Full-length transcript sequencing traces the brain isoform diversity in house mouse natural populations

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

Running title: Isoform diversity in house mouse populations

Manuscript type: Resource

Abstract

The ability to generate multiple RNA isoforms of transcripts from the same gene is a general phenomenon in eukaryotes. However, the complexity and diversity of alternative isoforms in natural populations remains largely unexplored. Using a newly developed full-length transcripts enrichment protocol, we sequenced full-length RNA transcripts of 48 individuals from outbred populations and subspecies of *Mus musculus*, as well as the closely-related sister species *Mus spretus* and *Mus spicilegus* as outgroups. This represents the largest full-length high-quality isoform catalog at the population level to date. In total, we reliably identify 117,728 distinct transcripts, of which only 51% were previously annotated. We show that the population-specific distribution pattern of isoforms is phylogenetically informative and reflects the segregating SNP diversity between the populations. We find that ancient house-keeping genes are the major source to the overall isoform diversity, and the recruiting of alternative first exon plays the dominant role in generating new isoforms. Given that our data allow to distinguish between population-specific isoforms and isoforms that are conserved across multiple populations, it is possible to refine the annotation of the reference mouse genome to a set of about 40,000 isoforms that should be most relevant for comparative functional analysis across species.

Key words

Isoform diversity, full-length transcript sequencing, Mus musculus, natural populations, LRS Special Issue

Introduction

The ability to generate multiple RNA isoforms (or transcripts) from the same gene increases vastly the complexity of transcriptome and proteome in eukaryotes (Keren et al. 2010; Nilsen and Graveley 2010). Alternative isoforms with different coding sequences can be generated through multiple types of alternative splicing events (Naftaly et al. 2021), including skipped exon (SE), retained intron (RI), mutually exclusive exons (MX), alternative 5' (A5'), and 3' (A3') splice sites. Isoform diversity can also be expanded via the inclusion of alternative transcription start sites (TSSs) or transcription termination sites (TTSs), resulting in differences in the untranslated regions (UTRs) (Landry et al. 2003; Tian and Manley 2017).

By analyzing high-throughput EST and short-read RNA-Seq data across species, a number of comparative studies on the complexity of isoform landscape have been performed in plants (Ling et al. 2019), insects

(Malko et al. 2006), vertebrates (Mudge et al. 2011; Barbosa-Morais et al. 2012), birds (Rogers et al. 2021), mice (Harr and Turner 2010), and primates (Lin et al. 2010). These initial comparative studies have provided important insights into the enormous diversity of alternative isoforms, but only at the local level of splicing junction events, due to the limitation of the short-read sequencing technology.

The recent development of long-read single-molecule sequencing technologies has enabled the capture of the full-length isoform diversity (Byrne et al. 2019), further facilitating the comparative analysis of alternative isoforms at the global transcript level in different species. Leung SK et al. (Leung et al. 2021) used long-read isoform sequencing to generate full-length transcript sequences in the human and mouse cortex. They showed not only a similar global pattern of isoform diversity between human and mouse, but also striking differences on certain genes between these two species. Another recent study from Ferrandez-Peral L (Ferrandez-Peral et al. 2022) generated an extensive full-length isoform catalog in primates, and found that a high isoform diversity in immune genes correlates with signals of positive selection in the coding regions of the genes. These between-species comparison studies have provided insights into the complexity of alternative isoforms at a deep evolutionary time scale. But the diversity and pattern of alternative isoform landscape within natural populations remain largely unexplored (Verta and Jacobs 2022).

Owing to its well-defined evolutionary history (Guenet and Bonhomme 2003; Phifer-Rixey and Nachman 2015), the house mouse (*Mus musculus*) has been shown to be a particularly suitable model system for studying the 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 (Harr et al. 2016): 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 (Harr et al. 2016). We have previously used the house mouse system to study the evolutionary pattern of single nucleotide variants (Staubach et al. 2012), gene copy number variants (Pezer et al. 2015), and gene retrocopy variants (Zhang et al. 2021; Zhang and Tautz 2022). This made the house mouse an ideal system to directly trace the origin and pattern of alternative isoforms in natural populations.

In this study, we exploit PacBio Iso-Seq and Illumina RNA-Seq technologies to explore the diversity of alternative isoforms in the house moue natural populations and closely-related outgroup species. 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 (Bekpen et al. 2018), but

89 with many more of these transcripts being likely to be functional in the brain compared to the testis, where
90 there is a lot of expression due to a transcriptionally permissive chromatin environment, especially in late
91 spermatogenic cell types (Murat et al. 2023). We develop an optimized protocol to capture predominantly
92 full-length transcripts and validated them with existing data. Applying this newly developed enrichment
93 protocol, we identify widespread brain isoform diversity in house mouse natural populations, with a
94 substantial fraction as novel transcripts not previously registered in the mouse genome annotations. We
95 show the individual isoform landscape can be used to successfully build the phylogenetic tree, and most of
96 the isoforms can serve as effective markers to differentiate house mouse natural populations. We further
97 analyze the origin of the isoform diversity, and find that ancient house-keeping genes contribute most to
98 the overall isoform diversity, and new 5'-exons from alternative promotors are the relatively most
99 abundant source of new transcripts. The transcriptomic data allow to define the set of conserved isoforms
100 across mouse species, which are of particular relevance for functional studies that use the mouse as a
101 biomedical model system in comparison to humans.

102

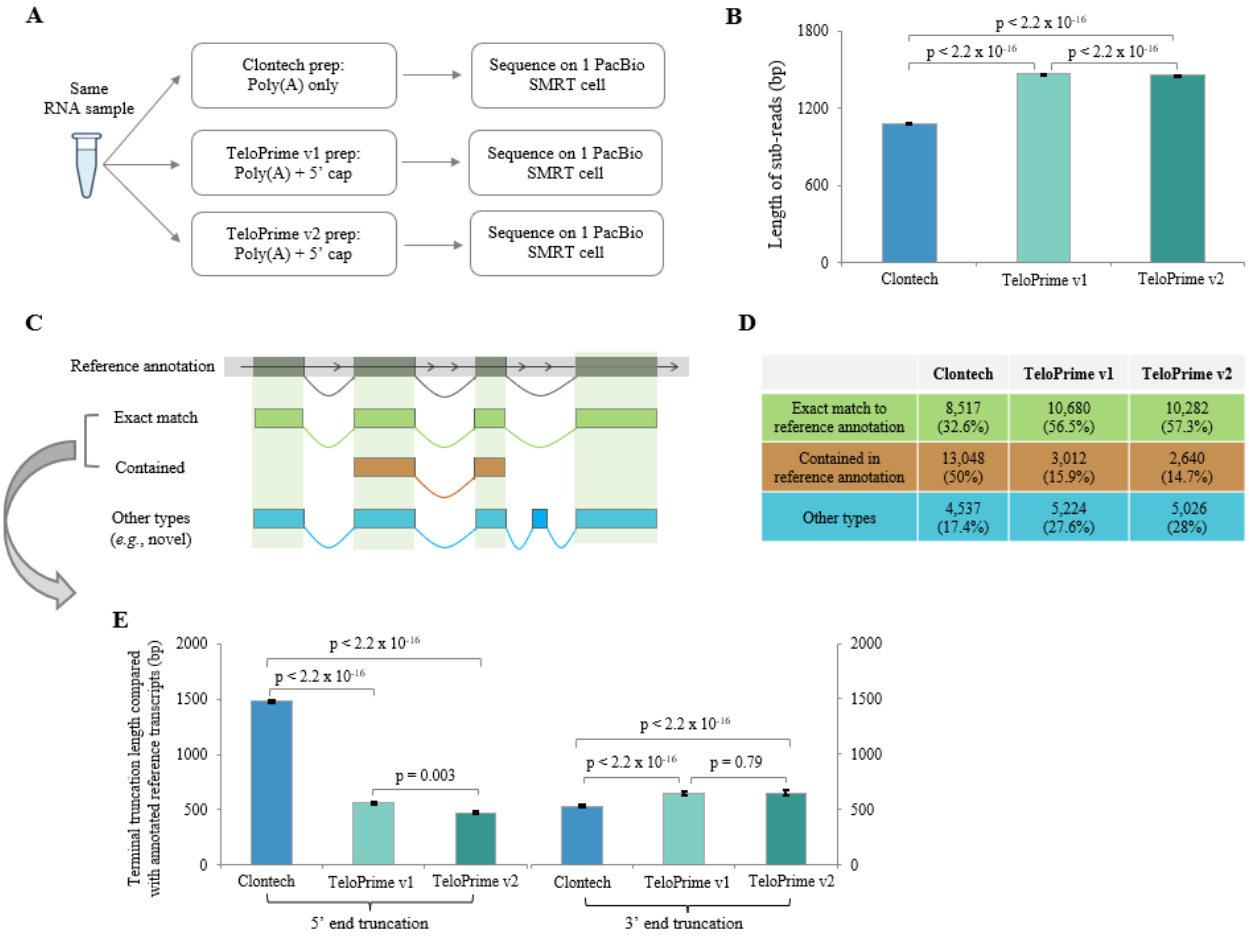
103 **Results**

104 *An optimized experimental protocol to enrich full-length transcript*

105 A key step for full-length RNA sequencing is the enrichment of intact transcripts without degradation,
106 ideally with hallmarks of both a 5' cap and a poly(A) tail. It has been reported that the TeloPrime full-
107 length cDNA Amplification kit could selectively synthesizes cDNA molecules from mRNAs carrying a 5'
108 cap, in complement to the standard PacBio cDNA library preparation protocol that could only selectively
109 synthesizes cDNA molecules from transcripts with a poly(A) tail (Cartolano et al. 2016). In this study, we
110 developed a new PacBio Iso-Seq library preparation protocol via combing the TeloPrime full-length
111 cDNA Amplification kit and an alternative oligo-dT primer from the standard PacBio cDNA synthesis kit,
112 thus can synthesize cDNA molecules from transcripts with both a 5' cap and a poly(A) tail.

