Stem cell transcriptional profiles from mouse subspecies reveal *cis*-regulatory evolution at translation genes

Noah M. Simon[1,2,3], Yujin Kim[4,5], Diana M. Bautista[6], James R. Dutton[4,5], Rachel B. Brem[2,3]*

[1]Biology of Aging Doctoral Program, Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA 90089

[2]Buck Institute for Research on Aging, Novato, CA 94945, USA

[3]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley CA 94720, USA

[4]Stem Cell Institute, University of Minnesota, Minneapolis, MN 55455, USA

[5]Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455, USA

[6]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley CA 94720

*Correspondence to rbrem@berkeley.edu

Word count of main text: 6,963

Abstract

A key goal of evolutionary genomics is to harness molecular data to draw inferences about selective forces that have acted on genomes. The field progresses in large part through the development of advanced molecular-evolution analysis methods. Here we explored the intersection between classical sequence-based tests for selection and an empirical expression-based approach, using stem cells from *Mus musculus* subspecies as a model. Using a test of directional, *cis*-regulatory evolution across genes in pathways, we discovered a unique program of induction of translation genes in stem cells of the Southeast Asian mouse *M. m. castaneus* relative to its sister taxa. As a complement, we used sequence analyses to find population-genomic signatures of selection in *M. m. castaneus,* at the upstream regions of the translation genes, including at transcription factor binding sites. We interpret our data under a model of changes in lineage-specific pressures across *Mus musculus* in stem cells with high translational capacity. Together, our findings underscore the rigor of integrating expression and sequence-based methods to generate hypotheses about evolutionary events from long ago.

Introduction

The central objective of molecular-evolution research is to draw inferences about changing selective pressures between lineages, based on clues from omics data. Since its inception, the field has relied in large part on model-fitting methods using DNA sequence (Kreitman 2000) and gene expression (Price et al. 2022). Complementing the latter, empirical genome ranking/bootstrapping approaches have emerged in the more recent literature (Ferguson and Chang, 2020), for which methods development and refinement remain an active area of research (Berg *et al.*, 2019; Sohail *et al.*, 2019; Johri *et al.*, 2020; Price *et al.*, 2022).

One clear-cut empirical molecular-evolution strategy (Fraser *et al.*, 2010, 2011; Bullard *et al.*, 2010; Martin *et al.*, 2012; York *et al.*, 2018; Gokhman *et al.*, 2021) takes as input measurements of *cis*-regulatory variation from expression profiles. The test identifies cases in which, among the unlinked genes of a pathway subject to *cis*-regulatory change, alleles in one taxon tend to drive expression mostly up, or mostly down, relative to another taxon. This pattern of independent genetic variants with similar effects at similar genes is unlikely under neutral expectations (Orr, 1998). It thus serves as a suggestive signature of changes in selective pressure on the pathway between lineages. At this point, rigorous evolutionary inference requires additional follow-up, including sequence-based tests to distinguish between positive and relaxed

selection as the driver of expression divergence. A number of studies in yeast have made this link (Martin *et al.*, 2012; Fraser *et al.*, 2012; Roop *et al.*, 2016); in metazoans, it remains an open question whether results from expression-based *cis*-regulatory pathway tests can be validated with sequence-based evidence for selection (though see Mack et al. 2023).

In the current work, we set out to harness the diversity among mouse lineages in pluripotent stem cell expression programs, to model the integration of expression- and sequence-based tests for selection in multi-gene pathways. We focused on the mouse *Mus musculus castaneus (M. m. castaneus)*. This subspecies is endemic to southeast Asia and diverged 0.5-1 MYA from other *Mus musculus* (Chapman and Ruddle, 1972; Sangster *et al.*, 1993). Previous surveys have established divergence between *M. m. castaneus* and laboratory strains in terms of gene expression (Fraser *et al.*, 2011; Xiong *et al.*, 2014; Chappell *et al.*, 2017; Tkatchenko *et al.*, 2019; Chou *et al.*, 2022) and phenotype (Johnson *et al.*, 1997, 2006; Beamer *et al.*, 1999; Yu *et al.*, 2007; Koturbash *et al.*, 2011; French *et al.*, 2015; Omura *et al.*, 2015; Chappell *et al.*, 2017; Hsiao *et al.*, 2020; Chou *et al.*, 2022). Our goal was to use stem cell transcriptomes to identify pathways subject to directional *cis*-regulatory change between *M. m. castaneus* and laboratory mice. We earmarked one pathway hit, a set of translation genes at which *M. m. castaneus* alleles acted in *cis* to drive uniquely high expression, for independent validation analyses with sequence data. At these loci, population-genomic tests revealed signals of

unique evolution in *M. m. castaneus* in the regions upstream of genes. Together, our results establish the utility of an expression-based molecular-evolution test with sequence-based follow-up, in a model mouse tissue.

Materials and Methods

RNA-seq data sources

For our initial screen for directional *cis*-regulatory variation in pathways, we used transcriptional profiles of CAST/EiJ male x 129/SvImJ female F1 hybrid embryonic stem cells (NCBI [National Center for Biotechnology Information] accession GSE60738, samples SRR1557132, SRR1557133, SRR1557112, and SRR1557123; Marks et al. 2015). For validation and follow-up we used additional transcriptional profiles from reciprocal crosses of CAST/EiJ x C57BL/6J hybrid pluripotent stem cells (NCBI accession GSE90516, samples SRR5054337-5054348 and SRR5054353-5054364; Werner et al. 2017); homozygous embryonic stem cells from a panel of *M. musculus* subspecies (EBI [European Bioinformatics Institute; Sarkans et al. 2018] accession E-MTAB-7730 [Skelly et al. 2020]); and in-house cultures of CAST/EiJ male x 129/SvImJ female F1 hybrid embryonic stem cells, and homozygous induced pluripotent stem cells (see below).

For validation of the role of candidate transcription factors in translation regulation in stem cells, we analyzed published transcriptional data from trophoblast stem cells knocked out for *Eed* (Weigert *et al.*, 2023; NCBI

accession GSE198076, samples SRR18260888, SRR18260892, SRR18260902, and SRR18260903) and E8.5 embryos knocked out for *Ehmt2* (Auclair *et al.*, 2016; NCBI accession GSE71500, samples SRR2133432, SRR2133433, SRR2133434, and SRR2133435), respectively.

RNA-seq read mapping

For analysis of CAST/EiJ x 129/SvImJ and CAST/EiJ x C57BL/6J hybrid stem cell transcriptomes from (Marks *et al.*, 2015), we downloaded raw reads and mapped with the STAR aligner (Dobin et al 2013) to a concatenated genome file containing chromosomal sequences from both parent strains (CAST/EiJ and either 129/SvImJ or C57BL/6J).

For analysis of transcriptomes of homozygous CAST/EiJ and C57BL/6J stem cells cultured in-house (see below), we downloaded raw reads and mapped them to the corresponding reference genome for the respective strain (CAST/EiJ or GRCm38). For validation of the role of candidate transcription factors in translation regulation in stem cells, we mapped RNA-seq from *Ehmt2* knockout embryos (NCBI accession GSE71500; Auclair *et al.*, 2016) and their corresponding wild-type to the GRCm38 reference, and we mapped RNA-seq from *Eed* knockout stem cells and their corresponding wild-type to the CD-1 reference genome (NCBI accession GSE209815; Jung *et al.*, 2023).

RNA-seq normalization and quantification

Reads that mapped ambiguously to more than one locus were discarded. Read counts were generated during the STAR alignment step using the '--quantMode GeneCounts' option. For analysis of transcriptomes of homozygous stem cells from (Skelly *et al.*, 2020) (genotypes C57BL/6J, A/J, 129S1/SvImJ, NZO/HILtJ, NOD/ShiLtJ, WSB/EiJ, CAST/EiJ, and PWD/PhJ), we downloaded mapped read counts from EBI's ArrayExpress database (Sarkans *et al.*, 2018). For each data set in turn, normalized (TPM, transcripts per million) counts were generated by dividing read counts per gene by transcript length (using annotations from the Ensembl database, build 102) and then dividing by library size. Normalized read counts for hybrids and homozygous strains are reported in Table S1 and S2, respectively.

Hybrid RNA-seq mapping quality control

To eliminate potential artifacts from allele-specific mapping errors in hybrid RNA-seq analyses, we performed a simulated RNA-seq experiment using the Polyester package in R (Frazee *et al.*, 2015) as follows. For the CAST x 129 hybrid stem cell transcriptome from (Marks *et al.*, 2015), we generated two replicates of simulated reads from the hybrid genome with ~200 reads for each annotated transcript. These simulated reads were then mapped back to the genome with STAR as above. For a given gene with called orthologs in the CAST and 129 genomes, for each allele in turn we tabulated the ratio between the number of successfully mapped simulated reads and the number of simulated reads that went into the mapping, as a report of the extent of artifact-

free mapping. We converted each such ratio to a percentage; we then took the absolute value of the difference between the ratio for the 129 and CAST allele as a report of the difference in mapping fidelity between them. We filtered out genes for which the latter parameter exceeded 5%, a total of 1,852 genes of 10,380 initial homologs in the data from (Marks *et al.*, 2015) (Table S3). We repeated this procedure for CAST x BL6 hybrid stem cell transcriptomes from (Werner *et al.*, 2017), resulting in exclusion of 2,886 genes out of an initial 8,887 homologs expressed in the RNA-seq data (Table S3).

