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: 4,867

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. We then mined population-genomic sequences to pursue underlying regulatory mechanisms for this expression divergence, finding robust evidence for alleles unique to *M. m. castaneus* at the upstream regions of the translation genes. We interpret our data under a model of changes in lineage-specific pressures across *Mus musculus* in stem cells with high translational capacity. 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 (Bullard et al. 2010; Fraser et al. 2010; Fraser et al. 2011; 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 conclusions require 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 (Fraser et al. 2012; Martin 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 molecular-evolution inference (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; Beamer et al. 1999; Johnson et al. 2006; 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), including a particularly avid differentiation phenotype by *M. m. castaneus* stem cells (Ortmann et al. 2020). 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 strategy with sequence-based follow-up, and they shed new light on evolutionary and regulatory mechanisms of an ancient divergence in mouse stem cell expression programs.

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 the embryonic stem

cell line 129Cas, from a male blastocyst from a cross between CAST/EiJ male x 129/SvImJ female, and homozygous induced pluripotent stem cells.

For validation of the role of candidate transcription factors in translation regulation in stem cells, we analyzed published transcriptional data from embryonic stem cells with *Ctr9* knocked down by shRNA (Ruan et al. 2023; NCBI accession GSE219206, samples SRR22493339, SRR22493340, SRR22493341, and SRR22493342) 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 (Marks et al. 2015) and CAST/EiJ x C57BL/6J (Werner et al. 2017) hybrid stem cell transcriptomes, 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). Data from our in-house cultured hybrid stem cells (see below) were mapped in the same manner.

For analysis of transcriptomes of homozygous CAST/EiJ and C57BL/6J stem cells cultured in-house (see below), we mapped the raw reads to the corresponding reference genome for the respective strain. For analysis 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 raw reads from EBI's ArrayExpress database (Sarkans et al. 2018) and mapped these to the respective reference genome. For validation of the role of candidate transcription factors in translation regulation in stem cells, we mapped RNA-seq from *Ctr9* knockdown embryonic stem cells (NCBI accession GSE219206; Ruan et al. 2023) and their corresponding shRNA control samples to the 129/SvImJ genome.

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 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

In-house stem cell expression profiling experiments used material as follows. A 129S6/SvEv embryonic stem cell line was obtained from Millipore Sigma,

(Catalog no. SCR012, Burlington, MA, USA). The hybrid embryonic stem cell line was 129Cas, from a male blastocyst from a cross between CAST/EiJ male x 129/SvImJ female, a kind gift from Joost Gribnau. 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.

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 of CAST, four replicates each of 129 and 129Cas) 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 make two haploid pseudogenomes from each individual mouse in the context of the C57BL/6J reference genome (GRCm38) using the ‘consensus’ command from bcftools software (Danecek et al. 2021), one incorporating the alternate allele at each heterozygous site using the default options, and the other incorporating the reference allele at each heterozygous site using the ‘--haplotype R' option. We used each pseudogenome separately as input into downstream analyses as described next.

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. We retained for analysis only those CDSs that 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.

In a public resource of chromatin immunoprecipitation sequencing data sets (Kolmykov et al. 2021), we tabulated binding sites for 680 transcriptional regulators across the mouse genome. For each regulator, we identified binding sites that were within 50kb upstream of a genic transcription start site (TSS). We analyzed these sites in a comparison of the pseudogenomes from each *M. m. domesticus* or *M. m. musculus* population in turn against the *M. m. castaneus* population as follows. For a given gene and regulator, using the complete set of pseudogenomes for which heterozygote positions had been assigned as the reference allele, we tabulated the number of nucleotide positions in the regulator's binding sites within 50kb of the TSS that were polymorphic across *M. m. castaneus* and/or the respective sister subspecies population, P_u. We repeated this calculation using the pseudogenomes in which the heterozygote positions had been assigned as the alternate allele, and we took the mean of the two P_u values as a final estimate of the polymorphism of the regulator's bound site. We used an analogous pipeline to count positions in the regulator's binding sites that were divergent between *M. m. castaneus* and the respective sister subspecies population and fixed within each, D_u. Next, we identified all nucleotide positions within 50kb of the TSS

that did not fall into binding sites for the regulator, and we calculated P_u and D_u as above. We carried out these calculations for all genes and all populations. We then used the resulting values, per regulator, as input into a two-factor ANOVA, testing for an interaction between binding site identity and gene membership in the translation GO term, using the populations of *M. m. domesticus* and *M. m. musculus* as replicates. We applied this approach to each transcriptional regulator and corrected for multiple testing with the Benjamini-Hochberg method (Benjamini and Hochberg 1995). Significant ANOVA results are reported in Table 1; full results are in Table S7. Multiple sequence alignment was edited for display using Jalview software (Waterhouse et al. 2009). Binding site data for the top four hits from this analysis with an odds ratio > 1 are reported in Table S8. To check whether the D_u/P_u enrichment in binding sites of our top-scoring regulator, Ctr9, was driven by patterns of divergence or polymorphism, we repeated the two-factor ANOVA test as above but used as input D_u by itself and, separately, P_u by itself (normalized by binding site length), rather than the ratio between them.

Results

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

As a 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; Yang et al. 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 expression divergence between *M. m. castaneus* 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 2A-B and Figure S1C-D). The trend persisted in comparisons between CAST homozygous stem cells and those of other mouse lineages (Figure 2A 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 stronger effect of divergence attributable to *trans*-acting regulators. Given these signatures of directional *cis*- and *trans*-acting variation between *Mus* subspecies, we considered the translation gene cohort to be an informative target for mechanistic and molecular-evolution follow-up.

