Phylogenetic analysis of NEAT1 and MALAT1 long non-coding RNAs highlights structure-function relationships in paraspeckle biology

Arkhipova, K.; Drukker, M.E.

PHYLOGENETIC ANALYSIS OF *NEAT1* AND *MALAT1* LONG NON-CODING RNAs HIGHLIGHTS STRUCTURE–FUNCTION RELATIONSHIPS IN PARASPECKLE BIOLOGY

AUTHORS

Ksenia Arkhipova[1*] and Micha Drukker[1,2*]

[1] Division of Drug Discovery and Safety, Leiden Academic Centre for Drug Research (LACDR), Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands

[2] Department of Internal Medicine, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands

* To whom correspondence should be addressed. Email: arkhipova.a.ksenia@gmail.com, m.drukker@lacdr.leidenuniv.nl

ABSTRACT

Paraspeckles are nuclear bodies essential for gene regulation and stress response, and they are built upon the long non-coding RNA NEAT1. Together with the syntenic MALAT1, these are the only lncRNAs that use the tRNA-processing machinery for maturation, yet they differ in function and evolutionary conservation. To investigate these differences, we identified NEAT1 and MALAT1 orthologs across 545 mammals. For NEAT1, we found that G-quadruplexes, short motifs interacting with DBHS proteins and TDP-43, long gene length, and self-complementary regions are highly conserved features that likely stabilize paraspeckle integrity. Transposable elements also contributed structural modules potentially recognized by DBHS proteins, underscoring their role in NEAT1 evolution. The NEAT1Short isoform was present in all orthologs, and the TDP-43–mediated isoform switch appears to be conserved. In contrast, MALAT1 function likely relies on its conserved primary sequence and regions under purifying selection. This is the first large-scale phylogenetic study of *NEAT1* – a lncRNA that lacks

sequence similarity between orthologs while maintaining functional and syntenic conservation.

INTRODUCTION

Retrieving functionally important regions from an analysis of conservation patterns of primary and secondary structures of proteins and non-coding RNAs is a common approach. The method is based on the identification of conserved regions between orthologs, highlighting the pressure of purifying selection, which ensures that deleterious mutations are not established in the population, thereby maintaining only functionally essential structures (Charlesworth et al. 1993). Detailed mechanisms of function for the two long non-coding RNAs, *NEAT1* and *MALAT1*, connected by the uniqueness of their maturation processes, are not yet clear. However, the association of these genes with neurodegenerative diseases and cancer highlights the urgent need to identify regions and properties crucial for their function.

NEAT1 and *MALAT1* share unique structural elements at their 3′-ends and the maturation processing machinery. Both genes are located on chromosome 11 in the human genome, positioned in close proximity to each other and coded on the same strand (Fig. 1A), with SCYL1 adjacent to *MALAT1* and FRMD8 bordering *NEAT1*. Similar localisation of the genes in the mouse genome suggests possible synteny in other mammalian genomes as well (Stadler 2010). *NEAT1* is notably longer than *MALAT1*, spanning around 23 kilobases compared to *MALAT1*'s 8 kilobases. Similar to other lncRNAs, *NEAT1* and *MALAT1* are transcribed by polymerase II, but unlike any other known lncRNAs, tRNA-processing machinery is involved in their maturation. Specifically, the 3′-end of the genes forms a tRNA-like structure, which is recognised by RNase P and

RNase Z, introducing two cuts before and after this structure, respectively (Fig. 1A) (Wilusz et al. 2008; Sunwoo et al. 2009). After the tRNA-like structure is cut out, the newly formed 3′-end folds into a triple helix, which stabilises the transcript (Brown et al. 2012). The tRNA-like structure (called mascRNA in the *MALAT1* gene) is further processed by another enzyme of tRNA maturation machinery - the CCA-adding enzyme, which can add two CCAs to the 3′-end of the structure instead of a single CCA, triggering its degradation. In five tested human cell lines it was demonstrated that *NEAT1*'s tRNA-like structures degrade in the cytoplasm, while mascRNA remains stable (Wilusz et al. 2008; Wilusz et al. 2011). The triple helix and tRNA-like structures of *MALAT1* are exceptionally conserved and have been detected across a wide range of vertebrates, including zebrafish, lizards, and reptiles (Stadler 2010; Zhang et al. 2017; Monroy-Eklund et al. 2023). Conservation of *NEAT1* tRNA-like structure has also been demonstrated among several mammals (Marz et al. 2014).

NEAT1 encodes two isoforms, the long and the short (Fig. 1A), which we refer to as *NEAT1Long* and *NEAT1Short*, also known as *NEAT1*_1 and *NEAT1*_2, respectively (Naganuma et al. 2012). These isoforms share the 5′-end of the *NEAT1* gene, with the *NEAT1Short* undergoing polyadenylation at approximately 3.7 kb of the gene. To date, it remains an open question whether *NEAT1Short* exists in all mammalian species encoding *NEAT1*.

NEAT1Long is an architectural nuclear-retained RNA, which is an essential component of paraspeckles (Sasaki et al. 2009). These nuclear bodies are built around *NEAT1* and stabilised by proteins of two main classes. Members of the Drosophila behaviour/human splicing (DBHS) family (NONO, SFPQ, PSPC1) are multidomain oligomerising proteins capable of binding nucleic acids (reviewed in Knott et al. 2016). NONO and SFPQ can also recognise secondary structures like stem loops, which can be formed from splice sites or inverted repeats of transposable Alu elements (IRAlu) or G-quadruplexes -

guanine tracks separated by loops organised in layers by Hoogsteen hydrogen bonds (Knott et al. 2016; Simko et al. 2020; Mou et al. 2022). These proteins can form dimers with each other, enriching the diversity of interactions within paraspeckles. Another group of proteins that stabilize paraspeckles contain prion-like domains (Hennig et al. 2015). FUS and RBM14 are examples of essential paraspeckle proteins of this type (Hennig et al. 2015; Fox et al. 2018). The conservation and importance of the individual elements of *NEAT1* recognised by these proteins remain unclear. It is also uncertain how interchangeable these proteins are, as not all proteins in these families have been identified as essential in specific types of cells, with many considered merely important (Fox et al. 2018).