113 To test the performance of TeloPrime-based protocol in an actual sample, we selected one mouse
114 individual from the Taiwan population (Supplementary Dataset S1A), and implemented three distinct
115 cDNA library preparation protocols for the RNA samples extracted from the brain (Figure 1A): i) standard
116 PacBio cDNA library preparation protocol; ii) TeloPrime-based protocol with cDNA Amplification Kit v1;
117 iii) TeloPrime-based protocol with cDNA Amplification Kit v2. TeloPrime-based kit v2 includes updated
118 reaction conditions in comparison with kit v1, while the method of cDNA generation remains unchanged.

Each of the three cDNA libraries was sequenced on one SMRT cell of the same sequencing run, and the sequencing reads were analyzed and compared using the same computational approach (see Methods).



	Clontech	TeloPrime v1	TeloPrime v2
Exact match to reference annotation	8,517 (32.6%)	10,680 (56.5%)	10,282 (57.3%)
Contained in reference annotation	13,048 (50%)	3,012 (15.9%)	2,640 (14.7%)
Other types	4,537 (17.4%)	5,224 (27.6%)	5,026 (28%)

Figure 1 Performance evaluation of three types of full-length transcript enrichment protocols. (A) shows a simplified flow chart of this evaluation experiment setup. (B) indicates the length distribution of sub-reads generated from three PacBio Iso-Seq libraries. (C) illustrates the major categories of Iso-Seq transcripts based on their alignment coordinates to the reference genome, and (D) for the distribution of these categories. (E) shows the terminal truncation length of Iso-Seq transcripts from both 5' and 3' ends, compared to the reference annotation in Ensembl v103. Only Iso-Seq transcripts with exact matches or wholly contained in reference annotation were taken for analysis. The error bars indicate the standard error of means (SEM) of the length values. Statistical p-values were computed using two-sided Wilcoxon rank sum tests.

As shown in Figure 1B, we found that the sub-reads generated from TeloPrime-based cDNA libraries are on average longer (TeloPrime v1: 1,467bp; TeloPrime v2: 1,459bp), compared with those from standard Clontech cDNA library (mean value: 1,085bp). Compared with that from the Clontech cDNA library (32.6%), a much higher fraction of isoforms from the TeloPrime v1 cDNA library (56.5%) and TeloPrime

135 v2 cDNA library (57.3%) showed a complete and exact match of the intron chain as the annotated
136 reference transcripts (Figure 1C and 1D). Moreover, 50% of the isoforms from the Clontech cDNA library
137 were contained in the annotated reference transcripts (a truncation of terminal exons, likely the
138 degradation of the transcripts). In comparison, this value is only 15.9% in the TeloPrime v1 cDNA library
139 and 14.7% in the TeloPrime v2 cDNA library, suggesting that TeloPrime-based cDNA enrichment
140 protocols are more likely to enrich full-length transcripts. An in-depth investigation of the truncation
141 lengths showed that the isoforms from the Clontech cDNA library are actually at least 2-fold more
142 severely truncated from 5'end than those from TeloPrime-based cDNA libraries (similarly less extent
143 from 3'end, Figure 1E), which confirmed the reliability of TeloPrime-based cDNA kits to selectively
144 enrich mRNA transcripts carrying a 5' cap.

145 Overall, these findings suggest that the newly developed TeloPrime-based protocol indeed provides a
146 better solution to enrich full-length transcripts. Given the optimized reaction conditions, we applied
147 TeloPrime-based protocol with cDNA Amplification Kit v2 to generate cDNA libraries for PacBio Iso-
148 Seq for the following analysis.

149

150 *Alternative isoform landscape in the wild mice brain transcriptome*

151 With the above optimized experimental protocol, we analyzed the alternative isoform landscape in the
152 whole brain transcriptomes for forty-eight unrelated outbred wild-type mice individuals raised under
153 tightly controlled laboratory conditions (Supplementary Dataset S1 and Dataset S2). They included forty
154 house mouse (*Mus musculus*) individuals derived from five natural populations in the three major
155 subspecies (*M. m. domesticus*, *M. m. musculus*, and *M. m. castaneus*), as well as eight individuals from
156 two closely related outgroup species (*Mus spicilegus* and *Mus spretus*) (Figure 2A). To define high-quality
157 transcriptomes, we performed both PacBio Iso-Seq and Illumina RNA-Seq for each brain sample. As a
158 result, an average of 60.1 (SD: 9.6) million raw sub-reads and 28.2 (SD: 3.3) million raw read pairs per
159 sample were produced for PacBio Iso-Seq and Illumina RNA-Seq, respectively (Supplementary Table S1).
160 This dataset represents the largest full-length isoform category at a comparative population level to date.

161 An overview of the methodology to generate the high-quality transcriptome is given in Supplementary
162 Figure S1. In brief, the raw PacBio Iso-Seq subread data were processed to produce circular consensus
163 sequences (CCS), and further refined to generate full-length non-chimeric (FLNC) reads using the
164 IsoSeq3 pipeline. To identify high-confidence isoforms, the FLNC reads were subject to *de novo*
165 clustering using the same IsoSeq3 pipeline to generate an average of 49,147 (SD: 4,554) non-redundant

166 isoforms per sample (Supplementary Dataset S1B), with the support from at least two FLNC reads. We
167 aligned these unique isoforms of each sample to the GRCm39/mm39 reference genome using minimap2
168 (Li 2018), and used the TAMA program (Kuo et al. 2020) with optimized parameters (Supplementary
169 Text of Methods) to collapse and merge the transcript models across all 48 samples into a single non-
170 redundant transcriptome. We further refined a computational pipeline to filter out low-quality transcripts
171 and those of potential artifacts (see Methods), resulting in a total of 117,728 high-confidence distinct
172 transcripts derived from 15,012 distinct loci (Supplementary Dataset S3).

173 To evaluate the reliability of these transcripts, we assessed their quality for three distinct criteria
174 separately (Supplementary Figure S4A): i) Transcription start site (TSS); ii) Transcription Termination
175 Site (TTS); iii) Splice Junctions (SJ). Reliable TSSs are defined as those supported by Ensembl
176 GRCm39/mm39 reference genome annotation (Howe et al. 2021), collection of experimentally validated
177 TSS peaks (Abugessaisa et al. 2019), and a significantly greater coverage ratio between inside isoform and
178 outside isoform than null hypothesis (Tardaguila et al. 2018) (Supplementary Figure S3). Similarly,
179 reliable TTSs are defined as those supported by the reference genome annotation (Howe et al. 2021), and
180 collection of experimentally validated poly(A) sites and motifs (Herrmann et al. 2020). The reliable SJs
181 are defined as those present in reference genome annotation (Howe et al. 2021), and those belonging to the
182 canonical splicing signals (GT-AG, GC-AG, and AT-AC) (Parada et al. 2014). A more detailed discussion
183 on evaluating the reliability of the detected transcripts can be found in Supplementary Text of Methods.
184 Based on the above-defined criteria, we found that 98.7% of the TSSs, 94.4% of the TTSs, and 99.1% of
185 the SJs from all the detected transcripts are supported by respective previous evidence (Supplementary
186 Dataset S3), and a comparable large extent (>96% for all three features) of transcripts could be validated
187 for all the seven tested populations (Supplementary Figure S6).

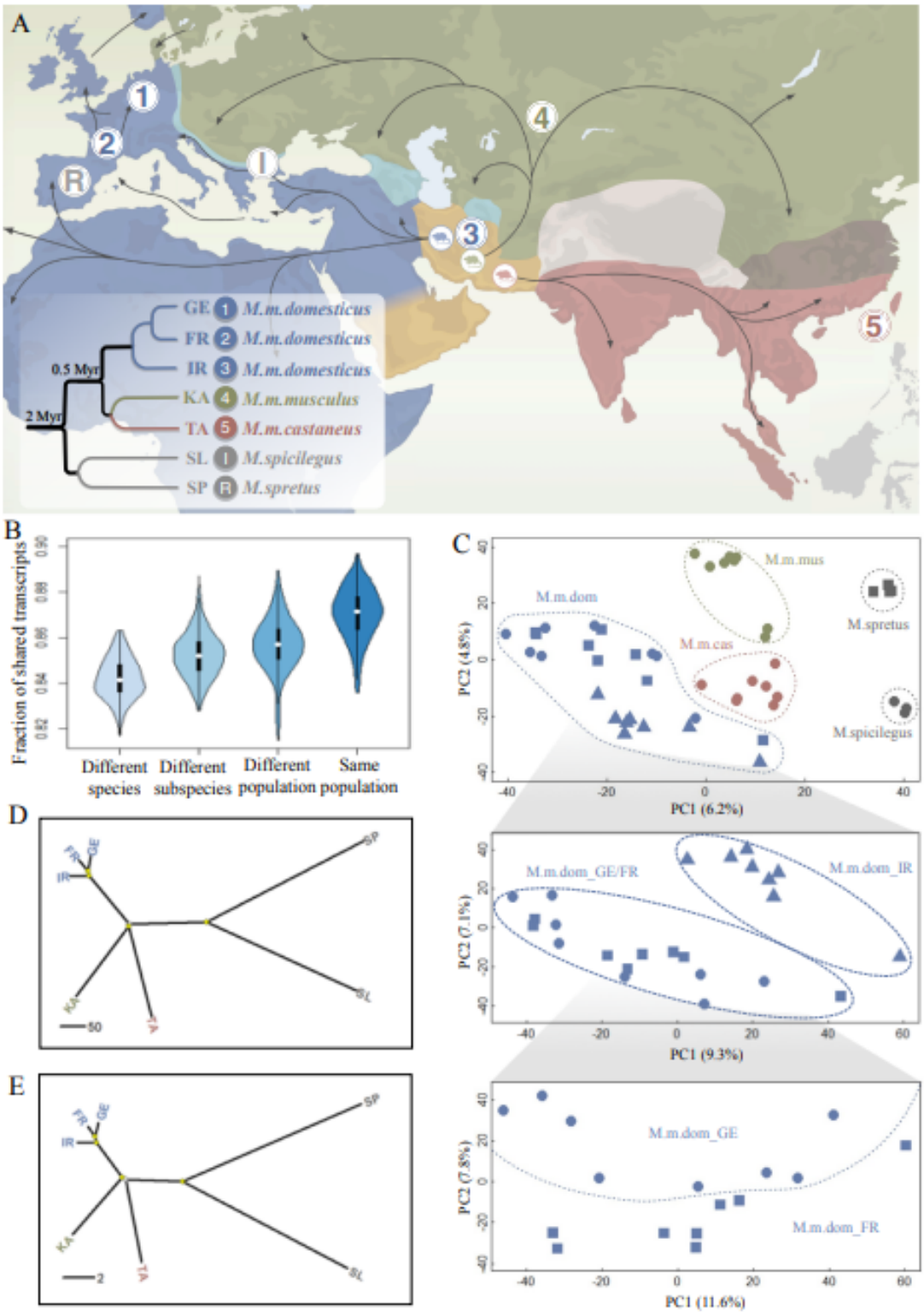


Figure 2 Geographic locations and phylogenetic relations of sampled mice individuals. (A) Geographic locations for the sampled mouse populations in this study. Territory areas for each house mouse subspecies: *M. m. domesticus* (blue), *M. m. musculus* (green), and *M. m. castaneus* (brown). Black arrows indicate possible migration