In silico **screen for directional** ***cis*****-regulatory variation**

For our initial screen of polygenic, directional cis-regulatory variation in pathways, we harnessed profiles from CAST/EiJ male x 129/SvImJ female F1 hybrid pluripotent stem cells (Marks et al. 2015). We generated a list of one-to-one orthologous genes (Table S4 between CAST/EiJ and 129/SvImJ from the Ensembl database (build 102) using biomaRt (Durinck *et al.*, 2009). At a given gene, we tested for differential expression between the CAST/EiJ and 129/SvImJ alleles using the reads mapping to each as input into DESeq2 (Love et al. 2014). We eliminated from further analysis genes with fewer than 10 total reads across all samples. Differential expression results are reported in Table S5.

For each gene with differential allele-specific expression at adjusted $P < 0.05$, we assigned a quantitative sign statistic s, equal to the $\log_2$(129 allele

expression/CAST allele expression). We assigned all genes without differential allele-specific expression to have $s = 0$. We downloaded gene annotations in Gene Ontology 'biological process' terms from the AmiGO database (Carbon *et al.*, 2009). We eliminated from further testing all terms containing fewer than 10 genes with significant differential allele-specific expression. For a given remaining term containing n genes, we summed the s values across the component genes to yield a summary statistic S_{true}. To evaluate significance by resampling, we randomly sampled n genes from the total set with expression data and summed their s values, generating a resampled summary statistic $S_{resample}$. We carried out this calculation 10,000 times and used as a two-sided P value the proportion of resamples in which $|S_{resample}| \geq |S_{true}|$. We corrected for multiple testing with the Benjamini-Hochberg method. P values are reported in Table S6. All further expression and molecular-evolution analyses focused on genes in the top-scoring term, GO:0006412, translation.

Induced pluripotent stem cell derivation and culture

To establish C57BL/6J and CAST/EiJ induced pluripotent stem cell lines, we used mouse embryonic fibroblasts E13.5 obtained from Jackson Laboratories (Bar Harbor, ME, USA) as input into a stem cell derivation protocol as previously described (Terzic *et al.*, 2016). Briefly, we used octamer-binding transcription factor-4 (Oct4), sex-determining region y-box 2 (Sox2), and Kruppel-like factor-4 (Klf4) as reprogramming factors, introduced using pMXs retroviral vectors. A 129S6/SvEv embryonic stem cell line was obtained from

Millipore Sigma, (Catalog no. SCR012), Burlington, MA, USA). A CAST/EiJ x 129/SvImJ mouse embryonic stem cell line was generously provided by Joost Gribnau (Marks *et al.*, 2015).

Stem cells were cultured on irradiated mouse embryonic fibroblasts (R and D Systems, Minneapolis, MN, USA) in miPSC medium: knockout DMEM with 4.5 g/ L d-glucose (Gibco, Grand Island, NY, USA), 10% knockout serum replacement (KSR) (Gibco), 10% fetal bovine serum (FBS) (HyClone, Logan, UT, USA), 1× MEM nonessential amino acids (MEM NEAA) (Gibco), 1× GlutaMAX (Gibco), 0.1 mM 2-mercaptoethanol (BME) (Life Technologies, Grand Island, NY, USA), and 0.02% ESGRO-LIF (Millipore, Billerica, MA, USA). Cells were incubated at 37°C in 5% CO_2.

RNA isolation and sequencing

RNA was extracted from undifferentiated pluripotent stem cell cultures (two replicates each of BL6 and CAST) following feeder depletion using the RNAqueous™-Micro Total RNA Isolation Kit (Thermo Fisher Scientific) and on-column DNase treatment (QIAGEN, Hilden, Germany). RNA samples were processed into mRNA libraries and sequenced on an Illumina NovaSeq 6000 Sequencing System, yielding ~20M paired-end 150 bp reads per sample. Read-mapping was as in RNA-seq normalization and quantification, above.

Population genomic analysis

We downloaded resequencing data from wild populations of *M. m. domesticus* (from France, Germany and Iran), *M. m. musculus* (from Afghanistan, Czech Republic and Kazakhstan), and *M. m. castaneus* (from northwest India) (Harr *et al.*, 2016). VCF files were used to map SNPs from each individual mouse onto the C57BL/6J reference genome (GRCm38) using the 'consensus' command from bcftools software (Danecek *et al.*, 2021). Default options for 'bcftools consensus' algorithm were used, meaning that in the case of the heterozygous loci, the reference GRCm38 allele would be used. Sequences upstream of the transcription start site for each gene were extracted utilizing the pybedtools Python package (Quinlan and Hall, 2010; Dale *et al.*, 2011). In order to assess only a single transcript per gene, gffread (Pertea and Pertea, 2020) was used to extract coding sequences (CDS) from each annotated transcript. A custom Python script was used to check whether each CDS contained an in-frame start and stop codon, signifying a valid open reading frame (ORF). For genes with multiple transcripts containing a valid ORF, the longest transcript was selected. Genes without an in-frame ORF were eliminated from further analysis.

We carried out selection tests separately in a comparison of each *M. m. domesticus* or *M. m. musculus* population in turn against the *M. m. castaneus* population. In each case, for a given gene, we generated an alignment of the 50kb region upstream of the gene start site and split it into ten 5kb windows. In each window of each gene, we tabulated the number of sites harboring

polymorphisms across one or both populations, and the number of sites with alleles fixed in each population and divergent between them (P_u and D_u respectively); these data are reported in Table S7. For a given window, we took the sum of D_u and, separately, the sum of P_u across the translation GO term genes, and we calculated the ratio $R_{translation} = D_u/P_u$. To evaluate significance by resampling, for a given window we randomly sampled 10 genes from the total set with sequence data; calculated P_u and D_u, their sums across genes, and their ratio $R_{resample}$ as above; and over 10,000 resamples, we used as a one-sided *P* value the proportion of resamples in which $R_{resample} \geq R_{translation}$. To check robustness of the signal, we also employed a leave-one-out strategy, where we repeated the resampling test over 10,000 permutations with each translation gene excluded in turn; these data are reported in Table S8.

Analysis of transcription factor binding sites

In a public resource of chromatin immunoprecipitation sequencing data sets (Kolmykov *et al.*, 2021), we tabulated binding sites for 680 transcription factors across the mouse genome. For each factor, D_u/P_u was calculated as above for binding sites within 50kb upstream of genes, and then separately in the remaining non-binding site regions. These values were then used as input for a two-factor ANOVA, testing for an interaction effect between binding site identity and membership in the translation GO term for each factor. D_u/P_u between the *M. m. castaneus* population and each *M. m. domesticus* or *M. m. musculus*

population were used as replicates in the ANOVA analysis. ANOVA results are reported in Table S9. Multiple sequence alignment was edited for display using Jalview software (Waterhouse *et al.*, 2009). Binding site data for the top 3 hits from this analysis are reported in Table S10.

Results

A screen for directional *cis*-regulatory change in mouse stem cell pathways

As a first testbed for analyses of pathway *cis*-regulatory change, we used transcriptional profiles (Marks *et al.*, 2015) of embryonic stem cells of an F1 hybrid background from a cross between two homozygous mouse strains: a male *M. m. castaneus* (CAST/EiJ, hereafter CAST), and a female of the 129/SvImJ laboratory strain genotype (hereafter 129), which is of admixed origin (Frazer *et al.*, 2007; Yang *et al.*, 2007, 2011). In the F1, because the two subspecies' alleles of a given gene are in the same nucleus, any difference in allele-specific expression between them can be attributed to genetic variation acting in *cis*, *i.e.,* not mediated through a soluble factor (Wittkopp *et al.*, 2004). We implemented a pipeline of allele-specific read-mapping taking account of the potential for mapping artifacts (see Methods); and we retained for analysis all genes exhibiting significant expression divergence between the species' alleles (962 genes at a 0.05 *p*-value threshold). We tabulated the directional effect for each gene—whether the CAST allele was more highly expressed

than the laboratory-strain allele, or *vice versa*—using the $\log_2$-transformed fold-change between expression of the 129 allele and CAST allele. Then, to formulate our test, we used as pathways groups of genes of common function, each comprised of a Gene Ontology biological process term. For each such group, we quantified the agreement in the direction of allelic expression differences between species across the gene members. We evaluated significance based on resampling (Table S6). Of the complete survey results, one pathway showed significant signal: a cohort of genes from the GO term for translation, at which the CAST allele was upregulated relative to that of 129 in the hybrid at a 2-fold excess of genes relative to those with higher expression from the 129 allele (Table S6, Figure 1A, and Figure S1A). This represented a potential case in which selective pressures on regulation of the respective loci had changed between the *M. m. castaneus* and laboratory-strain lineages.

***M. m. castaneus cis*-regulatory alleles drive high expression of translation genes in stem cells**

As a first verification of the trend for *cis*-regulatory alleles from *M. m. castaneus* driving high expression of translation genes, we repeated the culture and sequencing of CAST x 129 hybrid stem cells. The results confirmed the robust directional imbalance in allele-specific expression among translation genes, with the CAST allele expressed more highly across the set 2.54-fold more often than the 129 allele (Figure 1A-B). Similarly, we analyzed the transcriptome of hybrid stem cells from a cross between CAST and the

admixed C57BL/6J laboratory strain (hereafter BL6; Werner et al. 2017), and observed a 2.48-fold imbalance favoring high expression by the CAST allele among translation genes (Figure 1A and Figure S1B). Together, these data establish that *M. m. castaneus cis*-regulatory alleles at translation genes encode a unique activating program relative to those encoded by 129 and BL6 alleles, in stem cells.