The formation of paraspeckles is linked to the transcription of *NEAT1* molecules. Initially, these molecules coalesce and then recruit multidomain proteins and other paraspeckle components (Mao et al. 2011). The paraspeckle structure consists of two main parts: the inner 'core' and the outer 'shell' (Hirose et al. 2019). These are distinguished by the folding of *NEAT1*, where the 3' and 5' ends are located in the 'shell', while the middle part of the gene forms the 'core', and by the predominant localisation of resident proteins (Hirose et al. 2019). While the paraspeckle structure has been established in multiple cell types, the individual elements responsible for securing the distribution of resident proteins are less understood.

Current data about the conservation of *NEAT1* is inconsistent. In the early phylogenetic study on a diverse but limited set of around eight mammalian genomes, it was demonstrated that *NEAT1* orthologs are identifiable in Eutherians while absent in marsupials, likely due to incomplete assemblies (Stadler 2010). However, between human and mouse *NEAT1* orthologs, there are only a few patches of similarity, although both form functional paraspeckles. Thus, *NEAT1* is an example of an lncRNA where a lack of sequence similarity does not imply a lack of function, like some other lncRNAs

(Pang et al. 2006). This discrepancy — expecting that functional conservation necessarily implies primary sequence conservation — was confirmed by the identification and functional confirmation of *Neat1* in opossum cells (marsupials), where traces of sequence similarity could be found in only 6% of the gene's length. (Cornelis et al. 2016). The fourth mammal in which the *Neat1* gene has been identified and paraspeckles are confirmed is the naked mole-rat (Yamada et al. 2022), although, sequence homology of this ortholog was not analysed in detail. During our research, *NEAT1* orthologs were identified in koala and platypus genomes, as well as in several non-mammalian vertebrates, using a computational approach (Weghorst et al. 2024). The conserved secondary structure could explain the ability of *NEAT1* to form paraspeckles. However, a comparison of the secondary structures of mouse and human short isoforms of *NEAT1* , which include only the first ~4 kb, revealed predominantly different patterns, with only some regions of similarity (Lin et al. 2018). Therefore, the fundamental question of what elements are essential for *NEAT1* function remains open.

MALAT1 is one of the most highly expressed genes in human cells. Like *NEAT1*, it is a nuclear-retained lncRNA. *MALAT1* is located in speckles – another type of nuclear body in close proximity to paraspeckles. Unlike *NEAT1*, *MALAT1* is a highly conserved lncRNA, with orthologs identified in zebrafish and other vertebrates (Stadler 2010; Weghorst et al. 2024). Moreover, the conservation of a large part of the *MALAT1* secondary structure was demonstrated in 51 mammals (McCown et al. 2019).

In this study, we aimed to investigate the structure–function axis of *NEAT1* and *MALAT1* by identifying conserved regions, sequence features, structures, and regulatory elements within a new large collection of orthologs from 545 diverse mammals. Our prediction of *NEAT1Short* isoforms and alternative polyadenylation signals (PAS) underscores the universal presence of the short isoform across all orthologs examined. We also analysed the conservation of transcriptional regulation, triple helix elements, and tRNA-like structures, further consolidating previously known findings. After analysing the overall diversity of *NEAT1* orthologs, we selected 16 the most dissimilar ones, which we called archetypes, for the identification of shared features. The primary sequence of the orthologs was scrutinized for nucleotide composition and the presence and enrichment of various repeats, like transposable elements and short sequence motifs. Our analysis revealed the ubiquitous features likely most critical to paraspeckle function, including GU repeats, recognized by TDP-43, and G-quadruplexes. We identified specific patterns of TEs integration and their role in the evolutionary shaping of *NEAT1* and its function. Overall, our results suggest that certain domains, elements, structures, and RNA processing events in NEAT1 are universally crucial for the function of paraspeckles.

RESULTS

Defining genomic coordinates for *NEAT1* and *MALAT1* orthologs in 545 mammals

Although *NEAT1* orthologs have been reported in a limited number of mammalian species and vertebrates (Stadler 2010, Cornelis et al. 2016, Weghorst et al. 2024), the nucleotide sequences available to us at the start of our study were for only three species: human, mouse, and opossum. With these three dissimilar sequences of the *NEAT1* gene available, we undertook the challenge of identifying *NEAT1* orthologs in mammalian genomic assemblies (Fig. 1A). The developed algorithm relied on synteny of *NEAT1* and *MALAT1*, as well as the high degree of conservation of *MALAT1*. We first searched for orthologs of *MALAT1*. Then, the homology patches of *MALAT1* served as anchoring points for genomic contig selection, and the surrounding regions were explored to locate *NEAT1*. Due to the considerable length of *NEAT1* gene, we separately searched for similarities to fragments containing the TATA-box of the promoter region and the triple helix followed by a tRNA-like structure. The outstanding high degree of conservation of these structures allowed us reliably identify 5′- and 3′- ends of *NEAT1* orthologs, despite the variability in the primary sequence of the gene. We reconstructed

506 *NEAT1* and 469 *MALAT1* gene orthologs (Suppl. Fig. 1A). In total, the identified *NEAT1* and *MALAT1* orthologs originate from 545 mammalian genomes (487 species, 122 families, 24 orders; Suppl. Fig. 1A), 17 of which belong to four orders of marsupials.

To substantiate the gene predictions, we inspected profiles of mapped transcriptomic reads using Genome Browser (Raney et al. 2024, Fig. 1B, C, Suppl. Fig. 1B). We input the established coordinates of both genes, including the short isoform(s), and compared our predictions with the results of transcriptome read mapping. We performed this verification for the most divergent *NEAT1* orthologs (archetypes), for which Genome Browser data were available (Suppl. Fig. 1B), and observed a very good agreement between our predictions and the expression profiles of both genes. Since the remaining orthologs exhibit clear sequence homology to at least one of the archetypes, we assumed that the transcriptomic read mapping pattern would be comparable. To further support our findings, we compared our results to *NEAT1* and *MALAT1* orthologs from the naked mole-rat (Yamada et al. 2022) and koala (Weghorst et al. 2024), with which there was very good agreement (see Methods).

