Stem cell transcriptional profiles from mouse subspecies reveal *cis*-regulatory evolution at ribosomal large-subunit 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: 6,323

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 and benchmarking of advanced molecular-evolution analysis methods. Here we evaluated the rigor and performance of a test of directional, *cis*-regulatory evolution across genes in pathways, using stem cells from *Mus musculus* subspecies as a model. We discovered a unique program of induction of ribosomal large-subunit genes in stem cells of the Southeast Asian mouse *M. m. castaneus* relative to its sister taxa, driven in part by *cis*-regulatory variation. As a complement, we used sequence analyses to find population- and comparative-genomic signatures of selection in *M. m. castaneus,* at the upstream regions of the ribosomal gene cohort. The unique high-expression program for these loci in *M. m. castaneus* was recapitulated in stem cell-derived neurons. 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). As a complement, empirical genome ranking/bootstrapping approaches have emerged in the more recent literature (Ferguson and Chang 2020), with development and refinement remaining 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 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 (Fraser et al. 2012; Martin et al. 2012; Roop et al. 2016); in metazoans, the rigor and utility of expression-based *cis*-regulatory pathway analyses remain to be fully validated (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). 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 ribosomal 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, phylogenetic and population-genomic tests revealed signals of unique evolution in *M. m. castaneus*. Together, our results

validate 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 CAST/EiJ male x 129/SvImJ female embryonic stem cells (NCBI accession GSE114745, all samples; Rivera-Mulia et al. 2018); 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); purebred 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 purebred induced pluripotent stem cells and differentiated neurons from in-house cultures (see below).

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) and (Rivera-Mulia et al. 2018), we downloaded raw reads and mapped them using TopHat 2.1.1 (Kim and Salzberg 2011) 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 purebred 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). Reads that mapped ambiguously to more than one locus were discarded. Mapped reads were sorted with Samtools (Danecek et al. 2021) and gene counts were generated using htseq-count (Anders et al. 2015). For analysis of transcriptomes of purebred 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 from stem cell RNA-seq datasets are reported in Tables S1 and S2.

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 S3) 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; Table S4). We eliminated from further analysis genes with fewer than 10 total reads across all samples. For each of the remaining genes with differential allele-specific expression at adjusted $P < 0.05$, we assigned a qualitative sign statistic s ($s = 1$ if the CAST/EiJ allele had a higher read count than the 129/SvImJ allele, and $s = -1$ for the inverse case). 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 S5. All further expression and molecular-evolution analyses focused on the ten genes with significant differential allele-specific expression between the CAST/EiJ and 129/SvImJ alleles in the top-scoring term, ribosomal large subunit assembly (Figure 1).

Induced pluripotent stem cell derivation and culture

We utilized mouse induced pluripotent stem cell lines derived at the UMN Stem Cell Institute as previously described (Terzic et al. 2016) from primary cultures of C57BL/6J and CAST/EiJ E13.5 mouse embryonic fibroblasts obtained from Jackson Laboratories (Bar Harbor, ME, USA), using 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.

Induced pluripotent 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.

Pluripotent stem cell differentiation

Motor neuron differentiation from C57BL/6J and CAST/EiJ E13.5 induced pluripotent stem cells was performed essentially as described (Wu et al. 2012). Briefly, differentiation was conducted in embryoid bodies in ultra-low attachment plates (Corning, Corning, NY, USA) with orbital shaking at 65 rpm, with half media changes each day. Embryoid bodies were aggregated and cultured in neural differentiation media (1:1 knockout DMEM:DMEM/F12 (Sigma-Aldrich, St. Louis, MO, USA), 5% FBS, 5% KSR, 1% MEM NEAA, 1% GlutaMAX, 0.04 uM BME, and 0.05% Insulin-Transferrin-Selenium [Thermo Fisher Scientific, Waltham, MA, USA]) supplemented with 50 ng/ml Noggin (StemCell Technologies, Vancouver, Canada), 20 ng/ml FGF2 (Thermo Fisher Scientific), and 20 ng/ml FGF8 (Thermo Fisher Scientific) for two days and then with 1 μM retinoic acid (Millipore) and 1 μM smoothened agonist (Millipore) for five days. On day seven, embryoid bodies were dissociated with Accutase (Millipore) and cells were cryopreserved or plated for further differentiation on poly-ornithine (Millipore), laminin (StemCell Technologies)-coated tissue culture plates in neural differentiation media containing BDNF (Thermo Fisher Scientific), GDNF (GenScript, Piscataway, NJ, USA), CNTF (BioLegend, San Diego, CA, USA), and NT3 (PeproTech, Cranbury, NJ, USA), all at 10 ng/mL. 10 μM AraC (Sigma-Aldrich) was added to the media for the first 48 hours and half media changes were performed daily. DMSO (0.1% by volume) was added to the media from days 5 to 7.

RNA isolation and sequencing

RNA was extracted from undifferentiated pluripotent stem cell cultures (2 BL6 replicates, 2 CAST replicates) following feeder depletion and from pluripotent stem cell-derived motor neurons (4 BL6 replicates, 1 CAST replicate) following 7 days of culture on laminin. For pluripotent stem cell samples, we used the RNAqueous™-Micro Total RNA Isolation Kit (Thermo Fisher Scientific). For neuron samples, we used the RNAeasy-Micro Total RNA Isolation Kit (Thermo Fisher Scientific). In each case we used 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 above. Normalized read counts from these stem cell and neuronal RNA-seq datasets are reported in Table S2.