Homozygous *M. m. castaneus* stem cells are distinguished by high expression of translation genes

We expected that, if regulatory divergence between CAST and other lineages at translation genes had been important for fitness in the organismal and ecological context, it would be apparent in the context of homozygous strains, which integrate the effects of genetic factors acting both in *cis* and in *trans* (Signor and Nuzhdin, 2018). Consistent with this prediction, translation genes were more highly expressed in CAST homozygous stem cells relative to those of 129 and BL6, in a published data resource (Skelly *et al.*, 2020) and in our own culture and sequencing (Figure 1A and 1C and Figure S1C-D). The trend persisted in comparisons between CAST homozygous stem cells and those of other mouse lineages (Figure 1A and Figure S1E-I). Interestingly, the tendency for high expression of translation genes by CAST in homozygous stem cells was of larger magnitude than we had noted in our analyses of *cis*-regulatory variation in hybrids (Figure 1A), indicating that the latter was reinforced by a more pervasive effect of divergence in *trans*-acting regulators. Given these

signatures of directional *cis*- and *trans*-acting variation between *Mus* subspecies, we considered the translation gene cohort to be a compelling candidate for molecular-evolution follow-up, to investigate the underlying selective pressures.

M. m. castaneus **harbors unique and conserved alleles in spatial windows of translation gene upstream regions**

We next turned to sequence-based tests of selection to explore the evolutionary history of translation gene regulation between mouse lineages. We reasoned that analysis of multiple *M. musculus* subspecies could help polarize any sequence variation we would detect at the loci. Toward this end, we developed an analysis pipeline for single-nucleotide variants from wild-caught mice of the *Mus* genus (Harr *et al.*, 2016), in which we tabulated divergence between subspecies, normalized by intra-subspecies polymorphism, in spatial windows of upstream regions for every gene in the genome. We then used these values as input into resampling tests for enrichment of high values of this normalized quantity in translation genes relative to a genomic null, representing a hallmark of positive selection (Kreitman, 2000). In a first such approach, we classified variant sites based on their spatial positions in upstream loci (Figure 2A). The results revealed elevated sequence divergence between *M. m. castaneus* on the one hand and populations of *M. m. musculus* and *M. m. domesticus* on the other (Figure 2A-B). The signal was most marked in a window 25 kb upstream of translation

start among genes of the translation GO term and fell off gradually on either side of this peak (Figure 2A), as expected if *M. m. castaneus* had acquired unique alleles in *cis*-regulatory elements with positional preferences across this gene cohort. Such a trend was consistent with the unique expression program in *M. m. castaneus* that we had noted in stem cells as a product in part of *cis*-regulatory change (Figure 1A).

A role for transcription factors in divergently regulated translation genes

As a complement to our focus on the spatial positioning of *cis*-regulatory sequence changes in translation genes, we developed a separate molecular-evolution approach classifying variants on the basis of binding sites by transcriptional regulators (Kolmykov *et al.*, 2021). For the sites bound by a given regulator, we again tested for elevated divergence between *M. m. castaneus* and its relatives, relative to intra-subspecies polymorphism, across the genes of our translation cohort (Figure 3A-B). Three regulators (Ehmt2, Eed, and Msx1) exhibited robust signal in this scheme, with binding sites at translation genes enriched for divergence between *M. m. castaneus* and its sister lineages (Table S9 and Figure 3A-B). In *Eed* knockout stem cell and *Ehmt2* knockout early embryonic transcriptomes (Auclair *et al.*, 2016; Weigert *et al.*, 2023), we detected a directional impact of these factors on expression of translation genes (Figure 3C and Figure S2), consistent with an inference of direct regulatory function at translation genes. Interestingly, unique alleles in *M. m. castaneus* in binding sites of these regulators did not fall in the window

that had been top-scoring in our spatially resolved analysis of upstream regions (at 25 kb upstream of coding start; Figure 2 and Table S10), suggesting a role in the latter of mechanisms beyond Ehmt2, Eed, and Msx1 binding. As such, results from the two test schemes represent independent evidence for positive selection in distinct elements of the upstream regions of translation genes—which together represent compelling candidates for the mechanism of their non-neutral expression variation across mouse lineages, and its manifestation in stem cells.

Discussion

Empirical molecular evolution tests, once developed, must be validated to establish their rigor and utility for the field. A compelling means toward this end is to integrate results from an emergent test strategy with those of more classic tools, when they complement each other in support of an evolutionary inference. In the current work, we have detected sequence signatures of selection between mouse subspecies in a cohort of translation genes that also exhibits directional, polygenic *cis*- and *trans*-acting regulatory variation in stem cells. Against a backdrop of other case studies focused on *cis*-regulatory change as detected in transcriptomes (Fraser *et al.*, 2010, 2011; Bullard *et al.*, 2010; Martin *et al.*, 2012; York *et al.*, 2018; Agoglia *et al.*, 2021; Gokhman *et al.*, 2021; Mack *et al.*, 2023; Wang *et al.*, 2024), our results represent a proof of concept for sequence-based validation of this approach in mammals.

Our analyses implicate several transcriptional regulators in the divergence between mouse subspecies in translation genes. The variation we have detected in translation gene binding sites for the polycomb repressive complex 2 component Eed (Mozzetta *et al.*, 2014; Obier *et al.*, 2015; Li *et al.*, 2018), as a candidate determinant of stem cell expression change between mice, dovetails with the known role for Eed in development and stem cell differentiation (Young, 2011; Obier *et al.*, 2015). The polycomb repressive complex 2 more broadly has been shown to regulate pluripotency of stem cells through methylation of the DNA encoding ribosomal RNAs (Zhang *et al.*, 2020); our focused study of Eed suggests that its effects in stem cells are also mediated by other translation components, as reflected in binding at these loci in stem cells (Li *et al.*, 2018; Oksuz *et al.*, 2018; Kaaij *et al.*, 2019; Kriz *et al.*, 2021). Likewise, the evolutionary and genomic profiles of binding sites at translation genes for the methyltransferase Ehmt2, including their inferred impact on expression in stem cells, dovetail with the known function of this factor in stem cell maintenance and differentiation (Ikegami *et al.*, 2007; Leitch *et al.*, 2013; Boroviak *et al.*, 2014; Auclair *et al.*, 2016; Kim *et al.*, 2020) and direct binding by Ehmt2 to translation genes in stem cells (Mozzetta *et al.*, 2014). As such, our results suggest that the known roles of Eed and Ehmt2 in stem cell identity and differentiation may be mediated by regulation of protein synthesis, and that evolution used these binding sites in part to tune expression of translation factors between mouse lineages. A similar function

may ultimately also emerge for the development regulator Msx1, whose binding sites were also top-scoring in our molecular-evolution analyses across *Mus*. That said, we expect that our genomic approach affords only partial coverage and power in discovering elements of the mechanism of mouse subspecies divergence at translation genes, and thus that many other contributing transcription factors likely remain to be identified.

Our finding of divergence at translation genes between mouse lineages echoes reports of *cis*-regulatory change in the translation machinery of other organismal systems, from yeasts (Tanay *et al.*, 2005; Hogues *et al.*, 2008; Li and Fay, 2017; Sorrells *et al.*, 2018) to animal ancestors distributed over deep time (Brown *et al.*, 2008). This literature leaves open the question of what ecological forces might drive divergence of translation gene expression, and the phenotypes that would mediate such effects. Under a model in which ribosomal protein dosage governs the readiness of a cell to divide (Polymenis and Aramayo, 2015), adaptive regulatory variation in translation pathways may often reflect species-unique logic of cell growth decisions.

In metazoan stem cells in particular, translation plays a critical role in differentiation. Inducing (Easley *et al.*, 2010) or compromising (Khajuria *et al.*, 2018) stem cell translation can drive qualitative differences in differentiation behavior. According to current models, high expression of translation genes in stem cells (Sampath *et al.*, 2008) sets up a poised state to enable rapid protein production in their differentiated progeny (Gabut *et al.*, 2020). On the basis of

this tight link between translation and differentiation, it is tempting to speculate that the variation we have seen among *Mus musculus* lineages in stem-cell translation gene expression was driven by selection in distinct niches for unique differentiation properties. If so, the loci we study here would act in a complex genetic architecture of stem cell differentiation, alongside other variants characterized across mouse lineages (Ortmann *et al.*, 2020; Skelly *et al.*, 2020). Future work will explore the phenotypic relevance of translation gene divergence across *Mus*, and its role in the slew of traits (Johnson *et al.*, 1997; Beamer *et al.*, 1999; Johnson *et al.*, 2006; Yu *et al.*, 2007; Koturbash *et al.*, 2011; French *et al.*, 2015; Chappell *et al.*, 2017; Hsiao *et al.*, 2020; Chou *et al.*, 2022) that distinguish *M. m. castaneus* from the rest of its genus.

Acknowledgements

The authors thank Taekyu Kang and Matt Dean for helpful discussions. This work was supported by National Institutes of Health R01NS116992 to DMB, JRD, and RBB, and R01GM120430 to RBB.

