation is substantial in natural populations (Osada *et al.* 2006; Genissel *et al.*

2007; Campbell *et al.* 2008; Gruber and Long 2009; Lemmon *et al.* 2014), *trans*-acting variation contributes more to polymorphic expression variation within species (Wittkopp *et al.* 2008; Lemos *et al.* 2008; Coolon *et al.* 2014). *M. m. domesticus* and *M. m. musculus* diverged roughly 350,000 years ago and still share some ancestral variation. Thus, some of the regulatory differences observed between inbred strains could still be polymorphic in one or both subspecies. Finally, overlap between our data and those of Phifer-Rixey *et al.* (2014) is greatest for genes associated with *cis*- changes and *cis* + *trans* changes (where variants act in the same direction), suggesting that these two regulatory categories may contribute disproportionately to regulatory divergence between subspecies compared to within-subspecies variation.

Stabilizing selection has been identified as a dominant force underlying gene expression evolution (Gilad *et al.* 2006). A widespread reduction in gene expression variation compared to neutral expectations based on intra- and interspecific comparisons (Lemos *et al.* 2005; Rifkin *et al.* 2003) and mutation accumulation lines (Denver *et al.* 2005) suggests that changes in expression are frequently deleterious. The apparent reduction in expression divergence in these studies compared to neutral expectations could be the outcome of two separate processes: the elimination of *cis*- and *trans*- acting variants through purifying selection or compensatory evolution between regulatory elements that conserves expression levels. Our results favor the latter explanation. We identified a significantly greater proportion of instances where *cis* and *trans* variants acted in opposition than expected under neutrality**,** consistent with widespread lineage-specific compensatory evolution.

What drives this compensatory evolution? One possibility is that selection initially favors a mutation acting in *trans*, perhaps because selection favors a change in expression of some downstream gene. If the initial *trans*- change is highly pleiotropic, it may alter the expression of other downstream genes in a suboptimal way. Selection would then favor the restoration of optimal expression levels at these genes through compensatory *cis*- changes (Goncalves *et al.* 2012; Coolon *et al.* 2014).

Against this background of widespread compensatory evolution involving changes in both *cis*- and *trans*-, we also found evidence for lineage-specific positive selection on a subset of *cis*- only changes. Selection is predicted to act efficiently on *cis*-regulatory variants (Wray 2007) and simulations suggest that natural selection is more likely to drive *cis*-regulatory divergence than *trans*-regulatory divergence (Emerson *et al.* 2010). In our study, hundreds of genes related to transcriptional regulation with *cis*- changes showed biased directionality. It is clear from this result that positive, directional selection is contributing to a non-negligible proportion of regulatory divergence.

2.3.2. Mis-expression in sterile hybrids

In crosses between *M. m. musculus* (PWK/PhJ) females and *M. m. domesticus* (LEWES/EiJ) males, hybrid males have significantly smaller testes and lower sperm counts compared to hybrid males in the reciprocal cross (Good *et al.* 2008a) (see Table S16 for phenotypes of the mice in this study). We took advantage of the

asymmetrical nature of hybrid male sterility in this cross to identify genes that were uniquely mis-expressed in sterile hybrids. This approach allowed us to separate mis-expression that was associated with hybridization from mis-expression that was associated with sterility. For example, the 28 genes that were mis-expressed in fertile hybrids (Table S5) can be excluded as contributing to reproductive isolation.

Despite the power of this approach, it is important to recognize that it does not allow us to directly identify genes causing sterility. The set of genes that are mis-expressed only in sterile hybrids is expected to include causative genes, but it may also include genes that are mis-expressed as downstream effects of genes causing sterility. The latter category is likely inflated by differences in the cellular composition of testes in fertile and sterile animals. Testes contain a heterogeneous mixture of cell types; sterile and fertile hybrids contain different proportions of somatic, mitotic, early meiotic and postmeiotic cells. For example, in the well-studied cross between *M. m. domesticus*$^{C57BL/6J}$ and *M. m. musculus*^{PWD} in which *Prdm9* is implicated in hybrid male sterility, essentially complete meiotic arrest occurs in pachytene with spermatocytes undergoing apoptosis (Mihola *et al.* 2009). Nonetheless, several lines of evidence suggest that differences in cellular composition are not the main cause of the expression differences we have identified here. First, in contrast to the cross between *M. m. domesticus*$^{C57BL/6J}$ and *M. m. musculus*^{PWD}, meiotic arrest is incomplete in the cross performed here: cells from all stages of spermatogenesis can be found in the testes of sterile males although the proportions differ in sterile and fertile animals. Second, we would expect to see a greater effect of cellular composition on X-linked gene expression than autosomal expression in whole testis, since transcription on the X Chromosome is largely silenced from pachytene through the later stages of spermatogenesis (Turner 2007). Despite this expectation, there is close agreement between our finding of X-linked overexpression and the results of Campbell *et al.* (2013), who studied X-linked expression in flow-sorted germ cells for the same genotypes. Campbell *et al.* (2013) showed that the over-expression of the X Chromosome in sterile hybrid males from this cross is not an artifact of changes in cellular composition but reflects major shifts in gene expression in sterile animals that occur in individual cell types. Third, all patterns of allele-specific expression documented here are robust to cellular composition since they were determined only in fertile F1 males, which have the same cellular composition as the parents. Finally, the strong association between *cis-trans* compensatory evolution and mis-expression (Table 4) would not be expected if cellular composition is the primary driver of differences in gene expression between fertile and sterile hybrids (unless different cell types showed differences in the amount of compensatory regulatory evolution, a pattern that is not seen; supplementary methods and Tables S17 and S18).

Numerous studies have established a central role for the X Chromosome in hybrid male sterility in house mice. Quantitative trait locus mapping of sterility phenotypes (White *et al.* 2011; Bhattacharyya *et al.* 2014), phenotyping of introgression lines (Oka *et al.* 2004; Good *et al.* 2008b; Campbell *et al.* 2013; Oka *et al.* 2014), and studies of introgression across the hybrid zone (e.g. Tucker *et al.* 1992; Dod *et al.* 1993), have all suggested that loci on the *M. m. musculus* X Chromosome contribute to postzygotic isolation between the subspecies. Mis-

expression of *M. m. musculus* X-linked genes in sterile hybrids is associated with disruption of meiotic sex chromosome inactivation (MSCI), the process of transcriptional silencing of X and Y chromosomes during spermatogenesis (Good *et al.* 2010; Campbell *et al.* 2013; Bhattacharyya *et al.* 2013, 2014). The upregulation of X-linked genes in sterile hybrids seen here is consistent with this earlier work.

Previous studies have also implicated numerous autosomal loci in in reproductive isolation in mice (e.g., Forejt and Ivanyi 1974; Mihola *et al.* 2009; White *et al.* 2011; Janoušek *et al.* 2012; Forejt *et al.* 2012; Phifer-Rixey *et al.* 2014). In particular, Mihola *et al.* (2009) characterized a gene on chromosome 17, *Prdm9*, which interacts with the *M. m. musculus* X Chromosome to drive hybrid sterility; however, sterility in the cross studied here is not associated with known variants of this locus (Good *et al.* 2008b, Good *et al.* 2010). Nonetheless, we identified 202 autosomal genes with aberrant expression only in the sterile hybrid, consistent with the idea that autosomal genes contribute to hybrid male sterility. Interestingly, these were enriched for GO categories associated with gene regulation. Several of the aberrantly expressed genes where *cis* and *trans* variants act in opposition are known from previous work to play a role in spermatogenesis, cell cycle control, or to be expressed mainly in the testis (Table 3). Candidates of particular interest which deserve follow-up in future studies are *Myl10* and *Mtcl1,* both of which were identified in regions of overlap between our study, a study identifying peaks of differentiation between the subspecies, a QTL mapping study on markers of hybrid sterility, and regions of low introgression across the hybrid zone (Phifer-Rixey *et al.* 2014; White *et al.* 2011; Janoušek *et al.* 2012). More detailed characterization of the phenotype of hybrid sterility in this cross will be useful for elucidating the role of particular genes.

We also found a highly significant over-representation of genes showing positive selection among those that were aberrantly expressed only in sterile hybrids. Because the test we used was restricted to those genes showing *cis* changes alone, the nature and identity of the interacting loci, if any, are unknown. Nonetheless, an emerging pattern from studies of the genetics of postzygotic isolation is that most of the identified genes show signatures of positive selection (Presgraves 2010). Our results are certainly consistent with this emerging picture and further suggest that selection on regulatory changes contributes to the evolution of reproductive isolation.

Previous studies have identified an association between c*is*- and *trans*-changes favoring the expression of the opposite allele and mis-expression in hybrids (Landry *et al.* 2005; Tirosh *et al.* 2009; McManus *et al.* 2010; Schaefke *et al.* 2013). Landry *et al.* (2005) first identified an association between compensatory coevolution between *cis* and *trans* elements and mis-expression in hybrids. While this initial study made powerful predictions as to how regulatory divergence *could* result in reproductive incompatibilities between species, a phenotypic association with this pattern that is separable from expression differences associated with hybridization has been lacking until now.

The Dobzhansky-Muller model of postzygotic isolation is one of the cornerstones in our understanding of the genetics of speciation (Coyne and Orr 2004). Despite the fact that gene regulation necessarily involves interactions

between loci, there have been few systematic attempts to link disruptions in gene regulation across the genome to phenotypes underlying reproductive isolation (Turner *et al.* 2014). Here we showed that genes that are mis-expressed uniquely in sterile hybrid males are associated with opposing changes in *cis* and *trans.* Strictly compensatory changes (i.e., where expression levels in both subspecies are the same) were particularly enriched in genes with aberrant or mis-expression. These results provide strong evidence that compensatory regulatory evolution may underlie Dobzhansky-Muller incompatibilities and contribute to reproductive isolation between *M. m. musculus* and *M. m. domesticus*.

2.4. Materials and Methods

2.4.1. *Samples*

M. m. musculus was represented by whole testis from the wild-derived inbred strains PWK/PhJ and CZECHII/EiJ (hereafter, *M. m. musculus*^{PWK} and *M. m. musculus*^{CZII}), and *M. m. domesticus* was represented by whole testis from the LEWES/EiJ and WSB/EiJ strains (hereafter, *M. m. domesticus*^{LEWES} and *M. m. domesticus*^{WSB}).

Hybrids were generated from reciprocal crosses between *M. m. musculus*^{PWK} and *M. m. domesticus*^{LEWES}. Male hybrids in this cross are sterile when the mother is *M. m. musculus*^{PWK} and fertile when the mother is *M. m. domesticus*^{LEWES}. To circumvent the problem of inbreeding depression in pure species, we crossed *M. m. musculus*^{PWK} females to *M. m. musculus*^{CZII} males, and *M. m. domesticus*^{LEWES} females to *M. m. domesticus*^{WSB} males.

2.4.2. *Sequencing and mapping*

For each sample, 100 base-pair paired-end reads were sequenced from mRNA on the Illumina HiSeq2000 platform. A mean of 7.5 Gb of sequence was obtained for each sample.

Subspecies were mapped with the program Tophat (Kim *et al.* 2013) to the appropriate pair of reference genomes (either *M. m. musculus*^{PWK} and *M. m. musculus*^{CZII} or *M. m. domesticus*^{LEWES} and *M. m. domesticus*^{WSB}) as well as to the opposite maternal reference (*M. m. domesticus*^{LEWES} or *M. m. musculus*^{PWK}). Hybrids were mapped to *M. m. musculus*^{PWK} and *M. m. domesticus*^{LEWES}. Only reads that mapped preferentially to one subspecies were retained for further analysis. See supplemental methods for information on the reference genomes used for mapping.

On average, a greater proportion of reads mapped to *M. m. musculus*^{PWK} per sample than to *M. m. domesticus*^{LEWES} (see Table S1). This difference may be due to real differences in allelic expression or due to a mapping bias; to account for the difference in the number of allele-specific reads across samples, reads were later randomly downsampled across samples (see below).

2.4.3. *Regulatory divergence*

An equal number of reads from each parental sample were combined to create a mixed parental pool comparable to allele-specific counts in fertile hybrids. Downsampling was chosen to equalize power across comparisons as described in

Coolon *et al.* (2014). Reads were then pooled for the following categories: 1) *M. m. musculus* subspecies reads, 2) *M. m. domesticus* subspecies reads, 3) Fertile hybrid *M. m. musculus* allelic reads, and 4) Fertile hybrid *M. m. domesticus* allelic reads. Genes with fewer than 20 reads for any sample or allele were excluded. Genes were sorted into regulatory categories based on a binomial test between reads mapping to each parent, a binomial test between reads mapping to each allele in the fertile hybrid, and a Fisher's exact test comparing these values (see supplemental methods for details on regulatory divisions) (Wittkopp *et al.* 2004; McManus *et al.* 2010). As described by Goncalves *et al.* (2014), *cis + trans* can further be subdivided into genes where *cis* and *trans* are acting in the same direction (hybrid ratio < pure species ratio) or opposite directions (hybrid ratio > pure species ratio).

2.4.4. *Inheritance patterns*

After reads were mapped and counted, reads mapping to *M. m. domesticus*^{LEWES} and *M. m. musculus*^{PWK} were combined for each sample for total hybrid counts. Mapped reads from pure species and hybrids were downsampled to an equivalent number per sample and then pooled by genotype (metaseqR; Moulos and Hatzis 2014).

2.4.5. *Testing for enrichment of opposing or reinforcing* cis- *and* trans- *changes*

The expected numbers of *cis*- and *trans*- changes acting in the same or opposing directions were calculated based on the proportion of negative and positive *cis*- and *trans*- changes (Table S15). Expected numbers were calculated by multiplying the proportion of directional independent *cis*- and *trans*- changes together and then in opposition by the total number of genes with divergence in both *cis* and *trans.*

2.5. Data Access

The sequencing data generated for this study have been submitted to the NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject/) under BioProject ID PRJNA286765.

2.6. Chapter 2 Tables

Table 1. Studies that have identified regulatory divergence due to changes in *cis*- and *trans*- between species.

Species	Comparison	Divergence time	Tissue	*cis* v. *trans*[1]	Misexpression[2]	CAWM[3]	Citation
Insects							
Drosophila melanogaster x *D. simulans*	Interspecific	2.5 mya	whole fly	*cis*	No		Wittkopp et al. 2004
D. melanogaster x *D. simulans*	Interspecific	2.5 mya	whole fly	*cis*	Yes	Yes	Landry et al. 2005
D. melanogaster & *D. simulans*	Intra- and interspecific	2.5 mya	whole fly	*trans (intra-), cis (inter-)*	No		Wittkopp et al. 2008
D. melanogaster x *D. simulans*†	Interspecific	2.5 mya	head, body	*cis*	Yes	N/A[4]	Graze et al. 2009
D. melanogaster x *D. sechellia*†	Interspecific	2.3 mya	whole fly	*trans*	Yes	Yes	McManus et al. 2010
D. melanogaster; *D. simulans* x *D. sechellia; D. melanogaster* x *D. simulans*†	Intra- and interspecific	10,000; 250,000; 2.5 mya	whole fly	*trans*	Yes	No	Coolon et al. 2014
Fungi							
Saccharomyces cerevisiae x *S. paradoxus*†	Interspecific	5 mya	-	*cis*	Yes	Yes	Tirosh et al. 2009
S. cerevisiae†	Intraspecific	-	-	*trans*	No		Emerson et al. 2010
S. cerevisiae†	Intraspecific	-	-	*trans*	Yes	Yes	Schaefke et al. 2013
Plants							
Populus trichocarpa x *P. deltoides*	Interspecific		leaf, stem	*cis*	No		Zhuang & Adams 2007
Arabidopsis thaliana x *A. arenosa*†‡	Interspecific	6 mya	leaf	*cis*	No		Shi et al. 2012
Cirsium arvense†	Intraspecific	-	leaf	*trans*	Yes	No	Bell et al. 2013
Zea mays ssp. parviglumis x *Z. m.* ssp. mays†	Interspecific	9,00	ear, leaf, stem	*cis*	No		Lemmon et al. 2014
Coffea canephora x *C. eugenioides*†‡	Interspecific		leaf	*trans*	No		Combes et al. 2015
Mammals							
M. m. domesticus x *M. m. castaneus*†	Intersubspecific	350,000 - 1 mya	liver	*cis*	No		Goncalves et al. 2012
M. m. domesticus x *M. m. castaneus*†	Intersubspecific	350,000 - 1 mya	retina	*cis*	No		Shen et al. 2014

[1]Regulatory divergence primarily attributed to *cis* or *trans* variants in crosses.
[2]Mis-expression tested for in crosses.
[3]Compensatory evolution associated with mis-expression
[4]Association between mis-expression and compensatory evolution not formally tested.
†Genome-wide analysis (RNAseq or microarray).
‡Hybrids are allopolyploid

Table 2. An enrichment of *cis*- and *trans*- changes that act in opposition compared to changes that act in the same direction

	Negative fold change		Positive fold change	
Direction	**Expected**[3]	**Observed**	**Expected**[3]	**Observed**
Opposing[1]	1256	2257	931	1326
Same[2]	1069	352	1093	414

[1]Opposing refers to instances where *cis* and *trans* variants act in opposing directions. This includes genes categorized as *cis* x *trans*, compensatory, and a subset of *cis* + *trans* (where variants act in opposition)

[2]Same refers to instances where *cis* and *trans* variants act in the same direction. This includes genes categorized *cis* + *trans* where variants act in the same direction

[3]Expected values are based on the proportion observed when *cis* or *trans* changes occur by themselves (see Methods).

Table 3. Aberrantly expressed genes in the sterile hybrid with phenotypes of interest for hybrid incompatibilities.

Gene name	Associated function/Expression[1]	Direction[2]	Regulatory category
Arl8a	Cell cycle; chromosome segregation; mitotic nuclear division; cell division	+	*cis + trans*, opposing
Brd4	Positive regulation of G2/M transition of mitotic cell cycle	+	*cis + trans*, opposing
Cherp	Negative regulation of cell proliferation; RNA processing	+	Compensatory
Cib4	Highly expressed in testis	-	*cis + trans*, opposing
Cited2	Male gonad development	+	*cis + trans*, opposing
Crisp2	Testis-specific expression	+	*cis* by *trans*
Ctdsp1	Negative regulation of G1/S transition of mitotic cell cycle	+	Compensatory
Cul7	Mitotic cytokinesis; regulation of mitotic nuclear division	+	Compensatory
Gm5617	Testis-specific expression	+	Compensatory
Hspa8	Heat shock protein; regulation of cell cycle	+	*cis* by *trans*
Hspb1	Heat shock protein; negative regulation of apoptotic signaling pathway	+	Compensatory
Kat2a	Cell proliferation; chromatin binding	+	Compensatory
Mad1l1	Mitotic nuclear division, mitotic spindle assembly checkpoint	+	*cis + trans*, opposing
Map3k9	Apoptotic process; cell death	+	Compensatory
Morc2b	Testis-specific expression	-	*cis* by *trans*
Mtcl1*	Microtubule crosslinking factor	+	*cis* by *trans*
Myl10*	Testis-specific expression	-	*cis + trans*, opposing
Phactr4	Regulation of cell cycle	+	*cis + trans*, opposing
Plcz1	Testis-specific expression	-	Compensatory
Ppp1r42	Highly expressed in testis; microtubule organizing center	+	*cis* by *trans*
Prm1	Spermatogenesis; mutants associated with deformed	+	Compensatory

	sperm		
Prm2	Spermatogenesis; mutants associated with deformed sperm	+	*cis* by *trans*
Sh3bp4	Negative regulation of cell proliferation; positive regulation of autophagy; negative regulation of cell growth	+	Compensatory
Usf2	Homozygous null mutants males are usually infertile	+	Compensatory
Zbtb16	Male germ-line stem cell asymmetric division; homozygous mutants develop infertility	+	Compensatory

[1]Phenotype and expression data from Mouse Genome Informatics (Eppig *et al.* 2015) and Su *et al.* (2004), available through BioGPS (Wu *et al.* 2009)

[2]The direction of change between pure species and the sterile hybrid (i.e., genes designated with a "+" are expressed above the level of both pure species)

*Gene names have been identified in regions of overlap between a hybrid zone study, a differentiation study, and a QTL mapping study between *M. m. musculus* and *M. m. domesticus*

2.7. Chapter 2 Figures

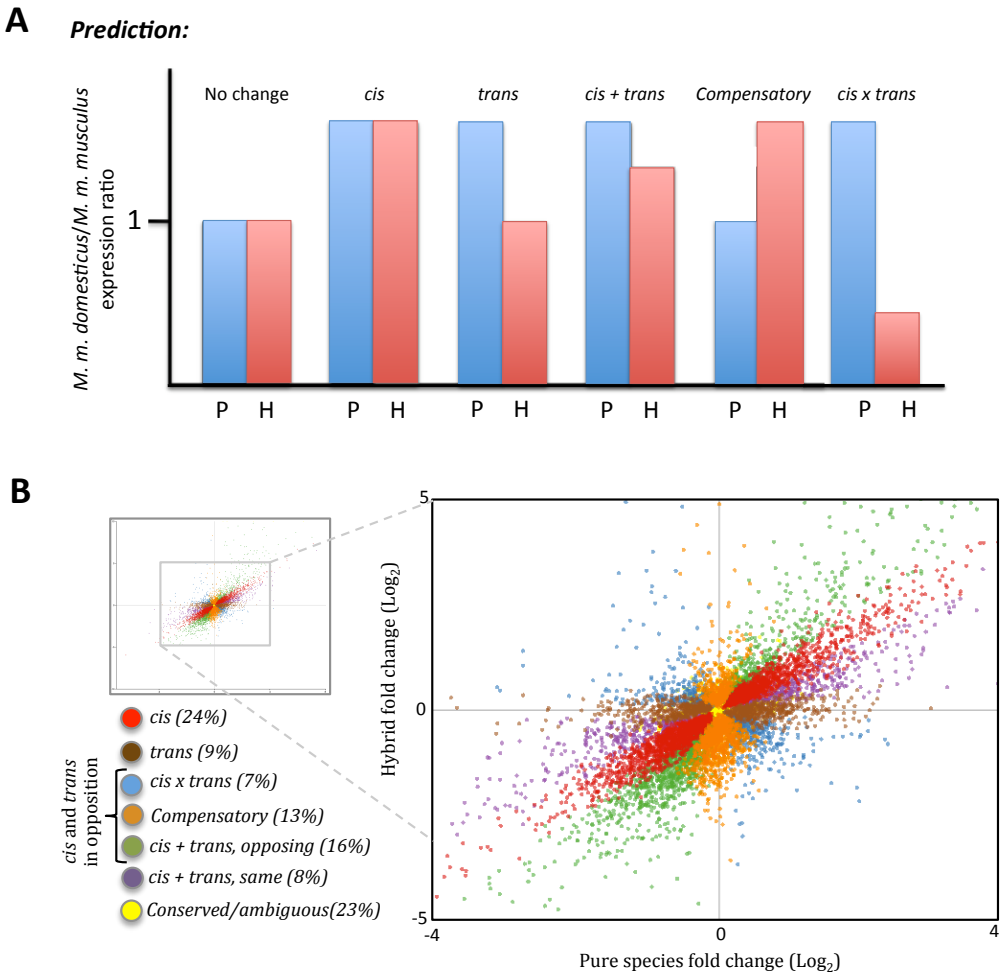


Figure 1. A) Categories of regulatory divergence between *M. m. musculus* and *M. m. domesticus* inferred from gene expression levels in pure subspecies and hybrids, where P and H are the ratio of reads mapping to *M. m. domesticus* versus *M. m. musculus* in the pure species and hybrids, respectively. B) The relative distribution of regulatory categories in this dataset. Each point represents one gene. Points represent $\log_2$ fold changes between reads mapping to each allele in the hybrid (*M. m. domesticus/M. m. musculus*) and the reads mapping to each subspecies (*M. m. domesticus/M. m. musculus*). Genes are color-coded based on their inferred regulatory category.

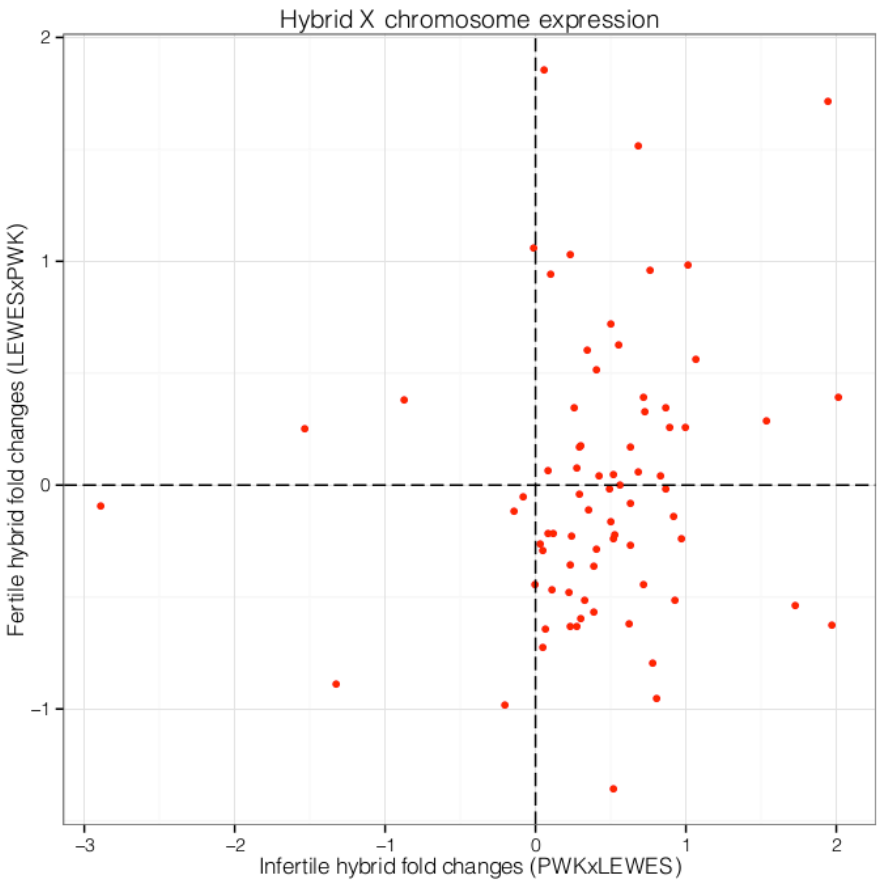


Figure 2. Expression on the X Chromosome in reciprocal hybrids. Each point represents one gene.

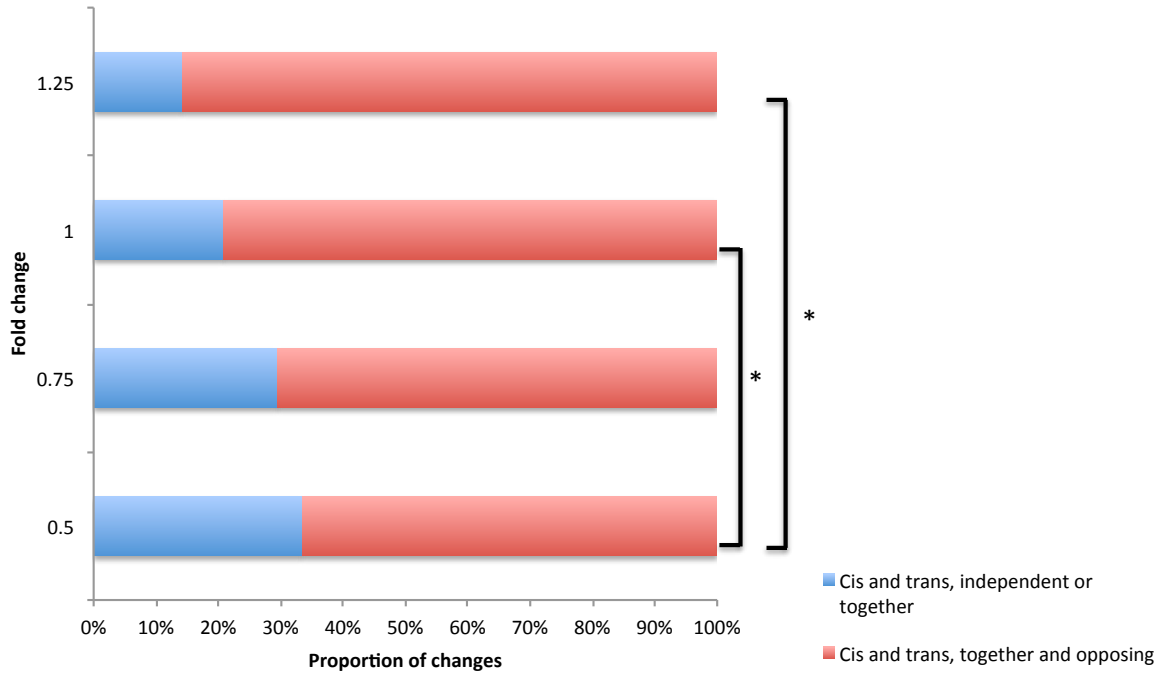


Figure 3. The relationship between the magnitude of expression differences and the number of genes in different regulatory categories. Larger fold changes between both subspecies and the sterile hybrid are associated with a greater proportion of genes where *cis* and *trans* variants act in opposition to one another.

2.8. Chapter 2 Supplemental material

2.8.1. Sampling and mouse husbandry

The wild-derived inbred strains used in this study were purchased from The Jackson Laboratory (http://jaxmice.jax.org), and maintained at the University of Arizona in accordance with Institutional Animal Care and Use Committee (IACUC) protocols. *M. m. musculus* was represented by the PWK/PhJ and CZECHII/EiJ strains (hereafter, *M. m. musculus*^{PWK} and *M. m. musculus*^{CZII}), and *M. m. domesticus* was represented by the LEWES/EiJ and WSB/EiJ strains (hereafter, *M. m. domesticus*^{LEWES} and *M. m. domesticus*^{WSB}).

Hybrids were generated from reciprocal crosses between *M. m. musculus* ^{PWK} and *M. m. domesticus*^{LEWES}. This cross results in infertile hybrid males in one direction and fertile hybrid males in the reciprocal direction. Infertile hybrid males in this cross have significantly reduced testis weight and sperm count compared to pure subspecies (Good *et al.* 2008). For simplicity, hereafter we refer to these hybrid males with lowered fertility as "sterile" though sterility is not complete in all individuals. Individuals for this experiment where chosen at random, without regard for testes weight, sperm count, or other reproductive phenotypes. Phenotypes for these mice are listed in Table S16. Sterile hybrid samples have a lower mean testes weight and sperm count than the fertile hybrid samples or the pure subspecies samples.

Since the parents in this cross are fully inbred while the hybrids are not, differences in expression between the parents and hybrids could be due to inbreeding depression. To circumvent this problem, we crossed *M. m. musculus*^{PWK} females to *M. m. musculus*^{CZII} males, and *M. m. domesticus*^{LEWES} females to *M. m. domesticus*^{WSB} males. The heterozygous progeny of these intrasubspecific crosses will henceforth be referred to simply as *M. m. musculus* and *M. m. domesticus.*

All animals were 70 days old, unmated, and singly housed at time of euthanasia. Testes were dissected under RNAse free conditions and placed in RNAlater at 4°C overnight and then moved to -80°C. RNA was then extracted with Qiagen's RNAeasy Plus Mini Kit. Three biological replicates were collected for each genotype.

2.8.2. Sequencing and Mapping

For each sample, 100 base-pair paired-end reads were sequenced from mRNA on the Illumina HiSeq2000 platform. A mean of 7.5 Gb of sequence was obtained for each sample. Illumina adaptors and the trailing 3 bases were clipped with Trimomatic (Bolger *et al.* 2014).

Tophat (v2.1, settings: --b2-sensitive) was used to map reads from each sample to the appropriate reference genomes (Kim *et al.* 2013). References for PWK and WSB, based on the Wellcome Trust's SNP and indel calls on C57bl/6, are publically available (Turro *et al.* 2011); consubspecific CZECHII and LEWES references were constructed by inserting SNPs called with SAMTOOLS mpileup into the PWK and WSB reference genomes, respectively. SNPs were called based on a pileup against the PWK and WSB genomes with these samples as well as transcriptome data from a recent study (see Phifer-Rixey *et al.* 2014 for SNP calling).

Subspecies were mapped to the appropriate pair of reference genomes (either *M. m. musculus*^{PWK} and *M. m. musculus*^{CZII} or *M. m. domesticus*^{LEWES} and *M. m. domesticus*^{WSB}) as well as to the opposite maternal reference (*M. m. domesticus*^{LEWES} or *M. m. musculus*^{PWK}). Reads that mapped preferentially or equally well to the incorrect subspecies were discarded at the 0, 1, and 2 mismatch thresholds. For example, if a read from a *M. m. musculus* sample mapped equally well to both subspecies reference genomes with zero mismatches, the read was discarded. Alternatively, if the read mapped to the *M. m. musculus* reference with 0 mismatches and to the *M. m. domesticus* reference with 1 mismatch, it was retained for further analysis.

Hybrids were mapped to *M. m. musculus*^{PWK} and *M. m. domesticus*^{LEWES}. Only reads that mapped preferentially to *M. m. musculus*^{PWK} or *M. m. domesticus*^{LEWES} were retained for further analysis. Reads that mapped equally well to both *M. m. musculus*^{PWK} and *M. m. domesticus*^{LEWES} at 0, 1, and 2 mismatch thresholds were discarded so as to retain only reads for which an allele specific to *M. m. musculus* or *M. m. domesticus* could be assigned. On average, a greater proportion of reads mapped to *M. m. musculus*^{PWK} per sample than to *M. m. domesticus*^{LEWES} (Table S1). This difference may be due to real differences in allelic expression or due to a mapping bias; to account for the difference in the number of allele specific reads across samples, reads were later randomly downsampled across samples (see below).

Gene annotations were converted by creating LiftOver chains between reference genomes and the C57BL/6J mouse reference genome using BLAT and liftover command line programs (personal script, available on request). Chains files were then used with the UCSC command line LiftOver tool to convert annotations for C57BL/6J (downloaded from Ensembl, v68) between assemblies (Kuhn *et al.* 2007). The HTSeq-count package was used to count reads mapped to features (settings: --stranded=no –mode=union) (Anders *et al.* 2015).

2.8.3. Preprocessing of mapped reads

A principle components (PC) analysis separated *M. m. musculus* samples from *M. m. domesticus* samples along PC1, explaining nearly 70% of the variation between samples. Spearman's Rank-Order Correlation was calculated between samples ($p < 2.2e\text{-}16$; Table S14). As expected based on the PC analysis, correlations between replicates within subspecies were considerably higher than comparisons between subspecies (0.95-0.99 for intrasubspecific comparisons, 0.77-0.83 for intersubspecific comparisons). Correlations between expression levels in the sterile and fertile hybrid and each parent were also high (Fig. S1).

2.8.4. Characterizing regulatory divergence

An equal number of reads from each parental sample were combined to create a mixed parental pool comparable to allele-specific counts in hybrids. Downsampling was chosen to equalize power across comparisons as described in Coolon *et al.* (2014). A similar method was also recently employed for downsampling allele-specific reads in humans (Lappalainen *et al.* 2013). Using simulations, Coolon *et al.* 2014 found that such a downsampling approach produces data sets with the same power to detect significant expression differences with Fisher's exact tests as data sets collected at smaller sample sizes.

Reads were then pooled for the following categories: 1) *M. m. musculus* subspecies reads, 2) *M. m. domesticus* subspecies reads, 3) Hybrid *M. m. musculus* allelic reads, and 4) Hybrid *M. m. domesticus* allelic reads. After excluding genes with fewer than 20 reads for any sample or allele (McManus *et al.* 2010), 9,851 autosomal genes could be compared. For each gene, we performed binomial tests between reads mapping to *M. m. domesticus* and *M. m. musculus*, and between reads mapping to the *M. m. domesticus* allele and *M. m. musculus* allele in hybrids. A Fisher's exact test was performed comparing the ratio of *M. m. domesticus* to *M. m. musculus* reads between subspecies and within hybrids. Binomial and Fisher's exact tests were implemented in R (v 3.1.1, CRAN) using the *binom.test* and *fisher.test*, respectively. *P*-values were corrected for a false discovery rate (FDR) of 0.05 with the R's *p.adjust*. More conservative cutoffs (0.01 *p*-value for the binomial and Fisher's exact test, 0.01 FDR correction) had minimal effects on the overall results of our analysis and are provided in Table S2.

To identify regulatory categories for each gene, we first calculated H, the ratio of the number of reads mapping to the *M. m. domesticus* allele compared to the number of reads mapping to the *M. m. musculus* allele in hybrids, and P, the ratio of the number of reads in the *M. m. domesticus* parent compared to the number of reads in the *M. m. musculus* parent. The significance for individual values of P and H was assessed using binomial tests. The significance of P/H values was assessed using Fisher's exact tests. Regulatory categories were then defined as follows (Fig. 1) (McManus *et al.* 2010): 1) *cis* only: H is significant, P, is significant, and P/H is not significant; 2) *trans* only: H is not significant, P is significant, and P/H is significant; 3) *cis + trans:* H is significant, P is significant, P/H is significant, and P and H have the same sign; 4) *cis x trans:* H is significant, P is significant, P/H is significant, and P and H have opposing signs; and 5) compensatory: H is significant, P is not significant, P/H is significant (Wittkopp *et al.* 2004; McManus *et al.* 2010). As described by Goncalves *et al.* (2014), *cis + trans* can further be subdivided into genes where *cis* and *trans* are acting in the same direction (H<P) or opposite directions (H>P). The latter reflects compensatory evolution that is not complete.

We note that differences in experimental methods (e.g., depth of sequencing, number of replicates) can affect measures of allelic imbalance with a single species. The power to detect allele-specific expression is largely based the number of informative transcripts sequenced. Mathematical modeling and computer simulations have demonstrated that at least 5-10 million reads are necessary to assess allele-specific expression in a genome where approximately 10,000 genes are expressed (Fontanillas *et al.* 2010), a threshold we surpass (see Table S1). Additionally, inter-individual variation due to polymorphisms, environmental factors (e.g., facility, method of euthanasia, presence of other individuals), and mRNA quality can affect gene expression (Buckland 2004). We attempted to mitigate these risks in our study by using wild-derived inbred lines, age matching individuals, and raising all animals under the same environmental conditions (facility, diet, singly housed at time of euthanasia, etc.).