192 routes, mainly during the spread of agriculture and trading. The inset figure shows the canonical phylogenetic
193 relationships among the house mouse populations and outgroup species (branch lengths not scaled)**.** Geographic
194 locations: 1, Germany (GE); 2, France (FR); 3, Iran (IR); 4, Kazakhstan (KA); 5, Taiwan (TA). The sampling
195 locations for the outgroup species are I, Slovakia (SL) for *M. spicilegus* and R, Spain (SP) for *M. spretus*. (B) shows
196 the fraction of shared isoforms for pair-wise individual comparison. The fraction of shared isoforms between each
197 pair of individuals is defined as the number of overlapping isoforms divided by the average number of detected
198 isoforms of the two individuals to compare. Black boxes represent the interquartile range (IQR, distance between the
199 first and third quartiles), with white dots in the middle to denote the median. The boundaries of the whiskers (also the
200 ranges of violins) are based on the 1.5 IQR values for both sides. (C) shows the projection of the top two PCs of
201 isoform variation in house mouse and outgroup individuals. Enlarged insets represent the results for the three
202 populations from the subspecies of *M. m. domesticus*, as they cannot be well distinguished in the main figure. (D)
203 and (E) show phylogenetic trees built on the basis of isoform and SNP variants fixed within each population,
204 respectively. Split nodes marked in yellow are the ones with bootstrap support value >70%.

205

206 *Alternative isoforms as population differentiation markers*

207 Given the high numbers of isoforms in wild mice brain transcriptomes across populations (Table 1), we
208 asked whether the isoform diversity landscape can distinguish individuals, *i.e.*, whether they could
209 contribute to individual differences. If this is the case, they could also be used to trace the divergence
210 trajectories of the populations.

211 A pairwise individual comparison of the sharing of isoforms reveals a higher fraction of shared isoforms
212 for individuals from the same species/subspecies/population (Figure 2B). The overall principal component
213 analysis (PCA) shows clear clusters of individuals according to their origin. Still, considerable diversity
214 within each cluster implies that isoform diversity indeed confers a form of individuality to each animal
215 (Figure 2C). We further used the presence/absence matrix of fixed isoforms within each population to
216 build an overall phylogeny (Figure 2D), which has the same general topology as a phylogeny based on the
217 SNP variants called from Illumina RNA-Seq datasets from the same set of mice individuals (Figure 2E).

218 Given these observations, we performed an isoform frequency-based approach to detect the isoforms as
219 population-differentiating markers (see Methods). Each isoform was assigned a population differentiation
220 index (PDI) to indicate its power to differentiate populations. After applying a false discovery rate (FDR)
221 threshold of 0.05, we detected in total 113,008 isoforms (96% of all isoforms) that could differentiate
222 populations from house mouse and outgroup species (with significantly larger PDI values compared with
223 null distribution from 1,000 times random shuffling; see Supplementary Dataset S3). We conclude that
224 isoforms can serve as effective markers to differentiate house mouse natural populations.

225 **Table 1 Statistics on the detected transcripts in the house mouse populations and outgroups.** The digit succeeding the character N indicates the number of
226 sequenced mouse individuals in each population. The graphic depiction of isoform categories is shown in Figure 3A. Isoforms categorized as FSM and ISM are
227 defined to be known transcripts, gene loci with any isoform not categorized as either Antisense or Intergenic to be annotated genes.

228

		M.m.dom		**M.m.mus**	**M.m.cas**	**M.spicilegues**	**M.spretus**
	GE (N: 8)	**FR (N: 8)**	**IR (N: 8)**	**KA (N: 8)**	**TA (N: 8)**	**SL (N: 4)**	**SP (N: 4)**
Unique genes	8,538 (SD:603)	8,282 (SD:640)	8,123 (SD:452)	8,304 (SD:200)	8,066 (SD:225)	7,840 (SD:254)	8,043 (SD:185)
Annotated	8,466 (SD:609)	8,215 (SD:651)	8,061 (SD:452)	8,226 (SD:198)	8,000 (SD:218)	7,745 (SD:261)	7,973 (SD:181)
Novel	72 (SD:13)	67 (SD:16)	62 (SD:18)	78 (SD:10)	67 (SD:16)	95 (SD:17)	70 (SD:6)
Unique transcripts	23,027 (SD:2,001)	21,609 (SD:2,315)	21,763 (SD:2,006)	22,380 (SD:1,553)	21,279 (SD:1,532)	20,644 (SD:858)	21,640 (SD:640)
Protein-coding	21,394 (SD:2,062)	20,024 (SD:2,431)	20,155 (SD:1,941)	20,798 (SD:1,486)	19,642 (SD:1,435)	18,786 (SD:929)	20,063 (SD:597)
Nonsense-mediated decay (NMD)	1,152 (SD:226)	1,071 (SD: 220)	1,098 (SD:209)	1,125 (SD:146)	1,142 (SD:216)	1,038 (SD:146)	1,131 (SD:46)
Non-protein-coding	1,634 (SD:239)	1,585 (SD:284)	1,608 (SD:324)	1,582 (SD:235)	1,637 (SD:261)	1,858 (SD:73)	1,577 (SD:63)
Known transcripts	17,266 (SD:1,216)	16,306 (SD: 1,419)	16,343 (SD:1,257)	16,487 (SD:861)	15,765 (SD:821)	15,202 (SD:512)	15,608 (SD:416)
Full Splice Match (FSM)	15,007 (SD:1,217)	14,245 (SD: 1,441)	13,654 (SD:1,053)	14,466 (SD:810)	13,561 (SD:602)	12,812 (SD:523)	13,621 (SD:297)
Incomplete Splice Match (ISM)	2,259 (SD:297)	2,060 (SD: 654)	2,689 (SD:531)	2,021 (SD:593)	2,204 (SD:313)	2,389 (SD:325)	1,988 (SD:154)
Novel transcripts	5,762 (SD:796)	5,304 (SD: 913)	5,421 (SD:778)	5,893 (SD:708)	5,514 (SD:768)	5,442 (SD:415)	6,032 (SD:237)
Not In Category (NIC)	3,765 (SD:599)	3,490(SD:651)	3,478 (SD:514)	3,611 (SD:459)	3,469 (SD:528)	3,204 (SD:318)	3,582 (SD:165)
Novel Not in Category (NNC)	1,811 (SD:202)	1,647 (SD: 271)	1,768 (SD:256)	2,066 (SD:254)	1,839 (SD:234)	2,015 (SD:120)	2,248 (SD:59)
Fusion (F)	74 (SD:11)	67 (SD:15)	78 (SD:11)	90 (SD:18)	93(SD:17)	65 (SD:7)	84 (SD:8)
Genic (G)	22 (SD:5)	19 (SD:4)	18 (SD:5)	22 (SD:4)	25 (SD:4)	24 (SD:1)	24 (SD:3)
Antisense (A)	35 (SD:8)	30 (SD:10)	33 (SD:12)	32 (SD:6)	30 (SD:10)	49 (SD:12)	31 (SD:3)
Intergenic (I)	54 (SD:7)	51 (SD:13)	44 (SD:14)	72 (SD:10)	59 (SD:15)	84 (SD:6)	62 (SD:10)

229 *A high number of novel transcripts absent in the reference annotation*

230 To further characterize the features of the 117,728 high-confidence distinct transcripts, we compared them
231 with those annotated in the GRCm39/mm39 reference genome from Ensembl v103 (Howe et al. 2021),
232 built largely based on a single C57BL/6 lab mouse inbred strain, which has predominantly a genomic *M.*
233 *m. domesticus* population background. On the basis of their alignment status to the Ensembl mouse
234 transcriptome v103 (Howe et al. 2021), these transcripts were classified into eight distinct structural
235 categories using SQANTI3 (Tardaguila et al. 2018) (Figure 3A): i) full splice match (FSM); ii) incomplete
236 splice match (ISM); iii) novel in category (NIC); iv) novel not in category (NNC); v) Fusion (F); vi) genic
237 (G); vii) antisense (A); viii) intergenic(I). The statistics on the number of transcripts from each structural
238 category at the population level can be found in Table 1.

239 In total, we found only 51% of the 117,728 distinct transcripts matching perfectly to a whole (FSM, 34.9%)
240 or subsection (ISM, 16.1%) of a reference annotated transcript, designated as known transcripts following
241 the convention in (Leung et al. 2021; Naftaly et al. 2021). The remaining 49% (57,730) of the identified
242 transcripts were missing in the Ensembl transcriptome (Figure 3B), including the ones deriving from the
243 rearrangement and modification of annotated exons (novel combination of known splice sites or NIC:
244 23.8%; novel splice sites or NNC: 22.5%; and merging of adjacent transcripts or Fusion: 1.0%), from
245 intronic regions (Genic: 0.2%), and from entirely novel gene loci (Antisense: 0.5%; Intergenic: 0.9%). Of
246 note, the novel transcripts exhibit comparable high quality and reliability as the known transcripts, based
247 on the confirmation of the transcription start site, splice junction, and transcription termination site
248 (Supplementary Figure S4C-E).

249 One may argue that the transcripts specific to outgroup species could contribute substantially to the novel
250 transcript category, as the GRCm39/mm39 reference genome was built largely based on a single C57BL/6
251 lab inbred strain. However, this is not the case, as we still found around 43% of the identified transcripts
252 were missing in the reference annotation after excluding those transcripts detected only in the individuals
253 of outgroup species (Supplementary Figure S8B).