Author Contributions

NMS, RBB, DMB, and JRD conceived of the study and provided direction. NMS processed the sequencing data and carried out all *in silico* analyses. JRD derived the iPSC lines used for sequencing and neuron differentiation. YK performed the cell cultures and RNA isolations. All authors contributed to and reviewed the final manuscript.

Competing Interests

The authors declare no competing financial interests.

Data Archiving

RNA-seq data generated for this study was deposited in the GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE234761).

References

Agoglia RM, Sun D, Birey F, Yoon S-J, Miura Y, Sabatini K, *et al.* (2021). Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* 592: 421–427.

Auclair G, Borgel J, Sanz LA, Vallet J, Guibert S, Dumas M, *et al.* (2016). EHMT2 directs DNA methylation for efficient gene silencing in mouse embryos. *Genome Res* 26: 192–202.

Beamer WG, Shultz KL, Churchill GA, Frankel WN, Baylink DJ, Rosen CJ, *et al.* (1999). Quantitative trait loci for bone density in C57BL/6J and CAST/EiJ inbred mice. *Mammalian Genome* 10: 1043–1049.

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, *et al.* (2019). Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* 8: e39725.

Boroviak T, Loos R, Bertone P, Smith A, Nichols J (2014). The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat Cell Biol* 16: 513–525.

Brown SJ, Cole MD, Erives AJ (2008). Evolution of the holozoan ribosome biogenesis regulon. *BMC Genomics* 9: 442.

Bullard JH, Mostovoy Y, Dudoit S, Brem RB (2010). Polygenic and directional regulatory evolution across pathways in Saccharomyces. *Proceedings of the National Academy of Sciences* 107: 5058–5063.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, *et al.* (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* 25: 288–289.

Chapman VM, Ruddle FH (1972). GLUTAMATE OXALOACETATE TRANSAMINASE (GOT) GENETICS IN THE MOUSE: POLYMORPHISM OF GOT-1. *Genetics* 70: 299–305.

Chappell GA, Israel JW, Simon JM, Pott S, Safi A, Eklund K, *et al.* (2017). Variation in DNA-Damage Responses to an Inhalational Carcinogen (1,3-Butadiene) in Relation to Strain-Specific Differences in Chromatin Accessibility and Gene Transcription Profiles in C57BL/6J and CAST/EiJ Mice. *Environmental Health Perspectives* 125: 107006.

Chou M-Y, Appan D, Chang K-W, Chou C-H, Lin C-Y, Gau SS-F, *et al.* (2022). Mouse hybrid genome mediates diverse brain phenotypes with the specificity of reciprocal crosses. *FASEB J* 36: e22232.

Dale RK, Pedersen BS, Quinlan AR (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* 27: 3423–3424.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, *et al.* (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008.

Durinck S, Spellman PT, Birney E, Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 4: 1184–1191.

Easley CA, Ben-Yehudah A, Redinger CJ, Oliver SL, Varum ST, Eisinger VM, *et al.* (2010). mTOR-Mediated Activation of p70 S6K Induces Differentiation of Pluripotent Human Embryonic Stem Cells. *Cellular Reprogramming* 12: 263–273.

Ferguson J, Chang J (2020). An empirical Bayesian ranking method, with applications to high throughput biology. *Bioinformatics* 36: 177–185.

Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, Mehrabian M, *et al.* (2011). Systematic detection of polygenic cis-regulatory evolution. *PLoS Genet* 7: e1002023.

Fraser HB, Levy S, Chavan A, Shah HB, Perez JC, Zhou Y, *et al.* (2012). Polygenic cis-regulatory adaptation in the evolution of yeast pathogenicity. *Genome Res* 22: 1930–1939.

Fraser HB, Moses AM, Schadt EE (2010). Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences* 107: 2977–2982.

Frazee AC, Jaffe AE, Langmead B, Leek JT (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* 31: 2778–2784.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, *et al.* (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448: 1050–1053.

French JE, Gatti DM, Morgan DL, Kissling GE, Shockley KR, Knudsen GA, *et al.* (2015). Diversity Outbred Mice Identify Population-Based Exposure Thresholds and Genetic Factors that Influence Benzene-Induced Genotoxicity. *Environ Health Perspect* 123: 237–245.

Gabut M, Bourdelais F, Durand S (2020). Ribosome and Translational Control in Stem Cells. *Cells* 9: 497.

Gokhman D, Agoglia RM, Kinnebrew M, Gordon W, Sun D, Bajpai VK, *et al.* (2021). Human-chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. *Nat Genet* 53: 467–476.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, *et al.* (2016). Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Sci Data* 3: 160075.

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, *et al.* (2008). Transcription Factor Substitution during the Evolution of Fungal Ribosome Regulation. *Mol Cell* 29: 552–562.

Hsiao K, Noble C, Pitman W, Yadav N, Kumar S, Keele GR, *et al.* (2020). A Thalamic Orphan Receptor Drives Variability in Short-Term Memory. *Cell* 183: 522-536.e19.

Ikegami K, Iwatani M, Suzuki M, Tachibana M, Shinkai Y, Tanaka S, *et al.* (2007). Genome-wide and locus-specific DNA hypomethylation in G9a deficient mouse embryonic stem cells. *Genes to Cells* 12: 1–11.

Johnson KR, Erway LC, Cook SA, Willott JF, Zheng QY (1997). A major gene affecting age-related hearing loss in C57BL/6J mice. *Hearing Research* 114: 83–92.

Johnson KR, Zheng QY, Noben-Trauth K (2006). Strain background effects and genetic modifiers of hearing in mice. *Brain Research* 1091: 79–88.

Johri P, Charlesworth B, Jensen JD (2020). Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. *Genetics* 215: 173–192.

Jung YH, Wang H-LV, Ali S, Corces VG, Kremsky I (2023). Characterization of a strain-specific CD-1 reference genome reveals potential inter- and intra-strain functional variability. *BMC Genomics* 24: 437.

Kaaij LJT, Mohn F, van der Weide RH, de Wit E, Bühler M (2019). The ChAHP Complex Counteracts Chromatin Looping at CTCF Sites that Emerged from SINE Expansions in Mouse. *Cell* 178: 1437-1451.e14.

Khajuria RK, Munschauer M, Ulirsch JC, Fiorini C, Ludwig LS, McFarland SK, *et al.* (2018). Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis. *Cell* 173: 90-103.e19.

Kim K, Ryu TY, Ryu J-W, Han T-S, Jung C-R, Son M-Y, *et al.* (2020). RNA-seq based transcriptome analysis of EHMT2 functions in breast cancer. *Biochemical and Biophysical Research Communications* 524: 672–676.

Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev VJ, *et al.* (2021). GTRD: an integrated view of transcription regulation. *Nucleic Acids Research* 49: D104–D111.

Koturbash I, Scherhag A, Sorrentino J, Sexton K, Bodnar W, Swenberg JA, *et al.* (2011). Epigenetic Mechanisms of Mouse Interstrain Variability in Genotoxicity of the Environmental Toxicant 1,3-Butadiene. *Toxicological Sciences* 122: 448–456.

Kreitman M (2000). Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet* 1: 539–559.

Kriz AJ, Colognori D, Sunwoo H, Nabet B, Lee JT (2021). Balancing cohesin eviction and retention prevents aberrant chromosomal interactions, Polycomb-mediated repression, and X-inactivation. *Mol Cell* 81: 1970-1987.e9.

Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N, *et al.* (2013). Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* 20: 311–316.

Li XC, Fay JC (2017). Cis-Regulatory Divergence in Gene Expression between Two Thermally Divergent Yeast Species. *Genome Biol Evol* 9: 1120–1129.

Li Y, Zheng H, Wang Q, Zhou C, Wei L, Liu X, *et al.* (2018). Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol* 19: 18.

Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15: 550.

Mack KL, Square TA, Zhao B, Miller CT, Fraser HB (2023). Evolution of Spatial and Temporal cis-Regulatory Divergence in Sticklebacks. *Mol Biol Evol* 40: msad034.

Marks H, Kerstens HHD, Barakat TS, Splinter E, Dirks RAM, van Mierlo G, *et al.* (2015). Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome Biology* 16: 149.

Martin HC, Roop JI, Schraiber JG, Hsu TY, Brem RB (2012). Evolution of a Membrane Protein Regulon in Saccharomyces. *Molecular Biology and Evolution* 29: 1747–1756.

Mozzetta C, Pontis J, Fritsch L, Robin P, Portoso M, Proux C, *et al.* (2014). The histone H3 lysine 9 methyltransferases G9a and GLP regulate polycomb repressive complex 2-mediated gene silencing. *Mol Cell* 53: 277–289.

Obier N, Lin Q, Cauchy P, Hornich V, Zenke M, Becker M, *et al.* (2015). Polycomb Protein EED is Required for Silencing of Pluripotency Genes upon ESC Differentiation. *Stem Cell Rev and Rep* 11: 50–61.

Oksuz O, Narendra V, Lee C-H, Descostes N, LeRoy G, Raviram R, *et al.* (2018). Capturing the Onset of PRC2-Mediated Repressive Domain Formation. *Mol Cell* 70: 1149-1162.e5.

Omura T, Omura K, Tedeschi A, Riva P, Painter MW, Rojas L, *et al.* (2015). Robust Axonal Regeneration Occurs in the Injured CAST/Ei Mouse CNS. *Neuron* 86: 1215–1227.

Orr HA (1998). Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. *Genetics* 149: 2099–2104.