For principal component analysis, we filtered expression data from purebred stem cell and neuron samples (Skelly et al. [2020] and data generated in this study; Table S2) to remove non-coding RNAs and mitochondrially-encoded genes. For the remaining genes, we calculated log2(TPM+1) values for input into the prcomp module in R.

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). 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 10kb region upstream of the gene start site and split it into ten 1kb 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; Table S6). For a given window, we took the sum of D_u and, separately, the sum of P_u across the 10 focal ribosomal large-subunit genes

(see above), and we calculated the ratio $R_{ribo} = D_u/P_u$. 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 $R_{resample} \geq R_{ribo}$. Resampling statistics are reported in Table S7.

Phylogenetic analysis

To infer phylogenetic branch lengths, we used the following reference genomes from wild-derived inbred mouse strains: *M. m. musculus*: PWK/PhJ, *M. m. castaneus:* CAST/EiJ, *M. m. domesticus:* WSB/EiJ, *M. spretus*: SPRET/EiJ. VCF files for each strain were used to map SNPs onto the C57BL/6J reference genome (GRCm38) using the 'consensus' command from the software bcftools (Danecek et al. 2021). 10kb-long upstream sequences for each gene were extracted utilizing the pybedtools Python package (Quinlan and Hall 2010; Dale et al. 2011). For each gene, sequences were aligned and split into ten 1kb windows.

To account for differences in phylogenetic history of *M. musculus* subspecies across the genome, we used posterior probabilities per locus (White et al. 2009) to assign genes into one of three phylogenetic topologies: *musculus*/*castaneus* (MC)*, castaneus*/*domesticus* (CD)*,* or *musculus*/*domesticus* (MD) as sister subspecies (Figure S1). Genes within a

given locus were assigned to topology with the highest probability; genes within loci with probabilities <50% were eliminated from further analysis. The most likely topology for each gene is reported in Table S8. Of the 10 focal ribosomal large-subunit genes (see above), 6 were assigned to the MC phylogenetic topology, 2 to the CD topology and 2 to the MD topology.

For each 1kb window of each gene, we used the baseml package of the PAML suite (Yang 1997) to infer maximum likelihood lengths of each branch of the phylogeny, and for each branch in turn we calculated L_{ribo}, the average length across the 10 focal ribosomal large-subunit genes (see above). To evaluate significance by resampling, for a given window we randomly sampled 10 genes from the total set with sequence data, matching the topologies of the 10 ribosomal genes (*i.e.,* 6 genes from the MC group, 2 from the CD group and 2 from the MD group), and we calculated the mean length for the focal window and branch, $L_{resample}$. Over 10,000 resamples, we used as a one-sided *P* value the proportion of resamples in which $L_{resample} \geq L_{ribo}$ (Table S8).

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 purebred 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). After allele-specific read-mapping and differential expression testing (Table S4), we earmarked all genes of the genome exhibiting significant expression divergence between the species' alleles (1238 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*. 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. Of the complete survey results (Table S5), one pathway showed robustly significant signal: a cohort of genes encoding ribosomal large subunit components and assembly factors (Figure 1). 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.

At most loci of the ribosomal gene group, the *M. m. castaneus* allele drove higher expression than the 129 allele in the hybrid, in our focal expression profiling stem cell data set (Figure 1) and in an additional study in stem cells of the same background (Figure S2A; Rivera-Mulia et al. 2018). The pattern was largely recapitulated in stem cells from crosses between CAST and the admixed C57BL/6J laboratory strain (hereafter BL6; Werner et al. 2017) (Figure S2B-C), indicating that the activating program encoded by *M. m. castaneus cis*-regulatory alleles at the ribosomal genes was unique relative to 129 and BL6, in the stem cell context. On this basis we considered the ribosomal gene cohort to be a compelling candidate for molecular-evolution follow-up, to investigate selective pressures on directional *cis*-regulatory change across *Mus*.

Unique and conserved alleles in *M. m. castaneus* in ribosomal large-subunit gene upstream regions

We turned to sequence-based tests of selection as an independent means to explore the evolutionary history of ribosomal large-subunit 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. We first 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, normalized by intra-species polymorphism, in discrete spatial windows of upstream regions for every gene in the genome. A resampling test

revealed enrichment of high values of this normalized quantity in ribosomal large-subunit genes relative to a genomic null, in comparisons of *M. m. castaneus* with all sampled populations of *M. m. musculus* (Figure 2A) and all but one of *M. m. domesticus* (Figure 2B). These data make clear that upstream regions of the genes of the ribosomal group are enriched for alleles unique to and conserved in *M. m. castaneus*, as expected under a model of positive selection in the history of this subspecies at the loci (Kreitman 2000).

