

1 **From de novo to ‘de novo’: most novel protein coding genes**
2 **identified with phylostratigraphy represent old genes or recent**
3 **duplicates**

4

5 **Claudio Casola^{1*}**

6

7 ¹ Department of Ecosystem Science and Management, Texas A&M University,
8 College Station, TX 77843-2138

9

10

11 **Abstract**

12 The evolution of novel protein-coding genes from noncoding regions of the
13 genome is one of the most compelling evidence for genetic innovations in nature.
14 One popular approach to identify de novo genes is phylostratigraphy, which
15 consists of determining the approximate time of origin (age) of a gene based on
16 its distribution along a species phylogeny. Several studies have revealed
17 significant flaws in determining the age of genes, including de novo genes, using
18 phylostratigraphy alone. However, the rate of false positives in de novo gene
19 surveys, based on phylostratigraphy, remains unknown. Here, I re-analyze the
20 findings from three studies, two of which identified tens to hundreds of rodent-
21 specific de novo genes adopting a phylostratigraphy-centered approach. Most of
22 the putative de novo genes discovered in these investigations are no longer
23 included in recently updated mouse gene sets. Using a combination of synteny
24 information and sequence similarity searches, I show that about 60% of the
25 remaining 381 putative de novo genes share homology with genes from other
26 vertebrates, originated through gene duplication, and/or share no synteny
27 information with non-rodent mammals. These results led to an estimated rate of
28 ~12 de novo genes per million year in mouse. Contrary to a previous study
29 (Wilson et al. 2017), I found no evidence supporting the preadaptation hypothesis
30 of de novo gene formation. Nearly half of the de novo genes confirmed in this
31 study are within older genes, indicating that co-option of preexisting regulatory
32 regions and a higher GC content may facilitate the origin of novel genes.

33

34

35 **Introduction**

36 Protein-coding genes can emerge through mechanisms varying from gene
37 duplication to horizontal transfer and the ‘domestication’ of transposable
38 elements, all of which involve pre-existing coding regions. Conversely, the
39 process of de novo gene formation consists of the evolution of novel coding
40 sequences from previously noncoding regions, thus generating entirely novel

41 proteins. The discovery of de novo genes is facilitated by extensive comparative
42 genomic data of closely related species and their accurate gene annotation.
43 Because of these requirements, de novo genes have been mainly characterized
44 in model organisms such as *Saccharomyces cerevisiae* (Carvunis et al. 2012; Lu
45 et al. 2017; Vakirlis et al. 2017), *Drosophila* (Begun et al. 2007; Reinhardt et al.
46 2013; Zhao et al. 2014) and mammals (Heinen et al. 2009; Knowles and
47 McLysaght 2009; Li et al. 2010; Murphy and McLysaght 2012; Neme and Tautz
48 2013; Ruiz-Orera et al. 2015; Guerzoni and McLysaght 2016; Neme and Tautz
49 2016).

50
51 Even among model organisms, the identification of de novo genes remains
52 challenging. One major caveat in de novo gene discovery is their association with
53 signatures of biological function. Because de novo genes described thus far tend
54 to be taxonomically restricted to a narrow set of species, their functionality is not
55 obvious as for older genes that are conserved across multiple taxa. Transcription
56 and translation are considered strong evidence of de novo genes functionality,
57 although the detection of peptides encoded from putative coding sequences is
58 not always an indication of protein activity (Xu and Zhang 2016). Given their
59 limited taxonomic distribution, testing sequence conservation and selective
60 regimes on de novo genes coding sequences is often unachievable. Population
61 genomic data can provide evidence of purifying selection on de novo genes but
62 are thus far limited to a relatively small number of species (Zhao et al. 2014;
63 Chen et al. 2015).

64
65 An important signature of de novo gene evolution is the presence of *enabler*
66 substitutions, which are nucleotide changes that alter a 'proto-genic' DNA region
67 to facilitate its transcription and/or ability to encode for a protein (Vakirlis et al.
68 2017). Enabler substitutions can be recognized only by comparing de novo
69 genes with their ancestral noncoding state, which can be identified by analyzing
70 syntenic regions in the genome of closely related species that do not share such
71 substitutions (Guerzoni and McLysaght 2016). It follows that putative de novo
72 genes with no detectable synteny in other species have likely originated through
73 other processes, including duplication of preexisting genes, horizontal gene
74 transfer, or transposon insertion and subsequent domestication. Loss of synteny
75 may also arise through deletion of orthologous genes in other species; however,
76 this scenario appears improbable when comparing a large number of genomes.

77
78 A third fundamental hallmark of de novo genes is the lack of homology of their
79 proteins with proteins from other organisms. This feature is often the first step in
80 comparative genome-wide surveys aimed at identifying putative de novo genes

81 in a given species or group of species. Similarly, de novo proteins should be
82 devoid of functional domains that occur in older proteins.

83

84 Previous studies on de novo gene evolution show a range of complexity in the
85 strategies used to characterize these genes. A common approach to assess the
86 evolutionary age of genes is known as *phylostratigraphy* and it has often been
87 applied to retrieve a set of candidate de novo genes. This method relies on
88 homology searches, usually consisting of BLAST surveys of a species proteome,
89 allowing the inference of a putative 'age' of each gene along the phylogeny of a
90 group of species (Domazet-Lošo et al. 2007). For example, mouse proteins that
91 share homology with sequences in the rat proteome, but are absent in other
92 mammals, would be categorized as 'rodent-specific'. Notably, genes that appear
93 to be lineage-specific may represent de novo genes, but could also be derived
94 from any of the other processes mentioned above.

95

96 Phylostratigraphy is theoretically simple and effective, yet it is known to contain
97 several methodological flaws (Elhaik et al. 2006; Moyers and Zhang 2015, 2016;
98 Moyers and Zhang 2017). For instance, both rapid sequence evolution and short
99 coding sequences lead to underestimating gene age (Moyers and Zhang 2015).
100 This increases the likelihood for rapidly evolving genes to be erroneously
101 recognized as novel species-specific genes when phylostratigraphy-only
102 approaches are used. It is known that after gene duplication one or both of the
103 two copies may experience accelerated sequence evolution, which may result in
104 an underestimate of their age. In agreement with this observation, a recent study
105 in primates has shown that genes that evolve faster also tend to duplicate more
106 (O'Toole et al. 2018). Importantly, phylostratigraphic studies that ignore synteny
107 data will be unable to provide evidence of enabler substitutions and are thus
108 uninformative of the mechanism by which those genes evolved.

109

110 It is perhaps not surprising that researches chiefly based on phylostratigraphy
111 have led to estimates of de novo gene formation rates that exceeds or are
112 comparable to those of gene duplication. For instance, it has been suggested
113 that the *S. cerevisiae* genome contains hundreds of de novo genes that emerged
114 during the Ascomycota evolution and that at least 19 genes are *S. cerevisiae*-
115 specific, compared to a handful of gene duplicates found only in *S. cerevisiae*
116 (Carvunis et al. 2012). Similarly, a relatively recent study reported that 780 de
117 novo genes emerged in mouse since its separation from the Brown Norway rat
118 around 12 million years ago, at a rate of 65 genes/million year (Neme and Tautz
119 2013). According to these estimates, de novo genes represent about half of all
120 young genes, the other half being formed by gene duplicates. Because the

121 overall gene number did not appear to have increased significantly during
122 mammal and yeast evolution, such a high pace of de novo gene formation must
123 be accompanied by rampant levels of gene loss. If true, this would represent a
124 'gene turnover paradox', given that most genes are maintained across mammals.
125 For instance, according to the Mouse Genome Database, 17,093/22,909 (~75%)
126 protein coding mouse genes share homology with human genes (Blake et al.
127 2017).

128
129 The 'gene turnover paradox' has been opposed by some authors. For example,
130 analyses by Moyers and Zhang used simulations to show that gene age is
131 underestimated in a significant proportion of cases based on phylostratigraphy
132 (Moyers and Zhang 2015, 2016; Moyers and Zhang 2017). These authors
133 pointed out that most putative *S. cerevisiae*-specific de novo genes overlap with
134 older genes and show no signature of selection operating on their coding
135 sequence (Moyers and Zhang 2016). These studies have proved critical to
136 address major pitfalls of phylostratigraphy. However, the exact proportion of false
137 positives in de novo gene studies remains unknown and it is unclear how many
138 putative de novo genes should instead be considered fast evolving gene
139 duplicates. A correct assessment of de novo genes is critical to establish their
140 evolutionary history and more broadly to identify genomic features, if any, that
141 may facilitate the emergence of novel genes.