Unique alleles in *M. m. castaneus* at regulatory loci of translation genes

To dissect regulatory and evolutionary mechanisms of the expression divergence in translation genes between *M. m. castaneus* and other lineages, we focused on sequence changes that could serve as candidate determinants of *cis*-regulatory variation, upstream of coding regions. For this purpose we

developed a screening pipeline for single-nucleotide variants from wild-caught mice (Harr et al. 2016), which we classified on the basis of upstream binding sites by transcriptional regulators (Kolmykov et al. 2021). For the sites bound by a given regulator, we evaluated divergence between *M. m. castaneus* and its relatives, normalized by intra-subspecies polymorphism, across the genes of our translation cohort, in light of the classic interpretation of this metric as a hallmark of positive selection. Results revealed robust differences in normalized divergence between the translation genes and a genomic null in the binding sites of each of seven regulators (Table 1). In four of these top-scoring cases, binding sites at translation genes were enriched for normalized divergence, indicative of an excess of derived alleles that distinguished *M. m. castaneus* from *M. m. domesticus* and *M. m. musculus* (Table 1). All four such screen hits were well-studied regulators of cell identity and differentiation (Ctr9, Rfx6, Hoxa11, and Ehmt2). An additional three regulators emerging from our population-genomic screen had significantly low normalized divergence between *M. m. castaneus* and its relatives in binding sites upstream of translation genes, reflecting especially limited variation between subspecies at these positions and/or relaxed constraint within them (Table 1). We conclude that evolutionarily interpretable changes, most notably spikes of alleles unique to *M. m. castaneus* suggestive of positive selection, can be resolved at *cis*-regulatory sites in translation genes. And these divergent variants emerge

across the binding sites for multiple regulators, as expected if evolutionary rewiring in this system involved a complex network of inputs.

Signatures of evolutionary and regulatory impact of Ctr9 site variants at translation genes

As a test case for deeper insights into mouse subspecies variation in the translation pathway, we focused on Ctr9, a core member of the Paf1/RNA polymerase II complex whose binding sites were top-scoring in terms of normalized divergence between subspecies at our focal genes (Table 1). The signal manifested in each comparison between *M. m. castaneus* and a given *M. musculus* relative (Figure 3A-B). Analyses of the components of our normalized metric made clear that the latter result was driven by elevated variation between subspecies per se: divergence between *M. m. castaneus* and its relatives was 1.7-fold higher in Ctr9 sites at translation genes than at control loci (Figure 3C), whereas polymorphism within populations at these sites was almost on par with that of controls (Figure 3D). Given these patterns of unique alleles in *M. m. castaneus* at Ctr9 sites at translation genes, we considered them as particularly likely to contribute to the *cis*-controlled *M. m. castaneus* expression program that we had noted in this pathway in stem cells (Figure 1). Consistent with such a function, Ctr9 sites with sequence divergence between *M. m. castaneus* and its relatives coincided with divergent expression by the *M. m. castaneus* allele in hybrid stem cells, among translation genes (Figure 3E). Furthermore, our inference of Ctr9 as an

activator for the translation pathway was borne out by expression profiles of *Ctr9* knockdown in stem cells of laboratory mice (Ruan et al. 2023), which established a marked directional influence of Ctr9 on translation genes (Figure 3F). Together, these data highlight Ctr9 as a case in which binding sites for a stem cell transcriptional regulator have evolved uniquely at translation genes in *M. m. castaneus*, representing candidate drivers of non-neutral expression divergence in this system.

Discussion

Empirical molecular evolution tests, once developed, must be validated to establish their rigor and utility for the field. A key 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 divergence 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 (Bullard et al. 2010; Fraser et al. 2010; Fraser et al. 2011; 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 findings in mouse subspecies echo 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 *M. m. castaneus* expression program in translation genes could underlie the uniquely avid differentiation behavior by stem cells of this genotype into definitive endoderm (Ortmann et al. 2020). Such a phenotype could well represent an adaptation, as suggested by the expression- and sequence-based signatures of selection we detect for the divergent regulation

at translation genes in *M. m. castaneus*. More broadly, we would expect translation factors to act as part of a complex genetic architecture of stem cell differentiation as it differs between mouse lineages, alongside other variants mapped in this system (Ortmann et al. 2020; Skelly et al. 2020).

Our population-genomic screen highlighted several transcriptional regulators whose binding sites harbor unique alleles in *M. m. castaneus* at translation genes, each of which represents a candidate determinant of the stem cell expression program in this lineage. The case for function of these regulators in stem cells is supported by a deep prior literature from laboratory mice. The top hit from our screen, Ctr9, is an integral member of the Pol II-associating factor 1 complex (Mueller and Jaehning 2002) with a well-characterized role in development and maintaining stem cell identity and pluripotency (Ding et al. 2009; Ruan et al. 2023). Direct binding by Ctr9 to translation genes in stem cells (Rahl et al. 2010; Ding et al. 2020) dovetails with our inference of Ctr9 regulation of this cohort, including its difference between lineages. Given the occupancy of Ctr9 at superenhancers as well as downstream of transcription termination sites in embryonic stem cells (Ding et al. 2020), such elements may ultimately prove to underlie Ctr9 function at translation loci.

Likewise, the methyltransferase Ehmt2, another hit in our screen for unique alleles in *M. m. castaneus* at translation genes, has well-known functions 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 binds

directly to translation genes in stem cells (Mozzetta et al. 2014). Indeed, early embryonic knockout of Ehmt2 (Auclair et al. 2016) had a directional impact on expression of translation genes in stem cell transcriptomes (Figure S2). Our screen hit Hox11a, a well-studied developmental regulator (Yamamoto and Kuroiwa 2003; Boulet and Capecchi 2004; Wong et al. 2004; Horvat-Switzer and Thompson 2005; Kherdjemil et al. 2016; Rux et al. 2016; Leclerc et al. 2023), has been characterized in a similar literature, including descriptions of its function in stem cells (Rux et al. 2016; Song et al. 2020; Leclerc et al. 2023). And Rfx6, also a hit in our screen, has been previously implicated in development in a number of animal systems (Smith et al. 2010; Trott et al. 2020; Lu et al. 2021), including in translation gene regulation (Cheng et al. 2019; Lu et al. 2021). The emerging picture from our analyses is thus that Ctr9, Rfx6, Hoxa11, and Ehmt2 may exert their effects on stem cell identity and differentiation in part via regulation of protein synthesis, and that binding sites for these regulators represent candidate players in the mechanism by which evolution has tuned expression of translation factors in *M. m. castaneus*. That said, we expect that our genomic approach affords only partial coverage and power in discovering elements of this mechanism, and that many other contributing regulators likely remain to be identified.

In summary, our discovery of a unique expression program in *M. m. castaneus* stem cells, at translation genes which also harbor divergent binding sites for a suite of regulators, represents an informative molecular-evolution case study.

And our results serve as a foundation for future work at the physiological level, to pursue the relevance of translation gene regulation in stem-cell differentiation (Ortmann et al. 2020) and other phenotypes (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 Kaylee Christensen for assistance with computational resources and 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, Pașca SP, Fraser HB. 2021. Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. Nature. 592(7854):421–427. doi:10.1038/s41586-021-03343-3.