2.8.5. Polygenic test for selection

Genes categorized as divergent in *cis* alone were subjected to a test for polygenic selection as proposed by Bullard *et al.* (2010) and Fraser *et al.* (2010, 2011). This approach is a modification of a sign test originally proposed by Orr (1998) for detecting selection on quantitative traits. Gene Ontology (GO) Consortium terms were chosen as grouping terms.

The null expectation of this test makes no assumptions about a given gene regulatory network other than that the *cis* changes in a gene set are independent. A total of 4,852 GO categories were represented by at least one gene. Each *cis*-regulated variant was assigned either a "+" or "-" depending on the direction of the $\log_2$-fold change. This designation is partially arbitrary as a "+" can either indicate upregulation in *M. m. domesticus* or downregulation in *M. m. musculus*. However, under neutrality we expect a 1:1 ratio of genes designated as "+" and "-" in each gene set. As directionality of genes categorized as divergent in *cis* was not exactly 1:1 (1,093 upregulated and 1,251 downregulated relative to *M. m. domesticus*), we adjusted our null expectation to accommodate this difference. When genes in the same GO category were within 100,000 base pairs of one another, the gene with the lower fold change was eliminated from the set. This was done so as to exclude non-independent loci; linkage disequilibrium in mice decays well within this distance (Laurie *et al.* 2007) so loci separated by more than 100 kb typically have independent evolutionary histories. This had no effect on the gene-sets we identified with biased directionality. A hypergeometric test was applied to each gene set and *p*-values were adjusted to a 5% false discovery rate.

2.8.6. Characterizing inheritance patterns and mis-expression

After reads were mapped and counted with HTSeq-count as described above, reads mapping to *M. m. domesticus*$^{\text{LEWES}}$ and *M. m. musculus*$^{\text{PWK}}$ were combined for each sample for total hybrid counts. Mapped reads from pure species and hybrids were downsampled to an equivalent number per sample to equalize between library sizes (metaseqR; Moulos and Hatzis 2014). After downsampling, we retain 17,535,821 reads for each genotype. Genes with fewer than 20 reads (for autosomes) or 10 reads (for the heterogametic X chromosome) per sample were eliminated, and remaining reads were pooled for each genotype. As this cutoff is less restrictive than that for allele-specific reads, more genes could be compared. Inheritance patterns were determined with a 1.25-$\log_2$ fold difference cut-off between parents and sterile and fertile hybrids (Gibson *et al.* 2004, McManus *et al.* 2010). Inheritance patterns were inferred as follows: 1) *additive:* different from both subspecies but intermediate to their expression levels, 2) *dominant:* different from one subspecies but similar to the other, 3) *mis-expressed:* different from both subspecies, expressed over or under both, and 4) *similar*: similar to expression in both subspecies.

Testes contain a heterogeneous mixture of cell types. Therefore, differences in expression between sterile and fertile animals could reflect differences in cellular composition (Good *et al.* 2010). However, several observations suggest that this is not the main determinant of observed differences in expression between sterile and fertile mice. First, in contrast to the cross with C57bl/6 where germ cells undergo meiotic arrest and apoptose, meiotic arrest in the PWKxLEWES cross is incomplete and germs cells do progress through meiosis II and spermiogenesis, although there is still germ cell loss after mid-pachytene. Second, the close agreement between our results for the X chromosome (see Results) and those of Campbell *et al.* (2013) who studied X-linked expression in flow-sorted cells for the same genotypes suggests that differences in cellular composition are not driving the main patterns reported here. Third, allele-specific patterns of expression are robust to differences in cellular composition, so inferences about regulatory divergence, compensatory evolution, and positively-selected *cis*-regulatory changes cannot arise from changes in cellular composition. Finally, the strong association between *cis*-

trans compensatory evolution and mis-expression (see Results and section entitled "Testing for associations between compensatory changes and cell type" below) is not expected if cellular composition is the primary driver of differences in gene expression between fertile and sterile hybrids.

2.8.7. Characterizing expression on the X chromosome

The expression levels of X-linked genes in the sterile and fertile hybrid were compared. As the X chromosome is hemizygous in males, we tried lowering the minimum number of reads per sample required for comparison to 10 (total >30 reads after pooling for both the fertile and sterile hybrid) as well as testing with the 20 read threshold (total >60 reads after pooling) used for autosomes. While fewer genes could be analyzed with the more stringent cutoff (75 vs. 93), it did not affect the expression patterns we identified. To characterize expression on the X chromosome, fold changes were calculated between the hybrid and the maternal parent (i.e., *M. m. musculus* for the sterile hybrid and *M. m. domesticus* for the fertile hybrid). As each hybrid is only compared to the appropriate parent (i.e., the one with which it shares an X-chromosome), the differences seen in X chromosome expression between fertile and sterile males will not be affected by any mapping bias. Using a $\log_2$ fold-change between the hybrid and the appropriate parent circumvents the issue of mapping bias to X-linked genes.

2.8.8. Testing for enrichment of opposing or reinforcing *cis*- and *trans*- changes

Under neutrality, an equal number of changes in *cis*- and *trans*- should act in opposing and reinforcing directions. Genes categorized as compensatory, *cis* x *trans,* and *cis* + *trans* (in opposition) show evidence of variants acting in opposition. Genes categorized as *cis* + *trans* (same direction) show evidence of variants acting in the same direction. We tested for deviations from this neutral expectation. The expected numbers of *cis*- and *trans*-changes acting in the same or opposing directions were calculated based on the proportion of negative and positive *cis*- and *trans*- changes (Table S15). Expected numbers were calculated by multiplying the proportion of independent *cis*- and *trans*- changes together and then in opposition by the total number of genes with divergence in both *cis* and *trans.*

2.8.9. Testing for associations between compensatory changes and cell type

Whole testes of sterile hybrids and pure species contain different proportions of particular cell types. We tested whether the observed association between mis-expression and compensatory evolution could be the result of these differences in cellular composition. For the observed association between mis-expression and cis:trans compensatory evolution to be driven by differences in cellular composition, there would have to be a greater amount of compensatory evolution in some cell populations than in others. Specifically, mitotic and somatic cells, which presumably represent a greater proportion of the cells in the testes of sterile hybrid males, would have to show more compensatory evolution (as inferred from allele-specific expression in the fertile F1 males) than meiotic or post-meiotic cells. To determine whether this was the case, the GermOnline database (Gattiker *et al.* 2007) was used to associate genes with expression in particular testis cell types. Cell type annotations were derived from microarray experiments on enriched cell populations (Chalmel *et al.* 2007). Testis expression clusters were defined as

follows: somatic (Sertoli cells), mitotic (spermatogonia), meiotic (spermatocytes), and post-meiotic (round spermatids). Genes with multiple annotations were eliminated from this analysis. After associating genes with particular cell types, we found that the proportion of opposing changes in somatic and mitotic cells was not higher than in the other cell types (Tables S17 and S18). The same result holds when only mis-expressed or aberrantly expressed genes were considered (data not shown). This analysis shows that differences in cellular composition are not driving the observed association between mis-expression and cis:trans compensatory evolution.

2.8.10. Identification of Imprinted Genes

Genomic imprinting is an epigenetic phenomenon where the expression of an allele is based on the sex of the parent from which it is inherited. Genomic imprinting is believed to arise as a result of differential methylation during male and female gametogenesis and is maintained through development and adult life (Li and Sasaki 2011). We tested for imprinting in the testes by comparing allele-specific autosomal expression in reciprocal crosses. As above, binomial tests were employed to test for allele-specific expression; resulting *p*-values were corrected for a 5% false discovery rate. To identify preferential expression of maternal alleles, we looked for cases where the maternally expressed allele was expressed significantly higher than the paternal allele in reciprocal crosses. We repeated this analysis with paternal alleles to identify preferential paternal expression. We identified 29 genes whose expression in our study is consistent with imprinting. Twenty-seven genes show preferential paternal (maternally imprinted) expression. Five of these 29 genes have been established as imprinted in previous analyses: *Peg3* (Kuroiwa *et al.* 1996), *Peg10* (Wertz and Herrmann 2000; Ono *et al.* 2003), *Impact* (Hagiwara *et al.* 1997), *Sgce* (Ono *et al.* 2003), and *Zrsr1* (Hatada *et al.* 1993).

We identified fewer putative imprinted loci than recent studies on the liver (Goncalves *et al.* 2014, Pinter *et al.* 2015), embryonic fibroblasts (Pinter *et al.* 2015), tropoblast stem cells (Calabrese *et al.* 2015), and brain tissue (Crowley *et al.* 2015) in mouse crosses. These analyses identified on the order of 50 to 100s of genes with parent-of-origin effects. As imprinting can be highly tissue specific (Prickett and Oakey 2012), this difference may reflect differences in the number of imprinted genes expressed in the testes versus other somatic tissues. Alternatively, this difference could be the result of reduced power in our analysis as a consequence of fewer biological replicates; the number of biological replicates has been shown to affect estimates of allelic balance in previous analyses (Goncalves *et al.* 2014). The presence of the imprinted genes in our study has no effect on any of the conclusions regarding expression evolution and hybrid sterility.

2.9. Chapter 2 Supplemental Tables

Table S1. The raw total number of reads mapping uniquely to each genotype for each sample (*M. musculus*^{PWK}, *M. m. musculus*^{CZII}, *M. m. musculus*^{WSB}, *M. m. musculus*^{LEWES}).

Species/Cross	ID	Total mapped reads	*M. m. d.*^{LEWES}	*M. m. m.*^{PWK}	*M. m. d.*^{CZII}	*M.m. d.*^{WSB}
M. m. domesticus	148	18325938	16801289			1524649
	149	30951560	28815359			2136201
	150	28439873	26450066			1989807
M. m. musculus	151	15737834		14312773	1425061	
	152	33072259		30968238	2104021	
	170	26362361		24687598	1674763	
Sterile hybrids	52	19116520	8347531	10768989		
	278	37107215	19942234	17164981		
	131	13004868	5545928	7458940		
Fertile hybrids	93	13589435	6133434	7456001		
	290	17402955	7603068	9799887		
	272	43177412	18697568	24479844		

Table S2. The number of genes inferred to fall into each regulatory category.

Category	*0.01 FDR*[1]	*0.05 FDR*[2]
cis only	2377	2349
cis + *trans*	1720	2392
Compensatory	1239	1309
cis x *trans*	450	648
trans only	795	883
Conserved	3270	2270
Total	9851	9851

[1]The number of genes inferred to fall into each category at a 1% false discovery rate
[2]The number of genes inferred to fall into each category at a 5% false discovery rate

Table S3. Comparisons of $\log_2$ fold changes between pure subspecies and hybrid genes on the X chromosome.

	Number of genes on the X chromosome	
	Positive fold change	Negative fold change
Sterile hybrid	78	15
Fertile hybrid	43	49

Table S4. Comparisons of $\log_2$ fold changes between pure species and hybrid genes on the X chromosome and pooled autosomes.

	Number of genes (sterile hybrid)	
Chromosomes	Positive fold change	Negative fold change
X	78	15
Pooled autosomes	4332	2387

	Number of genes (fertile hybrid)	
Chromosomes	Positive fold change	Negative fold change
X	43	49
Pooled autosomes	2754	3093

Table S5. Inferred inheritance patterns for autosomal genes based on a 1.25-$\log_2$ fold change cut-off between pure species and hybrids.

	Number of genes	
Inheritance pattern	Sterile hybrid	Fertile hybrid
Mis-expressed		
Over-expressed	55	14
Under-expressed	9	14
Dominant		
M. m. musculus dominant	932	583
M. m. domesticus dominant	404	327
Similar	8834	9300
Additive	37	32

Table S6. Comparisons of over- and under-expressed genes in the fertile and sterile hybrid on the X chromosome.

	Over-expressed[1]	Under-expressed[2]
Sterile hybrid	55	14
Fertile hybrid	9	14

[1]Expressed above the level of the appropriate pure species
[2]Expressed below the level of the appropriate pure species

Table S7. Genes identified as under positive selection are non-randomly associated with aberrant expression in sterile hybrids.

	Selection[1]	No selection[2]
Aberrant expression	12	5
Not aberrantly expressed	398	1951

[1]Cis only eQTLs identified as under selection in a sign test
[2]Cis only eQTLs not identified as under selection in a sign test

Table S8. Distribution of mis-expressed genes in regulatory categories.

	Number of mis-expressed genes[1]	
Regulatory category	Sterile hybrid	Fertile hybrid
cis only	3	3
trans only	0	0
cis + trans, same	3	0
cis + trans, opposing	10	3
Compensatory	15	3
cis x trans	4	2

[1]1.25 $\log_2$ difference in the same direction between the sterile F1 and both subspecies

Table S9. Mis-expression in the sterile hybrid is non-randomly associated with compensatory evolution.

	Number of genes (Sterile F1)	
Regulatory categories	Mis-expressed[2]	Total
Compensatory	15	1294
All other regulatory categories[1]	20	6252

[1]Includes the following regulatory categories: *cis* only, *trans* only, *cis* x *trans, cis+trans*
[2]1.25 $\log_2$ difference in the same direction between the sterile F1 and both subspecies

Table S10. Distribution of mis-expressed genes in regulatory categories in the fertile hybrid.

	Number of genes (Fertile F1)	
Regulatory categories	Mis-expressed[1]	Total
cis and *trans,* independent or same direction	3	3995
cis and *trans* together, opposing	8	3575

[1]1.25 $\log_2$ difference in the same direction between the fertile F1 and both subspecies

Table S11. Regulatory categories associated with aberrantly expressed genes in the sterile hybrid.

Aberrantly expressed genes[1]	
Regulatory categories	**Number of genes**
cis only	17
trans only	7
cis + trans, same direction	6
cis + trans, opposing	29
Compensatory	51
cis x trans	11

[1]Aberrantly expressed genes were defined as genes in the sterile hybrid with read counts greater than 1-$\log_2$ fold different from both subspecies and significantly different from the fertile hybrid based on a binomial test

Table S12. Distribution of aberrantly expressed genes in the sterile hybrid across regulatory categories.

Regulatory categories	Number of genes Aberrantly expressed[1]	Total
cis and *trans*, independent or same direction	30	3968
cis and *trans* together, opposing	91	3492

[1]Aberrantly expressed genes were defined as genes in the sterile hybrid with read counts greater than 1-$\log_2$ fold different from both subspecies and significantly different from the fertile hybrid based on a binomial test

Table S13. Aberrantly expressed genes non-randomly associated with compensatory evolution.

	Number of genes	
Regulatory categories	Aberrantly expressed[1]	Total
Compensatory	51	1258
All other regulatory categories[2]	70	6202

[1]Aberrantly expressed genes were defined as genes in the sterile hybrid with read counts greater than 1-$\log_2$ fold different from both subspecies and significantly different from the fertile hybrid based on a binomial test

[2]Includes the following regulatory categories: *cis* only, *trans* only, *cis* x *trans, cis+trans*

Table S14. Spearman's Rank-Order Correlations between intra- and intersubspecific replicates of *M. m. musculus* (designated *M. m. m.*) and *M. m. domesticus* (designated *M. m. d.*).

	M. m. m.[151]	*M. m. m.*[152]	*M. m. m.*[170]	*M. m. d.*[148]	*M. m. d.*[149]	*M. m. d.*[150]
M. m. m.[151]		0.96	0.96	0.79	0.8	0.8
M. m. m.[152]			0.99	0.77	0.83	0.83
M. m. m.[170]				0.78	0.83	0.83
M. m. d.[148]					0.95	0.95
M. m. d.[149]						0.99
M. m. d.[150]						

Table S15. The number of genes with positive and negative fold changes in each regulatory category.

Categories	Negative[1]	Positive[2]
cis x trans	372	276
cis + trans, opposing	1020	606
Compensatory	865	444
cis + trans, same direction	352	414
cis alone	1256	1093
trans alone	406	477

[1]Negative fold changes are associated with upregulation in *M. m. musculus* or downregulation in *M. m. domesticus*

[2]Positive fold changes are associated with down regulation in *M. m. musculus* or upregulation in *M. m. domesticus.*

Table S16. Reproductive phenotypes of animals used in this study.

Male #	Genotype[1]	Body mass (g)	Mean testes (mg)	RTW (mg/g)[2]	SV (g)[3]	RSVW (mg/g)[4]	Mean % motility[5]	Count[6]
93	LP	19.73	88.8	4.501	0.094	4.79	95.48	22.4
272	LP	16.56	76.4	4.61	N/A	N/A	N/A	19.6
290	LP	17.17	63	3.67	N/A	N/A	84.56	13.2
148	LW	16.99	86.9	5.11	0.067	3.93	90.48	35.6
149	LW	15.06	83.4	5.54	0.061	4.04	96.83	18.6
150	LW	15.88	92.1	5.8	0.095	5.96	94.11	31.4
151	PC	18.4	88.8	4.83	0.10	5.56	93.81	14.4
152	PC	13.09	77	5.88	0.056	4.31	94.16	23
170	PC	14.67	75	5.11	0.052	3.54	95.63	17.6
52	PL	13.78	54.1	3.93	0.071	5.18	74.18	3.8
131	PL	13.71	48	3.50	0.044	3.18	83.61	4.2
278	PL	17.07	61.1	3.58	N/A	N/A	57.78	3

[1]Genotypes of each sample: LP = LEWESxPWK (fertile hybrid), LW = LEWESxWSB (*M. m. domesticus*), PC = PWKxCZECHII (*M. m. domesticus*), PL = PWKxLEWES (sterile hybrid)
[2]Relative testis weight
[3]Seminal vesicle weight
[4]Relative seminal vesicle weight
[5]Refers to the percentage of live sperm
[6]Number of sperm x10^6

Table S17. Opposing *cis* and *trans* changes with expression in a particular cell type.

Expression cluster	Opposing *cis* and *trans*	Total number of genes	Proportion
Somatic and mitotic	606	3428	0.18
Meiotic and post-meiotic	876	2808	0.31

[1]The total number of genes associated with expression in a particular testis cell type

Table S18. Strictly compensatory changes with expression in a particular cell type.

Expression cluster	Compensatory	Total number of genes	Proportion
Somatic and mitotic	256	3428	0.07
Meiotic and post-meiotic	250	2808	0.09

[1]The total number of genes associated with expression in a particular testis cell type

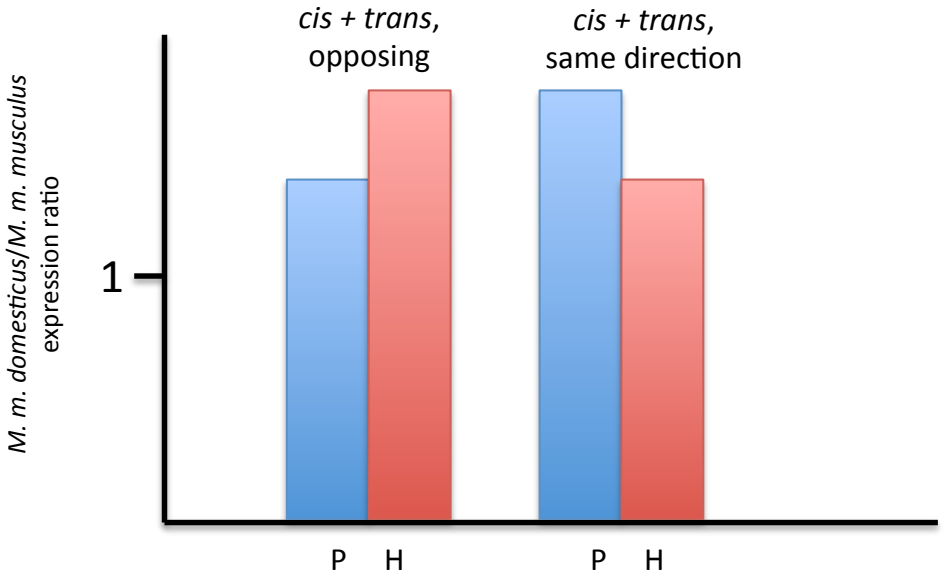


Figure S1. A simplified schematic of the division between the *cis + trans* categories.

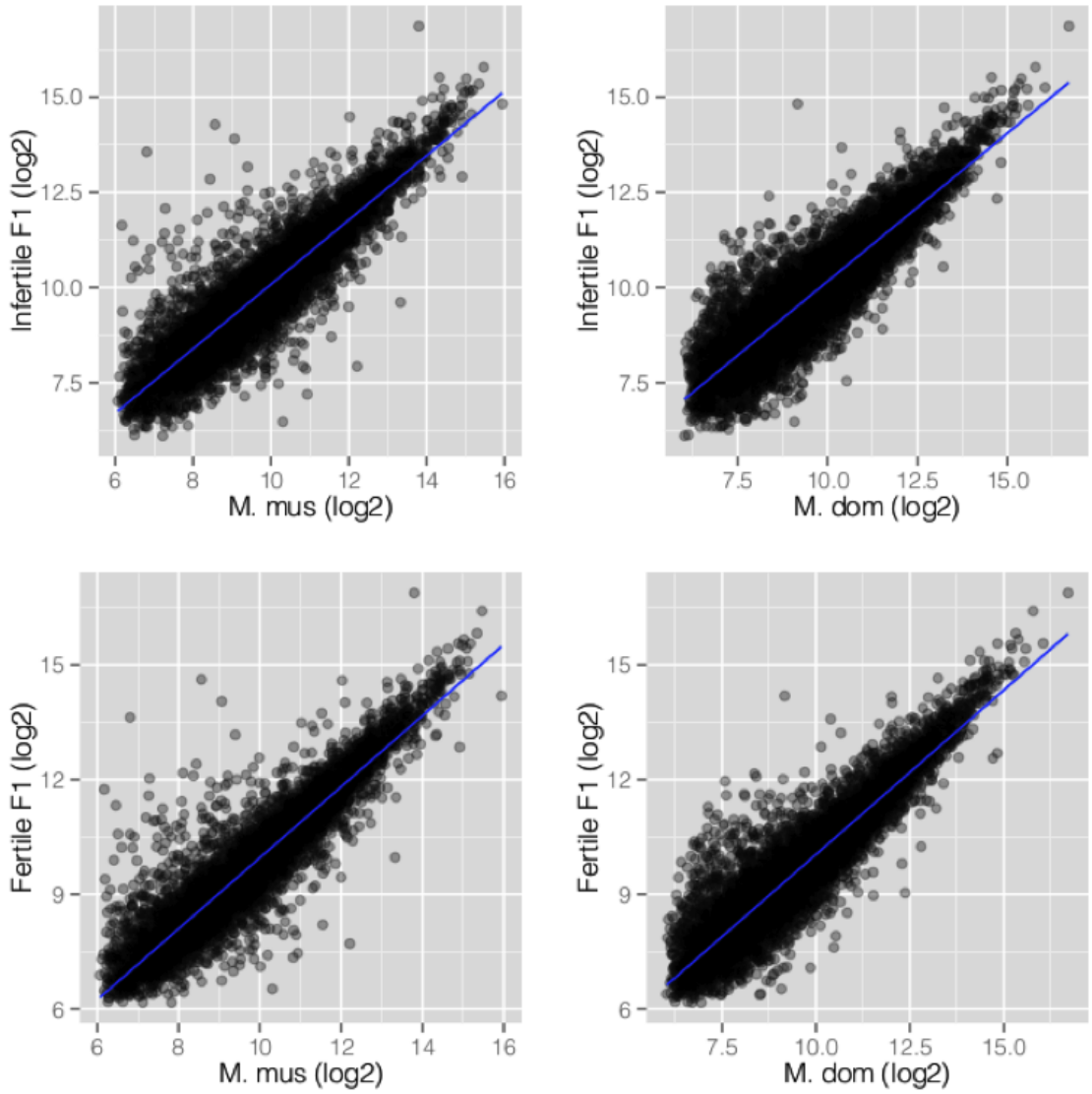


Figure S2. A) Correlation between gene expression levels in sterile hybrid and the *M. m. musculus* parent (designated M. mus) (*corr*=0.90). B) Correlation between gene expression in the sterile hybrid and the *M. m. domesticus* parent (designated M. dom) (*corr*=0.91). C) Correlation between gene expression in the fertile hybrid and the *M. m. musculus* parent (*corr*=0.93). D) Correlation between gene expression in the fertile hybrid and the *M. m. domesticus* parent (*corr*=0.92).

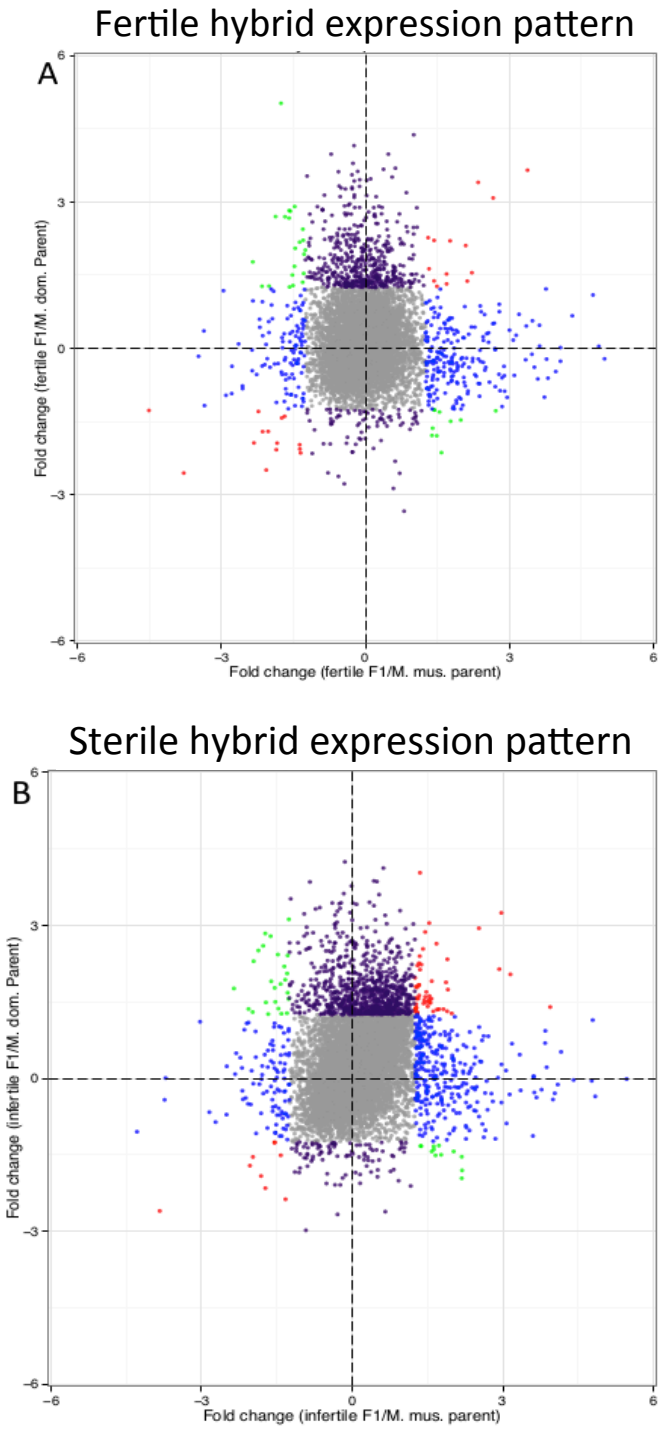


Figure S3. The inferred inheritance patterns for genes in fertile (A) and sterile (B) hybrids. Each point corresponds to a gene's $\log_2$ fold change between the hybrid and each pure species. Each point is colored according to its inferred expression pattern (Red: Mis-expression; Green: Additive; Purple: *M. m. musculus* dominant; Blue: *M. m. domesticus* dominant; Gray: Similar)

Chapter 3

Gene regulation underlies environmental adaptation in house mice

This chapter has been previously published and is reproduced here in accordance with the journal's article sharing policy:

> Mack KL, Ballinger MA, Phifer-Rixey M, Nachman MW. 2018. Gene regulation underlies environmental adaptation in house mice. *Genome Res.* 28:1636-1645.

Abstract

Changes in *cis*- regulatory regions are thought to play a major role in the genetic basis of adaptation. However, few studies have linked *cis*- regulatory variation with adaptation in natural populations. Here, using a combination of exome and RNA-seq data, we performed expression quantitative trait locus (eQTL) mapping and allele-specific expression analyses to study the genetic architecture of regulatory variation in wild house mice (*Mus musculus domesticus*) using individuals from 5 populations collected along a latitudinal cline in eastern North America. Mice in this transect showed clinal patterns of variation in several traits, including body mass. Mice were larger in more northern latitudes, in accordance with Bergmann's rule. We identified 17 genes where *cis*-eQTL were clinal outliers and for which expression level was correlated with latitude. Among these clinal outliers, we identified two genes (*Adam17* and *Bcat2*) with *cis*-eQTL that were associated with adaptive body mass variation and for which expression is correlated with body mass both within and between populations. Finally, we performed a gene co-expression network analysis to identify expression modules associated with measures of body size variation in these mice. These findings demonstrate the power of combining gene expression data with scans for selection to identify genes involved in adaptive phenotypic evolution and also provide strong evidence for *cis*- regulatory elements as essential loci of environmental adaptation in natural populations.

3.1. Introduction

Understanding the genetic basis of adaptation is a major goal in evolutionary biology. *Cis*- regulatory mutations, which can change the expression of proximal genes, have long been predicted to be important targets for adaptive phenotypic evolution (King and Wilson 1975, Wray 2007, Stern and Orgogozo 2008, Wittkopp and Kalay 2012). One reason for this is that *cis*- regulatory mutations may have fewer deleterious pleiotropic effects than protein-coding changes. While protein- coding mutations may affect protein products across tissues and developmental stages, *cis*- regulatory mutations can affect the expression of genes in spatially and temporally specific ways. In apparent support of this idea, several studies have identified positive selection on non-coding regions (e.g., Jenkins *et al.* 1995, Crawford *et al.* 1999, Kohn *et al.* 2004; Andolfatto 2005; MacDonald and Long

2005; Holloway *et al.* 2007; Jeong *et al.* 2008; Torgerson *et al.* 2009) and an important role for non-coding variation in local adaptation (e.g., Jones *et al.* 2012, Fraser 2013).

Despite the accumulating evidence that regulatory loci play an important role in adaptive evolution, there are still only a handful of cases where *cis*- regulatory mutations have been linked to ecologically important traits. Among the best examples are adaptive coat color differences in deer mice (Linnen *et al.* 2013), the ability to digest lactose in humans (Tishkoff *et al.* 2007), and pelvic reduction in sticklebacks (Chan *et al.* 2010). Most examples of adaptive gene expression have been identified through candidate gene approaches, which typically favor traits for which components of a pathway are already known and the genetic basis of the trait is relatively simple. However, most traits are influenced by many loci of small to modest effect. Thus, identifying genetic variants associated with adaptation at complex traits is key to understanding the genetic basis of adaptation.

One avenue for linking adaptive non-coding variation to either molecular or organismal phenotypes is through gene expression. In expression quantitative trait loci (eQTL) mapping, gene expression levels are tested for associations with genetic markers to identify variants that contribute to expression phenotypes. Expression quantitative trait mapping is an effective method for identifying regulatory variants because gene expression is frequently influenced by nearby *cis*-eQTL (Nica and Dermitzakis 2013). *Cis*-eQTL have been successfully detected with small sample sizes (Montgomery and Dermitzakis 2011, Tung *et al.* 2015) and in wild individuals from natural populations (Tung *et al.* 2015). Combining eQTL mapping with genomic scans for selection can be a powerful method for identifying the gene targets of adaptive genetic variation (Fraser 2013, Ye *et al.* 2013) and potentially linking this variation to adaptive organismal phenotypes.

House mice (*Mus musculus domesticus*) provide a useful model for studying the genetic basis of adaptation. House mice are an important biomedical model and have a distribution that mirrors that of human populations (Phifer-Rixey and Nachman 2015). In the eastern United States, house mice show latitudinal variation consistent with local adaptation. Mice collected at northern latitudes are heavier than mice at southern latitudes and their progeny also show differences in a common laboratory environment, indicating that this difference is genetic (Lynch 1992, Phifer-Rixey *et al.* 2018). This observation conforms to the classic ecogeographic observation known as Bergmann's rule that animals in colder climates have larger mass to reduce heat loss (Bergmann 1847). While Bergmann's rule has been observed in many groups, including humans (Ashton *et al.* 2000, Ruff 2002, Foster and Collard 2013), no study so far has linked this pattern to variation at specific genes. Consistent with energetic adaptation of mice from eastern North America, laboratory strains founded from northern and southern locations also show differences in aspects of blood chemistry, including leptin, glucose, and triglyceride levels (Phifer-Rixey *et al.* 2018).

Recent work with these populations identified hundreds of genes with environmental associations in North American (Phifer-Rixey *et al.* 2018). Here we combine a genomic scan for selection with expression quantitative trait (eQTL) mapping to identify regulatory variants that contribute to gene expression differences and show signals of selection in these populations, identifying two strong candidate genes for adaptive phenotypic variation. To our knowledge, this study represents the first case where genomic

scans have been combined with eQTL mapping to identify regulatory variants in natural populations that underlie an adaptive organismal phenotype.

3.2. Results

3.2.1. Cis- regulatory variation in wild house mice

To characterize regulatory variation in wild mice, we sequenced liver transcriptomes from 50 mice collected from five populations along a latitudinal transect on the east coast of North America (Figure 1)(Table S1, File S1). Mice were collected from 29-44°N degrees in latitude. Liver was collected in RNAlater and body mass and length were recorded for each individual. From these individuals, we produced a total of ~1.2 billion RNA-seq reads with an average of 15,473,949 uniquely mapped exonic reads per sample, which were used to quantify gene-wise mRNA abundance (hereafter, gene expression). We also analyzed DNA sequence data generated from exome-capture of the same individuals (Phifer-Rixey *et al.* 2018). Exome and RNA-seq data were used to identify variants segregating in *M. m. domesticus* (see Methods).

We identified *cis*- regulatory variation using two complementary approaches, expression quantitative trait loci (eQTL) mapping and allele-specific expression (ASE). To identify *cis*-eQTLs, we tested for associations between variants within 200-kb of a gene and expression level using a linear mixed model. Variants near a gene are more likely to act in *cis* to affect gene expression. *Cis*-eQTL typically have larger effect sizes than *trans*-eQTL, making then easier to detect in small sample sizes (Montgomery and Dermitzakis 2011). After filtering, a total of 406,999 variants were identified using exome data and tested for associations with expression at 13,080 genes. We identified *cis*-eQTL for 849 of these genes (6.5% of genes surveyed). Reflecting the probe set, the majority of *cis*-eQTL were identified in gene bodies (57%) and introns (18%)(Figure S1).

Allele-specific expression (i.e. differences in expression between parental alleles) can also be used to infer epigenetic or genetic variation acting in *cis* (Cowles *et al.* 2002). As the two parental alleles are exposed to the same *trans*- acting environment within an individual, differences in expression at heterozygous sites can be used to infer *cis*-regulatory variation. A total of 28,234 exonic heterozygous sites, corresponding to 6,738 genes, could be tested for ASE. Across all individuals, we found evidence for ASE for 442 genes at a false discovery rate of 5% (6.7% of genes surveyed)(Table S2).

In investigating the power to detect *cis*- regulatory variation, we found that *cis*-eQTL were more likely to be detected when SNP density is higher near and within the gene of interest (Mann-Whitney *U* test, $p<2.2 \times 10^{-16}$)(Figure S2). We were more likely to detect ASE for genes with higher expression and higher SNP density (Mann-Whitney *U* test, $p=3.1 \times 10^{-11}$ and $p<2.2 \times 10^{-16}$, respectively)(Figures S2,S3). While differences in the power to detect ASE and *cis*-eQTL can lead to the identification of different gene sets, we found significant overlap between the gene sets identified with these analyses (hypergeometric test, $p=5 \times 10^{-6}$, Table S2).

3.2.2. Evidence for adaptive regulatory variation

To assess whether the regulatory variation documented above underlies adaptive difference among populations, we studied sequence and gene expression variation along a latitudinal cline (Figure 1a). Clinal patterns of variation can reflect local adaptation as a response to spatially varying selection (Endler 1977). Regulatory variants with clinal

frequencies that mediate clinal patterns of gene expression would be strong candidates for adaptive regulatory evolution. To identify such variants, we searched for cases where (1) gene expression is clinal, (2) gene expression is associated with a *cis*-eQTL, and (3) allele frequencies of the *cis*-eQTL vary clinally (Figure 2). While geographic clines may alternatively be explained by isolation by distance, there is no evidence for isolation by distance for these populations (see Supplemental material and methods).

To identify clinal patterns of gene expression, we tested for correlations between latitude and expression levels in the liver transcriptomes of the 50 wild individuals. We identified 1,488 genes for which expression was significantly correlated with latitude (P<0.05), 132 of which were associated with a *cis*-eQTL (Figure 2). We also tested for differential expression between the most northern population (New Hampshire/Vermont) and the most southern population (Florida) and identified 458 genes with differential expression between the ends of the cline (Figure S4), 48 of which were associated with a *cis*-eQTL (at q<0.1)(Table S3).

To connect these patterns to clinal sequence variation, a genome scan using the program Latent Factor Mixed Models (LFMM) was performed to test for correlations between latitude and genetic variation while accounting for population structure (Frichot *et al.* 2013)(see methods). For this study, LFMM has an advantage over other methods because it does not assume a specific demographic model, but still accounts for demographic history by estimating genome-wide co-variance among allele frequencies. We focused on SNPs in the 5% tail of the distribution and considered these clinal outliers (|z-scores|>2)(Figure 2a). Blocks of linkage disequilibrium (LD)(Gabriel *et al.* 2002)(Figure S5) were then inferred to identify co-localization between outlier SNPs and *cis*-eQTL. Of *cis*-eQTL that fell within the same LD block as an outlier, 17 were associated with genes that also show significant clinal patterns of gene expression (Tables 1,S4)(Figure 2). When comparing the latitudinal extremes, average estimates of F_{st} for these candidate loci were significantly higher than that of the full list of loci (Full list average F_{st}=0.10, candidate average F_{st}=0.34; Permutation test, p=0.0014). Eight of these genes were also significantly differentially expressed between the ends of the cline (Table S5). These 17 genes represent cases where *cis*-eQTL contribute to expression differences between populations and show signals of local adaptation, making them strong candidates for adaptive regulatory variation.

