


RESEARCH ARTICLE

Open Access



De novo emergence, existence, and demise of a protein-coding gene in murids

Jan Petrzilek^{1,2†}, Josef Pasulka^{1†}, Radek Malik¹, Filip Horvat^{1,3}, Shubhangini Kataruka^{1,4}, Helena Fulka^{1,5} and Petr Svoboda^{1*} 

Abstract

Background: Genes, principal units of genetic information, vary in complexity and evolutionary history. Less-complex genes (e.g., long non-coding RNA (lncRNA) expressing genes) readily emerge de novo from non-genic sequences and have high evolutionary turnover. Genesis of a gene may be facilitated by adoption of functional genic sequences from retrotransposon insertions. However, protein-coding sequences in extant genomes rarely lack any connection to an ancestral protein-coding sequence.

Results: We describe remarkable evolution of the murine gene *D6Erd527e* and its orthologs in the rodent *Muroidea* superfamily. The *D6Erd527e* emerged in a common ancestor of mice and hamsters most likely as a lncRNA-expressing gene. A major contributing factor was a long terminal repeat (LTR) retrotransposon insertion carrying an oocyte-specific promoter and a 5' terminal exon of the gene. The gene survived as an oocyte-specific lncRNA in several extant rodents while in some others the gene or its expression were lost. In the ancestral lineage of *Mus musculus*, the gene acquired protein-coding capacity where the bulk of the coding sequence formed through CAG (AGC) trinucleotide repeat expansion and duplications. These events generated a cytoplasmic serine-rich maternal protein. Knock-out of *D6Erd527e* in mice has a small but detectable effect on fertility and the maternal transcriptome.

Conclusions: While this evolving gene is not showing a clear function in laboratory mice, its documented evolutionary history in *Muroidea* during the last ~40 million years provides a textbook example of how a several common mutation events can support de novo gene formation, evolution of protein-coding capacity, as well as gene's demise.

Keywords: De novo, Gene, Evolution, LTR, Retrotransposon, CAG, Polyserine, D6Erd527e, Oocyte

Background

Some are born to sweet delight

Some are born to endless night

[William Blake, 1863]

After a gene comes into being, it evolves and lasts for a variable period of time. The term gene historically denotes the basic physical and functional unit of heredity [1] but the definition has been evolving along with advancement of knowledge of genomes and transcriptomes (reviewed in [2]). In traditional gene definitions, a molecule encoded by a gene (an RNA or a protein) has a function. However, proving an absence of a function of a molecule encoded by a putative gene is an impossible task. Furthermore, the evolutionary theory implies that genetic traits emerge purposelessly before their function is established by means of natural selection. Thus, any DNA sequence encoding a defined RNA molecule may be considered a gene.

[†]Jan Petrzilek and Josef Pasulka contributed equally to this work.

*Correspondence: svobodap@img.cas.cz

¹Institute of Molecular Genetics of the Czech Academy of Sciences, Videnska 1083, 142 20 Prague 4, Czech Republic

Full list of author information is available at the end of the article



A distinct category of genes are protein-coding genes, where an RNA carries information for protein synthesis. Protein-coding genes usually evolve through some mechanism involving an existing protein-coding sequence as evidenced by clusters of orthologous groups and conserved protein domains [3, 4]. In fact, many extant protein-coding genes are descendants of genes of the last universal common ancestor [5]. In contrast, long-non-coding RNA (lncRNA) genes in complex eukaryotic genomes often emerge from more-or-less random genomic sequences and have rapid evolutionary turnover [6–8]. Notably, lncRNAs and protein-coding genes are not separately evolving discreet gene classes; protein-coding genes can lose the protein-coding capacity and become lncRNAs [9, 10]. At the same time, the coding potential of cytoplasmic lncRNAs is being probed by scanning ribosomes [11]. lncRNAs may occasionally become bona fide protein-coding genes, particularly when a processed pseudogene integrates into an existing lncRNA unit [12]. However, complete de novo emergence of protein-coding capacity in a lncRNA is rare [13].

Emergence of a novel gene from a random sequence does not require much as cryptic promoters, splice sites, and poly(A) sites emerge in a random sequence by a considerable chance. In addition, functional gene parts can be stochastically distributed across a genome by transposable elements [14, 15]. While amplification of transposable elements threatens genome integrity, they can also move around functional gene parts, such as promoters, enhancers, exons, terminators or splice junctions (reviewed in [16, 17]). A common source of functional genic elements are long terminal repeats (LTRs), identical sequences flanking internal sequences of transposable elements (TEs) and retroviruses. 5' LTRs serve as promoters while 3' LTRs provide a functional polyadenylation signal. Significance of functional sequences in LTRs is underscored by the fact that most LTR retrotransposon insertions in mammalian genomes become solo LTRs, which form when homologous recombination recombines out the internal sequence between LTRs [18, 19]. A solo LTR carrying a functional promoter and a poly(A) site is a versatile platform offering several ways how it could shape transcriptional landscape at the insertion site. Indeed, LTRs were often co-opted (exapted [20]) as promoters and exons on hundreds of occasions in the mammalian germline [12, 21] and dozens of cases were documented also in mammalian somatic cells (reviewed in [22]).

A specific feature of the mouse genome evolution has been a repeated expansion of the non-autonomous mammalian apparent LTR retrotransposon (MaLR) group, which generated ~340,000 insertions [23, 24]. Rodent MaLR elements evolved from ancestral MLT family

elements into two families denoted ORR and MT. In the lineage leading to mice, ORR and MT families underwent several amplifications during the last 60 million years, giving rise to specific subfamilies; for example, the MTD subfamily expanded in the mouse genome ~40–50 million years ago (MYA) and pre-dated the MTC subfamily, which expanded 30–40 MYA [23]. MT elements have oocyte-specific expression and their LTRs often carry a conserved splice donor [12]. Thus, an MT LTR essentially carries the first exon, which can be “plugged in” into an existing gene, or can create a novel transcriptional unit in the genome [12, 21]. In rodent germ cells and early embryos, there are hundreds of protein coding genes and lncRNAs utilizing MaLR-derived promoters and first exons [12]. MTD and MTC LTR insertions, which still function as promoters and first exons could be under positive selection. For example, an MTC subfamily LTR insert in *Dicer1* gene drives oocyte-specific isoform of the protein and is essential for normal oocyte function [25].

We previously reported that an MTD subfamily LTR insertion provided the promoter and the first coding exon in a de novo formed protein-coding gene annotated in the mouse genome as *D6Ertd527e* [12]. Here, we provide an extended analysis of evolutionary history of the *D6Ertd527e* locus in *Muroidea* rodents (mice, rats, gerbils, hamsters, voles, and relatives), and we show that the locus offers textbook example of events occurring during a protein-coding gene “life cycle.”