One of the open questions about *NEAT1* is whether a short isoform is present and expressed in other mammals. We attempted to identify *NEAT1Short* isoforms in the orthologs by searching for the positions of the canonical polyadenylation signal (PAS), which comprises an 'AATAAA' motif, and successfully identified a single PAS in all Eutherian. We noted that *NEAT1Short* forms a recognisable, twice-higher pattern in transcriptomic profiles (Fig. 1B, C), which we also observed in the phylogenetically oldest mammal in our collection, *Tachyglossus aculeatus* (order Monotremata, short-beaked echidna, Fig. 1C).

In the opossum (marsupial), Cornelis *et al.* found evidence for two active PASs approximately 500 bp apart (Cornelis et al. 2016), both of which we identified in all marsupials and Monotremata (Fig. 1C). Next, we asked whether an alternative PAS can be found in Eutherians as well. We showed that many orthologs (n = 275, 55%) have alternative PASs in close proximity, located on both sides of the 'main' PAS (± 600 bp, Fig. 1D) and the position of an alternative PAS is taxon-specific (Fig. 1D). Thus, *NEAT1Short* appears to be a ubiquitous mammalian isoform, with evidence for additional alternative isoforms in its vicinity.

Our search algorithm relied on the synteny between *NEAT1* and *MALAT1*. Indeed, 92% of the 428 mammalian genomes containing both *NEAT1* and *MALAT1* had the genes on the same contig. The genes were consistently encoded in close proximity, with the intergenic distance rarely exceeding 60 kb and averaging 36,755.3 ± 9,927.91 bp (Suppl. Fig. 1C). With the exception of two species (*Rousettus madagascariensis* and *Oryctolagus cuniculus*), both genes were encoded on the same strand of DNA. We suspect that the assembly quality may explain this observation, as neither of these assemblies belonged to the GenBank reference set. Overall, we successfully identified large set of *NEAT1* and *MALAT1* orthologs across mammalian taxa for phylogenetic analysis.

Conservation of the triple helix and tRNA-like structures of *NEAT1* and *MALAT1*

Next, we focused on the 3′-end elements—the triple helix and tRNA-like structures. While these structures of *MALAT1* are known to be highly conserved (Zhang et al. 2017), the situation for *NEAT1* was less clear. Overall, we found low divergence between the 3′-ends of both, *NEAT1* and *MALAT1,* orthologs across all mammals (Fig. 2A). The triple helix structure consists of three principal parts: the structure-forming motif itself, a hairpin loop and a linker (Fig. 2A). We found that the conservation of the structure-forming motif was exceptional, with no mismatches in any *NEAT1* or *MALAT1* ortholog (Fig. 2A). However, the sequences of the hairpin loop and the linker displayed clear specificity for *NEAT1* or *MALAT1* and had high sequence variability in *NEAT1*. In *NEAT1* orthologs, they were nearly equal in size (28.7 ± 0.98 bp and 29.8 ± 1.06 bp), whereas the linker of *MALAT1* was one-third shorter (31.36 ± 1.87 bp and 23.59 ± 1.8 bp, Fig. 2A).

Therefore, our results suggest that, for *NEAT1* and *MALAT1* RNA stability, the sequence of the triple-helix-forming motif is the most crucial element—possibly along with the length of the hairpin and the linker.

The conservation degree of the tRNA-like structures of *NEAT1* and *MALAT1* orthologs was high, although *NEAT1* orthologs exhibited slightly greater variation (Fig. 2A). We also analysed patterns of coordinated nucleotide changes in complementary pairs (coevolving) to assess the pressure of purifying selection on the secondary structure of the tRNA-like elements. We found that the secondary structures of both genes are well conserved (Fig. 2B), and the sizes of individual elements, such as hairpin loops, did not vary drastically. Our analysis clearly highlighted the strongest purifying selection on the third hairpin loop of the tRNA-like structures in both genes, suggesting it has higher functional importance. Taken together, the high degree of sequence conservation of these structural elements highlights their critical role in processing and maturation of *NEAT1* and *MALAT1*.

Analysis of promoter and transcriptional control of *NEAT1* and *MALAT1*

The conservation of promoter regions and transcription factors' binding sites across species can highlight the importance of a gene within certain physiological processes. Conversely, variability in transcriptional regulation can suggest functional differences. We began with an analysis of the TATA-box and the downstream promoter area. Overall, this region was more conserved in *NEAT1* orthologs than in *MALAT1*, which is surprising given the opposite, greater primary sequence variability of *NEAT1* compared to *MALAT1* (Fig. 3A). We found that *NEAT1* orthologs in all Eutherians possessed the classical TATA-box sequence 'TATAAA', with greater promoter area diversity observed in marsupials. The variability of the transcription initiation site in *MALAT1* was significantly higher, and it was less variable only within individual mammalian taxa (e.g., Primates, Chiroptera in Fig. 3A). We also noted a higher diversity of TATA-box motifs in *MALAT1*, such as

‘CATAAA’ in the Chiroptera order, and both ‘AATAAA’ and the classical ‘TATAAA’ in Primates.

As a next step, we predicted transcription factor (TF) binding sites within the 1 kb promoter area of *NEAT1* and *MALAT1* orthologs. An individual promoter of *NEAT1* and *MALAT1* orthologs had, on average, 216.4 ± 33 and 168.2 ± 29 TF binding sites, respectively. Although the average number of sites did not differ drastically, we investigated how many of these sites were identified between orthologs. Surprisingly, we observed that only a small number of TF binding sites were shared among the promoters of *MALAT1* orthologs. We applied a rather permissive threshold of 65% of orthologs per gene, resulting in 25 TFs for *MALAT1* and 123 TFs for *NEAT1* (Suppl. Table 1). Among the predicted TF binding sites for *NEAT1* and *MALAT1* orthologs, we identified 15 that overlapped, including EGR1 and SP1, which have been experimentally validated (Li et al. 2015; Che et al. 2021; Kumar and Mishra 2022; Binder et al. 2023; Tian et al. 2023). Additionally, analysis of GO terms suggested regulation by transcription factors associated with the processes many of which have been experimentally validated for both genes (Fig. 3B) supporting the findings of this unique analysis. Overall, our results indicate a higher degree of conservation of the regulatory elements of *NEAT1* transcription compared to *MALAT1*.