3.2.3. Linking adaptive regulatory variation to specific traits

The liver plays a central role in metabolic processes in the body, and regulatory changes in this tissue may contribute to latitudinal variation in traits related to metabolism. Body mass varies clinally (Figure 1a) and lab born progeny from populations at the ends of the transect also show differences in blood glucose, triglyceride, adiponectin, and leptin levels (Lynch 1992, Phifer-Rixey *et al.* 2018). Four of the 17 candidate genes identified as strong candidates also have mutant phenotypes related to body weight and metabolism. Laboratory mutants for *Cox7c*, and *Hmgb1* are associated with changes in glucose levels (Blake *et al.* 2017) and mutants for *Adam17* and *Bcat2* are also associated with changes in body mass (She *et al.* 2007, Wu *et al.* 2004, Gelling *et al.* 2008, Blake *et al.* 2017), glucose (She *et al.* 2007, Blake *et al.* 2017, Serino *et al.* 2007), leptin (She *et al.* 2007, Gelling *et al.* 2008), and adiponectin levels (Blake *et al.* 2017, Serino *et al.* 2007). Another gene identified in this analysis, *Iah1*, transcriptionally regulates genes with important roles

in lipid metabolism and triglyceride synthesis and falls under a QTL for fatty liver in mice (Kobayashi *et al.* 2016, Suzuki *et al.* 2016).

3.2.4. Adam17 and Bcat2 are candidates for adaptive differences in body mass

While knockout models can provide a link between genotypes and putative phenotypes, these models may not reflect the phenotypic consequences of mutations found in natural populations (Palopoli and Patel 1996). Changes in body weight are also among the most common effects of gene knockouts in mice, and may often reflect downstream consequences of other phenotypic changes (Reed *et al.* 2008, White *et al.* 2013). While identifying the genetic basis of complex adaptive traits is challenging, gene expression provides an intermediate phenotype that may link sequence variants to organismal traits. To connect adaptive variation in body mass in these populations to genetic variation, we asked whether body mass differences were associated with gene expression differences in the set of candidate genes (Table 1). Since latitude and body mass co-vary in this sample (Figure 1b), we controlled for latitude by regressing it out as a variable. We identified two genes, *A disintegrin and metallopeptidase domain 17*(*Adam17*)(Figure 3A-F) and *branched chain amino acid transaminase 2* (*Bcat2*), for which expression was significantly correlated with body mass, after accounting for latitude as a co-variable (*Adam17*: Pearson's correlation, $p=4.6 \times 10^{-4}$, $R^2=0.22$; *Bcat2*: $p=4.5 \times 10^{-3}$, $R^2=0.17$; see also Table S6, Figure S6). To further account for the possible confounding effects of population structure, we also looked at the correlation between expression level and body mass within each of the five populations. Replicating the pattern seen across populations, *Adam17* expression was negatively associated with body mass in 4 of the 5 populations, and *Bcat2* expression was positively associated with body mass in 4 of the 5 populations (Figure S7,S8). Despite a lack of power for within-population comparisons, the association between *Adam17* expression and body mass was significant in New Hampshire/Vermont (Pearson's correlation, $p=3.5 \times 10^{-3}$) and the association between *Bcat2* expression and body mass was significant in Pennsylvania (Pearson's correlation, $p=0.03$) and Georgia (Pearson's correlation, $p=1.8 \times 10^{-3}$).

The *cis*-eQTL for *Adam17* and *Bcat2* explain 34% and 29.7% of the variance in expression for theses genes, respectively. Genotypes at these sites were also associated with differences in body mass (Mann-Whitney *U* test, *Bcat2,* TT>CC, $p=0.024$; *Adam17*, CC>TT, $p=0.036$)(Figure S9). Again, co-variation between latitude and body mass can confound relationships between body mass and candidate genes. After regressing latitude from body mass to control for co-variation between these variables, the *Adam17 cis*-eQTL was significantly associated with body mass (Figure 3G)(Cochran-Armitage trend test, $p=0.034$), although the *Bcat2 cis*-eQTL was not (Cochran-Armitage trend test, $p=0.14$). The *Adam17* and *Bcat2 cis*-eQTLs explain an estimated 8.35% and 1.51% of the variation in body mass, respectively. These estimates should be treated as approximations since they may be influenced by (1) unmeasured environmental differences between populations, (2) population structure (even when population structure is accounted for using principle components, as was done here; see Browning and Browning 2011; Dandine-Roulland *et al.* 2011), (3) imperfect linkage disequilibrium between the surveyed SNPs and causal variants (Wray *et al.* 2013), and small sample size (Xu 2003).

Nonetheless, it is likely that the effect size for the *Adam17 cis*-eQTL is large compared to what is seen in most human GWAS for complex traits (Stranger *et al.* 2011).

Large-effect mutations may be favored in situations where populations are initially far from an optimum (Orr 1998, Dittmar *et al.* 2016). For example, variation at one gene accounts for a >2 kg weight difference between Europeans and Inuits (Fumagalli *et al.* 2015), and a single IGF1 allele in dogs accounts for 15% of variance in dog skeletal size (Sutter *et al.* 2007). House mice in this transect descended from mice in Western Europe adapted to a Mediterranean climate, and thus likely experienced strong selection pressures in a novel environment, potentially favoring some mutations of large effect.

To investigate regulatory variation at *Adam17* in Western Europe, we retrieved available liver RNA-seq and genomic data from European mice (Harr *et al.* 2016). We found that the *Adam17 cis*-eQTL is segregating within European populations (Figure S10A) and is significantly associated with liver expression in European individuals (Figure S10B, $P=3.2 \times 10^{-6}$; see Supplemental material and methods). This suggests that adaptation by the large-effect regulatory variation at *Adam17* in the United States is a product of selection on standing genetic variation.

Notably, *Adam17* and *Bcat2* are the two candidate genes from Table 1 with known lab mouse mutants that affect body mass (Wu *et al.* 2004; She *et al.* 2007; Gelling *et al.* 2008; Blake *et al.* 2017). *Bcat2* encodes a protein that catalyzes the first step of branched-chain amino acid (BCAA) metabolism, which affects metabolism and body mass in humans and rodents (Newgard *et al.* 2009). *Adam17* encodes a protein that regulates several signaling pathways. Adult *Adam17* heterozygous and null mutants show differences in metabolic phenotypes including body mass, susceptibility to diet induced obesity, and energy homeostasis (Serino *et al.* 2007; Gelling *et al.* 2008). ADAM17 and its physiological inhibitor, TIMP3, have also been reported to be involved in the glucose homeostasis and adipose, hepatic, and vascular inflammation in both genetic and nutritional models of obesity in mice (Fiorentino *et al.* 2010; Menghini *et al.* 2012; Matsui *et al.* 2014). In addition to its association with body mass and metabolism in mice, in humans variation at *ADAM17* has been linked to differences in body weight, BMI, waist circumference, and obesity risk (Junyent *et al.* 2010) and shows signatures of selection (Pickrell *et al.* 2009; Parnell *et al.* 2010; Fumagalli *et al.* 2011).

One target of ADAM17 activity is the epidermal growth factor receptor (EGFR) signaling pathway (Lee *et al.* 2003). Phenotypes observed in mice with mutant EGF receptors (including changes in body weight [Blake *et al.* 2017]) suggest that changes in EGFR signaling as a consequence of deficit ADAM17 activity may contribute to the metabolic phenotypes seen in *Adam17* mutants (Gelling *et al.* 2008). We tested for an overrepresentation of genes in the EGFR signaling pathway in the set of genes with clinal expression by annotating genes to pathways using the PANTHER database (Thomas *et al.* 2003). We saw a 1.57-fold enrichment of genes in this pathway compared to a background set of genes expressed in the liver (hypergeometric test, $p=0.018$). We also find that the gene that encodes the only known physiological inhibitor of *Adam17*, *Timp3* (Le Gall *et al.* 2010), is differentially expressed between the northern and southern populations (Figure S11, $q=0.09$) and has expression that is correlated with that of *Adam17* (Figure S11, Pearson's correlation, $p=0.02$, $R^2=0.09$). Unlike *Adam17, Timp3* expression is not associated with body mass (Pearson's correlation, $p=0.054$), although our sample size may not be sufficient to detect an association.

The data above clearly suggest that regulatory variation at *Adam17* and *Bcat2* underlies adaptive differences in body mass, but they do not identify the specific causal

mutations. To identify candidate casual mutations, we used annotations from the mouse ENCODE project (Mouse ENCODE Consortium *et al.* 2012) to search for putative regulatory elements near the *Adam17* and *Bcat2 cis*-eQTLs. The *Adam17 cis*-eQTL is in LD with SNPs through a proximal enhancer and in the *Adam17* promoter, both of which are active in the livers of adult mice. Low-coverage whole genome data show that there are variants segregating within this enhancer in these populations (Figures S12,S13)(whole genome data from [(Phifer-Rixey *et al.* 2018]). Two of the *Adam17* promoter variants are also clinal outliers (Figure S14). The *Bcat2 cis*-eQTL is within an intronic region and is not in LD with annotated regulatory elements that are active in liver tissue.

3.2.5. Expression modules are correlated with body size variation in natural populations of house mice

Next, we used a gene co-expression network approach to identify biologically related gene sets associated with phenotypic variation in these populations. Weighted Gene Co-expression Network Analysis (WGCNA) was used to identify groups of genes with highly correlated expression, called co-expression modules (Langfelder and Horvath 2008)(see Methods). Expression modules were assigned for male and female mice separately, and then male-female consensus modules were created to identify co-expression patterns shared across sexes.

Co-expression modules were then tested for correlations with measures of body size (Figures S15,S16,S17). Five expression modules in males and five expression modules in females were correlated with trait variation (Figure S18). Trait-associated modules were enriched for a number of Gene Ontology (GO) categories compared to the background set of genes expressed in the liver, including growth factor binding (q=5.3 × 10^{-8}) and lipid metabolic process (q=1.2 × 10^{-2}). None of the male-female consensus modules were significantly correlated with organismal traits, indicating that associations between co-expression modules and traits are sex-specific (Figure S17).

Focusing on the modules with the highest trait correlations (royalblue module in females, corr=0.92, p=2 × 10^{-8} and black module in males, corr=0.8, p=5 × 10^{-8}, for body mass index), we annotated genes with mutant phenotypes collected from Mouse Genome Informatics (MGI)(Blake *et al.* 2017). Supporting the association between these expression modules and phenotypic variation, we found that many of the genes with high connectivity in these modules have mutant phenotypes related to body size or metabolism (Figure 4). For example, the most connected gene in the female royalblue module is *Nr2c2*. Mutant phenotypes for *Nr2c2* include changes in eating behavior, energy homeostasis, body mass, size, and blood chemistry. Similarly, highly connected genes in the male black module (e.g., *Col3a1, Col1a1, Col1a2, Col5a2, Sparc, Bcam, Fstl1, Igfbp5, Cpe, Cav1, Lamc1, Ltbp3, Krt7*) show mutant phenotypes related to body mass and body size. Four of these genes (*Adamts2*, *Col1a1*, *Col1a2*, *Sparc*) were also identified as hub genes in the module most highly correlated with mouse body weight in another study utilizing an F2 laboratory cross (Ghazalpour *et al.* 2006).

Finally, we used the co-expression dataset to identify regulatory variation within modules associated with body size. Within the body size associated modules (Figure S18), we associated 189 genes with a *cis*-eQTL, including several highly connected genes in the sex-specific modules with the highest trait correlations (Figure 4). As in the previous analysis, we then searched for genes with a *cis*-eQTL that co-localized with a clinal

sequence variant. We identified 15 genes with clinally varying *cis*-eQTL in the body size associated modules (Table S7). We found that gene expression for 4 of these 15 genes was significantly correlated with BMI in one sex (Females: *Ube2q2*, p=0.0002; *3110082I17Rik*, p=0.0027; *Cep85*, p=0.017; Males: *Pygb* p=0.035). *Cis*-eQTL associated with these genes were not significantly associated with BMI, however, our study is also underpowered for identifying sex-specific associations. The correlation between gene expression and BMI and the presence of clinal *cis*-eQTL make these genes of interest for future study.

3.3. Discussion

Identifying loci and genes that underlie adaptive variation within and between populations is a major goal in evolutionary biology. One method used to identify such variants are genomic scans for selection. While many genomic scans attempt to link sequence variants to phenotypes through gene annotations and knockout models, most fail to connect genotypes to phenotypes in natural populations. Here, we used expression data from natural populations of house mice collected along an environmental gradient to link regulatory variation at two genes (*Adam17* and *Bcat2*) with body mass variation. We have linked these genes to body mass variation by 1) associating *cis*-eQTL with the expression of *Adam17* and *Bcat2*, 2) associating the *Adam17* and *Bcat2 cis*-eQTL with body mass variation, and 3) the associating the expression of these two genes with body mass variation. Supporting the association we see between these genes and body mass, mutant alleles for *Adam17* and *Bcat2* in laboratory mice are associated with changes in body mass and metabolism (Wu *et al.* 2004; She *et al.* 2007; Serino *et al.* 2007; Gelling *et al.* 2008; Blake *et al.* 2017). Interestingly, these two genes account for a substantial proportion of phenotypic variation in body mass among the mice studied here, with large effect sizes compared to those measured in GWAS for most complex traits. For traits under stabilizing selection within populations (as in virtually all human GWAS) effect sizes are expected to be much smaller than in comparisons between populations experiencing strong divergent selection, as is the case here. The effect size of mutations underlying traits under stabilizing selection within populations is expected to be smaller than the effect sizes of mutations in the early stages of an adaptive walk (Orr 1998; Remington 2015).

In addition to identifying regulatory variation at specific genes associated with body mass, we also used a systems biology approach to identify co-expression patterns associated with body size variation in wild mice. Gene co-expression networks capture biologically relevant relationships between genes that can be useful for understanding gene functions and interactions. Here we have used this information to characterize co-expression modules that were associated with body size and identified regulatory variation within these co-expressed gene sets that may play a role in body size variation.

The tendency for body size to increase with latitude (i.e., Bergmann's Rule) has been documented in many species, including humans (Ashton *et al.* 2000; Ruff 2002; Foster and Collard 2013), and reflects an evolved response to differences in temperature (Bergmann 1847). In humans, many candidate genes for metabolic disorders, such as obesity, also show evidence of climatic adaptation (Hancock *et al.* 2008). Intriguingly, in humans, both *ADAM17* and *BCAT2* have been implicated in metabolic disease (Arribas and Esselens 2009, Newgard *et al.* 2009; Junyent *et al.* 2010; Menghini *et al.* 2013), and variation at *Adam17* has been identified in genome scans for selection (Pickrell *et al.* 2009; Parnell *et al.* 2010;

Fumagalli *et al.* 2011) in addition to its association with body weight and obesity risk (Junyent *et al.* 2010).

Finally, this study provides evidence for the role of *cis*- regulatory variation in environmental adaptation in natural populations. While *cis*- regulatory variation has long been hypothesized to play a major role in adaptive phenotypic evolution, connecting regulatory variation with adaptive organismal phenotypes remains tricky. Combining eQTL mapping with genomic scans, as was done here, may be a fruitful approach for identifying adaptive regulatory variation in other natural systems.

3.4. Methods

3.4.1. Sampling

Mice used in this study were collected from 5 sampling locations (Table S1, File S1) along a latitudinal gradient in the eastern United States. Mice were sacrificed in the field and measurements (body weight, total body length, tail length) were taken at time of collection. Body mass index (BMI) was calculated as body weight/length^2 (g/mm^2). Liver tissue was collected in RNAlater and stored at 4^{o}C overnight and then frozen to -80^{o}C until RNA extraction with the Qiagen's RNeasy Mini Kit.

3.4.2. mRNA-sequencing and mapping

For each sample, 100 base-pair paired-end reads were sequenced on the Illumina HiSeq 4000 platform. RNA-seq reads were mapped with TopHat2 (Kim *et al.* 2013) to personal reference genomes, created by inserting variants into the mouse reference (GRCm38) and masking indels (see Supplemental material and methods). We removed genes with fewer than 500 reads across samples (i.e., an average of 10 reads per sample). Gene expression was then quantile normalized and corrected for hidden factors and known co-variates (individual sex and the first 6 principle components from genotype data to account for population structure) using a Bayesian approach (Stegle *et al.* 2010, Stegle *et al.* 2012)(Figure S19,S20).

3.4.3. Exome capture sequencing and identification of clinal outliers

The exome-sequence data was used to identify clinal outliers (Phifer-Rixey *et al.* 2018)(see also Supplemental material and methods). Libraries were enriched for exonic target regions and subsequently 100-bp paired end reads were sequenced on the Illumina HiSeq 2000 platform, resulting in 2 GB of raw sequence data per individual. Forty-one of the 50 individuals for which there is exome- sequence data have matched RNA-seq libraries (see Table S1). Reads were mapped with Bowtie 2 (Langmead and Salzberg 2012) and allele frequencies were estimated with ANGSD (Korneliussen *et al.* 2014). LFMM (Frichot *et al.* 2013) was used to identify covariance between environmental and genetic variation (see Supplemental material and methods).

3.4.4. cis-eQTL discovery

We performed *cis*-eQTL mapping using variant calls from RNA-seq and exome data (see Supplemental material and methods). One limitation of this method is that the genotyping dataset is limited to sites represented by these data (i.e., variant calls are largely limited to exomic regions of the genome). Consequently, many causal sites may not be typed and variants associated with expression may be tagging causal sites in LD. For the

exome dataset, depth per site of the targeted exome was ~15×. For genes represented in the analysis, on average per individual we had sufficient coverage for ~32% of bases within gene boundaries and ~15% of bases in the 200-kb boundary used as the cut-off for *cis*-eQTL mapping (Table S8).

To identify *cis*-acting eQTLs, we used a linear mixed model applied in the program GEMMA (Zhou and Stephens 2012) on expression residuals to associate expression with sequence variants (see Supplemental material and methods). A relatedness matrix was computed and included as a covariate. We retained the variant with the lowest *p*-value for each gene and then performed a Bonferroni's correction. Variants with Bonferroni corrected *p*-values of < 0.05 were considered significant.

3.4.5. Weighted gene co-expression analysis

We carried out a weighted gene co-expression network analysis (WGCNA) on expression residuals following WGCNA protocols (Langfelder and Horvath 2008) to create expression modules. Each module is summarized by a representative eigengene, the first principle component of a given module. Each gene's expression was correlated with the module eigengene as a measure of the gene's centrality to the module, called module membership.

3.5. Data access

Illumina sequencing data from this project has been submitted to the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject) under the accession number PRJNA407812.

3.6. Chapter 3 Table

Table 1. *cis*-eQTL that co-localize or are within the same LD block as a clinal outlier that also show expression changes correlated with latitude.

Symbol	Expression correlation co-efficient with latitude	*p*-value	Phenotypes[1]
Tcea1	0.6	3.66E-06	cardiovascular, embryo, growth/size/body, hematopoietic, homeostasis, limbs/digits/tail, liver/biliary, mortality/aging
Iah1	-0.43	0.0018	cardiovascular, limbs/digits/tail, skeleton
Lnx1	-0.41	0.0035	hematopoietic, immune
2810402E24Rik	0.38	0.0073	
Arl6ip4	0.36	0.0096	
Nsa2	-0.36	0.011	
Rpl3	0.35	0.014	
Bcat2	0.34	0.016	adipose, behavior, growth/size/body, homeostasis, renal/urinary
1810024B03Rik	-0.32	0.023	
Rplp0	0.32	0.023	hematopoietic, immune
Rpap2	-0.32	0.023	
F11	0.31	0.027	hematopoietic, homeostasis, nervous system
Hmgb1	0.31	0.031	endocrine/exocrine, homeostasis, immune, cellular, hematopoietic, mortality/aging, behavior, growth/size/body, mortality/aging, respiratory, vision/eye
Adam17	-0.3	0.032	cardiovascular, cellular, digestive/alimentary, embryo, growth/size/body, hematopoietic, homeostasis, immune, integument, mortality/aging, muscle, nervous system, pigmentation, respiratory, vision/eye
Cox7c	-0.3	0.035	homeostasis, mortality/aging
Ccdc137	0.29	0.041	
Nsfl1c	0.28	0.0496	

[1]Abnormal phenotypes in targeted gene mutants, collected from Mouse Genome Informatics database (MGI)

3.7. Chapter 3 Figures

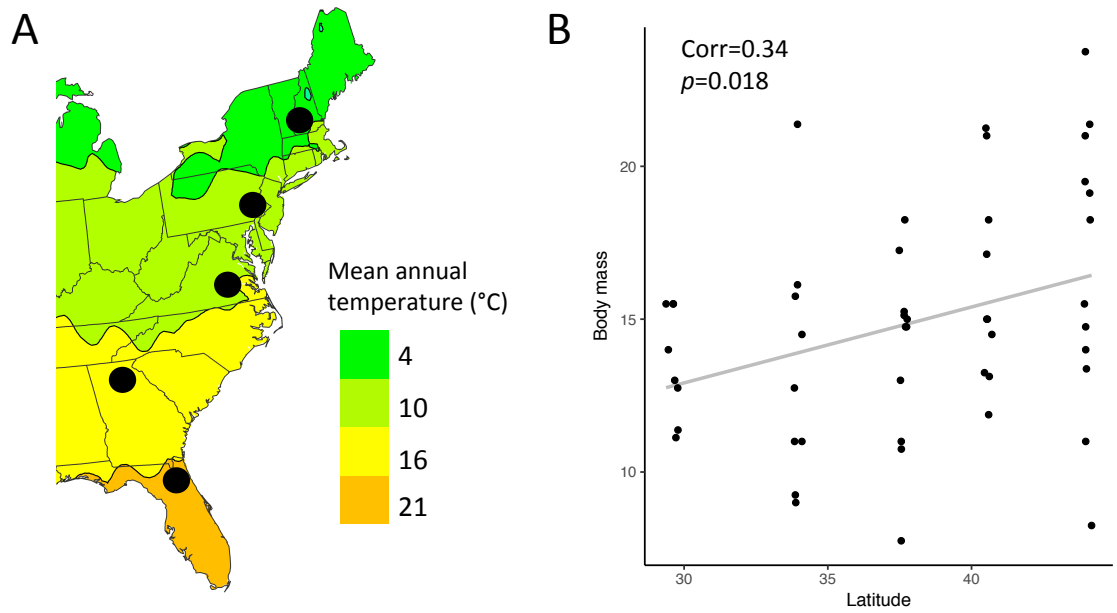


Figure 1. A. Sampling locations along the east coast of North America (climate map obtained from NOAA, National Weather Service). B. Consistent with Bergmann's Rule, body mass in mice increases with increasing latitude (Pearson's correlation=0.34, p=0.018)(see Table S9)(See Supplemental material and methods).

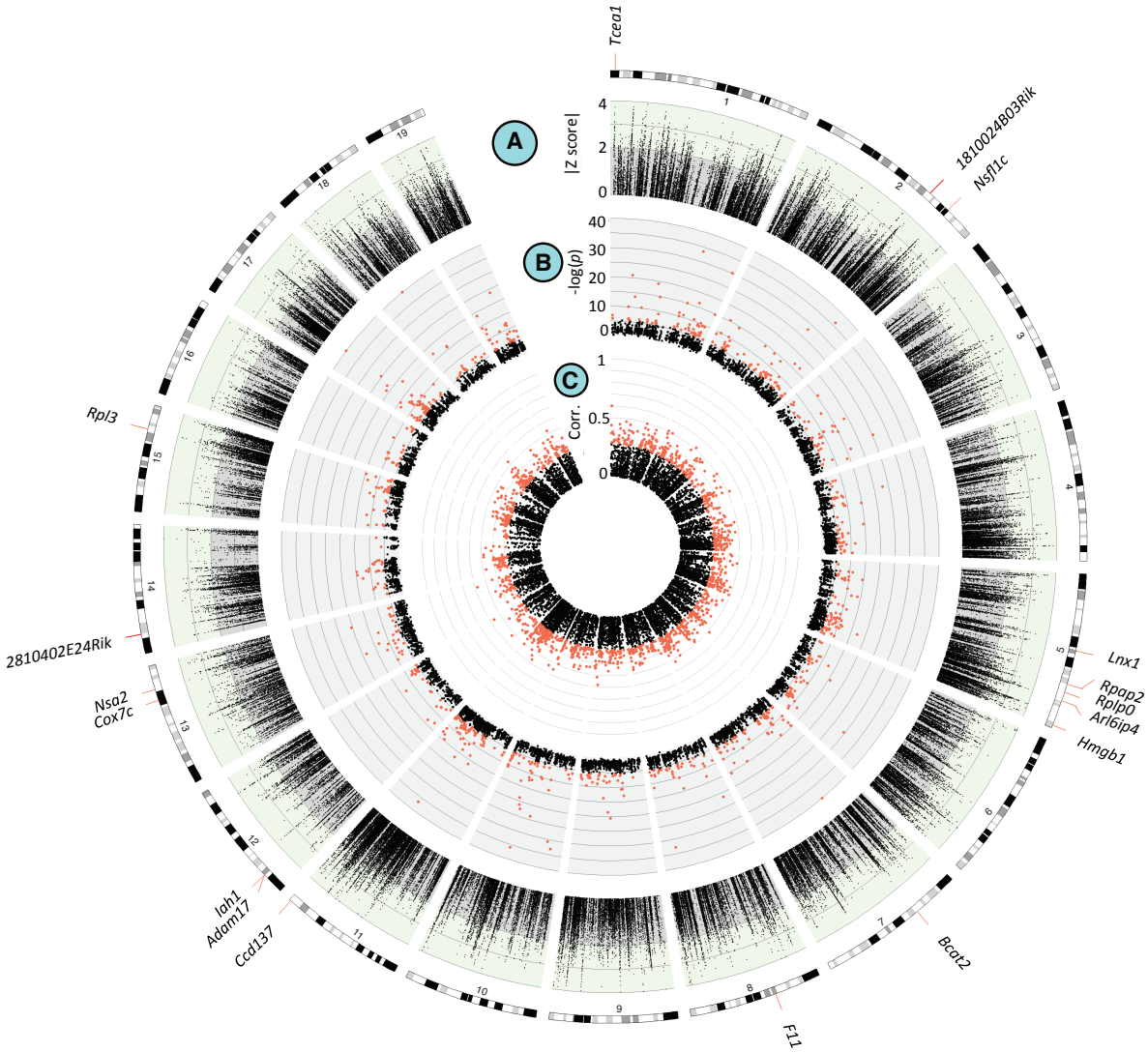


Figure 2. Overlap between genomic scans identifies regulatory variants that are candidates for clinal adaptation. A. The LFMM |z-scores| for each SNP vs. chromosome position. SNPs with |z-scores|>2 were considered clinal outliers. B. Manhattan plot of *cis*-eQTL. Shown in red are significant SNPs. C. Manhattan plot of gene starting position vs. the correlation between gene expression and latitude. Points labeled in orange are genes for which expression is significantly correlated with latitude (p<0.05). On the outside are ideograms with the location of genes for which these three signals (A,B,C) overlap. Figure created with Circos (Krzywinski *et al.* 2009).

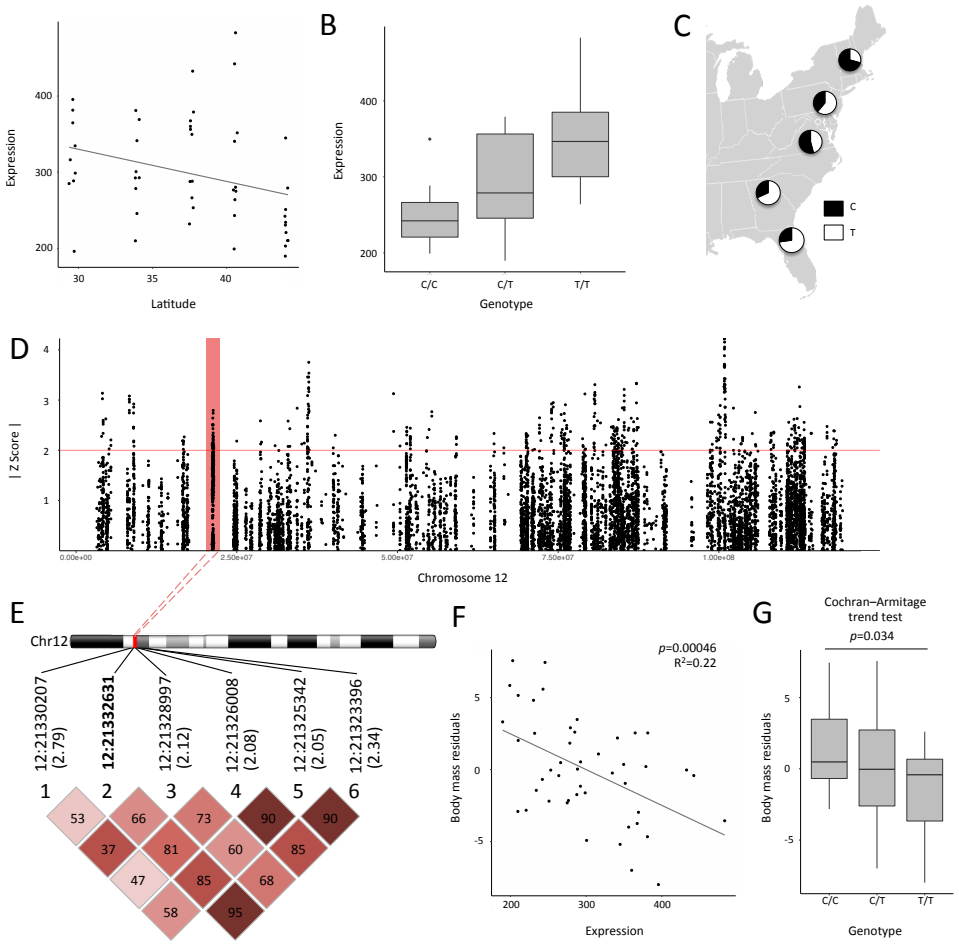


Figure 3. *Adam17* is a candidate for adaptive differences in body mass among mice in eastern North America. A. Expression of *Adam17* is correlated with latitude (p=0.032, Pearson's correlation=-0.30). Sex was not a significant predictor of *Adam17* expression. B. A SNP at Chr12:21332631 was identified as a *cis*-eQTL for *Adam17.* C. Allele frequencies of Chr12:21332631 in five populations. D. The LFMM |z-scores| for sites on Chromosome 12 versus position. Points above the red line were considered clinal outliers in this study. The red box represents the peak in which Chr12:21332631 is found. E. Nearby outlier SNPs in LD with Chr12:21332631. Correlations (r^2, %) are given in each block. The z-scores for each site's association with latitude are given in parentheses. F. *Adam17* expression is significantly associated with body mass when controlling for latitude (Pearson's correlation, $p=4.6 \times 10^{-4}$, R^2=0.22). G. Genotype at Chr12:21332631, the *cis*-eQTL for *Adam17*, significantly trends with body size when latitude is controlled for (Cochran-Armitage trend test, p=0.034).

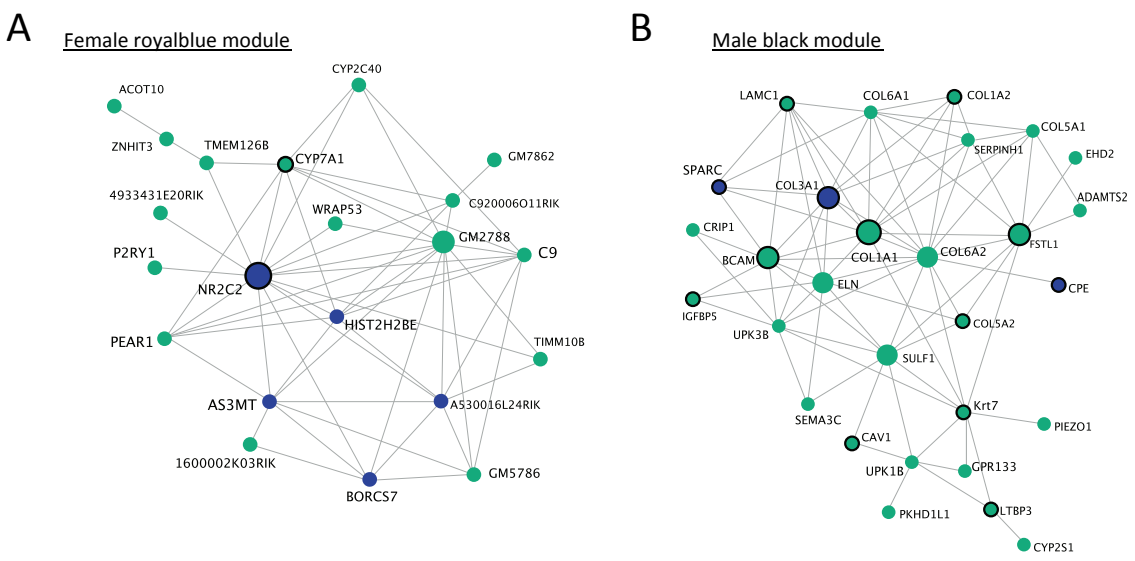


Figure 4. Visualization of the most connected genes in the female "royalblue" (A) and the male "black" with co-expression modules with VisANT (Hu *et al.* 2008) (B). The royalblue module is associated with BMI ($p=2 \times 10^{-8}$) and body length variation ($p=6 \times 10^{-6}$). The black module is associated with BMI ($p=5 \times 10^{-8}$), body mass ($p=0.001$), and body length variation ($p=3 \times 10^{-10}$). Blue circles represent genes for which we identified a *cis*-eQTL that explains a component of expression variation. Circles with black borders are genes with mutant phenotypes related to body size or metabolism. Phenotype information was collected from MGI (Blake at al. 2017).

3.8. Chapter 3 Supplemental material

3.8.1 Sampling

Mice were collected from 5 sampling localities in the eastern United States (see File S1 for individual localities). Museum specimens were prepared (skin and skull) and have been deposited in the U.C. Berkeley Museum of Vertebrate Zoology (Phifer-Rixey *et al.* 2018)(see File S1). At least 10 individuals were collected at each location. Individuals were collected at a minimum of 500 meters apart from one another to avoid collecting closely related individuals. Sex and reproductive status were assessed and recorded at the time of collection. Body weight, total body length, tail length, hindfoot length, and ear length were measured and recorded for each individual. Animals were sacrificed in accordance with IACUC protocols and tissues (liver, kidney, heart and spleen) were collected and stored at -80°C. Liver tissue was also collected in RNAlater and stored at 4°C overnight and then frozen to -80°C until subsequent RNA extraction.

3.8.2 mRNA-sequencing and library preparation

RNA was extracted with Qiagen's RNeasy Mini Kit, and the extracted RNA from each sample was quantified with a Qubit spectrophotometer. RNA libraries were built with KAPA's stranded mRNA-Seq kit and were sequenced at UC Berkeley's Vincent J. Coates Genomics Sequencing Laboratory. For each sample, 100 base-pair paired-end reads were sequenced on the Illumina HiSeq4000 platform across 3 lanes.

3.8.3 Quantifying gene expression

RNA-seq reads were trimmed with Trimmomatic (Bolger *et al.* 2014) and mapped with TopHat2 (Kim *et al.* 2013) to personal reference genomes (see "*Variant calling on exome and RNA-seq data"* for methods of constructing personalized references), allowing two mismatches. Mapping bias toward the reference allele can reduce the accuracy of allele-specific expression measurements. We attempted to mitigate the effects of mapping bias by (1) allowing mismatches during mapping, which has been shown to reduce reference bias (Stevenson *et al.* 2013), and (2) creating and mapping RNA-seq reads to personal reference genomes for each individual. While only heterozygous sites were tested for allele-specific expression, indels were masked in our personal reference genomes, as these have been shown to cause biased allele-specific assignment (Stevenson *et al.* 2013). We found that creating personalized reference genome also increased the number of reads mapped overall. Additionally, when looking at sites for which we find significant allele-specific expression, we found no bias in the number of reads mapped towards the reference for these sites.

We achieved an average of 1.7G mapped reads per sample. Reads that did not map uniquely were discarded. We used HTseq-count (Anders *et al.* 2015) to count reads overlapping exons to estimate mRNA abundance per gene. Read counts were subsequently quantile normalized to account for differences in sequencing depth between libraries. Information on the read depth per sample is available in File S1.

3.8.4 Exome capture sequencing and identification of clinal outliers

DNA was extracted from tissues from 50 individuals using the Qiagen Gentra Puregene Kit. Forty-one of these individuals have matched RNA-seq libraries (see Table S1). Libraries were prepared with unique barcodes for each individual and a NimbleGen in-solution capture kit (SeqCap EZ Mouse) was used to enrich libraries for exonic target regions. Individuals were pooled in groups of 16-17 and each 100-bp paired-end reads were sequenced for each pool on one lane of Illumina HiSeq2000. Sequence data were cleaned to remove adaptor sequences, filter out low-complexity reads, bacterial contamination, and PCR duplicates. Overlapping reads were merged. Reads were subsequently mapped with Bowtie 2 (Langmead and Salzberg 2012) to the C57BL/6J (GRCm38) mouse reference genome. Average coverage per site was approximately ~15×.

Allele frequencies at variable sites were estimated with the program ANGSD (Korneliussen *et al.* 2014). Individual sites were filtered based on (1) the posterior probability of an individual's genotype (>=0.50), (2) the *p*-value for the likelihood ratio test for the SNP being variable (<=0.001), and (3) minor allele frequency (>5%), resulting in a total of 281,361 sites that could be tested for clinal associations.

Clinal outliers were identified using the program LFMM (Frichot *et al.* 2013) using latitude as an environmental correlate. LFMM implements a Bayesian PCA to simultaneously infer background population structure and identify covariance between environmental and genetic variation.

Fifty runs of LFMM with 2 latent factors (-K 2) and 50,000 burnin cycles in the Gibbs sampler algorithm (-b 50000) were performed to obtain z-scores for each SNP. The median z-score for a variant was taken across all runs.

While isolation by distance can alternatively explain clinal patterns of variation, there is no evidence for isolation by distance for these populations. Principle component analyses on genotypes (Figure S19) and on expression (Figure S20) also do not cluster individuals based on latitude. Thus, clinal patterns are not a consequence of isolation by distance.