Ortmann D, Brown S, Czechanski A, Aydin S, Muraro D, Huang Y, *et al.* (2020). Naive Pluripotent Stem Cells Exhibit Phenotypic Variability that Is Driven by Genetic Variation. *Cell Stem Cell* 27: 470-481.e6.

Pertea G, Pertea M (2020). GFF Utilities: GffRead and GffCompare. *F1000Res* 9: ISCB Comm J-304.

Polymenis M, Aramayo R (2015). Translate to divide: control of the cell cycle by protein synthesis. *Microbial Cell* 2: 94–104.

Price PD, Palmer Droguett DH, Taylor JA, Kim DW, Place ES, Rogers TF, *et al.* (2022). Detecting signatures of selection on gene expression. *Nat Ecol Evol* 6: 1035–1045.

Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.

Roop JI, Chang KC, Brem RB (2016). Polygenic evolution of a sugar specialization trade-off in yeast. *Nature* 530: 336–339.

Sampath P, Pritchard DK, Pabon L, Reinecke H, Schwartz SM, Morris DR, *et al.* (2008). A Hierarchical Network Controls Protein Translation during Murine Embryonic Stem Cell Self-Renewal and Differentiation. *Cell Stem Cell* 2: 448–460.

Sangster MY, Heliams DB, MacKenzie JS, Shellam GR (1993). Genetic studies of flavivirus resistance in inbred strains derived from wild mice: evidence for a new resistance allele at the flavivirus resistance locus (Flv). *Journal of Virology* 67: 340–347.

Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, *et al.* (2018). The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Research* 46: D1266–D1270.

Signor SA, Nuzhdin SV (2018). The Evolution of Gene Expression in cis and trans. *Trends Genet* 34: 532–544.

Skelly DA, Czechanski A, Byers C, Aydin S, Spruce C, Olivier C, *et al.* (2020). Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency. *Cell Stem Cell* 27: 459-469.e8.

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, *et al.* (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* 8: e39702.

Sorrells TR, Johnson AN, Howard CJ, Britton CS, Fowler KR, Feigerle JT, *et al.* (2018). Intrinsic cooperativity potentiates parallel cis-regulatory evolution. *eLife* 7: e37563.

Tanay A, Regev A, Shamir R (2005). Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* 102: 7203–7208.

Terzic D, Maxon JR, Krevitt L, Dibartolomeo C, Goyal T, Low WC, *et al.* (2016). Directed Differentiation of Oligodendrocyte Progenitor Cells from Mouse Induced Pluripotent Stem Cells. *Cell Transplant* 25: 411–424.

Tkatchenko TV, Shah RL, Nagasaki T, Tkatchenko AV (2019). Analysis of genetic networks regulating refractive eye development in collaborative cross progenitor strain mice reveals new genes and pathways underlying human myopia. *BMC Medical Genomics* 12: 113.

Wang B, Starr AL, Fraser HB (2024). Cell-type-specific cis-regulatory divergence in gene expression and chromatin accessibility revealed by human-chimpanzee hybrid cells. *Elife* 12: RP89594.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.

Weigert R, Hetzel S, Bailly N, Haggerty C, Ilik IA, Yung PYK, *et al.* (2023). Dynamic antagonism between key repressive pathways maintains the placental epigenome. *Nat Cell Biol* 25: 579–591.

Werner RJ, Schultz BM, Huhn JM, Jelinek J, Madzo J, Engel N (2017). Sex chromosomes drive gene expression and regulatory dimorphisms in mouse embryonic stem cells. *Biol Sex Differ* 8: 28.

Wittkopp PJ, Haerum BK, Clark AG (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430: 85–88.

Xiong H, Morrison J, Ferris MT, Gralinski LE, Whitmore AC, Green R, *et al.* (2014). Genomic Profiling of Collaborative Cross Founder Mice Infected with Respiratory Viruses Reveals Novel Transcripts and Infection-Related Strain-Specific Gene and Isoform Expression. *G3 Genes|Genomes|Genetics* 4: 1429–1444.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F (2007). On the subspecific origin of the laboratory mouse. *Nat Genet* 39: 1100–1107.

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, *et al.* (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* 43: 648–655.

York RA, Patil C, Abdilleh K, Johnson ZV, Conte MA, Genner MJ, *et al.* (2018). Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes. *Proc Natl Acad Sci U S A* 115: E11081–E11090.

Young RA (2011). Control of Embryonic Stem Cell State. *Cell* 144: 940–954.

Yu H, Mohan S, Edderkaoui B, Masinde GL, Davidson HM, Wergedal JE, *et al.* (2007). Detecting Novel Bone Density and Bone Size Quantitative Trait Loci Using a Cross of MRL/MpJ and CAST/EiJ Inbred Mice. *Calcif Tissue Int* 80: 103–110.

Zhang H, Wu Z, Lu JY, Huang B, Zhou H, Xie W, *et al.* (2020). DEAD-Box Helicase 18 Counteracts PRC2 to Safeguard Ribosomal DNA in Pluripotency Regulation. *Cell Reports* 30: 81-97.e7.

Figure captions

Figure 1: Directional *cis*-regulatory variation in translation gene expression in stem cells between mouse subspecies. (A) The first three columns report results from analyses of differential allele-specific expression across genes from the Gene Ontology term GO:0006412, translation, in stem cells of F1 hybrids (left columns) between CAST and either C57BL/6 (from Werner et al. 2017) or 129 ((Marks *et al.*, 2015) or this study) as indicated. The remaining columns report results from analyses of differential expression across translation genes between homozygous mouse stem cells in comparisons between CAST and either 129 (this study or (Skelly *et al.*, 2020)), C57BL/6J, A/J, NZO/HILtJ, NOD/ShiLtJ, WSB/EiJ, CAST/EiJ, or PWD/PhJ (Skelly *et al.*, 2020) as indicated. The first and second rows report the percentage of translation genes in which the CAST allele is expressed higher or lower, respectively, than the allele of the other parent, and the third row reports the ratio of these quantities. **(B)** Each point reports allele-specific expression of a translation gene in CASTx129 hybrid stem cells cultured in this study: the *x*-axis reports the $\log_2$ ratio of expression of the respective strain alleles, and the *y*-axis reports the $\log_{10}$ of the significance of the difference. Point colors report significance of differential allele-specific expression (p-adj, adjusted *p*-value). Red and black text inlays report the percentage of translation genes where expression of the respective parental allele is higher with and without filtering for significance, respectively. **(C)** Data are as in (B),

but from a comparison between homozygous CAST and 129 stem cells cultured in this study.

Figure 2: Elevated nucleotide divergence between *M. m. castaneus* and other subspecies in regions upstream of translation genes. (A) Each bar reports results from a metagene analysis of nucleotide divergence between a *M. m. castaneus* population from India and a *M. m. musculus* population from Afghanistan in a 5000-bp window starting at the indicated distance upstream of transcription start sites. The *y*-axis reports the ratio of the average number of sites in the respective window exhibiting divergence between the subspecies (D_u) to the average number of polymorphic sites (P_u) across genes from the Gene Ontology term GO:0006412, translation, normalized against the analogous quantity for all other genes in the genome. *: $p > 0.05$ from a one-sided resampling test for elevated D_u/P_u in the translation gene cohort against a genomic null. **(B)** Summary of the analysis of D_u/P_u between the *M. m. castaneus* population and 6 *M. m. musuclus* and *M. m. domesticus* populations. Normalized D_u/P_u and resampling p-values are reported for the 5kb window from 25kb to 20kb upstream for each comparison.

Figure 3: Signatures of species divergence and regulatory function of Eed at translation genes. (A) In a given row, cells report the ratio of inter-subspecies divergence (D_u) to within-subspecies polymorphism (P_u) at Eed binding sites upstream of genes from the Gene Ontology term GO:0006412, translation, or all other genes of the genome, in a comparison between *M. m.*

castaneus and the indicated population of a sister taxon. Divergence is significantly enriched in Eed binding sites upstream of translation genes (see Table S9). **(B)** An Eed binding footprint upstream of the translation gene Eif4e2 as an example of mouse subspecies variation. Each row reports the sequence of one *M. m. castaneus, domesticus,* or *musculus* individual (*M. m. c.*, *M. m. d.*, or *M. m. m.* respectively). **(C)** Data and symbols are as in Figure 1B and 1C except that each point reports a comparison of expression of one translation gene between *Eed* knockout and wild-type mouse stem cells (Weigert et al 2023), and red and black text inlays report the percentage of translation genes where expression was higher in the indicated genotype with and without filtering for significance, respectively (DE, differential expression).

Supplementary figure captions

Figure S1: Directional *cis*-regulatory variation in translation gene expression in stem cells between mouse subspecies. Data are as in Figure 1B **(A-B)** or Figure 1C **(C-I)** of the main text except that transcriptomes analyzed were as follows. **(A)** CAST x 129 hybrid stem cells (Marks *et al.*, 2015). **(B)** CAST x BL6 hybrid stem cells (Werner *et al.*, 2017). **(C)** CAST and 129 homozygous stem cells (Skelly *et al.*, 2020). **(D)** CAST and BL6 homozygous stem cells (Skelly *et al.*, 2020). **(E)** CAST and WSB/EiJ homozygous stem cells (Skelly *et al.*, 2020). **(F)** CAST and PWD/PhJ homozygous stem cells (Skelly *et al.*, 2020). **(G)** CAST and A/J homozygous stem cells (Skelly *et al.*, 2020). **(H)** CAST and NOD/ShiLtJ homozygous stem cells (Skelly *et al.*, 2020). **(I)** CAST and NZO/HILtJ homozygous stem cells (Skelly *et al.*, 2020).