Gene length variation of *NEAT1* orthologs

NEAT1 is one of the longest known lncRNA in the human genome (Derrien et al. 2012), and its length may be a crucial parameter for its architectural function in facilitating phase separation and stabilising paraspeckles. However, the length of two studied lncRNAs have not been a primary focus in previous studies. We analysed the distribution of lengths of *NEAT1* orthologs and found that the average length was 21,114.1 ± 2,811.3 bp (only assemblies without gaps were used). However, the difference between the longest and shortest variants was more substantial: 14,505 bp in *Ochotona curzoniae* (plateau pika, Lagomorpha) and 36,456 bp in *Gymnobelideus leadbeateri* (Leadbeater's possum, Diprotodontia). Notably, the lengths of *NEAT1Short* isoforms varied within a much narrower range, 3,415.18 ± 218.9 bp (Fig. 3C), which suggests potential functional importance.

We observed that the length of the *NEAT1Long* isoform and its variation exhibited some taxon-specific patterns (Fig. 3C, D). Marsupials from the Microbiotheria, Diprotodontia, and Dasyuromorphia orders had the longest *Neat1* genes of all mammals, averaging 30,659.9 ± 4,575.1 bp. However, we did not find evidence for a general evolutionary trend of *NEAT1* shortening as an association between gene length and the phylogenetic distance of a species from *Tachyglossus aculeatus*, Monotremata was not pronounced (Spearman's rho = -0.06, p = 0.18). Additionally, *NEAT1* length varied more within some orders, such as Primates and Artiodactyla, compared to Carnivora. The length of *MALAT1* orthologs varied within a narrower range than that of *NEAT1*, 6,986.8 ± 326.78 bp (Fig. 3D), with a taxon-specific pattern. Marsupials, like *NEAT1* orthologs, had the longest *Malat1* gene (8,124.25 ± 449.08 bp), while rodents exhibited the shortest *Malat1* gene (6,653 ± 176.3 bp). Our findings indicate that the exceptional length of *NEAT1* is conserved across mammals, implying a functional role in paraspeckle biology.

***NEAT1* and *MALAT1* orthologs primary sequence diversity and *NEAT1* archetypes**

Our dataset of hundreds of *NEAT1* and *MALAT1* orthologs enabled a unique assessment of their sequence diversity across mammals and provided insight into their evolutionary patterns. In order to do this, we generated a heatmap (Fig. 4A,B) depicting the average nucleotide identity (ANI) between ortholog pairs in an all-vs-all comparison, with mammals ordered according to the phylogenetic tree. For *NEAT1*, this analysis revealed clusters of higher homology with a strong phylogenetic signal, as these clusters corresponded to mammalian orders (yellow arrows, Fig 4B). However, between clusters, the similarity of *NEAT1* orthologs was low, in some cases barely exceeding 20% ANI (highlighted clusters, Fig 4B). The high sequence diversity and low similarity levels limit the applicability of standard phylogenetic methods based on multiple sequence alignment, as such alignments become nearly random for the most divergent sequences.

To simplify the identification of shared gene features that may be functionally important, we selected *NEAT1* orthologs with the lowest sequence similarity to one another, which we refer to as archetypes (Fig. 4C, D). Some archetypes represented large groups of orthologs—for example, human *NEAT1* represented the cluster comprising those from Primates, Chiroptera, Carnivora, Artiodactyla, and Rodentia families other than Muridae and Cricetidae, while mouse *Neat1* served as an archetype for the Muridae and Cricetidae families (Rodentia order). The remaining archetypes originated from Monotremata, Rodentia (4 archetypes), the Lagomorpha order (2 archetypes), Marsupials (2 archetypes), Eulipotyphla (3 archetypes), Hyracoidea, and the Tenrecidae family (Afrosoricida order) (Fig. 4B,D).

Our results confirmed that *MALAT1* is much more conserved than *NEAT1*, with orthologs of Eutherians sharing 60% ANI or higher (Suppl. Fig. 2A, Fig. 4B) and only the orthologs of Marsupialia and Monotremata were more distinct. Overall, the clustering patterns of heatmaps for both genes were very similar, and the *MALAT1* orthologs in species encoding *NEAT1* archetypes were also among the most diverse (Suppl. Fig. 2B, C). Analysis of this subset of *MALAT1* orthologs revealed positions in multiple sequence alignments that were identical among the archetypes, covering approximately 13% of the *MALAT1* sequence (Suppl. Fig. 2D). These findings suggest a high functional importance for the primary sequence of *MALAT1*, particularly its 3′-end.

To estimate the degree of sequence variation of *NEAT1* and *MALAT1*, we compared the averaged ANI of the genes to the averaged ANI of coding sequences (CDSs) and 3′-UTRs of transcripts of orthologs of protein-coding genes in mammals (Fig. 4C). We found that *MALAT1* was nearly as conserved as CDSs, while NEAT1 exhibited conservation levels comparable to 3′-UTRs. Notably, the ANI of *NEAT1Short* was significantly higher than that of the *NEAT1_3.5kb+* region (*NEAT1Long*, downstream of 3.5 kb). The *NEAT1Short* isoform displayed some sequence similarity among archetypes, whereas similarity in the *NEAT1_3.5kb+* region was nearly absent. This is the first systematic analysis comparing the conservation level of *NEAT1Short* to the rest of the gene, with the higher conservation of *NEAT1Short* underscoring its potential functional significance.

Transposable elements integrate into specific regions of NEAT1 and are rarely detected in MALAT1, despite nucleotide composition

Due to the high diversity of the primary sequences of *NEAT1* orthologs, we focused on identifying shared features that could be detected without the use of multiple sequence alignment. We began with the analysis of transposable elements (TEs), which were detected in high numbers in human and mouse *NEAT1* orthologs previously (Vlachogiannis et al. 2021), and found their high diversity and enrichment in almost all *NEAT1* orthologs (Fig. 5A). We also observed that the distribution of TEs along *NEAT1* archetypes was predominantly species-specific (Suppl. Fig. 3B). While our TE identification method depends on how well TEs are studied in specific groups of mammals—which may affect the finding of exact TE types and frequencies—we can still gain a general impression of the importance of TEs in the evolution of *NEAT1*.