To investigate further the evidence for evolutionary change between species on upstream regions of our focal genes, we next took a phylogenetic approach. We inferred branch lengths of the *Mus* phylogeny from upstream-region sequences of each gene in the genome (Figure 3A), using a representative genome from each *Mus musculus* subspecies and *M. spretus* as an outgroup. Then, for each lineage in turn, we tested for enrichment of long branches among the ribosomal large-subunit genes relative to a genomic null. The results revealed significant signal in the *M. m. castaneus* lineage (Figure 3B), echoing our findings from population-genomic analyses (Figure 2). Separately, phylogenetic tests also revealed evidence for accelerated evolution in the ribosomal-gene group along the *M. spretus* lineage (Figure 3B), attesting to independent evolutionary changes in this species. These data, in conjunction with population-genomic tests (Figure 2), highlight patterns of derived, conserved alleles in *M. m. castaneus* in the upstream regions of ribosomal genes.

A unique high-expression program of ribosomal large-subunit genes in *M. m. castaneus* relative to *M. m. musculus*

Selection on gene regulation in the wild involves the evolutionary tuning of expression to optimize phenotype and fitness. As these derive from the total expression output of a relevant genotype in a tissue, they integrate the effects of genetic factors acting both in *cis* and in *trans* (Signor and Nuzhdin 2018). Having initiated our study with a dissection of *cis*-acting regulation in hybrid stem cells (Figure 1 and Figure S2), we now sought a holistic view of expression divergence between mouse subspecies in the context of purebreds. We first harnessed a resource of transcriptional profiles from stem cells of a panel of purebred mouse genotypes (Skelly et al. 2020). In the ribosomal gene group, we identified a coherent program of high expression in stem cells from purebred CAST relative to those of purebred BL6 (Figure 4A, red *versus* purple). This mirrored the inducing effects of *cis*-regulatory alleles from CAST at the ribosomal genes that we had identified in transcriptomes in hybrids with BL6 (Figure S2B-C). The ribosomal gene expression program in purebred BL6 stem cells largely paralleled that of the PWD wild-derived *M. m. musculus* background, as did that of the NOD lab strain (Figure 4A); we refer to this set of strains as *M. m. musculus*-like.

To pursue further validation of the expression divergence of ribosomal genes in purebreds, we chose purebred BL6 as a representative of *M. m. musculus*-like strains, for comparison against CAST. Independent establishment of a line

of induced pluripotent stem cells from CAST, followed by transcriptional profiling of these and an analogous line from BL6, confirmed high expression of the ribosomal gene group in the former (Figure 4B). We differentiated the stem cells of both genotypes into motor neurons and transcriptionally profiled these samples (Figure S3), finding again that CAST genotype drove high expression of the ribosomal gene cohort (Figure 4C). Thus, at the ribosomal genes, the overall expression output of the CAST genotype is robustly distinct from that of *M. m. musculus*-like lines in stem cells and their descendants. Such a pattern lends itself well to a classic interpretation of expression divergence with phenotypic and evolutionary relevance, with mRNA levels conserved across one population and distinct from another (Blekhman et al. 2008).

Interestingly, however, we noted no appreciable expression divergence between purebred CAST stem cells and those of the 129 laboratory strain, across the ribosomal gene cohort (Figure 4A, red *versus* green). This posed a contrast with the *cis*-regulatory variation we had detected between 129 and CAST alleles in hybrid stem cells (Figure 1), likely as a product of *trans*-acting variants between the strains counterbalancing those in *cis*. We refer to this high expression program for the ribosomal genes as *M. m. domesticus*-like, as we also detected it in stem cells from purebreds of the WSB *M. m. domesticus* strain, as well as the admixed lab strains AJ and NZO (Figure 4A). We conclude that any interpretation of phenotypic or evolutionary interest in this

system would not invoke expression differences between *M. m. domesticus*-like strains and CAST. Rather, the complete body of our analyses of ribosomal gene regulation in stem cells establishes a high-expression program driven by CAST alleles that is most salient in *cis*-regulatory profiles and in purebred comparisons against *M. m. musculus*-like strains.

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, for a given case study, 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 ribosomal genes that also exhibits directional, polygenic *cis*-regulatory variation in stem cells. Against a backdrop of other case studies of *cis*-acting expression change (Bullard et al. 2010; Fraser et al. 2010; Fraser et al. 2011; Martin et al. 2012; York et al. 2018; Gokhman et al. 2021; Mack et al. 2023), our results represent a proof of concept for the validation of this approach in mammals.

The particular patterns of evolutionary change that we have traced in ribosomal large-subunit genes highlighted unique trends by *M. m. castaneus* alleles. In expression profiles of stem cells, this species harbored *cis*-regulatory alleles upregulating the ribosomal cohort in hybrid backgrounds and also stood

out by virtue of high ribosomal gene expression in the purebred context, in comparisons against *M. m. musculus* though not *M. m. domesticus*. Likewise, *M. m. castaneus* sequence exhibited signatures of directional selection in the upstream regions of ribosomal genes, with respect to both other subspecies. These patterns are consistent with a complex history of parallel evolution across lineages, as has been inferred in other contexts within and between mouse subspecies (Phifer-Rixey and Nachman 2015; Ferris et al. 2021). We propose that the program of low ribosomal gene expression in stem cells represents the ancestral state for the *M. musculus* clade. Under this model, derived variants that act to induce the genes would have arisen and been maintained in *M. m. castaneus* under selection; separately, *trans*-acting factors of similar effect would have evolved along the *M. m. domesticus* lineage.