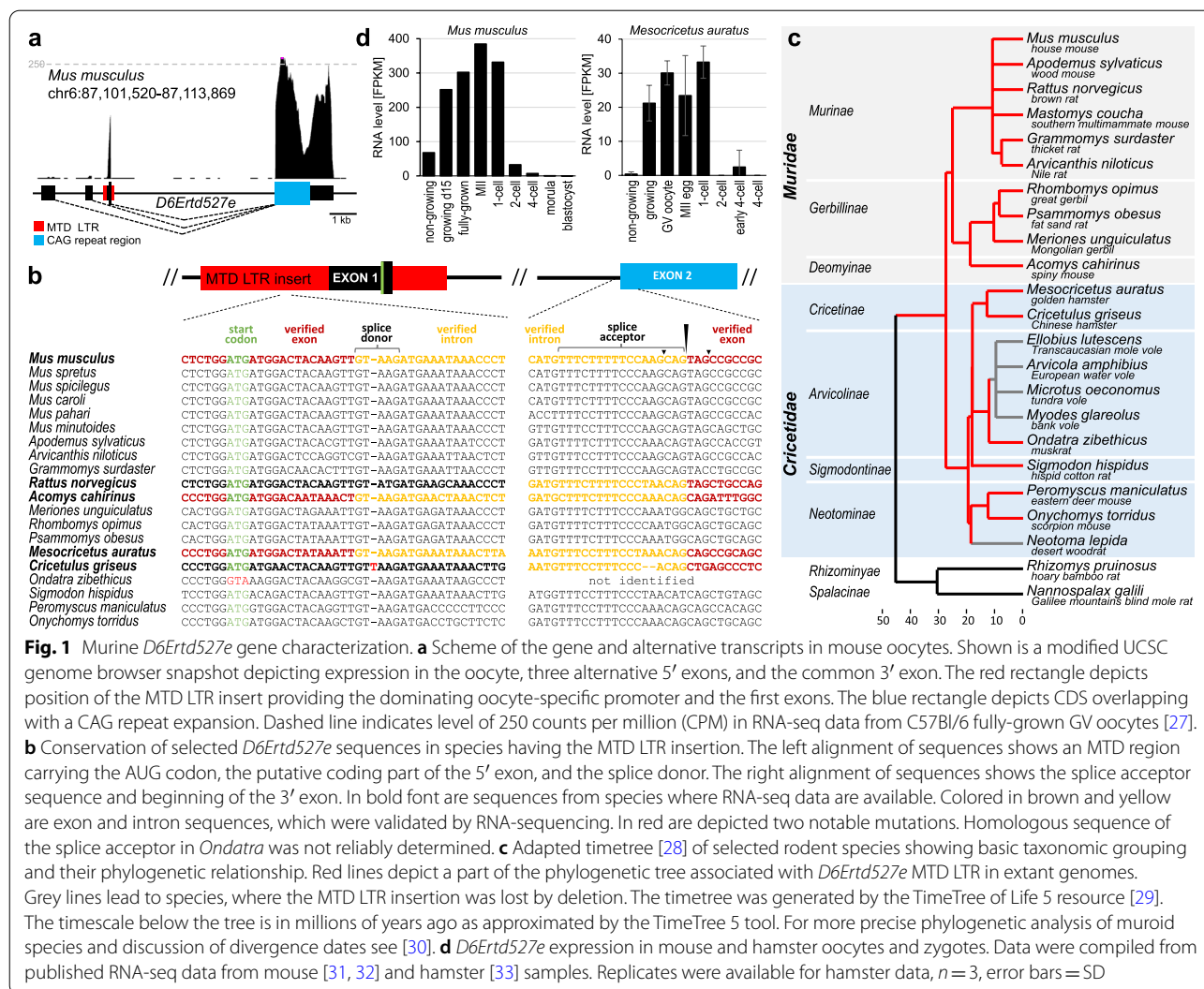
Results

Features of murine *D6Ertd527e*

D6Ertd527e was first annotated as an anonymous expressed DNA segment in mice [26] and later identified as a protein-coding gene expressed in mouse oocytes [12]. *D6Ertd527e* is localized in a syntenic intergenic region between *Gfpt1* and *Atrx* genes (Additional file 1: Fig. S1a): in the central part of the chromosome 6 and its gene structure has two remarkable features (Fig. 1a).

First, the main *D6Ertd527e* promoter and the first exon are an exaptation of an MTD LTR insertion [12]. Sequence conservation in *Muroidea* species suggests that the original MTD LTR insert in the common ancestor of mice and hamsters carried a full 5' terminal exon with an AUG codon (Fig. 1b, c). There are at least two weak upstream promoters with 5' terminal exons (Fig. 1a), but they do not seem to give rise to protein-coding transcripts in mice [12].

Second, *D6Ertd527e* encodes a serine-rich protein and its coding sequence (CDS) largely evolved from an expanding and mutating (CAG)_n repeat. While the (CAG)_n repeat expansion and the coding sequence considerably vary among *Muroidea* genomes, the unique 3'



UTR sequence of *D6Ert527e* is conserved across placental mammals (Additional file 1: Fig. S1b).

Consistent with oocyte-specific expression of MT elements, *D6Ert527e* is maternally expressed [12]. Expression of *D6Ert527e* is detectable in non-growing mouse and golden hamster oocytes and the transcript accumulates during oocyte's growth (Fig. 1d). In both species, *D6Ert527e* mRNA remains stable until the 1-cell stage but is degraded afterwards (Fig. 1d). Consistent with robust maternal expression, RNA-seq data from mouse organs [34] show the highest transcript level of *D6Ert527e* in the ovary (Additional file 1: Fig. S1c). RNA-seq suggested a low expression of *D6Ert527e* in several other organs (testes, intestine, colon, spleen, lung), but none of those transcripts originated from the MTD promoter (Additional file 1: Fig. S1d). Detailed inspection of RNA-sequencing data from the somatic

tissues exhibiting low *D6Ert527e* expression revealed the last exon of *D6Ert527e* can be rarely spliced with upstream *Gfpt1* exons, forming a rare alternative 3' end of *Gfpt1*.

Loss of the MTD promoter of D6Ert527e during evolution

Analysis of MTD insertions suggested that the MTD LTR promoter of *D6Ert527e* was lost at least three times during genome evolution in *Cricetidae* (Additional file 1: Fig. S2). Two distinct deletions were observed in genomes of *Neotominae* subfamily. One deletion was found in *Neotoma lepida* (desert woodrat) and the other one in a deer mouse species *Peromyscus leucopus* (white-footed mouse) while four other deer mouse genomes (*P. maniculatus*, *eremicus*, *californicus*, and *aztecus*) carried the MTD insertion. In the *Arvicolinae* family, all examined species except of *Ondatra zibethicus* (muskrat), carry

the same deletion (Fig. 1c and Additional file 1: Fig. S2). Off note is that *Ondatra* is the only species where the conserved AUG codon from the MTD LTR was lost and where we could not identify the syntenic splice acceptor in the exon 2 (Fig. 1b). These three independent losses in *Arvicolinae* and *Neotominae* imply absence of positive selective pressure on maintaining the MTD LTR promoter in *Cricetidae*.

D6Ertd527e expression in rodent oocytes

D6Ertd527e expression in rodent oocytes could be examined in transcriptomes of five *Muroidea* species. While the maternal transcriptome of *Cricetulus griseus* (Chinese hamster) was newly sequenced, RNA-seq data from *Mus musculus*, *Rattus norvegicus* and *Mesocricetus auratus* (golden hamster) were obtained from the literature [12, 27, 35]. From the *Acomys cahirinus* (spiny mouse) were available only zygotic samples [36], but they allowed for identifying the *D6Ertd527e* transcript. Analysis of RNA-seq data revealed diversity of *D6Ertd527e*

expression levels, functional gene elements, and expression (Fig. 2a). The MTD insert has become the main dominant promoter in *Mus musculus*, *Acomys cahirinus*, and *Mesocricetus auratus*. However, in *Mesocricetus auratus*, three additional promoters also yielded considerable amount of *D6Ertd527e* transcripts (Fig. 2a and Additional file 1: Fig. S3).

Although present in the genome, the MTD LTR promoter was essentially inactive in rat and had minimal activity in Chinese hamster oocytes. In rat oocytes, we have found only a single read in the *D6Ertd527e* locus unambiguously coming from a spliced transcript. But it clearly originated from a different promoter than the MTD LTR promoter. In fact, there was no evidence for transcriptional activity of the MTD LTR promoter (Fig. 2a). This is remarkable considering the presence of the MTD promoter and the fact that house mouse and spiny mouse express *D6Ertd527e* well (Fig. 2a).

In Chinese hamster oocytes, *D6Ertd527e* expression was negligible relative to expression in golden hamster