Figure S2: Ehmt2 knockout induces translation genes. Data are as in Figure 3C of the main text except that transcriptomes analyzed were wild-type and Ehmt2 knockout stem cells (Auclair et al. 2016).

Supplementary table captions

Table S1. Allele-specific normalized read counts from RNA-seq of stem cells from hybrids involving CAST. Each tab reports allele-specific expression measurements in one genotype of hybrid embryonic stem cells. Tab TPM_GSE234761_CAST_M_x_129_F contains results from four replicate transcriptomes of CAST/EiJ male x 129/SvImJ female F1 hybrid cells from (Marks *et al.*, 2015). Tab TPM_GSE60738_CAST_M_x_129_F contains results from four replicate transcriptomes of replicates of CAST/EiJ male x 129/SvImJ female F1 hybrid cells from this study. Tab TPM_GSE90516_CAST_M_x_BL6_F contains results from 12 replicate transcriptomes of CAST/EiJ male x C57BL/6J female cells (Werner *et al.*, 2017). Tab TPM_GSE90516_CAST_F_x_BL6_M contains results from 11 replicate transcriptomes of CAST/EiJ female x C57BL/6J male cells (Werner *et al.*, 2017). In each tab, the first column reports Ensembl gene accession numbers. The remaining columns are TPM (transcripts per million) values for reads mapped non-ambiguously to either parental allele for each gene (two columns for each biological sample).

Table S2. Normalized read counts from RNA-seq of stem cells from homozygous strains. Each tab reports expression measurements from one data source of experimental profiles of cell culture from homozygous mouse strains: pluripotent stem cells (PSCs) of BL6 and CAST/EiJ genotypes from

this study; embryonic stem cells of eight genotypes from (Skelly et al. 2020); *Eed* knockouts in trophoblast stem cells from (Weigert *et al.*, 2023); and *Ehmt2* knockouts in E8.5 embryos (Auclair *et al.*, 2016). In each tab, the first column reports Ensembl gene accession numbers. The remaining columns report the TPM (transcripts per million) values. Tab TPM_GSE234761_homozyg_CAST_129 contains results from two replicate transcriptomes of CAST PSCs and four replicate transcriptomes of 129 PSCs. Tab TPM_E_MTAB_7730_8_strains contains results from ESC replicate transcriptomes from eight strains (3 replicates CAST, 35 replicates PWD, 3 replicates WSB, 36 replicates BL6, 3 replicates 129, 3 replicates AJ, 22 replicates NOD, 3 replicates NZO). Tab TPM_GSE198076_EED_KO contains results from two replicate transcriptomes of CD-1 wildtype (WT) trophoblast stem cells (TSCs) and two replicate transcriptomes of *Eed* knockout TSCs. Tab TPM_GSE71500_EHMT2_KO contains results from two replicate transcriptomes of WT BL6 embryos (E8.5) and two replicate transcriptomes of *Ehmt2* knockout embryos (E8.5).

Table S3. Mapping results from simulated RNA-seq experiments using hybrid genomes. The first and second tabs report the mapped read counts from simulated RNA-seq of CAST x 129 and CAST x BL6 hybrids, respectively. In a given tab, for each row, the first column lists the Ensembl gene id for the gene analyzed; the second through fourth columns report the mapping percentage (the number of simulated reads mapped, normalized by the total

number of simulated reads) for the indicated allele in the indicated simulated replicate; the last column reports the absolute value of the difference of mapping percentages between the CAST allele and that of the indicated strain (129 or BL6).

Table S4. Accession ID numbers for gene homologs across strains. Gene and transcript accession numbers used to compare homologous alleles between strains. GRCm38_ensembl_gene_ID, Ensembl gene IDs for the C57BL/6 reference. 129_MGP_gene_ID, 129S1/SvImJ Mouse Genome Project IDs. CAST_MGP_gene_ID, CAST/EiJ Mouse Genome Project IDs. gene_name, gene names from the Mouse Genome Informatics database. GRCm38_ensembl_transcript_ID is the Ensembl transcript ID for the longest in-frame ORF for that gene; empty cells in this column are genes with no valid ORF in the GRCm38 genome using the Ensembl build 102 annotations.

Table S5. Differential expression test results in hybrid and homozygous stem cells. Each tab reports results of a genome-wide survey of differential expression between homologous alleles in interspecies hybrid stem cells; between homozygote stem cells of different species genotypes; or between a regulator gene knockout and its isogenic wild-type in stem cell or early embryo samples. Log2FoldChange, log-transformed ratio of normalized expression of the indicated alleles or homozygote samples; Pvalue, Wald test *p*-value; Padj, Benjamini-Hochberg adjusted *p*-value. Tab DESeq_GSE60738_CASTx129_hyb, comparison of CAST vs 129 allele-

specific expression in F1 hybrid stem cells from (Marks *et al.*, 2015). Tab DESeq_GSE234761_CASTx129_hyb, comparison of CAST vs 129 allele-specific expression in F1 hybrid stem cells from this study. Tab DESeq_GSE90516_CAST_x_BL6_hyb, comparison of CAST vs BL6 allele-specific expression in F1 hybrid stem cells from (Werner *et al.*, 2017). Tab DESeq_GSE234761_CASTvs129, comparison of CAST vs 129 expression in homozygous stem cells from this study. Tab DESeq_E_MTAB_7730_CASTvs129, comparison of CAST vs 129 expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_E_MTAB_7730_CASTvsBL6, comparison of CAST vs BL6 expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_E_MTAB_7730_CASTvsWSB, comparison of CAST vs WSB expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_E_MTAB_7730_CASTvsPWD, comparison of CAST vs PWD expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_E_MTAB_7730_CASTvsAJ, comparison of CAST vs AJ expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_E_MTAB_7730_CASTvsNOD, comparison of CAST vs NOD expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_E_MTAB_7730_CASTvsNZO, comparison of CAST vs NZO expression in homozygous stem cells from (Skelly et al. 2020). Tab DESeq_GSE198076_EED_KO, comparison of wildtype vs *Eed* knockout

expression in stem cells from (Weigert *et al.*, 2023). Tab DESeq_GSE71500_EHMT2_KO, comparison of wildtype vs *Ehmt2* knockout expression in E8.5 embryos from (Auclair *et al.*, 2016).

Table S6. A screen for directional allele-specific expression in CAST x 129 hybrid embryonic stem cells. Each row reports results of a statistical test for directional allele-specific expression variation in CAST/EiJ male x 129/SvImJ female F1 hybrid embryonic stem cells (Marks *et al.*, 2015) in the indicated Gene Ontology term. total_number_of_genes, number of genes in the term with analyzable expression measurements; sum_sign_statistic, sum across genes of the term of $\log_2$(129 allele expression/CAST allele expression); resampling_P_val, resampling-based significance of the enrichment for high absolute value of the sign statistic in the term; Adjusted_P, *p*-value after Benjamini-Hochberg correction for multiple testing.

Table S7. Counts of divergent and polymorphic sites in regions upstream of gene start sites, in comparisons between wild mouse populations. Shown are results of analyses of genome sequences from wild mouse populations taken from Harr et al. (2016). Each of the first six tabs reports results from a comparison between a *M. m. castaneus* population (captured in northwest India) and one population of another subspecies: *M. m. domesticus* populations from France (Mmd_FRA), Germany (Mmd_GER), and Iran (Mmd_IRA); or *M. m. musculus* populations from Kazakhstan (Mmm_KAZ), the Czech Republic (Mmm_CZE), and Afghanistan (Mmm_AFG). In a given one of

the first six tabs, the first three columns report Ensembl (build 102) transcript IDs, gene IDs, and gene names, and each of the remaining columns reports the number of divergent sites, D_u, or the number of polymorphic sites, P_u**,** from the 5-kb window at the indicated position upstream of the transcription start site for the indicated gene. In the final two tabs, each row reports the mean D_u and P_u values at the indicated upstream window across all genes in the genome and genes of the translation GO term respectively, in a comparison between *M. m. castaneus* and the indicated population of *M. m. domesticus* or *M. m. musculus*.

Table S8. Tests for enriched divergence relative to polymorphism in the upstream regions of translation genes, in comparisons between wild mouse populations. In the first tab, each column reports results of resampling tests for elevated values of the ratio between the number of divergent sites, D_u, and the number of polymorphic sites, P_u, between *M. m. castaneus* and the indicated population of *M. m. domesticus* or *M. m. musculus* from Table S6, in upstream regions of translation genes. The first column reports the 5-kb window upstream of the transcription start site subject to analysis. Each of the following pairs of columns reports, respectively, the mean D_u/P_u across the translation gene cohort in the indicated population in a comparison against *M. m. castaneus*, and the *p*-value from a resampling test for high mean D_u/P_u in translation genes relative to the genomic null. The second tab reports the results of a calculation in which we repeated the resampling analyses of the

first tab for the window at 25kb upstream but left out each gene in turn. In a given row, the first column reports the gene left out, and the remaining pairs of columns report D_u/P_u across the translation gene cohort and the *p*-value from a resampling test as in the first tab.