Auclair G, Borgel J, Sanz LA, Vallet J, Guibert S, Dumas M, Cavelier P, Girardot M, Forné T, Feil R, et al. 2016. EHMT2 directs DNA methylation for efficient gene silencing in mouse embryos. Genome Res. 26(2):192–202. doi:10.1101/gr.198291.115.

Beamer WG, Shultz KL, Churchill GA, Frankel WN, Baylink DJ, Rosen CJ, Donahue LR. 1999. Quantitative trait loci for bone density in C57BL/6J and CAST/EiJ inbred mice. Mammalian Genome. 10(11):1043–1049. doi:10.1007/s003359901159.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 8:e39725. doi:10.7554/eLife.39725.

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(6):513–525. doi:10.1038/ncb2965.

Boulet AM, Capecchi MR. 2004. Multiple roles of Hoxa11 and Hoxd11 in the formation of the mammalian forelimb zeugopod. Development. 131(2):299–309. doi:10.1242/dev.00936.

Brown SJ, Cole MD, Erives AJ. 2008. Evolution of the holozoan ribosome biogenesis regulon. BMC Genomics. 9(1):442. doi:10.1186/1471-2164-9-442.

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(11):5058–5063. doi:10.1073/pnas.0912959107.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, the AmiGO Hub, the Web Presence Working Group. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics. 25(2):288–289. doi:10.1093/bioinformatics/btn615.

Chapman VM, Ruddle FH. 1972. GLUTAMATE OXALOACETATE TRANSAMINASE (GOT) GENETICS IN THE MOUSE: POLYMORPHISM OF GOT-1. Genetics. 70(2):299–305. doi:10.1093/genetics/70.2.299.

Chappell GA, Israel JW, Simon JM, Pott S, Safi A, Eklund K, Sexton KG, Bodnar W, Lieb JD, Crawford GE, 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(10):107006. doi:10.1289/EHP1937.

Cheng C, Lu J, Cao X, Yang F, Liu J, Song L, Shen H, Liu C, Zhu X, Zhou J, et al. 2019. Identification of *Rfx6* target genes involved in pancreas development and insulin translation by ChIP-seq. Biochemical and Biophysical Research Communications. 508(2):556–562. doi:10.1016/j.bbrc.2018.11.133.

Chou M-Y, Appan D, Chang K-W, Chou C-H, Lin C-Y, Gau SS-F, Huang H-S. 2022. Mouse hybrid genome mediates diverse brain phenotypes with the specificity of reciprocal crosses. FASEB J. 36(3):e22232. doi:10.1096/fj.202101624R.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 27(24):3423–3424. doi:10.1093/bioinformatics/btr539.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience. 10(2):giab008. doi:10.1093/gigascience/giab008.

Ding L, Paszkowski-Rogacz M, Mircetic J, Chakraborty D, Buchholz F. 2020. The Paf1 complex positively regulates enhancer activity in mouse embryonic stem cells. Life Sci Alliance. 4(3):e202000792. doi:10.26508/lsa.202000792.

Ding L, Paszkowski-Rogacz M, Nitzsche A, Slabicki MM, Heninger A-K, Vries I de, Kittler R, Junqueira M, Shevchenko A, Schulz H, et al. 2009. A Genome-Scale RNAi Screen for Oct4 Modulators Defines a Role of the Paf1 Complex for Embryonic Stem Cell Identity. Cell Stem Cell. 4(5):403–415. doi:10.1016/j.stem.2009.03.009.

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(8):1184–1191. doi:10.1038/nprot.2009.97.

Easley CA, Ben-Yehudah A, Redinger CJ, Oliver SL, Varum ST, Eisinger VM, Carlisle DL, Donovan PJ, Schatten GP. 2010. mTOR-Mediated Activation of p70 S6K Induces Differentiation of Pluripotent Human Embryonic Stem Cells. Cellular Reprogramming. 12(3):263–273. doi:10.1089/cell.2010.0011.

Ferguson J, Chang J. 2020. An empirical Bayesian ranking method, with applications to high throughput biology. Bioinformatics. 36(1):177–185. doi:10.1093/bioinformatics/btz471.

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(3):e1002023. doi:10.1371/journal.pgen.1002023.

Fraser HB, Levy S, Chavan A, Shah HB, Perez JC, Zhou Y, Siegal ML, Sinha H. 2012. Polygenic cis-regulatory adaptation in the evolution of yeast pathogenicity. Genome Res. 22(10):1930–1939. doi:10.1101/gr.134080.111.

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(7):2977–2982. doi:10.1073/pnas.0912245107.

Frazee AC, Jaffe AE, Langmead B, Leek JT. 2015. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 31(17):2778–2784. doi:10.1093/bioinformatics/btv272.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 448(7157):1050–1053. doi:10.1038/nature06067.

French JE, Gatti DM, Morgan DL, Kissling GE, Shockley KR, Knudsen GA, Shepard KG, Price HC, King D, Witt KL, et al. 2015. Diversity Outbred Mice Identify Population-Based Exposure Thresholds and Genetic Factors that Influence Benzene-Induced Genotoxicity. Environ Health Perspect. 123(3):237–245. doi:10.1289/ehp.1408202.

Gabut M, Bourdelais F, Durand S. 2020. Ribosome and Translational Control in Stem Cells. Cells. 9(2):497. doi:10.3390/cells9020497.

Gokhman D, Agoglia RM, Kinnebrew M, Gordon W, Sun D, Bajpai VK, Naqvi S, Chen Coral, Chan A, Chen Chider, et al. 2021. Human-chimpanzee fused

cells reveal cis-regulatory divergence underlying skeletal evolution. Nat Genet. 53(4):467–476. doi:10.1038/s41588-021-00804-3.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, Babiker H, Linnenbrink M, Montero I, Scavetta R, et al. 2016. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci Data. 3(1):160075. doi:10.1038/sdata.2016.75.

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A, Whiteway M. 2008. Transcription Factor Substitution during the Evolution of Fungal Ribosome Regulation. Mol Cell. 29(5):552–562. doi:10.1016/j.molcel.2008.02.006.

Horvat-Switzer RD, Thompson AA. 2005. HoxA11 Is Expressed in the Developing Hemangioblast as Well as Early Hematopoietic Precursor Stem Cells. Blood. 106(11):4206. doi:10.1182/blood.V106.11.4206.4206.