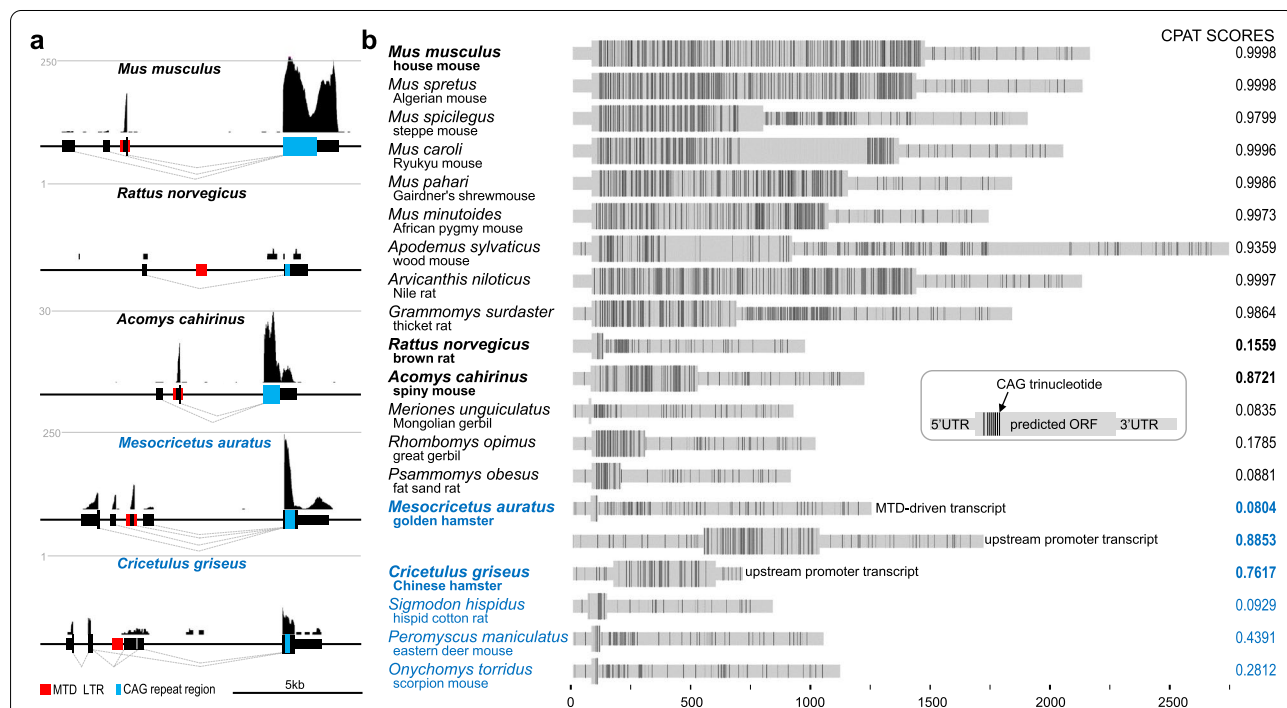


Fig. 2 *D6Ertd527e* transcript variability. **a** Variability of exon–intron structure of *D6Ertd527e* transcripts in oocytes of five different rodent species. Shown are modified UCSC genome browser snapshots depicting distribution of RNA-seq reads, level of expression and exon–intron structures inferred from analysis of spliced individual sequence reads. Position of the MTD LTR insert is indicated by red rectangles. Blue rectangles depict regions containing expanded CAG repeats. Full display of repetitive sequences from Repeatmasker is available in Additional file 1: Fig. S2. Dashed lines indicate normalized expression level in CPMs. *Rattus norvegicus* analysis revealed a single spliced read from > 120 million mapped reads from four independent libraries. **b** Distribution of AGC codons in predicted *D6Ertd527e* transcripts in rodent species carrying the MTD LTR insertion. In case of *Cricetulus griseus*, we used the most abundant transcript isoform transcribed from a promoter upstream of the MTD insert. In case of *Rattus norvegicus*, where the locus seems silent, we show a hypothetical transcript spliced between the conserved splice sites (Fig. 1c) to demonstrate that the putative coding sequence starting from the AUG codon in MTD is soon terminated. CPAT score [37] was calculated for predicted coding sequences represented by the thicker part of a transcript scheme. The recommended cut-off for the mouse coding probability for the CPAT release 3.00 was 0.44 [37]

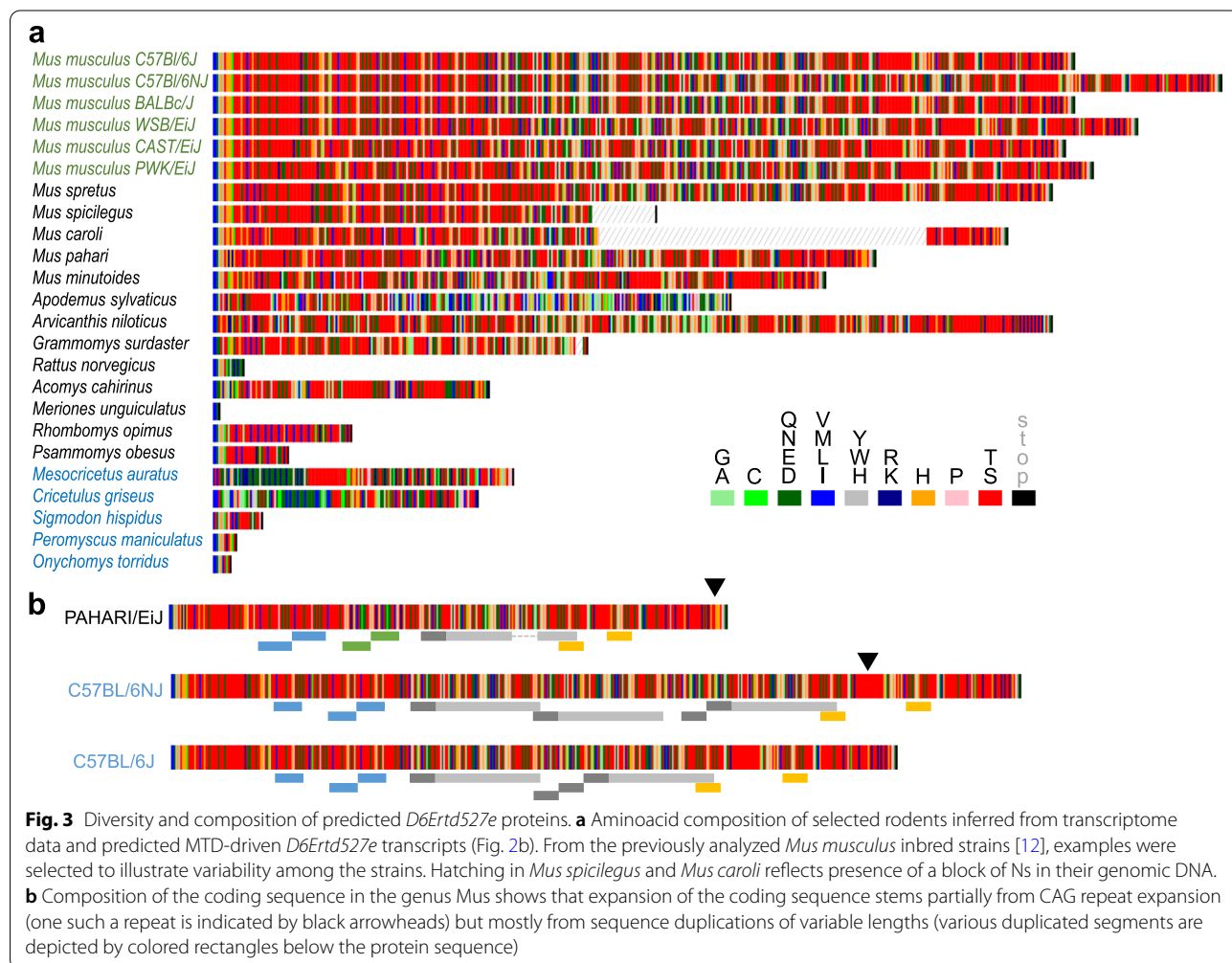
oocytes and originated mainly from a promoter upstream of the MTD LTR promoter (Fig. 2a). Taken together, *D6Ertd527e* expression varies greatly in *Muroidea* oocytes and highest observed levels of *D6Ertd527e* expression are supported by the MTD LTR promoter.

Protein-coding potential of D6Ertd527e

D6Ertd527e transcripts originating from the MTD promoter carry predicted coding sequences of variable lengths and amino acid composition (Fig. 3). MTD LTR carries a conserved AUG initiation codon (in fact, it is AUGAUG, Fig. 1a), which likely came with the original insert in the common ancestor of mice and hamsters, but its significance is difficult to interpret. When analyzing nucleotide-exchange rate of MTD LTR nucleotides (Additional file 1: Fig. S4), the first nucleotide of the first codon in AUGAUG sequence appears to have a higher nucleotide exchange rate than other nucleotides. Consequently, 58% of MTD carry at least 1 AUG: ~14% of MTD LTRs carry the first AUG, ~29% of MTD LTRs have the

second AUG present, and 15% of MTD LTRs carry both. Thus, the persistence of AUG in the *D6Ertd527e* MTD LTR might reflect functional significance. However, six examined species lost the entire MTD insert and two have minimal or no expression. Finally, golden hamster produces four *D6Ertd527e* transcripts differing in the 5' terminal exon and the MTD-driven transcript does have a strong coding potential while one of the other three isoforms has (Fig. 2).

Predicted coding sequences of *D6Ertd527e* transcripts are typically serine-rich as a consequence of (CAG)_n repeat expansion in the last exon. The presence of (CAG)_n clusters in the *D6Ertd527e* locus in *Cricetidae* species indicates that a short (CAG)_n repeat must have been present in the common ancestor of hamsters and mice (Fig. 2). *D6Ertd527e* transcripts have variable (CAG)_n repeat distribution but (CAG)_n repeats typically locate in the protein coding sequence, especially if the coding sequence is longer. In several rodent genomes, such as in *Mus spicilegus* (steppe mouse), *Apodemus sylvaticus*



(wood mouse), and *Grammomys surdaster* (thicket rat), there are large (CAG)_n repeat clusters also in predicted 3' UTRs (Fig. 2B). In the rat genome, the predicted CDS is minimal and most of CAG trinucleotides are downstream of it (Fig. 2b).

The predicted protein-coding sequences of *D6Ertd527e* in extant species reveal dynamic evolution of the protein coding sequence in the *Muroidea* superfamily (Figs. 2b and 3a). There is a highly variable length of the coding sequence even among closely related species (Fig. 2b). Many species (e.g., *Meriones unguiculatus* (Mongolian gerbil), *Rattus norvegicus*, *Onychomys torridus* (scorpion mouse), *Peromyscus maniculaculatus* (eastern deer mouse)) have minimal coding sequences, suggesting that MTD-driven *D6Ertd527e* homologs in these species are not protein coding.