Next, we analysed the integration positions of TEs in *NEAT1* orthologs by binning the orthologs into 5% length intervals and counting the number of TEs in each bin (Fig. 5B). Summing the data per taxon, we found that a few taxa exhibited a bimodal distribution of integration sites, around 30-40% and 70-80% of the gene length. These taxa included Carnivora, Artiodactyla, Primates, and Chiroptera orders. However, in Rodentia, TEs were broadly distributed, with a slight preference for the end of the gene (Fig. 5B). Although

NEAT1 is known to be enriched in TEs, this is the first indication that it contains two predominant regions permissive to TE integration without disrupting function.

Regions of self-complementarity can potentially contribute to *NEAT1's* secondary structure formation and paraspeckle stabilisation, however, have not been a focus of previous research. For example, IRAlu elements (SINE) of 3'-end of human *NEAT1*, which are regions of self-complementarity in close proximity, can form stem loops that contribute to *NEAT1* A-to-I modification and paraspeckle assembly via interaction with NONO and SFPQ (Knott et al. 2016; Vlachogiannis et al. 2021). Therefore, we studied the presence of self-complementary regions in *NEAT1* and *MALAT1* in the whole diversity of mammalian orthologs and found that these regions were common in *NEAT1* but not in *MALAT1* (Fig. 5C, D, Suppl. Fig. 3C). Specifically, we identified self-complementary regions in 71% of *NEAT1* orthologs, with 14.68 ± 20.85 regions per ortholog, and *Lophiomys imhausi* (Rodentia) exhibiting the maximum recorded number of 132 regions (Fig. 5D). We observed that some of these possible interactions occurred over long distances, while others were in close proximity, potentially resembling the function of IRAlu elements in human *NEAT1* (Vlachogiannis et al. 2021, Fig. 5D, Suppl. Fig. 3C). Additionally, the self-complementary interactions exhibited taxa-specific pattern highlighting potential evolutionary adaptations in certain mammalian groups (Fig. 5C). This diversity of interactions could be explained by the bimodal pattern of TE distribution, as we also noted that TEs were frequently the sources of these complementary regions. Overall, this is the first indication of the importance of the self-complementary regions associated with TE integration activity in *NEAT1* mammalian orthologs.

Importantly, TEs were rarely localised within *NEAT1Short* isoforms, highlighting their exposure to separate evolutionary pressures. We identified only 49 cases in six mammalian orders (Suppl. Fig. 4A). While it has been shown that mouse *Malat1* contains the SINE B2 element, we found this to be an exception, as our data revealed only 13 orthologs with a single TE (Suppl. Fig. 4B). Most of these TEs were found in Rodentia and they were localised in close proximity to the 5′-end (Suppl. Fig. 4B). Our original findings further highlighted the importance of *MALAT1*'s primary sequence for its function, and systematically showed that it is rarely affected by TE activity.

As SINEs typically integrate into A-T enriched regions (Daniels and Deininger 1985), we analysed nucleotide usage in *NEAT1* and *MALAT1* orthologs to gain mechanistic insight (Suppl. Fig. 5). We found a high enrichment of T and a depletion of C nucleotides in almost all orthologs of both genes. *MALAT1* orthologs additionally exhibited a high proportion of A nucleotides, demonstrating a nucleotide composition potentially more prone to TE integration (Fig. 5E, Suppl. Fig. 5). To determine how these nucleotide proportions relate to other genes, we compared them to CDS and 3′-UTR regions of protein-coding genes in mammals (Fig. 5E). This analysis showed enrichment of C and G nucleotides in CDSs and A and T nucleotides in 3′-UTRs. Additionally, it has been shown that 3′-UTRs are also prone to TE integration (Lagemaat et al. 2003), which aligns well with the nucleotide usage profile which we analysed. We found that *NEAT1* and *MALAT1* had similar composition to 3′-UTRs (genes were within the standard deviation), although *MALAT1* exhibited an even stronger depletion of C nucleotides. Therefore, our analysis uniquely demonstrated that from a sequence composition perspective, *MALAT1* exhibited an exceptionally low TE frequency.

Finally, we analysed nucleotide usage along the sequences of the two genes. We identified peaks of G nucleotide usage at both ends of the *NEAT1* gene, with a more pronounced peak at the 5′-end (Fig. 5F). This pattern was noticeable in almost all archetypes (Suppl. Fig. 6). Overall, the A-T enriched central region of *NEAT1* coincided well with the hot spots of TE integration. In *MALAT1* orthologs, the nucleotide usage pattern differed, showing a peak of A nucleotide usage at the 5′-end of the gene, which correlates with the integration sites of the infrequently detected TEs (Suppl. Fig. 7). In summary, we demonstrated the positional specificity of the high frequency of TE's integration in *NEAT1Long*, which corresponds well to A-T nucleotides enrichment and the presence of self-complementary interactions. In contrast, TE integration was exceptionally low in *NEAT1Short* isoforms and in *MALAT1* orthologs.

G-quadruplexes and binding sites for TDP-43 are common features in archetypes

The next group of features we analysed were short primary sequence patterns. Guanine tracks separated by loops can form G-quadruplexes—secondary structures which, in human *NEAT1* and *MALAT1*, facilitate interactions with NONO (Arun et al., 2020; Mou et al. 2022). We explored the universality of these structures in *NEAT1* and *MALAT1* orthologs beyond humans and predicted them in high numbers in *NEAT1* (19.2 ± 5.9 per ortholog) and *MALAT1* (9.1 ± 1.6 per ortholog).

In the *NEAT1* archetypes, they predominantly localised at both ends, within the 'shell' area of the paraspeckles (Fig. 6A). This observation aligns well with our finding of nucleotide usage at both ends of *NEAT1*, showing enrichment in G nucleotides. We compared frequencies of G-quadruplexes to CDSs and 3′-UTRs of orthologs of protein-coding genes in mammals (Fig. 6B), with length-normalization applied. *NEAT1* and *MALAT1* orthologs contained more G-quadruplexes than most transcripts' parts, especially in some individual orthologs. Our findings point to the significant importance of G-quadruplexes in both genes.