254 Given that the current reference annotation is based on an isogenic strain, it represents nominally only the
255 isoforms of a single individual, but not the isoforms typical for a population or across populations. At the
256 same time, it includes isoforms that are restricted to its population of origin. We have therefore compiled a
257 set of 39,537 isoforms that are conserved in the house mouse and the outgroup species (Supplementary
258 Dataset S3). A substantial fraction (30.6%) of these isoforms is not included in the reference annotation
259 (Supplementary Figure S8C). We propose that this conserved set of isoforms should be the best reference
260 when comparing between mice and other species, *e.g.*, humans.

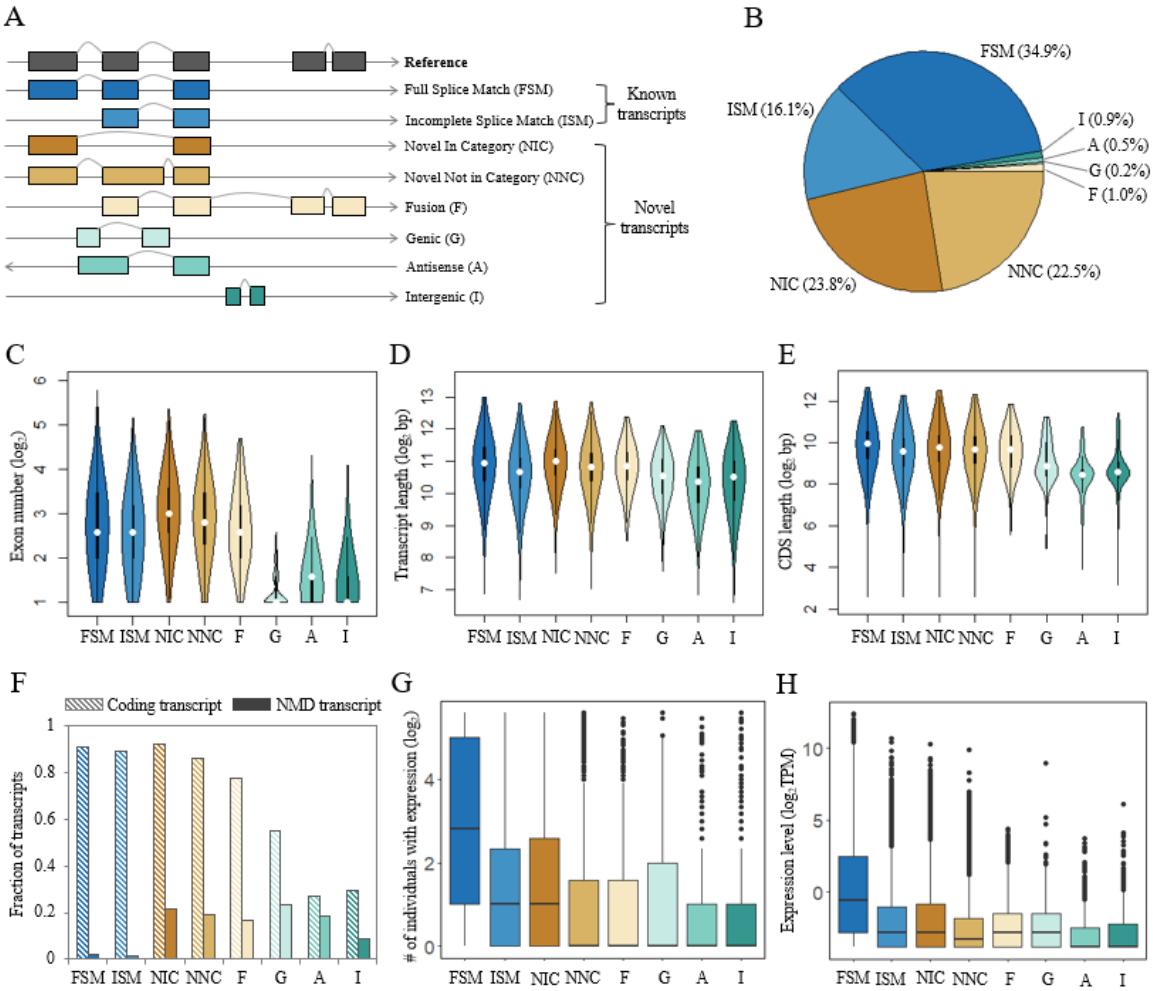


Figure 3 Characterization of all detected isoforms. (A) Types and illustrations of identified isoforms. (B) Fraction distribution of isoform structural categories. The detailed results across populations can be found in Supplementary Figure S9. (C)-(H) show the distributions of isoform types with respect 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 (Tardaguila et al. 2018). 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), novel transcripts deriving from annotated exonic regions (NIC, NNC, and F - which constitute the bulk of the new transcripts) show comparable exon numbers (Figure 3C), transcript length (Figure 3D), CDS length (Figure 3E). The novel transcripts from intronic (G) and unannotated gene loci (A, I) show generally lower values for all these features (Supplementary Table S2). Notably, all the novel transcripts with coding potential have a significantly higher probability to

276 become degraded via the nonsense-mediated decay (NMD) process (Chang et al. 2007) due to premature
277 translation-termination codons (PTCs - detected by SQANTI3) than known transcripts (Figure 3F, mean
278 0.18 vs. 0.02, Fisher's exact test, p-value < 2.2 x 10^{-16}). Most novel transcripts were found to be more
279 restrictively expressed in a smaller number of individuals, with the exception of NIC, which are
280 comparable to ISM in this respect (Figure 3G). The general expression levels of the novel transcripts,
281 except A and I, are at a similar level as the known ISM transcripts (Figure 3H). In terms of absolute
282 transcript abundance, the newly discovered isoforms contribute about 6% to the total number of expressed
283 transcripts in the overall dataset, in contrast to 49% of identified unique transcripts as novel ones (Figure
284 3B).

285

286 *Housekeeping genes as the major contributor to isoform diversity*

287 To analyze the characteristics of the gene loci that contribute significantly to the overall isoform diversity,
288 we classified all the detected gene loci into three roughly similar-sized classes on the basis of their
289 splicing isoform numbers (Figure 4A): i) lightly spliced (LS, with no more than two isoforms, n = 5,372),
290 ii) moderately spliced (MS, with isoform number between three and eight, n = 5,012), and iii) heavily
291 spliced (HS, with no less than nine isoforms, n = 4,628). Among all the 117,728 isoforms detected in the
292 wild mice brain transcriptome, 4.3% are derived from LS genes, 18.2% from MS genes, and the largest
293 fraction (77.4%) from HS genes.

294 Based on the gene expression analysis of matched Illumina RNA-Seq dataset, we found that genes with
295 more splicing isoforms have a significantly higher expression level in the brain (Figure 4B, MS vs. LS: p-
296 value = 8.7 x 10^{-294}, HS vs. MS: p-value = 3.2 x 10^{-82}, two-sided Wilcoxon rank sum tests)**.** Additionally,
297 through analyzing the transcriptomic sequencing dataset from thirteen tissues of the *M. m. domesticus*
298 derived C57BL/6 inbred line (Lin et al. 2014), we observed a clear trend that genes with more splicing
299 isoforms tend to be expressed significantly more broadly, *i.e.*, with expression in more tissues (Figure 4C,
300 MS vs. LS: p-value = 8.2 x 10^{-129}, HS vs. MS: p-value = 4.2 x 10^{-17}, two-sided Wilcoxon rank sum tests).
301 Furthermore, gene ontology (GO) enrichment analysis on the HS gene group showed that they are
302 involved in a variety of biological processes for the maintenance of cellular function (Supplementary
303 Dataset S4), a common feature of housekeeping genes (Eisenberg and Levanon 2003). These findings
304 suggest that housekeeping genes contribute most to the overall isoform diversity.

305 Previous studies have shown that evolutionarily old genes usually have a broader tissue expression and
306 higher expression levels (Zhang et al. 2015), thus it can be hypothesized that genes with more splicing

isoforms tend to be evolutionarily older. We set out to test this hypothesis via a comparative analysis of the gene age distribution pattern among the gene groups with different numbers of splicing isoforms, for which the mouse gene age assignment data was retrieved from (Neme and Tautz 2013). Our analysis indeed showed that there is a significant enrichment in evolutionarily older gene groups for those genes with more splicing isoforms (Figure 4D, MS vs. LS: p-value = 4.1×10^{-15}, HS vs. MS: p-value = 1.5×10^{-7}, two-sided Wilcoxon rank sum tests).

