1 **Entire genome transcription across evolutionary time exposes non-coding DNA**

2 **to *de novo* gene emergence**

3

4   Rafik Neme and Diethard Tautz*

5   Max-Planck Institute for Evolutionary Biology, August-Thienemann-Strasse 2, 24306 Plön, Germany

6   * corresponding author: tautz@evolbio.mpg.de

7

8   **Even in the best studied Mammalian genomes, less than 5% of the total genome length is annotated as exonic.**

9   **However, deep sequencing analysis in humans has shown that around 40% of the genome may be covered by**

10   **poly-adenylated non-coding transcripts occurring at low levels[1]. Their functional significance is unclear[2,3], and**

11   **there has been a dispute whether they should be considered as noise of the transcriptional machinery[4,5]. We**

12   **propose that if such transcripts show some evolutionary stability they will serve as substrates for *de novo* gene**

13   **evolution, i.e. gene emergence out of non-coding DNA[6–8]. Here, we characterize the phylogenetic turnover of**

14   **low-level poly-adenylated transcripts in a comprehensive sampling of populations, sub-species and species of**

15   **the genus *Mus*, spanning a phylogenetic distance of about 10 Myr. We find evidence for more evolutionary**

16   **stable gains of transcription than losses among closely related taxa, balanced by a loss of older transcripts across**

17   **the whole phylogeny. We show that adding taxa increases the genomic transcript coverage and that no major**

18   **transcript-free islands exist over time. This suggests that the entire genome can be transcribed into poly-**

19   **adenylated RNA when viewed at an evolutionary time scale. Thus, any part of the "non-coding" genome can**

20   **become subject to evolutionary functionalization via *de novo* gene evolution.**

21   Genes can emerge *de novo* from non-genic regions of the genome[9–11]. Newly arising transcripts are initially usually

22   non-coding, can later acquire functional open reading frames[12–14] and can quickly become essential[15]. A number of

23   possibilities have been discussed by which new transcripts can arise, including single mutational events[10],

24   stabilization of bi-directional transcription[16] and insertion of transposable elements with promotor activity[17]. These

25   events were initially thought to be rare[6], but an increasing number of studies show that *de novo* gene emergence

26   is a rather active mechanism[18–22]. Surveys across phylogenetic times have shown that the highest gene emergence

27   rates are found in youngest taxa[7]. This led to the prediction that high emergence rates must be balanced with high

28   loss rates, because gene numbers do not grow much over time[7]. A comparison of open reading frame turnover

29   rates of *de novo* evolved genes among *Drosophila* species has shown that this is indeed the case[22].

30   Here we assessed the numbers of new transcript gains in a comprehensive phylogenetic framework. Given that the

31   emergence of a new stable transcript is a prerequisite for evolving a new functional gene, we expect that the

32   transcript emergence rate is a key parameter in the process of *de novo* gene emergence. Stable transcripts can be

33   created out of combinations of cryptic functional sites, including a minimal promoter, splicing signals and poly-

34   adenylation sites. This applies to completely new transcripts and to modification of existing transcripts by adding

35   new exons from non-coding DNA. Given the widespread presence of cryptic functional sites in genome sequences,

36   a single mutational step can convert a non-transcribed or a spuriously transcribed genome region into a stable

37   transcript, as has been shown for *Pldi,* the first documented *de novo* gene in the mouse[10].

38   Importantly, once a genomic region becomes transcribed, most subsequent mutations within the transcribed

39   region will not lead to a loss of the transcript, since only a few sites are responsible for active and stable

40   transcription. Hence, one can predict that the *de novo* transcript emergence dynamics would show a higher rate of

41   gain than loss at short evolutionary time scales. Hence, transcriptional gain would constitute a powerful

42   mechanism to continuously expose new genome regions to evolutionary testing, providing the fuel for *de novo*

43   gene emergence.

44   To test this prediction, we selected species, subspecies and populations related to the house mouse (*Mus*

45   *musculus* - suppl. Table S1) as a phylogenetic framework for identifying the emergence and loss of new transcripts.

46   The taxa chosen span approximately 10.5 Myr of evolutionary divergence and represent up to ~5.6% overall

47   genomic divergence (Figure 1, suppl. Table S2). Using such closely related taxa ensures that most neutrally evolving

48   sequences can be reliably mapped across all species[23]. We generated genome sequences for species without

49   published genomes, and transcriptome sequences for brain, liver and testis for all taxa (suppl. Tables S3-S5).

50   For comparative transcriptome analysis, we identified all mappable regions of the *M. m. domesticus* reference

51   genome (C57Bl/6)[24] using genomic reads from all studied species. We call this the "common genome",

52   representing the total genome where transcript mapping across taxa is reliable. We used a mapping algorithm that

53   was specifically designed to deal with the polymorphisms occurring under cross-species mapping conditions[23].

54   We first focused on genome-wide signals of transcriptional activity to identify the origin of new transcripts within

55   the phylogeny (suppl. Table S8). For this purpose, we determined the base-wise transcriptome coverage from poly-

56   adenylated RNA for each species. This measure of coverage includes both annotated genes and previously un-

57   annotated transcripts, whereby the latter are the majority. We set single read coverage as the lower cutoff

58   because we were specifically interested in detecting low-level transcription as an early sign of *de novo* gene

59   emergence. However, we also report results using a stringent cutoff of five reads for comparison (the median

60   coverage across all transcripts is 3.7).

61   When comparing transcriptome coverage among taxa, we find that the overall proportion of shared transcripts is

62   higher for closely related taxa than for distantly related pairs. Consequently, a phylogenetic tree reconstructed

63   based on shared transcript coverage mirrors the species tree (Figure 1B, C). This detectable phylogenetic signal in

64   transcription coverage suggests that transcripts gained at a given point in evolutionary time are sufficiently stable

65   to be retained in sister taxa, implying that they can become exposed to evolutionary testing.