3.8.5 Variant calling on exome and RNA-seq data

Variant calling was performed with Genome Analysis Toolkit v3.6 (GATK) HaplotypeCaller (McKenna *et al.* 2010). After marking duplicate reads, we performed the Base Quality Recalibration using high quality variant calls for *Mus musculus* downloaded from the Wellcome Trust (ftp://ftp-mouse. sanger.ac.uk/current_snps). We then performed joint genotyping on exome data with HaplotypeCaller followed by Variant Quality Score Recalibration using the Wellcome Trust variant calls and a recently published variant call set from natural populations of *Mus musculus* (Harr *et al.* 2016). As joint genotyping is not recommended for RNA-seq reads, we performed variant calling separately on each of these samples with HaplotypeCaller. We then filtered variants based on variant confidence (if QD < 2.0), strand bias (if FS > 30.0), mapping quality, (if MQRankSum < -12.5 and if MQ < 35.0), and bias in read position (ReadPosRankSum < -8.0). Additionally, we excluded sites where fewer than 5 reads supported the genotype call. These genotype and indel calls were used to create personal reference genomes for each sample. Variant calls were inserted into the mouse reference (GRCm38) and indels were masked using bedtools (Quinlan and Hall 2010).

3.8.6 cis-eQTL discovery with GEMMA

To control for population structure, hidden factors, and sex, expression levels were corrected with the program PEER (Stegle *et al.* 2010; Stegle *et al.* 2012), which uses a Baysian approach to infer determinants using a factor analysis method. Population structure was accounted for by including the first 6 principle components from the genotype data, which accounted for a combined 24.18% of variation. Principal component analysis was performed using the package SNPRelate (Zheng *et al.* 2012).

The program GEMMA (Zhou and Stephens 2012) was used to identify putative *cis*-regulatory variation for autosomal genes. Sex chromosomes were excluded from the analysis. Linear mixed model approaches have demonstrated success in controlling for relatedness among samples and controlling for population stratification (e.g., Kang *et al.* 2008; Listgarten *et al.* 2010; Price *et al.* 2010; Zhang *et al.* 2010). While mice were collected for this study in a way to minimize the sampling of closely related individuals, individuals show different levels of relatedness. To account for this, a centered relatedness matrix was computed by GEMMA based on the input genotypes and included as a covariant.

GEMMA fits a linear mixed model in the following form:

$$y = W\alpha + x\beta + u + \epsilon;\ \ u \sim MVN_n(0, \lambda\tau^{-1}K), \epsilon \sim MVN_n(0, \tau^{-1}I_n)$$

where y represents a n-vector of qualitative traits for n individuals, W is a $n \times c$ matrix of covariates, α is a c-vector of the corresponding coefficients including the intercept, x is an n-vector of genotypes, β is the effect size, u is a vector of random effects, ϵ is an n-vector of errors and τ^{-1} is the variance of residual errors, λ represents is the ratio between the two variance components, K represents the $n \times n$ relatedness matrix, I_n is a n × n identity matrix, and finally MVN is multivariate normal distribution. In this case, y is an n by 1 vector of gene expression residuals for n individuals, x is the n by 1 vector of genotypes, and u is an n by 1 vector to control for relatedness and population structure, and ϵ represents residual errors as an n × 1 vector. We test the alternative hypothesis $H_1 : \beta \neq 0$ against $H_0 : \beta = 0$ for each variant within 200-kb of the gene of interest. The 200-kb distance was based on thresholds used in other studies to identify *cis*-eQTL (Pickrell *et al.* 2010, Sun 2012, Sun and Wu 2013; Tung *et al.* 2015). Sites with 1) a minor allele frequencies less than 0.01, 2) Hardy-Weinberg p-values below 0.001, or 3) Missing genotype calls for 10 or more individuals, were excluded from the *cis*-eQTL analysis.

Missing genotypes were imputed with BEAGLE (Browning and Browning 2009). The 9 individuals for which we had no exome data and only RNA-seq reads show a high rate of missing genotypes compared to other samples because of reduced coverage in some regions. Consequently, when missing genotypes were imputed using BEAGLE for these samples, we found that they clustered together in a group in a PCA (where these samples cluster with their source population when genotypes are not imputed; Figure S19). As a result, we only performed imputation for the 41 individuals for which we had exome-sequence in the *cis*-eQTL mapping analysis, and not for the individuals without matched exome libraries, and then subsequently mapped *cis*-eQTL using this set of variants.

3.8.7 Identifying allele-specific expression

To identify allele-specific expression, we focused on exonic heterozygous sites where both the reference and alternative allele were supported by more than 10 reads. To test for allele-specific expression, we considered a beta-binomial distribution to model allelic counts. We estimate dispersion by setting the likelihood at p=0.5 (assuming no allelic imbalance) and varying the dispersion parameter from 0 and 1 to minimize the likelihood function. This dispersion parameter was then used to estimate the beta binomial maximum likelihood. For each gene in a given individual, we took the site with the lowest p-value and performed a false discovery rate correction.

3.8.8. Associations between body mass and candidate genes

All statistical analyses of body mass associations were performed in R (v3.3.2) using individuals of both sexes. To account for co-variation between latitude and body mass, body mass was first adjusted for latitude with linear regression and subsequently tested for correlations with the expression of genes in Table 1. Sex is not a significant predictor for body mass in these data (Table S9) and the relationship between *Adam17* and *Bcat2* expression and body mass was significant whether or not individuals were excluded based on reproductive status (see Table S6). Associations between *Bcat2* and *Adam17* expression and body mass without adjustments for latitude are available in Figure S6. SNP contributions to the phenotypic variance were estimated using an ANOVA model after adjusting body mass for latitude, the first eigenvector of SNPs to control for population structure, their interaction, and sex using a linear regression.

3.8.9. Differential expression

We used the R package DESeq2 (Love *et al.* 2014) to identify differential expression between populations at the latitudinal extremes. We tested 13,635 genes for differential expression between 8 individuals from Florida and 12 individuals from New Hampshire and Vermont. We used DESeq2 to normalize raw gene read counts, estimate dispersion factors for each gene, and then test for differential expression based on a negative binomial distribution (Figure S4). The resulting p-values were then false discovery rate corrected.

3.8.10. Characterizing linkage disequilibrium

To characterize linkage disequilibrium (LD) in this dataset, we used PLINK v1.9 (Chang *et al.* 2015) to calculate squared inter-variant allele count correlations (r^2), normalized measure of allelic association (D'), and to create LD blocks (following the definition of haplotype block from [Gabriel *et al.* 2002]). We restricted the LD analysis to SNPs with a minor allele frequency greater than 0.01 (see Figure S5).

3.8.11. Weighted gene co-expression analysis

We performed weighted gene co-expression analysis (WGCNA) following WGCNA protocols (Langfelder and Horvath 2008) on non-pregnant females (n=19) and males (n=21). One male outlier was filtered based on sample clustering. Pearson's correlations for all gene pairs across samples were calculated to create similarity matrices. Soft thresholding power was selected based on scale-free topology. The minimum module size was set to 30 genes. We assigned 13,636 genes to 47 expression modules in male mice and 13,635 genes to 40 expression modules in female mice. Male-female consensus

modules were also created to identify co-expression patterns shared across sexes. Across males and females, we identified 44 expression modules comprising 9,359 genes. We tested for associations between a summary profile (called an eigengene) and external traits (BMI, tail length, body mass, and latitude) to identify modules associated with these external traits (Figure S15,S16,S17).

3.8.12. Replication of the Adam17 cis-eQTL in European populations

SNP calls and liver RNA-seq reads (mapped to mm10) from Harr *et al.* 2016 were retrieved (http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/) for mice from Germany (n=8) and France (n=8) (Harr *et al.* 2016). To quantify gene expression, we used HTseq-count (Anders *et al.* 2015) to count reads overlapping exons. Reads overlapping exonic regions were subsequently quantile normalized. The SNP identified as a *cis*-eQTL for *Adam17* (Chr12:21332631) in Eastern North America was found to be significantly correlated with *Adam17* expression in European mice. We were unable to validate the *Bcat2* cis-eQTL in European individuals as 11/16 individuals were fixed for the reference allele.

3.9. Chapter 3 Supplemental Tables

Table S1. Summary of samples used in this study (further metadata is available in File S1)

RNA-seq library ID	Exome available	Exome name	Population	Sex
MNKM01-19	x	MWNMVP001_108B	FL	F
MNKM01-40	x	MWNMVP001_115B	FL	F
MNKM01-38	x	MWNMVP003_113B	FL	F
MNKM01-39	x	MWNMVP003_114B	FL	F
MNKM01-37	o	N/A	FL	F
MNKM01-41	o	N/A	FL	F
MNKM01-20	x	MWNMVP001_110B	FL	M
MNKM01-21	x	MWNMVP003_112B	FL	M
MNKM01-22	x	MWNMVP001_124B	GA	F
MNKM01-26	x	MWNMVP001_131B	GA	F
MNKM01-44	x	MWNMVP001_132B	GA	F
MNKM01-45	x	MWNMVP001_133B	GA	F
MNKM01-1	x	MWNMVP002_130B	GA	F
MNKM01-24	o	N/A	GA	F
MNKM01-42	x	MWNMVP002_128B	GA	M
MNKM01-23	x	MWNMVP003_125B	GA	M
MNKM01-25	x	MWNMVP003_129B	GA	M
MNKM01-35	x	MWNMVP002_137B	NH	F
MNKM01-46	x	MWNMVP003_134B	NH	F
MNKM01-47	x	MWNMVP003_135B	NH	F
MNKM01-27	o	N/A	NH	F
MNKM01-28	x	MWNMVP001_138B	NH	M
MNKM01-29	x	MWNMVP003_140B	NH	M
MNKM01-2	o	N/A	NH	M
MNKM01-5	x	MWNMVP001_150B	PA	F
MNKM01-34	x	MWNMVP001_156B	PA	F
MNKM01-32	x	MWNMVP002_153B	PA	F
MNKM01-30	x	MWNMVP003_146B	PA	F
MNKM01-43	o	N/A	PA	F
MNKM01-6	x	MWNMVP001_148B	PA	M
MNKM01-31	x	MWNMVP001_151B	PA	M
MNKM01-7	x	MWNMVP001_152B	PA	M
MNKM01-33	x	MWNMVP002_155B	PA	M
MNKM01-8	o	N/A	PA	M
MNKM01-50	x	MWNMVP002_159B	VA	F
MNKM01-11	x	MWNMVP002_161B	VA	F
MNKM01-16	x	MWNMVP002_167B	VA	F
MNKM01-18	x	MWNMVP002_169B	VA	F
MNKM01-10	o	N/A	VA	F
MNKM01-12	x	MWNMVP002_162B	VA	M

MNKM01-13	x	MWNMVP002_163B	VA	M
MNKM01-14	x	MWNMVP002_164B	VA	M
MNKM01-15	x	MWNMVP002_165B	VA	M
MNKM01-17	x	MWNMVP002_166B	VA	M
MNKM01-9	o	N/A	VA	M
MNKM01-49	x	MWNMVP003_142B	VT	F
MNKM01-4	x	MWNMVP003_145B	VT	F
MNKM01-36	x	MWNMVP001_143B	VT	M
MNKM01-3	x	MWNMVP001_144B	VT	M
MNKM01-48	x	MWNMVP003_141B	VT	M

Table S2. Testing for allele-specific expression in wild mice

	Number
Heterozygous sites[1]	28,234
Genes tested	6,738
Genes with ASE	442
Genes with ASE in >1 individual	258
Genes with ASE and a *cis*-eQTL	40

[1]Heterozygous sites within exons with sufficient read depth to test for ASE
[2]q-value < 0.05

Table S3. Differential expression between latitudinal extremes (FL vs. NH/VT)

	Number of genes
Genes with differential expression	458
Genes with differential expression and a *cis*-eQTL[1]	48
Genes with differential expression and allele-specific expression[1]	14

[1]*q*-value < 0.05

Table S4. Candidate gene *cis*-eQTL (Table 1) states and allele frequencies

Gene	Reference[1]	Alternative	Ancestral[2]	Major	Minor	Minor allele frequency
Tcea1	T	A	N/A	A	T	0.49
Iah1	T	C	C	C	T	0.48
Lnx1	T	C	C	T	C	0.31
2810402E24Rik	A	C	C	A	C	0.26
Arl6ip4	G	A	G	G	A	0.32
Nsa2	T	C	T	C	T	0.38
Rpl3	G	G	A	G	A	0.26
Bcat2	T	C	C	C	T	0.17
1810024B03Rik	A	G	N/A	A	G	0.49
Rplp0	A	G	N/A	A	G	0.42
Rpap2	T	C	T	T	C	0.43
F11	A	G	G	G	A	0.45
Hmgb1	G	A	N/A	G	A	0.33
Adam17	T	C	C	T	C	0.49
Cox7c	G	A	G	G	A	0.28
Ccdc137	T	C	C	T	C	0.43
Nsfl1c	A	G	G	A	G	0.31

[1]Reference allele based on mm10

[2]Ancestral allele based on *Mus spretus* (SPRET/EiJ) alignment

Table S5. *cis*-eQTL that within an LD block of a clinal outlier that are also differentially expressed between Florida and New York

Symbol	Log_2 Fold Change	*q*-value	Phenotypes
Hmgb1	0.65	0.0026	behavior, cellular, growth/size/body, homeostasis, mortality/aging, respiratory, vision/eye, endocrine/exocrine
Lnx1	-1.39	0.0037	hematopoietic, immune
Ccdc137	0.73	0.0097	hematopoietic, immune
Bcat2	0.69	0.015	adipose, behavior, growth/size/body, homeostasis, renal/urinary
2810402E24Rik	0.91	0.021	
Cox19	-0.76	0.048	
AA465934	-0.92	0.048	
F11	0.62	0.0084	hematopoietic, homeostasis, nervous system
Iah1	-0.76	0.062	cardiovascular, limbs/digits/tail, skeleton
Oasl1	-0.95	0.032	homeostasis, immune, mortality/aging
Tcea1	0.63	0.095	cardiovascular, embryo, growth/size/body, hematopoietic, homeostasis, limbs/digits/tail, liver/biliary, mortality/aging

Table S6. Association between *Adam17* and *Bcat2* expression and body mass ware significant whether or not females assessed as pregnant were excluded.

	Adam17 expression vs. body mass residuals	*Bcat2* expression vs. body mass residuals
Sample	*p*-value	*p*-value
All Individuals	**4.6E-04**	**0.0041**
Non-pregnant individuals	**8.3E-04**	**0.041**

Table S7. Genes within body size associated co-expression modules that are associated with a cis-eQTL that co-localizes with a LFMM outlier.

Gene name	| Z Score |[1]
Females	
Ube2q2	3.04
3110082I17Rik	2.28
Cep85	3.01
Bcat2	2.19
F830016B08Rik	2.02
Rpl3	2.98
Iah1	2.13
Dhdh	2.47
Cib1	2.15
2810402E24Rik	2.03
Zc3h6	2.07
Males	
Pygb	2.42
Ccdc137	2.93
Nsun3	2.01
Pgghg	2.15

[1]LFMM |Z Score|

Table S8. The average proportion of bases with coverage within an individual

	Average proportion of bases with coverage in one individual
Genes surveyed	0.32
200-kb surrounding genes surveyed	0.15

Table S9. Analysis of body size. A linear model was fit in the following form: Body mass (g) ~ Sex + Latitude (N=50).

Response	Predictor	P value
Body Mass	Sex	0.46
	Latitude	**0.018**

3.10. Chapter 3 supplemental Figures

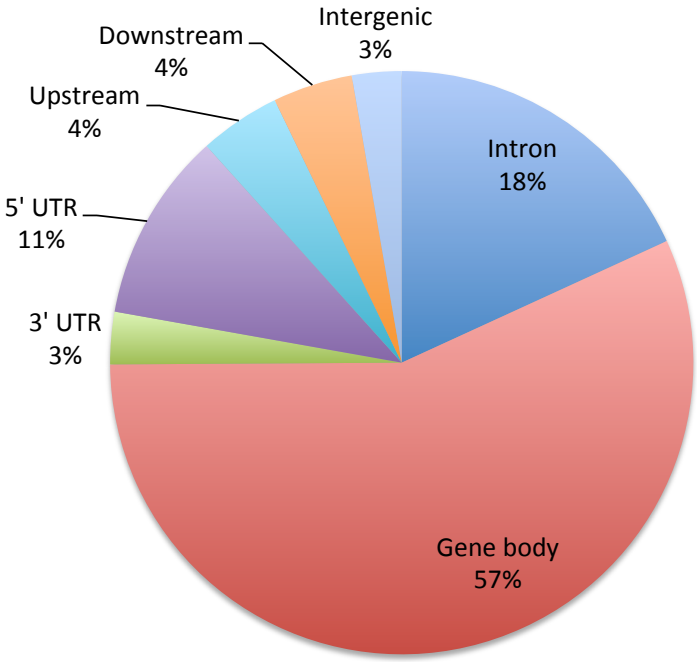


Figure S1. The locations of SNPs identified as *cis*-eQTL. RefSeq exon annotations were used to annotate the locations of cis-eQTL. Some genes are annotated to multiple elements (e.g., within an exon and intron) and are represented in multiple categories.

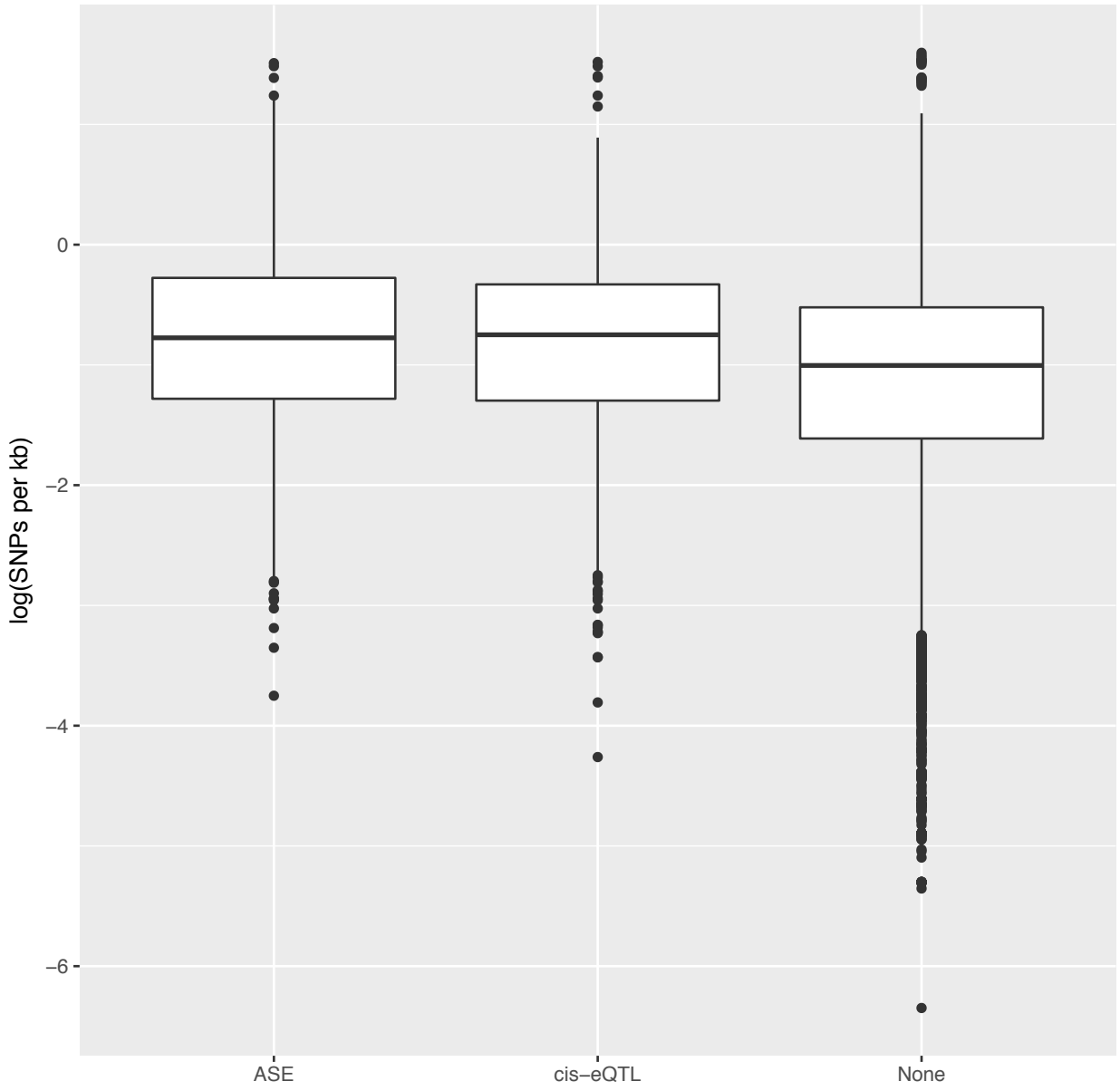


Figure S2. Allele-specific expression and *cis*-eQTL detection is more likely in genes when SNP density is higher (Wilcoxon test, p=3.1e-11 and p< 2.2e-16, respectively).

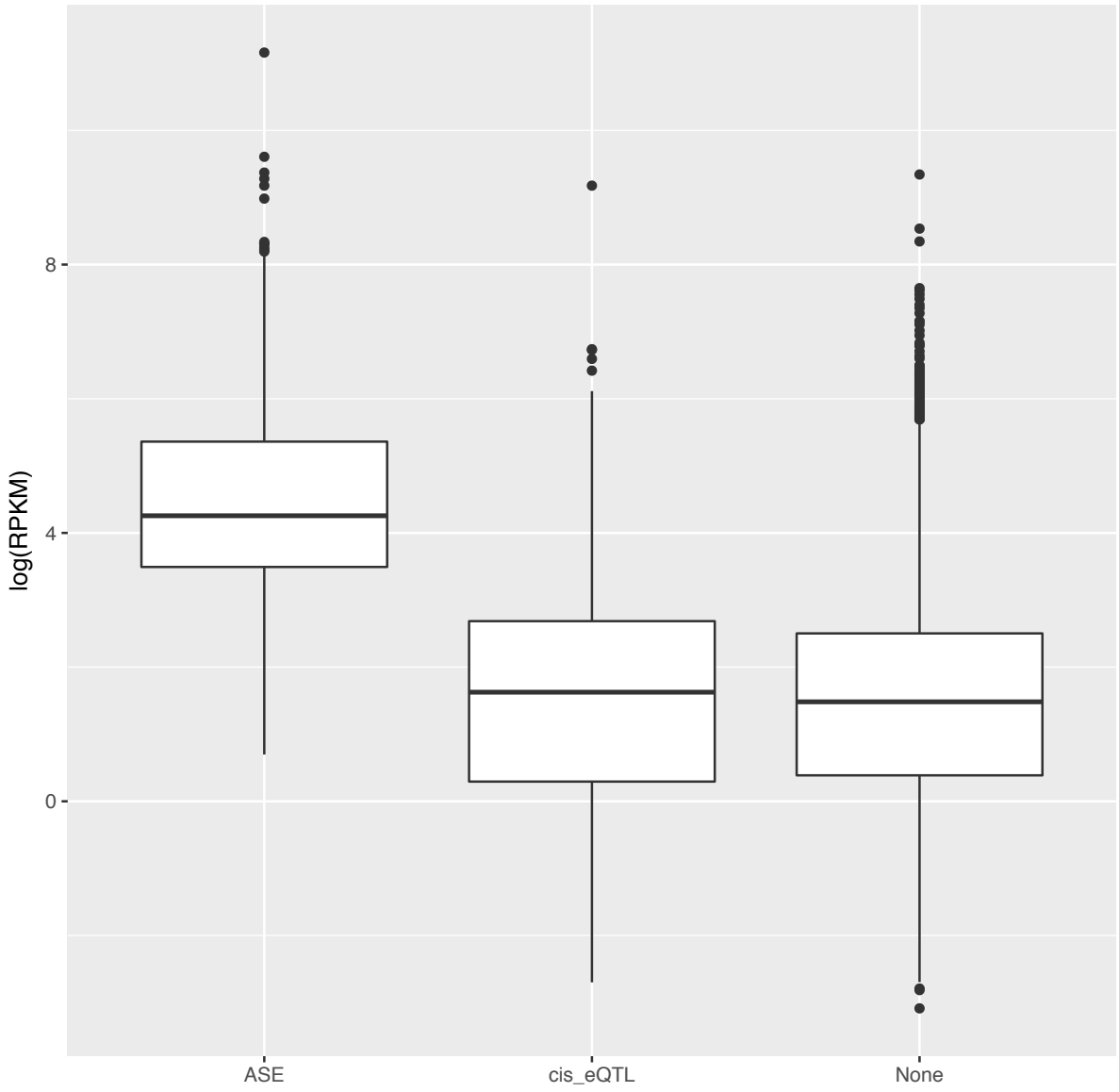


Figure S3. Allele-specific expression detection is more likely in genes with higher expression (Wilcoxon test, *p*-value < 2.2e-16).

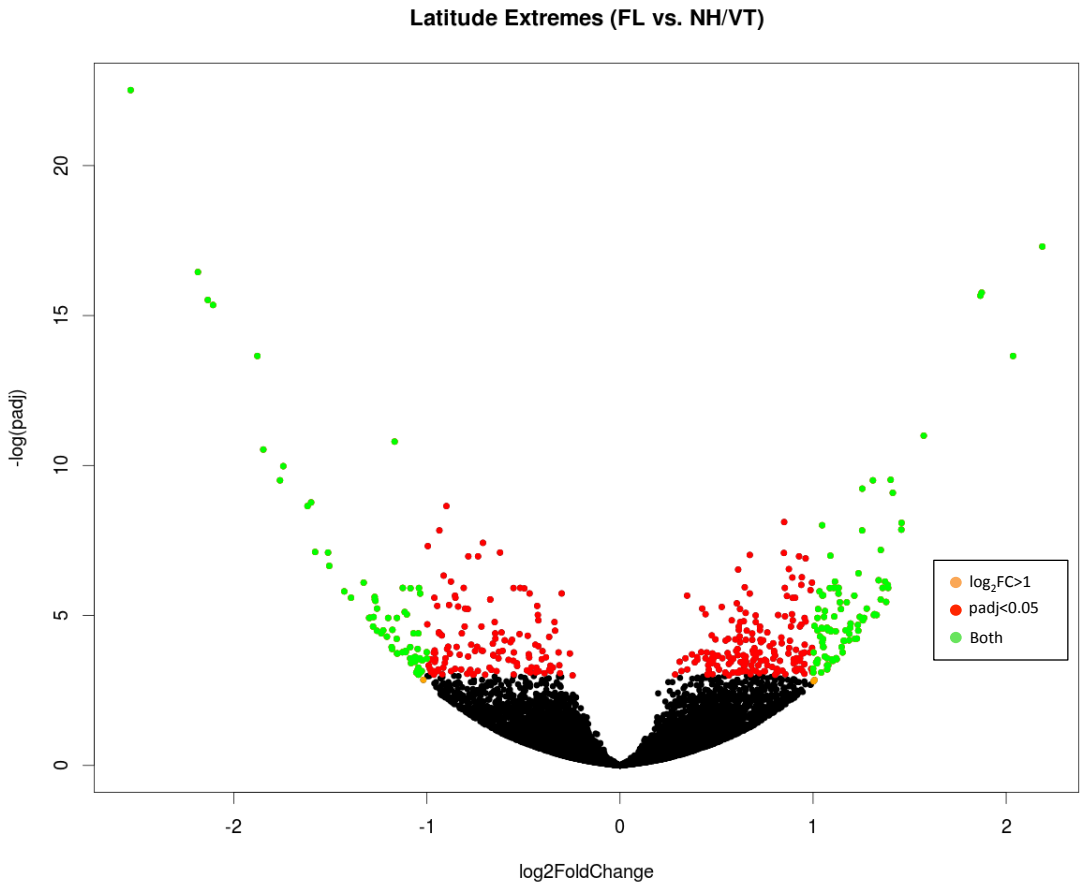


Figure S4. Differential expression between latitudinal extremes. We compared gene expression between wild collected mice in Florida and mice collected from New Hampshire and Vermont.

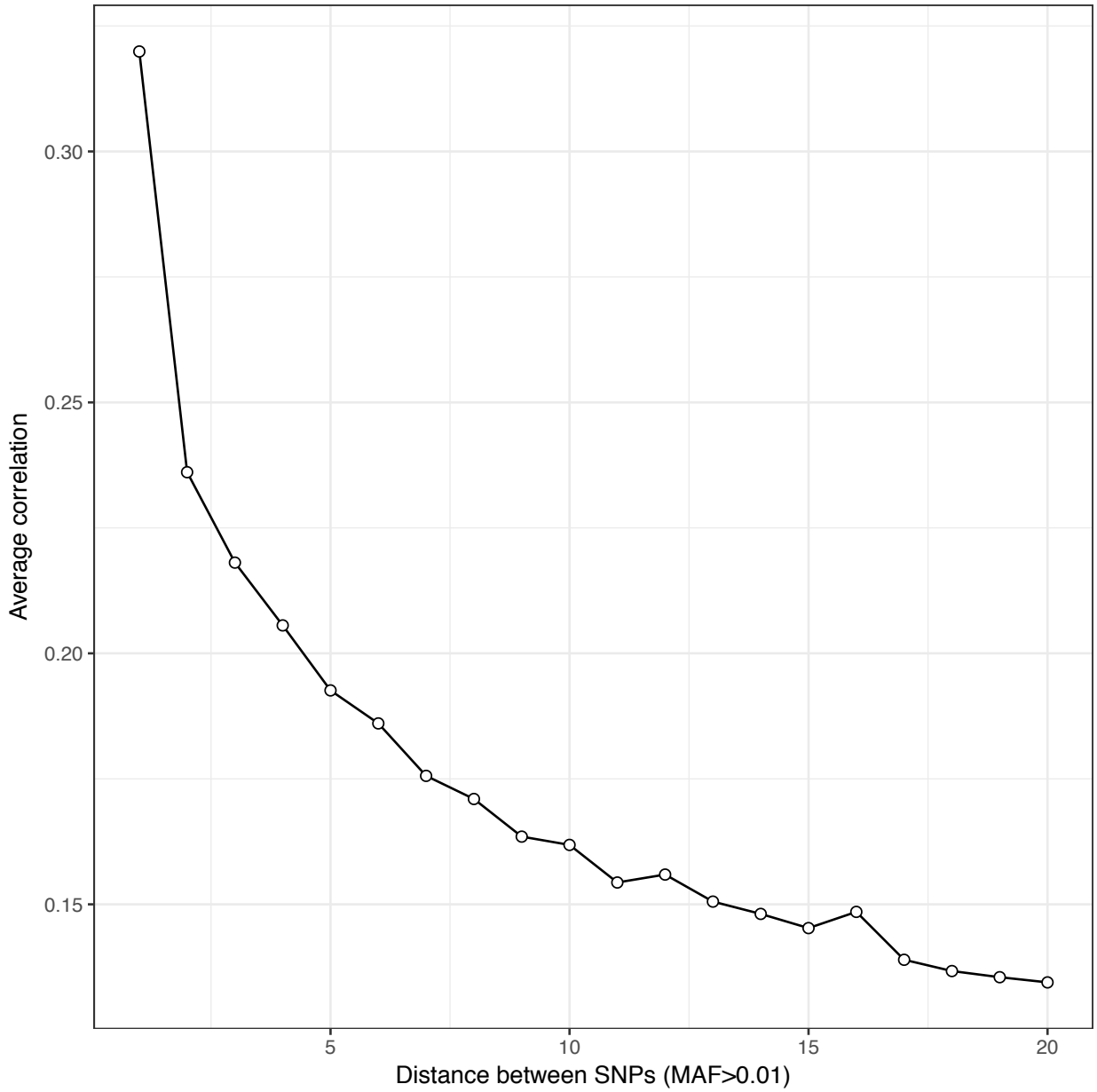


Figure S5. Average correlation (r^2) between variants against physical distance (in kb) for SNPs with a minor allele frequency greater than 0.01.

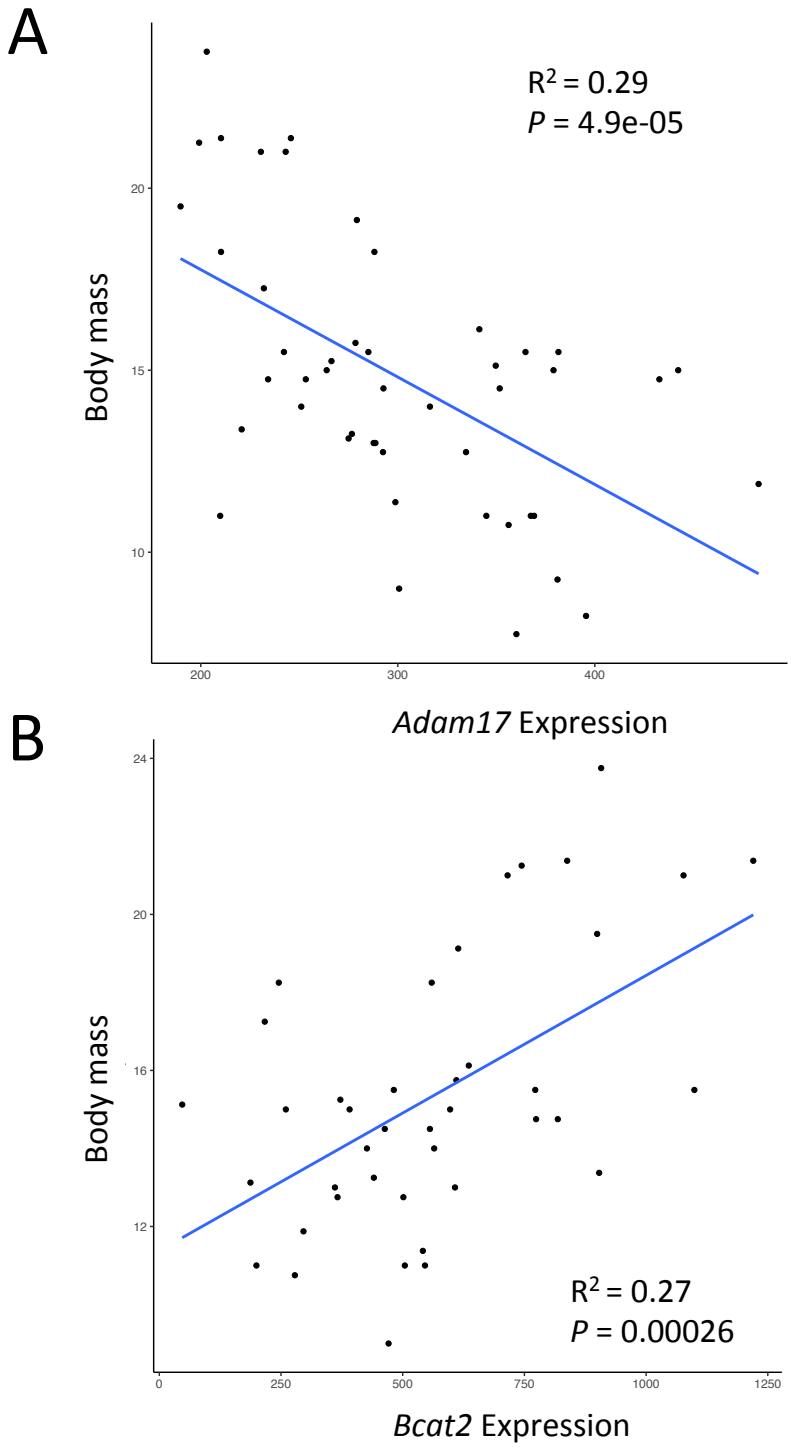


Figure S6. A. Correlation between *Adam17* expression and body mass without adjustment for latitude. B. Correlation between *Bcat2* expression and body mass without adjustment for latitude.

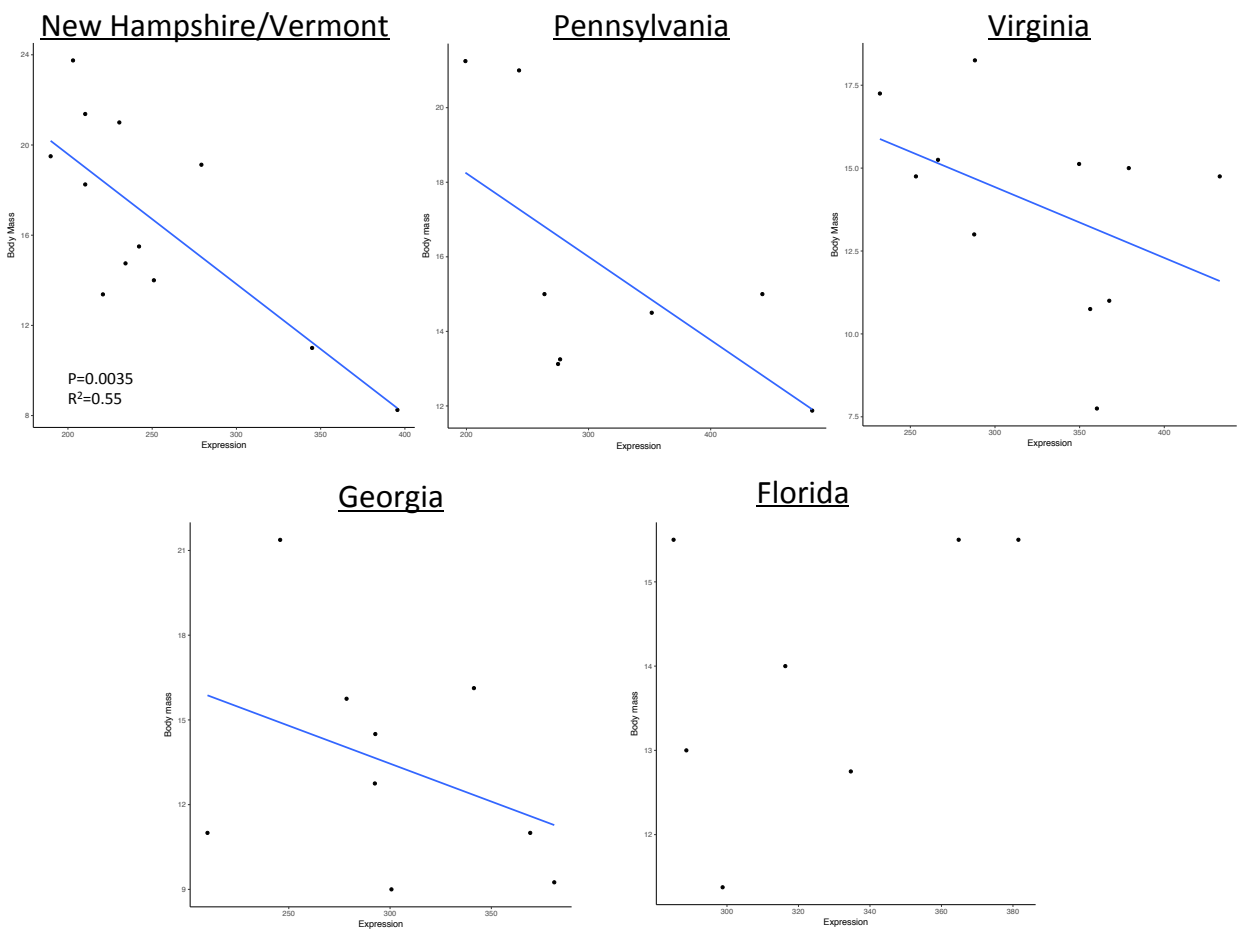


Figure S7. Correlations between *Adam17* and body mass in 5 populations. *Adam17* is significantly associated with body mass in one population (New Hampshire/Vermont, p=0.0035) and trends in the right direction in 4 of the 5 populations.