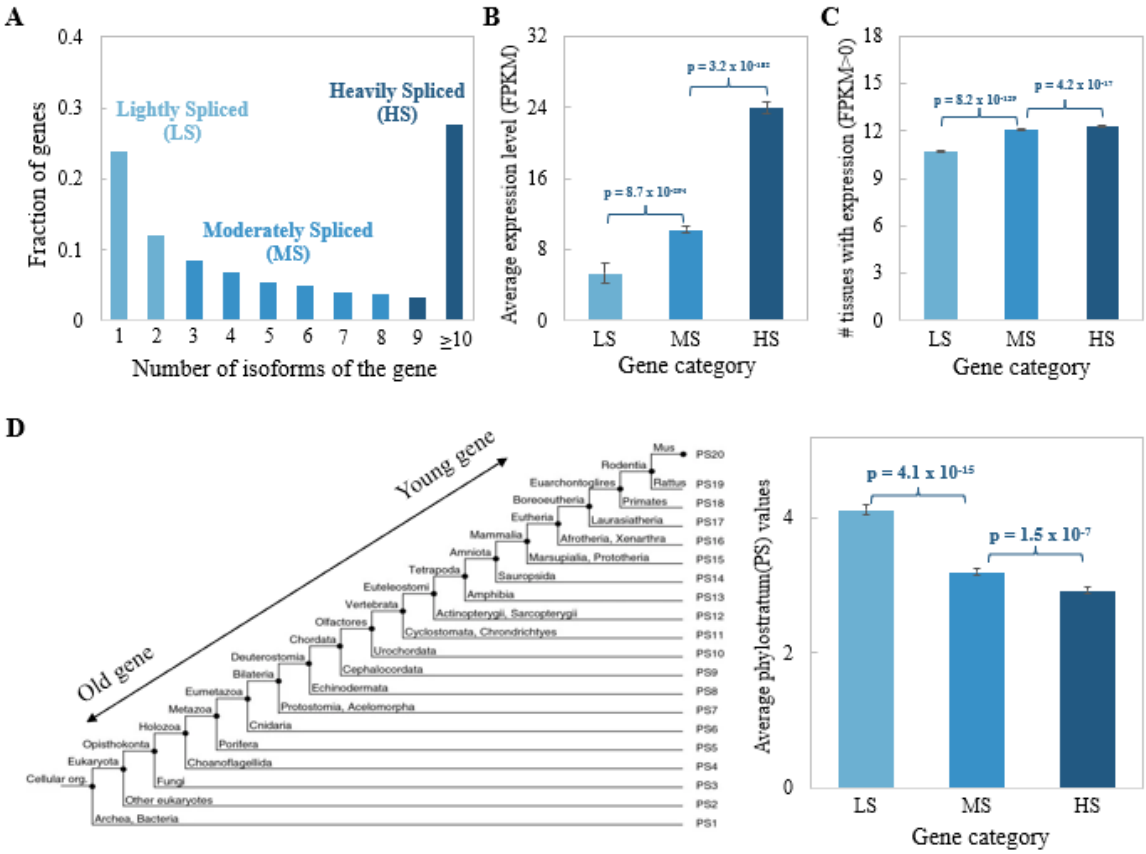


Figure 4 Characteristics of gene loci with different numbers of isoforms. (A) Gene categories were based on the number of isoforms of each gene. Three gene categories were classified with roughly the same number of genes in each group. (B) Comparison of the average gene expression level among the three gene categories. The average gene expression level was calculated based on the Illumina RNA-Seq generated in this study. (C) Comparison of the number of tissues with expression among the three gene categories. The transcriptomic sequencing data from 13 tissues of the *Mus musculus* C57BL/6 inbred line was retrieved from (Lin et al. 2014). The detectable gene expression in each tissue was defined as FPKM > 0. (D) Comparison of the average phylostratum (PS) values among the three gene categories. The PS assignment data for mouse protein-coding genes were retrieved from (Neme and Tautz 2013). The smaller PS values indicate the older gene groups. Error bars show standard errors of the mean (SEM) values of average gene expression levels, number of tissues with expression, and average PS values, respectively. Statistical p-values were computed using two-sided Wilcoxon rank sum tests.

326 *Local alternative splicing events contribute to isoform diversity*

327 The full-length isoforms appear as a combination of different types of local alternative splicing (AS)
328 events, and thus it is useful to disentangle the relative contribution of each type of AS event to the overall
329 isoform diversity. For this purpose, we used the SUPPA2 program (Trincado et al. 2018) to identify
330 different types of splicing events in the full set of high-confidence transcripts, including skipped exon
331 (SE), retained intron (RI), alternative 5' splice site (A5), alternative 3' splice site (A3), mutually exclusive
332 exon (MX), alternative first exon (AFE), and alternative last exon (ALE) (Figure 5A).

333 Among all the 94,581 splicing events detected in all the transcripts (Figure 5B), AFE events contribute
334 most to the overall isoform diversity (31.7%), followed by SE events (19.8%), RI events (14.1%), A5
335 (11.6%), A3 (10.1%), ALE (9.4%), and with MX as the least contributor (3.3%). This finding is in line
336 with the previous report on the AFE as the most prevalent splicing event for the overall transcriptome in
337 the inbred laboratory mouse cerebral cortex (Leung et al. 2021), suggesting the dominant role of using
338 AFE to generate alternative isoforms in mice brain transcriptomes. The "alternative first exon" transcripts
339 originate evidently from new promotors, suggesting that these can easily develop upstream of existing
340 genes. This is in line with the realization that enhancers as regulatory elements can also assume promotor
341 functions (Andersson and Sandelin 2020).

342 AFEs would not necessarily impact the coding sequences (Landry et al. 2003). To address this issue, we
343 performed an additional analysis by collapsing the transcripts with the same coding sequence (*i.e.*, only
344 transcripts with predicted ORFs were considered) into a single unique ORF. We indeed observed a much
345 lower number (less than two-thirds) of unique ORFs than transcripts in each mouse individual
346 (Supplementary Figure S10). Based on the landing position of the start codon in relation to local AS
347 events (Figure 5C), we further enumerated the number of local AS events causing the change of the
348 respective coding sequences. We found that only 46.5% (43,978) of all the local splicing events had an
349 impact on the ORF structures (Figure 5D). Compared to AFE, SE and RI events are more prevalent among
350 all the local AS types to contribute to ORF diversity. The majority of AFE events (55%) would cause no
351 change in the coding sequences, while acting as a major source to generate isoform diversity due to the
352 emergence of alternative promoters (Huang et al. 2021). These data illustrate the distinct roles of local AS
353 events in contributing to isoform and ORF diversity in wild mice brain transcriptomes.

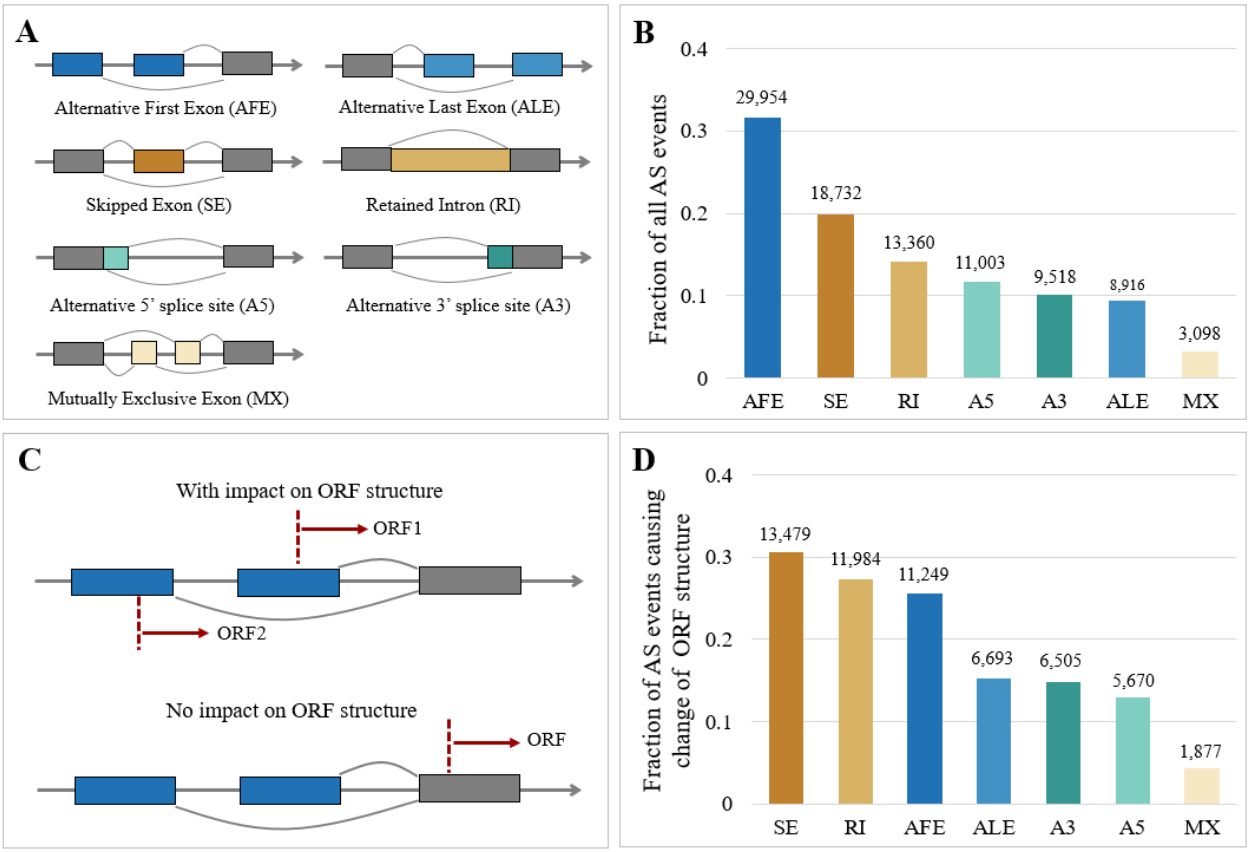


Figure 5 Distribution of different types of local AS events. (A) Types and illustration of AS local events. (B) The distribution of all types of local AS events. (C) An example of AFE events that change ORF structures, and similar situations for other types of local AS events. The dashed red lines indicate 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 types of local AS events.

Discussion

We used PacBio Iso-Seq and Illumina RNA-Seq to characterize the full-length isoform diversity of the brain transcriptomes in house mouse natural populations, resulting in the first and most comprehensive full-length isoform category representation at a comparative population level to date. Via implementation of a separate 5'cap selection step (Cartolano et al. 2016; Kuo et al. 2017), our optimized approach improved the performance to enrich genuine full-length transcripts.

Our overall results confirm the conclusions from previous studies that the diversity of alternatively spliced transcripts surpasses the current annotation level, even of exceptionally well-curated genomes, such as the one from the mouse (Leung et al. 2021). Our study finds double as many transcripts as are currently annotated, suggesting the widespread diversity of alternative isoforms in natural populations. Intriguingly,

these isoforms could be efficient markers to infer population relationships, and to resolve the phylogenetic tree of house mouse natural populations and outgroups. This suggests that the isoform diversity reflects directly the genetic polymorphisms segregating in these populations. Note that the mice used for this study were reared under tightly controlled laboratory conditions to cut out possible environmental influences on isoform diversity.

Interestingly, evolutionarily old housekeeping genes with overall high expression in many tissues show the largest number of new splice variants on average, likely due to the general longer gene length and larger number of exons for the evolutionarily old housekeeping genes (Zhang et al. 2015). Indeed, we found that heavily spliced genes are significantly longer in length (indicated by the longest transcript length) and have significantly more exons (Supplementary Figure S11), potentially allowing more space for alternative splicing machinery to act on.