66   The total transcript coverage of the common genome across all species combined for all three tissues is 67% in our

67   data set. Coverage was highest in testis (53.4%), intermediate in brain (41.5%) and lowest in liver (23.5%) (Figure

68   2A). When comparing all transcriptional gains versus losses across the surveyed phylogeny, we observe that gains

69   are indeed more frequent than losses (Figure 2B, C), thus confirming our initial hypothesis.

70   Recording base-wise coverage without paying attention to gene models entails the risk that one is also measuring

71   transcriptomic noise, i.e. spurious random transcriptional initiation in a subset of cells of the tissue under

72   investigation. We have specifically explored the noise issue through deeper sequencing of the brain samples of all

73   taxa. The brain is a complex tissue in which some transcripts are expected to occur only in a small subset of cells.

74   These rare transcripts should become detectable by deep sequencing and thus transcript coverage should increase

75   with more reads available. This is indeed the case; however rarefaction curves and their projections reach

76   saturation for each of the taxa (suppl. Table S6 and Figure S1). This observation argues against a significant amount

77   of transcriptional noise in our data, since noise should lead to a continuous increase of coverage with sequencing

78   depth, at least if noise is randomly distributed across the genome and across the cells of the tissue. Further, it rules

79   out a possible problem with DNA contamination, as this would also be expected to rise with increasing sequencing

80   depth.

81   We have further explored the rarefaction principle to assess whether adding more sequences or more taxa to the

82   data set leads to higher transcriptomic coverage of the common genome. Taking all aggregated reads (including

83   the additional deep sequencing data from brain) across all tissues for all taxa, saturation is reached at 78.5%

84   coverage for sequencing depth (Figure 2D), but no saturation is reached with the number of taxa used here (Figure

85   2E). Hence, adding more taxa within this phylogenetic framework, for example species and sub-species on the

86   *Apodemus* branch, would predictably lead to increasingly higher transcriptomic coverage of the common genome,

87   up to the entire genome when ~38 taxa were surveyed within the phylogeny (based on the intercept of the

88   regression curve).

89   This analysis suggests that there may be no regions that are not transcribed at some point in phylogenetic time.

90   However, genome annotations in a given species usually show an uneven distribution of transcripts; some regions

91   harbor many clustered transcripts and other regions are nearly devoid of transcripts ("gene deserts"). We

92   compared regions devoid of any transcriptomic coverage in our taxonomic sample to regions that show

93   transcription in at least one sample. We find that transcribed regions are more abundant and larger on average

94   than non-transcribed regions (Figure 2E). The maximum length of non-transcribed regions is ~ 20kb (at ≥1 reads)

95   or ~50kb (at ≥5 reads), suggesting that large gene-free "deserts" are in principle accessible to transcription and

3

96   possible regulatory constraints[25] do not fully prevent their transcription. In fact, the mouse de novo gene *Pldi* has

97   arisen within a gene desert[10].

98   Most *de novo* transcripts are expected to be neutral, but some may turn into more stable proto-genes[6,19] that can

99   eventually become functional, either as regulatory RNA, or by acquiring a functional reading frame. To identify

100   candidate proto-genes we used algorithms that are able to reconstruct transcriptional islands and splice junctions

101   (STAR[26]/cufflinks[27]) and join them into predicted gene models (suppl. Table S9). We did this for each taxon

102   separately and assessed the gain and loss patterns of these transcripts in a phylogenetic context. Excluding all

103   previously annotated transcripts, we find a total of 17,746 new candidate proto-genes, distributed across all taxa

104   (suppl. Figure S2). When looking only at gains of proto-gene transcripts in the terminal branches, we find that

105   about 1,300 new proto-genes are gained per million years (Figure 3A). Interestingly, at least 3,000 proto-genes are

106   already present at the youngest divergence level, implying within-species polymorphism that was also described

107   for *Drosophila*[21].

108   When counting gains versus losses of proto-genes, we find again higher numbers gained than lost over short

109   phylogenetic times. However, gains and losses balance out over longer evolutionary times when including the

110   whole phylogeny and all annotated genes (Figure 3B). This pattern confirms the two essential predictions we made

111   about *de novo* gene emergence: (1) newly acquired transcripts are not easily lost and thus have a life-time

112   sufficient for evolutionary testing and (2) genes do not accumulate over time because gain and loss rates are

113   balanced across longer time spans. Hence, when a given taxonomic lineage gains many *de novo* genes, it will lose

114   some of its older genes.

115   Our analysis is conservative in several respects. First, we focused on poly-adenylated transcripts, thereby avoiding

116   inclusion of RNA fragments that have been processed (i.e. excised introns) and randomly transcribed fragments.

117   However, this means we also exclude RNAs which are not transcribed by RNA polymerase II, such as tRNAs,

118   snRNAs and ribosomal RNAs. The human ENCODE data suggest that such non-poly-adenylated RNAs are abundant[1]

119   and it is likely that proto-genes can arise from those transcripts as well, i.e. we are likely underestimating the

120   proto-gene emergence rate. Second, we focused on three tissues and one developmental stage only. Although we

121   included testis and brain, which are known to have the highest diversity of transcripts[28] we can also expect that

122   including more tissues and developmental stages would further increase the transcriptomic coverage. Taking these

123   factors into account, as well as the fact that increased taxonomic representation shows no signs of saturation with

124   respect to transcriptomic coverage (Figure 2E), it seems reasonable to conclude that when measured at a

125   phylogenetic time scale, the entire genome can become subject to transcription.

