Plasticity and evolutionary dynamics of alternative RNA splicing

Wenyu Zhang[1,2,3,*], Anja Guenther[3,4], Yuanxiao Gao[5], Kristian Ullrich[3], Bruno Huettel[6] and Diethard Tautz[3,*]

[1] School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710129, China

[2] Shaanxi Key Laboratory of Qinling Ecological Intelligent Monitoring and Protection, School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710129, China

[3] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Ploen 24306, Germany

[4] Research Group Behavioural Ecology of Individual Differences, Max Planck Institute for Evolutionary Biology, Ploen, 24306, Germany

[5] School of Mathematics & Data Science, Shaanxi University of Science & Technology, Xi'an 710021, China

[6] Max-Planck-Genome-Centre Cologne, MPI for Plant Breeding Research, Cologne 50829, Germany

***Corresponding authors:**

Wenyu Zhang; Phone: +86 029-8843-1769; E-mail: wyzhang@nwpu.edu.cn

Diethard Tautz; Phone: +49 4522-763-390; E-mail: tautz@evolbio.mpg.de

ORCID:

Wenyu Zhang: 0000-0002-2507-3033

Anja Guenther: 0000-0002-2350-0185

Yuanxiao Gao: 0000-0002-6719-8299

Kristian Ullrich: 0000-0003-4308-9626

Bruno Huettel: 0000-0001-7165-1714

Diethard Tautz: 0000-0002-0460-5344

Abstract

Most eukaryotic genes are expressed in multiple RNA isoforms representing variants of the respective genes. Full-length RNA sequencing techniques have uncovered an extreme diversity of RNA isoforms, but a subset of them might be generated by noise in the splicing machinery. For some genes, it has been shown that environmental influences can lead to isoform switching, implying that isoform diversity could also be subject to plastic changes in response to environmental conditions. Further, it has been suggested that isoform diversity could be a basis for adaptive evolutionary novelty. However, explicit tests of all three of these assumptions are missing. To address these questions, we have analyzed here the variation of full-length brain RNA transcripts from natural populations and subspecies of *Mus musculus*, as well as the outgroup species *Mus spretus* and *Mus spicilegus*. We find a substantial influence of splicing noise in generating rare isoform variants. However, after filtering these out, we reliably identify more than 117,000 distinct isoforms in the dataset, about doubling the number of the currently annotated set. Comparisons with individuals raised under different environmental conditions show a very strong plasticity effect in shaping isoform expression, including major isoform switching in proteins that bind to splice site enhancers. Using site frequency spectra tests in comparison to SNP data from the same individuals, we find no evidence for lineage-specific isoforms to become frequently fixed. We conclude that lineage-specific isoforms do not contribute much to novel adaptations, either because they are generated mainly through noise in the splicing machinery or are subject to negative selection. However, isoform diversity is strongly shaped by environmental conditions, both for lineage-specific isoforms, as well as conserved ones. Therefore, the functional role of isoform diversity may mostly be related to trigger plastic responses to environmental changes.

Key words: Alternative isoforms, house mouse, natural population, full-length RNA sequencing, plasticity, splicing noise, site-frequency spectra, natural selection

Introduction

The ability to generate multiple RNA isoforms (or transcripts) from the same gene increases vastly the complexity of eukaryotic transcriptomes and it has been suggested that this may give rise to the evolution of phenotypic diversity and environmental adaptations [1-4]. Isoform switching can also be triggered through environmental temperature changes, for example in genes involved in sex-determination in reptiles [5] or flowering time and stress in plants [6,7]. Also, homoiotherm mammals can regulate splicing of some genes in

response to small changes in body temperature [8,9]. There are also a number of well-studied cases where the emergence of isoforms could be linked to evolutionary changes (see [4] for the most recent review).

The recent development of long-read single-molecule sequencing technologies has enabled the capture of the full-length isoform diversity [10], further facilitating the comparative analysis of alternative isoforms at the global transcript level in different species [11,12]. This has revealed a very high diversity of isoforms but it remains open how much of this is due to noise in the splicing machinery. Further, given that there are well-studied cases where environmental effects, such as temperature, can functionally regulate alternative splicing [9,13], a systematic assessment of the role of environment on isoform diversity is nonetheless missing. Noise and plasticity impact also our understanding of the evolutionary dynamics of recently emerged alternative isoforms, especially at the very early evolutionary stage when they are polymorphic within populations. It has been suggested that most of these isoforms may be neutral [4], but direct tests of this assumption are missing so far. We are addressing these questions here on the basis of samples from the natural diversity of house mouse populations, subspecies and species.

Owing to its well-defined evolutionary history [14,15], the house mouse (*Mus musculus*) has been shown as a particularly suitable model system for studying the evolutionary dynamics of polymorphisms and recently originated genetic elements in natural populations. Currently, three major lineages of the house mouse, which diverged roughly half a million years ago, are distinguished as subspecies [16]: the Western European house mouse *Mus musculus domesticus*, the Eastern European house mouse *Mus musculus musculus*, and the Southeast Asian house mouse *Mus musculus castaneus.* With a divergence time of fewer than 2 million years, closely related outgroup species (*e.g.*, *Mus spretus*) are also available to this model system [16]. The populations are subject to fast adaptations, as evidenced by the detection of high frequencies of selective sweeps and adaptative introgression [17,18]. We have also previously used the house mouse system to study the evolutionary pattern of gene retrocopy variants [19,20], which has shown that new retrocopies of genes are usually subject to negative selection.

Here we use brain as the source tissue for the transcriptome analysis, given that fact that it harbors the largest diversity of cell types with an overall transcript diversity comparable to testis [21], but with many more of these transcripts being likely to be functional in the brain compared to the testis, where there is a lot of expression due to a transcriptionally permissive chromatin environment, especially in late spermatogenic cell types [22]. We used an optimized protocol to capture predominantly full-length transcripts and validated them with existing data. The sequencing depth was chosen to ensure that the diversity of all transcripts and isoforms was captured in each individual when excluding all singletons that are likely to be generated by noise. A comparison with mice raised under different environmental conditions indicates that plasticity can

indeed substantially shape the isoform pattern. Using comparisons with SNP frequency distributions from Illumina RNA-Seq dataset of the same individuals, we show that the distribution of species-specific isoforms is strongly skewed towards being rarer among individuals than neutral SNPs. This indicates that they are either mostly generated by noise effects, or are subject to negative selection. Hence, while the population-level analysis does not support the notion that alternative splicing is a major contributor to the generation of adaptive novelty at the population level, it turns out as a major player in plastic responses to environmental conditions.

Results

An optimized approach to accurately detect full-length transcript transcriptome

We analyzed the alternative isoform landscape in the whole brain transcriptomes for forty-eight unrelated outbred wild-type mice individuals raised under tightly controlled laboratory conditions (Supplementary Dataset S1A and Dataset S2). They included forty house mouse (*Mus musculus*) individuals derived from five natural populations in the three major subspecies (*M. m. domesticus, M. m. musculus*, and *M. m. castaneus*), as well as eight individuals from two closely related outgroup species (*Mus spicilegus* and *Mus spretus*) (Figure 1A). Given that the implementation of 5'cap selection can significantly improve the enrichment of genuine full-length transcripts [23,24], we developed an optimized cDNA enrichment protocol with 5'cap selection for PacBio Iso-Seq library construction (Supplementary Text on Methods). To define high-quality transcriptomes, we performed both PacBio Iso-Seq and Illumina RNA-Seq for each brain sample (Figure 1C). Ten additional animals reared under different environmental conditions were sequenced with the same technique for the analysis of the role of plasticity (see below). Since we found a major influence of environmental conditions on the isoform pattern (see below), we used only the data from the 48 individuals that were raised under controlled laboratory conditions for the main comparative part of the analysis. The data for the ten additional individuals were only used for the plasticity comparisons.

An overview of the methodology to generate the high-quality transcriptome is given in Supplementary Figure S1 (see Methods). In brief, the raw PacBio Iso-Seq subreads were processed to produce circular consensus sequences (CCSs), and further refined to generate full-length non-chimeric (FLNC) reads. The FLNC reads were subject to *de novo* clustering to generate non-redundant isoforms. We implemented optimized parameters to align the unique isoforms of each sample to the GRCm39/mm39 reference genome, and to collapse and merge the transcript models across all 48 samples in the main experiment into a single

non-redundant transcriptome. We further refined a computational pipeline to filter out low-quality transcripts and those of potential artifacts (see Methods).