Hsiao K, Noble C, Pitman W, Yadav N, Kumar S, Keele GR, Terceros A, Kanke M, Conniff T, Cheleuitte-Nieves C, et al. 2020. A Thalamic Orphan Receptor Drives Variability in Short-Term Memory. Cell. 183(2):522-536.e19. doi:10.1016/j.cell.2020.09.011.

Ikegami K, Iwatani M, Suzuki M, Tachibana M, Shinkai Y, Tanaka S, Greally JM, Yagi S, Hattori N, Shiota K. 2007. Genome-wide and locus-specific DNA hypomethylation in G9a deficient mouse embryonic stem cells. Genes to Cells. 12(1):1–11. doi:10.1111/j.1365-2443.2006.01029.x.

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(1):83–92. doi:10.1016/S0378-5955(97)00155-X.

Johnson KR, Zheng QY, Noben-Trauth K. 2006. Strain background effects and genetic modifiers of hearing in mice. Brain Research. 1091(1):79–88. doi:10.1016/j.brainres.2006.02.021.

Johri P, Charlesworth B, Jensen JD. 2020. Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection. Genetics. 215(1):173–192. doi:10.1534/genetics.119.303002.

Khajuria RK, Munschauer M, Ulirsch JC, Fiorini C, Ludwig LS, McFarland SK, Abdulhay NJ, Specht H, Keshishian H, Mani DR, et al. 2018. Ribosome Levels Selectively Regulate Translation and Lineage Commitment in Human Hematopoiesis. Cell. 173(1):90-103.e19. doi:10.1016/j.cell.2018.02.036.

Kherdjemil Y, Lalonde RL, Sheth R, Dumouchel A, de Martino G, Pineault KM, Wellik DM, Stadler HS, Akimenko M-A, Kmita M. 2016. Evolution of Hoxa11 regulation in vertebrates is linked to the pentadactyl state. Nature. 539(7627):89–92. doi:10.1038/nature19813.

Kim K, Ryu TY, Ryu J-W, Han T-S, Jung C-R, Son M-Y, Kim D-S, Cho H-S. 2020. RNA-seq based transcriptome analysis of EHMT2 functions in breast cancer. Biochemical and Biophysical Research Communications. 524(3):672–676. doi:10.1016/j.bbrc.2020.01.128.

Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev VJ, Kulakovskiy IV, Kel A, Kolpakov F. 2021. GTRD: an integrated view of transcription regulation. Nucleic Acids Research. 49(D1):D104–D111. doi:10.1093/nar/gkaa1057.

Koturbash I, Scherhag A, Sorrentino J, Sexton K, Bodnar W, Swenberg JA, Beland FA, Pardo-Manuel deVillena F, Rusyn I, Pogribny IP. 2011. Epigenetic Mechanisms of Mouse Interstrain Variability in Genotoxicity of the Environmental Toxicant 1,3-Butadiene. Toxicological Sciences. 122(2):448–456. doi:10.1093/toxsci/kfr133.

Kreitman M. 2000. Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet. 1:539–559. doi:10.1146/annurev.genom.1.1.539.

Leclerc K, Remark LH, Ramsukh M, Josephson AM, Palma L, Parente PEL, Sambon M, Lee S, Lopez EM, Morgani SM, et al. 2023. Hox genes are crucial regulators of periosteal stem cell identity. Development. 150(6):dev201391. doi:10.1242/dev.201391.

Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N, Mansfield W, Nashun B, Knezovich JG, Smith A, et al. 2013. Naive pluripotency is associated with global DNA hypomethylation. Nat Struct Mol Biol. 20(3):311–316. doi:10.1038/nsmb.2510.

Li XC, Fay JC. 2017. Cis-Regulatory Divergence in Gene Expression between Two Thermally Divergent Yeast Species. Genome Biol Evol. 9(5):1120–1129. doi:10.1093/gbe/evx072.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 15(12):550. doi:10.1186/s13059-014-0550-8.

Lu J, Cheng C, Cheng Z-C, Wu Q, Shen H, Yuan M-X, Zhang B, Yang J-K. 2021. The dual role of RFX6 in directing β cell development and insulin production. J Mol Endocrinol. 66(2):129–140. doi:10.1530/JME-20-0119.

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(3):msad034. doi:10.1093/molbev/msad034.

Marks H, Kerstens HHD, Barakat TS, Splinter E, Dirks RAM, van Mierlo G, Joshi O, Wang S-Y, Babak T, Albers CA, et al. 2015. Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. Genome Biology. 16(1):149. doi:10.1186/s13059-015-0698-x.

Martin HC, Roop JI, Schraiber JG, Hsu TY, Brem RB. 2012. Evolution of a Membrane Protein Regulon in Saccharomyces. Molecular Biology and Evolution. 29(7):1747–1756. doi:10.1093/molbev/mss017.

Mozzetta C, Pontis J, Fritsch L, Robin P, Portoso M, Proux C, Margueron R, Ait-Si-Ali S. 2014. The histone H3 lysine 9 methyltransferases G9a and GLP regulate polycomb repressive complex 2-mediated gene silencing. Mol Cell. 53(2):277–289. doi:10.1016/j.molcel.2013.12.005.

Mueller CL, Jaehning JA. 2002. Ctr9, Rtf1, and Leo1 Are Components of the Paf1/RNA Polymerase II Complex. Mol Cell Biol. 22(7):1971–1980. doi:10.1128/MCB.22.7.1971-1980.2002.

Omura T, Omura K, Tedeschi A, Riva P, Painter MW, Rojas L, Martin J, Lisi V, Huebner EA, Latremoliere A, et al. 2015. Robust Axonal Regeneration Occurs in the Injured CAST/Ei Mouse CNS. Neuron. 86(5):1215–1227. doi:10.1016/j.neuron.2015.05.005.

Orr HA. 1998. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. Genetics. 149(4):2099–2104. doi:10.1093/genetics/149.4.2099.

Ortmann D, Brown S, Czechanski A, Aydin S, Muraro D, Huang Y, Tomaz RA, Osnato A, Canu G, Wesley BT, et al. 2020. Naive Pluripotent Stem Cells Exhibit Phenotypic Variability that Is Driven by Genetic Variation. Cell Stem Cell. 27(3):470-481.e6. doi:10.1016/j.stem.2020.07.019.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. F1000Res. 9:ISCB Comm J-304. doi:10.12688/f1000research.23297.2.

Polymenis M, Aramayo R. 2015. Translate to divide: control of the cell cycle by protein synthesis. Microbial Cell. 2(4):94–104. doi:10.15698/mic2015.04.198.