126   Pervasive transcription of the genome was noted soon after deep sequencing approaches became possible and

127   this pattern was systematically explored in the ENCODE projects[1,29]. While the functional significance of pervasive

128   transcription is a matter of continuous dispute[4,5,30], our results provide an evolutionary dynamics perspective on

129   this question where emergence, functionalization and decay of gene functions should be seen as an evolutionary

130   life cycle of genes[8]. *De novo* gene birth should no longer be considered as the result of unlikely circumstances, but

131   rather a mechanism of testing genome regions for their adaptive potential. Within this evolutionary perspective,

132   any part of the genome – "junk" DNA included – has the possibility to become useful.

133

134

135   **References**

136   1. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111,** 6131–6138
137   (2014).
138   2. Kutter, C. *et al.* Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. *PLoS Genet* **8,**
139   e1002841 (2012).
140   3. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological
141   implications. *Trends Genet.* **30,** 439–452 (2014).
142   4. Van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most 'dark matter' transcripts are associated with
143   known genes. *PLoS Biol.* **8,** e1000371 (2010).
144   5. Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E. & Taft, R. J. The reality of pervasive transcription.
145   *PLoS Biol* **9,** e1000625 (2011).
146   6. Siepel, A. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res.* **19,** 1693–1695 (2009).
147   7. Tautz, D. & Domazet- Loso, T. The evolutionary origin of orphan genes. *Nat Rev Genet* **12,** 692 – 702 (2011).
148   8. Neme, R. & Tautz, D. Evolution: Dynamics of De Novo Gene Emergence. *Curr. Biol.* **24,** R238–R240 (2014).
149   9. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA
150   in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci.*
151   *U. S. A.* **103,** 9935–9939 (2006).
152   10. Heinen, T. J. A. J., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region.
153   *Curr. Biol. CB* **19,** 1527–1531 (2009).
154   11. Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res* **19,** 1752 –
155   1759 (2009).
156   12. Cai, J., Zhao, R., Jiang, H. & Wang, W. De Novo Origination of a New Protein-Coding Gene in Saccharomyces
157   cerevisiae. *Genetics* **179,** 487–496 (2008).
158   13. Xie, C. *et al.* Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS*
159   *Genet.* **8,** e1002942 (2012).
160   14. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides.
161   *eLife* **3,** e03523 (2014).
162   15. Chen, S., Zhang, Y. E. & Long, M. New Genes in Drosophila Quickly Become Essential. *Science* **330,** 1682–1685
163   (2010).
164   16. Wu, X. & Sharp, P. A. Divergent Transcription: A Driving Force for New Gene Origination? *Cell* **155,** 990–996
165   (2013).
166   17. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory
167   networks. *Genome Res.* **24,** 1963–1976 (2014).
168   18. Wu, D.-D., Irwin, D. M. & Zhang, Y.-P. De Novo Origin of Human Protein-Coding Genes. *PLoS Genet* **7,**
169   e1002379 (2011).

170 19. Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A. & Yildirim, M. A. Proto-genes and de novo gene
171   birth. *Nature* **487,** 370 – 374 (2012).
172 20. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo
173   evolution. *BMC Genomics* **14,** 117 (2013).
174 21. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and Spread of de Novo Genes in Drosophila melanogaster
175   Populations. *Science* **343**, 769 (2014).
176 22. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3,** (2014).
177 23. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly
178   polymorphic genomes. *Bioinforma. Oxf. Engl.* **29,** 2790–2791 (2013).
179 24. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562
180   (2002).
181 25. Montavon, T. & Duboule, D. Landscapes and archipelagos: spatial organization of gene regulation in
182   vertebrates. *Trends Cell Biol.* **22,** 347–354 (2012).
183 26. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* bts635 (2012).
184   doi:10.1093/bioinformatics/bts635
185 27. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes
186   using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
187 28. Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev.*
188   *Genet.* **15,** 734–748 (2014).
189 29. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W. et al. A comparative encyclopedia of DNA elements in the
190   mouse genome. Nature 515, 355-364 (2014).
191 30. Jensen, T. H., Jacquier, A. & Libri, D. Dealing with Pervasive Transcription. *Mol. Cell* **52,** 473–484 (2013).
192

Supplementary information

194 Supplementary_files includes Tables S1-S7 and Figures S1 and S2. Supplementary_TableS8 - Excel file with gain loss
195 patterns of transcript coverage per branch of the phylogeny. Supplementary_TableS9 - Excel file with list of all
196 gene models and gain/loss pattern along branches.

Author contribution

204 DT and RN conceived the project, RN did the experimental work and data analysis, DT and RN discussed the data
205 interpretation and wrote the manuscript.

206

207   **Methods online**

208   ***Origin of the sampled taxa***
209   We selected ten taxa, ranging from population level through sister genera (Figure 1 A).

210   The youngest divergence point sampled, at about 3,000 years, corresponds to the split between two European
211   populations of *Mus musculus domesticus* [1] one from France (Massif Central = MC) and one from Germany
212   (Cologne-Bonn area = CB) [2]. These European populations in turn have diverged from an ancestral *M. m. domesticus*
213   population in Iran (Ahvaz = AH) about 12,000 years ago [1]. The European *M. m. domesticus* are also the closest
214   relatives of the reference genome, the C57BL/6J strain [3].

215   We included two populations of *Mus musculus musculus*; one from Austria (Vienna = WI) and one from Kazakhstan
216   (Almaty = KH). These two populations are supposed to have a longer divergence between then the European *M. m.*
217   *domesticus* populations, but a more accurate estimate is currently not available. We set the divergence for
218   analyses at around 10,000 years as an approximate estimate.