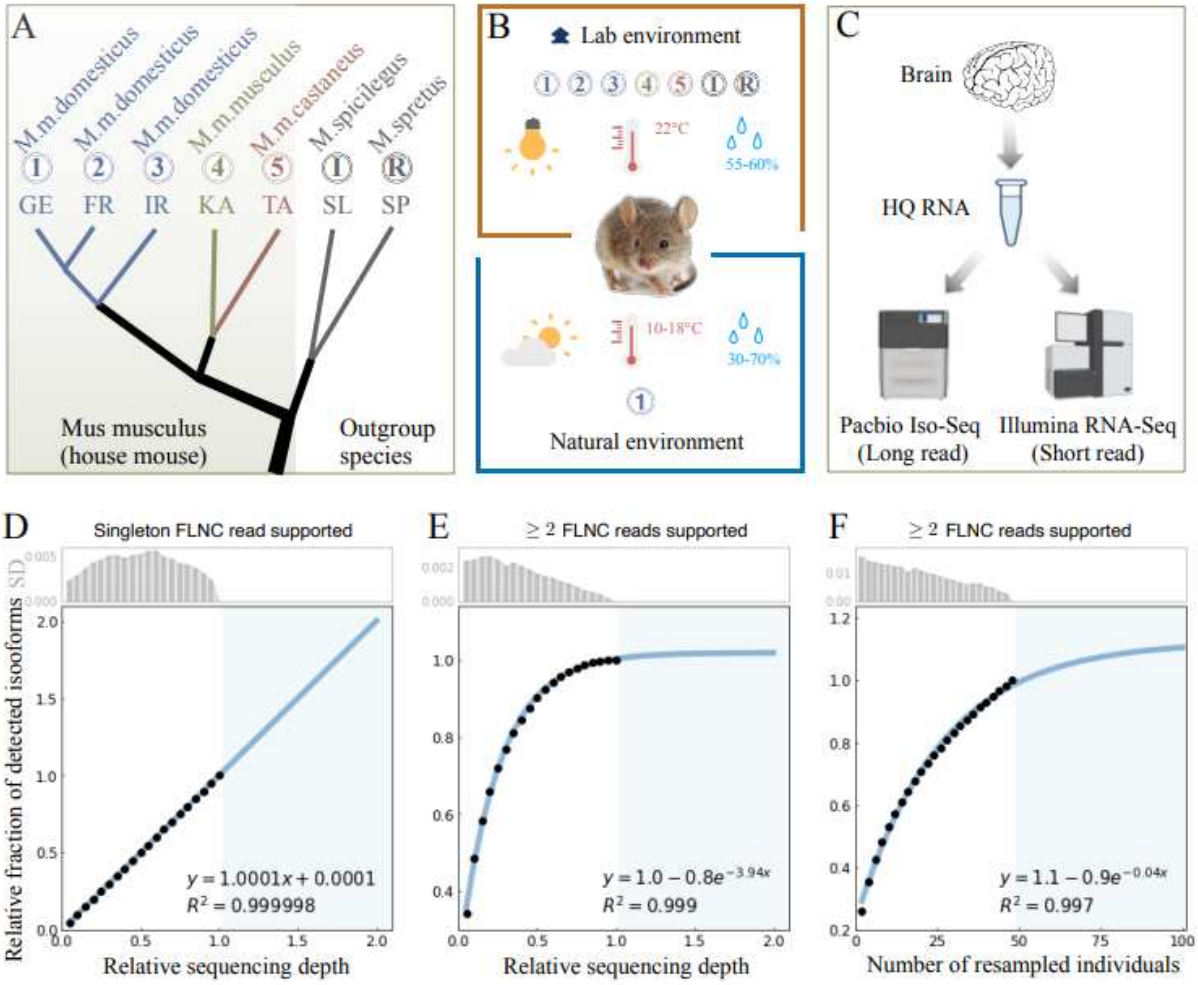

Figure 1 Overview of the study system. (A) Phylogenetic relationships among the house mouse populations and outgroup species (branch lengths not scaled). Abbreviation for population labels: 1, Germany (GE); 2, France (FR); 3, Iran (IR); 4, Kazakhstan (KA); 5, Taiwan (TA), I, Slovakia (SL), and R, Spain (SP). Judged from the evolutionary distances at the overall level of nucleotide difference, the distance of the *Mus musculus* subspecies is at the level of the human-chimpanzee divergence and the distance of the Mus species used here corresponds to the human-gibbon divergence [19,25,26]. (B) Depiction of the environmental conditions of mouse breeding. The 48 sampled mice from all seven populations in the main experiment were raised under tightly controlled laboratory conditions, and additional 10 sampled mice derived from GE population were under semi-natural environment. (C) Sequencing scheme in this study. Both PacBio Iso-Seq and Illumina RNA-Seq data were generated for each mouse brain sample. (D) and (E) show the relative fractions of detected isoforms supported with singleton FLNC read and two or more FLNC reads with increasing random resampling Iso-Seq sequencing depth, respectively. The resampling sequencing depths were selected from 0.05 to 1, with a step size 0.05. The blue area shows the prediction after doubling the current Iso-Seq sequencing depth. The illustration is based on one randomly selected individual (GE3) in the GE population, and the results for individuals from other populations can be found in Supplementary Figure S11 and S12. (F) Relative

fractions of detected isoforms with increasing random resampling sample sizes of individuals in the main experiment. The resampling sequencing sizes were selected from 2 to 48, with a step size of 2. The blue area shows the prediction after doubling the current sampling of mice individuals. The bar plots in the above panels from (D) to (F) show the standard deviation (SD) of each resampling analysis.

Estimating the influence of noise on isoform diversity

Since biochemical systems are never perfect, one should expect a certain number of errors in the splice reactions [27,28]. This could be considered as noise, rather than being regulated through genetic polymorphisms. We used two tests to assess the impact of such splicing errors to isoform diversity.

For the first test, we compared isoforms represented by singleton FLNC reads with those represented by two or more FLNC reads (called "high-confidence" in the following). In our data we find 324,960 singleton FLNC reads supported isoforms (Supplementary Dataset S4) versus 117,728 supported by two or more FLNC reads across all 48 individuals (Supplementary Datasets S5). Random resampling analysis of the sequencing depth at the individual level shows that the number of singletons does not reach saturation (Figure 1D), in contrast to those with two or more reads (Figure 1E). Because of this difference in saturation behavior, we conclude that singletons are mostly the product of splicing errors, which is consistent with the conclusion in previous studies [27,29]. Interestingly, random resampling analysis of individuals' subsamples shows that the number of detectable high-confidence transcripts remained unsaturated with the number of sampled individuals in our dataset (Figure 1F). That is to say, many more new isoforms are expected to show up when more individuals are analyzed, suggesting that they are more likely generated by genetic polymorphisms between the individuals than by noise.

In the second test, we asked whether genes that express one dominant isoform produce on average more additional isoforms when they are higher expressed in a given individual, with a special focus on the high-confidence isoforms. To test this, we selected a subset of genes with more than 10 isoforms where the average expression level for the top expressed transcript (T) is at least five times higher than the cumulative expression level for the other (O) isoforms from the same locus: $(T / \Sigma (O) \geqq 5)$. We find 448 genes that fulfill this condition, with isoform numbers ranging from 11 to 59 (Supplementary Dataset S6). Among them, 48% (214) show a significant positive correlation between isoform number and the top expression level between individuals (one-sided Kendall´s tau test, p-value < 0.05). This could suggest an influence of splicing error noise, but this proportion is actually lower than the corresponding values for the whole dataset. When including all the 3,450 genes with more than 10 isoforms in the correlation analysis, we find 63% (2,165) with significant positive correlation (one-sided Kendall´s tau test, p-value < 0.05). This analysis

suggests therefore that isoform diversity tends to rise with expression level (see also below for an extended analysis of this point). But it does not support that this effect is driven by noise, given that it is lower for genes with the highest expression level contrasts.

Based on these two tests, we suggest that by excluding the singleton reads from our further analysis, we are excluding most of the effects of splicing noise. Evidently, it is still possible that reads that occur more than once are generated through noise effects due to the chromatin context in which they are transcribed or special structures of their RNAs. However, in view of our saturation analysis and the analysis of the genes with highly contrasting expression alleles, we consider the set of high-confidence isoforms to be at least highly enriched in variants that are not simply generated by errors in the splicing machinery.

With the filter of calling only isoforms that were found at least twice in a given individual, we find a total of 117,728 high-confidence distinct transcripts derived from 15,012 distinct loci (Supplementary Dataset S5), which were used for the following analyses. The reliability of these high-confidence isoforms was further validated on the basis of reference annotation and empirical information (Supplementary Figure S6).

Comparison to the reference annotation

Among the 117,728 high-confidence transcripts, 55.4% (65,201) are lineage-specific for the house mouse and 33.6% (39,537) are conserved in both the house mouse and the outgroups (the remainder are specific to the outgroups only, Supplementary Dataset S5). Given that only the conserved isoforms are likely to have a functional role (see further analysis below), we restrict the comparison to the reference annotation to the set of conserved isoforms (Figure 2). A comparison to the full set of high-confidence isoforms is presented in Supplementary Analysis Results SAR1.

To further characterize the features of the conserved transcripts, we compared them with those annotated in the GRCm39/mm39 reference genome from Ensembl v103 [30], which was built largely based on a single C57BL/6 lab mouse inbred strain. On the basis of their alignment status to the Ensembl mouse transcriptome annotation, these transcripts were classified into eight distinct structural categories using SQANTI3 [31] (Figure 2A): i) full splice match (FSM); ii) incomplete splice match (ISM); iii) novel in category (NIC); iv) novel not in category (NNC); v) Fusion (F); vi) genic (G); vii) antisense (A); viii) intergenic(I). In total, we found 69.4% of the 39,537 distinct isoforms matching perfectly to a whole (FSM, 54.9%) or subsection (ISM, 14.5%) of a reference annotated transcript, designated as known transcripts following the convention in [11,32]. The remaining 30.6% of the identified isoforms are novel transcripts, currently not annotated in the Ensembl transcriptome (Figure 2B).

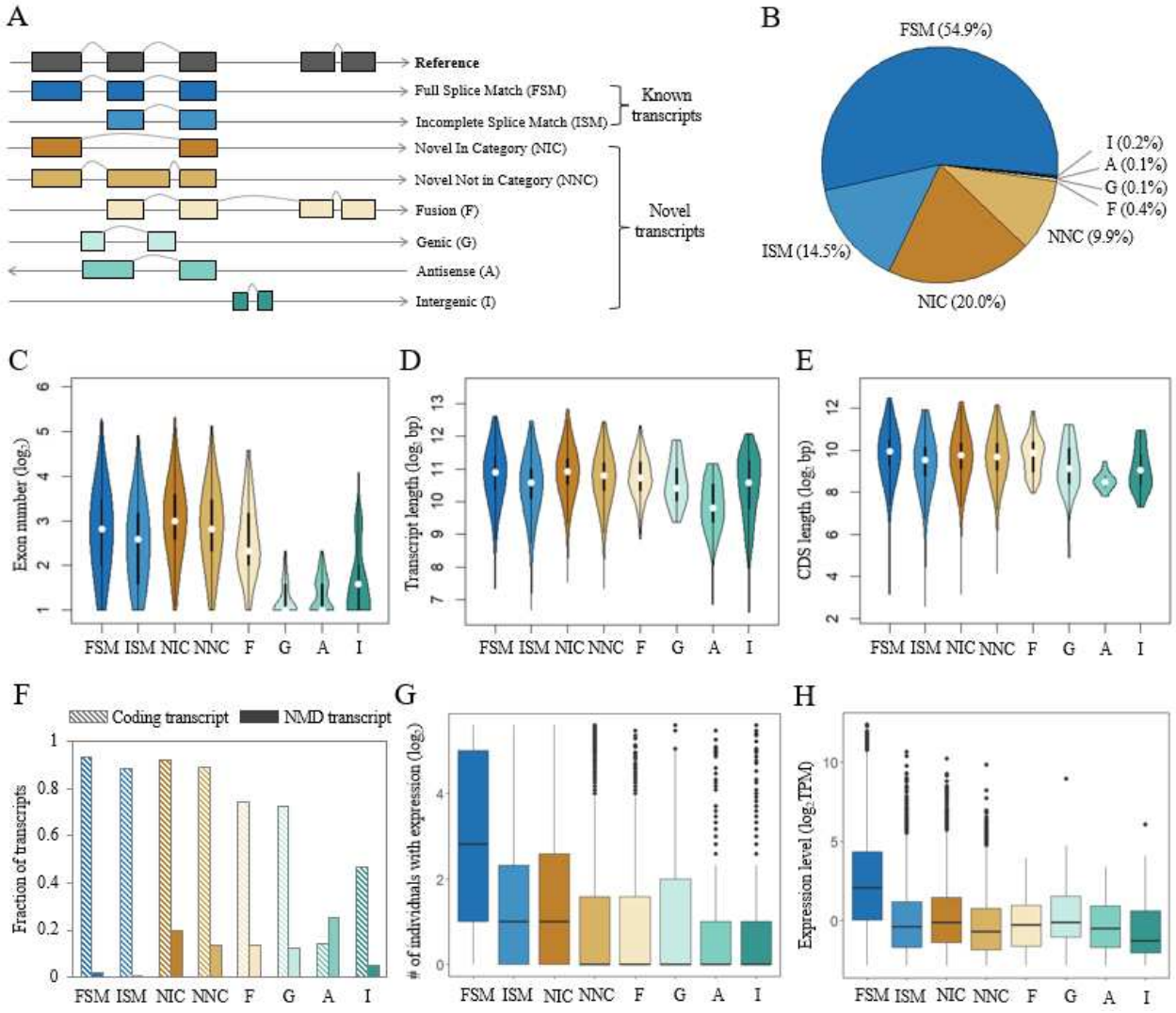

Figure 2 Characterization of the detected conserved isoforms. The conserved isoforms are defined as those detected in both the house mouse and outgroups. (A) Types and illustrations of identified isoforms. (B) Fraction distribution of isoform structural categories. (C)-(H) show the distributions of isoform types with respect to distinct features. Transcripts with expression were defined as the ones with non-zero TPM values, and the expression levels were computed on the basis of the number of supported FLNC reads using SQANTI3 [31]. Boxes represent the interquartile range (IQR, distance between the first and third quartiles), with white dots (or black lines) in the middle to denote the median. The boundaries of the whiskers (also the ranges of violins for panels C-E) are based on the 1.5 IQR values for both sides; black dots in G and H represent outliers.