However, in the absence of maternal transcriptome data, it is unclear whether there could be alternative transcripts with longer CDS. This issue is exemplified in Golden hamster where the *D6Ertd527e* locus is well expressed and four different 5' terminal exons are spliced to the 2nd (last) exon (Fig. 2a). Notably, the MTD-driven transcript coding sequence has a short CDS terminated at the beginning of the 2nd exon and is likely non-coding. Two other transcripts from the *D6Ertd527e* do not carry significant CDS as well but the most upstream promoter drives expression of a putative protein-coding transcript with a reasonably large CDS (Fig. 2).

Reading frames in a (CAG)_n repeat encode three possible peptide chains: polyQ, polyS, or polyA. Notably, long *D6Ertd527e* reading frames in *Muridae* species are devoid of frameshifting and generally remain within the polyS frame while their polyQ reading frames are riddled with stop codons. The polyQ reading frame accumulates stop codons naturally because a single point mutation in (CAG)_n can form a stop codon only in the polyQ frame (C to T conversion resulting in TAG). In contrast, single point mutations in polyS or polyA frames of (CAG)_n are either silent or cause an amino acid change. Consequently, the serine-rich, *D6Ertd527e* CDS among *Muridae* species show high divergence of predicted protein coding sequences (Fig. 3a), which stems from a combination of (CAG)_n expansion, duplications/recombination events, and point mutations (Fig. 3b). Variability exists even among laboratory strains, where is particularly remarkable CDS expansion in the C57BL/6NJ strain, which has the longest CDS. C57BL/6NJ is an NIH subline of C57BL/6. It was separated from C57BL/6 J (mouse reference sequence) in 1951 [38] and carries an extra duplicated segment unlike the closely related C57BL/6 J or more distant BALB/cJ (Fig. 3a, b).

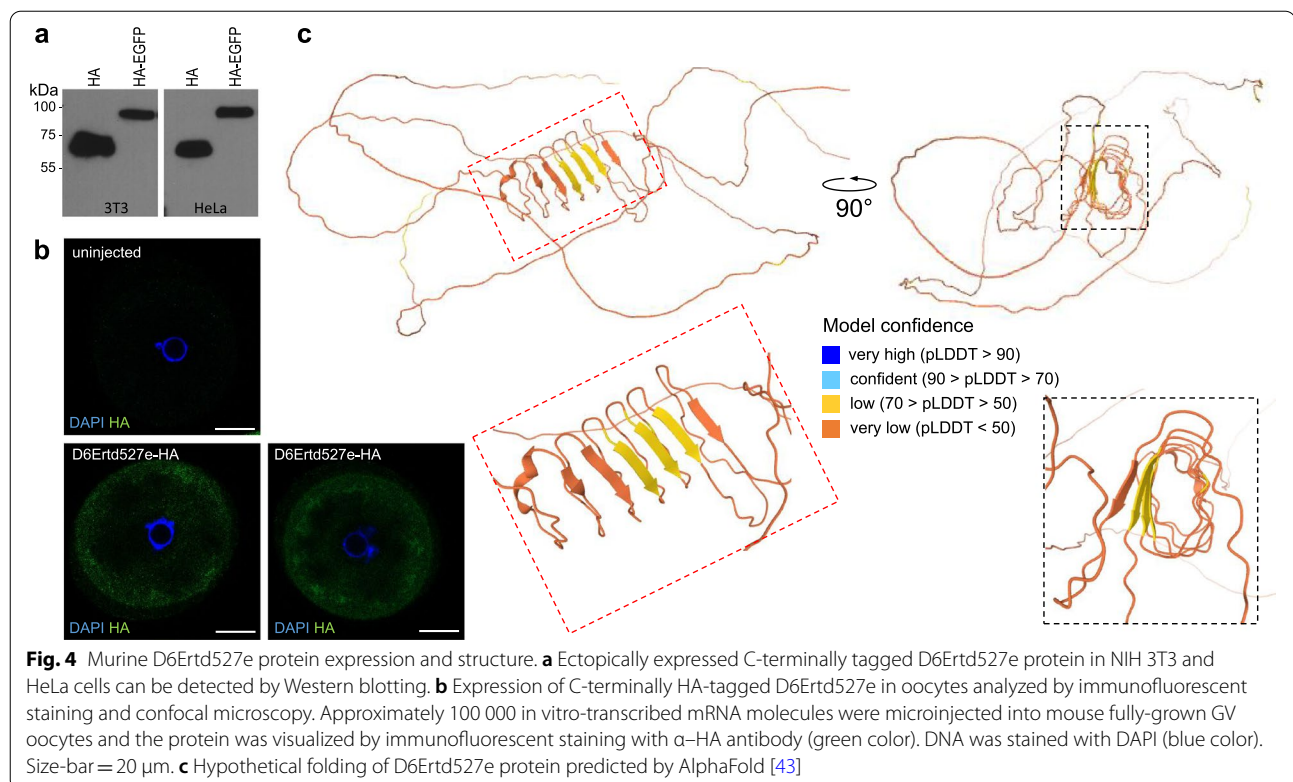
Deviations from the utilization of the serine reading frame in the predicted protein sequences were found in

Apodemus sylvaticus and hamsters (Fig. 3a). In *Apodemus*, the polyS frame at the N-terminus is rather short and approximately 2/3 of the protein are not rich in (CAG)_n repeats (Fig. 3a). In hamsters, the MTD-driven transcript does not carry a long ORF (golden hamster) or is not expressed at all (Chinese hamster). However, both hamsters utilize an upstream promoter driving expression of a transcript with longer CDS. The predicted Chinese hamster *D6Ertd527e* CDS identified in this transcript is relatively long and with a glutamine-rich segment, which would give the predicted encoded protein entirely different properties than exhibit other *D6Ertd527e* homologs in the *Muridae* family (Fig. 3a). However, as mentioned above, expression of *D6Ertd527e* in Chinese hamster oocytes is minimal. This contrasts with golden hamster, where this transcript with protein coding potential is well expressed. The predicted protein carries a glutamine-rich segment at the N-terminus but also a frameshift into the polyS frame in the central part of the predicted protein (Fig. 3a). While significance of this *D6Ertd527e* remains unknown, it provides another example of significant divergence of a potentially encoded protein from the *D6Ertd527e* locus.

Taken together, many *Muridae* species evolved relatively long MTD-driven *D6Ertd527e* CDS and their predicted protein sequences considerably diverged from a homopolymeric amino acid sequence that would be encoded by a perfect trinucleotide repeat. Relative absence of frameshifts between polyS, polyQ, and polyA frames of (CAG)_n repeats in mice implies that their *D6Ertd527e* homologs might encode serine-rich proteins that are under neutral or positive selection. In contrast, *Cricetidae* did not evolve longer MTD-driven *D6Ertd527e* CDS. Instead, the MTD LTR was repeatedly lost and two *D6Ertd527e* CDS transcripts with predicted longer CDS originate from a different promoter.

Murine D6Ertd527e protein features

Murine *D6Ertd527e* is a *bona fide* protein-coding gene, as its protein product was reported in proteomic studies of mouse oocytes [39–42], and six different *D6Ertd527e* peptides can be identified in released data [41]. Likewise, *D6Ertd527e* fused to a C-terminal hemagglutinin (HA) tag can be ectopically expressed in cultured mammalian cells and detected as a protein of the expected size by immunoblot analysis (Fig. 4a, Additional file 2). In cultured cells, the *D6Ertd527e* protein diffusely localized to the cytoplasm, with no apparent effect on the expressing cells [12]. Ectopic expression of *D6Ertd527e*–HA demonstrated that the MTD LTR provides a functional 5' UTR and translation initiation start and the (CAG)_n repeat-derived CDS can be translated into a detectable non-aggregating protein. We also expressed *D6Ertd527e*–HA



protein from mRNA microinjected into the oocyte. We microinjected $\sim 100,000$ molecules of *D6Ertd527e*-HA mRNA into fully-grown GV oocytes, cultured them for 20 h and stained them with α -HA antibody. Staining appeared stronger in oocyte's periphery and some denser areas (Fig. 4b). Taken together the CDS gives rise to a stable protein, which does not exhibit strong aggregation propensity.

Recent advances in protein structure predictions enables to build protein models in silico with a good accuracy [43]. Structural prediction of *D6Ertd527e* revealed a largely unstructured protein with an unusual central beta-sheet barrel (Fig. 4c). However, this central structure was predicted with a low confidence and was not observed in structural predictions of *D6Ertd527e* homologs in *Mus pahari* (shrew mouse) and *Acomys cahirinus* (spiny mouse, Additional file 1: Fig. S5) suggesting that it is not a conserved functional element of *D6Ertd527e*. We thus conclude that *D6Ertd527e*-encoded proteins are typically intrinsically disordered serine rich proteins.