Next, we used *NEAT1* archetypes to identify frequent or systematically recurring sequence motifs that are universally important for potential paraspeckle formation and function. We chose hexamers as an optimum between diversity and uniqueness, given that the 4,096 possible combinations of letters in hexamers are theoretically diverse enough to appear only once or twice in the longest *NEAT1* ortholog, which contains

6,075 hexamers. Longer motifs are more diverse (16,384 combinations of 7-mers), making it less likely to find the same motif in all orthologs.

As a result of hexamer profiling, we identified two groups of motifs that are both frequent and common to all *NEAT1* archetypes. The first group comprised 'GU'-based hexamers ('GUGUGU' and 'UGUGUG'), which are known TDP-43 binding sites (Rot et al. 2017; Modic et al. 2019). These hexamers displayed largely ortholog-specific distribution patterns, with some showing a preference for the 3'-end in certain archetypes (Fig. 6A, Suppl. Fig. 8A). TDP-43, known to localize to the 'shell' region of paraspeckles (West et al. 2016), may bind these motifs. The second group of motifs included 'UCUGUG' and 'CUGUGU' and was found at higher frequency and lower variability in the central region of *NEAT1*, corresponding to the paraspeckle 'core'. While these motifs may also be recognised by TDP-43 (Rot et al. 2017), the difference in distribution patterns suggests distinct regulatory mechanisms and possibly varying binding affinities for TDP-43. Additionally, these motifs can be recognised by other RNA-binding proteins. We noticed that some of the identified hexamers and G-quadruplexes were located within TEs (Suppl. Fig. 8A), emphasising the special role of TEs in shaping *NEAT1*'s biology. Both groups of hexamers, as well as G-quadruplexes, were also observed in non-mammalian *NEAT1* orthologs (Suppl. Fig. 8B, Weghorst et al. 2024), though with greater variability in distribution and abundance.

We summarised the key features of *NEAT1* sequences that are potentially important for paraspeckle function in Figure 6C. This underscores the importance of G-quadruplexes, TDP-43 binding motifs, and self-complementary regions, which can potentially determine the functional interactions with proteins essential for paraspeckle assembly, even in the absence of primary sequence conservation.

Taxa-specific speed of *NEAT1* evolution

This uniquely large collection of *NEAT1* orthologs enabled us to uncover previously unrecognized patterns in its evolutionary development. The divergence of primary sequences of *NEAT1* orthologs cannot be explained solely by the phylogenetic tree and the evolutionary time since the taxa split. We observed this by examining the ANI of orthologs within mammalian orders (Fig. 4A). For example, orthologs of Carnivora or Artiodactyla are highly similar to each other (60–70% ANI), whereas Lagomorpha or Eulipotyphla include several archetypes with ANI lower than 10%. Another notable observation is that Rodentia and Lagomorpha are phylogenetically closer to Primates than to Carnivora, yet orthologs of Primates are much more similar to Carnivora than to those of Rodentia and Lagomorpha. We focused on the Rodentia order, as it comprised the largest number of identified orthologs and exhibited high diversity in their primary sequences (Fig. 7A, B). Within this order, we observed that different families contributed orthologs with varying levels of similarity within a taxon (e.g. Muridae and Cricetidae families). The similarity between taxa could be high for some families (red dashed cluster, Fig. 7A) and very low for others (yellow arrows, Fig. 7A). Thus, taxonomic borders within the Rodentia order also did not adequately explain the variability of *NEAT1* sequences.

Next, we sought to identify possible drivers of *NEAT1* evolution and analysed orthologs originating from six Rodentia genera for which we had at least three species (Fig. 7B). The highest primary sequence variation was detected in the genera *Mus* and *Acomys*, while other genera exhibited relatively high levels of conservation. This difference in the rate of evolution was also visible between the genera. Specifically, the evolutionary divergence of *Sciurus* and *Marmota* occurred earlier than that of *Mus*, *Acomys*, and *Microtus*, yet the similarity of orthologs between *Sciurus* and *Marmota* was the highest (67.3% ANI, Fig. 7B). Although *Mus* and *Acomys* diverged later than either of them from *Microtus*, the similarity level was the same (51% ANI) between *Mus* and *Acomys* and between *Acomys* and *Microtus*. This suggests a high rate of evolution in *Mus* and *Acomys* and a greater conservation in *Microtus*.

In the analysed graphical alignments of *NEAT1* orthologs, we noticed that the most varied regions are frequently associated with sites enriched in TEs (Fig. 7B, see also Suppl. Fig. 9, 10). Our collection of orthologs included several species (e.g. *Mus musculus*), for which multiple genome assemblies existed, resulting in several *Neat1* variants (Suppl. Fig. 9). In these cases, sequence divergence was minimal, and only TE integration events accounted for the differences. Therefore, our results clearly demonstrate that an accelerated mutational process, accompanied by high TE integration activity, was a major driver in the evolutionary shaping of *NEAT1*, with a clear taxon-specific pattern.

One key difference between SINE and LINE elements is that SINEs depend on LINEs for amplification. Moreover, a specific mechanism for SINE excision has not been identified (Batzer and Deininger 2002), suggesting that SINEs remain at their integration site and erode through mutational process (Richardson et al. 2015). However, we identified three rare but clear examples of SINE excision. For example, a SINE shared by the entire *Microtus* genus was excised in *Microtus ochrogaster* (highlighted area, Fig. 7A, Suppl. Fig. 10). This observation suggests the existence of a mechanism for SINE excision and indicates that, overall, TE dynamics is one of the major factors shaping *NEAT1* evolution.