In comparison to known transcripts (FSM, ISM - see Figure 2A for acronyms), novel transcripts deriving from annotated exonic regions (NIC, NNC, F - which constitute the bulk of the new transcripts) show comparable exon numbers (Figure 2C), transcript length (Figure 2D), and CDS length (Figure 2E). The novel transcripts from intronic (G) and unannotated gene loci (A, I) show generally lower values for all

these features. Notably, all the novel transcripts with coding potential have a significantly higher probability to become degraded via the nonsense-mediated decay (NMD) process [33] due to premature translation-termination codons (PTCs - detected by SQANTI3) than known transcripts (Figure 2F, mean 0.17 vs. 0.02, Fisher's exact test, p-value < 2.2×10^{-16}). Most novel transcripts were found to be more restrictively expressed in a smaller number of individuals, with the exception of NIC, which are comparable to ISM in this respect (Figure 2G). The general expression levels of the novel transcripts are at a similar level as the known ISM transcripts (Figure 2H).

Local alternative splicing events contribute to isoform diversity

The full-length isoforms appear as a combination of different types of local alternative splicing (AS) events, and thus it is useful to disentangle the relative contribution of each variety of AS event to the overall isoform diversity. For this purpose, we used the SUPPA2 program [34] to identify different types of splicing events in the full set of high-confidence transcripts, including skipped exon (SE), retained intron (RI), alternative 5' splice site (A5), alternative 3' splice site (A3), mutually exclusive exon (MX), alternative first exon (AFE), and alternative last exon (ALE) (Figure 3A).

Among all the 94,581 splicing events detected in the merged transcripts (Figure 3B), AFE events contribute most to the overall isoform diversity (31.7%), followed by SE events (19.8%), RI events (14.1%), A5 (11.6%), A3 (10.1%), ALE (9.4%), and with MX as the least contributor (3.3%). This finding is in line with the previous report on the AFE as the most prevalent splicing event for the overall transcriptome in the inbred laboratory mouse cerebral cortex [11], suggesting the dominant role of using AFE to generate alternative isoforms in mice brain transcriptomes. The "alternative first exon" transcripts originate evidently from new promotors, suggesting that these can easily develop upstream of existing genes. This is in line with the realization that enhancers as regulatory elements can also assume promotor functions [35].

AFEs would not necessarily impact the coding sequences [36]. To address this issue, we performed an additional analysis by collapsing the transcripts with the same coding sequence (*i.e.*, only transcripts with predicted ORFs were considered) into a single unique ORF. We indeed observed a much lower number (less than two-thirds) of unique ORFs than transcripts in each mouse individual (Supplementary Figure S14). Based on the landing position of the start codon in relation to local AS events (Figure 3C), we further enumerated the number of local AS events causing the change of the respective coding sequences. We found that only 46.5% (43,978) of all the local splicing events had an impact on the ORF structures (Figure 3D). Compared to AFE, SE and RI events are more prevalent among all the local AS types to contribute to ORF

diversity. The majority of AFE events (55%) would cause no change in the coding sequences, while acting as a major source to generate isoform diversity due to the emergence of alternative promoters [37]. These data illustrate the distinct roles of local AS events in contributing to isoform and ORF diversity in wild mice brain transcriptomes.

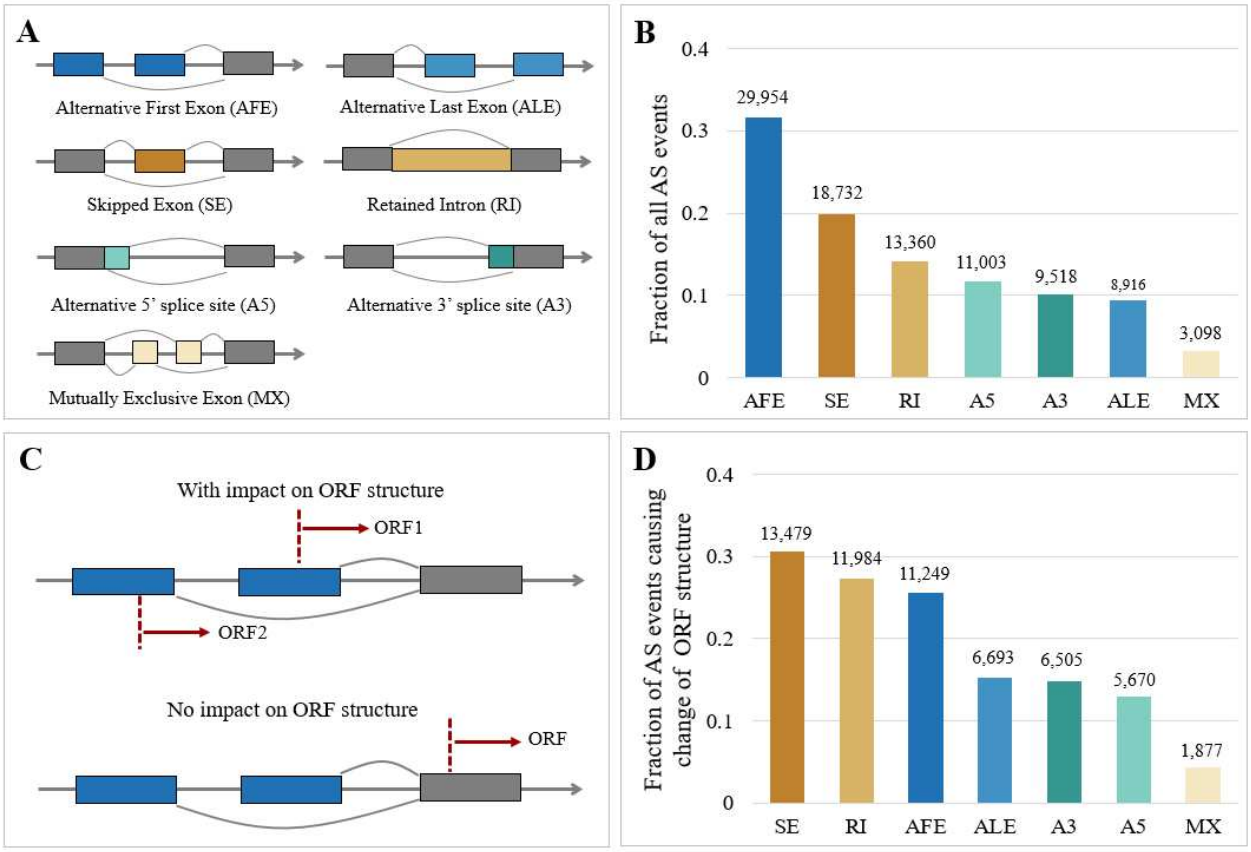

Figure 3 Distribution of different types of local AS events. (A) Types and illustration of AS local events. (B) The distribution of all types of local AS events. (C) An example of AFE events that change ORF structures, and similar situations for other types of local AS events. The dashed red lines indicate the in-frame start codon positions. (D) The distribution of all types of local AS events that impact respective ORF structures. The value above each bar in (B) and (D) indicates the number of respective type of AS events.

Plasticity of isoform diversity

Plasticity is a general term for epigenetic effects on the phenotype of individuals. It becomes evident when individuals are subjected to different environmental conditions, where metabolism including gene regulation is adjusted. In mouse experiments, one is therefore striving to maintain the individuals under constant conditions as much as possible and to sample them at the same time of the diurnal cycle. We have done this

also for the animals used for the main experiment here, hence we would expect only a minor role for plasticity in isoform generation.

To specifically test for the role of plasticity, we have generated data from ten animals that were part of a behavioral study that was done under different environmental conditions (Supplementary Dataset S1B). These mice were derived from the GE population and kept for five generations in semi-natural enclosures [38], *i.e.*, under natural temperature, humidity, and daylight cycles (Figure 1B). The data were merged with the data described above which resulted in an additional 16,171 isoforms (Supplementary Dataset S8), due to the addition of new individuals, as predicted by the rarefaction analysis above (Figure 1F). We have used these data to compare their overall transcriptome differences at the level of presence/absence of isoforms. We find that they deviate substantially from the transcriptomes obtained for the GE animals raised under laboratory-controlled conditions (Figure 4A). While the SNP spectrum of the GE animals has only very slightly changed under the breeding conditions of the semi-natural enclosure (Figure 4B), the transcriptomes diverge strongly with respect to isoform diversity, also in comparison to the other taxa in the study (Figure 4A). The GE animals harbor 25,220 isoforms that are absent in the NE animals, 8,679 (34.4%) of which are also found in animals of the outgroup species. On the other hand, the NE animals harbor 23,714 isoforms that are absent in GE, 3,093 (13%) of which are also found in animals of the outgroups. Hence, while the majority of changes in isoforms between the environmental conditions concerns lineage-specific isoforms there are also substantial numbers that are conserved and can therefore be expected to be functional (see discussion).

Linear modelling of the PC-scores for the most important PC-axes for these data show that the difference is most pronounced in PC1 (Supplementary Figure S15), and a significant distance was observed between NE and GE animals based on isoform landscape, but not on SNPs (Supplementary Table S3). We further used the presence of isoforms fixed within each population to build an overall phylogeny (Figure 4C), which has again a distinct topology as a phylogeny based on the SNP variants (Figure 4D), further confirming the strong impact of plasticity induced by environmental factors on isoform diversity. The overall findings remain valid when focusing only on the isoforms that have not changed their general expression levels (FDR >0.05) between NE and GE individuals (Supplementary Figure S18), and when controlling for the sequencing depth in all the sampled individuals (Supplementary Figure S19). Note that the isoform sharing patterns from the animals kept under constant laboratory conditions allow to generate a phylogeny that conforms to the expected relationships of the populations (Supplementary Analysis Results SAR4). This implies that the slight environmental differences that might still exist between them even under laboratory conditions should indeed have no major overall impact on the comparative analysis.