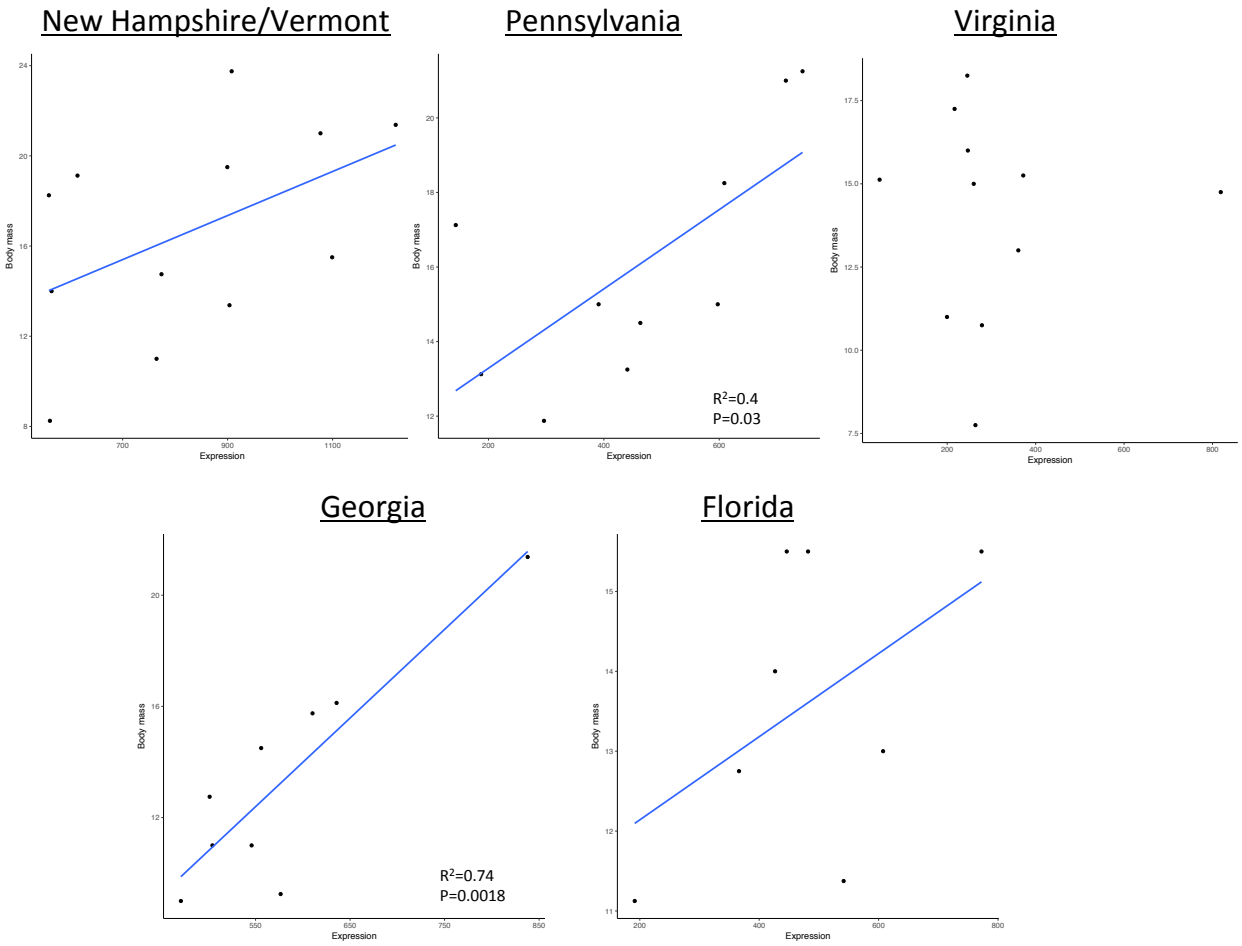


Figure S8. Correlations between *Bcat2* and body mass in 5 populations. *Bcat2* is significantly associated with body mass in two population (Pennsylvania and Georgia, p=0.03 and p= 0.0018, respectively) and trends in the right direction in 4 of the 5 populations.

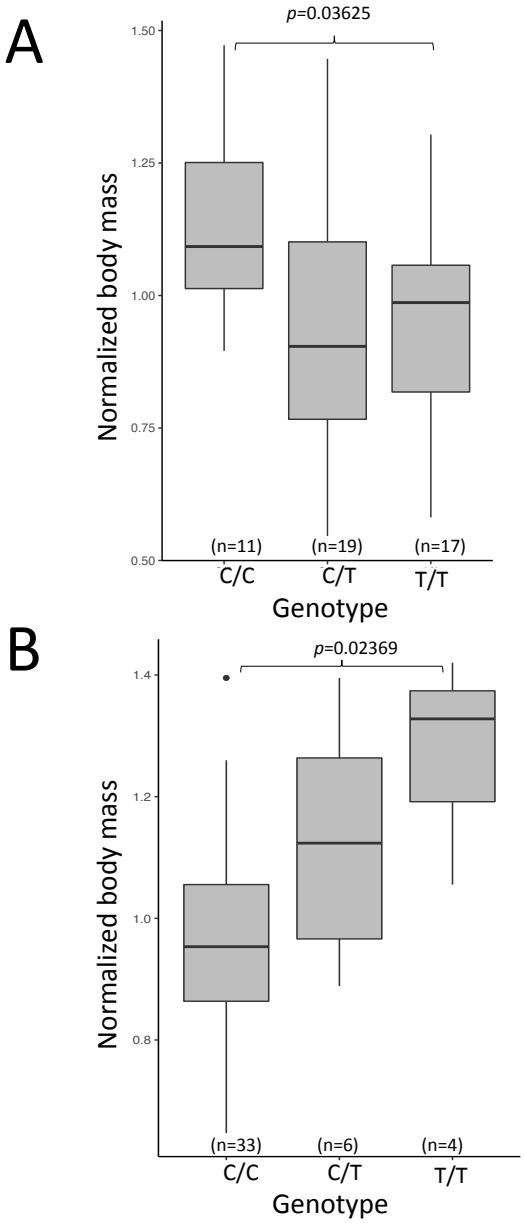


Figure S9. *Cis*-eQTL for (A) *Adam17* and (B) *Bcat2* are associated with differences in body mass.

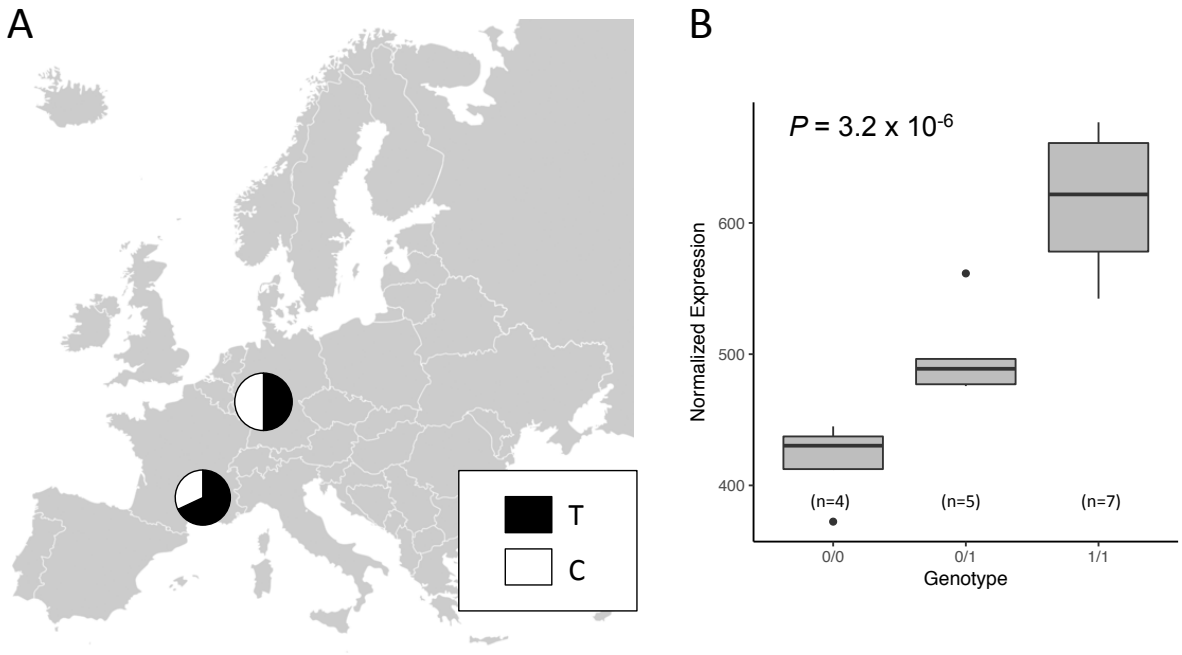


Figure S10. A. The *Adam17 cis*-eQTL (Chr12:21332631) in Europe. B. The Adam17 *cis*-eQTL is correlated with the expression of *Adam17* in European individuals.

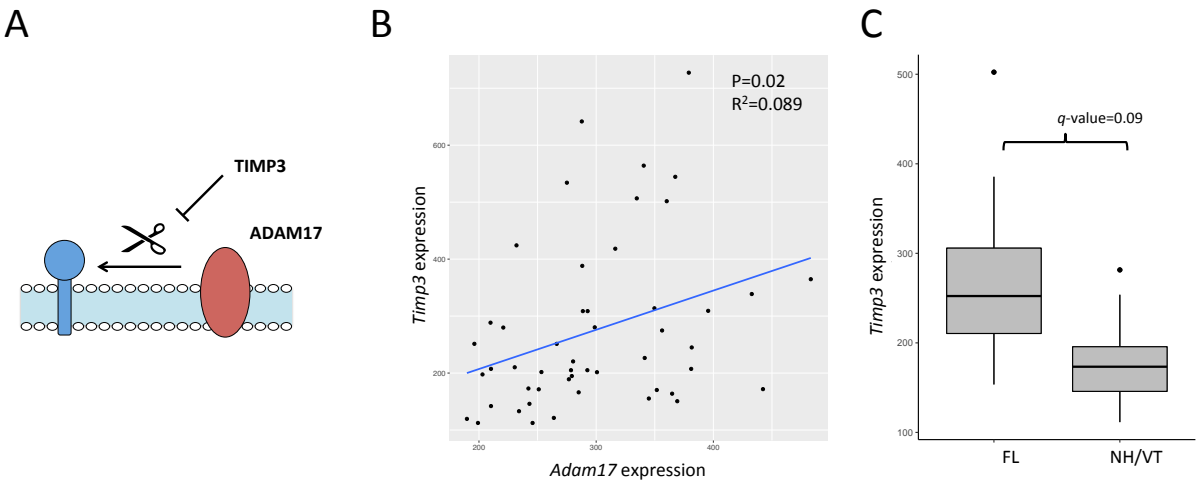


Figure S11. A. ADAM17 is a sheddase that is physiologically inhibited by TIMP3. B. *Timp3* expression is correlated with *Adam17* expression. C. *Timp3* is differentially expressed between the Florida and New Hampshire/Vermont populations.

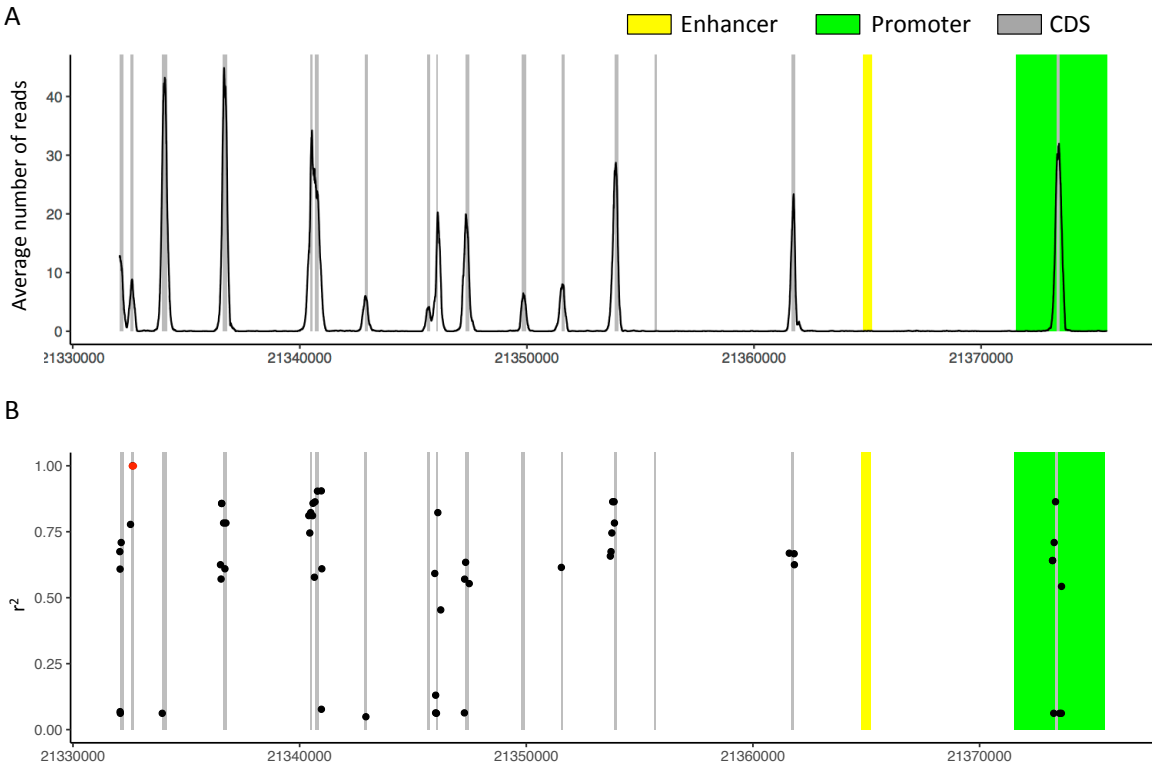


Figure S12. The *Adam17 cis*-eQTL is in LD with sites in a proximal promoter region. A. Average number of reads across individuals per site in exome data. B. r^2 between the *Adam17 cis*-eQTL (in red) and proximal sites.

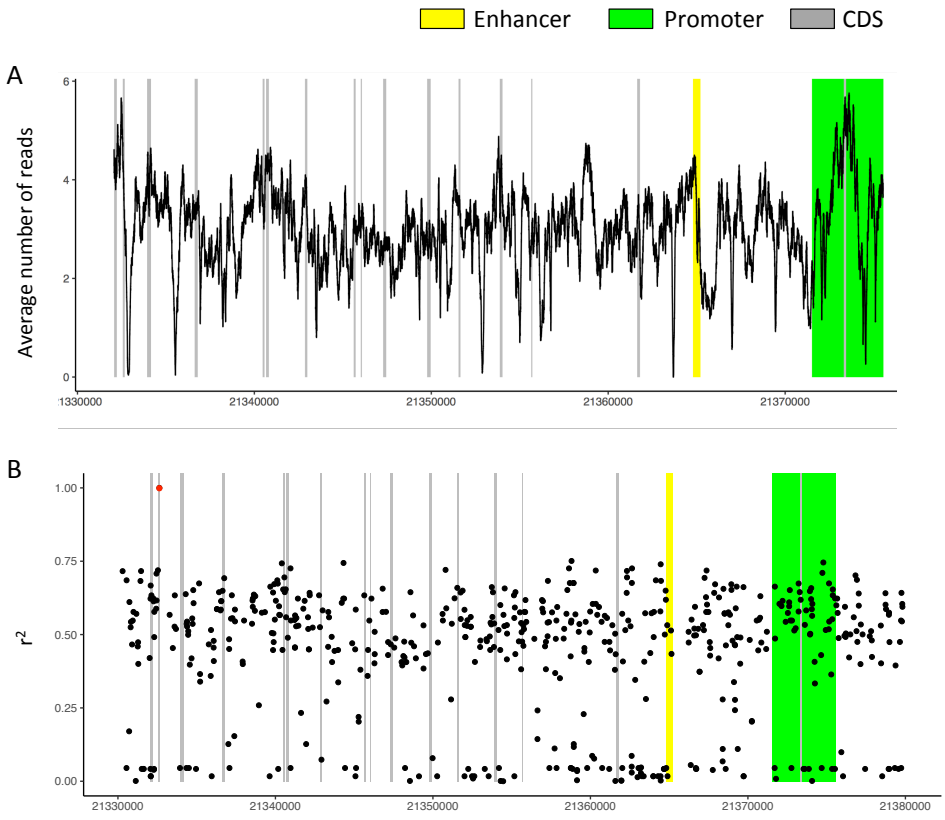


Figure S13. A. Average number of reads across individuals per site in low coverage whole genome data. B. r^2 between the *Adam17 cis*-eQTL (in red) and proximal sites.

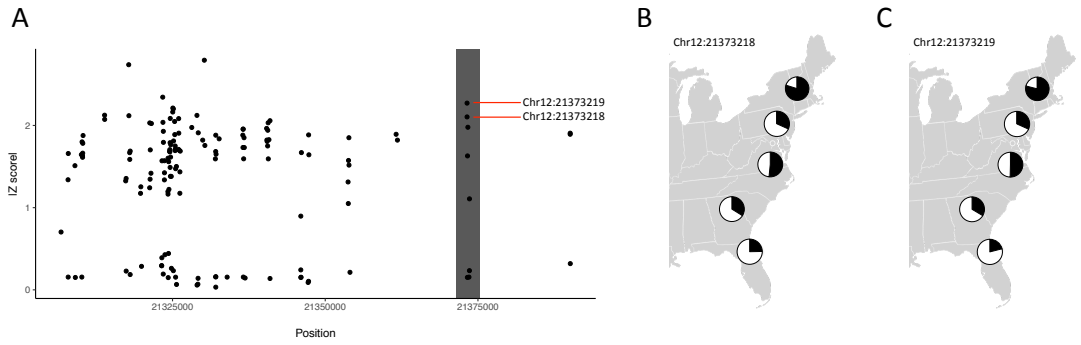


Figure S14. A. Two SNPs (Chr12:21373219 and Chr12:21373218) in the *Adam17* promoter (highlighted in grey) are clinal outliers. B. Clinal variation in allele frequencies at Chr12:21373218. C. Clinal variation in allele frequencies at Chr12:21373219.

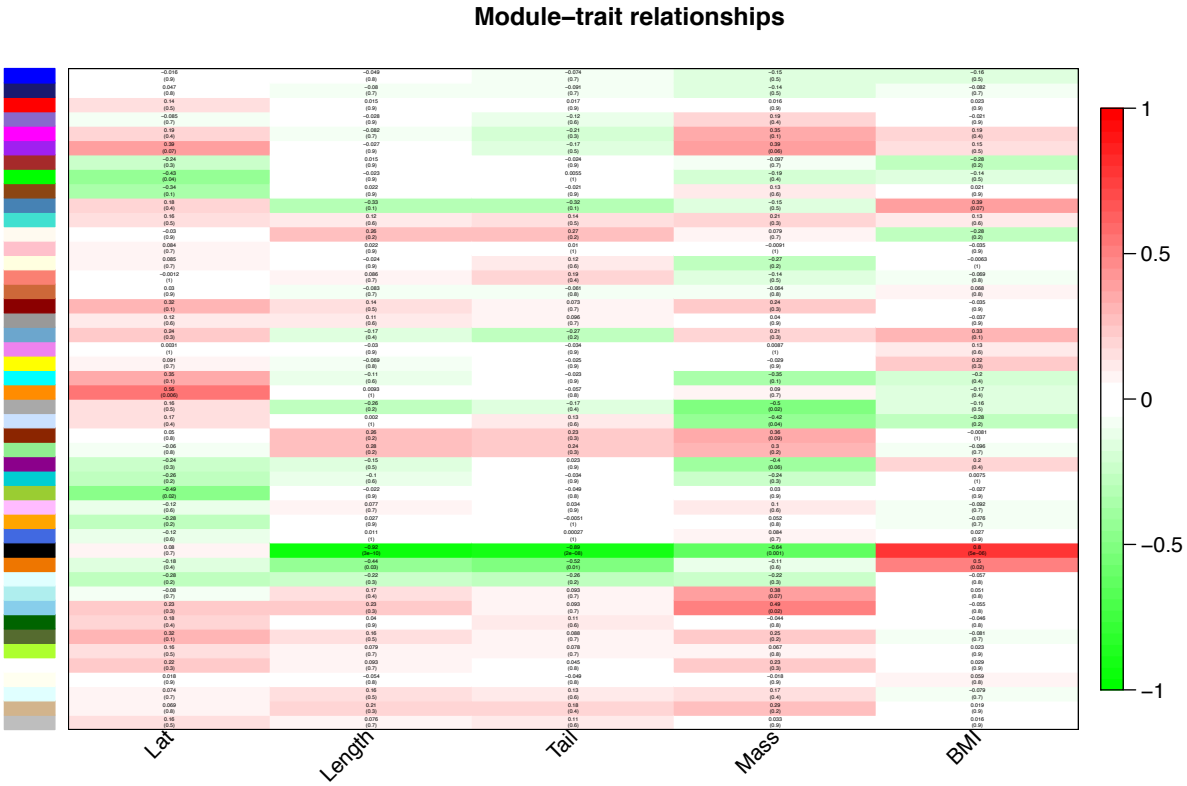


Figure S15. Relationship between co-expression modules and external traits in male mice. Each number represents the correlation between the module eigengene and the external trait and in parentheses is the associated *p*-value.

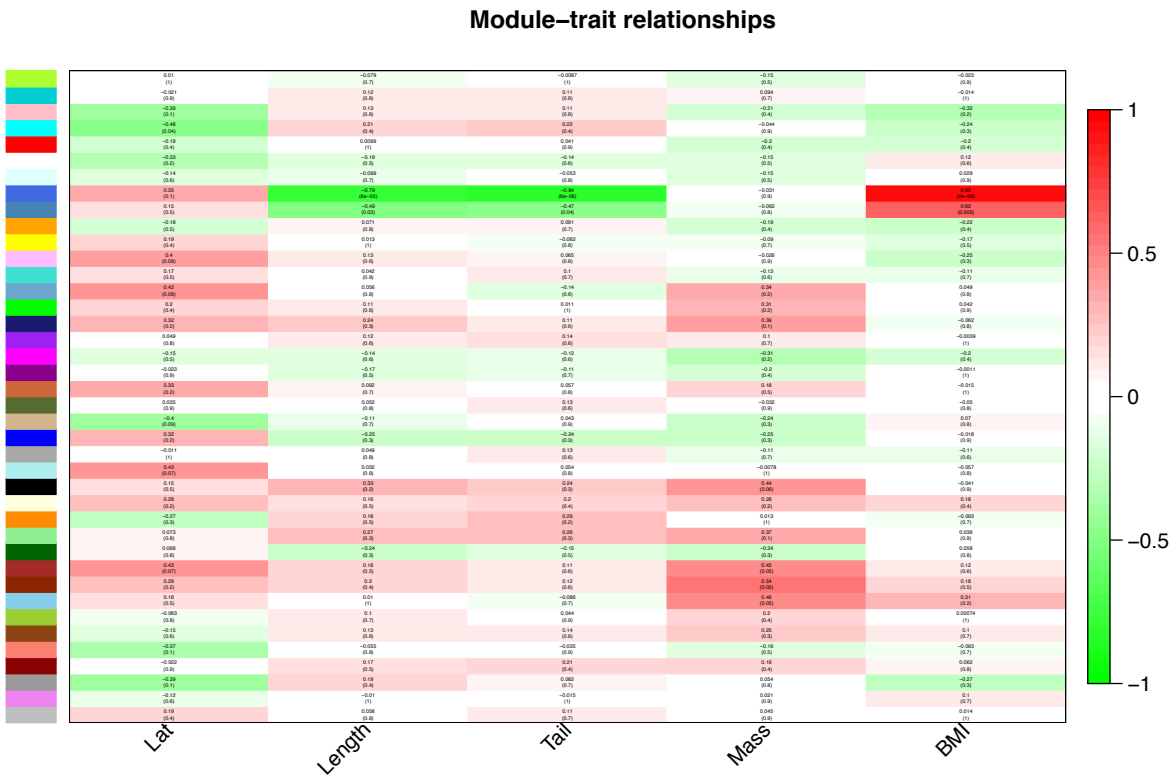


Figure S16. Relationship between co-expression modules and external traits in female mice. Each number represents the correlation between the module eigengene and the external trait and in parentheses is the associated *p*-value.

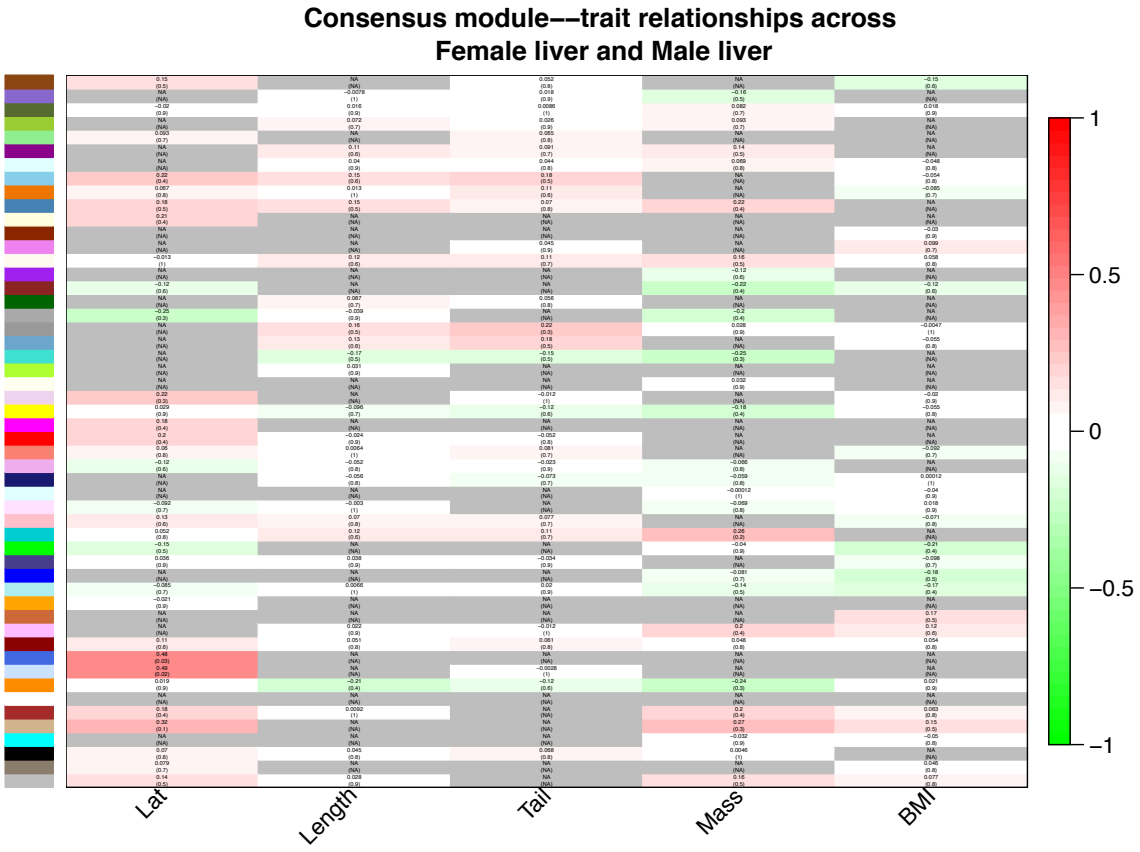


Figure S17. Relationship between consensus co-expression modules between males and female mice and external traits. Each number represents the correlation between the module eigengene and the external trait and in parentheses is the associated *p*-value.

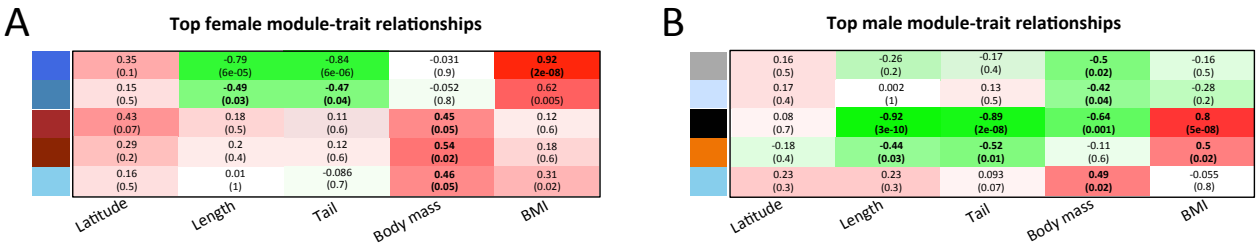


Figure S18. Co-expression modules with the top trait associations for female (A) and male (B) mice.

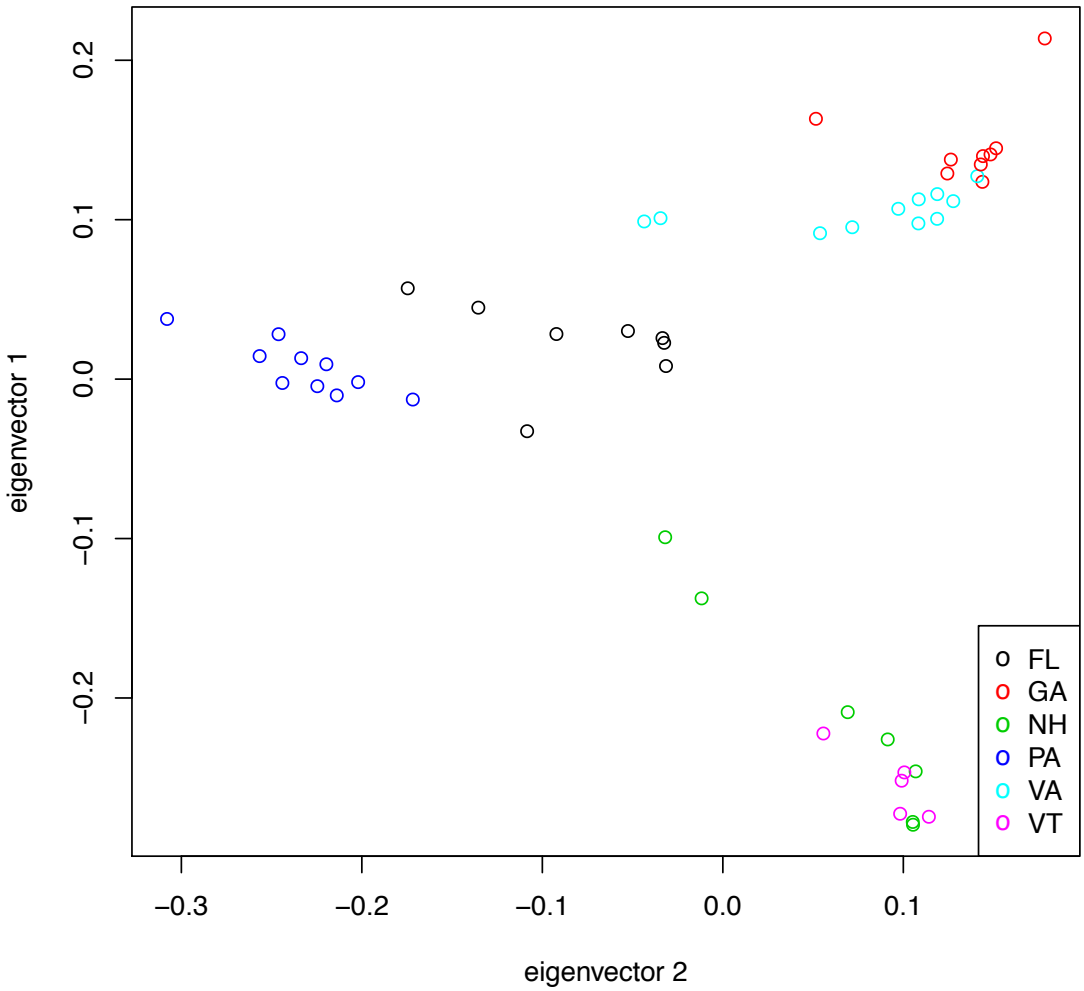


Figure S19. Principle-component analysis of genotype data from 50 individuals. Colors denote the population of origin.

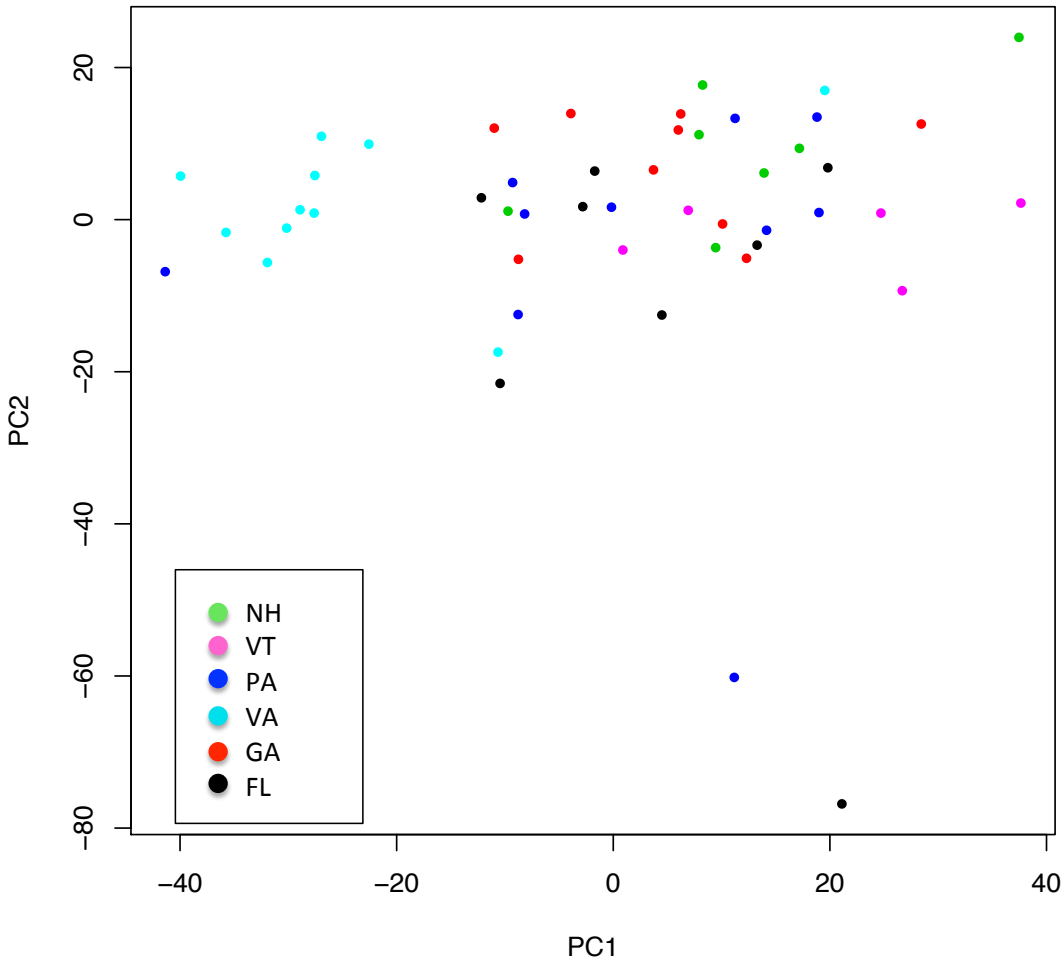


Figure S20. Principle-component analysis of expression data from 50 individuals. Colors denote the population of origin.

Chapter 4

Copy number variation in natural populations of house mice (*Mus musculus domesticus*) along an environmental gradient

Anticipated co-authorship: Mallory A. Ballinger[1], Megan Phifer-Rixey[2], Michael W. Nachman[1]

[1]Department of Integrative Biology and Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720, USA
[2]Department of Biology, Monmouth University, West Long Branch, NJ 07764, USA

Abstract

Copy number variants (CNVs) are thought to account for a substantial proportion of total genetic variation and have been associated with phenotypic differences between individuals that can impact fitness. Despite this, there are still few examples of copy number variants that contribute to local adaptation. We apply a read-depth based approach to characterize copy number variation using low-coverage whole genome data in wild-caught individuals of house mice (*Mus musculus domesticus*) collected from five populations along a latitudinal cline in the eastern United States. Consistent with a role for CNVs in local adaptation, we identified two regions where copy number is significantly correlated with latitude. These two regions overlap with 7 genes, whose functions include immunity and cold reception. One of these genes, *Trpm8*, has previously been shown to affect physiological responses to environmental cold in mice, ground squirrels, and hamsters. These results suggest that copy number variation is a significant contributor to genetic variation in North American populations and plays an important role in environmental adaptation.

4.1. Introduction

A major goal in evolutionary biology is characterizing the genetic variation that underlies adaptation. Copy number variants (CNVs) are a type of structural variation where segments of DNA vary in number between individuals. This class of structural variation is known to represent a major source of genetic diversity in mammals (Graubert *et al.* 2007, Perry *et al.* 2008). CNVs can also have phenotypic consequences and affect organismal fitness by altering the coding sequences or the expression of genes. In humans, CNVs have been linked to complex diseases including autism (Pagnamenta *et al.* 2009), schizophrenia (International Schizophrenia Consortium 2008), and diabetes (Jeon *et al.* 2010), among others (Consortium TWTCC 2010, Girirajan *et al.* 2011). Copy number variation may also provide an important source of genetic variation that selection can act on to

promote adaptation. However, there are still relatively few examples of copy number variants that have been demonstrated to be locally adaptive (Iskow *et al.* 2012).