219   *M. m. domesticus* has diverged from *M. m. musculus* and *Mus musculus castaneus* about 0.4 to 0.5 million years
220   ago, with a subsequent divergence, not long after, between *M. m. musculus* and *M. m. castaneus* [4]. We included
221   *M. m. castaneus* from Taiwan as a representative of the subspecies.

222   To account for longer divergence times, we included *Mus spicilegus* (estimated divergence of 1.2 million years);
223   *Mus spretus* (estimated divergence of 1.7 million years) [4]; *Mus matteyii* (subgenus *Nannomys*), the North African
224   miniature mouse (estimated divergence of 6.6 million years) [5,6], and *Apodemus uralensis*, the ural field mouse
225   (estimated divergence of 10.6 million years) [6].

226   The population-level samples (*M. m. domesticus* and *M. m. musculus*) included are maintained under outbreeding
227   schemes, which allows for natural polymorphisms to be present in the samples. All other non-population samples
228   are kept as more or less inbred stock, and therefore fewer polymorphisms are expected. All mice were obtained
229   from the mouse collection at the Max Planck Institute for Evolutionary Biology, following standard rearing
230   techniques which ensure a homogeneous environment for all animals. Mice were maintained and handled in
231   accordance to FELASA guidelines and German animal welfare law (Tierschutzgesetz § 11, permit from Veterinäramt
232   Kreis Plön: 1401-144/PLÖ-004697).

233   A total of 60 mice were sampled, as follows: Eight male individuals from each population-level sample (outbreds),
234   Iran (AH), France (MC), and Germany (CB) of *Mus musculus domesticus*, and Austria (WI) and Kazakhstan (KH) of
235   *Mus musculus musculus*. Four male individuals from the remaining taxa (partially inbred): *Mus musculus castaneus*
236   (TA), *Mus spretus* (SP), *Mus spicilegus* (SC), *Mus mattheyi* (MA) and *Apodemus uralensis* (AP). Mice were sacrificed
237   by CO_2 asphyxiation followed immediately by cervical dislocation. Mice were dissected and tissues were snap-
238   frozen within 5 minutes post-mortem. The tissues collected were liver (ventral view: front right lobe), both testis
239   and whole brain including brain stem.

240   ***Genome sequencing***
241   One individual from each of *M. spicilegus*, *M. mattheyi*, and *Apodemus uralensis* were selected for genome
242   sequencing. DNA was extracted from liver samples. DNA extraction was performed using a standard salt extraction
243   protocol. Tagged libraries were prepared using the Genomic DNA Sample preparation kit from Illumina, following
244   the manufacturers' instructions: After library preparation, the three genome samples were pooled together and

7

245   run in a whole IlluminaHiSeq 2000 flow cell (8 lanes, approximately 2.6 lanes per sample). Library preparation and

246   sequencing was performed at the Cologne Center for Genomics.

247   The genome from the strain SPRET/EiJ derived from *Mus spretus* was taken from [7,8], and was downloaded from the

248   European Nucleotide Archive (ENA) - accessions ERS076388 and ERS138732.

Transcriptome sequencing

250   The sampled tissues of each taxon were used for RNA extraction with the RNAeasy Mini Kit (QUIAGEN) and pooled

251   at equimolar concentrations. RNA quality was measured with the Agilent RNA Nano Kit, for the individual samples

252   and pools. Samples with RIN values above 7.5 were used for sequencing. Library preparation was done using the

253   Illumina TruSeq library preparation, with mRNA purification (PolyA selection), following manufacturers'

254   instructions. Sequencing was done in Illumina HiSeq 2000 sequencer. Libraries for each group were tagged, pooled

255   and sequenced in a single lane, corresponding to approximately one third of a HiSeq2000 lane. Additional

256   sequencing of the brain samples was performed to identify potential limitations in depth of sequencing. For this,

257   each brain library was sequenced on a full lane of a HiSeq2000. All library preparation and sequencing was done at

258   the Cologne Center for Genomics (CCG).

Raw data processing

260   All raw data files were trimmed for adaptors and quality using Trimmomatic [9]. The quality trimming was performed

261   basewise, removing bases below quality score of 20 (Q20), and keeping reads whose average quality was of at

262   least Q30. Reads whose trimmed length was shorter than 60 bases were excluded from further analyses, and pairs

263   missing one member because of poor quality were also removed from any further analyses.

Mapping

265   The reconstruction of transcriptomes using high-throughput sequencing data is not trivial when comparing

266   information across different species to a single reference genome. This is due to the fact that most of the tools

267   designed for such tasks do not work in a phylogenetically aware context. For this reason, any approximation which

268   deals with fractional data (i.e. any high-throughput sequencing setup available to this date) is limited by the

269   detection abilities of the software of choice and by the quality of the reference (transcriptome and genome).

270   Given the high quality state of the mouse genome repositories, we decided to take a reference-based approach, in

271   which all analyses are centered in the reference genome of the C57BL/6 laboratory strain of *Mus musculus*

272   *domesticus*. This enables direct comparisons across all species, with an obvious cost introduced by the mapping of

273   distantly related genomes.

274   For general comparisons, transcriptome and genome sequencing reads were aligned against the mm10 version of

275   the mouse reference genome from UCSC [10] using NextGenMap which performs extremely well with divergences of

276   over 10% compared to other standard mapping software [11]. The program was run under default settings, except

277   for --strata 1 and --silent-clip. The first option enforces uniquely mapping reads and the second drops the

278   unmapped portion of the reads, to avoid inflating coverage statistics. This is particularly relevant around exon-

279   intron boundaries, where exonic reads are forced into intronic regions unless this option is set.