The focus on full-length transcripts revealed that new 5'-exons from alternative promotors are the relatively most abundant source of new transcripts, followed by the "skipped exon" and "retained intron" variants, which were previously thought to be the most prevalent ones (Grau-Bove et al. 2018). The "alternative first exon" transcripts originate from new promotors that can easily develop upstream of existing genes. This is in line with the realization that enhancers as regulatory elements can also assume promotor functions (Andersson and Sandelin 2020). Only about half of the new splice variants affect the annotated ORF of the respective genes, implying that the ORF diversity generated through alternative splicing is lower than the one at the transcript level. However, new short ORFs or alternative ORFs that may arise in the alternatively spliced transcripts are yet to study (Leong et al. 2022).

Our dataset allows to define isoforms that should be of particular interest for functional comparisons across larger evolutionary distances, including humans. The currently annotated mouse genome includes also the isoforms that are population-specific for the subspecies from which the C57Bl/6 was derived. Most of these will have specifically arisen in this lineage and are therefore of less interest for broader functional comparisons. With our comparative data, however, we can define the set of isoforms that are typically found across multiple mouse species, including new ones that were so far missing from the annotation. We have compiled this more informative and more inclusive set of isoforms (Supplementary Dataset S3), which will be of particular use for broader comparative studies between mice and other species.

402 **Methods**

403 *Sample collection and RNA extraction*

404 A total of 49 male mice individuals were sampled in this study (Supplementary Dataset S1), including one
405 individual being used in the Iso-Seq library preparation protocol optimization experiment. Eight
406 individuals were chosen for each of the following populations covering all three major subspecies of the
407 house mouse (*Mus musculus*): Germany, France, and Iran populations from *Mus mus domesticus*,
408 Kazakhstan population from *Mus mus musculus*, and Taiwan population from *Mus mus castaneus*. We
409 also included four individuals from each of the two outgroup species, *Mus spretus* and *Mus spicilegus*. All
410 mice were derived from previously wild-caught founder mice, and were maintained in an outbred scheme
411 under standard rearing techniques, which ensure a homogeneous environment (Harr et al. 2016). Each
412 mouse individual was obtained from one distinct, unrelated breeding family, thus the relatedness between
413 individuals, especially from the same population, should be minimized (see Supplementary Dataset S2 for
414 pair-wise relatedness estimates).

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

422

423 *Selection of cDNA library enrichment protocol for PacBio Iso-Seq*

424 It has been reported that the TeloPrime full-length cDNA Amplification kit, which selectively synthesizes
425 cDNA molecules from mRNAs carrying a 5' cap in comparison to the standard PacBio cDNA library
426 preparation protocol, provides a better solution to enrich for actual full-length transcripts (Cartolano et al.
427 2016). In this study, we selected one male mouse from the Taiwan population (Supplementary Dataset
428 S1A) to test the performance of TeloPrime-based protocols in the real world.

429 We applied three distinct cDNA library enrichment protocols for the RNA sample extracted from the same
430 mice individual: i) standard PacBio Clontech SMARTer PCR cDNA Synthesis kit (Clontech Laboratories,
431 Inc.); ii) TeloPrime Full-Length cDNA Amplification Kit v1 (Lexogen GmbH), with the combination of

432 an alternative oligo-dT primer from the SMARTer PCR cDNA Synthesis kit; iii) TeloPrime Full-Length
433 cDNA Amplification Kit v2 (Lexogen GmbH), with the combination of an alternative oligo-dT primer
434 from the SMARTer PCR cDNA Synthesis kit. The former protocol selectively synthesizes cDNA
435 molecules from transcripts with a poly(A) tail. In contrast, the latter two protocols synthesize cDNA
436 molecules from transcripts with both a 5' cap and a poly(A) tail. TeloPrime-based kit V2 includes updated
437 reaction conditions in comparison with kit V1, while the method of cDNA generation remains unchanged.
438 For each protocol, we used 1 ug homogeneous RNA as input to generate the cDNA library. The cDNAs
439 were not size-selected, and PacBio libraries were prepared with the SMRTbell Template Prep Kit 1.0
440 (Pacific Biosciences) and sequenced on the PacBio Sequel I with Sequel DNA polymerase and binding kit
441 3.0 and sequencing chemistry version 3.0 for 1200 min. Each cDNA library was sequenced on one SMRT
442 cell of the same sequencing run.

443 We analyzed the sub-reads for each SMRT cell separately following the IsoSeq3 pipeline (v3.4.0;
444 https://github.com/PacificBiosciences/IsoSeq), and aligned the resulting isoforms to the mouse genome
445 (GRCm39/mm39) using minimap2 (v2.24-r1122) (Li 2018), with the computational protocol indicated in
446 the below text. The alignment results from the above three cDNA libraries were compared to the
447 coordinates of transcripts annotated in Ensembl (v103) using gffcompare (v0.12.2) (Pertea and Pertea
448 2020).

449

450 *PacBio Iso-Seq and Illumina RNA-Seq sequencing*

451 The initial experimental tests showed that our newly developed TeloPrime-based cDNA library
452 preparation protocol could provide a better solution to enrich for actual full-length transcripts, compared
453 to the standard PacBio cDNA library preparation protocol (Figure 1). Hence, the TeloPrime Full-Length
454 cDNA Amplification Kit v2 (Lexogen GmbH) was utilized to construct PacBio IsoSeq cDNA libraries.
455 One µg total RNA from each individual was used as input, and double-strand cDNA was produced by
456 following the manufacturer's instructions, except that an alternative oligo-dT primer from the SMARTer
457 PCR cDNA Synthesis kit (Clontech Laboratories, Inc.), which also included a random 10mer sequence as
458 a unique molecule identifier (UMI) after each sequence. The cDNAs were not size-selected, and PacBio
459 libraries were prepared with the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences) and sequenced on
460 the PacBio Sequel I with Sequel DNA polymerase and binding kit 3.0 and sequencing chemistry version
461 3.0 for 1200 min. Each library was sequenced on three SMRT cells to achieve sufficient coverage.

462 Poly(A) RNA from each sample was enriched from 1 µg total RNA by the NEBNext® Poly(A) mRNA
463 Magnetic Isolation Module (Catalog #: E7490, New England Biolabs Inc.). RNA-Seq libraries were
464 prepared using NEBNext Ultra™ II Directional RNA Library Prep Kit for Illumina (Catalog #: E7760,
465 New England Biolabs Inc.), according to manufacturer's instructions. A total of eleven PCR cycles were
466 applied to enrich library concentration. Sequencing-by-synthesis was done at the HiSeq3000 system in
467 paired-end mode 2 x 150bp. Raw sequencing outputs were converted to fastq files with bcl2fastq
468 (v2.17.1.14).

469

470 *RNA-Seq read QC and data processing*

471 We trimmed and filtered the low-quality raw fastq reads for each sample separately using the fastp
472 program (v0.20.0; --cut_front --average_qual 20 --length_required 50) (Chen et al. 2018), and only
473 included the paired-end reads with a minimum length of 50bp and average quality score of 20 for further
474 analysis.

475 The filtered fastq reads were aligned to mouse GRCm39/mm39 reference genome sequence with STAR
476 aligner v2.7.0e (Dobin et al. 2013), taking the mouse gene annotation in Ensembl v103 (Howe et al. 2021)
477 into account at the stage of building the genome index (--runMode genomeGenerate --sjdbOverhang 149).
478 The STAR mapping procedure was performed in two-pass mode, and some of the filtering parameters
479 were tweaked (personal communication with STAR developer; --runMode alignReads –twopassMode
480 Basic --outFilterMismatchNmax 30 --scoreDelOpen -1 --scoreDelBase -1 --scoreInsOpen -1 --
481 scoreInsBase -1 --seedSearchStartLmax 25 --winAnchorMultimapNmax 100), in order to compensate the
482 sequence divergences of individuals from various populations and species (Zhang et al. 2021). With this
483 optimized mapping pipeline, a similar alignment rate was reached for all the samples (Supplementary
484 Dataset S1C). The alignment bam files were taken for further analysis.

485

486 SNP variants calling and individual relatedness analysis

487 We followed the general GATK version 4 Best Practices to call genetic variants from Illumina RNA-seq
488 data. We first sorted the above alignment bam data using samtools v1.9 (Li et al. 2009), and marked
489 duplicates by using PICARD v2.8.0 (http://broadinstitute.github.io/picard). Reads with N in the cigar were
490 split into multiple supplementary alignments and hard clips mismatching overhangs using the
491 SplitNCigarReads function in GATK v4.1.9. By using BaseRecalibrator and ApplyBQSR functions in
492 GATK, we further recalibrated base quality scores with SNP variants that were called with the genomic

493 sequencing dataset of the mice individuals from the same populations (Zhang et al. 2021) to get analysis-
494 ready reads. Following, we called raw genetic variants for each individual using the HaplotypeCaller
495 function in GATK, and jointly genotyped genetic variants for all the individuals using the
496 GenotypeGVCFs function. We only retained genetic variants that passed the hard filter "QD < 2.0 || FS >
497 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 3.0" for further analysis
498 (Zhang et al. 2021).

499 We used the relatedness2 option of VCFtools v0.1.14 (Danecek et al. 2011) to assess pairwise individual
500 relatedness for all possible pairs of mice individuals, based on only bi-allelic SNP variants with no more
501 than 20% missing data and locating within complete linear chromosomes, thinned to 1 SNP every 1 Mb
502 (Harr et al. 2016). We defined that relatedness category based on the expected ranges of kinship
503 coefficients ('Phi'), according to the main reference of KING method (Manichaikul et al. 2010). The
504 detailed relatedness information of all possible individual pairs can be found in Supplementary Dataset S2.

505

506 *Iso-Seq read QC and data processing*

507 We analyzed the raw sub-reads for each SMRT cell separately following the IsoSeq3 pipeline (v3.4.0).
508 Circular consensus sequences (CCS) were generated from sub-reads using the CCS module in polish
509 mode (v6.0.0; --minPasses 3 --minLength 50 --maxLength 1000000 --minPredictedAccuracy 0.99) of the
510 IsoSeq3 pipeline, and the CCS reads generated in three SMRT cells for the same sample were merged. We
511 trimmed cDNA primers (5' TGGATTGATATGTAATACGACTCACTATAG; 3'
512 GTACTCTGCGTTGATACCACTGCTT) and orientated the CCS reads using the lima program with the
513 specialized IsoSeq mode (v2.0.0; --isoseq). The 10mer UMI following each CCS read was tagged and
514 removed using the tag module of the IsoSeq3 pipeline (--design T-10U). Following this, we identified the
515 processed CCS reads as full-length and non-chimeric (FLNC), based on the presence of a ploy(A) tail and
516 absence of concatemer using the refine module of the IsoSeq3 pipeline (--require-polya). We further
517 performed PCR deduplication based on the UMI tag information using the dedup module of the IsoSeq3
518 pipeline (default parameters). After this deduplication step, only one consensus FLNC sequence per
519 founder molecule in the sample was kept. We then performed *de novo* clustering of the above reads using
520 the cluster module of the IsoSeq3 pipeline, and kept only the high-confidence isoforms supported by at
521 least 2 FLNC reads for further analysis.