House mice (*Mus musculus domesticus*) are a useful model for studying the role of structural variation in adaptation. House mice are the premier mammalian biomedical model system and have recently expanded into novel habitats worldwide in association with humans (Phifer-Rixey and Nachman 2015). Studies of CNVs within house mice have found that these variants are a major source of genetic diversity and are often differentiated between inbred lines and between populations (Locke *et al.* 2015, Bryk and Tauz 2014, Pezer *et al.* 2015).

In the eastern United States, house mice show clines in body size and behavioral variation consistent with thermoregulatory adaptations. Mice at higher latitudes are larger than mice at more southern latitudes, and this difference persists in the lab indicating that these differences are genetic (Lynch 1992, Phifer-Rixey *et al.* 2018). Laboratory strains founded from the northern and southern locations also show differences in aspects of blood chemistry, including leptin, glucose, and triglyceride levels (Phifer-Rixey *et al.* 2018), as well as behavioral differences (Lynch 1992, Phifer-Rixey *et al.* 2018).

Recent work with these populations identified clinal sequence variation associated with gene expression and phenotypic differences (Phifer-Rixey *et al.* 2018, Mack *et al.* 2018). Here we use whole-genome data from individuals in these populations to characterize copy number variation and search for copy number variation that varies clinally with latitude. Divergence in copy number among populations along this cline may be evidence of a role for these structural variants in environmental adaptation in house mice.

4.2. Results and Discussion

4.2.1. *Distribution of shared copy number variants*

Low coverage whole-genome data were used to characterize copy number variation in wild populations of *Mus musculus domesticus.* Fifty wild individuals were collected from 5 populations (New Hampshire/Vermont, Pennsylvania, Virginia, Georgia, Florida) along a ~15° latitudinal gradient in Eastern North America (Figure 1A; Phifer-Rixey *et al.* 2018).

To identify CNVs, reads were mapped to the mouse reference genome (mm10/GRCm38) and copy number variant deviations were identified in 20-kb windows using a read-depth based approach implemented in the program FREEC (Boeva *et al.* 2011, Boeva *et al.* 2012)(see methods for discussion of this approach). In order to study population-level variation and reduce false positive calls, we focused on windows with copy number calls in at least 8 of the 50 individuals included in this study (~16.3% of individuals). After this filtering, we identified 117 shared copy number variants distributed throughout the genome (Figure 1B). A dendrogram generated by clustering individuals based on the presence or absence of CNV calls does not group individuals by population (Figure 1C), indicating that much of the variation is shared between among populations. We also found that genetic distance between individuals based on CNVs alone is not associated with

geographic distance (Mantel test, p=0.51, 999 permutations). The median tract length of these regions was 100kb (mean=289.28kb, Figure S1). These regions overlapped 280 genes, and 36% of these genes were partially or fully deleted in at least one individual. We found that genes that intersect copy number variants in this analysis also strongly overlap genic copy number variants identified in a study of wild European mice (hypergeometric test, $p = 9.57 \times 10^{-60}$) (Pezer *et al.* 2015), suggesting much of the variation we observed is also present in European populations.

In comparing average genic copy number between populations, we identified 43 genes with differences in copy number between at least two populations (Kruskal-wallis, P<0.05)(File S1), but did not identify any cases of population-specific amplifications or deletions. Genes with high differentiation between populations were enriched for Gene Ontology (GO) terms including G-protein coupled receptor activity ($q = 9.83 \times 10^{-4}$), signal transducer activity ($q = 2.36 \times 10^{-5}$), and transferase activity ($q = 2.88 \times 10^{-3}$).

4.2.2. *Copy number variation and gene expression in liver tissue*

One mechanism through which CNV can result in phenotypic variation is through changes in gene dosage. For example, gene duplications may result in increased expression of a gene's product, which can directly impact organismal fitness.

To assess whether copy number influenced expression in these populations, we used liver RNAseq data collected from the same individuals to survey the expression of genes overlapping copy number calls (see methods). If increased copy number results in increased expression increased expression of a gene in this tissue, we expect a positive correlation between the number of copies in an individual and the expression of that gene. Of the 108 genes overlapping copy number variants that also were expressed in the liver, 24 (22.22%) showed significant positive correlations between copy number and gene expression (Spearman's rank correlation, p<0.05)(Table S1). Two of these genes fall within a CNV on chromosome 17 (*Glo1* and *Dnah8*) and also show differences in average copy number between populations (Figure S2), suggesting that this copy number deviation results in expression differences between populations. Variation in *Glo1* copy number has previously been linked to anxiety-like behavior in inbred and outbred mice (Hovatta *et al.* 2005, Williams *et al.* 2009), as well as disease phenotypes in humans, including panic disorder and autism (Junaid *et al.* 2004, Sacco *et al.* 2007, Politi *et al.* 2006).

4.2.3. *Clinal variation in copy number*

A classic approach to identifying genetic variation that reflects local adaptation is to search for allele frequency changes that co-vary with an environmental gradient. Differences in average copy number between populations of house mice in eastern North America may reflect adaptations to environments that differ in temperature and other aspects of climate. Although geographic clines may also be explained by isolation by distance, there is no evidence of isolation by distance in these populations (Phifer-Rixey *et al.* 2018) nor do CNVs cluster based on geographic distance between populations (see above).

To search for CNVs that vary clinally in eastern North America, Pearson's correlation was used to test 20kb windows of variable copy number for correlations with latitude. *P*-values were then subjected to a false discovery rate correction and widows with a *q*-value < 0.10 were considered significant. To prevent copy number expansions or contractions in any one population from creating a spurious clinal signal, each population was dropped in turn and the significant regions were re-tested for correlations with latitude. Two regions, one on Chromosome 1 and one on chromosome 4, remained significant after this re-testing procedure (Figure 2C)(File S1). The region on Chromosome 1 encompasses the full transcriptional units for the genes *Mroh2a and Hjurp* and one complete transcript and the promoter region for the gene *Trpm8*. The region on Chromosome 4 encompasses the full transcriptional units for the genes *Skint9* and *Skint3* and partial transcripts of *Skint4 and Skint2.*

All genes within these two clinally varying regions have been identified as CNVs in previous studies surveying wild mice from other populations (Locke *et al.* 2015, Pezer et al. 2015). To compare average copy number for these regions in the mice surveyed here to other wild mice, we downloaded whole genome data from French, German and Iranian populations of *M. m. domesticus* (Harr *et al.* 2015). We found that the average copy number in the Eurasia populations is more similar to that of FL than the NH/VT populations for both regions (File S1).

Genes within these regions are candidates for environmental adaptation. Latitude on the east coast of North America is highly associated with several climatic variables, including mean annual temperature (Phifer-Rixey *et al.* 2018). Given the latitudinal variation in temperature and the metabolic differences between populations, a gene of particular interest is *Trpm8. Trpm8* is the primary molecular transducer of cold somatosensation and plays an essential role in physiological thermoregulation (Bautista *et al.* 2007, Dhaka *et al.* 2007, Milenkovic et al. 2014, Peier et al. 2002, Voets *et al.* 2004, McKemy *et al.* 2002). The TRPM8 protein is primarily expressed in sensory neurons where it is activated by cold temperatures (Peier *et al.* 2002, Bautista *et al.* 2007). *Trpm8* deficient mice exhibit no preference for optimum ambient temperature and impaired cold avoidance (Bautista *et al.* 2007, Dhaka *et al.* 2007, Colburn *et al.* 2007). A recent study from Matos-Cruz *et al.* (2017) demonstrated that activity-reducing substitutions within the *Trpm8* gene in thirteen-lined ground squirrels and Syrian hamsters increase cold tolerance. *Trpm8* sequence variation has also been implicated in cold adaptation in humans (Key *et al.* 2018) and woolly mammoths (Lynch *et al.* 2015, Smith *et al.* 2017, Chigurapati *et al.* 2018). While the tissue-specificity of *Trpm8* expression means we are unable to connect the variation at this gene with phenotypic differences between individuals in these populations, the clinal variation at this gene, in combination with its molecular function, makes it an exciting candidate for thermoregulatory adaptation in house mice.

4.3. Conclusion

In this study, we (1) identified copy number variation in natural populations of house mice, (2) identified copy number variants associated with differentiation between populations, (3) identified copy number variants associated with expression variation, and (4) identified clines in copy number variants consistent

with local adaptation, including a cline at *Trpm8*, the gene encoding a cold receptor that has been previously implicated in adaptive physiological response to cold in other systems (Key *et al.* 2018, Matos-Cruz *et al.* 2017, Lynch *et al.* 2015, Smith *et al.* 2017, Chigurapati *et al.* 2018). The work described here adds to a growing number of studies that have used whole genome approaches to identify population level copy number divergence (e.g., Pezer *et al.* 2015) and copy number variation consistent with local adaptation (Schrider *et al.* 2013, Schrider *et al.* 2016, Bryk and Tauz 2014). These studies highlight the importance of CNVs to genetic variation and population divergence, as well as a possible role for these elements as substrates for adaptive evolution.

4.4. Methods

4.4.1. *Samples and sequencing*

Mice were sampled from five localities along a latitudinal cline as described in Phifer-Rixey *et al.* (2018). In brief, mice were sacrificed in the field and tissue was collected for 10 individuals from each population. Libraries from two or three individuals were sequenced on one lane of an Illumina HiSeq2000 (100bp paired-end reads), resulting in ~9-10Gb of raw sequence data. Reads were mapped to the mm10/GRCm38 reference with Bowtie2 (Langmead and Salzberg 2012). Average coverage across the whole genome was approximately 2.5X. For sites where each individual had a least one mapped read, coverage was 3.3X. These data were previously published and additional details can be found in Phifer-Rixey *et al.* (2018).

4.4.2. *CNV detection*

The program FREEC was used to detect copy number variation relative to the mm10 (GRCm38) reference genome in 20kb windows. FREEC employs a read-depth based approach to estimate copy number using next-generation sequencing data. This method was chosen because it has been demonstrated to detect copy number variants with high reliability (Duan *et al.* 2013).

Read depth based approaches can be biased by GC content and region mappability (Magi *et al.* 2011). GC content varies across the genome and affects read coverage. To correct for GC-bias, we used FREEC to normalize read counts based on a GC-content profile of the mouse reference genome. Mappability, or how well reads map to a region, can be influenced by the presence of repetitive regions in the reference genome. Aligning reads to repetitive regions results in ambiguous mapping, which can skew estimates of coverage. To account for biases introduced by mappability, a mappability profile for the mm10 reference was created with GEM (Derrien *et al.* 2012) and used to correct read counts in FREEC. The resulting CNV predictions were tested for significance using a Wilcoxon test implemented in FREEC. One individual was determined to be an outlier based on sample clustering and excluded from further analyses. CNVs were summarized between individuals using the Bedtools (Quinlan *et al.* 2010) *intersect* command, requiring at least one 20-kb window of overlap. Notably, this strategy of using large fixed windows to infer copy number variation can result in higher confidence for CNV calls but can bias against the detection of small copy number variants (Pirooznia *et al.* 2015).

4.4.3. *Assessing call accuracy in low coverage data*

To assess our ability to detect copy number variation in our low coverage data with this method, we downloaded high coverage whole genome data from 5 European *Mus musculus domesticus* from Harr *et al.* (2016). We called CNV variation in 20kb windows at high coverage and then subsequently downsampled the read depth to match the approximate depth of our low-coverage libraries and then re-called CNV variation with the same samples. At low coverage, we were able to detect 81% of the copy number variants identified at high coverage. For copy number variants called at high and low coverage, 91% of the copy number calls (i.e., number of copies estimated for a region per individual) were the same. When copy number calls differed between high and low coverage test sets, the difference in calls was always 1. This suggests that while we have likely underestimated the number of copy number variants segregating within populations, the number of copies at a given region can be estimated with relatively high precision in the low coverage dataset.

4.4.4. *Gene expression and copy number analysis*

Liver RNAseq data were generated for 41 of the 50 individuals used in this study and previously published (Mack *et al.* 2018). In brief, liver tissue was taken in the field collected in RNAlater, stored overnight at 4^{o}C overnight and then frozen at -80^{o}C. RNA was extracted from liver tissue with the Qiagen's RNeasy Mini Kit. For each individual, 100bp paired-end reads were sequences on an Illumina HiSeq4000. Trimmomatic (Bolger *et al.* 2014) was used to trim adaptors and then trimmed reads were mapped with Tophat2 (Kim *et al.* 2013) to personal reference genomes based on the mm10 reference. Reads mapping to exonic regions were counted with HTSeq-count (Anders *et al.* 2015). Expression was quantile normalized and genes with low read counts (<10 reads on average per individual) were removed (see Mack *et al.* 2018 for details).

4.5. Chapter 4 Figures

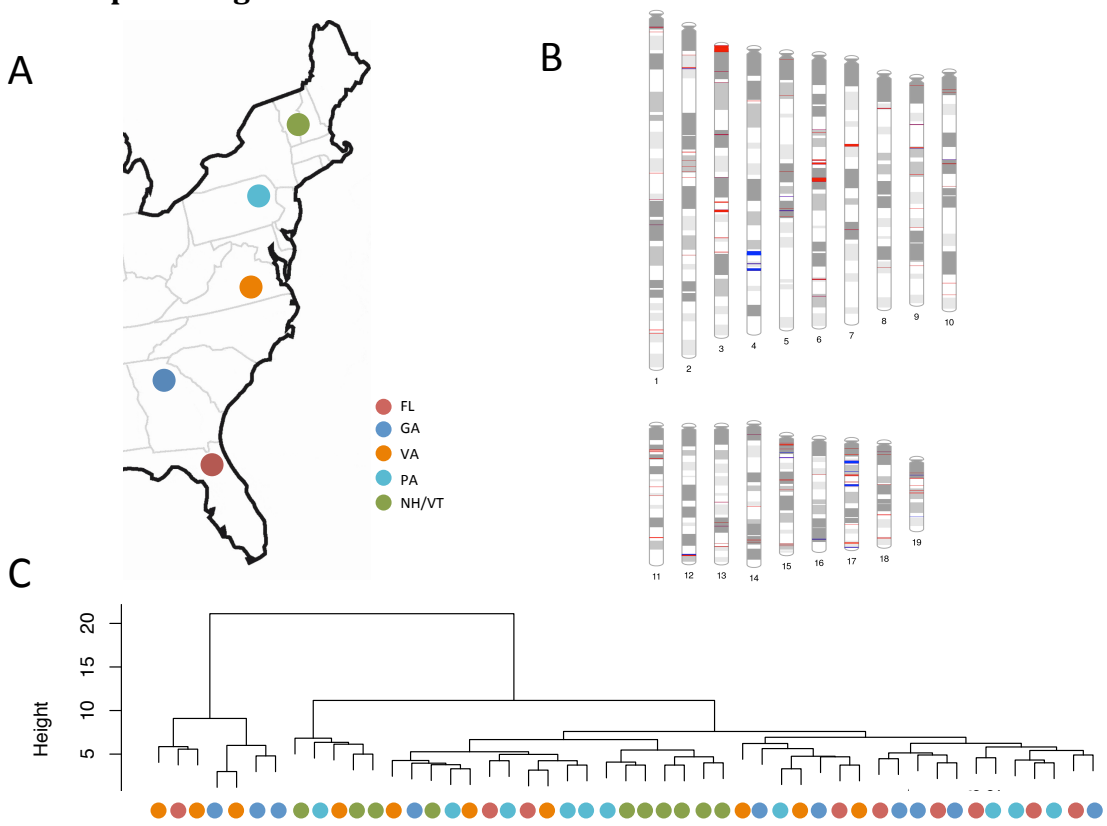


Figure 1. A. Sampling locations across the east coast of North America. Ten mice were collected for whole genome sequencing from each location. B. Distribution of shared CNVs across the House mouse genome. Red indicates gains and blue indicates losses compared to the house mouse reference. Purple indicate losses in some individuals and gains in others. C. Ward's hierarchical clustering of individuals based on the presence or absence of CNV calls does not group individuals by population.

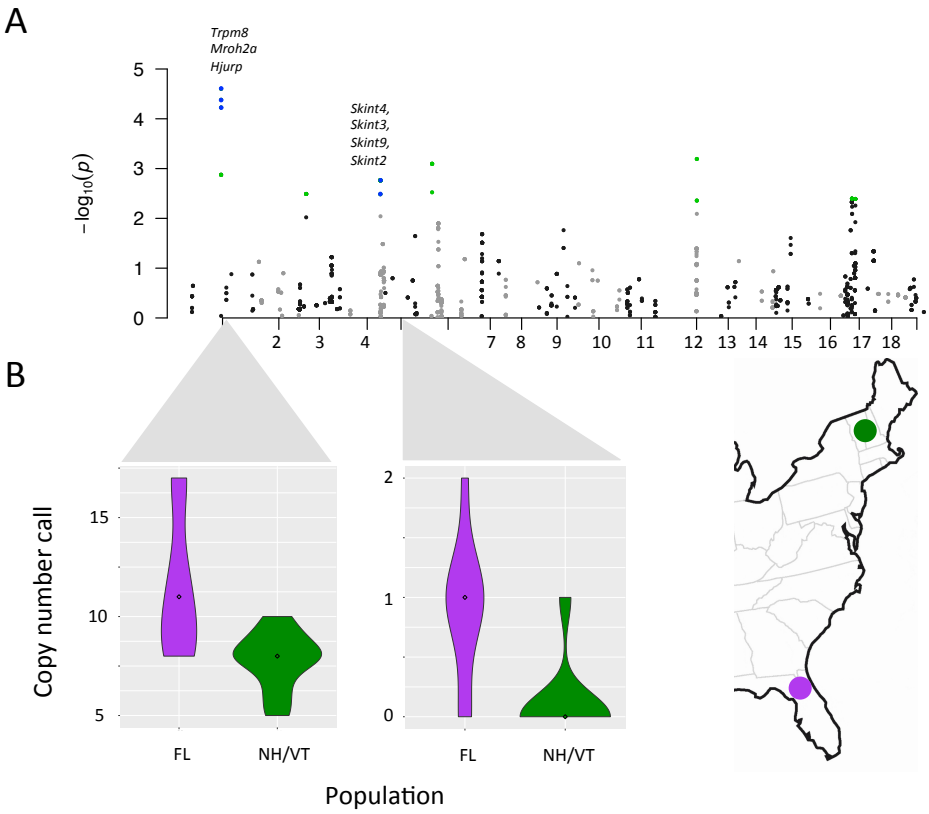


Figure 2. A. Correlations between copy number and latitude. We identified two regions (highlighted in blue), overlapping 7 genes, where copy number was significantly associated with latitude. B. Differences in average copy number between FL and NH/VT for the top latitude-associated windows. The diamonds indicate population means.

4.6. Chapter 4 Supplemental Tables

Table S1. Genes where liver expression is positively correlated with copy number state.

Gene name	*rho*[1]	*P*
Znrd1as	0.66	4.22E-06
Znrd1	0.65	4.75E-06
Btbd9	0.64	1.03E-05
Dnah8	0.61	3.08E-05
Glo1	0.60	4.98E-05
Slc22a3	0.50	1.09E-03
Nphp3	0.48	1.61E-03
Vwa5a	0.48	1.71E-03
Rpp21	0.48	1.85E-03
Arfgef1	0.46	2.55E-03
Trpm7	0.46	3.01E-03
Cwc22	0.42	7.65E-03
Ipo11	0.41	8.63E-03
Stap1	0.39	1.18E-02
Ugt1a6b	0.39	1.23E-02
Ubr3	0.38	1.57E-02
Uba5	0.36	2.37E-02
Wdr7	0.35	2.60E-02
Rictor	0.35	2.71E-02
Trim23	0.34	3.33E-02
Zfp874a	0.33	3.78E-02
Usp34	0.33	3.94E-02
Ppwd1	0.32	4.13E-02
Pds5a	0.32	4.29E-02

[1]Spearman's rank correlation *rho*

4.7. Chapter 4 Supplemental Figures

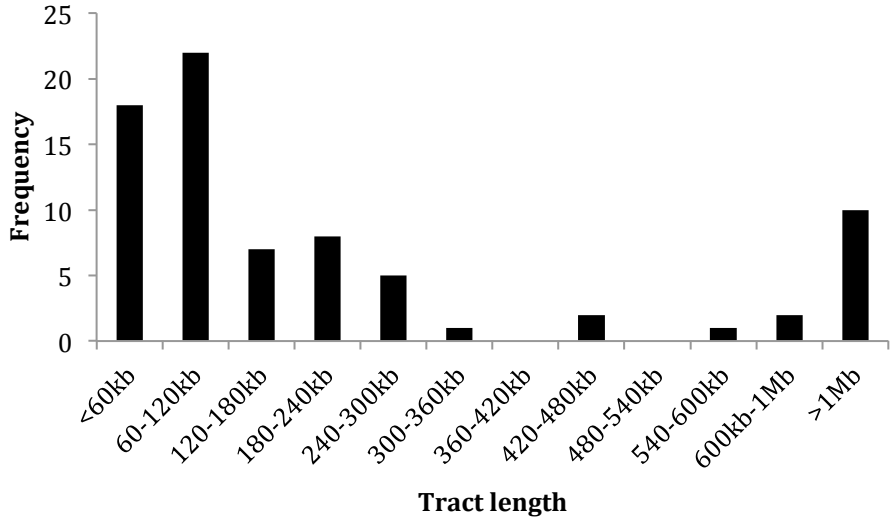


Figure S1. Length distribution of regions of variable copy number.

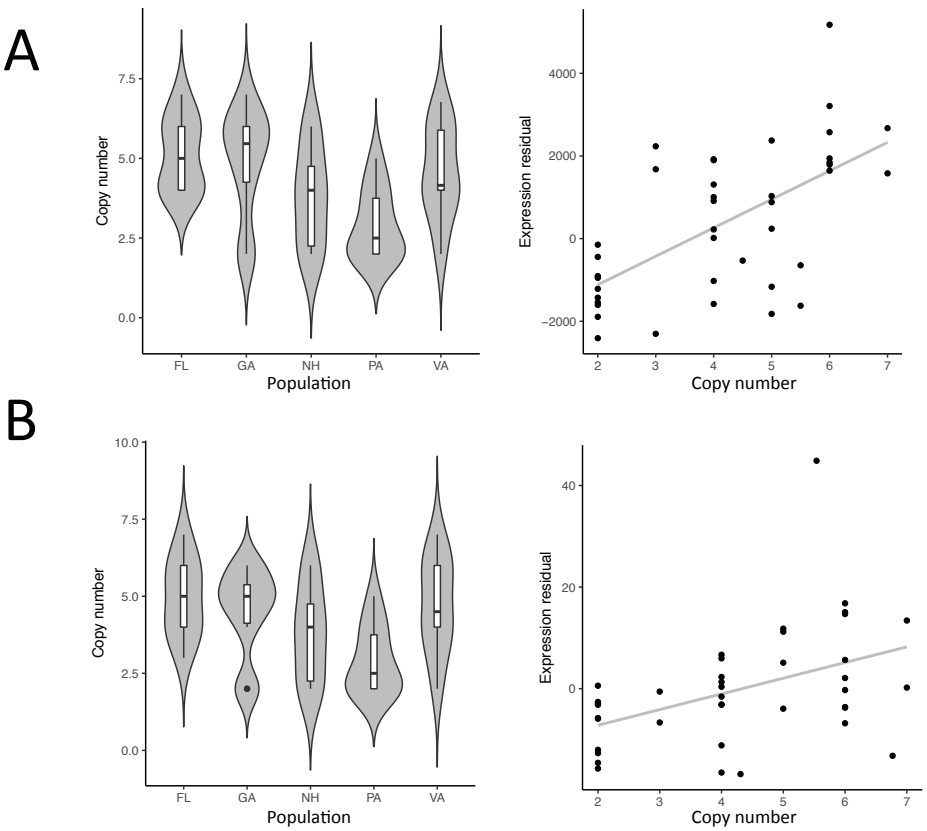


Figure S2. Two genes, *Glo1* (A) and *Dnah8* (B), in the same region for which genic copy number showed significant differences between populations (p=1.45x 10-2 and p=2.6 x 10^{-2}, respectively) and for which gene expression was significantly correlated with gene copy number (p=4 x 10^{-5} and *rho*=0.59; p=0.0059 and *rho*=0.43, respectively).

Chapter 5

Network connectivity, pleiotropy, regulatory variation, and constraint in natural populations of *M. m. domesticus*

Anticipated co-authorship: Megan Phifer-Rixey[1], Bettina Harr[2], Michael W. Nachman[3]
[1]Department of Biology, Monmouth University, West Long Branch, NJ 07764, USA
[2]Max-Planck-Institute for Evolutionary Biology, Plön, Germany
[3]Department of Integrative Biology and Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720, USA

Abstract

Interactions between genes can influence how selection acts on sequence variation. In gene regulatory networks, genes that affect the expression of many other genes may be under stronger evolutionary constraint than genes whose expression affects fewer partners. While this has been studied for individual tissue types, we know less about the effects of regulatory networks on gene evolution across different tissue types. We use RNAseq and genomic data collected from *Mus musculus domesticus* to construct and compare gene co-expression networks for 10 tissue types. We identify tissue-specific expression and local regulatory variation, and we associate these components of gene expression variation with sequence polymorphism and divergence. We found that genes with higher connectivity across tissues and genes associated with a greater number of cross-tissue modules showed significantly lower genetic diversity and lower rates of protein evolution. Consistent with this pattern, "hub" genes across multiple tissues also showed evidence of greater evolutionary constraint. Using allele-specific expression, we found that genes with *cis*-regulatory variation had lower average connectivity and higher levels of tissue specificity. Taken together, these results are consistent with strong purifying selection acting on genes with high connectivity both within and across tissues.

5.1. Introduction

Understanding the forces that govern genetic and phenotypic variation within and between species is an enduring problem in evolutionary biology. The number of interactions between genes and the phenotypic consequences of these interactions may be important determinants of evolutionary constraint (Fraser *et al.* 2002, Fraser *et al.* 2003). For example, a gene with many interactions in a gene regulatory network common across cells may be more pleiotropic than genes in the periphery of that network, or genes with tissue-specific expression (Stern and Orgogozo 2008, MacNeil and Walhout 2011). Such highly connected genes are expected to be under strong negative selection, as any change to these genes could affect their downstream partners (Stern and Orgogozo 2008). One approach to

studying relationships between genes across the genome is to identify gene co-expression networks, summarizing relationships between genes based on their coordinated expression across samples. Genes whose expression is more highly correlated with other genes in the network are thus more "connected" within a co-expression network. Gene co-expression is of biological interest as co-expressed genes are expected to be controlled by the same transcriptional regulatory program or otherwise be functionally related. Gene co-expression network analysis has been used to co-expressed gene sets, compare patterns across tissues (Pierson *et al.* 2015), between species (Stuart *et al.* 2003, Nowick *et al.* 2009, Eidsaa *et al.* 2017), and to identify sets of functionally related genes associated with quantitative or disease phenotypes (Ghazalpour *et al.* 2006, Chen *et al.* 2017, Yuan *et al.* 2017, Zhou *et al.* 2018).

A general feature of co-expression networks is that they are scale-free, with a small number of highly connected genes and many genes with very few connections (Barabasi and Oltvai 2004). The few highly connected genes are expected to show higher levels of pleiotropy compared to genes with fewer connections, and consequently are predicted to be more constrained both in terms of changes in gene expression and in protein sequence. Consistent with this, a number of studies have found that more connected genes exhibit lower genetic diversity and lower rates of molecular evolution (Masalia *et al.* 2017, Josephs *et al.* 2017, Mähler *et al.* 2017). These findings parallel what has been seen in protein-protein interaction networks, where genes encoding proteins with more protein-protein interactions have been shown to evolve more slowly than genes with fewer interactions (e.g., Fraser *et al.* 2002, Fraser *et al.* 2003).

The interplay of co-expression network topology and gene expression across tissues has received less attention. However, differences in co-expression networks between tissues may result in emergent properties of gene connectivity that affect sequence evolution. All cells carry out a combination of common and tissue-specific processes associated with their unique phenotypes. Consequently, genes that are highly connected in one tissue type may be more peripheral in others. Comparisons of co-expression networks across tissues can be used to characterize these differences (Pierson et al, 2015, Sonawane *et al.* 2017), and to investigate how such differences affect sequence evolution.

There are extensive genomic resources available for house mice (*Mus musulus domesticus*), making them a powerful system for studying co-expression networks. To investigate the relationship between cross-tissue co-expression networks and molecular evolution, we constructed co-expression networks for 10 tissue types in mice collected from natural populations. We used these data to compare co-expression network topology between tissues, identify tissue specific expression and local regulatory variation, and associate these components of gene expression variation with sequence variation and evolution.

5.2. Results and Discussion

We analyzed genome-wide expression data generated by Harr *et al.* (2016) for 224 tissue samples from 24 *M. m. domesticus*. These samples correspond to 10

different tissue types (muscle, thyroid, brain, testis, spleen, liver, gut, heart, lung, kidney) collected from lab-born progeny of wild house mice of diverse genotypes captured in Iran (N=8), France (N=8), and Germany (N=8), and raised in a common environment (see File S1).

5.2.1. Properties of gene connectivity within and across tissues

To characterize properties of gene connectivity within and across tissue, we used Weighted Gene Co-expression Network Analysis (WGCNA)(Langfelder and Horvath 2008) to construct co-expression networks, identify co-expression modules, and estimate gene connectivity. In a gene co-expression network analysis, the expression of each pair of genes is compared across samples to create a co-expression network. A gene's connectivity is defined as the sum of connection strengths between a focal gene and all other genes in a network. Genes with similar expression patterns can then be grouped into co-expression modules (see methods) (Langfelder and Horvath 2008)(Figure 1).

First, we investigated general properties of co-expression network topology within and across tissue types. Consistent with previous studies (Josephs *et al.* 2017), we found a significant positive correlation between connectivity and gene expression level for each tissue type (Spearman's rank correlation, Table S1). Gene connectivity was also correlated between different tissue types (Spearman's rank correlation, Table S2), with correlation coefficients ranging between 0.06-0.35 in pairwise comparisons between tissues. Testis, brain, and spleen showed the lowest average correlation coefficients in pairwise comparisons between these and other tissues.

To investigate how properties of gene expression correspond to the preservation of co-expression relationships across tissues, we also used WGCNA to identify modules that are shared across two tissue types, known as consensus modules (Figure 1)(Langfelder and Horvath 2008). For each pairwise comparison, we restricted our analysis to genes expressed across all tissue types (10,780 genes), built co-expression networks for these genes for each tissue, and then constructed consensus networks across each tissue pair (45 comparisons total). We then counted how many consensus modules with which each gene was significantly associated. For example, a gene that is significantly associated with a co-expression module between every pair of tissues would be found in 45 consensus modules, whereas a gene that is only found in a consensus module between the liver and spleen would be found in one consensus module. Average expression was significantly positively associated with the number of pairwise consensus modules in which a gene was found (Spearman's rank correlation, $rho = 0.88$, $p < 2.2e\text{-}16$), as was average gene connectivity across tissues (Spearman's rank correlation, $rho = 0.79$, $p < 2.2e\text{-}16$). We also observed a significant, but weaker, negative association with tissue-specificity (see below)(Spearman's rank correlation, $rho = -0.088$, $p = 1.17e\text{-}15$), where genes that had higher tissue-specificity values were found in fewer consensus modules.

5.2.2. Tissue specific expression and connectivity

To characterize properties of gene connectivity within and across tissue, we used Weighted Gene Co-expression Network Analysis (WGCNA)(Langfelder and Horvath 2008) to construct co-expression networks, identify co-expression modules, and estimate gene connectivity. In a gene co-expression network analysis, the expression of each pair of genes is compared across samples to create a co-expression network. A gene's connectivity is defined as the sum of connection strengths between a focal gene and all other genes in a network. Genes with similar expression patterns can then be grouped into co-expression modules (see methods) (Langfelder and Horvath 2008)(Figure 1).

First, we investigated general properties of co-expression network topology within and across tissue types. Consistent with previous studies (Josephs *et al.* 2017), we found a significant positive correlation between connectivity and gene expression level for each tissue type (Spearman's rank correlation, Table S1). Gene connectivity was also correlated between different tissue types (Spearman's rank correlation, Table S2), with correlation coefficients ranging between 0.06-0.35 in pairwise comparisons between tissues. Testis, brain, and spleen showed the lowest average correlation coefficients in pairwise comparisons between these and other tissues.

To investigate how properties of gene expression correspond to the preservation of co-expression relationships across tissues, we also used WGCNA to identify modules that are shared across two tissue types, known as consensus modules (Figure 1)(Langfelder and Horvath 2008). For each pairwise comparison, we restricted our analysis to genes expressed across all tissue types (10,780 genes), built co-expression networks for these genes for each tissue, and then constructed consensus networks across each tissue pair (45 comparisons total). We then counted how many consensus modules with which each gene was significantly associated. For example, a gene that is significantly associated with a co-expression module between every pair of tissues would be found in 45 consensus modules, whereas a gene that is only found in a consensus module between the liver and spleen would be found in one consensus module. Average expression was significantly positively associated with the number of pairwise consensus modules in which a gene was found (Spearman's rank correlation, $rho = 0.88$, $p < 2.2e\text{-}16$), as was average gene connectivity across tissues (Spearman's rank correlation, $rho=0.79$, $p < 2.2e\text{-}16$). We also observed a significant, but weaker, negative association with tissue-specificity (see below)(Spearman's rank correlation, $rho=-0.088$, $p = 1.17e\text{-}15$), where genes that had higher tissue-specificity values were found in fewer consensus modules.

5.2.3. Relationship between regulatory variation and connectivity

Previous studies have found that genes with local regulatory variation also show lower average connectivity in gene expression networks (Mähler *et al.* 2017, Josephs *et al.* 2017). To investigate the relationship between connectivity and regulatory variation in house mice, we identified genes with allele-specific expression in each tissue type. Allele-specific expression, the difference in expression between parental alleles, can be used to identify *cis*- acting epigenetic or genetic variation in heterozygous individuals (Cowles *et al.* 2002). In each tissue, we

tested exonic heterozygous sites for differences in expression between parental alleles (see Methods)(Table S3). We identified 4,146 genes with allele-specific expression across all 10 tissue types (False-discovery rate < 0.1; Table S4), many of which (28.48%) showed allele-specific expression in more than one tissue-type.

We then tested whether genes with allele-specific expression showed lower average connectivity within a tissue. As the power to detect allele-specific expression increases with expression level (Fontanillas *et al.* 2010; Figure S2), connectivity scores were adjusted for average expression level within each tissue (see methods). We found that in all tissues, genes with regulatory variation had lower average connectivity than genes without regulatory variation (permutation tests, all comparisons $p < 0.0001$). We also found that genes with allele-specific expression had higher levels of tissue-specificity on average (permutation test, $p < 0.0001$; S_{max} adjusted for average expression level across tissues).

Genes with local regulatory variation may have lower average connectivity if genes with higher connectivity are under stronger purifying selection and thus less tolerant of regulatory variation. Consistent with this, we find that genes with allele-specific expression in any tissue have higher dN/dS values (Mann-Whitney U, p=0.03). We also downloaded protein interaction data from STRING (Szklarczyk *et al.* 2017) and found that genes with regulatory variation encoded proteins that have fewer interacting partners on average (Mann-Whitney U, $p < 2.2e-16$). Finally, we found that genes with allele-specific expression were less likely to encode transcription factors (χ^2 test, $p < 0.0001$). This was also observed for transcription factors that were considered tissue-specific (χ^2 test, $p < 0.0001$).

5.2.4. Relationship between connectivity and sequence evolution

To examine the relationship between sequence evolution and characteristics of expression, we performed pairwise tests between aspects of gene expression across tissues (average connectivity, average expression level, and variance in expression and connectivity across tissues) and measures of sequence variation (SNP density) and protein evolution (dN/dS ratio) (Table 1). To control for the relationship between these measures and other variables, we then performed partial Spearman correlations between characteristics of gene expression and sequence evolution. We found that average connectivity and average expression level across tissues showed highly significant negative associations with dN/dS ratio (Figure 2A) and SNP density (Figure 2B). Variance in gene expression level and connectivity across tissues, represented by the interquartile range of a gene's expression or connectivity, were also found be a significant predictor of dN/dS ratio and SNP density (Table 1). We also performed 1,000 permutations in which the relationship between the predictors and dN/dS ratio and SNP density was randomized. None of correlations in the permutated datasets were more extreme than the observed partial correlations.

Modules that are preserved across tissues are expected to have functions that are common across tissues (Pierson *et al.* 2015). To assess whether the

preservation of module relationships across tissues was also associated with rates of sequence evolution, we asked whether genes found in a greater number of consensus modules between pairs of tissue types showed greater sequence conservation. We predicted that genes that were found in more modules across tissues would show greater sequence constraint, as these genes may also show higher levels of pleiotropy. Consistent with prediction, we found that dN/dS (Figure 3A; Spearman's rank correlation, rho = -0.22, $p < 2.2 \times 10^{-16}$) and SNP density (Figure 3B; Spearman's rank correlation, rho = -0.16, $p < 2.2 \times 10^{-16}$) were significantly negatively correlated with the number of consensus modules in which a gene was found. As in the previous analysis, we also performed a partial Spearman correlation to account for average expression level, expression variance, gene connectivity, and variance in connectivity across tissues. We found that the association between dN/dS ratio (Partial Spearman correlation, rho=-0.11, $p < 2.2 \times 10^{-16}$) and SNP density (Partial Spearman correlation, rho = -0.039, p = 0.00036) were still significant when accounting for these variables. In 1000 permutations in which the relationship between pair number and dN/dS ratio and SNP density was randomized, no correlation was more extreme than that observed for the dN/dS ratio and only one permutation was more extreme than that observed for SNP density.

5.2.5. Constraint on cross-tissue hub genes

Co-expression analyses have been widely applied to identify "hub" genes, or genes whose expression is highly correlated with their expression module. Hub gene analysis has also become a popular method for identifying genes whose expression is related to variation in quantitative traits (Ghazalpour *et al.* 2006) or disease phenotypes (e.g., Chen *et al.* 2017, Yuan *et al.* 2017, Zhou *et al.* 2018). As hub genes represent genes most highly associated with their module's expression, we expected genes that were annotated as hubs in more tissues would be genes that were more essential and would show greater sequence constraint.