Price PD, Palmer Droguett DH, Taylor JA, Kim DW, Place ES, Rogers TF, Mank JE, Cooney CR, Wright AE. 2022. Detecting signatures of selection on gene expression. Nat Ecol Evol. 6(7):1035–1045. doi:10.1038/s41559-022-01761-8.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841–842. doi:10.1093/bioinformatics/btq033.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc regulates transcriptional pause release. Cell. 141(3):432–445. doi:10.1016/j.cell.2010.03.030.

Roop JI, Chang KC, Brem RB. 2016. Polygenic evolution of a sugar specialization trade-off in yeast. Nature. 530(7590):336–339. doi:10.1038/nature16938.

Ruan Y, Wang Jiaqi, Yu M, Wang F, Wang Jiangjun, Xu Y, Liu L, Cheng Y, Yang R, Zhang C, et al. 2023. A multi-omics integrative analysis based on CRISPR screens re-defines the pluripotency regulatory network in ESCs. Commun Biol. 6:410. doi:10.1038/s42003-023-04700-w.

Rux DR, Song JY, Swinehart IT, Pineault KM, Schlientz AJ, Trulik KG, Goldstein SA, Kozloff KM, Lucas D, Wellik DM. 2016. Regionally Restricted Hox Function in Adult Bone Marrow-Multipotent Mesenchymal Stem/Stromal Cells. Dev Cell. 39(6):653–666. doi:10.1016/j.devcel.2016.11.008.

Sampath P, Pritchard DK, Pabon L, Reinecke H, Schwartz SM, Morris DR, Murry CE. 2008. A Hierarchical Network Controls Protein Translation during Murine Embryonic Stem Cell Self-Renewal and Differentiation. Cell Stem Cell. 2(5):448–460. doi:10.1016/j.stem.2008.03.013.

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(1):340–347. doi:10.1128/jvi.67.1.340-347.1993.

Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, Minguet J, Rada JC, Snow C, Tikhonov A, et al. 2018. The BioStudies database—one stop shop for all data supporting a life sciences study. Nucleic Acids Research. 46(D1):D1266–D1270. doi:10.1093/nar/gkx965.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. Trends Genet. 34(7):532–544. doi:10.1016/j.tig.2018.03.007.

Skelly DA, Czechanski A, Byers C, Aydin S, Spruce C, Olivier C, Choi K, Gatti DM, Raghupathy N, Keele GR, 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(3):459-469.e8. doi:10.1016/j.stem.2020.07.005.

Smith SB, Qu H-Q, Taleb N, Kishimoto NY, Scheel DW, Lu Y, Patch A-M, Grabs R, Wang J, Lynn FC, et al. 2010. Rfx6 directs islet formation and insulin production in mice and humans. Nature. 463(7282):775–780. doi:10.1038/nature08748.

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N, et al. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 8:e39702. doi:10.7554/eLife.39702.

Song JY, Pineault KM, Dones JM, Raines RT, Wellik DM. 2020. Hox genes maintain critical roles in the adult skeleton. Proceedings of the National Academy of Sciences. 117(13):7296–7304. doi:10.1073/pnas.1920860117.

Sorrells TR, Johnson AN, Howard CJ, Britton CS, Fowler KR, Feigerle JT, Weil PA, Johnson AD. 2018. Intrinsic cooperativity potentiates parallel cis-regulatory evolution. eLife. 7:e37563. doi:10.7554/eLife.37563.

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(20):7203–7208. doi:10.1073/pnas.0502521102.

Terzic D, Maxon JR, Krevitt L, Dibartolomeo C, Goyal T, Low WC, Dutton JR, Parr AM. 2016. Directed Differentiation of Oligodendrocyte Progenitor Cells from Mouse Induced Pluripotent Stem Cells. Cell Transplant. 25(2):411–424. doi:10.3727/096368915X688137.

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(1):113. doi:10.1186/s12920-019-0560-1.

Trott J, Alpagu Y, Tan EK, Shboul M, Dawood Y, Elsy M, Wollmann H, Tano V, Bonnard C, Eng S, et al. 2020. Mitchell-Riley syndrome iPSCs exhibit reduced pancreatic endoderm differentiation due to a mutation in RFX6. Dunwoodie S, Wallingford J, editors. Development. 147(21):dev194878. doi:10.1242/dev.194878.

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. doi:10.7554/eLife.89594.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25(9):1189–1191. doi:10.1093/bioinformatics/btp033.

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. doi:10.1186/s13293-017-0150-x.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. Nature. 430(6995):85–88. doi:10.1038/nature02698.

Wong KHH, Wintch HD, Capecchi MR. 2004. Hoxa11 Regulates Stromal Cell Death and Proliferation during Neonatal Uterine Development. Molecular Endocrinology. 18(1):184–193. doi:10.1210/me.2003-0222.

Xiong H, Morrison J, Ferris MT, Gralinski LE, Whitmore AC, Green R, Thomas MJ, Tisoncik-Go J, Schroth GP, Pardo-Manuel de Villena F, 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(8):1429–1444. doi:10.1534/g3.114.011759.

Yamamoto M, Kuroiwa A. 2003. Hoxa-11 and Hoxa-13 are involved in repression of MyoD during limb muscle development. Development, Growth & Differentiation. 45(5–6):485–498. doi:10.1111/j.1440-169X.2003.00715.x.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. 2007. On the subspecific origin of the laboratory mouse. Nat Genet. 39(9):1100–1107. doi:10.1038/ng2087.

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH-T, Nachman MW, Pialek J, et al. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet. 43(7):648–655. doi:10.1038/ng.847.

York RA, Patil C, Abdilleh K, Johnson ZV, Conte MA, Genner MJ, McGrath PT, Fraser HB, Fernald RD, Streelman JT. 2018. Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes. Proc Natl Acad Sci U S A. 115(47):E11081–E11090. doi:10.1073/pnas.1810140115.

Yu H, Mohan S, Edderkaoui B, Masinde GL, Davidson HM, Wergedal JE, Beamer WG, Baylink DJ. 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(2):103–110. doi:10.1007/s00223-006-0187-z.

Figure captions

Figure 1. Directional *cis*-regulatory variation in translation gene expression in stem cells between mouse subspecies. (A) Results from analyses of differential allele-specific expression across genes from the Gene Ontology term GO:0006412, translation, in stem cells of F1 hybrids between CAST and either C57BL/6 (from Werner et al. 2017) or 129 (Marks et al. 2015 or this study) 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 (p-adj, adjusted *p*-value). Point colors report significance of differential allele-specific expression (red, adjusted $p < 0.05$; black, adjusted $p > 0.05$). 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.