522 We aligned these isoforms of each sample to the GRCm39/mm39 reference genome sequence using the
523 minimap2 (v2.24-r1122; -ax splice:hq -uf --secondary=no -C5 -O6,24 -B4) (Li 2018), with parameter

524 setting following the best practice of Cupcake pipeline (Gordon et al. 2015). The alignment bam files were
525 further sorted using the samtools program (v1.9) (Li et al. 2009). We collapsed redundant transcript
526 models for each sample based on the above sorted alignment coordinate information using the collapse
527 module of the TAMA program (-d merge_dup -x capped -m 5 -a 1000 -z 30 -sj sj_priority) (Kuo et al.
528 2020). The rationales for defining redundant transcript models are shown in Supplementary Text of
529 Methods. Finally, isoforms across all 48 samples were merged into a single non-redundant transcriptome
530 using the merge module of the TAMA program (-d merge_dup -m 5 -a 1000 -z 30) (Kuo et al. 2020), with
531 the same parameter setting as the above collapse step.

532

533 *Iso-Seq transcriptome classification and filtering*

534 We performed the quality control analysis for the above merged non-redundant PacBio Iso-Seq
535 transcriptome using SQANTI3 v4.2 (Tardaguila et al. 2018), with the input datasets of Ensembl v103
536 (Howe et al. 2021) GRCm39/mm39 reference genome and gene annotation, FLNC read counts, Isoform
537 expression levels and STAR output alignment bam files and splice junction files from RNA-Seq short
538 reads, mouse transcription start sites (TSS) collected in refTSS database v3.1 (Abugessaisa et al. 2019),
539 and curated set of poly(A) sites and poly(A) motifs in PolyASite portal v2.0 (Herrmann et al. 2020).

540 We filtered out isoforms of potential artifacts mainly by following (Tardaguila et al. 2018). Mono-exonic
541 transcripts were excluded, as they tend more likely to be experimental or technical artifacts (Tardaguila et
542 al. 2018). Isoforms with unreliable 3'end because of a possible intrapriming event (intrapriming rate
543 above 0.6) were also removed from the dataset. We kept the remaining isoforms that met both of the
544 following criteria: 1) no junction is not labeled as RT-Switching; 2) all junctions are either canonical
545 (GT/AG; GC/AG; AT/AC) or supported by at least 3 spanning reads based on STAR junction output file.
546 All isoforms that passed the above filters were taken for further analysis.

547 Given their matching status to the Ensembl mouse transcriptome v103 (Howe et al. 2021), the above
548 transcripts were classified into eight distinct categories using SQANTI3 (Tardaguila et al. 2018): i) Full
549 Splice Match (FSM, matching perfectly to a known transcript); ii) Incomplete Splice Match (ISM,
550 matching to a subsection of a known transcript); iii) Novel In Category (NIC, with known splice sites but
551 novel splice junctions); iv) Novel Not in Category (NNC, with at least one unannotated splice site); v)
552 Fusion (F, fusion of adjacent transcripts); vi) Genic (G, overlapping with intron); vii) Antisense (A, on the
553 antisense strand of an annotated gene); viii) Intergenic (I, within the intergenic region). The transcripts

554 matching perfectly to a whole (FSM) or subsection (ISM) of reference annotated transcripts are designated
555 known transcripts, and the others as novel transcripts.

556 We evaluated the reliability of isoforms from each category, on the basis of reference annotation and
557 empirical information from three distinct aspects separately (Supplementary Text of Methods): i)
558 Transcription start site (TSS); ii) Transcription Termination Site (TTS); iii) splice junction (SJ). In
559 comparison to the well-support FSM transcripts, the transcripts from other categories show reduced
560 confidence levels in terms of TSS (Supplementary Figure S4B), but no reduction for TTS and SJ
561 (Supplementary Figure S4D and 4E). This might hint at a failure to capture accurate TSS for those
562 transcripts (Tardaguila et al. 2018). For instance, it is still possible that some ISM transcripts came from
563 partial fragments, due to the imperfect targeting of 5'-cap or the degradation of transcripts in the later
564 steps. To address this concern, we further excluded the non-FSM transcripts without support for TSSs
565 (Supplementary Figure S4C).

566

567 *Association between transcript pattern and population divergence*

568 We explored the association between individual transcript presence patterns and population divergence,
569 based on three lines of analyses. Firstly, we calculated and compared the fraction of shared transcripts
570 between all possible pairwise combinations from different species, from different subspecies but the same
571 species (n = 448), from different populations but the same subspecies, and from the same populations.
572 Secondly, we performed PCA on the individual transcript landscape using the R package ggfortify v0.4.16
573 (https://cran.r-project.org/web/packages/ggfortify/index.html). Additional PCAs were conducted for the
574 subspecies *M. m. domesticus*, given that more closely-related populations were assayed in this study.

575 Lastly, we built a phylogenetic tree for the seven assayed populations using the R package "ape" v5.7-1
576 (Paradis and Schliep 2019), based on the presence/absence matrix of fixed transcript in each population.
577 Euclidean distance was used as the distance measure between each pair of populations, and the neighbor-
578 joining tree estimation function was used to build the phylogenetic relations. The boot.phylo function
579 implemented in the same R package was used to perform 1,000 bootstrap replications, and population split
580 nodes of high confidence were taken as the ones with at least 70% bootstrap support values. For
581 comparison, we constructed another phylogenetic tree with the SNP variants that were called from the
582 Illumina RNA-Seq dataset generated in this study (shown in the above text), for which the same set of
583 mice individuals was used. To reduce computation complexity, we performed LD pruning on the SNP
584 data set by using PLINK v1.90b4.6 (Purcell et al. 2007), removing one of a pair of SNPs with LD $\geqslant$0.2 in

585 sliding window of 500 SNPs and step wise of 100 SNPs. The same procedure was followed to build the
586 phylogenetic tree, as it was done for the analysis of transcripts.

587

588 *Detection of population differentiation transcript markers*

589 We calculated population differentiation index (PDI) values for each transcript using a similar approach to
590 analyze population differentiating retroCNVs (Zhang and Tautz 2022). A larger PDI value generally
591 indicates greater power to use the presence status of the respective transcript to distinguish populations.
592 The computation of the PDI value for each transcript follows:

593 $$\mathrm{PDI} = \frac{\mathrm{P}(1-\mathrm{P}) - \sum_{i=1}^{\mathrm{N}} \mathrm{R}_i * \mathrm{P}_i * (1-\mathrm{P}_i)}{\mathrm{P}(1-\mathrm{P})} \quad (1)$$

594 $$\mathrm{R}_i = \mathrm{N}_i \,/\, \sum_{i=1}^{\mathrm{N}} \mathrm{N}_i \quad (2)$$

595 In the first equation, N represents the number of assayed populations (*i.e.*, 7), P_i is the frequency of the
596 focal transcript in the *i*th population, and P is the overall frequency in all the assayed populations. The
597 definition of R_i is shown in the second equation, and is computed as the number of individuals of the *i*th
598 population (N_i) divided by the total number of individuals from all populations. We tested the significance
599 of the PDI value for each transcript with a randomization test in which individual labels were randomly
600 shuffled 1,000 times, while population sizes remained the same. We calculated the p-value by comparing
601 the observed PDI for each transcript against the PDI values from the randomized null distribution.
602 Transcripts with significant PDI values after multiple testing corrections (FDR < 0.05) were considered
603 population differentiation transcript markers.

604

605 *Feature characterization of all the transcripts*

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

610 In the SQANTI3 pipeline (Tardaguila et al. 2018), the potential coding capacity and ORFs from the
611 transcript sequences were predicted using GeneMarkS-T (GMST) algorithm (Tang et al. 2015). An NMD
612 transcript is designated if there's a predicted ORF, and the CDS ends at least 50bp before the last junction

613 for the respective transcript. The expression level of each transcript for each sample was computed, on the
614 basis of the number of supported FLNC reads, and normalized in the unit of TPM (transcript per million).
615 Transcripts with expression in respective tissues were defined as the ones with non-zero TPM values. In
616 addition, we quantified the expression of each transcript using another program named Kallisto v0.46.2
617 (Bray et al. 2016), for which the expression quantification was based on the alignment of Illumina RNA-
618 Seq data to the merged isoform dataset derived from PacBio Iso-Seq data, as shown in the above text.

619

620 *Feature characterization of gene loci with different number of isoforms*

621 We classified all the detected gene loci into three roughly similar-sized classes, based on their splicing
622 isoform numbers: i) lightly spliced (LS, with no more than two isoforms, n = 5,372), ii) moderately
623 spliced (MS, with isoform number between three and eight, n = 5,012), and iii) heavily spliced (HS, with
624 no less than nine isoforms, n = 4,628). We characterized the features of the gene loci with different
625 numbers of isoforms from three aspects, *i.e.*, gene expression pattern, gene age distribution pattern, and
626 gene ontology (GO) functional term enrichment.

627 Firstly, we exploited two Illumina RNA-seq datasets for the gene expression pattern analysis: i) the one
628 generated in this study for the whole brain of 48 mice individuals; ii) a published dataset from thirteen
629 tissues (adipose tissue, adrenal gland, brain, heart, kidney, liver, lung, pancreas, sigmoid colon, small
630 intestine, spleen, ovary, and testis) of the *M .m. domesticus* C57BL/6 inbred line (Lin et al. 2014). The raw
631 fastq data for both datasets were trimmed, filtered, and aligned to the mouse GRCm39/mm39 reference
632 genome, as shown in the above text. We counted the fragments uniquely mapped to the annotated genes in
633 Ensembl v103 using featureCounts v1.6.3 (Liao et al. 2014) in reversely stranded mode, excluding multi-
634 mapping reads and those with alignment quality below 5 (-s 2 –Q 5). Gene expression levels were
635 calculated and transformed into a log2 scale of FPKM (Fragments Per Kilobase of transcript per Million
636 mapped fragments), on the basis of gene's transcript length and the number of mapped fragments in the
637 data set from each sample. For the first dataset, the average gene expression level was computed as the
638 mean value in all the assayed 48 samples. For the second dataset, the number of tissues with expression
639 for each gene was defined as the count of tissues with non-zero FPKM values. The statistical significances
640 of the expression patterns between pair-wise groups of gene sets were computed using two-sided
641 Wilcoxon rank sum tests.