142
143 Here, I address these issues by re-analyzing putative mouse de novo genes from
144 three recent articles (Murphy and McLysaght 2012; Neme and Tautz 2013;
145 Wilson et al. 2017) using a combination of sequence similarity searches and
146 synteny information. I show that more than half of the 874 putative de novo
147 genes previously described in mouse are absent in current versions of three
148 major mouse gene annotation databases, an indication of how gene annotation
149 volatility can affect de novo gene studies even among model organisms. Of the
150 remaining putative de novo genes, only ~40% could be validated. The dismissed
151 putative de novo genes either shared homology with genes found in multiple non-
152 rodent vertebrates, derived from duplication of pre-existing mouse genes, and/or
153 lacked synteny information with non-rodent mammals. I collectively refer to the
154 putative de novo genes that failed to pass the validation criteria as the 'de novo'
155 genes. These findings also indicate that false positives in phylostratigraphy
156 studies of de novo genes exceed previous estimates of type I error rates based
157 on simulations in *S. cerevisiae* (Moyers and Zhang 2016). Contrary to what was
158 suggested in a recent study (Wilson et al. 2017), I found no evidence of
159 preadaptation in the validated mouse de novo genes. Instead, I observed that the
160 trend reported by Wilson and collaborators, an inverse correlation between

161 intrinsic structural disorder (ISD) of proteins and gene age suggestive of a lower
162 tendency towards aggregation in proteins encoded by younger genes, is primarily
163 due to high ISD levels in de novo genes whose coding region overlap exons of
164 older genes.

165

166

167 **Results and Discussion**

168

169 *Putative de novo gene annotation status*

170

171 In this study, I re-analyzed Putative De Novo Genes (hereafter: PDNGs) from
172 three articles focused on rodent genomes (Murphy and McLysaght 2012; Neme
173 and Tautz 2013; Wilson et al. 2017). Hereafter, I will refer to these works as
174 M2012 (Murphy and McLysaght 2012), N2013 (Neme and Tautz 2013) and
175 W2017 (Wilson et al. 2017). I specifically focused on mouse-specific genes from
176 the M2012 and N2013 studies, and on the rodent-specific genes from the W2017
177 study. A total of 491 previously reported rodent PDNGs, particularly those from
178 the M2012 and N21023 studies, are not annotated as protein-coding genes in the
179 updated versions of three major mouse gene annotation databases: GENCODE
180 M16, RefSeq and UCSC Genome Browser 'known' genes (Table 1).

181

182

183 **Table 1. Original PDNG sets, currently annotated PDNGs and de novo**
184 **genes assessed in this study**

	M2012	N2013	W2017	Total
Mouse PDNG	69	773	84	874
Annotated in mm10	9	331	72	381
de novo genes	3 (7)	74 (139)	13 (18)	82 (152)

185 Numbers in parenthesis refer to de novo genes remaining when genes from automatic annotation
186 pipelines are included.

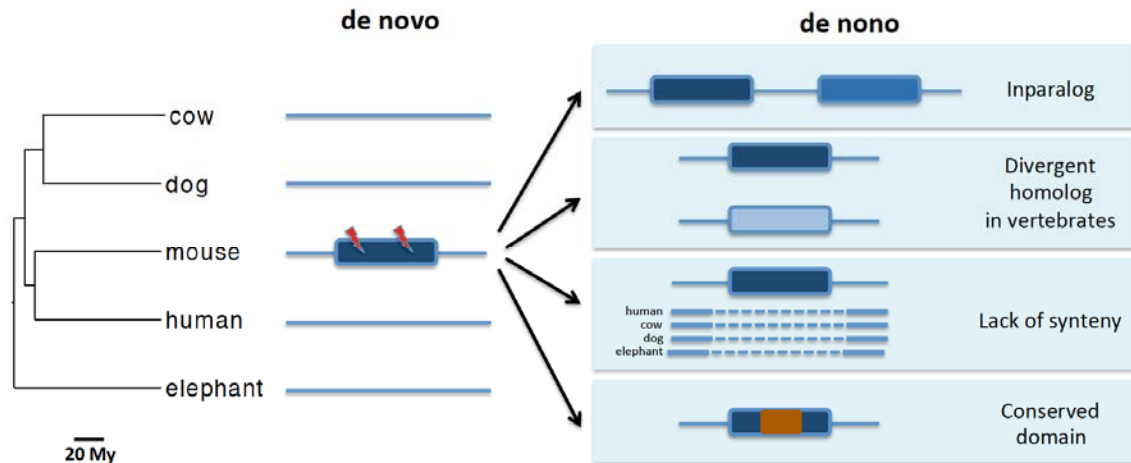
187

188

189 This is expected given that the three studies were based on less-well curated
190 gene sets. Surprisingly, the N2013 and W2017 datasets contained several
191 PDNGs lacking a start codon. These genes were excluded from further analysis
192 in this work. The final count of PDNGs with confirmed annotation as protein-
193 coding genes was 9, 331, and 72 from the M2012, N2013, and W2017 studies,
194 respectively (Table 1). After excluding overlap between the three gene sets, I
195 retrieved 381 PDNGs, which represent 44% of the 874 genes originally reported
196 as de novo genes. Below, I describe the four criteria I used to assess the
197 proportion of PDNGs that represent 'de novo' genes: presence of paralogous

198 genes in mouse (inparalogs); homology with genes found in multiple non-rodent
199 vertebrates; lack of synteny information with non-rodent mammals; presence of
200 conserved domains found in non-rodent proteins (Figs. 1, 2).

201
202



203
204

205 **Figure 1.** Features distinguishing de novo and 'de novo' genes. Rectangles, solid lines
206 and dashed lines represent genes, nongenic syntenic regions and nonsyntenic regions,
207 respectively. Presence of enabler substitutions (lightening bolts), absence of inparalogs
208 and homologs in other species, conserved synteny and lack of conserved domains
209 characterize de novo genes. Putative de novo genes that fail to conform to one or more
210 of these criteria represent 'de novo' genes. My: million years.

211
212

213 *Sequence similarity analyses to identify homologous genes in non-rodent* 214 *genomes*

215

216 According to the phylostratigraphic approach, de novo proteins in a focal
217 taxonomic group are recognized as such because they share no significant
218 sequence similarity with proteins from other taxa (Fig. 1). This assumption can be
219 violated under two scenarios. First, novel proteins are routinely added to existing
220 sequence databases, thus expanding the sequence space available to search for
221 possible homologous sequences of PDNGs. To explore this possibility I carried
222 out tBLASTn searches against the NCBI vertebrate nucleotide non-redundant
223 database using mouse de novo proteins. Second, alternative sequence similarity
224 search algorithms than those used in the original studies may reveal yet
225 unrecognized homologs of PDNGs. For instance, profile-based approaches such
226 as phmmer can be more accurate than non-profile methods, including BLAST, in
227 sequence homology searches (Saripella et al. 2016). A combination of these

228 methods has recently been applied to detect de novo genes in two yeast
 229 genomes (Vakirlis et al. 2017). I therefore interrogated a reference proteome
 230 database available through the EMBL-EBI phmmer server to identify PDNG
 231 homologs that are not recognized using BLASTP (see Methods). Finally, I
 232 visually inspected all PDNGs with synteny information to find possible
 233 orthologous in the human genome using the UCSC Genome Browser net-
 234 alignment track (Schwartz et al. 2003). Combining the results of both analyses I
 235 identified 98 PDNGs (26% of all PDNGs) with homologs in two or more
 236 vertebrate species (Table 2), including fourteen PDNGs with orthologous genes
 237 in human (Fig. 2d; Table S1).