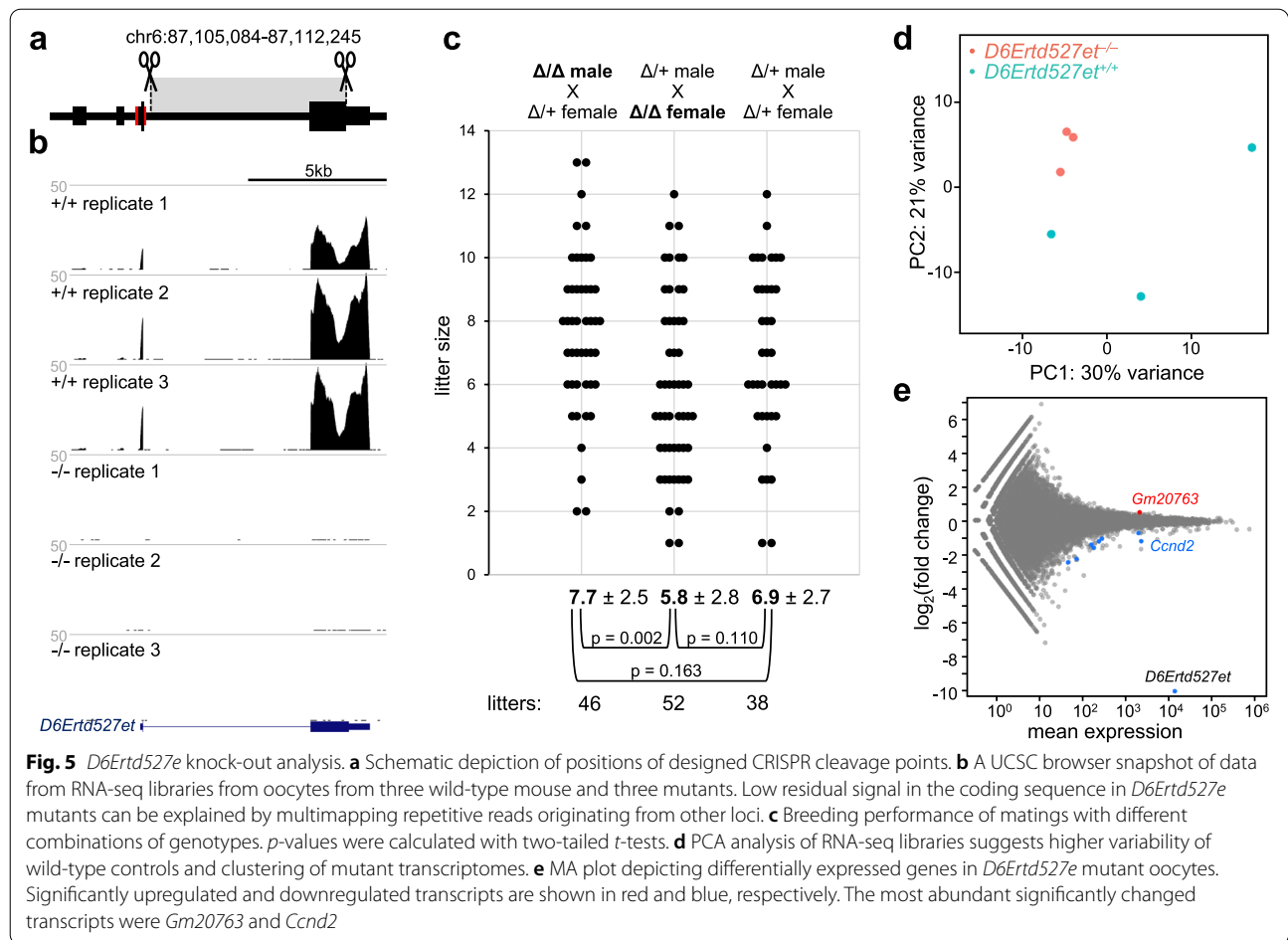
Functional analysis of *D6Ertd527e*

To examine biological significance of *D6Ertd527e*, we used CRISPR guided nucleases to generate a mouse deletion model for *D6Ertd527e*. CRISPR-cleavage positions were intended to delete the intron and the coding sequence from the last exon while retaining the MTD

LTR and the conserved 3' UTR sequence (Fig. 5a). We successfully produced mutant mice carrying deletion of the intron and the coding part of the 3' terminal exon (Additional file 1: Fig. S6), which resulted in the loss of *D6Ertd527e* expression (Fig. 5b). Mutant mice were fertile but had slightly smaller average litter size (Fig. 5c).

The litter size was smaller by approximately one pup but this difference was not significant when we compared all results from mutant and heterozygous females mated with heterozygous males ($p=0.110$, two-tailed t-test, Fig. 5c). However, inspection of breeding data suggested that litter size has variability affected by lower sizes of the first litters, which is a known phenomenon [44]. When this genotype-independent variability of first litters was removed from the analysis by not including them, the litter size of knock-out and heterozygous females mated to heterozygous males became statistically significant ($p=0.007$, two-tailed t-test).

This difference was not accompanied by any remarkable transcriptome change. Transcriptome profiling of mutant oocytes by RNA sequencing showed negligible changes in gene expression (Fig. 5d, e). The most abundant differentially expressed transcripts were *Gm20763* (increased abundance) and *Cyclin D2* (*Ccnd2*, reduced abundance). However, these changes do not seem to be functionally significant. *Gm20763* is an ORR1A2 LTR-driven lncRNA carrying an antisense *Kif1c* pseudogene.



This class of maternal lncRNAs would bind *Kif1c* mRNA and trigger endogenous RNAi [45]. However, the *Kif1c* is not among significantly downregulated genes indicating that the observed increase in *Gm20763* level does not affect RNAi-mediated repression of its target. Similarly, significance of reduced mRNA levels of *Ccnd2* in prophase-arrested fully-grown oocytes is questionable as this gene is important for cumulus cells but not the oocytes as shown in the *Ccnd2* knock-out where oocytes lacking CCND2 meiotically mature and develop to the blastocyst stage after fertilization at normal rates [46].

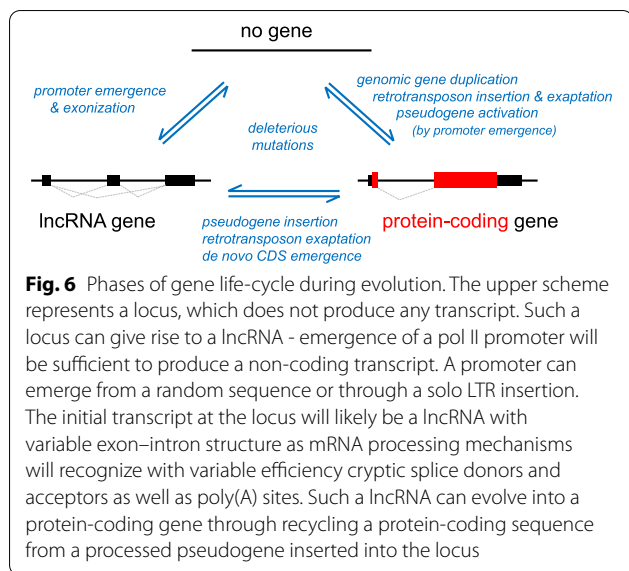
Taken together, absence of *D6Ertd527e* appears to have minor effect on fertility of mutant oocytes while transcriptome changes are minimal and do not provide any explanation for the reduced litter size of *D6Ertd527e*^{-/-} females.

Discussion

Here we report evolutionary history and functional analysis of a de novo formed murine gene, which offers an outstanding example of numerous phases of gene emergence, evolution and demise (Fig. 6) on an evolutionary

scale of tens of millions of years. The gene *D6Ertd527e* emerged in an intergenic region in the common ancestor of mice and hamsters. The same syntenic locus in porcine and bovine genomes does not produce a known transcript while the human locus carries an MLT1A0 LTR-derived promoter and the first exon of an oocyte-specific unannotated lncRNA that does not share any feature with *D6Ertd527e* and clearly represents an independent evolutionary event [12].

The critical event in the *D6Ertd527e* locus was insertion of an MTD LTR, which provided an oocyte-specific promoter and a full 5' terminal exon. The locus may have contained or evolved additional promoters, as suggested by alternative upstream transcription start sites in mouse and hamster genomes. However, the initial MTD LTR insertion already contained an AUG codon, which appears as the start codon in most predicted *D6Ertd527e* coding sequences in *Muridae* and its functionality has been validated in the murine *D6Ertd527* (Fig. 4a, b and [12]). The conserved splice donor in the MTD LTR is spliced with a downstream splice acceptor, which presumably evolved from a cryptic splice acceptor. At least



there is no evidence outside of murids that this splice acceptor would be a part of another functional transcript.

Evolution of the *D6Ertd527* in mice and hamsters took diverse paths. The MTD LTR driving *D6Ertd527e* expression was lost during *Cricetidae* evolution at least three times due to internal deletions (Additional file 1: Fig. S2). Admittedly, the analysis relied on rodent genome assemblies obtained by different methods and variable quality of assembly. However, existence of the deletion in *Arvicolinae* is supported by presence of the same apparent deletion in four different species. Deletion in a single species of genus *Peromyscus* is supported only indirectly. The *D6Ertd527e* gene region in nine *Peromyscus* species is assembled well. *Peromyscus leucopus*, which lacks the MTD insertion, and its closest sequenced relative *Peromyscus maniculatus* have their genomes (GCA_004664715.2 and GCF_003704035.1) assembled to the chromosomal level, which are higher quality assemblies [47, 48].