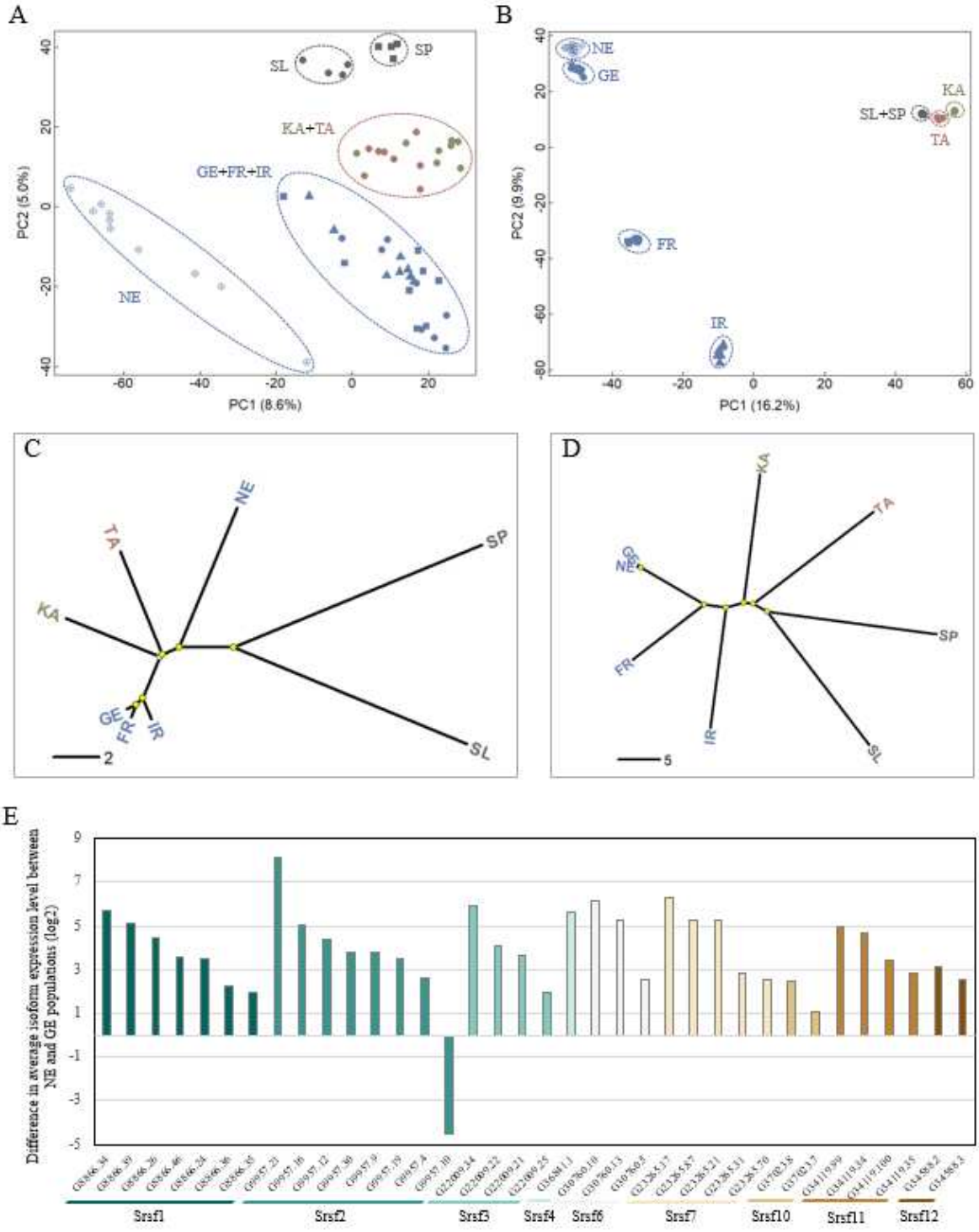

Figure 4 The plasticity influence on isoform diversity. (A) and (B) show the projection of the top two PCs based on isoform and SNP variants, separately. The SNP variants were called from matched Illumina RNA-Seq dataset. Extended results for the populations that cannot be well distinguished in the main figure are presented in Supplementary Figure S16. (C) and (D) show phylogenetic trees built on the basis of isoform and SNP variants fixed

297 within each population, respectively. The fixation is defined as presence in all the individuals with the give population,
298 and the results for a relaxed fixation criteria (present in at least 80% of the individuals) are shown in Supplementary
299 Figure S17. Split nodes marked in yellow are the ones with bootstrap support value >70%. Abbreviations for
300 geographic regions follow Figure 1, and NE indicates the mice individuals derived from GE population but reared
301 under natural environment. (E) Difference in the expression levels for isoforms of Srfs genes. The expression level
302 differences (log2-based) were calculated by subtracting the average expression level in GE individuals under
303 laboratory condition from the average of those under semi-natural environment. Only isoforms with significant
304 expression level differences after multiple testing correction (FDR < 0.05) and conserved in the outgroups are shown.
305 The statistics on the full dataset are shown in Supplementary Dataset S9.

306

307 Environmentally correlated alternative splicing is likely regulated by proteins binding to splicing enhancers,
308 especially the family of SR proteins. These proteins share a domain rich in serine and arginine residues and
309 they are commonly called Srfs proteins [39]. The mouse has 11 members in this protein family and we
310 surveyed all of them for changes of expression between the laboratory animals and the animals living under
311 semi-natural conditions (Supplementary Dataset S9). We found indeed major isoform expression changes
312 for most of these genes, including Srfs1, Srfs2, Srfs3, Srfs4, Srfs5, Srfs7, Srfs10, Srfs11 and Srfs12 (Figure
313 4E). Each of these genes has some isoforms that are more highly expressed under laboratory conditions, and
314 Srfs2 has also one isoform that is more highly expressed in the semi-natural environment. These are
315 candidates for mediating the observed plastic response.

316

317 *Fast turnover of alternative isoforms in house mouse natural populations*

318 To study the turnover rate of alternative splicing at a microevolutionary scale, we focused on the recently
319 emerged isoforms in the house mouse lineage, *i.e.*, detectable in at least one of the five house mouse natural
320 populations surveyed here but absent in outgroup species samples (*Mus spretus* and *Mus spicilegus*). We
321 identified 65,201 house mouse specific isoforms (derived from 13,207 distinct loci) across all the 40
322 surveyed house mouse individuals under laboratory breeding conditions (Supplementary Dataset S10). On
323 average, 4,661 (SD: 808) and 37,291 (SD: 2,741) recently emerged isoforms were found separately in each
324 house mouse individual and each house mouse population.

325 Our population-level analysis allows to apply a frequency-spectrum test to distinguish the effects of
326 selection (positive or negative) and drift on polymorphic characters. The general assumption is that the
327 number of individuals carrying a polymorphic variant depends only on the mutation rate and the fixation

rate in an ideal population. The expected allele frequency spectrum under neutral conditions can be calculated by coalescent or diffusion approaches, but this is affected by the demographic history of the populations, as well as the effects of positive or negative selection on the variants. To avoid simulations with more or less realistic assumptions, we use here a comparison with the SNP variants from the same populations to assess whether isoforms might have a higher or lower rate of fixation. We have previously used the same approach to estimate selection effects on newly arisen retrogenes [19].

We first focused on the SNP variants that were called based on the Illumina RNA-seq dataset from the same set of individuals (see Methods). Among the 871,512 house mouse specific SNP variants, around 22.4% are found in all three house mouse subspecies (Figure 5A), about 20.8% segregate in all five populations (Figure 5B), and 17.5% are found in all 40 analyzed house mouse individuals (Figure 5C). In contrast, of the 65,201 house mouse specific isoforms, only 8.1% are found in all the three subspecies, 3.6% in all the five populations, and 0.07% in all the 40 tested house mouse individuals. Hence the fixation rate of newly emerged isoforms is more than 200 times lower than that for SNPs in the same populations.

For neutrally segregating variants, one can expect that a fraction becomes fixed in a population-specific manner. This is indeed observed for the SNP variants. We find accentuated frequency peaks at the intervals of individuals' numbers of 8 (each of the five populations), 24 (subspecies-level: *M. m. domesticus*), and 40 (species-level: *Mus musculus*) (Figure 5C). In contrast, such a pattern is not observed for the isoforms, implying that most are not neutral. Instead, these recently derived transcripts show a more skewed pattern (Figure 5C), with a surprisingly high fraction of them being individual private (Fraction 0.55 vs. 0.11; Fisher's exact test, p-value $< 2.2 \times 10^{-16}$). Given that at our sequencing level, the number of detected isoforms had reached saturation for each individual (Figure 1E), it is unlikely that the absences in other individuals are due to failure of detection. Note that this pattern still holds when variants only found in more than two animals were analyzed (Supplementary Figure S20).

To avoid possible bias introduced by SNP variants from highly transcribed regions, as they were called from the Illumina RNA-seq dataset [40], we retrieved another SNP dataset called from genomic sequencing data of the same populations (equal number of individuals for each population, but different individuals) [19]. We analyzed this new SNP dataset in the same manner aforementioned, and found that the overall distribution pattern of house mouse specific transcripts is robust to the choices of SNP datasets (Supplementary Figure S21). A similar pattern is observed when focusing only on the ORF level via collapsing the isoforms forming the identical ORFs (Supplementary Figure S21).