238
 239

240 **Table 2. Summary of homology searches and synteny analysis for the three**
 241 **PDNG sets.**

	M2012	N2013	W2017	Combined PDNGs
Inparalogs (BLAST) ¹	0	63	27	81
Inparalogs (HMMER) ¹	0	80	31	102
PDNGs w/ inparalogs	0	88	32	110
Homology in Vertebrates (BLAST) ²	1	60	5	62
Homology in Vertebrates (HMMER) ²	0	33	15	43
PDNGs w/ homologs in Vertebrates	1	86	20	98
Presence of protein domain	1	23	19	39
Lack of synteny ³	1	110	30	131
Overall Total	3	192	55	229

242 ¹ Significant similarity (BLAST: e-value≤0.001; HMMER: i-Value≤0.001) with GENCODE M16,
 243 RefSeq and/or UCSC Genome Browser mouse genes

244 ² Homologous sequences (BLAST: e-value≤0.001; HMMER: i-Value≤0.001) found in at least two
 245 non-rodent vertebrate species

246 ³ No synteny conservation of PDNGs coding regions across mammals

247
 248

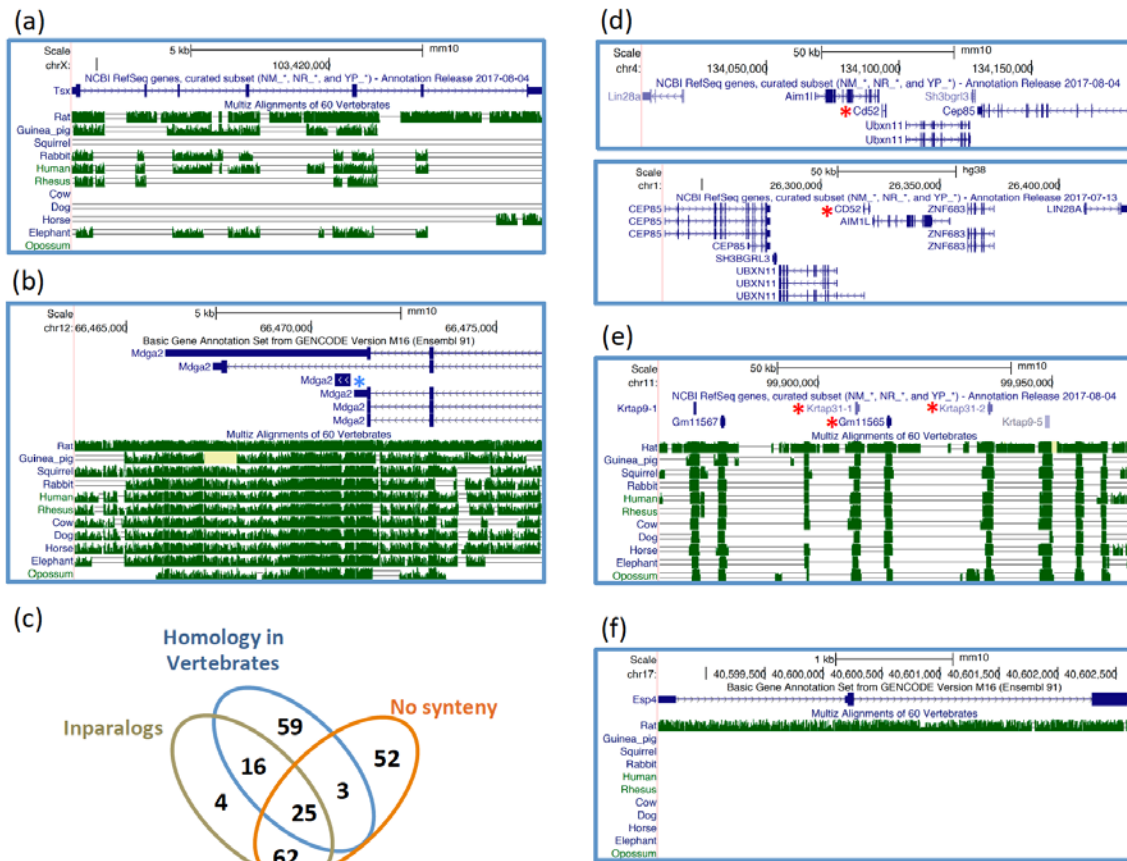
249 As expected, alignments of proteins from some of these orthologs showed short
 250 regions of sequence conservation (Fig. S1).

251 For five of these PDNGs with human orthologs, no inparalogs or vertebrate
 252 homologs were detected; however, genes with the same name were present in
 253 human and a visual inspection of their syntenic region, including nearby genes,
 254 confirmed that they were present in the human genome (Fig. S2).

255 Orthology relationships of PDNGs that were part of tandem arrays were not
 256 established given that some tandem arrays tend to experience high rates of gene
 257 turnover. Thus, the number of PDNGs with orthologs in human could be higher.
 258 In approximately one third of these PDNGs (36/107) homology to non-rodent

259 genes was uniquely detected through the phmmer search, indicating that a
 260 significant proportion of false positives in de novo surveys will be undetected
 261 using BLAST-only approaches.

262
 263



264
 265
 266
 267
 268
 269
 270
 271
 272
 273
 274
 275
 276
 277
 278
 279

Figure 2. Examples of de novo and 'de novo' rodent genes visualized through the UCSC Genome Browser. (a) An intergenic de novo gene with relatively low syntenic conservation across several non-rodent mammals. (b) A de novo gene (blue asterisk) overlapping with the 3'UTR of an older gene. Notice that the gene symbol is the same for the two genes, which share no coding or protein similarity. (c) Summary of 'de novo' genes features. (d) A 'de novo' gene (*Cd52*, red asterisk) with conserved flanking genes in mouse (top) and human (bottom). (e) A tandem array of keratin-associated genes including three 'de novo' genes (red asterisks). (f) A 'de novo' gene with no syntenic conservation beyond rat. Coding exons, UTRs and introns are shown as thick blue bars, thin blue bars and lines with arrows, respectively. When annotated, alternative transcripts are shown.

280

281 *Similarity analyses to identify paralogous genes in mouse*

282

283 De novo genes can undergo duplication and form novel gene families in a
284 genome. However, it is reasonable to consider the formation of large gene
285 families of de novo genes unlikely in the relatively short evolutionary time period
286 since mouse-rat divergence, which occurred as late as ~12 million years (my)
287 ago (Kimura et al. 2015). Perhaps more noteworthy, a conservative approach
288 should be applied to the study of de novo genes by discarding candidates with
289 paralogous genes in the same genome, or inparalogs. To assess the frequency
290 of PDNGs with inparalogs, I performed similarity searches based on tBLASTn,
291 BLASTn, and phmmer. Using the same threshold commonly applied in
292 phylostratigraphy studies with BLAST, namely a maximum e-value of 0.001, and
293 stringent criteria for phmmer results (see Methods), I found that 110 genes (29%
294 of all PDNGs) have at least one paralogous gene in GENCODE M16, RefSeq, or
295 UCSC known gene murine data sets (Fig. 2d; Table 2). Non-genic homologous
296 sequences to PDNGs were also found using BLASTn searches against the
297 mm10 mouse genome assembly, but were not included in the validation of
298 PDNGs.

299

300 Fifty-four 'de novo' genes clustered in 20 tandem arrays, defined as groups of 'de
301 novo' genes less than 100kb apart (Fig. 2e; Table S1). At least one gene pair
302 from each array was found using a 10kb distance cutoff. Several lines of
303 evidence suggested that these arrays were not entirely formed by de novo
304 genes. Many arrays contained other paralogs that were not annotated as PDNGs
305 in the first place (Fig. 2e). Moreover, PDNGs in some arrays belonged to known
306 gene families present in other mammals. For instance, two arrays contained
307 keratin-associated genes and another one was found within a cluster of defensin
308 genes. Finally, most 'de novo' genes in arrays (37/54) showed no synteny
309 conservation with other mammals, as expected in the presence of lineage-
310 specific duplications rather than de novo gene formation. This finding
311 underscores the importance of implementing more rigorous homology searches
312 *within* the focal genome in de novo gene studies in order to remove false
313 positives due to young gene duplicates.