The protein-coding capacity of the *D6Ertd527e* transcript expressed from the MTD LTR likely have evolved stochastically. The functional AUG in the original MTD insert would prime evolution of a protein-coding transcript but extreme variability of *D6Ertd527e* coding sequences starting from that MTD AUG suggests that *D6Ertd527e* either initially were lncRNAs or repeatedly evolved into ones. It is important to point out that there is no strict disjunction between pol II-transcribed spliced polyadenylated lncRNAs and protein-coding mRNAs as ribosomes usually scan lncRNAs and might translate their short putative coding sequences [49, 50].

Expansion of simple nucleotide repeats in genomic DNA is a common phenomenon (reviewed in [51]).

In a coding sequence, expanding nucleotide repeats will give rise to amino acid repeats (homorepeats in case of expanding trinucleotides) in affected proteins. Amino acid homorepeats in proteins are diverse in terms of the amino acid type, length, and biological effect (reviewed in [52, 53]). Particularly (CAG)_n trinucleotide repeat expansion translated in the polyQ frame, which would exceed certain thresholds in specific proteins, has been associated with a number of pathologies known as polyglutamine diseases [54]. However, amino acid homorepeats also have physiological roles [53] and trinucleotide repeat expansion offers a mechanism for devolution of homopolymeric intrinsically disordered proteins or their regions. A well-established example is an expansion of a low-complexity alanine-rich sequence during convergent evolution of antifreeze proteins in fish [55, 56]. In any case, combination of trinucleotide repeat expansion combined with larger recombination events and point mutations, which change serine residues (ACG codon) into other amino acids, offer an interesting model for stochastic evolution of protein coding sequence. A single point mutation in the repeat can convert an ACG codon into a codon encoding one of six other amino acids (Gly, Arg, Cys, Asn, Thr, Ile) but not into a stop codon. Two simultaneous point mutations in a codon further increase potential for amino acid changes while having only 3.7% chance of creating a stop codon and disrupting the evolving CDS. This is consistent with the appearance of murine *D6Ertd527e* CDS where the (CAG)_n repeat has been eroded by point mutations and recombinations while pure (CAG)_n repeats are restricted to specific regions, which seem to expand independently (Fig. 3a).

At the same time, the polyQ frame appears to be avoided in putative *D6Ertd527e* proteins translated from the start codon in the MTD LTR-derived exon. The proteins are serine rich, with only one major switch to the alanine-encoding (CAG)_n reading frame. As this phenomenon stretches across the entire *Muridae* family, it suggests some positive selection for its maintenance might exist.

As the CDS analysis depends on the quality of genome assembly, it should be pointed out that most of the *D6Ertd527e* locus sequences were assembled without any gaps within the repetitive coding sequence. This is not surprising as the coding (CAG)_n repeats are typically eroded by mutations, which facilitates sequence assembly, and perfect (CAG)_n repeats in sequences are typically not long enough to interfere with sequencing assembly. The only exceptions are *Mus spicilegus* and *Mus caroli*, which each have a single gap within the assembled *D6Ertd527e* CDS. Furthermore, when transcriptome

data were available, their mapping onto the assembled genomes did not reveal any issues with assembled genomic sequences of *D6Ertd527e*.

Did *D6Ertd527e* evolve some function as a protein-coding gene? The length and preservation of the CDS translated in the serine frame in the lineage leading to house mouse indicates that *D6Ertd527e* may indeed have some function. This notion is supported by slightly reduced litter size (15–20%) of *D6Ertd527e*^{-/-} females. On evolutionary scale, such an effect on reproductive fitness might represent a significant factor. At the same time, we were not able to pinpoint the function of *D6Ertd527e* protein. Sequence composition and structural analysis suggests that the encoded serine-rich protein is intrinsically disordered. But what biological role could it play?

One interesting example of a low-complexity serine-rich sequence is the Phosvitin (Pv) domain/protein from Vitellogenin, an egg-yolk precursor (reviewed in [57]). One function of Vitellogenin and yolk proteins is antioxidant activity providing protection against oxidative damage [58, 59]. Phosvitin is a serine-rich polypeptide (>50% serine), which is highly phosphorylated [60], attracts multivalent cations as calcium, magnesium, zinc, and iron [61] and its iron chelation ability was shown to reduce DNA damage [62]. Interestingly, Vitellogenin genes were lost during mammalian evolution during transition from yolk-dependent nourishment toward lactation and placentation [63]. That *D6Ertd527e* protein could contribute to reduction of oxidative damage and substitute function of Vitellogenin sounds attractive but it is not fully consistent with all data as the level of phosphorylation of *D6Ertd527e* is unclear. Mass-spec analysis supporting murine *D6Ertd527e* peptides in oocytes detected non-phosphorylated peptides. Likewise, a discrete band of ectopically expressed HA-tagged *D6Ertd527e* in 3T3 and HeLa cells (Fig. 4a) does not seem to support the notion of a highly phosphorylated *D6Ertd527e*. In any case, some cytoplasmic function of *D6Ertd527e* stemming from its biophysical features is a likely one that could purposelessly emerge during evolution of *D6Ertd527e*.

Conclusions

There is a number of characterized de novo protein-coding genes in vertebrates or elsewhere, which are described in the literature and are of a comparable age or even younger than *D6Ertd527e* (reviewed in [64, 65]). However, uniqueness of *D6Ertd527e* is that its documented evolution makes it an excellent textbook example of stochastic events, which bring into being a transcriptional unit in the genome, which can either evolve into

a protein-coding gene, remain, or disappear during evolution.

Methods

Animals

Animal experiments were approved by the Institutional Animal Use and Care Committees (Approval no. 58–2015) and were carried out in accordance with the law.

Oocyte and embryo collection

Fully grown, germinal vesicle (GV)-intact oocytes were obtained from C57Bl/6NCrl mice as described previously [66]. Oocytes were collected and microinjected in M2 medium supplemented with 0.2 mM 3-isobutyl-1-methyl-xanthine (IBMX; Sigma) and cultured in M16 medium (Sigma-Aldrich) supplemented with 0.2 mM 3-isobutyl-1-methyl-xanthine (IBMX; Sigma), at 37 °C in a 5% CO₂ atmosphere.

Oocyte microinjection

RNA for injection was diluted in pure water such that 100,000 molecules would be present in 5 picoliters (pl). Microinjections were done using a FemtoJet microinjector (Eppendorf). Femtojet injection pressure was set to maintain injection volume of 5 pl for all microinjections. Reliability of the estimated amount of microinjected molecules was experimentally addressed previously [67]. Injected mouse oocytes were cultured in M16 media (Merck) supplemented with IBMX in 5% CO₂ at 37 °C for 20 h.

CRISPR-mediated deletion of *D6Ertd527e*

The deletion mutant model was produced in the Czech Centre for Phenogenomics at the Institute of Molecular Genetics ASCR using Cas9-mediated deletion of *D6Ertd527e* intron 1 and the protein-coding sequence of exon 2 (Additional file 1: Fig. S6). Sequences of guide RNAs were sgRNA T5 5'-CCTCGAGATGAGCCATCC-3' and sgRNA E2 5'-CTTAGGAAATCATTCCCA-3'. To produce guide RNAs, synthetic 128 nt guide RNA templates including T7 promoter, 18nt sgRNA, and tracrRNA sequences were amplified using T7 and tracrRNA primers. Guide RNAs were produced in vitro using the Ambion mMACHINE T7 Transcription Kit and purified using the mirPremier™ microRNA Isolation Kit (Sigma). The Cas9 mRNA was synthesized from pSpCas9-puro plasmid using Ambion mMACHINE T7 Transcription Kit and purified using the RNeasy Mini kit (Qiagen). A sample for microinjection was prepared by mixing two guide RNAs in water (25 ng/μl for each) together with Cas9 mRNA (100 ng/

μl). Five picoliters of the mixture were microinjected into male pronuclei of C57Bl/6 zygotes and transferred into pseudo-pregnant recipient mice. PCR genotyping was performed on tail biopsies from 4-week-old animals. We obtained a positive founder which transmitted the mutant allele to F₁, and after two generations of breeding with C57Bl/6NCrl animals, the heterozygotes were used for breeding *D6Erd527e1^{-/-}* animals for phenotype analysis.

For detection of the knock-out allele were used D6Erd_gen_Fwd5: 5'-CCTGACACTCAAGAGACA CGGTCA and D6Erd_1_Rev: 5'-CACCTTTCTGTG CTTGTGCTGAAC giving a 877 bp (wild-type allele is absent, too long to amplify). For detecting the wild type allele were used D6Erd_gen_Fwd5 and D6Erd_MTD_Rev4: 5'-GAACTGCAAGCTGAGGCTCACAAG, yielding a 812 bp product.