280   Genomic reads were used to as empiric mapability, i.e. to identify which regions can be reliably detected. We

281   limited our analyses to regions in the reference genome which could be mapped at least 5 times from genomic

282   reads from all other species (5x coverage). This is the portion we call the 'common genome' in downstream

283   analyses. It is important to highlight that this is not the same as synteny, since we did not perform any co-linearity

284   analyses between fragments, but rather represents the mere presence in the species, in any possible order.

285 Furthermore having genomic reads enables the detection of true absences in transcriptional activity from absences
286 of genome regions, which would show similar patterns in transcriptome-only analyses.

Reconstruction of gene models

288 Due to the fact that NextGenMap is unable to perform split read analyses we opted for more standard tools to
289 reconstruct gene models from the data. For this we used STAR [12] to map reads to the reference, followed by
290 cufflinks [13] to obtain automated gene and transcript annotations for each species. The annotation file contains
291 models for expressed transcripts with splicing information (exon annotation) when available. All annotations were
292 merged using cuffmerge to generate a final annotation that includes gene models present at least once in the total
293 sample. Mono-exonic models shorter than 500 bases or contained within introns of multi-exonic transcripts were
294 excluded from analyses.

Parsimony gain and loss mapping

296 We estimated gain and loss events given the phylogenetic distribution of presence and absence of transcription at
297 a given position or for a given gene model using maximum parsimony (based on GLOOME, [14], the assumption that
298 gains and losses are equally likely, and a fixed tree describing the relationships between taxa.

Genome-wide estimation of transcriptional gains and losses

300 Genome-wide estimates of gain and loss of transcription were done at the nucleotide level, considering only
301 regions within the common genome.

302 Normalized versions of each set of aligned reads were generated by subsampling (samtools view –s x [15]; where x is
303 the proportion of each individual sample that matches the least abundant sample). Normalization was done only
304 across tissues and not between them. Normalized samples were merged at the species level to obtain a species-
305 wide transcription sample. Aligned reads in BAM format were converted to BedGraph format for phylogenetic
306 comparisons of format using bedtools [16]. Two parallel sets of comparisons were made: i) using all coverage
307 information from uniquely mapping reads, thus representing 'absolute' (minus normalization) coverage, and ii)
308 using a threshold of at least 5 uniquely mapped reads, thus representing 'stringent' coverage. Coverage files were
309 compared between groups using multiIntersectBed [16], to obtain the portions shared between all possible
310 combinations.

311 Each possible combination can be also interpreted as a binary presence/absence pattern. We summarized the total
312 amount of nucleotides in each specific pattern. Each pattern received a fixed amount of gains and losses,
313 consistent with the parsimony assumptions using GLOOME [14] in maximum parsimony mode. For example, the
314 pattern that indicates presence in German *M. m. domesticus* (CB), but absence in all other groups, corresponds to
315 one very recent gain and zero losses. The pattern that indicates presence in German and Iranian *M. m. domesticus*
316 (CB and AH), but absence in all other groups, corresponds to one gain (ancestor of *M. m. domesticus*) and one loss
317 (after divergence of French *M. m. domesticus*). In this context, we identify monophyletic gains as stable, i.e.
318 transcription is present in all derived groups after estimated gain, and unstable, i.e. transcription lost in at least
319 one derived group after gain.

Gene gain and loss rates from gene models

321 Gene models derived from STAR alignments and cufflinks reconstructions were used to calculate the rate at which
322 gene-like entities are gained or lost along the phylogeny. A single unified annotation for all species was generated
323 with cuffmerge (see "Gene model reconstruction" above) and FPKM values were obtained from mapped reads for
324 each tissue and species. FPKM of 0.1 was set as a threshold to define the presence or absence of a gene model in a

325   given sample. Similar to the reconstruction of genome-wide patterns, a maximum parsimony framework was

326   employed, assuming that gains and losses have equal probabilities of occurrence.

327   Rates of gain and loss of gene models were estimated from linear regressions between time and gain or loss using

328   R, with the lm() function from the stats package [17]. Rates were calculated for each tissue and each possible

329   combination of tissues (a gene can be differentially present or absent in a species), to obtain tissue-specific gains

330   and losses of models.

331   ***Reconstruction of phylogenetic relationships***

332   Using a manhattan distance matrix from the summarized transcriptional coverage (suppl. Table S8), we

333   constructed a neighbor-joining (NJ) tree that describes the proximity between any two taxa based on the shared

334   transcriptomic coverage between them. In this representation, closely related organisms have more shared

335   transcriptomic coverage than distantly related organisms. Analyses were performed in R, using the function dist()

336   from the stats package and nj() from the ape package [18].

337   Additionally, whole mitochondrial genomes were obtained for each taxon as consensus sequences from mapped

338   reads using samtools mpileup [15]. The sequences were aligned with MUSCLE [19], and a NJ tree was constructed with

339   the dist.dna() and nj() functions from the ape package [18]. All NJ trees were tested with 1000 bootstraps with the

340   boot.phylo() function from the ape package [18]. Reported trees have a support of 60% or greater.

341   ***Rarefaction and subsampling***

342   Transcriptome experiments tend to be limited by the depth of sequencing, with highly expressed genes being

343   relatively easy to sample, and rare transcripts becoming increasingly difficult to find. Given the large amount of

344   data generated, we investigated if our data shows signals of coverage saturation from subsets of the data of

345   different sizes. The total experiment, comprising ten taxa, corresponds to 6.4×10^9 reads (or 6.4 billion reads). We

346   subsampled (samtools view -s) portions of mapped reads for each taxon, ranging between 10% to 90%, at 10%

347   intervals. The observation of coverage saturation in this case would indicate that our sequencing efforts likely

348   cover most of the transcribed regions of the common genome.