Our discovery of divergence at ribosomal genes in mice echoes reports of *cis*-regulatory change at homologous loci between yeasts (Tanay et al. 2005; Hogues et al. 2008; Li and Fay 2017; Sorrells et al. 2018) and between animal ancestors distributed over deep time (Brown et al. 2008). This literature leaves open the question of what ecological forces might drive *cis*-regulatory change at ribosomal loci, 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 *cis*-regulatory variation at ribosomal-protein genes may often reflect species-unique logic of cell growth decisions.

Additionally, in 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 ribosome-protein 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 ribosome large-subunit 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 ribosomal 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; Omura 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 D.M.B., J.R.D., and R.B.B.

Author Contributions

NMS, RBB, DMB, and JRD conceived of the study and provided direction. NMS processed the sequencing data used 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

Novel 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

Anders S, Pyl PT, Huber W (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

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.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y (2008). Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet 4: e1000271.

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 ocaloacetate 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.

Edgar R, Domrachev M, Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.

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

Ferris KG, Chavez AS, Suzuki TA, Beckman EJ, Phifer-Rixey M, Bi K, et al. (2021). The genomics of rapid climatic adaptation and parallel evolution in North American house mice. PLOS Genetics 17: e1009495.

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.

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.

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.

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 D, Salzberg SL (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biology 12: R72.

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.

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

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.

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.

Phifer-Rixey M, Nachman MW (2015). Insights into mammalian biology from the wild house mouse Mus musculus. eLife 4: e05959.

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.

Rivera-Mulia JC, Dimond A, Vera D, Trevilla-Garcia C, Sasaki T, Zimmerman J, et al. (2018). Allele-specific control of replication timing and genome organization during development. Genome Res 28: 800–811.

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.

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.

White MA, Ané C, Dewey CN, Larget BR, Payseur BA (2009). Fine-Scale Phylogenetic Discordance across the House Mouse Genome. PLOS Genetics 5: e1000729.

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

Wu C-Y, Whye D, Mason RW, Wang W (2012). Efficient Differentiation of Mouse Embryonic Stem Cells into Motor Neurons. JoVE: 3813.

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 Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13: 555–556.

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.

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.

Figure captions

Figure 1. Directional *cis*-regulatory evolution of ribosomal large subunit assembly genes in cultured stem cells. Each pair of bars reports allele-specific expression of the indicated gene from the Gene Ontology term GO:0000027, ribosomal large subunit assembly, in embryonic stem cells of an F1 hybrid from a cross of a *M. m. castaneus* (CAST) male with a female of the *M. m. domesticus*-derived strain 129/SvImJ (129) background (Marks et al. 2015). Bar heights report means and points report individual replicates. TPM, transcripts per million. Only genes from the term showing significant differential allele-specific expression (adjusted $p < 0.05$) are shown. A directional *cis*-regulatory trend, in which the CAST allele of a given gene is expressed more highly than the 129 allele, is robustly significant for this term relative to a genomic null. Raw data are reported in Tables S1, S4, and S5.

Figure 2. Elevated nucleotide divergence between *M. m. castaneus* and other subspecies in regions upstream of ribosomal large subunit assembly genes. In a given panel, each bar reports a metagene analysis of nucleotide divergence between *M. m. castaneus* and the indicated population of *M. m. musculus* **(A)** or *M. m. domesticus* **(B)**, in a 1000-bp window centered 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:0000027, ribosomal large subunit assembly, normalized against the analogous quantity for all other genes in the genome. *: Benjamini-Hochberg adjusted $p > 0.05$ from a one-sided resampling test for elevated D_u/P_u in the ribosomal gene cohort against a genomic null. KAZ, Kazakhstan; AFG, Afghanistan; CZE, Czech Republic; GER, Germany; IRA, Iran; FRA, France. Raw data are reported in Tables S6 and S7.

Figure 3. Rapid evolution in the *M. m. castaneus* lineage of regions upstream of ribosomal large subunit assembly genes. (A) One of three phylogenies describing the relatedness of *Mus* species and subspecies at distinct genomic regions (see Figure S3 and White et al. (2009)). **(B)** Each cell reports results from a metagene analysis of inferred lengths of the indicated branch on the *Mus* phylogeny, in a 1000-bp window centered at the indicated distance upstream of transcription start sites. The *y*-axis reports the significance from a one-sided resampling test for elevated branch length in genes from the Gene Ontology term GO:0000027, ribosomal large subunit assembly, against a genomic null. Raw data are reported in Table S8.

Figure 4. Ribosomal large subunit assembly gene expression programs in stem cells and neurons of mouse subspecies purebreds. (A) Each set of bars reports expression in embryonic stem cells from (Skelly et al. 2020) of the indicated gene from the Gene Ontology term GO:0000027, ribosomal large subunit assembly, in purebred admixed lab strains (C57BL/6J [BL6], A/J [AJ],

129S1/SvImJ [129], NZO/HILtJ [NZO], NOD/ShiLtJ [NOD]) and purebred strains derived from wild-caught *M. m. domesticus* (WSB/EiJ [WSB]), *castaneus* (CAST/EiJ [CAST]), and *M. m.musculus* (PWD/PhJ [PWD]). **(B-C)** Symbols are as in (A), except that measurements were from independent transcriptional profiles of embryonic stem cells **(B)** or stem cell-derived motor neurons **(C)** of the purebred BL6 and CAST genotypes; raw data are reported in Table S2. Bar heights report means and points report individual replicates. TPM, transcripts per million.