642 Secondly, we analyzed and compared the evolutionary gene age pattern among three gene groups, taking
643 advantage of the mouse gene age assignment data from Neme *et al.* (Neme and Tautz 2013). In brief, each

644 mouse protein-coding gene was dated and given phylostratum (PS) assignment by inferring the absence
645 and presence of orthologs along the phylogenetic tree. The statistical significances of PS values between
646 pair-wise groups of gene set were computed using two-sided Wilcoxon rank sum tests.

647 Lastly, we performed GO enrichment analysis on the genes from the HS group using GOSeq v1.50.0
648 (Young et al. 2010), without correction of gene length differences. Only significant BP (biological process)
649 terms after multiple testing corrections (FDR < 0.05) were listed in this study.

650

651 *Analysis of local AS events*

652 We used the SUPPA2 program (Trincado et al. 2018) to identify local alternative splicing (AS) events in
653 the transcriptome. These local AS events are categorized into seven groups, including skipped exon (SE),
654 retained intron (RI), alternative 5' splice site (A5), alternative 3' splice site (A3), mutually exclusive exon
655 (MX), alternative first exon (AFE), and alternative last exon (ALE).

656 For the transcripts with predicted ORFs, we analyzed how the local AS events change the ORF structures.
657 In case the start codon position of the ORF lands in the region of a local AS event (Figure 5C), the focal
658 AS event is defined to cause a change in the respective coding sequence. On the other hand, in case the
659 start codon position of the ORF falls downstream of the local AS event, it does not influence the ORF
660 structure.

661

662 **Data access**

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

670

671 **Competing interest statement**

672 The authors declare that they have no conflicts of interests.

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.

Acknowledgements

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

References

Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, Kasukawa T. 2019. refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites. *J Mol Biol* **431**: 2407-2422.

Andersson R, Sandelin A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**: 71-87.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**: 1587-1593.

Bekpen C, Xie C, Tautz D. 2018. Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. *BMC Evol Biol* **18**: 121.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525-527.

Byrne A, Cole C, Volden R, Vollmers C. 2019. Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc Lond B Biol Sci* **374**: 20190097.

Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. 2016. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLoS One* **11**: e0157779.

Chang YF, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**: 51-74.

709 Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*
710 **34**: i884-i890.
711 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth
712 GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.
713 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013.
714 STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.
715 Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet* **19**: 362-365.
716 Ferrandez-Peral L, Zhan X, Alvarez-Estape M, Chiva C, Esteller-Cucala P, Garcia-Perez R, Julia E,
717 Lizano E, Fornas O, Sabido E et al. 2022. Transcriptome innovations in primates revealed by
718 single-molecule long-read sequencing. *Genome Res* doi:10.1101/gr.276395.121.
719 Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, Kang D, Underwood J, Grigoriev IV,
720 Figueroa M et al. 2015. Widespread Polycistronic Transcripts in Fungi Revealed by Single-
721 Molecule mRNA Sequencing. *PLoS One* **10**: e0132628.
722 Grau-Bove X, Ruiz-Trillo I, Irimia M. 2018. Origin of exon skipping-rich transcriptomes in animals
723 driven by evolution of gene architecture. *Genome Biol* **19**: 135.
724 Guenet JL, Bonhomme F. 2003. Wild mice: an ever-increasing contribution to a popular mammalian
725 model. *Trends Genet* **19**: 24-31.
726 Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Z, Babiker H, Linnenbrink M, Montero I,
727 Scavetta R et al. 2016. Genomic resources for wild populations of the house mouse, Mus
728 musculus and its close relative Mus spretus. *Sci Data* **3**: 160075.
729 Harr B, Turner LM. 2010. Genome-wide analysis of alternative splicing evolution among Mus subspecies.
730 *Mol Ecol* **19 Suppl 1**: 228-239.
731 Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2020. PolyASite 2.0: a
732 consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**: D174-
733 D179.
734 Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R,
735 Bhai J et al. 2021. Ensembl 2021. *Nucleic Acids Res* **49**: D884-D891.
736 Huang KK, Huang J, Wu JKL, Lee M, Tay ST, Kumar V, Ramnarayanan K, Padmanabhan N, Xu C, Tan
737 ALK et al. 2021. Long-read transcriptome sequencing reveals abundant promoter diversity in
738 distinct molecular subtypes of gastric cancer. *Genome Biol* **22**: 44.
739 Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition
740 and function. *Nat Rev Genet* **11**: 345-355.
741 Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating the dark
742 side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**: 751.
743 Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. 2017. Normalized long read RNA
744 sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**:
745 323.
746 Landry JR, Mager DL, Wilhelm BT. 2003. Complex controls: the role of alternative promoters in
747 mammalian genomes. *Trends Genet* **19**: 640-648.
748 Leong AZ, Lee PY, Mohtar MA, Syafruddin SE, Pung YF, Low TY. 2022. Short open reading frames
749 (sORFs) and microproteins: an update on their identification and validation measures. *J Biomed*
750 *Sci* **29**: 19.
751 Leung SK, Jeffries AR, Castanho I, Jordan BT, Moore K, Davies JP, Dempster EL, Bray NJ, O'Neill P,
752 Tseng E et al. 2021. Full-length transcript sequencing of human and mouse cerebral cortex
753 identifies widespread isoform diversity and alternative splicing. *Cell Rep* **37**: 110022.
754 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.
755 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome
756 Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools.
757 *Bioinformatics* **25**: 2078-2079.

758 Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning
759 sequence reads to genomic features. *Bioinformatics* **30**: 923-930.
760 Lin L, Shen S, Jiang P, Sato S, Davidson BL, Xing Y. 2010. Evolution of alternative splicing in primate
761 brain transcriptomes. *Hum Mol Genet* **19**: 2958-2973.
762 Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC et
763 al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc*
764 *Natl Acad Sci U S A* **111**: 17224-17229.
765 Ling Z, Brockmoller T, Baldwin IT, Xu S. 2019. Evolution of Alternative Splicing in Eudicots. *Front*
766 *Plant Sci* **10**: 707.
767 Malko DB, Makeev VJ, Mironov AA, Gelfand MS. 2006. Evolution of exon-intron structure and
768 alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res* **16**: 505-509.
769 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship
770 inference in genome-wide association studies. *Bioinformatics* **26**: 2867-2873.
771 Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R,
772 Hubbard T, Harrow J. 2011. The origins, evolution, and functional potential of alternative splicing
773 in vertebrates. *Mol Biol Evol* **28**: 2949-2959.
774 Murat F, Mbengue N, Winge SB, Trefzer T, Leushkin E, Sepp M, Cardoso-Moreira M, Schmidt J,
775 Schneider C, Mossinger K et al. 2023. The molecular evolution of spermatogenesis across
776 mammals. *Nature* **613**: 308-316.
777 Naftaly AS, Pau S, White MA. 2021. Long-read RNA sequencing reveals widespread sex-specific
778 alternative splicing in threespine stickleback fish. *Genome Res* **31**: 1486-1497.
779 Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de
780 novo evolution. *BMC Genomics* **14**: 117.
781 Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**:
782 457-463.
783 Parada GE, Munita R, Cerda CA, Gysling K. 2014. A comprehensive survey of non-canonical splice sites
784 in the human transcriptome. *Nucleic Acids Res* **42**: 10564-10578.
785 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses
786 in R. *Bioinformatics* **35**: 526-528.
787 Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**.
788 Pezer Z, Harr B, Teschke M, Babiker H, Tautz D. 2015. Divergence patterns of genic copy number
789 variation in natural populations of the house mouse (Mus musculus domesticus) reveal three
790 conserved genes with major population-specific expansions. *Genome Res* **25**: 1114-1124.
791 Phifer-Rixey M, Nachman MW. 2015. Insights into mammalian biology from the wild house mouse Mus
792 musculus. *Elife* **4**.
793 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI,
794 Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based
795 linkage analyses. *Am J Hum Genet* **81**: 559-575.
796 Rogers TF, Palmer DH, Wright AE. 2021. Sex-Specific Selection Drives the Evolution of Alternative
797 Splicing in Birds. *Mol Biol Evol* **38**: 519-530.
798 Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. 2012. Genome patterns of selection and
799 introgression of haplotypes in natural populations of the house mouse (Mus musculus). *PLoS*
800 *Genet* **8**: e1002891.
801 Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts.
802 *Nucleic Acids Res* **43**: e78.
803 Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M,
804 Macchietto M, Verheggen K et al. 2018. SQANTI: extensive characterization of long-read
805 transcript sequences for quality control in full-length transcriptome identification and
806 quantification. *Genome Res* doi:10.1101/gr.222976.117.

807 Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**:
808 18-30.
809 Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. 2018. SUPPA2: fast,
810 accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome*
811 *Biol* **19**: 40.
812 Verta JP, Jacobs A. 2022. The role of alternative splicing in adaptation and evolution. *Trends Ecol Evol* **37**:
813 299-308.
814 Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq:
815 accounting for selection bias. *Genome Biol* **11**: R14.
816 Zhang W, Landback P, Gschwend AR, Shen B, Long M. 2015. New genes drive the evolution of gene
817 interaction networks in the human and mouse genomes. *Genome Biol* **16**: 202.
818 Zhang W, Tautz D. 2022. Tracing the Origin and Evolutionary Fate of Recent Gene Retrocopies in
819 Natural Populations of the House Mouse. *Mol Biol Evol* **39**.
820 Zhang W, Xie C, Ullrich K, Zhang YE, Tautz D. 2021. The mutational load in natural populations is
821 significantly affected by high primary rates of retroposition. *Proc Natl Acad Sci U S A* **118**.

822