Each gene's module membership was estimated based on the correlation between that gene's expression and the expression of the module eigengene (Langfelder and Horvath 2008). Genes where module membership was greater than 0.8 were considered "hub genes" for subsequent analyses, a cut-off selected because of its usage in previous studies (e.g., Yuan *et al.* 2017). Consistent with what has been seen in human populations (Sonawane *et al.* 2017), we found that genes that encode transcription factors were more likely to be hub genes (χ^2 test, p = 0.0002).

We then compared hub genes across tissues. We found that a large proportion of the hub genes we identified in our analysis are unique to one tissue type (61%), and only 9.2% of these genes were annotated as hubs in 3 or more tissues (Figure S1). Consistent with the idea that cross-tissue hub genes represent genes with essential biological functions, we also found that genes that were identified in hubs in 3 or more tissues were highly enriched for mutant phenotypes related to mortality/aging ($q = 2.93 \times 10^{-12}$; including significant enrichment of the mortality/aging subcategories abnormal survival, preweaning lethality, prenatal lethality and embryonic lethality), abnormal cell physiology ($q = 5.66 \times 10^{-5}$), and abnormal homeostasis ($q = 1.98 \times 10^{-4}$). These genes were also enriched for several

GO terms, including positive regulation of biological process (q = 1.03 x 10^{-26}) and regulation of cellular processes (q = 3.01 x 10^{-22}). Genes annotated as hubs in just two tissues were also significantly enriched for mutant phenotypes related to mortality/aging, but this enrichment was less significant (q = 0.01).

Parallel to the previous analyses, where we asked whether more highly connected genes or genes found in a greater number of cross-tissue modules were more constrained, we also asked whether genes annotated as hubs in more tissue types were under greater evolutionary constraint by comparing the dN/dS ratios and SNP densities for genes identified as hubs in no tissues, 1 tissue (n=6,632), 2 tissues (n=2,532), and 3 or more tissues (n=1,001). We found that genes identified as hubs in more tissues showed lower average dN/dS and SNP density (Figure 4).

5.3. Conclusion

Here we have used natural populations of house mice to characterize co-expression networks for 10 tissue types and associate components of gene expression variation with sequence variation and evolution. Genes with higher connectivity across tissues showed significantly lower genetic diversity and lower rates of protein evolution. We also found that genes in more pairwise consensus modules show significantly lower genetic diversity and lower rates of protein evolution. Genes that were hubs across more tissues showed the same evidence for evolutionary constraint and were significantly enriched for mutant phenotypes related to mortality and aging. Finally, we found that genes with allele-specific expression had lower connectivity on average, lower dN/dS values, and fewer connections in protein-protein interaction networks. In this regard, regulatory variation at peripheral genes may provide variation that can act as a substrate for adaptive evolution. Altogether, these results are consistent with purifying selection acting on pleiotropic genes and suggest that gene connectivity is an important determinant of evolutionary constraint.

5.4. Methods

5.4.1. *Expression data*

RNAseq data was downloaded from Harr *et al.* (2016). These samples correspond to lab-born progeny of *M. m. domesticus* collected from Germany, Iran, and France and up to 10 tissue types per individual (muscle, thyroid, brain, testis, spleen, liver, gut, heart, lung, kidney). As reported in Harr *et al.* (2016), samples for DNA and RNA-sequencing were obtained from the first or second generation of out-breeding in an animal facility and are expected to represent full wild-type variation. Individuals used for RNA-sequencing were age-matched males. We downloaded RNAseq reads mapped with Tophat2 (Kim *et al.* 2013) to the mm10 reference genome (http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/). We then counted reads that mapped to exonic regions using HTSeq-count (Anders *et al.* 2015).

5.4.2. *Co-expression analysis*

First, samples that were tissue-specific outliers were identified through a principle component analysis and were removed from subsequent analyses. Three

individuals were removed because of relatedness (first or second degree relatives). For individual co-expression analyses, genes with fewer than 20 reads on average per tissue were removed. Gene expression was then quantile-normalized and corrected for the known co-variate of relatedness using a Baysian approach (Stegle *et al.* 2010, Stegle *et al.* 2012). The program Weighted Gene Co-expression Network Analysis was then used to construct co-expression networks for all tissue-types for all individuals, following WGCNA protocols (Langfelder and Horvath 2008). In short, we first constructed a gene co-expression network, represented by an adjacency matrix, which denotes co-expression similarity between pairs of genes across different samples, for each tissue. Then, modules were identified using hierarchical clustering. Dissimilarly between clusters is measured based on topological overlap. Each module is summarized by a representative eigengene, or the first principle component of the module. Each gene's total connectivity within a tissue was then retrieved using the command *intramodularConnectivity*.

To compare co-expression patterns across tissues, we then restricted our analysis to genes that were expressed across tissue types (10,780 genes). As described above, the program Weighted Gene Co-expression Network Analysis was used to build co-expression networks for each tissue type and then build consensus networks across each pair of tissues (45 pairs total).

5.4.3. *Tissue Specificity*

To compare gene expression across tissue types and identify genes with tissue specific expression, mapped reads were downsampled across samples/ tissues types to account for differences in average library size between individual samples. Genes with fewer than an average of 50 reads across all samples were discarded. Tissue specificity was subsequently defied as in Sonawane *et al.* (2017):

$$S_j^{(t)} = \left(med\left(e_j^{(t)}\right) - \ med\left(e_j^{(all)}\right)\right) - IQR\left(e_j^{(all)}\right)$$

Where the specificity (S) of gene j in tissue t corresponds to (the median (med) expression (e) of the gene in that tissue (t) - the median expression of the gene in all tissues (all)) - interquartile range (IQR) of expression of that gene across all tissues. A gene's highest S value across all 10 tissues was designed $S_{\max.}$ Genes in a tissue for which S>2 were considered tissue-specific. Under this definition, genes can be tissue specific in more than one tissue. The number of tissues a gene was considered for is the gene's multiplicity value. A total of 4902 genes were found to be tissue specific in just one tissue type, meaning these genes have a multiplicity of 1.

5.4.4. *Allele-specific expression*

To identify allele-specific expression, we downloaded genome-wide SNP calls from Harr *et al.* (2016) (http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/) for these individuals. Individuals that did not have corresponding genomic data were not included in this analysis. RNAseq reads mapped to the reference and alternative allele for heterozygous sites were counted using GATK ASEReadCounter (McKenna *et al.* 2010). Sites where fewer than 20 reads supported either the

reference or the alternative allele were discarded. Allele-specific expression was then called as described in Mack *et al.* (2018). The number of SNPs that could be tested in each tissue is listed in Table S3, corresponding to a total of 15,390 genes across all tissue types. We retained the variants with the lowest *p*-values per gene and then performed a false-discovery rate correction using R's *p.adjust* (See Table S4).

5.4.5. *Measures of sequence evolution*

Estimates of dN and dS between mouse and rat were downloaded from Ensembl (Zerbino *et al.* 2018). SNP density was estimated based on genome-wide SNP calls for from Harr *et al.* (2016), counting SNPs that fell within the boundaries of each gene and correcting for the length of a gene using gene start and stop annotations downloaded from Ensembl.

5.4.6. *Enrichment analyses*

Tests for enrichments of mutant phenotypes were done using modPhEA (Weng *et al.* 2017). All GO category enrichment analyses were performed with PANTHER (Mi *et al.* 2016).

5.4.7. *Protein interaction networks*

Protein networks were downloaded from STRING (Szklarczyk *et al.* 2017) for *Mus musculus domesticus.* Interactions were filtered for "high confidence" interactions (>0.7).

5.5. Chapter 5 Table

Table 1. Spearman's rank correlation coefficient between gene expression-related measures and sequence evolution

	dN/dS		SNP density	
	Pairwise	**Partial**	**Pairwise**	**Partial**
Average expression level across tissues	-0.26***	-0.15***	-0.15***	-0.14***
Expression IQR across tissues	-0.22***	0.042**	-0.05***	0.17***
Average connectivity across tissues	-0.18***	-0.045***	-0.16***	-0.09***
Connectivity IQR across tissues	-0.12***	0.04**	-0.11***	-0.04***

***P<0.0001

**P<0.001

5.6. Chapter 5 Figures

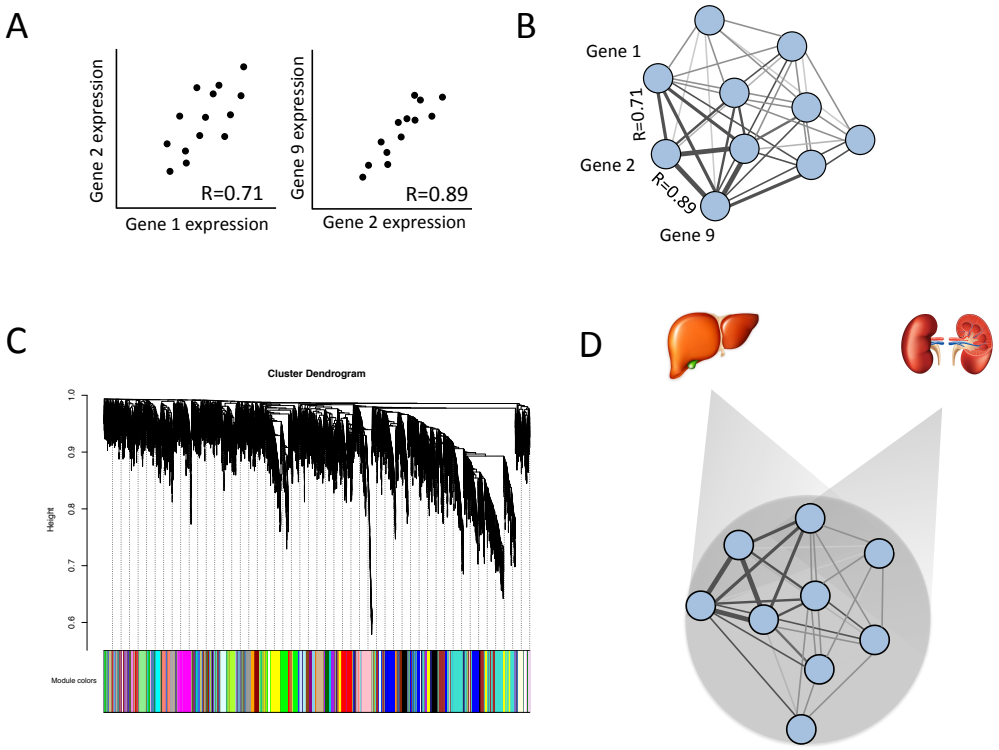


Figure 1. Constructing gene co-expression networks (Langfelder and Horvath 2008). (A) Co-expression similarity is compared between pairs of genes across different samples in order to build (B) a co-expression network. (C) Co-expression modules are identified using hierarchical clustering. (D) Consensus networks across each pair of tissues are then created to identify modules that are conserved across tissues.

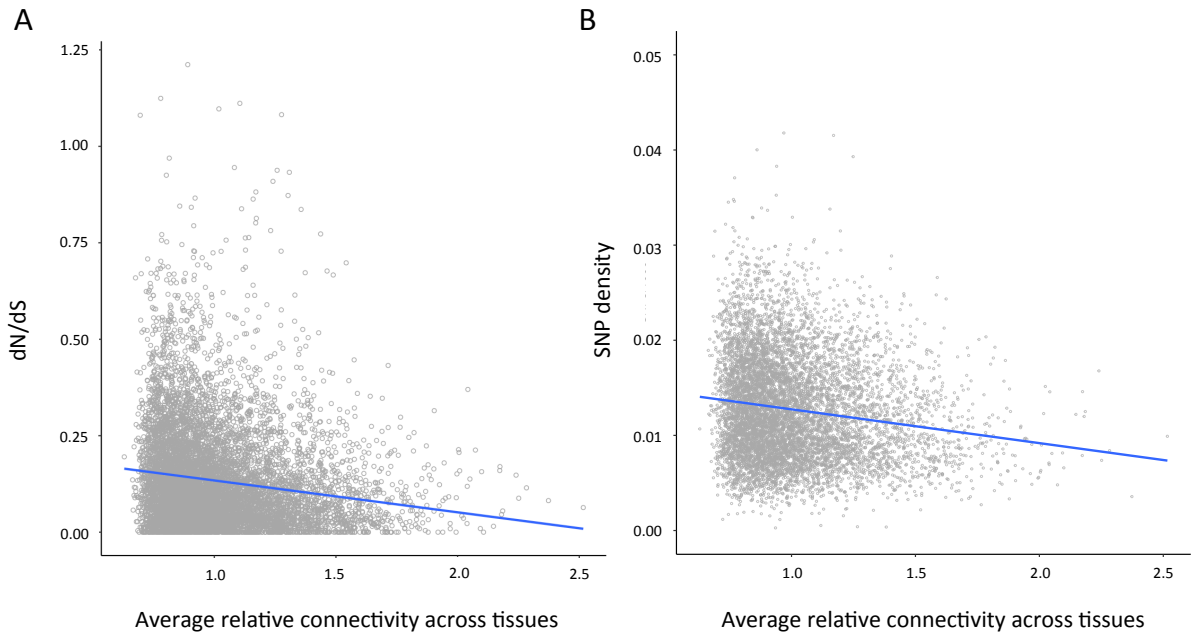


Figure 2. A. Average connectivity across tissues is significantly correlated with dN/dS ratio (Pairwise Spearman's rank correlation *rho*=-0.18, *P*<0.0001; Partial Spearman *rho*=-0.045, *P*<0.0001). B. Average connectivity across tissues is significantly correlated with SNP density (Pairwise Spearman's rank correlation *rho*=-0.16, *P*<0.0001; Partial Spearman *rho*=-0.09, *P*<0.0001).

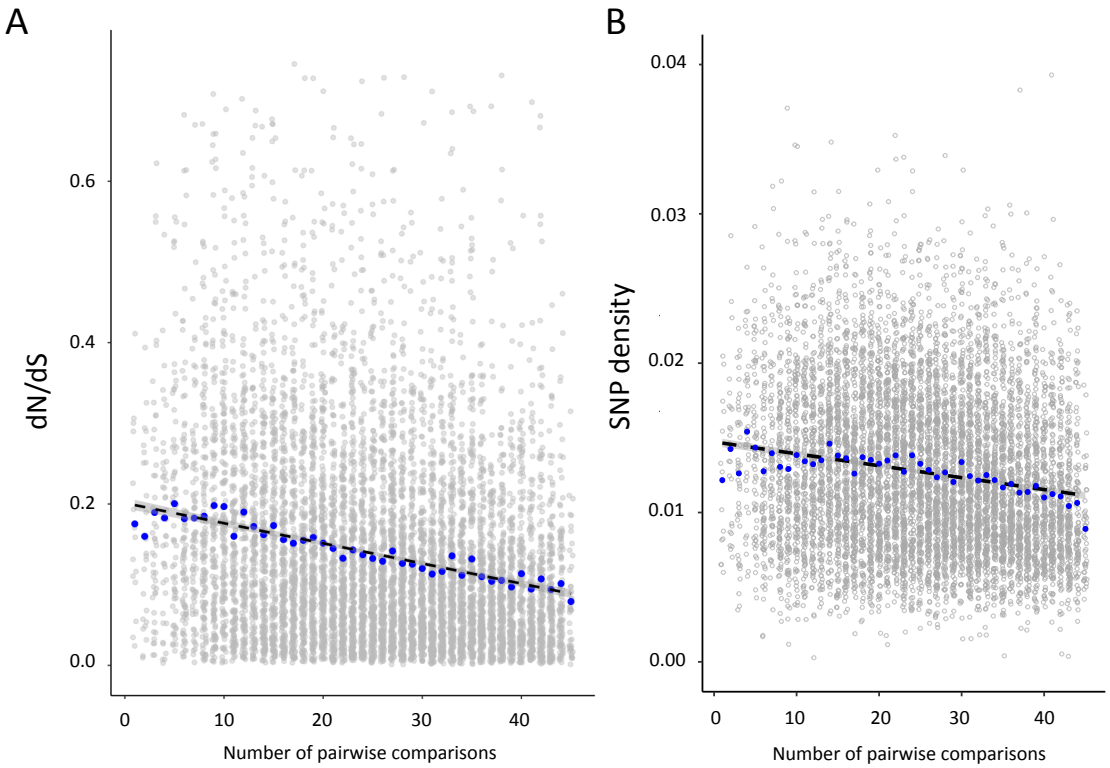


Figure 3. A. Genes in more pairwise consensus modules show significantly lower dN/dS values (Pairwise Spearman's rank correlation *rho* = -0.22, $P < 2.2 \times 10^{-16}$; Partial Spearman *rho* = -0.11, $P < 2.2 \times 10^{-16}$). B. Genes in more pairwise consensus modules also show significantly lower SNP density (Pairwise Spearman's rank correlation *rho* = -0.16, $P < 2.2 \times 10^{-16}$; Partial Spearman *rho*=-0.039, $P = 0.00036$).

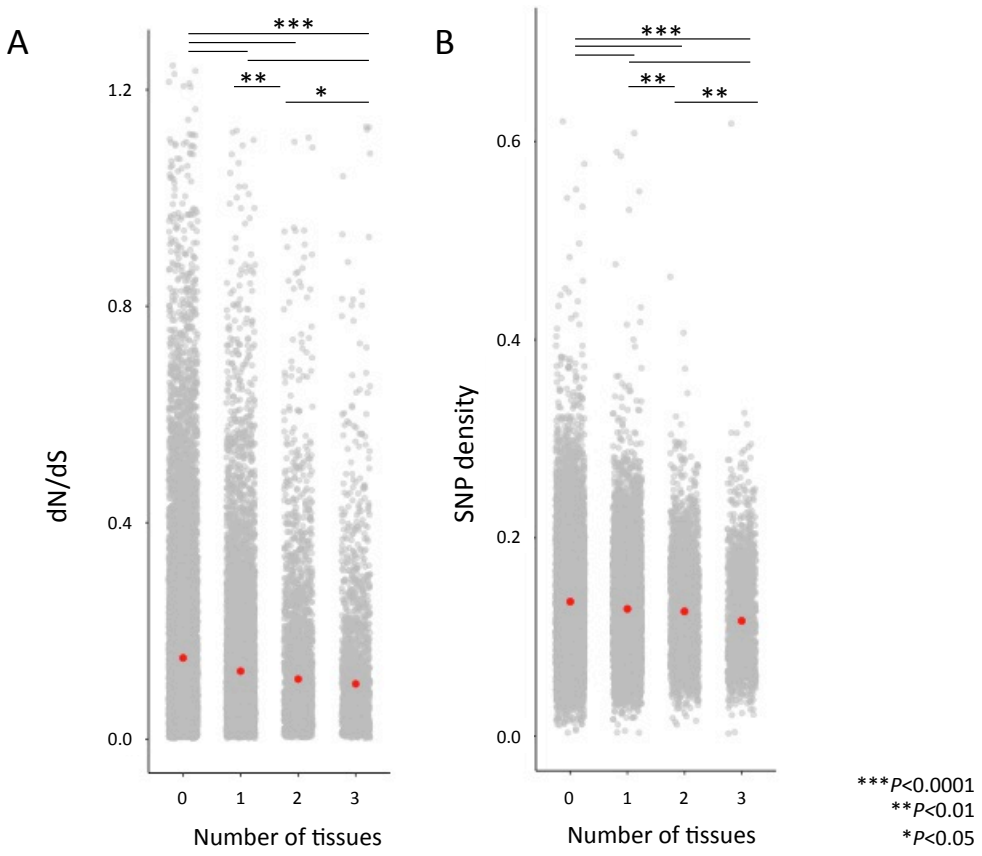


Figure 4. Genes that are "hubs" in more tissues are associated with lower dN/dS values (A) and lower SNP density (B). Comparisons were performed with permutation tests.

5.7. Chapter 5 Supplemental tables

Table S1. The relationship between gene expression and connectivity within tissues.

Tissue	*rho*[1]	*P*
Thyroid	0.31	< 2.2e-16
Lung	0.38	< 2.2e-16
Spleen	0.49	< 2.2e-16
Muscle	0.48	< 2.2e-16
Brain	0.55	< 2.2e-16
Testis	0.55	< 2.2e-16
Kidney	0.62	< 2.2e-16
Gut	0.63	< 2.2e-16
Liver	0.55	< 2.2e-16
Heart	0.44	< 2.2e-16

[1]Spearman's rank correlation *rho*

Table S2. Pairwise comparisons of gene connectivity between tissues (Spearman's rank correlation).

	Lung	Kidney	Muscle	Liver	Thyroid	Testis	Brain	Gut	Heart	Spleen
Lung		0.25	0.18	0.12	0.2	0.13	0.14	0.17	0.24	0.14
Kidney			0.27	0.29	0.35	0.23	0.23	0.33	0.29	0.21
Muscle				0.21	0.3	0.13	0.17	0.23	0.28	0.13
Liver					0.25	0.06	0.13	0.24	0.25	0.14
Thyroid						0.16	0.17	0.27	0.4	0.17
Testis							0.15	0.15	0.13	0.12
Brain								0.15	0.16	0.14
Gut									0.23	0.19
Heart										0.18

Table S3. Number of genes with a SNP that could be tested for allele-specific expression (ASE)

Tissue	Number of samples	Number of genes that could be tested for ASE
Spleen	18	8,389
Brain	20	7,973
Heart	20	6,351
Kidney	20	7,464
Gut	14	6,933
Thyroid	20	5,756
Testis	20	8,673
Muscle	18	5,347
Liver	20	6,151
Lung	14	9,048

Table S4. Number of genes with allele-specific expression.

	FDR	
Tissue	q <0.1	q<0.05
Spleen	1057	884
Brain	840	659
Heart	729	582
Kidney	1000	784
Gut	750	639
Thyroid	403	354
Testis	1255	971
Muscle	563	457
Liver	956	814
Lung	892	730

Table S5. Tissue-specific transcription factor enriched for tissue-specific mutant phenotypes

Tissue	Tissue-specific mutant phenotypes	*q*-value
Brain	Abnormal brain size	1.725×10^{-4}
	Abnormal brain weight	0.003
	Abnormal cerebellar cortex morphology	1×10^{-3}
Testis	Abnormal testis weight	0.027
	Small testis	0.01
	Abnormal seminiferous tubule size	0.016
Liver	Abnormal liver morphology	0.01
	Abnormal liver size	0.018
Spleen	Abnormal spleen size	0.002
	Small spleen	0.002
	Enlarged Spleen	0.017
	Abnormal splenocyte apoptosis	0.004
	Abnormal spleen physiology	0.006
	Abnormal splenocyte physiology	0.002
	Abnormal splenocyte proliferation	0.001

5.8. Chapter 5 Supplemental Figures

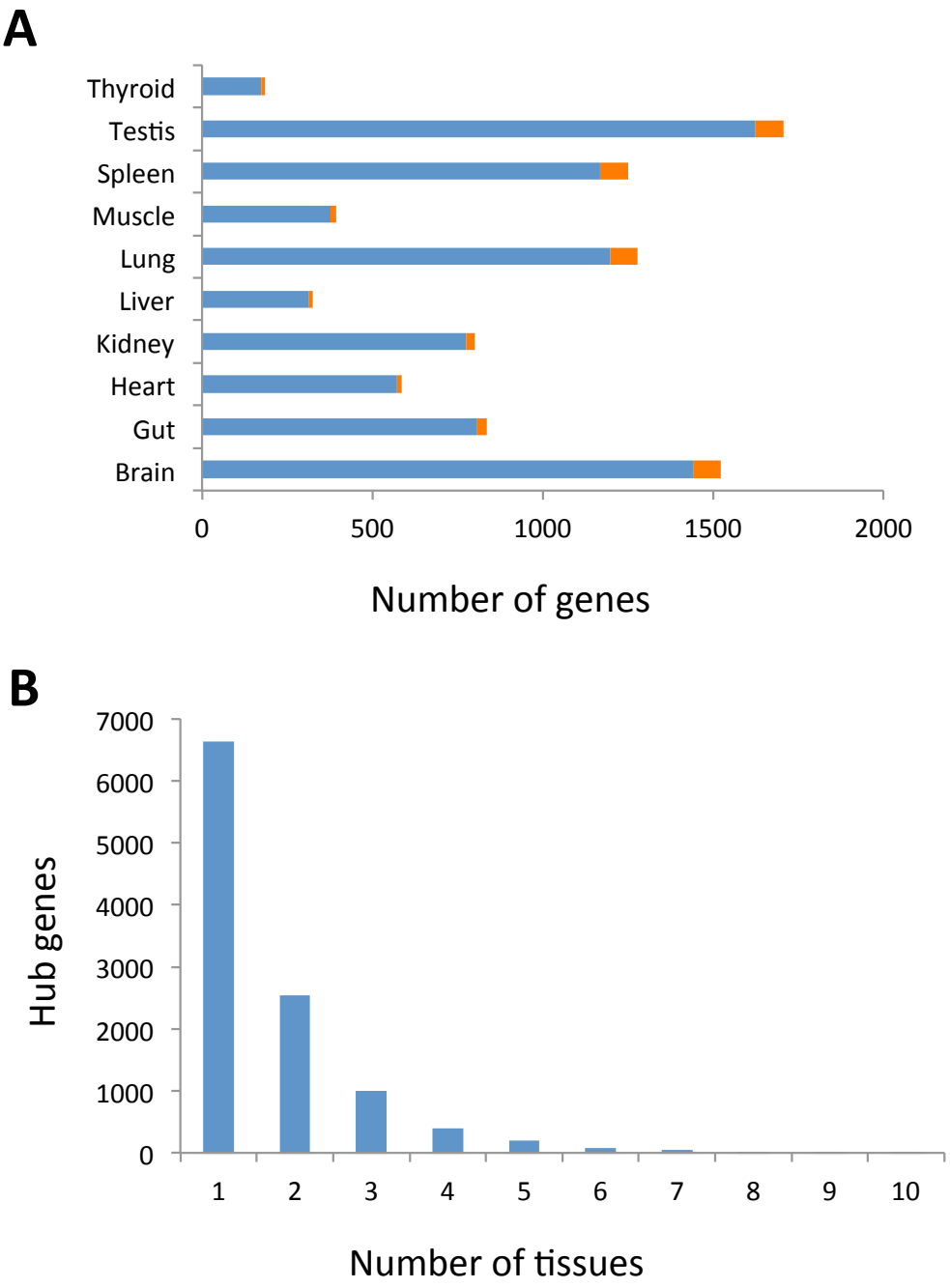
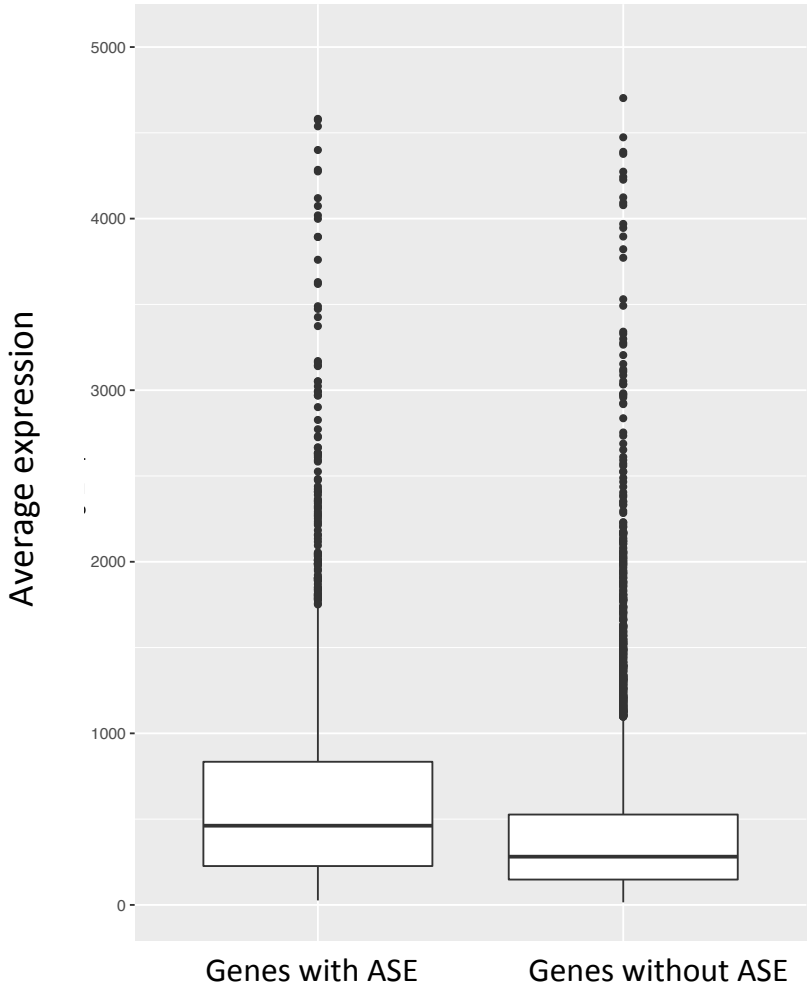


Figure S1. (A) The number of genes in each tissue that were classified as tissue-specific. In orange are genes that encode transcription factors. (B) The number of hub genes that are found across different numbers of tissues

Figure S2. Genes for which we could detect allele-specific expression have higher expression on average (permutation test, $p<0.0001$).



References

Albert R, Jeong H, Barabasi AL. 2000. Error and attack tolerance of complex networks. *Nature* 406:378–82.

Anders S, Pyl PT, Huber W. 2015. HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166-9.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437:1149–1152.

Arribas J, Esselens C. 2009. ADAM17 as a therapeutic target in multiple diseases. *Curr Pharm Des* 15:2319-2335.

Artieri CG, Fraser HB. 2014. Evolution at two levels of gene expression in yeast. *Genome Res.* 24, 411–421.

Ashton KG, Tracy MC, Queiroz AD. 2000. Is Bergmann's rule valid for mammals? *Am Nat.* 156: 390-415.

Bader DM, Wilkening S, Lin G, Tekkedil MM, Dietrich K, Steinmetz LM, Gagneur J. 2015. Negative feedback buffers effects of regulatory variants. *Mol Sys Biol.* 11: 785.

Balcova M, Faltusova B, Gergelits V, Bhattacharyya T, Mihola O, Trachtulec Z, Knopf C, Fotopulosova V, Chvatalova I, Gregorova S, Forejt J. 2016. Hybrid sterility locus on chromosome X controls meiotic recombination rate in mouse. *PLoS Genet.* 12: e1005906.

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2015. Impact of regulatory variation from RNA to protein. *Science* 347: 664–667.

Bateson W. 1909. Heredity and variation in modern lights. Pp. 85–101 in A. C. Steward, ed. Darwin and modern science. Cambridge Univ. Press, Cambridge, U.K.

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nature reviews genetics 5*:101.

Barbash DA, Siino DF, Tarone AM, Roote J. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. *Proc Natl Acad Sci USA* 100: 5302–5307.

Barrière A, Gordon KL, Ruvinsky I. 2012. Coevolution within and between regulatory loci can preserve promoter function despite evolutionary rate acceleration. *PLoS Genet*. 8: e1002961.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, De Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327: 836-840.

Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science* 326: 1538-1541.

Bedford T and Hartl DL. 2009. Optimization of gene expression by natural selection. *Proc Natl Acad Sci USA* 106: 1133-1138.

Bell GD, Kane NC, Rieseberg LH, Adams KL. 2013. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol and Evol.* 5: 1309-1323.

Bergmann C. 1847. Ueber die Verhaltnisse der warmeokonomie der thiere zu ihrer

grosse. Gottinger Studien 3: 595–708.
Bhattacharyya T, Gregorova S, Mihola O, Anger M, Sebestova J, Denny P, Simecek P, Forejt J. 2013. Mechanistic basis of infertility of mouse intersubspecific hybrids. *Proc Natl Acad Sci USA* 110: E468-E477.
Bhattacharyya T, Reifova R, Gregorova S, Simecek P, Gergelits V, Mistrik M, Martincova I, Pialek J, Forejt J. 2014. X chromosome control of meiotic chromosome synapsis in mouse inter-subspecific hybrids. *PLoS Genet*. 10: e1004088.
Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, and the Mouse Genome Database Group. 2017. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucl Acids Res.* 45: D723-D729.
Boeva V, Popova T, Bleakley K, Chiche P, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next generation sequencing data. *Bioinformatics* 28: 423-425.
Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27: 268-269.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* 30: 2114-20.
Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, Weigel D. 2007. Autoimmune response as a mechanism for a Dobzhansky–Muller-type incompatibility syndrome in plants. *PLoS Biol.* 5: e236.
Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478:343-348.
Brideau NJ, Barbash DA. 2011. Functional conservation of the *Drosophila* hybrid incompatibility gene *Lhr*. *BMC Evol Biol.* 11:57.
Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang XU, Barbash DA. 2006. Two Dobzhansky–Muller genes interact to cause hybrid lethality in *Drosophila*. *Science* 314: 1292–1295.
Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet.* 84: 210-223.
Browning SR, Browning BL. 2011. Population structure can inflate SNP-based heritability estimates. *Am J Hum Genet.* 89: 191.
Britton-Davidian J, Fel-Clair F, Lopez J, Alibert P, Boursot P. 2005. Postzygotic isolation between the two European subspecies of the house mouse: estimates from fertility patterns in wild and laboratory bred hybrids. *Biol J Linnean Soc*. 84: 379-393.
Bryk J, Tautz D. 2014. Copy number variants and selective sweeps in natural populations of the house mouse (*Mus musculus domesticus*). *Front in Genet.* 5: 153.
Buckland PR. 2004. Allele-specific gene expression differences in humans. *Hum Mol Gen.* 13: R255-R260.
Bullard JH, Mostovoy Y, Dudoit S, Brem RB. 2010. Polygenic and directional

regulatory evolution across pathways in Saccharomyces. *Proc Natl Acad Sci USA* 107: 5058-5063.
Calabrese JM, Starmer J, Schertzer MD, Yee D, Magnuson T. 2015. A Survey of Imprinted Gene Expression in Mouse Trophoblast Stem Cells. *G3* 5: 751-759.
Campbell P, Good JM, Nachman MW. 2013. Meiotic sex chromosome inactivation is disrupted in sterile hybrid male house mice. *Genetics* 193: 819-828.
Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134: 25-36.
Castillo-Davis CI, Kondrashov FA, Hartl DL. Kulathinal RJ. 2004. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res.* 14: 802–811.
Cattani MV, Presgraves DC. 2012. Incompatibility between X chromosome factor and pericentric heterochromatic region causes lethality in hybrids between *Drosophila melanogaster* and its sibling species. *Genetics* 191:549-559.
Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SS, Demougin P, Gattiker A, Moore J, Patard JJ, Wolgemuth DJ, Jégou B, Primig M. 2007. The conserved transcriptome in human and rodent male gametogenesis. *Proc Natl Acad Sci USA* 104: 8346–8351.
Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, Myers RM. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327: 302–305.
Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7.
Chen P, Wang F, Feng J, Zhou R, Chang Y, Liu J, Zhao Q. 2017. Co-expression network analysis identified six hub genes in association with metastasis risk and prognosis in hepatocellular carcinoma. *Oncotarget* 8:48948.
Chigurapati S, Miller W, Lynch VJ. 2018. Relaxed constraint and thermal desensitization of the cold-sensing ion channel TRPM8 in mammoths. bioRxiv 1:397356.
Colburn RW, Lubin ML, Stone DJ Jr, Wang Y, Lawrence D, D'Andrea MR, *et al.* 2007. Attenuated cold sensitivity in TRPM8 null mice. *Neuron* 54: 379–86.
Coolon JD, Stevenson KR, McManus CJ, Yang B, Graveley BR, Wittkopp PJ. 2015. Molecular mechanisms and evolutionary processes contributing to accelerated divergence of gene expression on the Drosophila X chromosome. *Mol Biol Evol.* 10: 2605-2615.
Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. *Genome Res.* 24: 797-808.
Combes MC, Hueber Y, Dereeper A, Rialle S, Herrera JC, Lashermes P. 2015. Regulatory divergence between parental alleles determines gene expression patterns in hybrids. *Genome Biol and Evol.* 7:1110-1121.
Consortium TWTCC. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464: 713–20.
Cowles CR, Hirschhorn JN, Altshuler D, Lander ES. 2002. Detection of regulatory

variation in mouse genes. *Nature Genet.* 32: 432-437.
Crawford DL, Segal JA, Barnett JL. 1999. Evolutionary analysis of TATA-less proximal promoter function. *Mol Biol Evol.* 16: 194–207.
Crowley JJ, Zhabotynsky V, Sun W, Huang S, Pakatci IK, Kim Y, Wang JR, Morgan AP, Calaway JD, Aylor DL, Yun Z. 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genet*. 47: 353–360.
Coyne JA. 1992. Genetics and speciation. *Nature* 355:511-515.
Coyne JA, Orr HA. 1989. Two rules of speciation. in Speciation and its consequences (D. Otte and J.A. Endler, eds. Sinauer, Inc.), pp. 180-207.
Coyne JA, Orr HA. 2004. *Speciation*. Sinauer & Associates, Sunderland, Massachusetts.
Cutter AD. 2012. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. *Trends Ecol Evolut.* 27: 209-218.
Dandine-Roulland C, Bellenguez C, Debette S, Amouyel P, Génin E, Perdry H. 2016. Accuracy of heritability estimations in presence of hidden population stratification. *Sci Rep.* 6: 26471.
Davidson JH, Balakrishnan CN. 2016. Gene regulatory evolution during speciation in a songbird. *G3* 6: 1357-1364.
Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R, Preece C. 2016. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* 530: 171-176.
Dean R, Harrison PW, Wright AE, Zimmer F, Mank JE. 2015. Positive selection underlies Faster-Z evolution of gene expression in birds. *Mol Biol Evol.* 64: 663–674.
Denby CM, Im JH, Richard CY, Pesce CG, Brem RB. 2012. Negative feedback confers mutational robustness in yeast transcription factor regulation. *Proc Natl Acad Sci USA* 109: 3874-3878.
Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nature Genet.* 37: 544-548.
Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PloS one* 7: e30377.
Din W, Anand R, Boursot P, Darviche D, Dod B, Jouvin-Marche E, Orth A, Talwar GP, Cazenave P-A, Bonhomme F. 1996. Origin and radiation of the house mouse: Clues from nuclear genes. *J Evolution Biol.* 9: 519–539.
Dhaka A, Murray AN, Mathur J, Earley TJ, Petrus MJ, Patapoutian A. 2007. TRPM8 is required for cold sensation in mice. *Neuron* 54: 371–8.
Dittmar EL, Oakley CG, Conner JK, Gould BA, Schemske DW. 2016. Factors influencing the effect size distribution of adaptive substitutions. *Proc R Soc B* 283: 20153065.
Dobzhansky T. 1937. Genetics and the Origin of Species. Columbia University Press: New York.
Dod B, Jermiin LS, Boursot P, Chapman VH, Nielsen JT, Bonhomme F. 1993. Counterselection on sex chromosomes in the *Mus musculus* European hybrid zone. *J Evolution Biol.* 6: 529-546.