Supplementary figure captions

Figure S1. Topologies of the *Mus musculus* subspecies phylogeny. Each panel shows one possible phylogeny of the *M. musculus* subspecies; each represents the most-likely descriptor of the evolutionary history of a subset of genomic regions, for use in phylogenetic inference per White et al. (2009).

Figure S2. Evidence for directional *cis*-regulatory evolution of ribosomal large subunit assembly genes in hybrid stem cells from replicated experiments and cross directions. Data are as in Figure 1 of the main text except that measurements are from **(A)** F1 hybrid CAST/EiJ female x 129/SvImJ male embryonic stem cells from Rivera-Mulia et al. (2018); **(B)** CAST/EiJ male x C57BL/6J female (Werner et al. 2017); or **(C)** CAST/EiJ female x C57BL/6J male (Werner et al. 2017). Raw data are reported in Table S1.

Figure S3. Purebred stem cell samples cluster separately from purebred neurons based on gene expression. Shown are results from principal component analysis of transcriptional profiles of purebred mouse stem cells and neurons. Each point reports results from one replicate profile of one genotype. The *x*- and *y*-axes report values of the first and second principal components respectively, and their labels report the variance across samples explained by each principal component. Stem cell genotypes profiled in (Skelly

et al. (2020)) are as in Figure 4 of the main text; stem cell genotypes from this study were CAST/EiJ and C57BL/6J (see Methods).

Supplementary table captions

Table S1. Allele-specific normalized read counts from RNAseq of stem cells from hybrids involving CAST. Each tab reports expression measurements from one experimental profile of allele-specific expression in hybrid embryonic stem cells: CAST/EiJ male x 129/SvImJ female F1 hybrid from (Marks et al. 2015); 129/SvImJ male x CAST/EiJ female F1 hybrid from (Rivera-Mulia et al. 2018); CAST/EiJ male x C57BL/6J female (Werner et al. 2017); CAST/EiJ female x C57BL/6J male (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 RNAseq of stem cells and neurons from purebred strains. Each tab reports expression measurements from one data source of experimental profiles of cell culture from purebred mouse strains: embryonic stem cells of eight genotypes from (Skelly et al. 2020); embryonic stem cells of BL6 and CAST/EiJ genotypes from this study; and stem-cell-derived motor neurons of BL6 and CAST/EiJ genotypes from this study. In each tab, the first column reports Ensembl gene accession numbers. The remaining columns report the TPM (transcripts per million) values: tab TPM_E-MTAB-7730 contains 3 biological replicates per strain; tab

TPM_purebred_ESC contains 2 biological replicates per sample; tab TPM_purebred_motor_neuron contains 4 BL6 biological replicates and 1 CAST biological replicate.

Table S3. 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 for which no valid ORF was extracted from the GRCm38 genome using the Ensembl build 102 annotations.

Table S4. Allele-specific expression in CAST x 129 hybrid embryonic stem cells. RNAseq data from CAST/EiJ x 129S1/SvImJ F1 hybrid stem cells (Marks et al. 2015) were analyzed using DESeq2 (Love et al. 2014) to test for differential expression between homologous CAST and 129 alleles. Log2FoldChange, log-transformed ratio of 129-to-CAST allele normalized counts. Pvalue, Wald test *p*-value. Padj, Benjamini-Hochberg adjusted P-value.

Table S5. 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. GO_term, Gene Ontology term analyzed. total_number_of_genes, number of analyzable genes in the GO term. sum_sign_statistic, sum across genes of the term of the qualitative sign statistic reporting which parental allele was more highly expressed (+1 for higher expression of the CAST allele, -1 for higher expression of the 129 allele, 0 for no significant expression difference between alleles). 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.

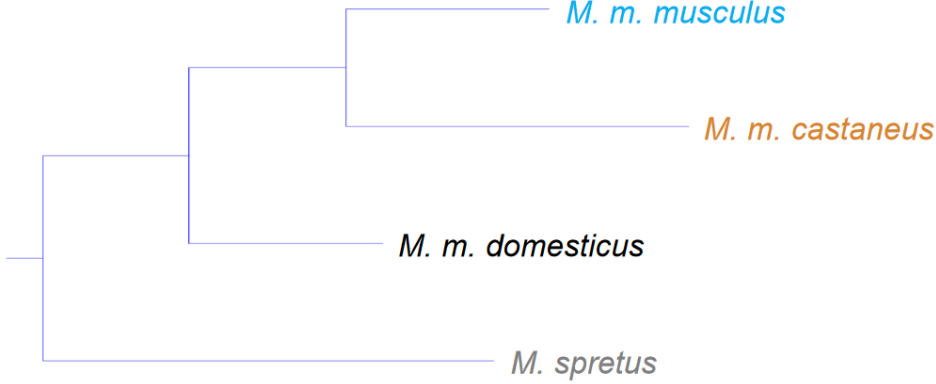
Table S6. Counts of divergent and polymorphic sites in regions upstream of gene start sites, in comparisons between wild mouse populations. Shown are results of analyses of genome sequences from wild mouse populations taken from Harr et al. (2016). Each tab reports results from a comparison between a *M. m. castaneus* population (captured in northwest India) and one population of another subspecies: *M. m. musculus* populations from Afghanistan (Mmm_AFG), the Czech Republic (Mmm_CZE), and Kazakhstan (Mmm_KAZ); *M. m. domesticus* populations from Germany (Mmd_GER), Iran (Mmd_IRA) and France (Mmd_FRA). The first three