314

315

316 *Synteny analyses*

317

318 Synteny information in de novo gene investigations is crucial to detect enabler
319 substitutions by comparing putative novel coding sequences with noncoding

320 orthologous regions in sister taxa (McLysaght and Guerzoni 2015; Vakirlis et al.
321 2017). To assess synteny conservation of PDNGs, I used genome-wide
322 alignment data available throughout the Galaxy portal. Approximately 34%
323 PDNGs (131/381) exhibited no synteny conservation in a 60-vertebrate
324 alignment, which includes 40 mammalian genomes (Fig. 2f; Table S1). It is
325 arguable that some of these 'de novo' genes with no synteny information may
326 represent true de novo genes that evolved in genomic region that have been lost
327 through deletions in non-rodent mammals. However, this is unlikely given that the
328 synteny information relies on alignments of a large number of mammalian
329 genomes (Blankenberg et al. 2011). Given the phylogenetic distribution of the
330 species present in the multialignment, loss of syntenic regions should have
331 occurred independently in no less than three mammalian lineages. This also
332 doesn't account for possible further losses of synteny within glires (lagomorphs
333 and rodents), which were not assessed in this study.

334
335 Many 'de novo' genes with no apparent orthologs outside rodents could
336 represent lineage-specific copies of older parent genes. Indeed, 76/131 'de novo'
337 genes with no synteny conservation shared homology with at least one inparalog.
338 Some of the remaining 55 'de novo' genes with no synteny conservation may
339 constitute rapidly evolving young gene duplicates with no detectable homology
340 with their parent genes; for example, four of them belong to tandem arrays. As
341 expected, the majority of 'de novo' genes with conserved synteny across
342 mammals (76/98) shared homology with genes found outside rodents.

343
344

345 *Conserved domains in PDNGs*

346

347 Conserved domains were found in protein sequences encoded by 39 PDNGs
348 (Table S1). As expected, some of these domains belong to gene families
349 identified in tandem arrays, such as defensin and keratin. Some conserved
350 domains were not functionally characterized; for instance, two PDNGs encoded
351 peptides with DUFs (domains of unknown function) and two other proteins
352 contained a proline-rich domain. Arguably, these domains might belong to novel
353 proteins present in multiple rodents. However, all PDNGs encoding proteins with
354 less well-characterized domains failed to pass one or multiple other criteria to be
355 considered valid 'de novo' genes (Table S1).

356
357

358 *New estimates of rodents de novo genes*

359

360 The results of homology searches and synteny analysis showed that only 152 of
361 the 381 PDNGs (~40%) annotated as protein coding genes represent de novo
362 genes (Fig. 2a-b). Notably, 70/152 (46%) of these PDNGs were automatically
363 annotated and might represent pseudogenes (Table 1). Excluding these genes
364 from the annotation list brings down the number of validated PDNGs to 85/314
365 (~27%). Moreover, this set of 152 de novo genes is likely including several 'de
366 novo' cases for three reasons. First, the criteria applied in the homology
367 searches were particularly stringent. Non-genic paralogous sequences were
368 excluded from inparalog searches and I required at least two non-rodent species
369 to show significant similarity with PDNGs to identify 'de novo' genes. A stringent
370 threshold was also applied in the HMMER searches (see Methods). Second,
371 several PDNGs showed synteny with non-rodent genomes for as little as 10% of
372 their coding regions. Third, enabler substitutions have not been searched for in
373 the N2013 and W2017 PDNG sets (Neme and Tautz 2013; Wilson et al. 2017).
374 Some de novo genes are also likely absent in the three studies analyzed here.
375 Accordingly, analyses of lncRNAs in human and mouse have shown several
376 potential protein coding de novo genes that were not been reported before (Xie
377 et al. 2012; Chen et al. 2015; Ruiz-Orera et al. 2015; Neme and Tautz 2016).

378
379 Overall, errors in de novo gene detection depended on lack of synteny (131
380 genes, 34% of PDNGs), presence of mouse paralogous genes (110 genes,
381 29%), and/or homology with genes from at least two non-rodent vertebrates (98
382 genes, 26%), as summarized in Table 2. In line with results from the re-analysis
383 of budding yeast PDNGs (Carvunis et al. 2012; Moyers and Zhang 2016), these
384 findings call for implementing more rigorous strategies to validate PDNGs
385 identified through phylostratigraphy using a combination of synteny analysis and
386 extensive sequence similarity searches. Both strategies are readily implemented
387 using existing databases and software, particularly in model taxa with established
388 synteny data. In non-model organisms, validation of PDNGs is more problematic
389 given the paucity of multiple closely related genomes and genome-wide
390 alignments necessary to retrieve synteny information.

391
392 The number of confirmed de novo genes varies significantly across the three
393 studies. Seven out of ten still annotated PDNGs from the M2012 paper were
394 validated in this re-analysis, compared to 139/331 (42%) and 18/72 (25%)
395 PDNGs reported in the N2013 and the W2017 studies, respectively. The two
396 latter works were based on a phylostratigraphy-only approach, whereas Murphy
397 and McLysaght integrated phylostratigraphy with an analysis of synteny and
398 enabler substitutions (Murphy and McLysaght 2012). The ten annotated PDNGs
399 all showed enabler substitutions (Murphy and McLysaght 2012). In spite of a

400 large number of detected false positives, that is, 'de novo' genes, a significant
401 difference remains in the number of validated de novo genes identified in the
402 three studies. This is especially striking in the M2012 and N2013 studies, which
403 focused on mouse-specific genes. Two factors seem to have contributed to the
404 observed discrepancy between these works. First, the two analyses relied on
405 different versions of the mouse Ensembl gene set, v56 and v66. Although both
406 versions have been discontinued, the closest available datasets from v54 and
407 v67 differ significantly in the number of annotated protein sequences (v54:
408 40,341; v67: 80,007). However, 108/117 validated de novo genes from the
409 N2013 dataset were already present in the mouse Ensembl v54 proteome. More
410 importantly, Murphy and McLysaght developed a pipeline incorporating several
411 stringent filtering steps that appear to be absent in the Neme and Tautz work
412 (Murphy and McLysaght 2012; Neme and Tautz 2013). Specifically, Murphy and
413 McLysaght analyzed only mouse PDNGs with orthologous noncoding regions in
414 rat and showed experimental evidence of both transcription and translation. None
415 of these criteria were applied in the Neme and Tautz study. Therefore, the N2013
416 PDNGs set contains genes with weak annotation support and genes lacking
417 synteny data. Thus, the 142 validated de novo genes from the N2013 study
418 represent an upper boundary of the number of potential mouse de novo genes,
419 with the caveats that some of them might be present also in the rat genome, but
420 could not be detected in BLAST searches due to high levels of divergence. More
421 than $\frac{3}{4}$ of PDNGs from the W2017 study have been reclassified as 'de novo'
422 genes in this study. The majority of these genes (43/54) showed significant
423 homology with other vertebrates. Seven of them correspond to human functional
424 orthologs and the pseudogene *Snhg11* (Fig. 2d; Fig. S2).

425

426

427 *Rate of de novo gene formation in mouse*

428

429 In their 2013 paper, Neme and Tautz identified 780 mouse-specific putative de
430 novo genes that, given their apparent absence in rat, would have emerged in the
431 past ~12 million years (Kimura et al. 2015). This corresponds to a rate of ~65
432 genes/my, which is similar to mouse-specific gene duplications estimates of 63
433 genes/my obtained comparing mouse and Brown Norway rat (Gibbs et al. 2004)
434 and 106 genes/my assessed using a phylogeny of 10 complete mammalian
435 genomes (Marmoset Genome 2014). To re-calculate the rate of de novo gene
436 formation in mouse according to my analysis, I first determined the orthology of
437 the 152 validated de novo genes in the rat genome. Only thirteen de novo genes
438 shared similarity to rat proteins. Thus, 139/152 validated de novo genes appear
439 to be mouse-specific. This result implies that the maximum rate of de novo gene

440 formation during the mouse lineage evolution correspond to ~11.6 gene/my,
441 assuming a mouse-rat divergence time of ~12 million years. This indicates that
442 de novo genes in mouse emerged at a pace that is at least about 5.4-9.1 times
443 slower compared to gene duplicates. Notably, a recent study has shown that de
444 novo genes originated at only ~2.1 gene/million year in the great apes (Guerzoni
445 and McLysaght 2016). Lower numbers of de novo genes were identified in
446 mouse by one of these authors in the M2012 paper, suggesting that
447 methodological differences might be largely responsible for the discrepancy in
448 the estimates of de novo gene formation between rodents and primates.