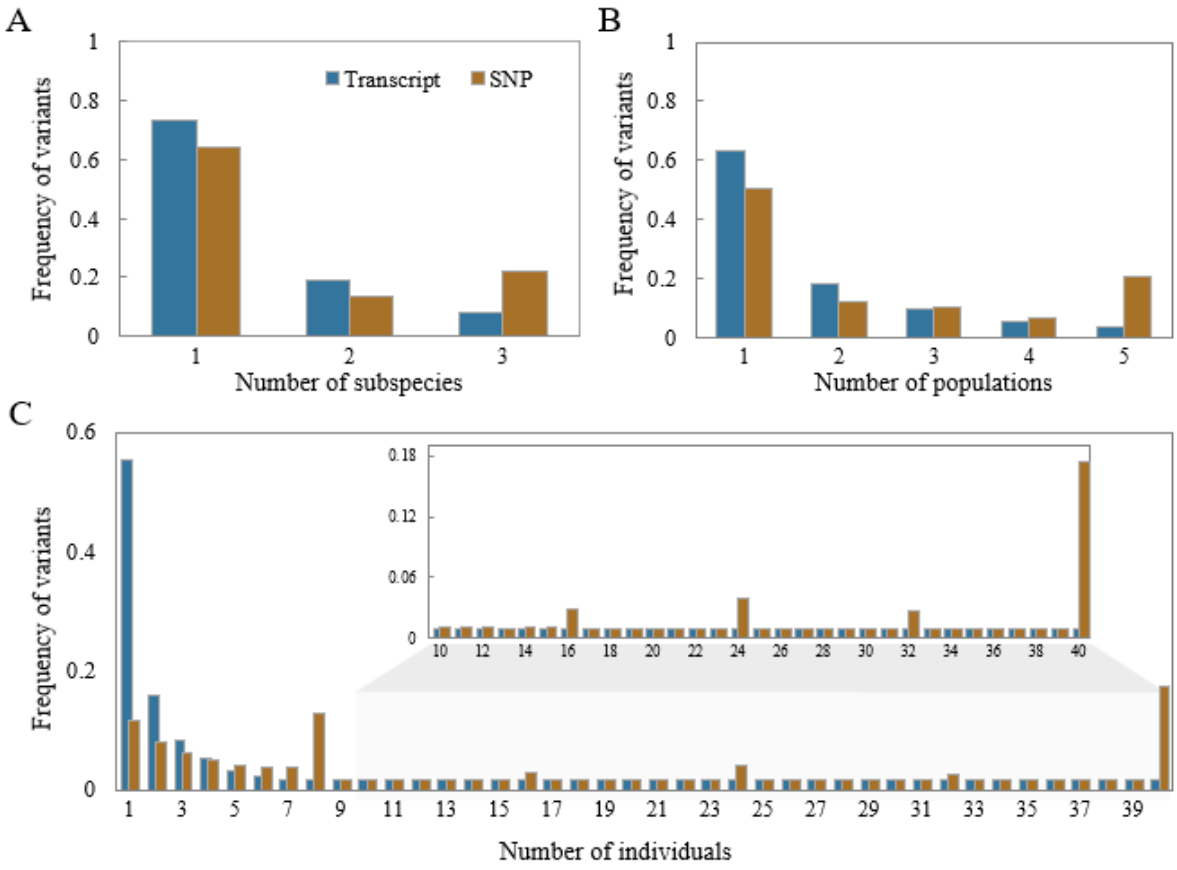

Figure 5 Frequency distribution of house mouse specific transcripts and SNPs. Distribution of the frequency of transcripts and RNA-Seq data-based SNPs across different house mouse (A) subspecies, (B) populations, and (C) individuals. Inset in (C) represents an enlargement with a focus on the frequencies of transcripts and SNPs present in larger numbers of individuals.

In a further analysis, we compared the transcript site frequency spectra with the corresponding frequency spectra of different SNP categories (Supplementary Figure S22). On the basis of the functional effects predicted using the Ensembl Variant Effect Predictor [41], the SNP variants were classified into four distinct groups: i) high-effect SNPs changing the coding gene structure (stop codons or splice sites), ii) moderate-effect SNPs with nonsynonymous changes, iii) low-effect SNPs with synonymous changes, and iv) modifier-effect SNPs occurring in noncoding regions. We used two-sided Kolmogorov–Smirnov tests to compare the overall similarity of the distributions, and found the most similar distribution between transcripts and the most constrained high-effect SNP category (Kolmogorov's D statistic for transcripts vs. high-effect SNPs: D = 0.27; transcripts vs. moderate-effect SNPs: D = 0.42; transcripts vs. low-effect SNPs: D = 0.56; transcripts vs. modifier-effect SNPs: D = 0.54).

374 This overall pattern is very different when one compares the frequency distribution of transcript variants
375 shared with at least one of the outgroup species. These are found on average in 21.4 individuals across the
376 populations, while the house mouse specific ones are found on average in 3.4 individuals (after
377 normalization on the numbers of assayed individuals in two groups, two-sided Wilcoxon rank sum test, p-
378 value < 2.2×10^{-16}, Supplementary Analysis Results SAR3A). This suggests that isoforms shared across
379 larger evolutionary distances are subject to purifying selection and are therefore more likely to be retained
380 in the populations over time.

381 An in-depth analysis of expression levels of house mouse specific isoforms across different categories
382 showed that the isoforms that are found in few individuals only tend to be lowly expressed, with a trend that
383 isoforms with higher frequencies in the populations are expressed at a higher level (Supplementary Analysis
384 Results SAR3B-D).

385

386 **Discussion**

387 The huge diversity of transcriptomes that is created by alternative splicing has long been recognized. But a
388 systematic comparative analysis in a natural population context has only now become possible through full-
389 length RNA sequencing techniques. We used here PacBio Iso-Seq to characterize the full-length isoform
390 diversity of the brain transcriptomes in house mouse natural populations, resulting in the first and most
391 comprehensive full-length isoform category representation at a comparative population level to date. Via
392 implementation of a separate 5'cap selection step [23,24], our optimized approach improved the performance
393 to enrich genuine full-length transcripts. Most importantly, we applied a sequencing depth at which
394 saturation of different isoforms was achieved at the individual level.

395 Our overall results confirm the conclusions from previous studies that the diversity of alternatively spliced
396 transcripts surpasses the current annotation level, even of exceptionally well-curated genomes, such as the
397 one from the house mouse [11]. We have detected double as many transcripts as are currently annotated and
398 we showed that this number keeps increasing when more individuals are sampled from the populations.

399 This unique dataset allowed us to tackle very general questions that arise in the context of the observation
400 of an exuberant isoform diversity that has been found in many studies. How much is caused by noise in the
401 splicing machinery? Can environmental conditions significantly change the isoform diversity? How does
402 the diversity translate into adaptive novelty?

403

The role of splicing noise

The saturation analysis via a rarefaction approach shows a strong contrast between isoforms that are detected only as single reads, versus isoforms that are detected as at least two reads in a given individual. For the former we reach by far no saturation, for the latter we see complete saturation at our sequencing depth. We conclude from this that at least a large fraction of singleton reads reflect aberrant splicing that is not repeatable. Given this clear distinction pattern, it is easy to simply remove the singleton fraction from the further analysis - and we recommend that this should become a standard procedure in comparable studies.

Noise should be particularly evident for highly expressed transcripts and we see also an overall correlation between expression level and isoform diversity. Intriguingly, however, our direct test for the role of expression level on noise patterns within a given locus does not show a strong tendency that the highest expressed alleles have more isoforms in the same individual, as it would be expected for a noise effect. Instead, highly expressed alleles have actually relatively fewer additional isoforms than one would expect at that expression level, indicating a rather strict control of splicing efficiency. Hence, highly expressed loci, which are often also evolutionary old genes, appear to be less sensitive to the influence of noise, probably because their trans-regulation has been optimized [42]. This would also explain why they can maintain on average more splice variants than low expressed, evolutionarily younger genes.

The role of plasticity

While it is well known that environmental conditions can influence splicing patterns (reviewed in [4]), we were surprised to see that the simple shift of a given population from constant laboratory conditions to more natural environmental conditions results already in a major change in isoform expression, including roughly a third that are conserved across the taxa and are therefore likely to be functional (see below).

The factors that had changed between the environments were the ambient living temperature, the natural day-light cycles and natural humidity. One would probably have considered these as relatively minor changes. Interestingly, it has previously been shown that already small body temperature changes associated with diurnal cycles can also cause alternative splicing patterns for thousands of transcripts [8]. It appears that splicing regulating proteins play a major role in this and we find indeed changes in transcript abundances for most of these genes. Among the best studied regulators in this context are Srsf2 and Srsf10 [8,43,44]. We found major expression changes for both Srsf2 and Srfs10 (Figure 4E), with Srfs2 showing the previously described alternative splice patterns associated with temperature changes. However, it is known that the

functions of Srfs2 are also broader, including regulating genomic stability, gene transcription, mRNA stability, and translation [45].

We conclude from our data that there can indeed be a strong influence of environmental conditions on isoform diversity, supporting the notion that environmental conditions need to be fully controlled to allow a detailed comparison of isoform diversity between different taxa.

No signal for adaptive evolution

Another strength of our dataset lies in the possibility of systematic comparisons of fixation probabilities between natural populations, subspecies, and species. Within such a framework, one can make inferences on microevolutionary patterns that were not possible in previous comparative studies on alternative splicing in more or less distantly related species [11,12,46-52]. In particular, our data allow us to directly assess whether the large isoform diversity generated through alternative splicing could be a major mechanism to create adaptive genetic novelty [2-4]. Our data do not support such a model.

Although we found a vast number of newly arisen alternatively spliced transcripts, most of them occur only in one or few individuals. Such a pattern is typical for polymorphic markers that evolve neutrally, or are under negative selection. In the case of isoform diversity, they could also include variants generated by the noise effects which would evidently not contribute to fixation patterns. We cannot fully exclude this possibility, but given that we have restricted our analysis to high-confidence isoforms we consider the noise component as small.

A direct comparison with the frequency distributions of different functional classes of SNPs from the same individuals showed that the isoform distribution is closest to the distribution of highly constrained SNPs. This implies that many novel isoforms are not even neutrally segregating, but are under negative selection. This inference is also supported by the second observation in the comparison with the SNPs. For SNPs we find patterns of random fixation for each population, typical for neutral markers over time, while such patterns are absent for the isoform frequency distributions. Their overall fixation probability is at least a factor of 200 lower than the one for the SNPs, implying that the negative selection is actually relatively strong for most of them. This is also in line with the observation that they are usually only lowly expressed. Only 45 out of 65,201 *Mus musculus* specific isoforms are fixed in all populations analyzed (Supplementary Dataset S10). These could be rare adaptive fixations, but could also represent random fixations of slightly deleterious variants.

Most interestingly, there is a strong contrast in the frequency distribution of isoforms shared with at least one outgroup species. They are shared by many more individuals, and the largest fraction is shared by all of them. This indicates that they serve active functions for the individuals, *i.e.*, are under stabilizing selection. This observation also fits the findings of Leung *et al.* [11] of some shared overall patterns of alternative splicing between the mouse and the human cortex.

We note that more than half of the isoforms annotated in Ensembl are lineage-specific isoforms in our analysis and therefore less likely to be broadly functional. It could be useful to annotate them as a different class for future comparisons with other species.

Ferrandez-Peral *et al.* [12] found a correlation between fast-evolving immune genes and high isoform diversity in their analysis of isoform diversity in lymphoblastoid cell lines from primates. They suggested that this could point to an adaptive role for isoforms, but this inference is too indirect to provide a direct link. Hence, while this may still be true for the particular gene class of immune genes, it does not invalidate our findings.

Wright *et al.* [4] listed several examples of alternative splicing of individual genes that may have had a role in adaptation or speciation events. But such individual cases cannot be used to support a model of frequent adaptive evolution through alternative splicing.

Overall, each individual harbors around 4,600 private isoforms. When most of them have a negative selection coefficient, this could substantially impact the overall genetic load of the individual [29]. In humans, mis-splicing events are causative for many disease phenotypes, including cancer, neurodegenerative diseases, and muscular dystrophies [53-55]. In a recent GWAS for human brain-related complex traits, Qi *et al.* [56] discovered cis QTLs affecting splicing in more than 12,000 genes, with a subset of them related to disease phenotypes. This corroborates our conclusion that most reproduceable splicing variants are genetically controlled and that they can have negative effects on the phenotype. But given our finding that isoform diversity in populations is strongly influenced by environmental conditions, the main effect of the observed high level of alternative splicing may be in conveying plastic responses of populations to changing environments.