Figure 2. Directional *trans*-regulatory variation in translation gene expression in stem cells between mouse subspecies. (A) Data are as in Figure 1A except that each column reports results from comparisons between homozygous CAST and homozygous 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. **(B)** Data are as in Figure 1B except that each point reports expression of a translation gene in a comparison between homozygous CAST and 129 stem.

Figure 3. Signatures of species divergence and regulatory function of Ctr9 at translation genes. (A) In a given row, each colored cell reports normalized divergence, the ratio of inter-subspecies divergence to within-subspecies polymorphism, at Ctr9 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. FRA, France; GER, Germany; IRA, Iran; AFG, Afghanistan; CZE, Czech Republic; KAZ, Kazakhstan. Cells are colored based on relative values. Values at the bottom report the results of an ANOVA testing for differences in normalized divergence between translation genes and controls (see Table 1 and Table S7). *P-adj*, screen *p*-value after correction for multiple testing; OR, odds ratio. **(B)** A portion of a Ctr9 binding footprint upstream of the translation gene Rps29 as an example of mouse subspecies variation. The *x*-axis reports position relative to the transcription start site. Each row reports the sequence of one mouse individual; at heterozygote loci, only the allele representing the alternate to the reference is shown, although our analysis methods accounted for both (see Methods). **(C)** Data are as in A, except that each value reports the number of sites divergent between *M. m.*

castaneus and the indicated sister population in the region of interest, standardized by the length of the latter. **(D)** Data are as in A, except that each value reports the number of sites polymorphic within *M. m. castaneus* and/or the indicated sister population in the region of interest, standardized by the length of the latter. **(E)** Shown is an analysis of mouse subspecies divergence variation at Ctr9 binding sites, in genes stratified by membership in the translation gene cohort and allele-specific expression differences between CAST and 129 alleles in F1 hybrid stem cells (ASE; Marks et al. 2015; Table S5). For a given bar, the *y*-axis reports the percentage of genes in the indicated category containing variants in Ctr9 binding sites divergent between *M. m. castaneus* and one or more sister taxon (Table S8). *, Fisher's exact test $p = 0.02$. **(F)** Data and symbols are as in Figure 1B except that each point reports a comparison of expression of one translation gene between *Ctr9* knockdown and shRNA control mouse stem cells (Ruan 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.

Table captions

Table 1. Patterns of variation between mouse subspecies in binding sites of transcriptional regulators upstream of translation genes. Shown are results of analysis of normalized sequence divergence, at the regions upstream of genes, between *M. m. castaneus* on the one hand and *M. m. musculus* and *M. m. domesticus* on the other. Each row reports results of an ANOVA testing for an interaction effect between two factors on the divergence metric: position of sequence variants in binding sites of the indicated transcriptional regulator and gene membership in the translation GO term (see Figure 3A). Normalized divergence was calculated as the number of sites divergent between *M. m. castaneus* and a given *M. musculus* relative, normalized by the number of sites polymorphic within the populations. The second and third columns report the odds ratio (OR) and nominal *p*-value, respectively, and the third reports the *p*-value after Benjamini-Hochberg correction for multiple testing. OR > 1 indicates enrichment of normalized divergence at regulator binding sites relative to the other three categories, and OR < 1 indicates depletion. Full results are shown in Table S7.

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_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 (Marks et al. 2015). Tab TPM_GSE234761_CAST_M_x_129_F contains results from four replicate transcriptomes of 129Cas (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 129 and CAST/EiJ genotypes from this

study; embryonic stem cells of eight genotypes from (Skelly et al. 2020); *Ctr9* knockdown in embryonic stem cells (Ruan 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_GSE219206_CTR9_KD contains results from two replicate transcriptomes of control shRNA ESC samples and two replicate transcriptomes of shRNA knockdown of *Ctr9,* Tab TPM_GSE71500_EHMT2_KO contains results from two replicate transcriptomes of wild-type 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_GSE219206_Ctr9_KD, comparison of control shRNA vs *Ctr9*

knockdown expression in stem cells from (Ruan 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. Patterns of variation between mouse subspecies in binding sites of transcriptional regulators upstream of translation genes. Data are as in Table 1 of the main text, except that results from all tested regulators are shown.

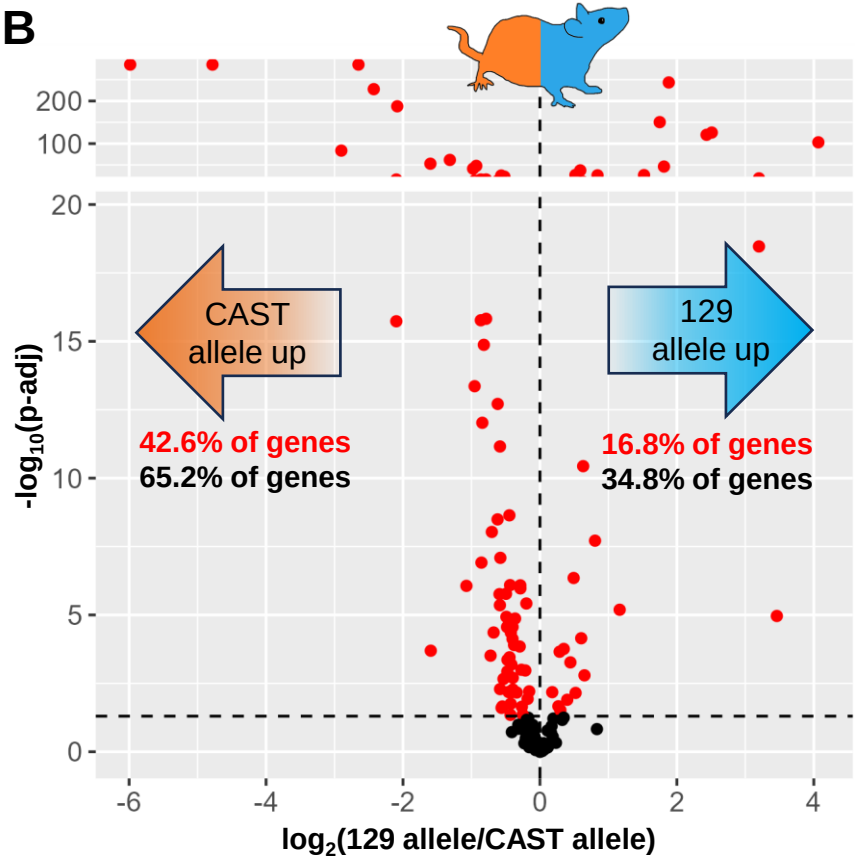
Table S8. Counts of divergent and polymorphic sites in binding sites of top four candidate transcriptional regulators. Each row reports counts upstream of a gene, of sequence variation between wild mouse populations in a binding site for Ctr9, Rfx6, Hoxa11, or Ehmt2 from the GTRD database