449

450

451 *Characteristics of mouse de novo genes*

452

453 The 152 validated de novo genes can be divided in two groups according to the
454 quality of their annotation. The manually annotated group contained 82 de novo
455 genes, compared to 70 automatically annotated genes (Table S2). I will refer to
456 these two groups as MA and AA, respectively. Gene length and number of exons
457 all increased significantly from the AA group to the MA group and for both of
458 them in comparison to 20,391 other mouse transcripts (Table S3). These
459 features have been observed in previous de novo gene studies, including some
460 of the data re-examined here (Murphy and McLysaght 2012; Neme and Tautz
461 2013; Guerzoni and McLysaght 2016).

462

463 Seventy-one de novo genes overlapped with coding exons (20), 5'UTRs (19),
464 3'UTRs (9) or introns (27) of older genes (Table S4). Except for the 3'UTR cases,
465 most de novo genes overlapped on the opposite strand of the older gene (Table
466 S4). Given that coding exons occupy only ~1% of mammalian genomes, the
467 occurrence of twenty de novo genes in overlap with coding exons is especially
468 notable. The emergence of novel ORFs on the complementary strand of
469 preexisting genes, known as overprinting, is relatively common in viruses but is
470 considered rare among eukaryotes (Pavesi et al. 2013). In bacteria, long ORFs
471 tend to be present on the opposite strand of genes (Yomo and Urabe 1994). The
472 presence of widespread long ORFs on the opposite strand of mammalian coding
473 exons could thus accelerate the origin of de novo genes through overprinting.

474 Overall, de novo genes tend to overlap with other genes almost six times more
475 often than older genes (46.7% vs. 8.4%; $P < 0.00001$, Fisher exact test). This
476 tendency has been documented in rodents and primates (Murphy and McLysaght
477 2012; Neme and Tautz 2013; Ruiz-Orera et al. 2015; Guerzoni and McLysaght
478 2016) and implies that the evolution of de novo genes may be facilitated near
479 older genes due to the high density of regulatory motifs, open chromatin and

480 elevated GC content (McLysaght and Hurst 2016). The evolution of de novo
481 genes should be facilitated in genomic regions with elevated GC content
482 because they tend to harbor fewer AT-rich stop codons (Oliver and Marin 1996).
483 Some of these features are also associated with de novo gene formation in
484 intergenic regions (Vakirlis et al. 2017). Long noncoding RNAs (lncRNAs) also
485 appear to represent another source of de novo genes, possibly because they are
486 associated with transcriptionally active regions (Xie et al. 2012; Chen et al. 2015;
487 Ruiz-Orera et al. 2015; Guerzoni and McLysaght 2016; Neme and Tautz 2016).

488
489

490 *Levels of intrinsic disorder in de novo genes and older genes*

491

492 It has been argued that de novo genes encoding for proteins that show low
493 propensity to form aggregates, and thus are less prone to induce cytotoxicity,
494 should be more likely to be fixed (Wilson et al. 2017). Wilson and colleagues
495 calculated the intrinsic structural disorder (ISD), a proxy for protein solubility
496 (Monsellier and Chiti 2007; Pallares and Ventura 2016), in all mouse proteins
497 and found that: 1) de novo genes showed the highest level of ISD, which
498 suggested they were preadapted to become novel genes because they encode
499 proteins with low tendency toward aggregation; 2) ISD levels increased
500 throughout mouse genes phylostrata.

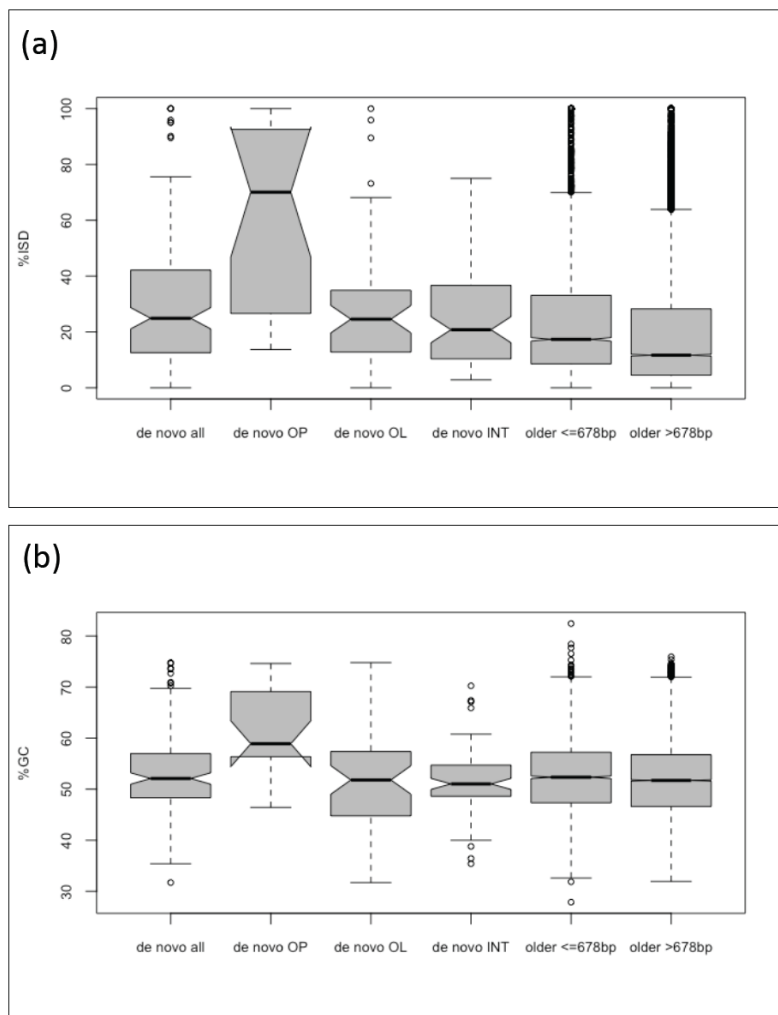
501

502 Here, I calculated ISD levels for the 152 validated rodent de novo genes and
503 20,391 older mouse genes using the algorithm implemented in the software
504 PASTA (Walsh et al. 2014). Validated de novo genes showed a significantly
505 higher proportion of ISD regions than older genes, including genes with
506 comparable length ($P < 0.0001$, Mann-Whitney U test. Fig. 3a; Table S3).
507 However, this derives from the particularly high levels of disorder in proteins
508 encoded by the twenty overprinted de novo genes. ISD levels are significantly
509 higher in these proteins compared to any other group of de novo or older proteins
510 (all $P < 0.002$, Mann-Whitney U test). On the contrary, I found no significant
511 difference in ISD levels between proteins from short older genes and proteins
512 encoded by either intergenic or overlapping but not overprinted de novo genes (P
513 = 0.051 and 0.105, respectively, Mann-Whitney U test).

514

515 Recent works have shown that high disorder levels in orphan and de novo
516 proteins are associated with the elevated GC content of their genes (Basile et al.
517 2017; Vakirlis et al. 2017). In agreement with these findings, overprinted de novo
518 genes showed significantly higher %GC compared to any other group of genes
519 (all $P < 0.0004$, Mann-Whitney U test; Fig. 3b; Table S3), contrary to intergenic or

520 overlapping but not overprinted de novo genes. Furthermore, the GC content and
521 ISD levels were highly correlated in the complete set of analyzed mouse genes (r
522 = 0.92, Pearson correlation). The particularly elevated GC content in overprinted
523 de novo genes can be explained by several factors. As already mentioned, the
524 diminished frequency of stop codons in GC-rich regions allows longer ORFs to
525 form. Additionally, the GC content is positively correlated with the transcriptional
526 activity in mammalian cells (Kudla et al. 2006), which could increase the
527 likelihood of proto-genes to be spuriously expressed and eventually evolve into
528 functional genes.
529
530



531
532
533 **Figure 3.** Comparison of intrinsic structural disorder percentage (a) and GC content (b)
534 between de novo genes and older genes. Gray boxes shows values between first and
535 third quartile. Medians are shown as black lines. Whiskers: minimum and maximum
536 values excluding outliers. OP: overprinting. OL: overlapping with non-coding regions of
537 older genes. INT: intergenic.