Methods

Sample collection and RNA extraction

A total of 58 adult male mice individuals were sampled in this study (Supplementary Dataset S1), and all these mice were derived from previously wild-caught founder mice, maintained in an outbreeding scheme [16]. For the main experiment, eight adult individuals were chosen for each of the following populations covering all three major subspecies of the house mouse (*Mus musculus*): Germany, France, and Iran populations from *Mus musculus domesticus*, Kazakhstan population from *Mus musculus musculus*, and Taiwan population from *Mus musculus castaneus*. We also included four adult individuals from each of the two outgroup species, *Mus spretus* and *Mus spicilegus*. These mice were reared under standard lab conditions, with well-controlled environmental factors: temperature 22°C, humidity (55-60%) and 12h:12h light scheme [16]. To test the plasticity effects of alternative splicing, we chose another ten adult individuals derived from the same *Mus musculus domesticus* (Germany) population that were reared in semi-natural enclosures, *i.e.*, under fluctuating natural temperature (10-18°C), humidity (30-70%), and daylight cycles.

Mice were sacrificed at approximately ten weeks of age by CO_2 asphyxiation followed immediately by cervical dislocation. The whole brain was dissected and immediately frozen in liquid nitrogen within 5 minutes post-mortem. Total RNAs were extracted and purified using RNeasy lipid tissue kits (Qiagen, The Netherlands). RNA was quantified using Qubit Fluorometers (Invitrogen, Thermo Scientific, USA), and RNA quality was assessed with 2100 Bioanalyzer (RNA Nanochip, Agilent Technologies, USA). All samples were with RIN values above 8.5 and then used for both PacBio and Illumina transcriptome sequencing at the Max Planck-Genome-Centre Cologne.

PacBio Iso-Seq and Illumina RNA-Seq library preparation and sequencing

Our initial experimental tests showed that the TeloPrime full length cDNA amplification kit, which selectively synthesizes cDNA molecules from mRNAs carrying a 5' cap, could provide a better solution to enrich for actual full-length transcripts, compared to the standard PacBio cDNA library preparation protocol (Supplementary Text of Methods). Hence, the TeloPrime full-length cDNA amplification kit v2 (Lexogen GmbH) was utilized to construct PacBio IsoSeq cDNA libraries. One µg total RNA from each individual was used as input, and double-strand cDNA was produced by following the manufacturer's instructions, except that an alternative oligo-dT primer from the SMARTer PCR cDNA synthesis kit (Clontech Laboratories, Inc.), which also included a random 10mer sequence as a unique molecule identifier (UMI)

after each sequence. The cDNAs were not size-selected, and PacBio libraries were prepared with the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). To get a similar number of clustered high-confidence isoforms, each library was sequenced on three 1M-ZMW SMRT cells on the PacBio Sequel I for the main experiment, and one 8M-ZMW SMRT cell on the PacBio Sequel II platform for the plasticity effect experiment, respectively (Supplementary Dataset S3A and S3C).

Poly(A) RNA from each sample was enriched from 1 µg total RNA by the NEBNext® Poly(A) mRNA Magnetic Isolation Module (Catalog #: E7490, New England Biolabs Inc.). RNA-Seq libraries were prepared using NEBNext Ultra™ II Directional RNA Library Prep Kit for Illumina (Catalog #: E7760, New England Biolabs Inc.), according to manufacturer's instructions. A total of eleven PCR cycles were applied to enrich library concentration. Sequencing-by-synthesis was done at the HiSeq3000 system in paired-end mode 2 x 150bp. Raw sequencing outputs were converted to fastq files with bcl2fastq v2.17.1.14. An average of 28.2 (SD: 3.3) and 56.7 (SD: 2.7) million raw fastq read pairs were generated for each sample in the main experiment and the plasticity effect experiment, separately (Supplementary Dataset S3B and S3D).

Iso-Seq read QC and data processing

We analyzed the raw sub-reads for each SMRT cell separately following the IsoSeq3 pipeline (v3.4.0; https://github.com/PacificBiosciences/IsoSeq). Circular consensus sequences (CCS) were generated from sub-reads using the CCS module in polish mode (v6.0.0; --minPasses 3 --minLength 50 --maxLength 1000000 --minPredictedAccuracy 0.99) of the IsoSeq3 pipeline, and the CCS reads generated in three SMRT cells for the same sample were merged. We trimmed cDNA primers (5' TGGATTGATATGTAATACGACTCACTATAG; 3' GTACTCTGCGTTGATACCACTGCTT) and orientated the CCS reads using the lima program with the specialized IsoSeq mode (v2.0.0; --isoseq). The 10mer UMI following each CCS read was tagged and removed using the tag module of the IsoSeq3 pipeline (--design T-10U). Following this, we identified the processed CCS reads as full-length and non-chimeric (FLNC), based on the presence of a ploy(A) tail and absence of concatemer using the refine module of the IsoSeq3 pipeline (--require-polya). We further performed PCR deduplication based on the UMI tag information using the dedup module of the IsoSeq3 pipeline (default parameters). After this deduplication step, only one consensus FLNC sequence per founder molecule in the sample was kept. We then performed *de novo* clustering of the above reads using the cluster module of the IsoSeq3 pipeline, and kept only the high-confidence isoforms supported by at least 2 FLNC reads for the main analysis. In addition, the isoforms supporting by singleton FLNC reads were utilized for the part of analysis of the noise influence on isoform diversity (see below).

We aligned these isoforms of each sample to the GRCm39/mm39 reference genome sequence using the minimap2 (v2.24-r1122; -ax splice:hq -uf --secondary=no -C5 -O6,24 -B4) [57], with parameter setting following the best practice of Cupcake pipeline [58]. The alignment bam files were further sorted using the samtools program v1.9 [59]. We collapsed redundant transcript models for each sample based on the above sorted alignment coordinate information using the collapse module of the TAMA program (-d merge_dup -x capped -m 5 -a 1000 -z 30 -sj sj_priority) [60]. The rationales for defining redundant transcript models are shown in Supplementary Text of Methods. Finally, isoforms across all 48 samples in the main experiment were merged into a single non-redundant transcriptome using the merge module of the TAMA program (-d merge_dup -m 5 -a 1000 -z 30) [60], with the same parameter setting as the above collapse step.

RNA-Seq read QC and data processing

We trimmed and filtered the low-quality raw fastq reads for each sample separately using the fastp program (v0.20.0; --cut_front --average_qual 20 --length_required 50) [61], and only included the paired-end reads with a minimum length of 50bp and average quality score of 20 for further analysis.

The filtered fastq reads were aligned to mouse GRCm39/mm39 reference genome sequence with STAR aligner v2.7.0e [62], taking the mouse gene annotation in Ensembl v103 [30] into account at the stage of building the genome index (--runMode genomeGenerate --sjdbOverhang 149). The STAR mapping procedure was performed in two-pass mode, and some of the filtering parameters were tweaked (personal communication with STAR developer; --runMode alignReads –twopassMode Basic --outFilterMismatchNmax 30 --scoreDelOpen -1 --scoreDelBase -1 --scoreInsOpen -1 --scoreInsBase -1 --seedSearchStartLmax 25 --winAnchorMultimapNmax 100), in order to compensate the sequence divergences of individuals from various populations and species [19]. With this optimized mapping pipeline, a similar alignment rate was reached for all the samples (Supplementary Dataset S3B and 3D). The alignment bam files were taken for further analysis.

SNP variants calling based on RNA-Seq dataset

We followed the general GATK version 4 Best Practices to call genetic variants from Illumina RNA-seq data. We first sorted the above alignment bam data using samtools v1.9 [59], and marked duplicates by using PICARD v2.8.0 (http://broadinstitute.github.io/picard). Reads with N in the cigar were split into multiple supplementary alignments and hard clips mismatching overhangs using the SplitNCigarReads function in

583 GATK v4.1.9. By using BaseRecalibrator and ApplyBQSR functions in GATK, we further recalibrated
584 base quality scores with SNP variants that were called with the genomic sequencing dataset of the mice
585 individuals from the same populations [19] to get analysis-ready reads. Following, we called raw genetic
586 variants for each individual using the HaplotypeCaller function in GATK, and jointly genotyped genetic
587 variants for all the individuals using the GenotypeGVCFs function. We only retained genetic variants that
588 passed the hard filter "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -
589 8.0 || SOR > 3.0" for further analysis [19].

590

591 *Iso-Seq transcriptome classification and filtering*

592 We performed the quality control analysis for the above merged non-redundant PacBio Iso-Seq
593 transcriptome using SQANTI3 v4.2 [31], with the input datasets of Ensembl v103 [30] GRCm39/mm39 reference
594 genome and gene annotation, FLNC read counts, Isoform expression levels and STAR output alignment
595 bam files and splice junction files from RNA-Seq short reads, mouse transcription start sites (TSS) collected
596 in refTSS database v3.1 [63], and curated set of poly(A) sites and poly(A) motifs in PolyASite portal v2.0 [64].

597 We filtered out isoforms of potential artifacts mainly by following [31]. Mono-exonic transcripts were
598 excluded, as they tend more likely to be experimental or technical artifacts [31]. Isoforms with unreliable 3'end
599 because of a possible intrapriming event (intrapriming rate above 0.6) were also removed from the dataset.
600 We kept the remaining isoforms that met both of the following criteria: 1) no junction is not labeled as RT-
601 Switching; 2) all junctions are either canonical (GT/AG; GC/AG; AT/AC) or supported by at least 3
602 spanning reads based on STAR junction output file. All isoforms that passed the above filters were taken
603 for further analysis.

604 Given their matching status to the Ensembl mouse transcriptome v103 [30], the above transcripts were
605 classified into eight distinct categories using SQANTI3 [31], as depicted in Figure 2A: i) Full Splice Match
606 (FSM, matching perfectly to a known transcript); ii) Incomplete Splice Match (ISM, matching to a
607 subsection of a known transcript); iii) Novel In Category (NIC, with known splice sites but novel splice
608 junctions); iv) Novel Not in Category (NNC, with at least one unannotated splice site); v) Fusion (F, fusion
609 of adjacent transcripts); vi) Genic (G, overlapping with intron); vii) Antisense (A, on the antisense strand of
610 an annotated gene); viii) Intergenic (I, within the intergenic region). The transcripts matching perfectly to a
611 whole (FSM) or subsection (ISM) of reference annotated transcripts are designated known transcripts, and
612 the others as novel transcripts.