(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		
	CAST x129 this study	CAST x129 Marks et al	CAST xBL6 Werner et al
% sig genes CAST up	42.6	15.3	43.7
% sig genes CAST down	16.8	7.6	17.6
(CAST up/ CAST down)	**2.54**	**2**	**2.48**

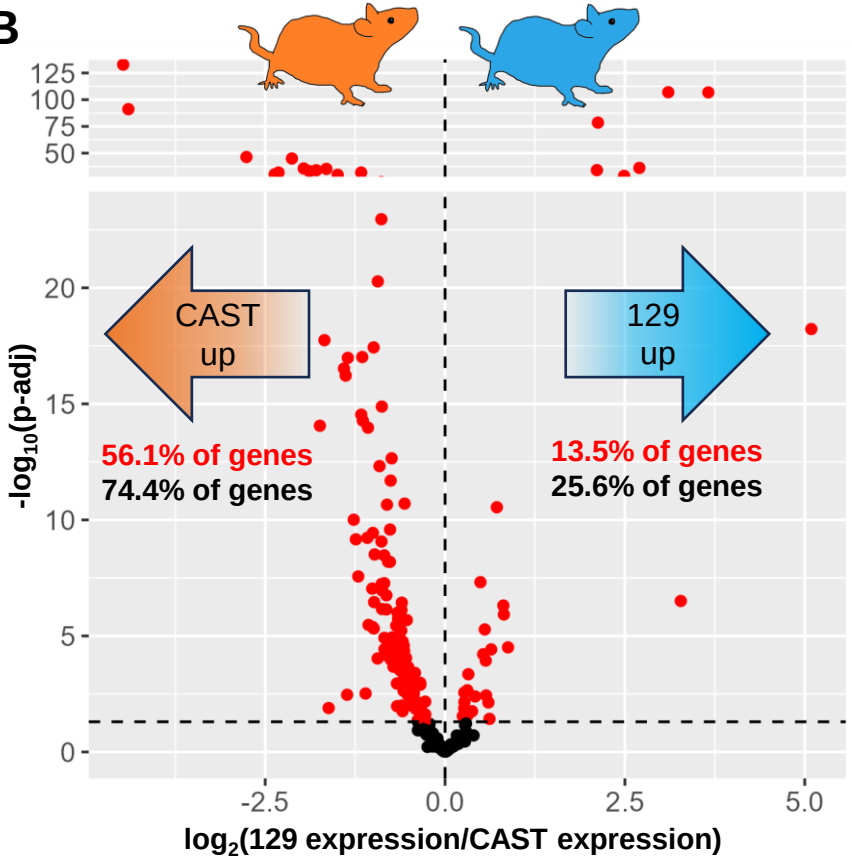
B



A

	Homozygous stem cells							
	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	56.1	65.7	66.1	87.2	36.1	48.0	64.4	51.6
% sig genes CAST down	13.5	8.4	19.2	8.8	8.4	9.8	16.4	11.6
(CAST up/ CAST down)	**4.17**	**7.85**	**3.43**	**9.90**	**4.32**	**4.91**	**3.92**	**4.46**

B



A Divergence/polymorphism

M.m.c. vs:		Ctr9 target sites: Transl. genes	Ctr9 target sites: Rest of genome	Rest of upstream: Transl. genes	Rest of upstream: Rest of genome
M.m.d.	FRA	0.113	0.074	0.070	0.069
	GER	0.095	0.067	0.068	0.067
	IRA	0.083	0.047	0.050	0.046
M.m.m.	AFG	0.103	0.050	0.059	0.053
	CZE	0.098	0.063	0.071	0.067
	KAZ	0.107	0.059	0.070	0.067