349   In parallel, we estimated the individual and combined contribution of each taxon to the transcriptomic coverage of

350   the common genome. Not all samples have the same phylogenetic distance to each other (some species have

351   more representatives than others). To account for this we generated one hundred arrays of the ten taxa with

352   random order, and recorded the coverage after the addition of each taxon in each array. The observation of

353   coverage saturation in this setup would indicate that taxonomic sampling is sufficient to cover most of the

354   potentially transcribed regions of the common genome. In order to estimate whether our data continued to

355   increase or approached saturation, we tested two alternative models: a generalized linear model with logarithmic

356   behavior (ever increasing) or a self-starting nonlinear regression model (saturating). Best fit was decided based on

357   the lowest AIC and BIC values. Analyses were performed in R, using the functions glm(), nls(), SSasymp(), BIC() and

358   AIC() from the stats package [17].

359   ***Analysis of transcribed and non-transcribed regions across the genome***

360   Transcribed and non-transcribed regions larger than 100 nucleotides were defined by the continuous presence or

361   absence of transcriptomic coverage from mapping information of each taxon and tissue. Combined transcribed

362   regions across species were obtained as mentioned before, and combined non-transcribed regions across species

363   were generated by subtracting transcribed regions from the common genome.

364 ***References for Methods online***

365 1. Cucchi, T., Vigne, J.-D. & Auffray, J.-C. First occurrence of the house mouse (Mus musculus domesticus
366 Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil
367 occurrences. *Biol. J. Linn. Soc.* **84,** 429–445 (2005).

368 2. Ihle, S., Ravaoarimanana, I., Thomas, M. & Tautz, D. An analysis of signatures of selective sweeps in natural
369 populations of the house mouse. *Mol. Biol. Evol.* **23,** 790–797 (2006).

370 3. Didion, J. P. & de Villena, F. P.-M. Deconstructing Mus gemischus: advances in understanding ancestry,
371 structure, and variation in the genome of the laboratory mouse. *Mamm. Genome Off. J. Int. Mamm. Genome*
372 *Soc.* **24,** 1–20 (2013).

373 4. Suzuki, H. *et al.* Evolutionary and dispersal history of Eurasian house mice Mus musculus clarified by more
374 extensive geographic sampling of mitochondrial DNA. *Heredity* **111,** 375–390 (2013).

375 5. Catzeflis, F. M. & Denys, C. The African Nannomys (Muridae): An early offshoot from the Mus lineage -
376 evidence from scnDNA hybridization experiments and compared morphology. *Isr. J. Zool.* **38,** 219–231 (1992).

377 6. Lecompte, E. *et al.* Phylogeny and biogeography of African Murinae based on mitochondrial and nuclear gene
378 sequences, with a new tribal classification of the subfamily. *BMC Evol. Biol.* **8,** 199 (2008).

379 7. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,**
380 289–294 (2011).

381 8. Yalcin, B., Adams, D. J., Flint, J. & Keane, T. M. Next-generation sequencing of experimental mouse strains.
382 *Mamm. Genome* **23,** 490–498 (2012).

383 9. Lohse, M. *et al.* RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics.
384 *Nucleic Acids Res.* **40,** W622–627 (2012).

385 10. Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S. & Karolchik, D. The UCSC Genome Browser database:
386 update 2011. *Nucleic Acids Res* **39,** D876 – D882 (2011).

387 11. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly
388 polymorphic genomes. *Bioinforma. Oxf. Engl.* **29,** 2790–2791 (2013).

389 12. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* bts635 (2012).
390 doi:10.1093/bioinformatics/bts635

391 13. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes
392 using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).

393 14. Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D. & Pupko, T. GLOOME: gain loss mapping engine.
394 *Bioinformatics* **26,** 2914–2915 (2010).

395 15. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).

396 16. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*
397 **26,** 841–842 (2010).

398 17. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical
399 Computing, Vienna, Austria, 2014). at <http://www.R-project.org/>

400 18. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language.
401 *Bioinformatics* **20,** 289–290 (2004).

402 19. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*
403 **32,** 1792–1797 (2004).