538

539

540 **Conclusions**

541 The discovery of de novo genes in eukaryotes has revealed how evolutionary
542 tinkering of noncoding regions can lead to novel protein sequences from scratch.
543 Previous analyses relying uniquely on phylostratigraphic methods suggested that
544 de novo genes are fixed at rates comparable to those of gene duplicates
545 (Carvunis et al. 2012; Neme and Tautz 2013). This conclusion cannot be
546 reconciled with the observed levels of interspecific gene homology and gene loss
547 rates, what I referred to as the 'gene turnover paradox'. Here, I used data from
548 three previous studies to show that the majority of putative de novo genes thus
549 far detected in rodents and still annotated in mouse represent either lineage-
550 specific gene duplicates or rapidly evolving genes shared across mammals. The
551 improved estimate of mouse-specific de novo genes points to a rate of novel
552 gene formation that is several times lower than the gene duplication rate, a
553 possible resolution of the 'gene turnover paradox'. Importantly, these results also
554 imply that the known homology detection bias in phylostratigraphy is *not*
555 minimized by focusing on the youngest genes in a given species, as previously
556 suggested (Wilson et al. 2017). However, as shown in this and other studies
557 (Murphy and McLysaght 2012; Vakirlis et al. 2017), false positive rates in de
558 novo gene surveys can be significantly reduced by utilizing a combination of
559 more sensitive homology search approaches and synteny analyses.

560

561 In one of the re-examined studies, Wilson and co-authors (2017) found that
562 putative de novo proteins have the highest levels of intrinsic structural disorder
563 (ISD), a measure that negatively correlates with protein toxicity, among mouse
564 proteins. This would suggest that de novo genes evolve more frequently from
565 proto-genes that are preadapted because they encode peptides with low level of
566 toxicity. Mouse de novo proteins validated in my study also show higher ISD
567 levels than older genes; however, I found that this is due to a subset of de novo
568 genes that share high GC content and overlap with coding exons of older genes.
569 In agreement with recent observations (Basile et al. 2017; Vakirlis et al. 2017),
570 this shows that the elevated disorder of mouse de novo proteins represent a
571 mere consequence of the high %GC of some de novo genes, rather than
572 supporting the preadaptation hypothesis.

573

574

575 **Methods**

576

577 *Putative de novo genes*

578

579 **Murphy and McLysaght 2012:** Mouse de novo gene IDs were retrieved from
580 Table 1 of the Murphy and McLysaght study (Murphy and McLysaght 2012) .
581 These genes were found using the Ensembl version 56. Protein-coding genes
582 from the two closest available Ensembl versions, v54 and v67, were downloaded
583 from the Ensembl archives
584 (<https://www.ensembl.org/info/website/archives/index.html>). Out of the 69
585 putative mouse de novo genes, only twenty-six were still annotated as protein-
586 coding genes in v67.

587

588 **Neme and Tautz 2013:** Neme and Tautz identified de novo gene using the
589 mouse Ensembl version 66. We retrieved all the 80,007 mouse transcript and
590 protein IDs and sequences annotated in the closest available data set, the
591 archived Ensembl version 67. We found a match for 779 out of 780 mouse
592 putative de novo genes in the Ensembl v67 version and selected the longest
593 protein isoform for these genes for subsequent analyses. Six PDNGs were
594 removed from the 779 gene set after applying a minimum protein length
595 threshold of 30 amino acids leading to a total of 773 analyzed PDNGs in the
596 N2013 data set. The presence of N2013 PDNGs in the mouse Ensembl v54
597 proteome was determined using a tBLASTn search with an evaluate threshold of
598 0.001. Hits that shared at least 90% sequence identity over at least half of the
599 query were considered orthologous sequences.

600

601 **Wilson et al. 2017:** Ensembl gene IDs and sequences of the 84 mouse young
602 genes were obtained from the supplementary table 2 of the Wilson et al. paper
603 (Wilson et al. 2017). Transcript and protein IDs and sequences annotated in the
604 closest available data set, the archived Ensembl version 75
605 (<https://www.ensembl.org/info/website/archives/index.html>), the same used in the
606 W2017 paper.

607

608

609 *Updated annotation of PDNGs*

610

611 The UCSC Genome Browser and Table Browser have been used to retrieve
612 genome coordinates of PDNGs from the three papers' datasets
613 (<http://genome.ucsc.edu/cgi-bin/hgTables>). However, the Ensembl gene track
614 from the most recent mouse genome assembly (GRCm38/mm10, December
615 2011) does not contain all Ensembl IDs corresponding to PDNGs. Genome
616 coordinates of PDNGs were therefore retrieved using the previous mouse
617 genome assembly (NCBI37/mm9, July 2007). These coordinates were then

618 transformed into coordinates of the mouse mm10 assembly using the LiftOver
619 tool (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>). The genome coordinates of
620 most PDNG coding exons were successfully lifted to the mm10 assembly and
621 uploaded to the Galaxy portal (<https://usegalaxy.org>). Genome coordinates of the
622 coding exons of mouse RefSeq, GENCODE M16 (same gene set as Ensembl
623 v91) and UCSC 'known' genes were also uploaded to Galaxy. PDNG coding
624 exons were then joined using the Galaxy tool 'Join' in the Menu 'Operate on
625 Genomic Intervals' applying 'All records of first dataset' to return their overlap
626 with coding exons of each of the three gene sets with a minimum of 50 bp
627 overlap (all PDNGs had at least one coding exon longer than 50bp). Coding
628 exons of 381 PDNGs overlapped with coding exons of at least one gene from the
629 three used gene sets. Notice that overlapping genes on opposite strands of
630 PDNGs were not excluded. PDNG whose IDs were not available in the mm9
631 Ensembl track were re-annotated by querying their coding sequences against the
632 mm10 genome assembly using blat (<http://genome.ucsc.edu/cgi-bin/hgBlat>) to
633 visually find matches with annotated GENCODE M16, RefSeq and UCSC known
634 genes. The M2012 and W2017 PDNGs were visually inspected on the UCSC
635 Genome Browser for overlap with annotated genes. Annotation was confirmed
636 for 9/69 M2012 PDNGs (Table 1).

637

638

639 *Sequence similarity analyses to identify paralogs*

640

641 Several types of BLAST searches on multiple mouse databases were carried out
642 using a consistent e-value threshold of 0.001. The mouse genome assembly
643 mm10 was searched locally using tBLASTn (Camacho et al. 2009). In searches
644 against the mouse genome only hits against the coding sequence of known
645 genes were considered valid paralogous genes of putative de novo genes. The
646 combined mouse proteomes from GENCODE M16 genes (Harrow et al. 2012),
647 the RefSeq genes (O'Leary et al. 2016) and the UCSC Genome Browser 'known
648 genes' (<http://genome.ucsc.edu/cgi-bin/hgTables>) databases were searched
649 locally using BLASTP. BLAST results were parsed and filtered using perl scripts
650 and Unix commands. Matches over less than 50% of the query sequence were
651 removed to increase stringency. Matches of PDNGs with multiple proteins were
652 carefully inspected to ensure that they corresponded to multiple loci rather than
653 alternative transcripts of the same gene. Alignments of PDNGs with a single
654 match were also inspected to determine if these hits represented paralogous
655 genes rather than self-hits.

656

657 Similar searches were performed locally using the algorithm phmmer in the
658 HMMER suite (<http://hmmmer.org/>). Each PDNG protein set was queried against
659 the three combined GENCODE, RefSeq and USCS proteomes using default
660 parameters. Results were visually inspected to identify matches between protein
661 sets. Hits with c-Evalue and i-Evalue below 0.001 were considered positive
662 matches.

663

664

665 *Sequence similarity analyses to identify homologs in vertebrates*

666

667 The vertebrate (taxid:7742) NCBI nr protein database was interrogated between
668 September 2017 and January 2018 in the NCBI BLAST portal
669 (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) using default settings except an e-value
670 threshold of 0.001, excluding Rodents (taxid:9989). The reference proteomes
671 database (https://www.ebi.ac.uk/reference_proteomes) was interrogated
672 between September 2017 and January 2018 using phmmer with a higher than
673 default stringency e-value of $1e^{-05}$ and excluding rodents from the search
674 (<https://www.ebi.ac.uk/Tools/hmmer/search/phmmer>). Only PDNG proteins with
675 significant hits with proteins from at least two vertebrates were considered
676 positive matches in both BLAST and HMMER searches.