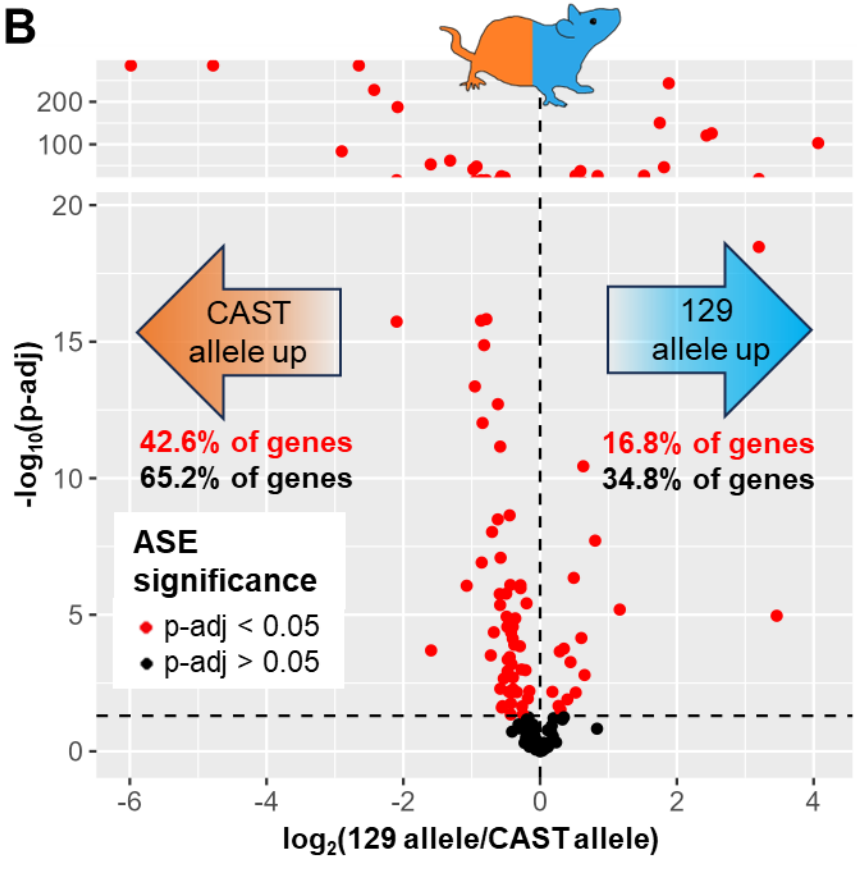
Table S9. Signatures of selection between mouse subspecies in transcription factor binding sites upstream of translation genes. Each row reports results of a two-factor ANOVA testing, across *M. m. musculus* and M. *m. domesticus* populations, for an elevated value of the ratio of sites divergent from *M. m. castaneus*, D_u, to the number of sites polymorphic within populations, P_u, in binding sites of the indicated transcription factor upstream of genes of the translation GO term, relative to non-binding sites and non-translation genes (see Figure 3A of the main text). The second and third columns report the ANOVA F statistic and nominal *p*-value, respectively, and the third reports the *p*-value after Benjamini-Hochberg correction for multiple testing.

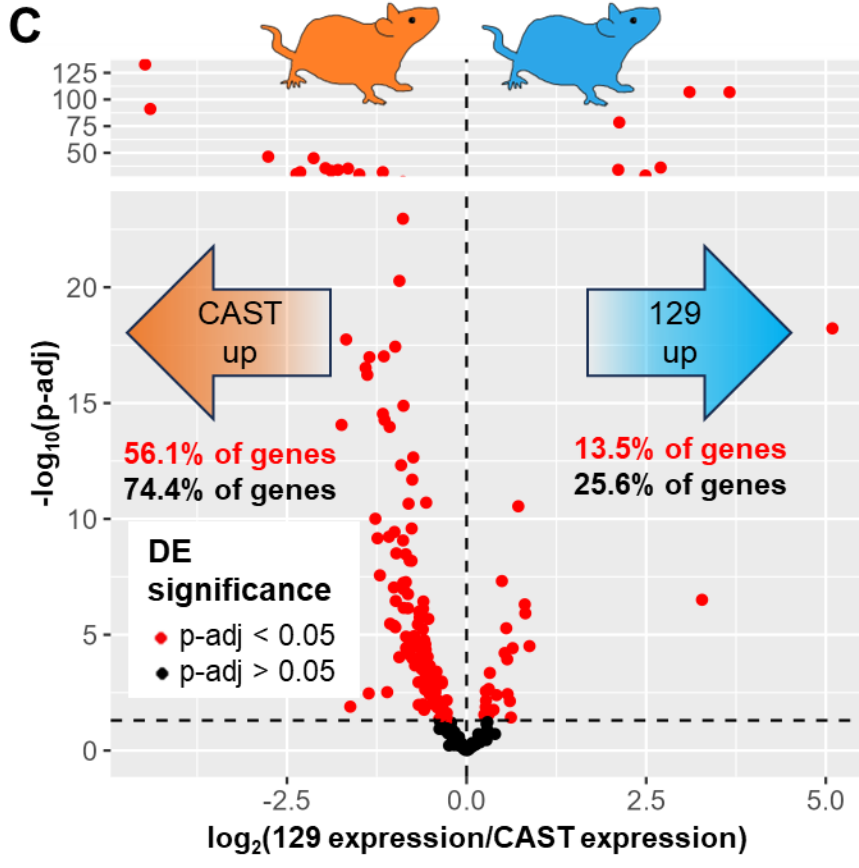
Table S10. Counts of divergent and polymorphic sites in binding sites of top three candidate transcriptional regulators. Each row reports counts upstream of a gene, of sequence variation between wild mouse populations in a binding site for Eed, Ehmt2, or Msx1 from the GTRD database (Kolmykov *et al.*, 2021). The first column reports the Ensembl transcript ID. The next three columns report the chromosome, strand and transcription start site (TSS) position for that gene, respectively. Column E reports whether or not the gene

is part of the translation GO term. Column F reports the transcription factor to the binding site for that row. Columns G-H report the chromosomal positions of the start and end of the binding site, respectively. Columns I-J report the upstream distance of the binding site start and end, respectively, relative to the TSS. Column K reports the length of the binding site which fall within the first 50kb upstream of the TSS. The remaining pairs of columns report the divergent sites within the binding site, d_{TF}, or the number of polymorphic sites within the binding site, p_{TF}**,** from a comparison between a *M. m. castaneus* population (captured in northwest India) and one population of another subspecies: *M. m. domesticus* populations from France (Mmd_FRA), Germany (Mmd_GER), and Iran (Mmd_IRA); or *M. m. musculus* populations from Kazakhstan (Mmm_KAZ), the Czech Republic (Mmm_CZE), and Afghanistan (Mmm_AFG).

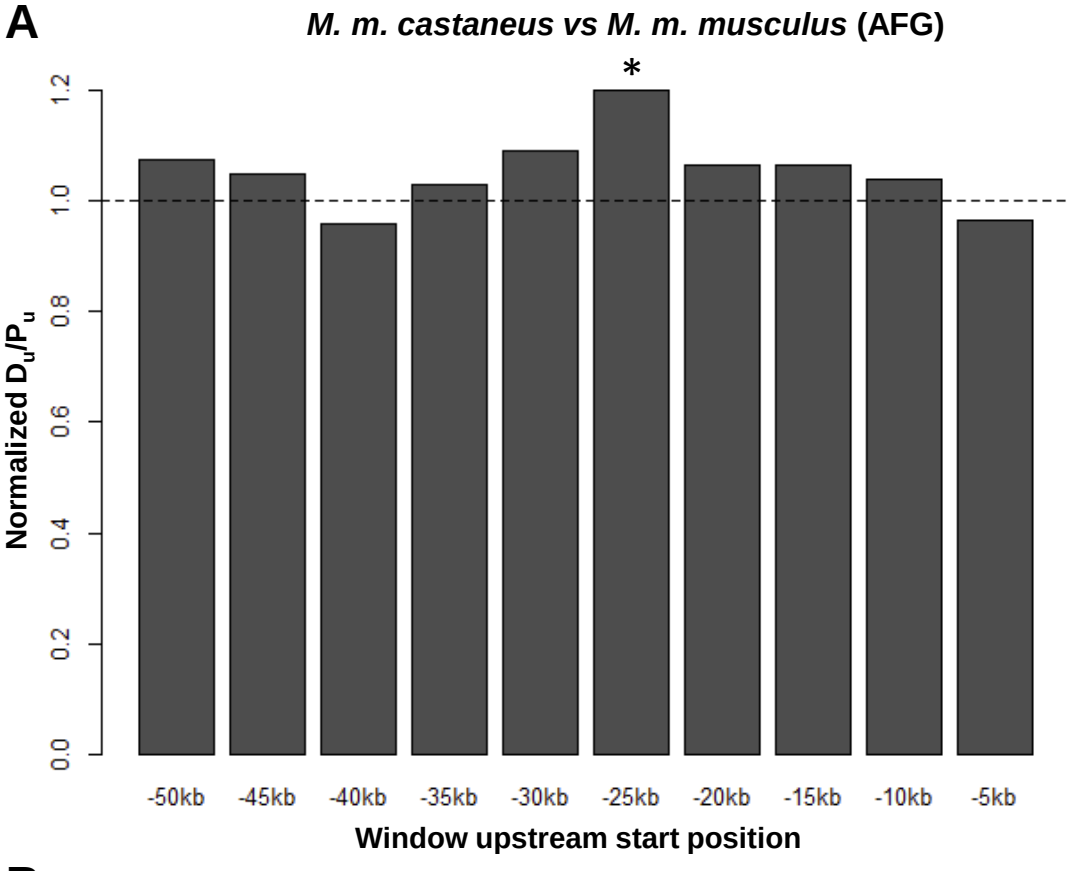
A

	Hybrid stem cells			Homozygous stem cells							
	CAST x129 this study	CAST x129 Marks et al	CAST xBL6 Werner et al	CAST vs 129 this study	CAST vs 129 Skelly et al	CAST vs BL6 Skelly et al	CAST vs PWK Skelly et al	CAST vs WSB Skelly et al	CAST vs AJ Skelly et al	CAST vs NOD Skelly et al	CAST vs NZO Skelly et al
% sig genes CAST up	42.6	15.3	43.7	56.1	65.7	66.1	87.2	36.1	48.0	64.4	51.6
% sig genes CAST down	16.8	7.6	17.6	13.5	8.4	19.2	8.8	8.4	9.8	16.4	11.6
(CAST up/ CAST down)	**2.54**	**2**	**2.48**	**4.17**	**7.85**	**3.43**	**9.90**	**4.32**	**4.91**	**3.92**	**4.46**