D6Erd527e expression vector

Full-length mouse *D6Erd527e* with the C-terminal HA-tag was synthesized by GENEWIZ in pUC57-Kan plasmid. The coding sequence was cleaved out by NheI and EagI and transferred into pSV40 plasmid backbone. EGFP coding sequence was PCR-amplified and inserted into XbaI and EagI restriction sites to produce D6Erd527e-HA-EGFP fusion. The final constructs were confirmed by sequencing. The plasmids are available from Addgene: pSV40_mD6Erd527e-HA as #192,222, pSV40_mD6Erd527e-HA-EGFP as #192,223.

Cell culture and transfection

Mouse NIH 3T3 cells were maintained in DMEM (Sigma-Aldrich) supplemented with 10% fetal calf serum (Sigma-Aldrich), penicillin (100 U/ml; Invitrogen), and streptomycin (100 μg/ml; Invitrogen) at 37 °C and 5% CO₂ atmosphere.

For transfection, the cells were plated on a 24-well plate, grown to 80% density, and transfected with 1 μg plasmid DNA using the Lipofectamine 3000 (ThermoFisher Scientific) according to the manufacturer's protocol. The cells were collected for analysis 48 h post-transfection.

Western blotting

Transfected NIH 3T3 cells were washed with PBS and lysed in RIPA buffer (50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% NP-40 (Igepal CA-630), 0.5% Na-deoxycholate, 0.1% SDS) supplemented with 1 × protease inhibitor cocktail set (Millipore). Proteins were separated on 10% polyacrylamide gel and transferred onto a PVDF membrane (Millipore). Anti-HA primary antibody (High affinity, #11867423001, Roche, dilution 1:1000) and HRP-conjugated goat anti-Rat secondary antibody (#31470,

ThermoFisher Scientific, dilution 1:50,000) were used for signal detection with SuperSignal West Femto Chemiluminescent Substrate (Pierce).

RNA sequencing

Total RNA was extracted from 25 wild-type or *D6Erd527e1^{-/-}* fully-grown oocytes from 8–10-week-old animals using PicoPure RNA Isolation Kit with on-column genomic DNA digestion according to the manufacturer's instruction (Qiagen). RNA-Seq libraries were constructed using the Ovation RNA-Seq system V2 (NuGEN) followed by Ovation Ultralow Library system (DR Multiplex System, NuGEN). RNA-Seq libraries were pooled and sequenced using 65-nt single-end-sequencing using Illumina HiSeq. *D6Erd527e1^{-/-}* oocyte sequencing data were deposited in GEO (<https://www.ncbi.nlm.nih.gov/geo/>) as GSE213820. *Cricetulus griseus* oocyte sequencing data were also deposited under GSE213820. Remaining RNA-seq data were published previously and were obtained from GEO database: *Mus musculus* data accession: GSE116771 [27], *Mesocricetus auratus* data accession: GSE116771 [12] and GSE169528 [33], *Rattus norvegicus* data accession: GSE137562 [35], and *Acomys cahirinus* data accession: PRJNA436818 [36].

RNA-seq mapping and expression analysis

All RNA-seq data were mapped onto indexed *Mus musculus* (mm10, GCA_000001635.2), *Rattus norvegicus* (rn7, GCA_015227675.2), *Acomys cahirinus* (AcoCah_v1_BIUU, GCA_004027535.1), *Mesocricetus auratus* (MesAur1.0, GCA_000349665.1), and *Cricetulus griseus* (criGriChoV2, GCA_900186095.1) genomes using STAR 2.5.3a [68] as previously described [27]. Read mapping coverage was visualized in the UCSC Genome Browser by constructing bigWig tracks using the UCSC tools [69]. Differential expression analysis was done in R software environment [70] using DESeq2 package [71] as previously described [27].

Abbreviations

3T3: 3-Day transfer, inoculum 3 × 10⁵ cells; CDS: Coding sequence; CPAT: Coding-Potential Assessment Tool; CRISPR: Clustered regularly interspaced short palindromic repeats; DMEM: Dulbecco's Modified Eagle Medium; DNA: Deoxyribonucleic acid; EDTA: Ethylene diamine tetraacetic acid; EGTA: Ethylene glycol tetraacetic acid; GV: Germinal vesicle; HA: Hemagglutinin; IBMX: 3-Isobutyl-1-methyl-xanthine; LINE: Long interspersed element; LTR: Long terminal repeat; lncRNA: Long non-coding RNA; MaLR: Mammalian apparent LTR retrotransposon; MLT: Mammalian LTR retrotransposon (retrotransposon family); MT: Mouse transcript (retrotransposon family); MTC: Mouse transcript C subfamily; MTD: Mouse transcript D subfamily; MYA: Million years ago; NIH: National Institutes of Health; ORR: Origin-region repeat (retrotransposon family); PBS: Phosphate-buffered saline; PCA: Primary component analysis; RIPA: Radioimmunoprecipitation assay buffer; RNA: Ribonucleic acid; sgRNA: Single guide RNA; SINE: Short interspersed element; TE: Transposable element; UCSC: University of California Santa Cruz; UTR: Untranslated region.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-022-01470-5>.

Additional file 1: Figs. S1–S5. Fig. S1. Additional data on D6Ert527e gene and its expression. **Fig. S2.** Analysis of deletions in genomic sequences of D6Ert527e in Cricetidae. **Fig. S3.** Expression of D6Ert527e in different rodents. **Fig. S4.** Nucleotide exchange rates along the MTD LTR. **Fig. S5.** D6Ert527e appears to encode an intrinsically disordered protein. **Fig. S6.** Production of D6Ert527e mutant allele.

Additional file 2. Uncropped western blot image used for Fig. 4a.

Acknowledgements

We thank Marian Novotny from the Faculty of Sciences of the Charles University for comments on the structural analysis in silico.

Authors' contributions

Conception, P.S.; study design R.M., J.Pe., J.Pa., P.S.; experimental work, H.F., F.H., R.M., J.Pe., J.Pa., S.K., data analysis F.H., R.M., J.Pe., J.Pa., S.K., P.S.; writing—original draft preparation, R.M. and P.S.; writing—review and editing, all co-authors; funding acquisition, P.S. All authors have read and agreed to the published version of the manuscript.

Funding

This work was funded from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant 647403, D-FENS) and RVO 68378050-KAV-NPUI. Financial support of S.K. and F.H. was in part provided by the Charles University in a form of a PhD student fellowship; this work will be in part used to fulfil requirements for a PhD degree and hence can be considered "school work." The authors used services of the Light Microscopy Core Facility, IMG CAS, Prague, Czech Republic, supported by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS, LM2018129 and CZ.02.1.01/0.0/0.0/18_046/0016045) and the Czech Centre for Phenogenomics at the Institute of Molecular Genetics supported by the Czech Academy of Sciences RVO 68378050 and by the project LM2018126 Czech Centre for Phenogenomics provided by MEYS. Computational resources for J.P. included support by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by MEYS. Computational resources for F.H. included support from the European Structural and Investment Funds grants for the Croatian National Centre of Research Excellence in Personalized Healthcare (contract #KK.01.1.1.01.0010) and Croatian National Centre of Research Excellence for Data Science and Advanced Cooperative Systems (contract #KK.01.1.1.01.0009).

Availability of data and materials

Expression data were deposited in the GEO database as GSE213820 [72], and plasmids are available through Addgene (pSV40_mD6Ert527e-HA as #192,222, and pSV40_mD6Ert527e-HA-EGFP as #192,223); the D6Ert527e deletion mutant mouse line has been archived and is available upon request as a frozen sperm sample.

Declarations

Ethics approval and consent to participate

Animal experiments were approved by the Institutional Animal Use and Care Committees (Approval no. 58–2015) and were carried out in accordance with the law.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Institute of Molecular Genetics of the Czech Academy of Sciences, Videnska 1083, 142 20 Prague 4, Czech Republic. ²Present address: Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical

University of Vienna, Vienna, Austria. ³Bioinformatics Group, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia. ⁴Present address: Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. ⁵Current address: Institute of Experimental Medicine of the Czech Academy of Sciences, Videnska 1083, 142 20 Prague 4, Czech Republic.