677

678

679 *Synteny analyses*

680

681 Mouse genome coordinates in BED format of the PDNGs coding exon were
682 retrieved from the UCSC Genome Browser table browser tool
683 (<http://genome.ucsc.edu/cgi-bin/hgTables>) using either Ensembl identifiers or
684 novel RefSeq/UCSC identifiers from the re-annotation of the three datasets
685 (Tables S1-2). The BED coordinates were then sent to the Galaxy portal
686 (<https://usegalaxy.org>).

687

688 I generated a workflow on Galaxy ([https://usegalaxy.org/u/claudiocasola/w/maf-](https://usegalaxy.org/u/claudiocasola/w/maf-blocks-for-mouse-mm10-sequences)
689 [blocks-for-mouse-mm10-sequences](https://usegalaxy.org/u/claudiocasola/w/maf-blocks-for-mouse-mm10-sequences)) to obtain MAF blocks (Multiple Alignment
690 Format blocks) from aligned sequences in the mouse genome assembly mm10
691 (Blankenberg et al. 2011). Briefly, the workflow utilizes genome coordinates to
692 extract MAF blocks from the 100-way multiZ alignment based on the human
693 genome assembly hg19. Overlapping MAF blocks were merged, filtered to retain
694 only mouse blocks, and joined to the coordinates of each coding exon of the
695 putative de novo genes. A few remaining overlapping MAF blocks were manually
696 removed from the MAF datasets.

697

698

699 *Protein domain analyses*

700

701 The NCBI Conserved Domain repository (Marchler-Bauer et al. 2017) was
702 interrogated with proteins encoded by PDNGs between 09-2017 and 01-2018
703 using default parameters except inclusion of retired sequences in the batch
704 search portal (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>).
705 Conserved domains of PDNGs with no evidence paralogy and lack of synteny
706 were also searched throughout the InterPro server
707 (<https://www.ebi.ac.uk/interpro/>) in March 2018.

708

709

710 *Gene structure*

711

712 Gene length, coding length, intron length, UTRs length and exon number of
713 mouse genes were obtained from the GENCODE M16 dataset through the
714 UCSC Table Browser. The 64,506 transcripts were filtered to remove noncoding
715 sequences and genes with only automatic annotation. Transcripts matching
716 PDNGs were also removed. All except the shortest transcripts of the remaining
717 genes were removed, leaving 20,470 genes.

718

719

720 *Quality of gene annotation*

721

722 The annotation quality of validated de novo genes was assessed by retrieving
723 data from the GENCODE M16 Basic gene set and the UCSC known genes from
724 the UCSC Table Browser. GENCODE transcript support levels range from 1 (all
725 splice junctions of the transcript are supported by at least one non-suspect
726 mRNA) to NA (the transcript was not analyzed). Additionally, high-quality
727 GENCODE genes are manually annotated in HAVANA
728 (<https://www.gencodegenes.org/gencodeformat.html>). Fourty-two de novo genes
729 with either no HAVANA ID or transcript support level equal NA were considered
730 low-quality genes. Similarly, 31 UCSC known transcript that has not been
731 reviewed or validated by the RefSeq, SwissProt or CCDS staff were considered
732 low quality.

733

734

735 *Protein disorder and protein aggregation analyses*

736

737 The software PASTA 2.0 (Walsh et al. 2014) with default settings was used to
738 estimate intrinsic structural disorder (ISD) in proteins encoded by the three
739 PDNG sets and 20,391 non-de novo Ensembl v91 proteins.

740

741

742 *Estimates of gene duplication rates*

743

744 Gene duplication events estimated to have occurred in mouse since its
745 divergence from the Brown Norway rat have been obtained from Worley et al.
746 (2014). Gene duplications and losses were modeled using the maximum-
747 likelihood framework implemented in the CAFE package (Han et al. 2013). A total
748 of 1,052 mouse-specific gene duplications were calculated based on gene family
749 data totaling 18,215 genes (supplementary figure 7 in (Marmoset Genome
750 2014)). Assuming ~22,000 genes in the mouse genome, the actual overall
751 number of gene duplicates is 1,275, leading to a rate of duplication of ~106/my.

752

753

754 *Overlap between genes*

755 Each validated de novo genes was visually inspected through the UCSC
756 Genome Browser to identify possible overlap with other genes. To find the
757 genome-wide proportion of overlapping genes, transcripts from all genes in the
758 mouse GENCODE M16 basic gene set were downloaded from the UCSC Table
759 Browser. Transcripts from the mitochondrial genome, non-protein coding
760 transcripts and transcripts from de novo genes were removed. For each
761 remaining gene, only the longer transcript was retained, leaving a total of 22,396,
762 of which 1876 (~8.4%) overlapped other genes.

763

764

765 *Rat de novo gene orthologs*

766

767 The genome coordinates of the 152 validated de novo genes from the mouse
768 assembly mm10 were used to retrieve syntenic regions in the rat rn6 assembly
769 with the LiftOver tool in the UCSC Genome Browser. A total of 29,107 Ensembl
770 and transcript coding regions were downloaded using the UCSC Table Browser
771 (Data last updated: 2017-06-09). Proteins and CDS were searched against the
772 retrieved rat genomic regions syntenic with mouse de novo genes using
773 tBLASTn, e-value threshold = 0.001. The BLAST results were parsed using an
774 in-house perl script and filtered to retain hits longer than 30bp and with at least
775 97% DNA sequence identity between CDS and genome. This step left 67
776 putative orthologous proteins to mouse de novo genes. Some of these proteins

777 were orthologous to proteins that in mouse overlap to the validated de novo
778 genes. Thus, I manually inspected these proteins against the mouse mm10
779 assembly using BLAT in the UCSC Genome Browser and by running a BLASTP
780 search between the 152 mouse de novo proteins and the 67 candidate rat
781 orthologs. The same approach was used to retrieve 17,619 rat RefSeq proteins
782 and CDS. I obtained 168 candidates that were screened against the 133 mouse
783 validated de novo genes. Additionally, I searched the combined 235 candidate
784 Ensembl and RefSeq proteins against the 152 mouse de novo proteins using
785 phmmer locally with default settings.

786

787

788 **Acknowledgments**

789

790 I am grateful to Aaron Quinlan and Ryan Layer for allowing me to use the term
791 'de novo'. I thank Michelle Lawing for help with statistical analyses and for
792 comments on the manuscript. This work has been supported by the National
793 Institute of Food and Agriculture, U.S. Department of Agriculture, under award
794 number TEX0-1-9599, the Texas A&M AgriLife Research, and the Texas A&M
795 Forest Service.

796

797

798 **References**

799

- 800 Basile W, Sachenkova O, Light S, Elofsson A. (2017). High GC content causes
801 orphan proteins to be intrinsically disordered. *PLoS Comput Biol*, **13**(3),
802 e1005375.
- 803 Begun DJ, Lindfors HA, Kern AD, Jones CD. (2007). Evidence for de novo
804 evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila*
805 *erecta* clade. *Genetics*, **176**(2), 1131-1137.
- 806 Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ, the Mouse
807 Genome Database G. (2017). Mouse Genome Database (MGD)-2017:
808 community knowledge resource for the laboratory mouse. *Nucleic Acids*
809 *Res*, **45**(D1), D723-D729.
- 810 Blankenberg D, Taylor J, Nekrutenko A, Galaxy T. (2011). Making whole genome
811 multiple alignments usable for biologists. *Bioinformatics*, **27**(17), 2426-
812 2428.
- 813 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden
814 TL. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*,
815 **10**421.