DISCUSSION

NEAT1 is a paradoxical lncRNA: it lacks sequence similarity between orthologs yet retains functionality, as confirmed for the four *NEAT1* archetypes of human, mouse, naked-mole rat and opossum. Here, by leveraging the extensive and diverse dataset of poorly conserved *NEAT1* orthologs, we investigated the factors contributing to its functional conservation by applying a strategy to identify smaller structural elements in phylogenetically diverse orthologs. A conserved feature of the *NEAT1* gene, as indicated by our research, is that it gives rise to three molecules—*NEAT1Long*, *NEAT1Short*, and a tRNA-like structure, which we discuss separately.

Architectural long *NEAT1* isoform

NEAT1 is one of the longest known lncRNA gene in the human genome and possibly in all mammals, as our results suggest. Functionally, longer RNAs enable faster and more efficient condensate formation, as demonstrated using synthetic RNAs (reviewed in Van Treeck and Parker 2018; Garcia-Jove Navarro et al. 2019). Similarly, studies have shown that RNAs isolated from stress granules are significantly longer than cytoplasmic RNAs which do not localize to stress granules (Khong et al. 2017). The remarkable length of *NEAT1*, which has not previously been the focus in relation to its functional impact, may also explain why paraspeckle formation begins immediately at the transcription site, with multidomain stabilising proteins recruited later (Mao et al. 2011). The substantial variation in the length of *NEAT1* orthologs raises an open question about the potential variation in the physical properties of paraspeckles across species, such as differences in the speed of paraspeckle assembly, their linear size and stiffness.

We discovered that GU repeats are a universal feature of *NEAT1* archetypes and are frequently localised near the 3′ end. Previously, we demonstrated that the isoform switch of human and murine *NEAT1* is regulated by TDP-43 (Modic et al. 2019), which is typically localized in the 'shell' region of paraspeckles (West et al. 2016). A decrease in TDP-43 availability, also caused by its sequestration into paraspeckles, prevents polyadenylation of the short isoform (Modic et al. 2019). In the same study, we also showed that GU repeats are the predominant mechanism for TDP-43 sequestration in paraspeckles. Thus, the interaction between *NEAT1* and TDP-43 is clearly a crucial functional aspect of paraspeckles, and our findings provide the first indication that the mechanism of isoform switching via TDP-43 association with GU repeats in *NEAT1* is likely conserved across all mammals.

G-quadruplexes are secondary structures common to all *NEAT1* orthologs in their 3′ and 5′ regions. Functionally, G-quadruplexes are involved in almost all aspects of gene expression regulation, from transcription to translation, in the modification of mRNAs and miRNAs, and in phase separation processes (Asamitsu and Shioda 2021; Dumas et al. 2021). Importantly, we noted that the list of RNA-binding proteins capable of interacting with G-quadruplexes overlaps with paraspeckle proteins (Fox et al. 2018; Bourdon et al. 2023). For example, paraspeckle proteins—HNRNPH3, HNRNPK, RBM14, SMARCA4, NONO, SFPQ, and TDP-43—along with 17 other non-essential proteins, including PSPC1, are all capable of binding to G-quadruplexes. The diversity of paraspeckle proteins that recognise G-quadruplexes suggests the potential for interchangeability in maintaining paraspeckle integrity, which may explain the importance but non-essentiality of certain proteins (Fox et al. 2018). Another protein that potentially binds to G-quadruplexes of *NEAT1* is SRSF1—a protein actively involved in splicing regulation, predominantly localized in nuclear speckles and regulated by *MALAT1* (Romero-Barrios et al. 2018; Yamada et al. 2022; De Silva et al. 2024). Additionally, SRSF1 binds to and stabilizes *NEAT1* RNA, which consequently affects the cell cycle (Zhou et al. 2019). Overall, G-quadruplexes provide a potential mechanism for the recruitment of mRNAs, miRNAs, and proteins to paraspeckles. We also speculate that G-quadruplexes, which are formed by DNA as well (Asamitsu and Shioda 2021; Dumas et al. 2021), may be utilised by G-quadruplex-binding proteins to cross-stitch paraspeckles to DNA.

We found that many *NEAT1* orthologs are characterised by reverse complementary regions, frequently originating from diverse TEs. In human *NEAT1*, IRAlu can form stem-loop structures that attract ADAR enzymes, which modify A-bases to-I, and are potentially bound by NONO (Elbarbary and Maquat 2015; Vlachogiannis et al. 2021). By extension, we postulate that the complementary regions, which we identified in high abundancies in many orthologs, may have the potential to form stem-loop structures and interact with NONO and/or ADAR enzymes. Another possibility is that in cases where complementary regions are interspersed, they might contribute to paraspeckle stabilisation, particularly in the early phase of paraspeckle assembly before the recruitment of multidomain proteins.

Short isoform of *NEAT1*

Our original results highlight the universality of *NEAT1Short* and the higher conservation of its primary sequence and isoform length compared to *NEAT1Long*. We detected only a small number of cases where *NEAT1Short* contained TEs, and overall, *NEAT1Short* was depleted of both simple and more complex repeats. These findings indicate a distinct functional trajectory for *NEAT1Short*, separate from *NEAT1Long*, about which little is currently known. For example, *NEAT1Short* has recently been associated with TIRR, an RNA-binding protein that interacts directly with 53BP1, restricting its access to DNA double-strand breaks and its association with p53 (Kilgas et al. 2024). It has been shown that *NEAT1Short* can be located outside of paraspeckles and concentrated in much smaller foci known as 'microspeckles,' the function of which remains unclear (Li et al. 2017). In experiments conducted by Naveed et al., it was demonstrated that *NEAT1Short* can have an effect on cell proliferation that is opposite to that of *NEAT1Long* (Naveed et al. 2021).

tRNA-like structure

The primary sequences of tRNA-like structures are highly conserved not only within *NEAT1* or *MALAT1* orthologs but also between the two genes. Our dataset, which significantly expands the number and diversity of known *NEAT1* and *MALAT1* sequences and their structural elements, allows for improved identification of the most conserved regions. Comparing coevolving structures in our analysis of 545 species with those identified in a smaller dataset (Marz et al. 2014) highlights the broader diversity of tRNA-like primary sequences within mammals. This higher sequence diversity, in turn, helps pinpoint the most functionally important structural components—specifically, the highly conserved hairpin III (Fig. 2A,B) and the overall tRNA-like conformation, which are likely key elements in the maturation processes of both *NEAT1* and *MALAT1*.