We evaluated the reliability of isoforms from each category, on the basis of reference annotation and empirical information from three distinct aspects separately (Supplementary Text on Methods and Figure S6): i) Transcription start site (TSS); ii) Transcription Termination Site (TTS); iii) splice junction (SJ). In comparison to the well-support FSM transcripts, the transcripts from other categories show reduced confidence levels in terms of TSS (Supplementary Figure S6B), but no reduction for TTS and SJ (Supplementary Figure S6D and 6E). This might hint at a failure to capture accurate TSS for those transcripts [31]. For instance, it is still possible that some ISM transcripts came from partial fragments, due to the imperfect targeting of 5'-cap or the degradation of transcripts in the later steps. To address this concern, we further excluded the non-FSM transcripts without support for TSSs (Supplementary Figure S6C).

Feature characterization of all the transcripts

We characterized the features of all the detected transcripts from seven different aspects: i) exon number; ii) transcript length; iii) CDS length; iv) fraction of coding transcripts; v) fraction of NMD transcripts; vi) the number of individuals with expression; vii) transcript expression level. All these feature results were extracted from the output of SQANTI3 analysis.

In the SQANTI3 pipeline [31], the potential coding capacity and ORFs from the transcript sequences were predicted using GeneMarkS-T (GMST) algorithm [65]. An NMD transcript is designated if there's a predicted ORF, and the CDS ends at least 50bp before the last junction for the respective transcript. The expression level of each transcript for each sample was computed, on the basis of the number of supported FLNC reads, and normalized in the unit of TPM (transcript per million). Transcripts with expression in respective tissues were defined as the ones with non-zero TPM values. In addition, we quantified the expression of each transcript using another program named Kallisto v0.46.2 [66], for which the expression quantification was based on the alignment of Illumina RNA-Seq data to the merged isoform dataset derived from PacBio Iso-Seq data, as shown in the above text.

Analysis of local AS events

We used the SUPPA2 program [34] to identify local alternative splicing (AS) events in the transcriptome. These local AS events are categorized into seven groups, including skipped exon (SE), retained intron (RI), alternative 5' splice site (A5), alternative 3' splice site (A3), mutually exclusive exon (MX), alternative first exon (AFE), and alternative last exon (ALE).

For the transcripts with predicted ORFs, we analyzed how the local AS events change the ORF structures. In case the start codon position of the ORF lands in the region of a local AS event (Figure 4C), the focal AS event is defined to cause a change in the respective coding sequence. On the other hand, in case the start codon position of the ORF falls downstream of the local AS event, it does not influence the ORF structure.

Rarefaction and subsampling

We investigated whether the PacBio Iso-Seq data provide sufficient coverage to detect all the isoform diversity for the given sequencing depth of a single individual and the number of sampled individuals in the main experiment. Concerning the sequencing depth at the individual level, we randomly selected one sample from each of the seven assayed populations, and subsampled portions of FLNC reads from each sample chosen for 100 times, ranging from 5% to 100%, at 5% intervals, and computed the fraction and variance of detected isoforms for each round of subsampling. Regarding the number of sampled individuals, we subsampled subsets of all the 48 assayed individuals for 100 times, ranging from 2 to 48, at an interval of 2, and computed the fraction and variance of detected isoforms for each round of subsampling.

We tested two alternative models to determine whether the number of detected isoforms would continue to increase or has approached saturation [67]: a generalized linear model with logarithmic behavior (ever-increasing) or a self-starting nonlinear regression model (saturating). The best fit was decided based on the minimum BIC value between the two models, and the saturating model was the best fit for both lines of analysis. The sequencing depth at the individual level has reached saturation with the given sequencing data, while the number of sampled individuals has not. All the analyses were performed in R v4.2.3, using the functions glm, nls, SSasymp, and BI from the "stats" package [68].

Test on the influence of noise on isoform diversity

We analyzed the influence of noise on isoform diversity using two different tests. For the first test, we extracted the list of isoforms that were supported by singleton FLNC reads in the *de novo* clustering step, and filter out potential artifacts using the same procedure as shown above. We performed the rarefaction analysis to test whether the number of singleton-supported isoforms has reached saturation with the sequencing depth at individual level, following the aforementioned pipeline. We further compared the saturation curves between the isoforms represented by singleton FLNC read and those represented by two or more FLNC reads.

For the second test, we tested whether genes that express one dominant isoform produce on average more additional isoforms when they are higher expressed in a given individual, for the isoform dataset supported by two or more FLNC reads. We selected 448 genes that fulfil the following criteria: 1) more than 10 isoforms; 2) the average expression level (in the unit of TPM) for the top expressed transcript (T) is at least five times higher than the cumulative expression level for the other (O) isoforms from the same locus: ($T / \Sigma (O) \geqq 5$; averaged across all individuals from the ingroup populations, *i.e.*, excluding *Mus spretus* and *Mus spicilegus,* because too many genes show major overall expression changes between the species). For comparison, we selected a set of 3,450 genes with more than 10 isoforms, without considering their expression level properties. We exploited one-sided Kendall´s tau tests to calculate the significance levels on the positive correlation between isoform number and the top expression level between individuals, and compared the fraction of genes with significant correlation (p-value < 0.05) between the two gene sets.

Test on the plasticity effect on isoform diversity

Following the above procedure, we generated the list of filtered high-confidence isoform and SNP variants for all the 58 mice individuals under both laboratory and natural environmental conditions. The SNP variants were called from the Illumina RNA-Seq dataset generated in this study (shown in the above text), for which the same set of mice individuals was used. To reduce computation complexity, we performed LD pruning on the SNP data set by using PLINK v1.90b4.6 [69], removing one of a pair of SNPs with LD $\geqslant$0.2 in sliding window of 500 SNPs and step wise of 100 SNPs.

Firstly, we performed principal component analysis (PCA) on the individual SNP and isoform landscape using the R package "ggfortify" v0.4.16. The PC-score of the top PCs-axes was extracted and analyzed according to species/subspecies and population differentiation across all samples. We used linear models implemented in the R package "stats" for analysis and conducted pair-wise post-hoc comparison with bonferroni correction for multiple testing in case of significant main effects.

Secondly, we built a phylogenetic tree for all the assayed population using the R package "ape" v5.7-1 [70], based on the presence matrix of fixed (*i.e.*, present in 100% of the individuals) or almost-fixed (*i.e.*, present in >80% of the individuals) isoform and SNP variants in each population. Euclidean distance was used as the distance measure between each pair of populations, and the neighbor-joining tree estimation function was used to build the phylogenetic relations. The boot.phylo function implemented in the same R package was used to perform 1,000 bootstrap replications, and population split nodes of high confidence were taken as the ones with at least 70% bootstrap support values.

We further performed two lines of analysis to bypass the possible bias: 1) excluding the isoforms derived from loci with significant expression levels (FDR < 0.05) between the GE and NE populations; 2) controlling the sequencing depth via randomly choosing the same number of PacBio FLNC reads from the datasets of other individuals in the main experiment as the average in the NE datasets and the same number of Illumina RNA-Seq reads from the NE datasets as the average in the datasets of main experiment. The same procedures were applied to call high-confidence isoform and SNP variants and to perform the population divergence analysis.

Frequency spectrum analysis of house mouse specific transcripts

We defined house mouse specific transcripts (n = 65,201) as the ones that are detectable in the house mouse natural populations, but absent in the outgroup species samples. In addition, we defined a set of conserved transcripts (n = 39,537), *i.e.*, present both in the house mouse and outgroup species samples. We compared the transcript density with respect to the number of individuals with an expression between two groups of transcripts. The individual number sizes for two groups of transcripts were normalized on the basis of the total number of tested individuals (n = 48).

We retrieved two SNP datasets for comparison analysis: one was called based on the Illumina RNA-Seq datasets generated in this study (same set of individuals for Iso-Seq data to generate transcripts, as shown above), and the other from genomic sequencing data of the same populations (equal number of individuals for each population, but different individuals) [19]. For both datasets, we only kept the house mouse specific SNP variants with unambiguous ancestral states in outgroup species (*i.e.*, the same homozygous genotype for 2 outgroup species), while with alternative alleles in house mouse individuals. In addition, we generated a house mouse specific ORF dataset, by collapsing the transcripts forming the identical ORFs. For all four types of variants (Transcripts, two types of SNPs, ORFs), we calculated their frequency distribution at three different levels (subspecies, population, and individual) by counting individuals with positive evidence of each variant, without distinguishing the homozygous and heterozygous status.

For the former SNP dataset, we further predicted the functional effects of each SNP by using Ensembl VEP v103 [41], based on the gene annotation data from Ensembl version 103. Consistent with Ensembl variation annotation [41], we categorized these SNPs into four groups given their predicted impacts: i) High effect - SNPs causing the gain/loss of start/stop codon or change of the splicing acceptor/donor sites; ii) Moderate effect - SNPs resulting in a different amino acid sequence; iii) Low effect - SNPs occurring within the region of the splice site, changing the final codon of an incompletely annotated transcript, changing the bases of

start/stop codon (while start/terminator remains), or where there is no resulting change to the encoded amino acid; iv) Modifier effect - SNPs occurring within the genes' non-coding regions (*e.g.*, UTR and intron). The frequency spectrum of house mouse specific transcripts was further compared with the site frequency spectrum of SNPs from the four above-defined categories. We quantified the distances between spectrum distributions by using two-sided Kolmogorov-Smirnov tests, and calculated the statistical significances of the fraction of individual private variants by using Fisher's exact tests.

Data availability

The raw Illumina RNA-Seq data and PacBio Iso-Seq data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under study accession number PRJEB54000 and PRJEB53988 (the main experiment), PRJEB67296 and PRJEB67298 (the plasticity test experiment), and PRJEB54001 (the Iso-Seq protocol optimization experiment). Alignment bam files, GTF track data, and SNP VCF files, and supplementary datasets are stored at the ftp site: https://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/mouse_brain_isoform/. All the essential computing codes, parameters, and related data sets are available at GitLab: https://gitlab.gwdg.de/wenyu.zhang/mouse_brain_transcriptome/.

Ethics statement

All the mice were kept according to FELASA (Federation of European Laboratory Animal Science Association) guidelines, with the permit from the Veterinäramt Kreis Plön: 1401-144/PLÖ-004697. Since only organ retrieval, but no animal experiments were involved, an ethical permit was not required, but the respective animal welfare officer at the University of Kiel (Prof. Schultheiss) was informed about the sacrifice of the mice individuals for this study, as required by law.

Author contributions

W.Z. and D.T. designed the study, interpreted the data, and wrote the paper. W.Z collected the animals for the comparative study, generated the materials for the sequencing dataset, analysed the data, and performed the statistical analysis. A.G designed the semi-natural environment study, collected the respective animals

763 and contributed to the statistical analysis of the data. Y.G, K.U, and B.H. contributed to the bioinformatic
764 data analysis and the sequencing runs. All authors read and approved the final manuscript.