A



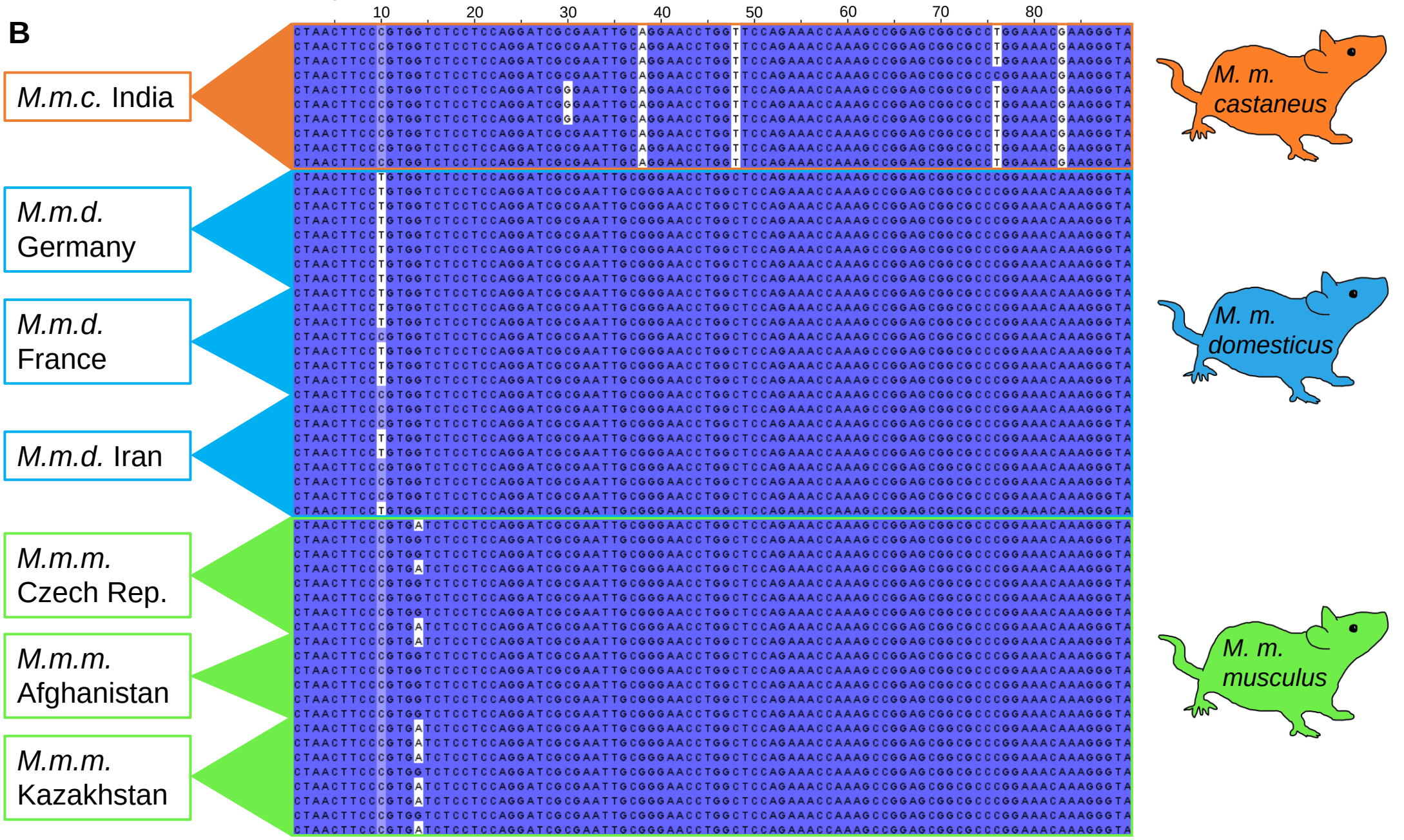
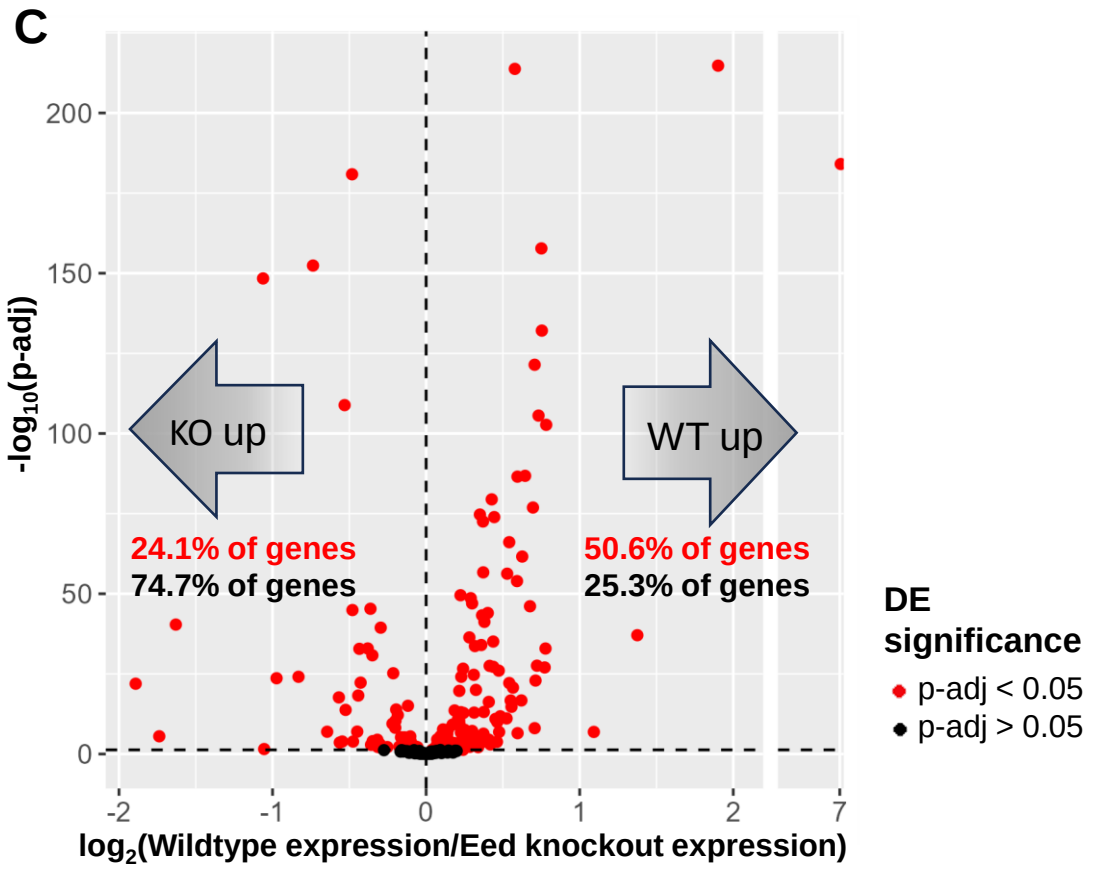
B

M. m. castaneus population vs other population:		-25kb to -20kb Normalized D_u/P_u	Resampling *P*
M. m. musculus	Czech Rep.	***1.168***	***0.027***
	Afghanistan	***1.202***	***0.013***
	Kazakhstan	***1.153***	***0.034***
M. m. domesticus	Iran	***1.176***	***0.032***
	Germany	1.012	0.430
	France	1.076	0.140

A

M. m. castaneus population vs other population:		D_u/P_u in Eed binding sites: Translation GO genes	D_u/P_u in Eed binding sites: Rest of genome	D_u/P_u in remaining sequence: Translation GO genes	D_u/P_u in remaining sequence: Rest of genome
M. m. domesticus	France	0.133	0.061	0.063	0.060
	Germany	0.128	0.061	0.057	0.060
	Iran	0.080	0.040	0.035	0.035
M. m. musculus	Afghanistan	0.106	0.031	0.048	0.036
	Czech Rep.	0.136	0.047	0.057	0.045
	Kazakhstan	0.119	0.043	0.056	0.043

B



C