404

405

11
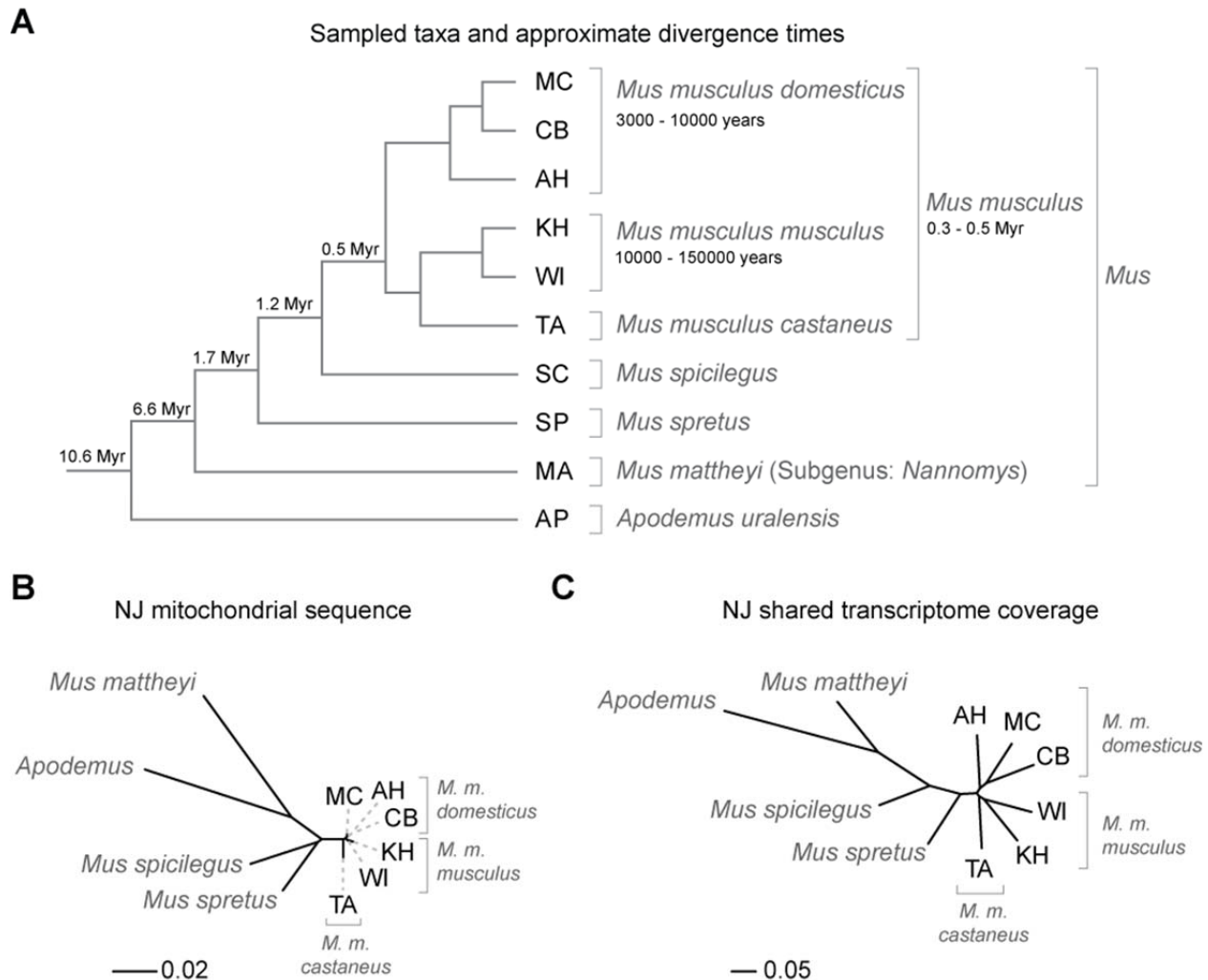
Figure 1. (A) Schematic relationships and approximate divergence times (see Methods) of the taxa under study. Tree branches are not shown to scale. (B) Molecular phylogeny based on whole mitochondrial genome sequences as a measure of molecular divergence (black lines represent the branch lengths, dashed lines serve to highlight short branches). (C) Tree based on shared transcriptome coverage of the genome. The percentage of shared transcripts mirrors the phylogenetic relationships between the studied taxa. MC: *M. m. domesticus* from France. CB: *M. m. domesticus* from Germany. AH: *M. m. domesticus* from Iran. KH: *M. m. musculus* from Kazakhstan. WI: *M. m. musculus* from Austria. TA: *M. m. castaneus* from Taiwan.
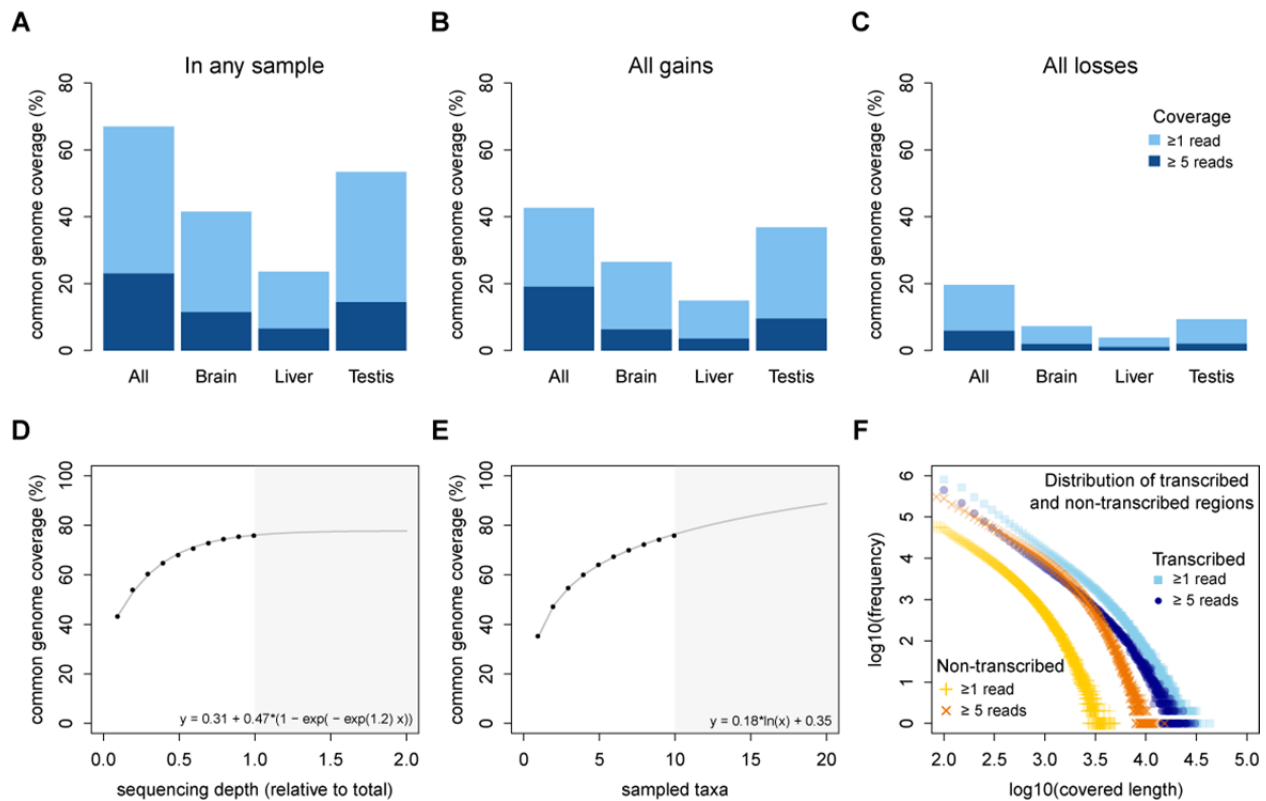
**Figure 2**. Coverage and phylogenetic turnover of base-wise transcription of the common genome. (A) Coverage across all taxa based on similar sequencing depth for each tissue (suppl. Table xx). (B, C) All gains (B) and all losses (C) along the phylogeny, assuming maximum parsimony (equal gain and loss probability). Light blue represents regions with base-wise coverage of least one uniquely mapping read, dark blue represents regions of base-wise coverage of at least five uniquely mapping reads. (D, E) Rarefaction, subsampling and saturation patterns using all available samples, including deeper sequencing of the brain samples. (D) sequencing depth saturation as estimated from a non-linear regression with asymptotic behavior, (E) sequencing depth saturation as estimated from an increase in the number of taxa. Black dots indicate increases per sub-sampled sequence fraction or taxon added from our dataset. Gray line indicates the predicted behavior from the indicated regression, and gray area shows the prediction after doubling the current sampling either in sequencing effort (D) or additional taxa (E). Each analysis was tested for logarithmic and asymptotic models and best fit was selected by AIC and BIC (see Methods). Standard deviations are too small to become visible in the plots. (F) Comparative analysis of lengths of regions transcribed or not transcribed across all data (including deeper brain sequencing) in all samples. Size distribution of regions not covered in any transcript (yellow) or with less of five transcripts (orange) compared to size distribution of regions with at least one transcript (light blue) or at least five transcripts (dark blue).