- 816 Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N,
817 Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. (2012). Proto-
818 genes and de novo gene birth. *Nature*, **487**(7407), 370-374.
- 819 Chen JY, Shen QS, Zhou WZ, Peng J, He BZ, Li Y, Liu CJ, Luan X, Ding W, Li S,
820 et al. (2015). Emergence, Retention and Selection: A Trilogy of Origination
821 for Functional De Novo Proteins from Ancestral LncRNAs in Primates.
822 *PLoS Genet*, **11**(7), e1005391.
- 823 Domazet-Loso T, Brajkovic J, Tautz D. (2007). A phylostratigraphy approach to
824 uncover the genomic history of major adaptations in metazoan lineages.
825 *Trends Genet*, **23**(11), 533-539.
- 826 Elhaik E, Sabath N, Graur D. (2006). The "inverse relationship between
827 evolutionary rate and age of mammalian genes" is an artifact of increased
828 genetic distance with rate of evolution and time of divergence. *Mol Biol
829 Evol*, **23**(1), 1-3.
- 830 Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S,
831 Scott G, Steffen D, Worley KC, Burch PE, et al. (2004). Genome
832 sequence of the Brown Norway rat yields insights into mammalian
833 evolution. *Nature*, **428**(6982), 493-521.
- 834 Guerzoni D, McLysaght A. (2016). De Novo Genes Arise at a Slow but Steady
835 Rate along the Primate Lineage and Have Been Subject to Incomplete
836 Lineage Sorting. *Genome Biol Evol*, **8**(4), 1222-1232.
- 837 Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. (2013). Estimating gene gain
838 and loss rates in the presence of error in genome assembly and
839 annotation using CAFE 3. *Mol Biol Evol*, **30**(8), 1987-1997.
- 840 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F,
841 Aken BL, Barrell D, Zadissa A, Searle S, et al. (2012). GENCODE: the
842 reference human genome annotation for The ENCODE Project. *Genome
843 Res*, **22**(9), 1760-1774.
- 844 Heinen TJ, Staubach F, Haming D, Tautz D. (2009). Emergence of a new gene
845 from an intergenic region. *Curr Biol*, **19**(18), 1527-1531.
- 846 Kimura Y, Hawkins MT, McDonough MM, Jacobs LL, Flynn LJ. (2015). Corrected
847 placement of *Mus-Rattus* fossil calibration forces precision in the
848 molecular tree of rodents. *Sci Rep*, **5**14444.
- 849 Knowles DG, McLysaght A. (2009). Recent de novo origin of human protein-
850 coding genes. *Genome Res*, **19**(10), 1752-1759.
- 851 Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. (2006). High guanine and
852 cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*,
853 **4**(6), e180.
- 854 Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q,
855 Zheng X, et al. (2010). A human-specific de novo protein-coding gene

- 856 associated with human brain functions. *PLoS Comput Biol*, **6**(3),
857 e1000734.
- 858 Lu TC, Leu JY, Lin WC. (2017). A Comprehensive Analysis of Transcript-
859 Supported De Novo Genes in *Saccharomyces sensu stricto* Yeasts. *Mol*
860 *Biol Evol*.
- 861 Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire
862 MK, Geer RC, Gonzales NR, et al. (2017). CDD/SPARCLE: functional
863 classification of proteins via subfamily domain architectures. *Nucleic Acids*
864 *Res*, **45**(D1), D200-D203.
- 865 Marmoset Genome SA, Consortium. (2014). The common marmoset genome
866 provides insight into primate biology and evolution. *Nat Genet*, **46**(8), 850-
867 857.
- 868 McLysaght A, Guerzoni D. (2015). New genes from non-coding sequence: the
869 role of de novo protein-coding genes in eukaryotic evolutionary innovation.
870 *Philos Trans R Soc Lond B Biol Sci*, **370**(1678), 20140332.
- 871 McLysaght A, Hurst LD. (2016). Open questions in the study of de novo genes:
872 what, how and why. *Nat Rev Genet*, **17**(9), 567-578.
- 873 Monsellier E, Chiti F. (2007). Prevention of amyloid-like aggregation as a driving
874 force of protein evolution. *EMBO Rep*, **8**(8), 737-742.
- 875 Moyers BA, Zhang J. (2016). Evaluating Phylostratigraphic Evidence for
876 Widespread De Novo Gene Birth in Genome Evolution. *Mol Biol Evol*,
877 **33**(5), 1245-1256.
- 878 Moyers BA, Zhang J. (2015). Phylostratigraphic bias creates spurious patterns of
879 genome evolution. *Mol Biol Evol*, **32**(1), 258-267.
- 880 Moyers BA, Zhang JZ. (2017). Further Simulations and Analyses Demonstrate
881 Open Problems of Phylostratigraphy. *Genome Biology and Evolution*, **9**(6),
882 1519-1527.
- 883 Murphy DN, McLysaght A. (2012). De novo origin of protein-coding genes in
884 murine rodents. *PLoS One*, **7**(11), e48650.
- 885 Neme R, Tautz D. (2016). Fast turnover of genome transcription across
886 evolutionary time exposes entire non-coding DNA to de novo gene
887 emergence. *Elife*, **5**e09977.
- 888 Neme R, Tautz D. (2013). Phylogenetic patterns of emergence of new genes
889 support a model of frequent de novo evolution. *BMC Genomics*, **14**117.
- 890 O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, Rajput B,
891 Robbertse B, Smith-White B, Ako-Adjei D, et al. (2016). Reference
892 sequence (RefSeq) database at NCBI: current status, taxonomic
893 expansion, and functional annotation. *Nucleic Acids Res*, **44**(D1), D733-
894 745.

- 895 O'Toole AN, Hurst LD, McLysaght A. (2018). Faster Evolving Primate Genes Are
896 More Likely to Duplicate. *Mol Biol Evol*, **35**(1), 107-118.
- 897 Oliver JL, Marin A. (1996). A relationship between GC content and coding-
898 sequence length. *Journal of Molecular Evolution*, **43**(3), 216-223.
- 899 Pallares I, Ventura S. (2016). Understanding and predicting protein misfolding
900 and aggregation: Insights from proteomics. *Proteomics*, **16**(19), 2570-
901 2581.
- 902 Pavesi A, Magiorkinis G, Karlin DG. (2013). Viral proteins originated de novo by
903 overprinting can be identified by codon usage: application to the "gene
904 nursery" of Deltaretroviruses. *PLoS Comput Biol*, **9**(8), e1003162.
- 905 Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. (2013). De
906 novo ORFs in *Drosophila* are important to organismal fitness and evolved
907 rapidly from previously non-coding sequences. *PLoS Genet*, **9**(10),
908 e1003860.
- 909 Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R,
910 Marques-Bonet T, Alba MM. (2015). Origins of De Novo Genes in Human
911 and Chimpanzee. *PLoS Genet*, **11**(12), e1005721.
- 912 Saripella GV, Sonnhammer EL, Forslund K. (2016). Benchmarking the next
913 generation of homology inference tools. *Bioinformatics*, **32**(17), 2636-
914 2641.
- 915 Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D,
916 Miller W. (2003). Human-mouse alignments with BLASTZ. *Genome Res*,
917 **13**(1), 103-107.
- 918 Vakirlis NN, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon
919 JJ, Lafontaine I. (2017). A molecular portrait of de novo genes in yeasts.
920 *Mol Biol Evol*.
- 921 Walsh I, Seno F, Tosatto SC, Trovato A. (2014). PASTA 2.0: an improved server
922 for protein aggregation prediction. *Nucleic Acids Res*, **42**(Web Server
923 issue), W301-307.
- 924 Wilson BA, Foy SG, Neme R, Masel J. (2017). Young Genes are Highly
925 Disordered as Predicted by the Preadaptation Hypothesis of De Novo
926 Gene Birth. *Nat Ecol Evol*, **1**(6), 0146-0146.
- 927 Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, Zhang M, Zhang R, Wei L, Li
928 CY. (2012). Hominoid-specific de novo protein-coding genes originating
929 from long non-coding RNAs. *PLoS Genet*, **8**(9), e1002942.
- 930 Xu J, Zhang J. (2016). Are Human Translated Pseudogenes Functional? *Mol Biol*
931 *Evol*, **33**(3), 755-760.
- 932 Yomo T, Urabe I. (1994). A frame-specific symmetry of complementary strands
933 of DNA suggests the existence of genes on the antisense strand. *J Mol*
934 *Evol*, **38**(2), 113-120.

935 Zhao L, Saelao P, Jones CD, Begun DJ. (2014). Origin and spread of de novo
936 genes in *Drosophila melanogaster* populations. *Science*, **343**(6172), 769-
937 772.
938