columns report Ensembl (build 102) transcript IDs, gene IDs, and gene names. Each of the remaining columns reports the number of divergent sites, D_u, or the number of polymorphic sites, P_u**,** from the 1-kb window at the indicated position upstream of the transcription start site for the indicated gene.

Table S7. Summary statistics of divergence and polymorphism in the upstream regions of ribosomal large subunit assembly genes, in comparisons between wild mouse populations. Shown are results of metagene analyses of the number of divergent sites, D_u, and the number of polymorphic sites, P_u, between *M. musculus* subspecies populations from Table S6, in upstream regions of ribosomal large subunit assembly genes. The first column reports the 1-kb window at the indicated position upstream of the transcription start site. In the following columns, each pair reports the mean D_u/ P_u across the ribosomal 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 relative to the genomic null.

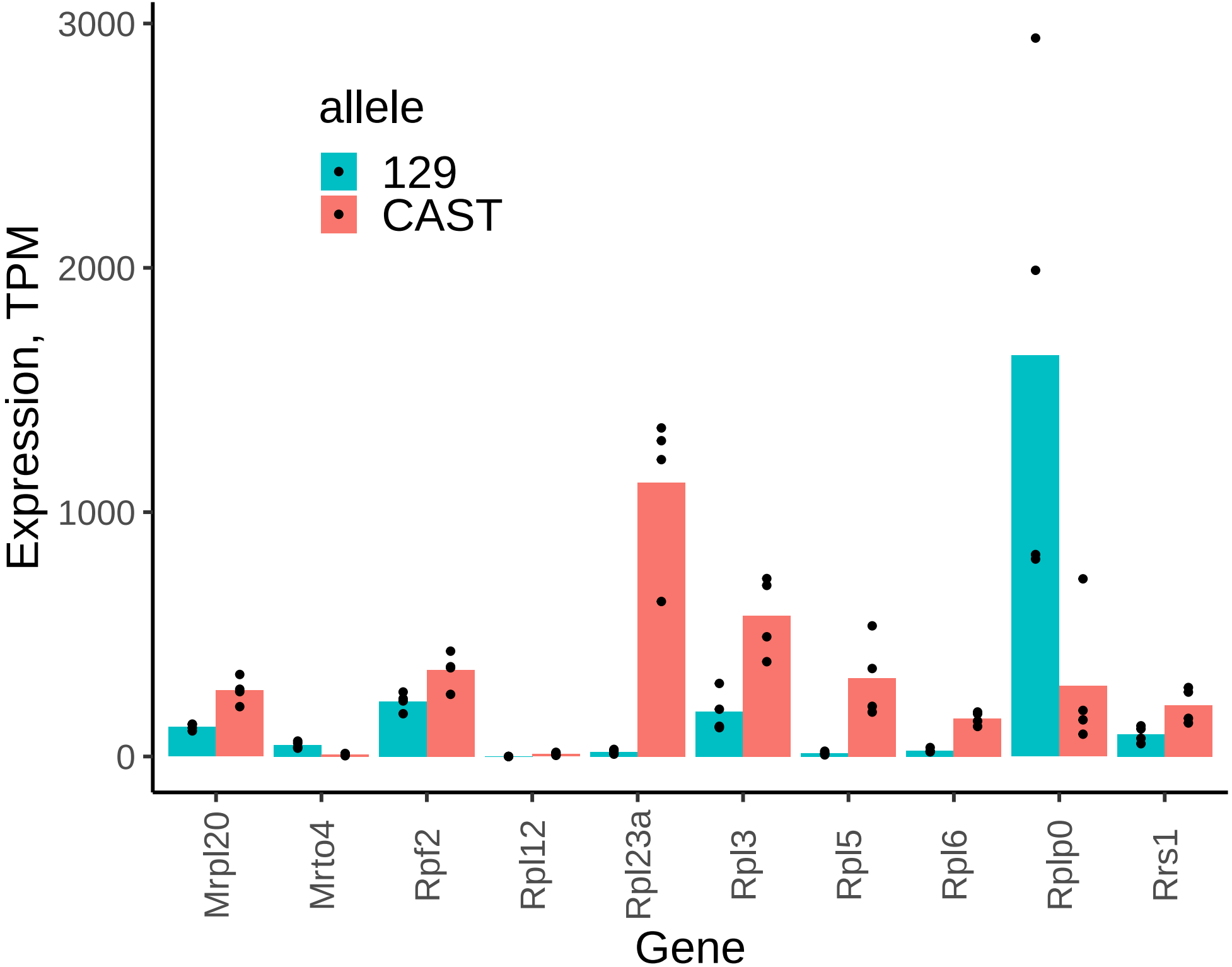
Table S8. Inferred phylogenetic branch lengths for sequences upstream of genes. Each row reports the inferred branch lengths on the *Mus* phylogeny, from the 1-kb window at the indicated position upstream of the transcription start site of the indicated gene, for the lineages leading to *M. m. castaneus* (CAS), *M. m. musculus* (MUS), *M. m. domesticus* (DOM), and *M. spretus* (SPR). Sister_topology, topology from Figure S1 used in phylogenetic branch length inference.