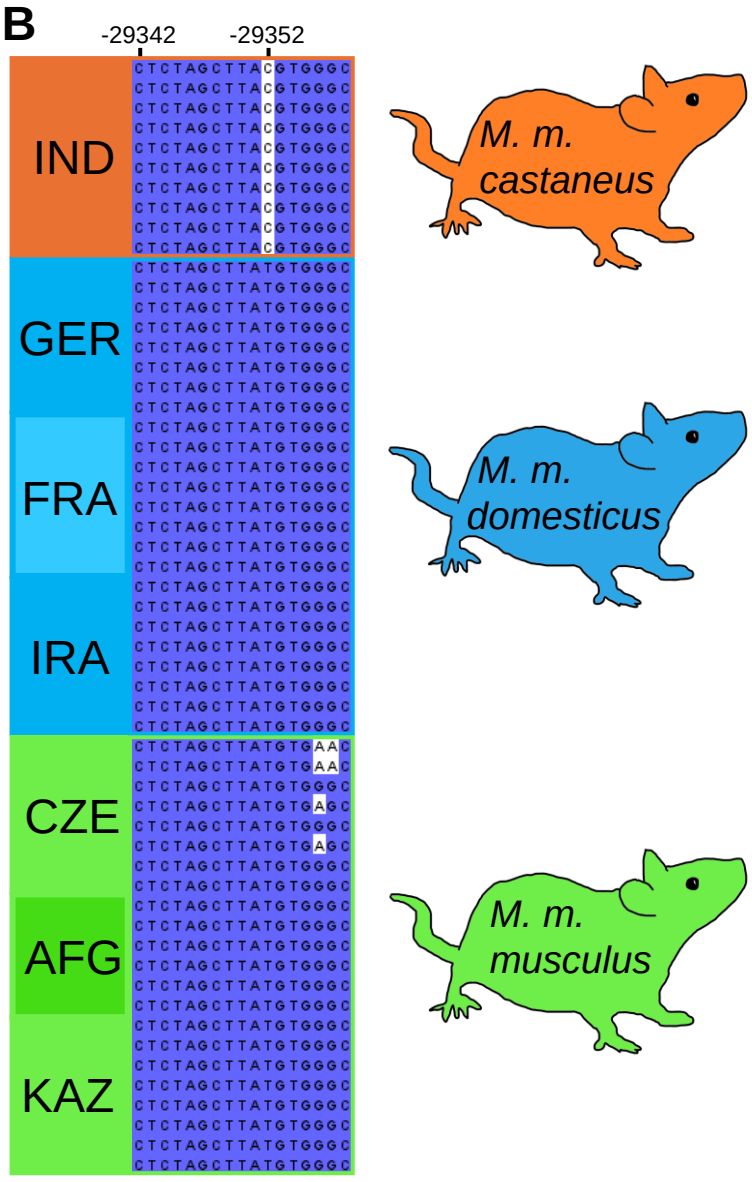
P-adj = 0.033
OR = 1.47

C Divergence/length

M.m.c. vs:		Ctr9 target sites: Transl. genes	Ctr9 target sites: Rest of genome	Rest of upstream: Transl. genes	Rest of upstream: Rest of genome
M.m.d.	FRA	0.00209	0.00131	0.00120	0.00121
	GER	0.00186	0.00121	0.00118	0.00118
	IRA	0.00177	0.00090	0.00092	0.00087
M.m.m.	AFG	0.00200	0.00090	0.00102	0.00094
	CZE	0.00186	0.00112	0.00120	0.00116
	KAZ	0.00200	0.00106	0.00119	0.00116

$P < 10^{-6}$
OR = 1.68

B

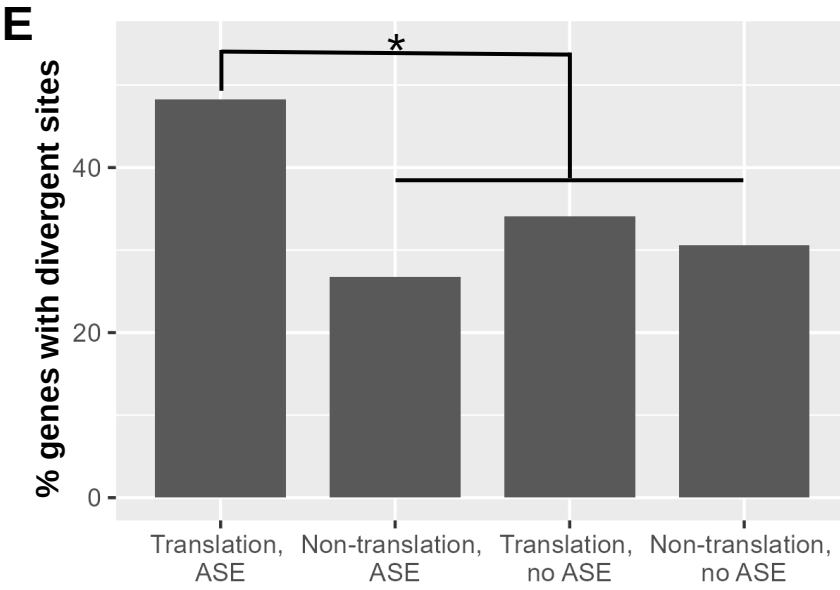
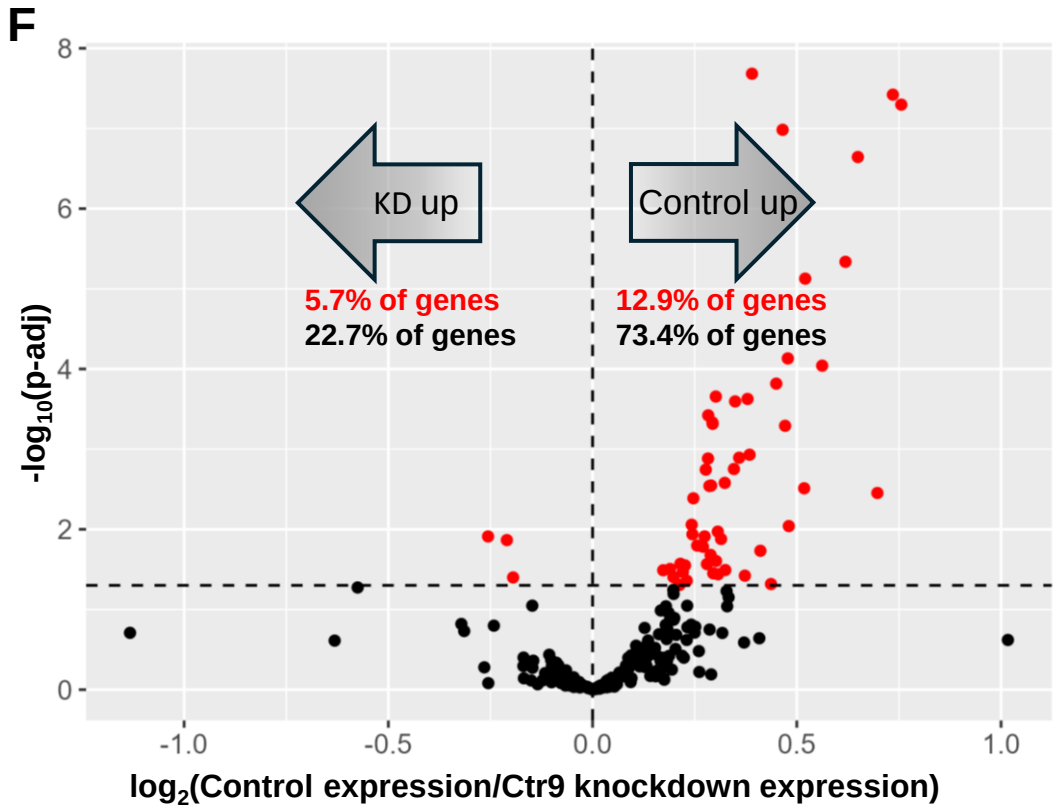


D Polymorphism/length

M.m.c. vs:		Ctr9 target sites: Transl. genes	Ctr9 target sites: Rest of genome	Rest of upstream: Transl. genes	Rest of upstream: Rest of genome
M.m.d.	FRA	0.0185	0.0178	0.0171	0.0176
	GER	0.0195	0.0181	0.0173	0.0177
	IRA	0.0212	0.0193	0.0186	0.0189
M.m.m.	AFG	0.0194	0.0181	0.0174	0.0177
	CZE	0.0191	0.0176	0.0169	0.0173
	KAZ	0.0187	0.0179	0.0170	0.0174

P = 0.0097
OR = 1.14

E



F