Dover GA, Flavell, RB. 1984. Molecular coevolution: DNA divergence and the maintenance of function. *Cell* 38: 622-623.
Duan J, Zhang JG, Deng HW, Wang YP. 2013. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS one 8*: e59128.
Dzur-Gejdosova M, Simecek P, Gregorova S, Bhattacharyya T, Forejt J. 2012. Dissecting the genetic architecture of F1 hybrid sterility in house mice. *Evolution* 66:3321-3335.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists*. BMC Bioinformatics* 10:48.
Eidsaa M, Stubbs L, Almaas E. 2017. Comparative analysis of weighted gene co-expression networks in human and mouse. *PloS One 12*: e0187611.
Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HH, Lu MY, Wu SH, Li WH. 2010. Natural selection on *cis* and *trans* regulation in yeasts. *Genome Res.* 2: 826-836.
Endler JA. *Geographic variation, speciation, and clines* (No. 10). Princeton University Press 1977.
Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. 2015. The Mouse Genome Database Group. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res*. 28: D726-36.
Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A. 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics 13*:397-406.
Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 312: 276-279.
Fear JM, León-Novelo LG, Morse AM, Gerken AR, Van Lehman K, Tower J, Nuzhdin SV, McIntyre LM2016. Buffering of Genetic Regulatory Networks in *Drosophila melanogaster*. *Genetics* 203: 1177–1190.
Ferguson J, Gomes S, Civetta A. 2013. Rapid male-specific regulatory divergence and down regulation of spermatogenesis genes in *Drosophila* species hybrids. *PLoS one* 8: e61575.
Fiorentino L, Vivanti A, Cavalera M, Marzano V, Ronci M, Fabrizi M, Menini S, Pugliese G, Menghini R, Khokha R, Lauro R. 2010. Increased tumor necrosis factor α–converting enzyme activity induces insulin resistance and hepatosteatosis in mice. *Hepatology* 51: 103-110.
Fontanillas P, Landry CR, Wittkopp PJ, Russ C, Gruber JD, Nusbaum, Hartl DL. 2010. Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Mol Ecol.* 19: 212–227.
Forejt J. 1985. Chromosomal and genic sterility of hybrid type in mice and men. *Exp Clin Immunogenet.* 2: 106–119.
Forejt J. 1996. Hybrid sterility in the mouse. *Trends Genet.* 12: 412–417.
Forejt J, Iványi P. 1974. Genetic studies on male sterility of hybrids between

laboratory and wild mice (*Mus musculus* L.). *Genet Res.* 24: 189-206.
Forejt J, Pialek J, Trachtulec Z. 2012. Hybrid male sterility genes in the mouse subspecific crosses. *Evolution of the house mouse*, 482-503.
Foster F, Collard M. 2013. A reassessment of Bergmann's rule in modern humans. *PloS ONE* 8: e72269.
Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol. 3*:11.
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science 296*:750-752.
Fraser HB. 2013. Gene expression drives local adaptation in humans. *Genome Res*. 23: 1089-1096.
Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, Schadt EE. 2011. Systematic detection of polygenic *cis*-regulatory evolution. *PLoS Genet.* 7:e1002023.
Fraser HB, Moses A, Schadt EE. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci USA* 107:2997.
Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol.* 30: 1687-1699.
Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, Korneliussen, TS, Gerbault P, Skotte L, Linneberg A, Christensen C. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349: 1343-1347.
Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet* 7: e1002355.
Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN. 2002. The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229.
Gattiker A, Niederhauser-Wiederkehr C, Moore J, Hermida L, Primig M. 2007. The GermOnline cross-species systems browser provides comprehensive information on genes and gene products relevant for sexual reproduction. *Nucleic Acids Res.* 35: D457–462.
Gelling RW, Yan W, Al-Noori S, Pardini A, Morton GJ, Ogimoto K, Schwartz MW, Dempsey PJ. 2008. Deficiency of TNFα converting enzyme (TACE/ADAM17) causes a lean, hypermetabolic phenotype in mice. *Endocrinology* 149: 6053-6064.
Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, Farber CR, Sinsheimer J, Kang HM, Furlotte N, Park CC. 2011. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* 7: e1001393.
Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* 2: e130.
Gibson G, Riley-Berger R, Harshman L, Kopp A, Vacha S, Nuzhdin S, Wayne M. 2004.

Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics* 167: 1791-1799.
Gilad Y, Oshlack A, Rifkin, SA. 2006. Natural selection on gene expression. *Trends Genet.* 22: 456-461.
Girirajan S, Campbell CD, Eichler EE. 2011. Human copy number variation and complex genetic disease. *Annu. Rev. Genet.* 45:203-226.
Genissel A, McIntyre LM, Wayne ML, Nuzhdin SV. *Cis* and *trans* regulatory effects contribute to natural variation in transcriptome of *Drosophila melanogaster*. *Mol Biol and Evol.* 25: 101-110.
Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, Brazma A, Odom DT, Marioni JC. 2012. Extensive compensatory *cis-trans* regulation in the evolution of mouse gene expression. *Genome Res.* 22: 2376-2384.
Good JM, Giger T, Dean MD, Nachman MW. 2010. Widespread over-expression of the X chromosome in sterile F1 hybrid mice. *PLoS Genet.* 6: e1001148.
Good JM, Handel MA, Nachman MW. 2008a. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* 62: 50-65.
Good JM, Dean MD, Nachman MW. 2008b. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* 179: 2213-2228.
Gomes S, Civetta A. 2015. Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Sci Rep.* 5: 11976.
Gordon KL, Ruvinsky I. 2012. Tempo and mode in evolution of transcriptional regulation. *PLoS Genet.* 8: e1002432.
Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, Ley TJ. 2007. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet.* 3: e3.
Graze RM, McIntyre LM, Main BJ, Wayne ML, Nuzhdin SV. 2009. Regulatory divergence in *Drosophila melanogaster* and *D. simulans*, a genomewide analysis of allele-specific expression. *Genetics* 183: 547-561.
Gregorova S, Mňuková-Fajdelová M, Trachtulec Z, Čapková J, Loudova M, Hoglund M, Hamvas R, Lehrach H, Vincek V, Klein J, Forejt J. 1996. Sub-milliMorgan map of the proximal part of mouse Chromosome 17 including the hybrid sterility 1 gene. *Mammalian Genome 7*: 107-113.
Gruber JD, Long AD. 2009. *Cis*-regulatory variation is typically polyallelic in Drosophila. *Genetics* 181: 661-670.
Haerty W, Singh RS. 2006. Gene regulation divergence is a major contributor to the evolution of Dobzhansky–Muller incompatibilities between species of *Drosophila*. *Mol Biol and Evol.* 23: 1707-1714.
Haldane JBS. 1922. Sex ratio and unisexual sterility in hybrid animals. *J Genet.* 12: 101-109.
Hagiwara Y, Hirai M, Nishiyama K, Kanazawa I, Ueda T, Sakaki Y, Ito T. 1997. Screening for imprinted genes by allelic message display: identification of a paternally expressed gene impact on mouse chromosome 18. *Proc Natl Acad Sci USA* 94: 9249-9254.
Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A.

2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* 4: e32.
Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet*. 4:e1000106.
Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, Babiker H, Linnenbrink M, Montero I, Scavetta R, Abai MR. 2016. Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus. Sci Data* 3: 160075.
Hedrick PW. 1981. The establishment of chromosomal variants. *Evolution* 35, 322-332.
Hodgins-Davis A, Rice DP, Townsend JP. 2015. Gene expression evolves under a House-of-Cards model of stabilizing selection. *Mol Biol Evol.* 32:2130-2140.
Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61:995-1016.
Holloway AK, Lawniczak MKN, Mezey JG, Begun DJ, Jones CD. 2007. Adaptive gene expression divergence inferred from population genomics. *PLoS Genet*. 3: e187.
Hou J, Wang X, McShane E, Zauber H, Sun W, Selbach M, Chen W. 2015 Extensive allele-specific translational regulation in hybrid mice. *Mol Syst Biol.* 11: 825.
Hovatta I, Tennant RS, Helton R, Marr RA, Singer O, Redwine JM, Ellison JA, Schadt EE, Verma IM, Lockhart DJ, Barlow C. 2005. Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. *Nature* 438: 662.
Hu Z, Snitkin ES, DeLisi C. 2008. VisANT: an integrative framework for networks in systems biology. *Brief Bioinform.* 9: 317–325.
Hutter P and Ashburner M. 1987. Genetic rescue of inviable hybrids between *Drosophila melanogaster* and its sibling species. *Nature* 327: 331–333.
Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218–223.
International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 4:237.
Iskow RC, Gokcumen O, Lee C. 2012. Exploring the role of copy number variants in human adaptation. *Trends in Genet.* 28: 245-57.
Janoušek V, Wang L, Luzynski K, Dufková P, Vyskočilová MM, Nachman MW, Munclinger P, Macholán M, Piálek J, Tucker PK. 2012. Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. *Mol Ecol*. 21: 3032-3047.
Jeon JP, Shim SM, Nam HY, Ryu GM, Hong EJ, Kim HL, *et al.* 2010. Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. *BMC Genomics* 11: 426.
Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. 2008. The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. *Cell* 132: 783–793.

Jenkins DL, Ortori CA, Brookfield JFY. 1995. A test for adaptive change in DNA sequences controlling transcription. *Proc R Soc Lond B* 261: 203-207.
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55-61.
Johnson NA, Lachance J. 2012. The genetics of sex chromosomes: evolution and implications for hybrid incompatibility. *The Year in Evolutionary Biology. New York Academy of Sciences* 1256: e1-e22.
Johnson NA, Porter AH. 2000. Rapid speciation via parallel, directional selection on regulatory genetic pathways. *J Theor Biol.* 205: 527-542.
Johnson NA, Porter AH. 2007. Evolution of branched regulatory genetic pathways: directional selection on pleiotropic loci accelerates developmental system drift. *Genetica* 129: 57-70.
Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. 2017. The relationship between selection, network connectivity, and regulatory variation within a population of *Capsella grandiflora*. *Genome Biol Evol. 9*:1099-1109.
Junaid MA, Kowal D, Barua M, Pullarkat PS, Sklower Brooks S*, et al.* 2004. Proteomic studies identified a single nucleotide polymorphism in glyoxalase I as autism susceptibility factor. *Am J Med Genet A* 131: 11–17.
Junyent M, Parnell LD, Lai CQ, Arnett DK, Tsai MY, Kabagambe EK, Straka RJ, Province M, An P, Smith E, Lee YC. 2010. ADAM17_i33708A>G polymorphism interacts with dietary n-6 polyunsaturated fatty acids to modulate obesity risk in the Genetics of Lipid Lowering Drugs and Diet Network study. *Nutr Metab Cardiovasc Dis.* 20: 698-705.
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709-1723.
Kayserili MA, Gerrard DT, Tomancak P, Kalinka AT. 2012. An excess of gene expression divergence on the X chromosome in *Drosophila* embryos: implications for the faster-X hypothesis. *PLoS Genet*. 8: e1003200.
Key FM, Abdul-Aziz MA, Mundry R, Peter BM, Sekar A, D'Amato M, Dennis MY, Schmidt JM, Andrés AM. 2018. Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLoS Genet.* 14:e1007298.
Khatri BS, Goldstein RA. 2015. Simple biophysical model predicts faster accumulation of hybrid incompatibilities in small populations under stabilizing selection. *Genetics* 201: 1525-1537.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.
King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107-116.
Kingman JFC. 1978. Simple model for balance between selection and mutation. *J Appl Probab.* 15: 1-12.
Kobayashi M, Suzuki M, Ohno T, Tsuzuki K, Taguchi C, Tateishi S, Kawada T, Kim YI, Murai A, Horio F. 2016. Detection of differentially expressed candidate genes for a fatty liver QTL on mouse chromosome 12. *BMC Genet.* 17: 73.

Kohn MH, Fang S, Wu C-I. 2004. Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes. *Mol Biol Evol* 21:374–383.
Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15: 356.
Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR. 2007. The UCSC genome browser database: update 2007. *Nucleic Acids Res.* 35: D668-73.
Kummerfeld SK, Teichmann SA. 2006. DBD: a transcription factor prediction database. *Nucleic Acids Res.* 34: D74-D81.
Kuo D, Licon K, Bandyopadhyay S, Chuang R, Luo C, Catalana J, Ravasi T, Tan K, Ideker T. 2010. Coevolution within a transcriptional network by compensatory trans and *cis* mutations. *Genome Res.* 20: 1672-1678.
Kuroiwa Y, Kaneko-Ishino T, Kagitani F, Kohda T, Li LL, Tada M, Suzuki R, Yokoyama M, Shiroishi T, Wakana S. 1996. *Peg3* imprinted gene on proximal chromosome 7 encodes for a zinc finger protein. *Nature Genet.* 12: 186-190.
Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19: 1639-1645.
Lande R. 1979. Effective deme sizes during long-term evolution estimated from rates of chromosomal rearrangement. *Evolution* 33: 234-251.
Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL. 2005. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics* 171:1813-1822.
Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559.
Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. 2012. *Nat Methods* 9: 357-359.
Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, Smith KL, Schadt EE, Nachman MW. 2007. Linkage disequilibrium in wild mice. *PLoS Genet.* 3:e144.
Larson EL, Vanderpool D, Keeble S, Zhou M, Sarver BA, Smith AD, Dean MD, Good JM. 2016. Contrasting Levels of Molecular Evolution on the Mouse X Chromosome. *Genetics* 204:1841-57.
Linnen CR, Poh YP, Peterson BK, Barrett RD, Larson JG, Jensen JD, Hoekstra HE. 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339: 1312-1316.
Le Gall, SM, Maretzky T, Issuree PD, Niu XD, Reiss K, Saftig P, Khokha R, Lundell D, Blobel, CP. 2010. ADAM17 is regulated by a rapid and reversible mechanism that controls access to its catalytic site. *J Cell Sci.* 123: 3913-3922.
Lee HY, Chou JY, Cheong L, Chang NH, Yang SY, Leu JY. 2008. Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell* 135: 1065-1073.
Lee DC, Sunnarborg SW, Hinkle CL, Myers TJ, Stevenson M, Russell W, Castner BJ, Gerhart, MJ, Paxton RJ, Black RA, Chang A. 2003. TACE/ADAM17 processing of EGFR ligands indicates a role as a physiological convertase. *Ann NY Acad Sci.* 995: 22-38.

Lemos B, Araripe LO, Fontanillas P, Hartl DL. 2008. Dominance and the evolutionary accumulation of *cis*-and *trans*- effects on gene expression. *Proc Natl Acad Sci USA* 105:14471-14476.
Lemos B, Meiklejohn CD, Cáceres M, Hartl DL. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* 59:126-137.
Lemmon ZH, Bukowski R, Sun Q, Doebley JF. 2014. The Role of *cis* Regulatory Evolution in Maize Domestication. *PLoS Genet.* 10:e1004745.
Swain Lenz D, Riles L, Fay JC. 2014. Heterochronic meiotic misexpression in an interspecific yeast hybrid. *Mol Biol Evol.* 31:1333-1342.
Li Y and Sasaki H. 2011. Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. *Cell Res.* 21:466-473.
Liénard MA, Araripe LO, Hartl DL. 2016. Neighboring genes for DNA-binding proteins rescue male sterility in *Drosophila* hybrids. *Proc Natl Acad Sci USA* 113: E4200-E4207.
Listgarten J, Kadie C, Schadt EE, Heckerman D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci USA* 107: 16465–16470.
Lim FL, Soulez M, Koczan D, Thiesen HJ, Knight JC. 1998. A KRAB-related domain and a novel transcription repression domain in proteins encoded by SSX genes that are disrupted in human sarcomas. *Oncogene* 17: 2013–2018.
Llopart A. 2012. The rapid evolution of X-linked male-biased gene expression and the large-X effect in *Drosophila yakuba*, *D. santomea*, and their hybrids. *Mol Biol Evol.* 29: 3873-3886.
Locke ME, Milojevic M, Eitutis ST, Patel N, Wishart AE, Daley M, Hill KA. 2015. Genomic copy number variation in Mus musculus. *BMC Genomics* 16:497.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550.
Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403: 564-567.
Lynch CB. 1992. Clinal variation in cold adaptation in *Mus domesticus*: verification of predictions from laboratory populations. *Am Nat.* 139: 1219-1236.
Lynch VJ, Bedoya-Reina OC, Ratan A, Sulak M, Drautz-Moses DI, Perry GH, Miller W, Schuster SC. 2015. Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the Arctic. *Cell Rep.* 12:217-28.
Lynch VJ, Tanzer A, Wang Y, Leung FC, Gellersen B, Emera D, Wagner GP. 2008. Adaptive changes in the transcription factor HoxA-11 are essential for the evolution of pregnancy in mammals. *Proc Natl Acad Sci USA* 105: 14928-14933.
MacDonald SJ, Long AD. 2005. Prospects for identifying functional variation across the genome. *Proc Natl Acad Sci USA* 102: 6614–6621.
Mack KL, Campbell P, Nachman MW. 2016. Gene regulation and speciation in house mice. *Genome Res.* 26: 451-461.
Mack KL, Ballinger MA, Phifer-Rixey M, Nachman MW. 2018. Gene regulation underlies environmental adaptation in house mice. *Genome Res.* 28:1636-1645.

MacNeil L, Walhout AJ. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.* 21: 645-657.
Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Gent. 13*:e1006402.
Masalia RR, Bewick AJ, Burke JM. 2017. Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PloS One 12*:e0182289.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297-1303.
Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. 2011. Read count approach for DNA copy number variants detection. *Bioinformatics* 28:470-8.
Maheshwari S, Barbash DA. 2011. The genetics of hybrid incompatibilities. *Annu. Rev. Genet.* 45: 331-355.
Maheshwari S, Barbash DA. 2012. *Cis*-by-*trans* regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene. *PLoS Genet.* 8: e1002597.
Malone JH, Chrzanowski TH, Michalak P. 2007. Sterility and gene expression in hybrid males of *Xenopus laevis* and *X. muelleri. PloS One* 2: e781.
Margolin JF, Friedman JR, Meyer WK, Vissing H, Thiesen HJ, Rauscher FJ. 1994. Krüppel-associated boxes are potent transcriptional repression domains. *Proc Natl Acad Sci USA* 91: 4509-4513.
Matos-Cruz V, Schneider ER, Mastrotto M, Merriman DK, Bagriantsev SN, Gracheva EO. 2017. Molecular prerequisites for diminished cold sensitivity in ground squirrels and hamsters. *Cell rep.* 21:3329-37.
Matsui Y, Tomaru U, Miyoshi A, Ito T, Fukaya S, Miyoshi H, Atsumi T, Ishizu A. 2014. Overexpression of TNF-α converting enzyme promotes adipose tissue inflammation and fibrosis induced by high fat diet. *Exp Mol Pathol.* 97: 354-358.
McKemy DD, Neuhausser WM, Julius D. 2002. Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature* 416: 52–8.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-1303.
McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp, PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* 20: 816-825.
McManus CJ, May GE, Spealman P, Shteyman A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* 24: 422–430.
Meisel RP, Malone JH, Clark AG. 2012. Faster-X evolution of gene expression in *Drosophila*. *PLoS Genet*. 8: e1003013.
Menghini R, Casagrande V, Menini S, Marino A, Marzano V, Hribal ML, Gentileschi P,

Lauro D, Schillaci O, Pugliese G, Sbraccia P. 2012. TIMP3 overexpression in macrophages protects from insulin resistance, adipose inflammation, and nonalcoholic fatty liver disease in mice. *Diabetes* 61: 454-462.
Menghini R, Menini S, Amoruso R, Fiorentino L, Casagrande V, Marzano V, Tornei F, Bertucci P, Iacobini C, Serino M, Porzio. 2009. Tissue inhibitor of metalloproteinase 3 deficiency causes hepatic steatosis and adipose tissue inflammation in mice. *Gastroenterology* 136: 663-672.
Menghini R, Fiorentino L, Casagrande V, Lauro R, Federici M. 2013. The role of ADAM17 in metabolic inflammation. *Atherosclerosis* 228: 12-17.
Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2016. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 45: D183-D189.
Michalak P, Noor MA. 2003. Genome-wide patterns of expression in Drosophila pure species and hybrid males. *Mol Biol Evol.* 20:1070–1076.
Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. 2009. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* 323:373-375.
Moehring AJ, Teeter KC, Noor MA. 2007. Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Mol Biol Evol.* 24:137-145.
Montgomery SB, Dermitzakis ET. 2011. From expression QTLs to personalized transcriptomics. *Nat Rev Genet.* 12: 277-282.
Moulos P and Hatzis P. 2014. Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Res.* 43: e25.
Muller HJ. 1940. Bearing of the *Drosophila* work on systematics. In: Huxley J (ed). The New Systematics. Clarendon Press: Oxford.
Muller HJ and Pontecorvo G. 1942. Recessive genes causing interspecific sterility and other disharmonies between *Drosophila melanogaster* and *simulans*. *Genetics* 27: 157.
Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327: 876-879.
Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF, Haqq AM, Shah SH, Arlotto, M, Slentz CA, Rochon J. 2009. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metab.* 9: 311-326.
Nica AC, Dermitzakis ET. 2013. Expression quantitative trait loci: present and future. *Phil Trans R Soc B* 368: 20120362.
Nowick K, Gernat T, Almaas E, Stubbs L. 2009. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci USA* 22358–22363.
Oka A, Aoto T, Totsuka Y, Takahashi R, Ueda M, Mita A. 2007. Disruption of genetic interaction between two autosomal regions and the X chromosome causes reproductive isolation between mouse strains derived from different subspecies. *Genetics* 175:185-197.

Oka A, Mita A, Sakurai-Yamatani N, Yamamoto H, Takagi N, Takano-Shimizu T, Toshimori K, Moriwaki K, Shiroishi T. 2004. Hybrid breakdown caused by substitution of the X chromosome between two mouse subspecies. *Genetics* 166: 913-924.
Oka A, Mita A, Takada Y, Koseki H, Shiroishi T. 2010. Reproductive isolation in hybrid mice due to spermatogenesis defects at three meiotic stages. *Genetics* 186:339-351.
Oka A, Shiroishi T. 2014. Regulatory divergence of X-linked genes and hybrid male sterility in mice. *Genes Genet Syst.* 89:99-108.
Oka A, Takada T, Fujisawa H, Shiroishi T. 2014. Evolutionarily diverged regulation of X-chromosomal genes as a primal event in mouse reproductive isolation. *PLoS Genet*. 10:e1004301.
Ono R, Shiura H, Aburatani H, Kohda T, Kaneko-Ishino T, Ishino F. 2003. Identification of a large novel imprinted gene cluster on mouse proximal chromosome 6. *Genome Res.* 13:1696-1705.
Ortíz-Barrientos D, Counterman BA, Noor MA. 2007. Gene expression divergence and the origin of hybrid dysfunctions. *Genetica* 129:71-81.
Orr HA. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* 139:1805-1813.
Orr HA. 1998. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* 52: 935-949.
Osada M, Kohn MH, Wu C-I. 2006. Genomic inference of *cis*- regulatory nucleotide polymorphism underlying gene expres- sion differences between Drosophila melanogaster mating races. *Mol Biol Evol.* 23:1585-1591.
Pagnamenta AT, Wing K, Akha ES, Knight SJ, Bölte S, Schmötzer G, Duketis E, Poustka F, Klauck SM, Poustka A, Ragoussis J. 2009. A 15q13. 3 microdeletion segregating with autism. *Eur J Hum Genet.* 17: 687.
Palmer ME, Feldman MW. 2009. Dynamics of hybrid incompatibility in gene networks in a constant environment. *Evolution* 63:418-431.
Palopoli MF, NH Patel. 1996. Neo-Darwinian developmental evolution: can we bridge the gap between pattern and process? *Curr Op in Gen and Dev.* 6: 502-508.
Parnell LD, Lee YC, Lai CQ. 2010. Adaptive genetic variation and heart disease risk. *Curr Opin Lipidol.* 21: 116.
Parvanov ED, Petkov PM, Paigen K. 2010. *Prdm9* controls activation of mammalian recombination hotspots. *Science* 327: 835-835.
Payseur BA. 2016. Genetic Links between Recombination and Speciation. *PLoS Genet.* 12: e1006066.
Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution* 58: 2064-2078.
Payseur BA, Nachman MW. 2005. The genomics of speciation: investigating the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus. Biol J Linn Soc.* 84: 523-534.
Peier AM, Moqrich A, Hergarden AC, Reeve AJ, Andersson DA, Story GM, *et al.* 2002. A TRP channel that senses cold stimuli and menthol. *Cell* 108:705–15.

Pezer Ž, Harr B, Teschke M, Babiker H, Tautz D. 2015 Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three conserved genes with major population-specific expansions. *Genome Res.* 25: 1114-1124.

Pierson E, Koller D, Battle A, Mostafavi S, GTEx Consortium. 2015. Sharing and specificity of co-expression networks across 35 human tissues. *PLOS Comput Biol. 11*:e1004220.

Phadnis N, Baker EP, Cooper JC, Frizzell KA, Hsieh E, De La Cruz AF, Shendure J, Kitzman JO, Malik HS.2015. An essential cell cycle regulation gene causes hybrid inviability in *Drosophila*. *Science* 350: 1552-1555.

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS. 2008. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet.* 82: 685-695.

Phadnis N, Orr HA. 2008. A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science* 323: 376-379.

Phifer-Rixey M, Bomhoff M, Nachman MW. 2014. Genome-Wide Patterns of Differentiation Among House Mouse Subspecies. *Genetics* 198: 283-297.

Phifer-Rixey M, Nachman MW. 2015. The Natural History of Model Organisms: Insights into mammalian biology from the wild house mouse *Mus musculus. eLife* 4: e05959.

Phifer-Rixey M, Bi K, Ferris KG, Sheehan MJ, Lin D, Mack KL, Keeble SM, Suzuki TA, Good JM, Nachman MW. 2018. The genomic basis of environmental adaptation in house mice. *PLoS Genet.* 14: e1007672.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh, GS, Myers RM, Feldman MW, Pritchard JK. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19: 826-837.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768–772.

Pinter SF, Colognori D, Beliveau BJ, Sadreyev RI, Payer B, Yildirim E, Wu CT, Lee JT. 2015. Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome. *Genetics* 200: 537-49.

Pirooznia M, Goes FS, Zandi PP. 2015. Whole-genome CNV analysis: advances in computational approaches. *Frontiers in genetics* 6:138.

Politi P, Minoretti P, Falcone C, Martinelli V, Emanuele E. 2006. Association analysis of the functional Ala111Glu polymorphism of the glyoxalase I gene in panic disorder. *Neurosci Lett.* 396: 163–166.

Porter AH, Johnson NA, Tulchinsky AY. 2016. Competitive binding of transcription factors drives Mendelian dominance in regulatory genetic pathways. arXiv preprint arXiv:1606.06668. Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat Rev Genet.* 11:175-180.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA. 2003. Adaptive evolution drives

divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* 423:715–719.
Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet.* 11: 459–463.
Prickett AR and Oakey RJ. 2012. A survey of tissue-specific genomic imprinting in mammals. *Mol Genet Genomics* 287:621-630.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841-842.
Ranz JM, Namgyal K, Gibson G, Hartl DL. 2004. Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans. Genome Res.* 14:373–379.
Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. *Nature Genet.* 33:138-144.
Sacco R, Papaleo V, Hager J, Rousseau F, Moessner R, *et al.* 2007. Case-control and family-based association studies of candidate genes in autistic disorder and its endophenotypes: TPH2 and GLO1. *BMC Med Genet.* 8: 11.
Sawamura K, Yamamoto MT, Watanabe TK. 1993. Hybrid lethal systems in the *Drosophila* melanogaster species complex. II. The Zygotic hybrid rescue (*Zhr*) gene of *D. melanogaster*. *Genetics* 133:307-313.
Satyaki PR, Cuykendall TN, Wei KH, Brideau NJ, Kwak H, Aruna S, Ferree PM, Ji S, Barbash DA. 2014. The *Hmr* and *Lhr* hybrid incompatibility genes suppress a broad range of heterochromatic repeats. *PLoS Genet*. 10: e1004240.
Schaefke B, Emerson JJ, Wang TY, Lu MY, Hsieh LC, Li WH. 2013. Inheritance of gene expression level and selective constraints on *trans*- and *cis*-regulatory changes in yeast. *Mol Biol and Evol.* 30:2121-33
Schrider DR, Begun DJ, Hahn MW. 2013. Detecting highly differentiated copy-number variants from pooled population sequencing. *Biocomputing 2013* 344-355.
Schrider DR, Hahn MW, Begun DJ. 2016. Parallel evolution of copy-number variation across continents in Drosophila melanogaster. *Mol Biol Evol. 33*: 1308-1316.
Shen SQ, Turro E, Corbo JC. 2014. Hybrid Mice Reveal Parent-of-Origin and *Cis*-and *Trans*- Regulatory Effects in the Retina. *PLoS One* 9:e109382.
Shi X, Ng DW, Zhang C, Comai L, Ye W, Chen ZJ. 2012. *Cis*-and *trans*-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat Commun.* 3:950.
Shirata M, Araye Q, Maehara K, Enya S, Takano-Shimizu T, Sawamura K. 2011. Two types of *cis-trans* compensation in the evolution of transcriptional regulation. *Proc Natl Acad Sci USA* 108:15276-15281.
Shirata M, Araye Q, Maehara K, Enya S, Takano-Shimizu T, Sawamura K. (2014) Allelic asymmetry of the Lethal hybrid rescue (*Lhr*) gene expression in the hybrid between *Drosophila melanogaster* and *D. simulans*: confirmation by using genetic variations of *D. melanogaster*. *Genetica* 142:43-48.
Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, Johansson M, Jaschob D,

Graczyk B, Shulman NJ, Wakefield J, Cooper SJ. 2013. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* 23:1496–1504.
Smith SD, Kawash JK, Karaiskos S, Biluck I, Grigoriev A. 2017. Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants. *DNA Research* 24:359-69.
Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, DeMeo DL, Quackenbush J, Glass K, Kuijjer ML. 2017. Understanding tissue-specific gene regulation. *Cell rep.* 21:1077-1088.
Stegle O, Parts L, Durbin R, Winn J. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol.* 6: e1000770.
Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc.* 7: 500-507.
Stern DL, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62: 2155-2177.
Stevenson KR, Coolon JD, Wittkopp PJ. 2013. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* 14: 536.
Storchová, R, Gregorová S, Buckiova D, Kyselova V, Divina P, Forejt J. 2004. Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm Genome* 15: 515-524.
Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249–55.
Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101:6062-6067.
Sugama T, Mukai T. 1993. A new imprinted gene cloned by a methylation-sensitive genome scanning method. *Nucleic acids res.* 21:5577-5582.
Sun W. 2012. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* 68: 1-11.
Sun W, Hu Y. 2014. Mapping of Expression Quantitative Trait Loci Using RNA-seq Data. In *Statistical Analysis of Next Generation Sequencing Data* (pp. 145-168). Springer, Cham.
Sundararajan V, Civetta A. 2011. Male sex interspecies divergence and down regulation of expression of spermatogenesis genes in *Drosophila* sterile hybrids. *J Mol Evol.* 72: 80-89
Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, von Mering C. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45: D362-68.
Tang S, Presgraves DC. 2009. Evolution of the Drosophila nuclear pore complex results in multiple hybrid incompatibilities. *Science* 323: 779–782.
Thomae AW, Schade GO, Padeken J, Borath M, Vetter I, Kremmer E, Heun P, Imhof A.

2013. A pair of centromeric proteins mediates reproductive isolation in *Drosophila* species. *Dev cell* 27: 412-424.
Ting CT, Tsaur SC, Wu ML, Wu CI.1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282: 1501-1504.
Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324: 659-662.
Trachtulec Z, Mňuková-Fajdelová M, Hamvas RM, Gregorová S, Mayer WE, Lehrach HR, Vincek V, Forejt J, Klein J. 1997. Isolation of candidate hybrid sterility 1 genes by cDNA selection in a 1.1 megabase pair region on mouse chromosome 17. *Mamm Genome* 8:312-316.
Trachtulec, Z. Mihola O, Vlcek C, Himmelbauer H, Pacces V, Forejt J. 2005. Positional cloning of the Hybrid sterility 1 gene: fine genetic mapping and evaluation of two candidate genes. *Biol J Linnean Soc.* 84, 637-641.
Trachtulec Z, Vlcek C, Mihola O, Gregorova S, Fotopulosova V, Forejt J. 2008. Fine haplotype structure of a chromosome 17 region in the laboratory and wild mouse. *Genetics* 178: 1777-1784.
True JR, Haag ES. 2001 Developmental system drift and flexibility in evolutionary trajectories. *Evol Dev.* 3: 109-119.
Tucker PK, Sage RD, Warner J, Wilson AC, Eicher EM. 1992. Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. *Evolution* 46: 1146-1163.
Tulchinsky AY, Johnson NA, Watt WB, Porter AH. 2014. Hybrid incompatibility arises in a sequence-based bioenergetic model of transcription factor binding. *Genetics* 198, 1155-1166.
Tulchinsky AY, Johnson NA, Porter AH. 2014b. Hybrid incompatibility despite pleiotropic constraint in a sequence-based bioenergetic model of transcription factor binding. *Genetics* 198: 1645-1654.
Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. 2015. The genetic architecture of gene expression levels in wild baboons. *eLife* 4: e04729.
Turelli M. 1984 Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theor Popul Biol.* 25: 138-193.
Turner JM 2007. Meiotic sex chromosome inactivation. *Development* 134: 1823-1831.
Turner LM, White MA, Tautz D, Payseur BA. 2014. Genomic networks of hybrid sterility. *PLoS Genet.* 10: e1004162.
Turro E, Su S-Y, Goncalves A, Coin L, Richardson S and Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Bio.* 12: R13.
Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature Rev Genet.* 15:221-233.
Voets T, Droogmans G, Wissenbach U, Janssens A, Flockerzi V, Nilius B. 2004. The principle of temperature-dependent gating in cold-and heat-sensitive TRP channels. *Nature* 430: 748–54.
Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, Strengman E, Sabatti C, van Kessel AG, *et al.* 2008. Recurrent CNVs disrupt three candidate genes in

schizophrenia patients. *Am J Hum Genet. 83*: 504-510.
Vyskočilová M, Pražanová G, Piálek J. 2009. Polymorphism in hybrid male sterility in wild-derived Mus musculus musculus strains on proximal chromosome 17. *Mamm Genome* 20: 83-91.
Walsh JB. 1982. Rate of accumulation of reproductive isolation by chromosome rearrangements. *Am Nat.* 120: 510-532.
Watanabe TK. 1979. A gene that rescues the lethal hybrids between *Drosophila melanogaster* and *D. simulans*. *Jpn J Genet.* 54:325–331.
Wei KH, Clark AG, Barbash DA. 2014. Limited gene misregulation is exacerbated by allele-specific upregulation in lethal hybrids between *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol.* 31: 1767-1778.
Weng MP, Liao BY. 2017. *mod*PhEA: *mod*el organism *P*henotype *E*nrichment *A*nalysis of eukaryotic gene sets. *Bioinformatics 33*:3505-3507.
Wertz K, Herrmann BG. 2000. Large-scale screen for genes involved in gonad development. *Mech Dev.* 98: 51-70
Williams IV R, Lim JE, Harr B, Wing C, Walters R, Distler MG, Teschke M, Wu C,
Wiltshire T, Su AI, Sokoloff G. 2009. A common and unstable copy number variant is associated with differences in *Glo1* expression and anxiety-like behavior. *PLoS One* 4:e4649.
Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* 43: 85-88.
Wittkopp PJ, Haerum BK, Clark A. 2008. Genetic basis of regulatory variation within and between Drosophila species. *Nature Genet.* 40: 346-50.
Wittkopp PJ and Kalay G. 2012. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 13: 59–69.
White MA, Steffy B, Wiltshire T, Payseur BA. 2011. Genetic dissection of a key reproductive barrier between nascent subspecies of house mice, *Mus musculus domesticus* and *Mus musculus musculus*. *Genetics* 169: 289-304.
Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet.* 8:206-216.
Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, Su AI. 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* 10:R130.
Wu JY, Kao HJ, Li SC, Stevens R, Hillman S, Millington D, Chen YT. 2004. ENU mutagenesis identifies mice with mitochondrial branched-chain aminotransferase deficiency resembling human maple syrup urine disease. *J Clin Invest.* 113: 434-440.
Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, Tang H, Snyder M. 2013. Variation and genetic control of protein abundance in humans. *Nature* 499:79-82.
Xu S. 2003. Theoretical basis of the Beavis effect. *Genetics* 165: 2259-2268.
Ye K, Lu J, Raj SM, Gu Z. 2013. Human expression QTLs are enriched in signals of environmental adaptation. *Genome Biol Evol.* 5: 1689-1701.
Yuan L, Chen L, Qian K, Qian G, Wu CL, Wang X, Xiao Y. 2017. Co-expression network

analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC). *Genomics data* 14: 132-140.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L. 2003. *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* 35:57-64

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genet* 42: 355-360.

Zheng X, Levine D, Shen J, Gogarten S, Laurie C and Weir B. 2012. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* 28: 3326-3328.

Zerbino DR *et al.* Ensembl 2018. *Nucleic acids research* 46.D1 (2017): D754-D761.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genet.* 44: 821-824.

Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, Zhao Q. 2018. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. *Int J Biol Sci.* 14:124.

Zhu J, Chen G, Zhu S, Li S, Wen Z, Li B, Zheng Y, Shi L. 2016. Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using RNA-seq. *Sci Rep. 6*: 28400.

Zhuang Y, Adams KL. 2007. Extensive allelic variation in gene expression in *Populus* F1 hybrids. *Genetics* 177: 1987-1996.