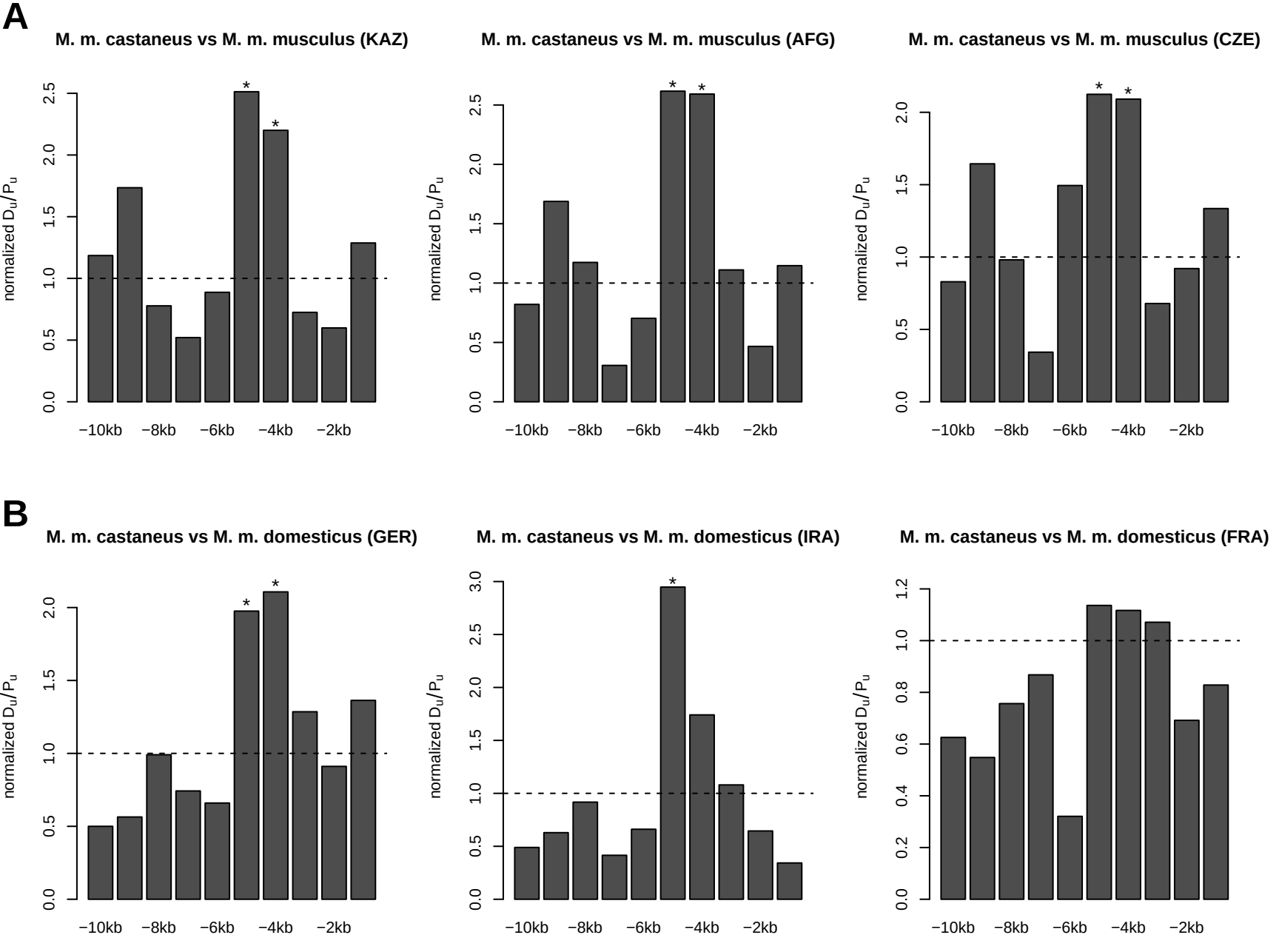
A

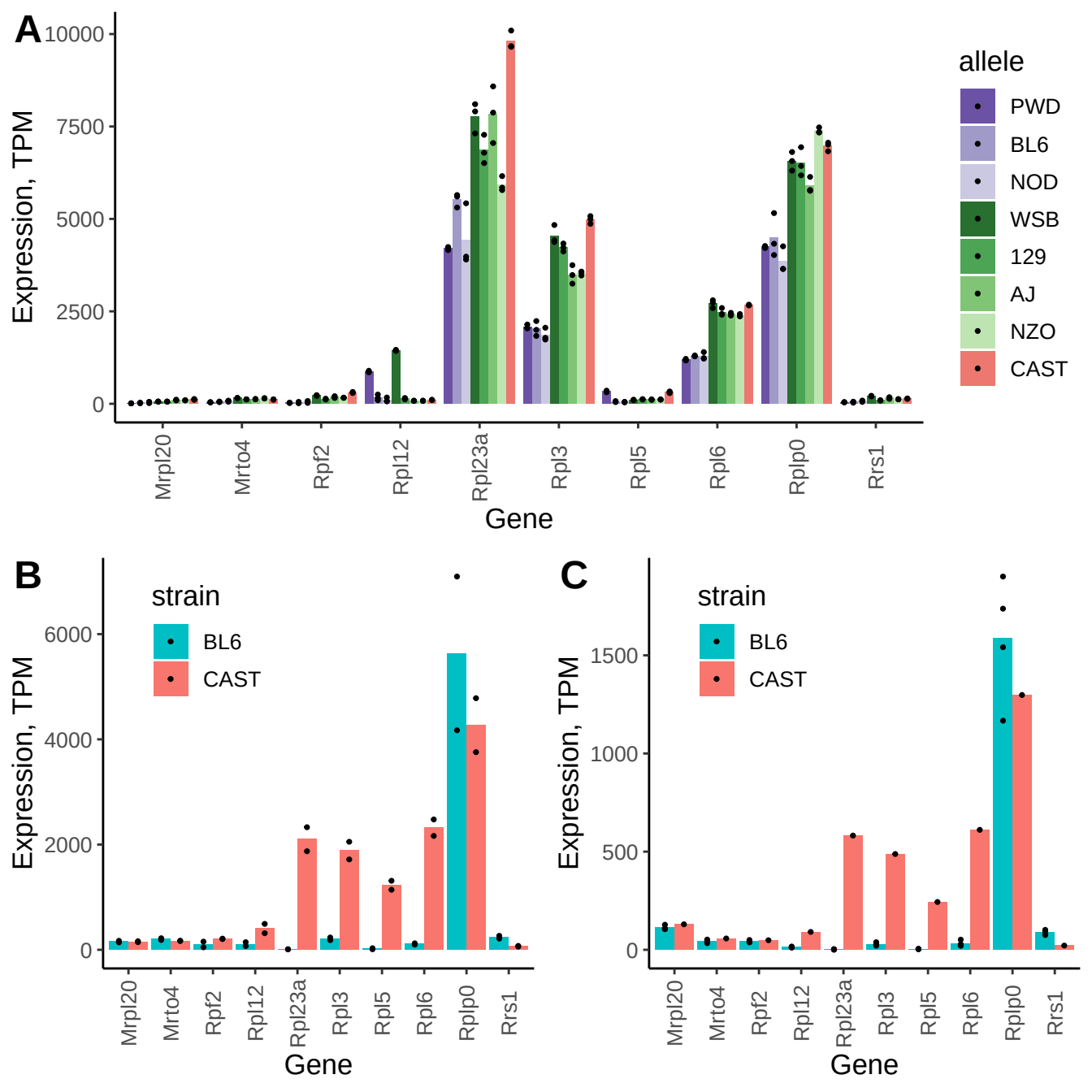


B

	-10kb – -9kb	-9kb – -8kb	-8kb – -7kb	-7kb – -6kb	-6kb – -5kb	-5kb – -4kb	-4kb – -3kb	-3kb – -2kb	-2kb – -1kb	-1kb – TSS
M. m. castaneus *p* value	0.252	0.465	0.252	0.166	0.062	0.315	0.252	0.068	0.119	**0.034**
M. m. musculus *p* value	0.527	0.889	0.803	0.979	0.325	0.849	0.597	0.419	0.399	0.458
M. m. domesticus *p* value	0.246	0.136	0.428	0.678	0.152	0.307	0.583	0.640	0.069	0.333
M. spretus *p* value	0.485	**0.048**	0.212	0.691	0.673	0.685	0.739	0.439	0.163	0.148


allele
129
CAST
Expression, TPM
3000
2000
1000
0
Mrpl20
Mrto4
Rpf2
Rpl12
Rpl23a
Rpl3
Rpl5
Rpl6
Rplp0
Rrs1
Gene


A
M. m. castaneus vs M. m. musculus (KAZ)
M. m. castaneus vs M. m. musculus (AFG)
M. m. castaneus vs M. m. musculus (CZE)
B
M. m. castaneus vs M. m. domesticus (GER)
M. m. castaneus vs M. m. domesticus (IRA)
M. m. castaneus vs M. m. domesticus (FRA)
normalized Du/Pu
−10kb
−8kb
−6kb
−4kb
−2kb


A
Expression, TPM
10000
7500
5000
2500
0
Mrpl20
Mrto4
Rpf2
Rpl12
Rpl23a
Rpl3
Rpl5
Rpl6
Rplp0
Rrs1
Gene
allele
PWD
BL6
NOD
WSB
129
AJ
NZO
CAST
B
strain
BL6
CAST
6000
4000
2000
0
C
strain
BL6
CAST
1500
1000
500
0