13

436

437

438
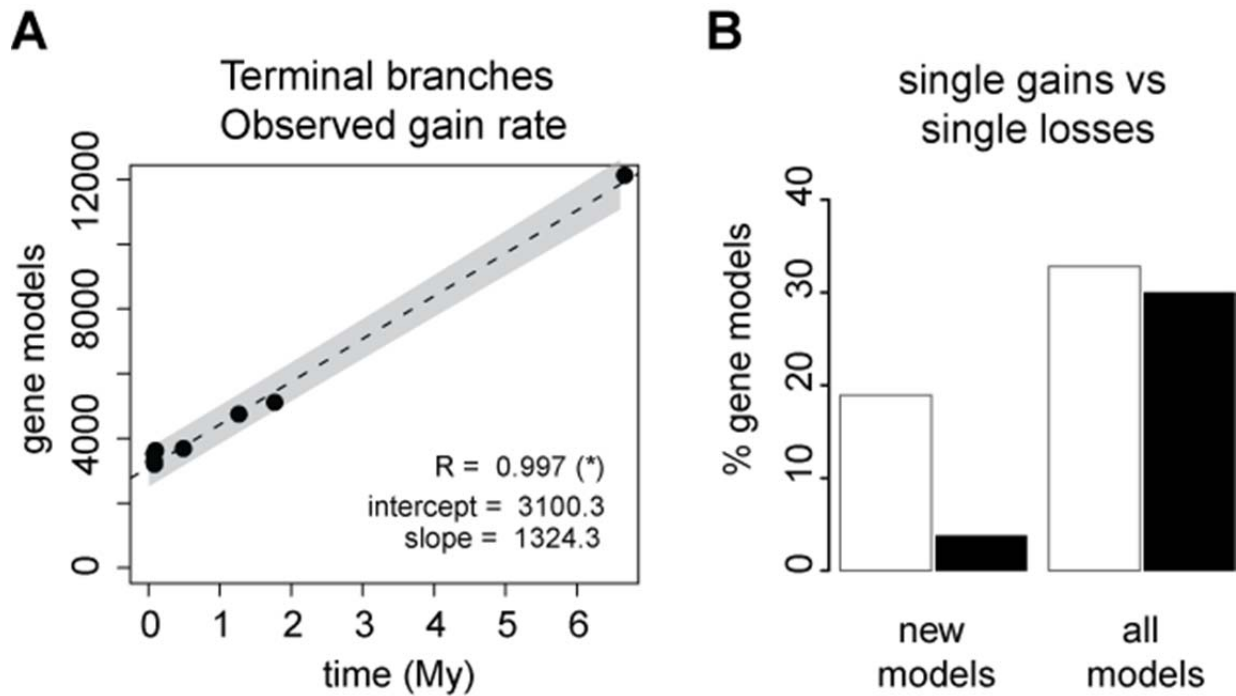
Figure 3: Turnover of proto-gene candidates. (A) Linear regression of observed gains across the phylogeny, using *Apodemus* as outgroup, and counting only gains in terminal branches. Regression coefficient is significant at p < 0.01. The grey area shows the 95% confidence interval of the regression. (B) Single gains versus single losses of a proto-gene candidate, for 'new' gene models, i.e. absent in *Apodemus*, and versus all detected models, i.e. including the whole gene set. The ratio of gain to loss of new models is significantly different from the observed for all models (Fisher's exact test, p < 0.01).

445