765

766 ## Acknowledgements

767 We appreciate Christine Pfeifle, Heike Harre, Milan Jovicic, and Mustafa Al-Ameer for mice breeding and
768 handling. We thank Carsten Fortmann-Grote and Chen Xie for computing assistance, and other lab members
769 for helpful discussion. We thank Henrik Kaessmann for helpful suggestions on the manuscript. Computing
770 was supported by the Wallace high-performance computing cluster of the Max Planck Institute for
771 Evolutionary Biology. This work was supported by institutional funding through the Max Planck Society to
772 D.T. and W.Z., and the starting research funds by Northwestern Polytechnical University (grant number:
773 G2022KY05106), and National Natural Science Foundation of China grants (grant number: 32370665) to
774 W.Z.

775

776 ## References

777 1 Bush, S. J., Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Alternative splicing and the evolution
778 of phenotypic novelty. *Philos Trans R Soc Lond B Biol Sci* **372**, doi:10.1098/rstb.2015.0474 (2017).
779 2 Salisbury, S. J., Delgado, M. L. & Dalziel, A. C. Alternative splicing: An overlooked mechanism
780 contributing to local adaptation? *Mol Ecol* **30**, 4951-4954, doi:10.1111/mec.16177 (2021).
781 3 Verta, J. P. & Jacobs, A. The role of alternative splicing in adaptation and evolution. *Trends Ecol*
782 *Evol* **37**, 299-308, doi:10.1016/j.tree.2021.11.010 (2022).
783 4 Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic
784 diversity. *Nat Rev Genet*, doi:10.1038/s41576-022-00514-4 (2022).
785 5 Martinez-Juarez, A. & Moreno-Mendoza, N. Mechanisms related to sexual determination by
786 temperature in reptiles. *J Therm Biol* **85**, 102400, doi:10.1016/j.jtherbio.2019.102400 (2019).
787 6 Lutz, U. *et al.* Natural haplotypes of FLM non-coding sequences fine-tune flowering time in
788 ambient spring temperatures in Arabidopsis. *Elife* **6**, doi:10.7554/eLife.22114 (2017).
789 7 Rosenkranz, R. R. E., Ullrich, S., Lochli, K., Simm, S. & Fragkostefanakis, S. Relevance and
790 Regulation of Alternative Splicing in Plant Heat Stress Response: Current Understanding and
791 Future Directions. *Front Plant Sci* **13**, 911277, doi:10.3389/fpls.2022.911277 (2022).
792 8 Preussner, M. *et al.* Body Temperature Cycles Control Rhythmic Alternative Splicing in Mammals.
793 *Mol Cell* **67**, 433-446 e434, doi:10.1016/j.molcel.2017.06.006 (2017).
794 9 Haltenhof, T. *et al.* A Conserved Kinase-Based Body-Temperature Sensor Globally Controls
795 Alternative Splicing and Gene Expression. *Mol Cell* **78**, 57-69 e54,
796 doi:10.1016/j.molcel.2020.01.028 (2020).
797 10 Byrne, A., Cole, C., Volden, R. & Vollmers, C. Realizing the potential of full-length transcriptome
798 sequencing. *Philos Trans R Soc Lond B Biol Sci* **374**, 20190097, doi:10.1098/rstb.2019.0097 (2019).
799 11 Leung, S. K. *et al.* Full-length transcript sequencing of human and mouse cerebral cortex identifies
800 widespread isoform diversity and alternative splicing. *Cell Rep* **37**, 110022,
801 doi:10.1016/j.celrep.2021.110022 (2021).

12 Ferrandez-Peral, L. *et al.* Transcriptome innovations in primates revealed by single-molecule long-read sequencing. *Genome Res*, doi:10.1101/gr.276395.121 (2022).
13 Pose, D. *et al.* Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* **503**, 414-417, doi:10.1038/nature12633 (2013).
14 Guenet, J. L. & Bonhomme, F. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet* **19**, 24-31, doi:10.1016/s0168-9525(02)00007-0 (2003).
15 Phifer-Rixey, M. & Nachman, M. W. Insights into mammalian biology from the wild house mouse Mus musculus. *Elife* **4**, doi:10.7554/eLife.05959 (2015).
16 Harr, B. *et al.* Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Sci Data* **3**, 160075, doi:10.1038/sdata.2016.75 (2016).
17 Teschke, M., Mukabayire, O., Wiehe, T. & Tautz, D. Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans. *Genetics* **180**, 1537-1545, doi:10.1534/genetics.108.090811 (2008).
18 Staubach, F. *et al.* Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (Mus musculus). *PLoS Genet* **8**, e1002891, doi:10.1371/journal.pgen.1002891 (2012).
19 Zhang, W., Xie, C., Ullrich, K., Zhang, Y. E. & Tautz, D. The mutational load in natural populations is significantly affected by high primary rates of retroposition. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2013043118 (2021).
20 Zhang, W. & Tautz, D. Tracing the Origin and Evolutionary Fate of Recent Gene Retrocopies in Natural Populations of the House Mouse. *Mol Biol Evol* **39**, doi:10.1093/molbev/msab360 (2022).
21 Bekpen, C., Xie, C. & Tautz, D. Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. *BMC Evol Biol* **18**, 121, doi:10.1186/s12862-018-1232-z (2018).
22 Murat, F. *et al.* The molecular evolution of spermatogenesis across mammals. *Nature* **613**, 308-316, doi:10.1038/s41586-022-05547-7 (2023).
23 Cartolano, M., Huettel, B., Hartwig, B., Reinhardt, R. & Schneeberger, K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLoS One* **11**, e0157779, doi:10.1371/journal.pone.0157779 (2016).
24 Kuo, R. I. *et al.* Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323, doi:10.1186/s12864-017-3691-9 (2017).
25 Chimpanzee, S. & Analysis, C. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:10.1038/nature04072 (2005).
26 Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195-201, doi:10.1038/nature13679 (2014).
27 Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**, e1001236, doi:10.1371/journal.pgen.1001236 (2010).
28 Wan, Y. & Larson, D. R. Splicing heterogeneity: separating signal from noise. *Genome Biol* **19**, 86, doi:10.1186/s13059-018-1467-4 (2018).
29 Saudemont, B. *et al.* The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol* **18**, 208, doi:10.1186/s13059-017-1344-6 (2017).
30 Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884-D891, doi:10.1093/nar/gkaa942 (2021).
31 Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res*, doi:10.1101/gr.222976.117 (2018).
32 Naftaly, A. S., Pau, S. & White, M. A. Long-read RNA sequencing reveals widespread sex-specific alternative splicing in threespine stickleback fish. *Genome Res* **31**, 1486-1497, doi:10.1101/gr.274282.120 (2021).

33 Chang, Y. F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**, 51-74, doi:10.1146/annurev.biochem.76.050106.093909 (2007).

34 Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* **19**, 40, doi:10.1186/s13059-018-1417-1 (2018).

35 Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**, 71-87, doi:10.1038/s41576-019-0173-8 (2020).

36 Landry, J. R., Mager, D. L. & Wilhelm, B. T. Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* **19**, 640-648, doi:10.1016/j.tig.2003.09.014 (2003).

37 Huang, K. K. *et al.* Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. *Genome Biol* **22**, 44, doi:10.1186/s13059-021-02261-x (2021).

38 Prabh, N., Linnenbrink, M., Jovicic, M. & Guenther, A. Fast adjustment of pace-of-life and risk-taking to changes in food quality by altered gene expression in house mice. *Ecol Lett* **26**, 99-110, doi:10.1111/ele.14137 (2023).

39 Manley, J. L. & Krainer, A. R. A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes Dev* **24**, 1073-1074, doi:10.1101/gad.1934910 (2010).

40 Lopez-Maestre, H. *et al.* SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res* **44**, e148, doi:10.1093/nar/gkw655 (2016).

41 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

42 Zou, X. *et al.* Mammalian splicing divergence is shaped by drift, buffering in trans, and a scaling law. *Life Sci Alliance* **5**, doi:10.26508/lsa.202101333 (2022).

43 Meinke, S. *et al.* Srsf10 and the minor spliceosome control tissue-specific and dynamic SR protein expression. *Elife* **9**, doi:10.7554/eLife.56075 (2020).

44 Neumann, A. *et al.* Alternative splicing coupled mRNA decay shapes the temperature-dependent transcriptome. *Embo Reports* **21**, doi:10.15252/embr.202051369 (2020).

45 Li, K. & Wang, Z. Q. Splicing factor SRSF2-centric gene regulation. *International Journal of Biological Sciences* **17**, 1708-1715, doi:10.7150/ijbs.58888 (2021).

46 Malko, D. B., Makeev, V. J., Mironov, A. A. & Gelfand, M. S. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res* **16**, 505-509, doi:10.1101/gr.4236606 (2006).

47 Harr, B. & Turner, L. M. Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Mol Ecol* **19 Suppl 1**, 228-239, doi:10.1111/j.1365-294X.2009.04490.x (2010).

48 Lin, L. *et al.* Evolution of alternative splicing in primate brain transcriptomes. *Hum Mol Genet* **19**, 2958-2973, doi:10.1093/hmg/ddq201 (2010).

49 Mudge, J. M. *et al.* The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol Biol Evol* **28**, 2949-2959, doi:10.1093/molbev/msr127 (2011).

50 Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587-1593, doi:10.1126/science.1230612 (2012).

51 Ling, Z., Brockmoller, T., Baldwin, I. T. & Xu, S. Evolution of Alternative Splicing in Eudicots. *Front Plant Sci* **10**, 707, doi:10.3389/fpls.2019.00707 (2019).

52 Rogers, T. F., Palmer, D. H. & Wright, A. E. Sex-Specific Selection Drives the Evolution of Alternative Splicing in Birds. *Mol Biol Evol* **38**, 519-530, doi:10.1093/molbev/msaa242 (2021).

53 Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-909 e821, doi:10.1016/j.cell.2018.02.011 (2018).

54 Montes, M., Sanford, B. L., Comiskey, D. F. & Chandler, D. S. RNA Splicing and Disease: Animal Models to Therapies. *Trends Genet* **35**, 68-87, doi:10.1016/j.tig.2018.10.002 (2019).

55 Jbara, A. *et al.* RBFOX2 modulates a metastatic signature of alternative splicing in pancreatic cancer. *Nature* **617**, 147-153, doi:10.1038/s41586-023-05820-3 (2023).

56 Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat Genet* **54**, 1355-1363, doi:10.1038/s41588-022-01154-4 (2022).

57 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

58 Gordon, S. P. *et al.* Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. *PLoS One* **10**, e0132628, doi:10.1371/journal.pone.0132628 (2015).

59 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

60 Kuo, R. I. *et al.* Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**, 751, doi:10.1186/s12864-020-07123-7 (2020).

61 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

62 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

63 Abugessaisa, I. *et al.* refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J Mol Biol* **431**, 2407-2422, doi:10.1016/j.jmb.2019.04.045 (2019).

64 Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**, D174-D179, doi:10.1093/nar/gkz918 (2020).

65 Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78, doi:10.1093/nar/gkv227 (2015).

66 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

67 Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife* **5**, e09977, doi:10.7554/eLife.09977 (2016).

68 Team, R. C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2022).

69 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).

70 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528, doi:10.1093/bioinformatics/bty633 (2019).