Received: 30 September 2022 Accepted: 15 November 2022

Published online: 08 December 2022

References

- Johannsen W. Elemente der exakten erblichkeitslehre. Deutsche wesentlich erweiterte ausgabe in fünfundzwanzig vorlesungen. Jena: G. Fischer; 1909. p. 534. https://www.archive.org/download/elementederexakt00joha/page/n4_w509.
- Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, et al. What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 2007;17(6):669–81.
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* 2003;4:41.
- Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, et al. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* 2013;41(Database issue):D348–52.
- Mushegian A. Gene content of LUCA, the last universal common ancestor. *Front Biosci.* 2008;13:4657–66.
- Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* 2012;8(7):e1002841.
- Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 2015;11(7):1110–22.
- Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.* 2014;30(10):439–52.
- Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, et al. A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS ONE.* 2008;3(6):e2521.
- Hezroni H, Ben-Tov Perry R, Meir Z, Housman G, Lubelsky Y, Ulitsky I. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol.* 2017;18(1):162.
- Housman G, Ulitsky I. Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochim Biophys Acta.* 2016;1859(1):31–40.
- Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res.* 2017;27(8):1384–94.
- Van Oss SB, Carvunis AR. De novo gene birth. *PLoS Genet.* 2019;15(5):e1008160.
- Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. *Nat Commun.* 2018;9(1):1530.
- Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome Biol.* 2016;17:100.
- de Souza FS, Franchini LF, Rubinstein M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol.* 2013;30(6):1239–51.
- Goke J, Ng HH. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO Rep.* 2016;17(8):1131–44.
- Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* 2013;9(4):e1003470.
- Ganesh S, Svoboda P. Retrotransposon-associated long non-coding RNAs in mice and men. *Pflugers Arch.* 2016;468(6):1049–60.
- Brosius J, Gould SJ. On "nomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA." *Proc Natl Acad Sci U S A.* 1992;89(22):10706–10.

21. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell*. 2004;7(4):597–606.
22. Thompson PJ, Macfarlan TS, Lorincz MC. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol Cell*. 2016;62(5):766–76.
23. Smit AF. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res*. 1993;21(8):1863–72.
24. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420(6915):520–62.
25. Flemr M, Malik R, Franke V, Nejepinska J, Sedlacek R, Vlahovicek K, et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell*. 2013;155(4):807–16.
26. Piao Y, Ko NT, Lim MK, Ko MS. Construction of long-transcript enriched cDNA libraries from submicrogram amounts of total RNAs by a universal PCR amplification method. *Genome Res*. 2001;11(9):1553–8.
27. Horvat F, Fulka H, Jankele R, Malik R, Jun M, Solcova K, et al. Role of Cnot6l in maternal mRNA turnover. *Life Sci Alliance*. 2018;1(4):e201800084.
28. Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. *Nature*. 1998;392(6679):917–20.
29. Kumar S, Suleski M, Craig JM, Kasprowitz AE, Sanderford M, Li M, et al. TimeTree 5: an expanded resource for species divergence times. *Mol Biol Evol*. 2022;39(8):msac174. <https://doi.org/10.1093/molbev/msac174>
30. Steppan SJ, Schenk JJ. Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates. *PLoS ONE*. 2017;12(8):e0183070.
31. Abe K, Yamamoto R, Franke V, Cao M, Suzuki Y, Suzuki MG, et al. The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3' processing. *EMBO J*. 2015;34(11):1523–37.
32. Gahurova L, Tomizawa SI, Smallwood SA, Stewart-Morgan KR, Saadeh H, Kim J, et al. Transcription and chromatin determinants of de novo DNA methylation timing in oocytes. *Epigenetics Chromatin*. 2017;10:25.
33. Zhang H, Zhang F, Chen Q, Li M, Lv X, Xiao Y, et al. The piRNA pathway is essential for generating functional oocytes in golden hamsters. *Nat Cell Biol*. 2021;23(9):1013–22.
34. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014;515(7527):355–64.
35. Ganesh S, Horvat F, Drutovic D, Efenberkova M, Pinkas D, Jindrova A, et al. The most abundant maternal lncRNA Sirena1 acts post-transcriptionally and impacts mitochondrial distribution. *Nucleic Acids Res*. 2020;48(6):3211–27.
36. Mamrot J, Gardner DK, Temple-Smith P, Dickinson H. Embryonic gene transcription in the spiny mouse (*Acomys cahirinus*): an investigation into the embryonic genome activation. *bioRxiv*. 2018:280412. <https://doi.org/10.1101/280412>.
37. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41(6): e74.
38. Simon MM, Greenaway S, White JK, Fuchs H, Gailus-Durner V, Wells S, et al. A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol*. 2013;14(7):R82.
39. Wang S, Kou Z, Jing Z, Zhang Y, Guo X, Dong M, et al. Proteome of mouse oocytes at different developmental stages. *Proc Natl Acad Sci U S A*. 2010;107(41):17639–44.
40. Pfeiffer MJ, Siatkowski M, Paudel Y, Balbach ST, Baeumer N, Crossetto N, et al. Proteomic analysis of mouse oocytes reveals 28 candidate factors of the "reprogrammome." *J Proteome Res*. 2011;10(5):2140–53.
41. Wang B, Pfeiffer MJ, Drexler HC, Fuellen G, Boiani M. Proteomic analysis of mouse oocytes identifies PRMT7 as a reprogramming factor that replaces SOX2 in the induction of pluripotent stem cells. *J Proteome Res*. 2016;15(8):2407–21.
42. Israel S, Ernst M, Psathaki OE, Drexler HCA, Casser E, Suzuki Y, et al. An integrated genome-wide multi-omics analysis of gene expression dynamics in the preimplantation mouse embryo. *Sci Rep*. 2019;9(1):13356.
43. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–9.
44. Weber EM, Algers B, Wurbel H, Hultgren J, Olsson IA. Influence of strain and parity on the risk of litter loss in laboratory mice. *Reprod Domest Anim*. 2013;48(2):292–6.
45. Karlic R, Ganesh S, Franke V, Svobodova E, Urbanova J, Suzuki Y, et al. Long non-coding RNA exchange during the oocyte-to-embryo transition in mice. *DNA Res*. 2017;24(2):129–41.
46. Sicinski P, Donaher JL, Geng Y, Parker SB, Gardner H, Park MY, et al. Cyclin D2 is an FSH-responsive gene involved in gonadal cell proliferation and oncogenesis. *Nature*. 1996;384(6608):470–4.
47. Long AD, Baldwin-Brown J, Tao Y, Cook VJ, Balderrama-Gutierrez G, Corbett-Detig R, et al. The genome of *Peromyscus leucopus*, natural host for Lyme disease and other emerging infections. *Sci Adv*. 2019;5(7):eaaw6441.
48. Harringmeyer OS, Hoekstra HE. Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat Ecol Evol*. 2022:1–15. <https://doi.org/10.1038/s41559-022-01890-0>.
49. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*. 2014;8(5):1365–79.
50. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*. 2015;4:e08890.
51. Kim JC, Mirkin SM. The balancing act of DNA repeat expansions. *Curr Opin Genet Dev*. 2013;23(3):280–8.
52. Mier P, Alanis-Lobato G, Andrade-Navarro MA. Context characterization of amino acid homorepeats using evolution, position, and order. *Proteins*. 2017;85(4):709–19.
53. Chavali S, Singh AK, Santhanam B, Babu MM. Amino acid homorepeats in proteins. *Nat Rev Chem*. 2020;4(8):420–34.
54. Shao J, Diamond MI. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. *Hum Mol Genet*. 2007;16 Spec No. 2:R115–23.
55. Chen L, DeVries AL, Cheng CH. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci U S A*. 1997;94(8):3811–6.
56. Chen L, DeVries AL, Cheng CH. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc Natl Acad Sci U S A*. 1997;94(8):3817–22.
57. Carducci F, Biscotti MA, Canapa A. Vitellogenin gene family in vertebrates: evolution and functions. *Eur Zoological J*. 2019;86(1):233–40.
58. Sun C, Zhang S. Immune-relevant and antioxidant activities of vitellogenin and yolk proteins in fish. *Nutrients*. 2015;7(10):8818–29.
59. Li H, Zhang S. Functions of vitellogenin in eggs. *Results Probl Cell Differ*. 2017;63:389–401.
60. Taborsky G. Phosvitin. *Adv Inorg Biochem*. 1983;5:235–79.
61. Finn RN. Vertebrate yolk complexes and the functional implications of phosvitins and other subdomains in vitellogenins. *Biol Reprod*. 2007;76(6):926–35.
62. Ishikawa S, Yano Y, Arihara K, Itoh M. Egg yolk phosvitin inhibits hydroxyl radical formation from the fenton reaction. *Biosci Biotechnol Biochem*. 2004;68(6):1324–31.
63. Brawand D, Wahli W, Kaessmann H. Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biol*. 2008;6(3):e63.
64. Long M, Betran E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 2003;4(11):865–75.
65. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet*. 2016;17(9):567–78.
66. Nagy A. In: *Manipulating the mouse embryo: a laboratory manual*. 3rd ed. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2003. p. x, 764.
67. Kataruka S, Modrak M, Kinterova V, Malik R, Zeitler DM, Horvat F, et al. MicroRNA dilution during oocyte growth disables the microRNA pathway in mammalian oocytes. *Nucleic Acids Res*. 2020;48(14):8050–62.
68. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

69. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. 2010;26(17):2204–7.
70. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013. <http://www.r-project.org/>.
71. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
72. Horvat F. De novo emergence, existence, and demise of a protein-coding gene in murids. NCBI GEO accession GSE213820. 2022. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213820>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