Differences in the conservation levels of tRNA-like structures in *NEAT1* and *MALAT1* orthologs may indicate functional divergence. It has been shown that *MALAT1*'s mascRNA may additionally play a role in cellular metabolism within the cytoplasm. For example, it can contribute to increased protein translation and cell proliferation by binding to the multi-tRNA synthetase complex (Lu et al. 2020). Dissimilarly, *NEAT1*'s tRNA-like molecules were shown to degraded in human cell lines (Wilusz et al. 2008; Wilusz et al. 2011). Based on these differences, it is important to systematically analyse the functions of tRNA-like molecules in different cell types and animals, as they may have been adapted for specific functions.

From conserved transcriptional regulation to *NEAT1*'s role in cell biogenesis

The identification of TF motifs shared by hundreds of mammalian species in the *NEAT1* and *MALAT1* promoters suggests their involvement in specific cellular and molecular pathways. Although our study presents the first large-scale computational prediction of potential biological processes for both genes, we observed a strong concordance between our results and previously reported experimental findings. For example, *NEAT1* has been implicated in apoptosis and proliferation (Adriaens et al. 2016; Kilgas et al. 2024), as well as in diverse neurodegenerative diseases (An et al. 2018), potentially via the same TFs involved in CNS development. The number of paraspeckles (and *NEAT1* expression levels) oscillates with circadian rhythms, releasing IRAlu-containing mRNAs (Torres et al. 2016; Torres et al. 2017) and regulating 53BP1 availability in a cell-cycle-dependent manner (Kilgas et al. 2024). Moreover, *NEAT1* directly binds approximately 30% of all mRNAs located in paraspeckles, most of which are also involved in circadian rhythm cycles (Jacq et al. 2021). Additionally, NONO and SFPQ are known to be involved in circadian rhythm regulation (Kowalska et al. 2012; Knott et al. 2016). Our analysis also suggests a potential role for *NEAT1* and *MALAT1* in spermatogenesis and gonad development, which aligns well with the findings of Zhang *et al.*, demonstrating that many *MALAT1*-like genes in *Anolis carolinensis* are highly expressed in the testis and enriched in the nuclei of round spermatocytes (Zhang et al. 2017).

NEAT1 and *MALAT1*: uniquely similar but different lncRNAs

Our study confirms the synteny of *NEAT1* and *MALAT1* across the full range of mammalian species. The uniqueness and similarity of their gene maturation processes along with their roles in spatially associated nuclear bodies, raise the expectation of similar regulation, conservation, and function for *NEAT1* and *MALAT1*. However, this is not the case: *MALAT1* is a highly conserved lncRNA, while *NEAT1* is more variable.

This difference in conservation is possibly associated with the frequency of TEs integration, as *NEAT1* is more prone to such integrations compared to *MALAT1*. However, our analysis of nucleotide usage highlighted an opposite trend: *MALAT1* has, on average, a more favourable nucleotide composition for TE integration. This further underscores the functional importance of conserved primary sequence of *MALAT1*. It has been shown that two regions in *MALAT1*, located approximately at 2-3 kb and 6-7 kb, are responsible for its localisation in nuclear speckles (Miyagawa et al. 2012), which aligns with our results showing a high level of sequence conservation in these regions. The accumulation of mutations in another conserved region of *MALAT1* (3–4.3 kb) has been associated with breast cancer progression (Ellis et al. 2012), highlighting the importance of an intact primary sequence for proper function under normal physiological conditions. Together, these findings suggest that *MALAT1*'s primary sequence plays a major role in its function, while for *NEAT1*, secondary structural elements appear to be more crucial.

The analysis of the conservation of promoter regions, TATA-boxes, and transcription TF binding sites revealed another key difference between *NEAT1* and *MALAT1*. Although *MALAT1* showed greater gene conservation than *NEAT1*, the variability in *MALAT1*'s promoter region and potential transcriptional regulation was higher. This provides an indication that *MALAT1* may have adapted to different gene networks across species, while *NEAT1* remains a consistent player in the same biological processes.

Uneven speed of *NEAT1* evolution

Our research identified two main mechanisms driving *NEAT1* evolution: divergence due to the accumulation of mutations and the high frequency of TEs integration and excision. It is widely accepted that TEs play a significant role in mammalian evolution (Senft and Macfarlan 2021). Intergenic lncRNAs are much more enriched in TEs compared to protein-coding genes (Hezroni et al. 2015) and the most common TE type in lncRNAs is ERVs, while SINEs and LINEs are depleted (Kelley and Rinn 2012). *NEAT1* is known to be enriched in repeats (Souquere et al. 2010) and here we demonstrate both the diversity and the impact of TEs on *NEAT1* evolution.

TEs influence gene length in both directions—making it longer through integration or shorter through excision—explaining the considerable variation in gene length across mammals. TEs also introduce self-complementary regions, stabilising paraspeckles, as well as repeats and G-quadruplexes, which serve as interaction sites for key resident proteins. This observation highlights the benefits of TEs integration for *NEAT1* function within paraspeckles. However, the bimodal pattern of TEs integration hot spots supports the idea that *NEAT1* cannot tolerate insertions throughout its sequence—particularly not within the 5′-end shared with the *NEAT1Short* isoform. Therefore, TEs play a crucial role in *NEAT1* evolution overall.

The consequences of TE integration into lncRNAs are variable, and *NEAT1* is not unique in being shaped by TEs. For example, TE insertions in the *ANRIL* lncRNA have been linked to increased gene conservation in primates (He et al. 2013). TEs also support *ANRIL*'s function in the trans-activation of a range of target genes, some of which are contributing to coronary artery disease (Holdt et al. 2013; Alfeghaly et al. 2021). *XIST*, another lncRNA enriched in TEs, provides further evidence of functional adaptation—where TEs have contributed to the formation of specific exons (Elisaphenko et al. 2008).

Our analysis highlighted taxa with accelerated *NEAT1* evolution, such as Eulipotyphla, Lagomorpha, and the *Mus* and *Acomys* genera of the Rodentia order. This phenomenon of varied evolutionary speed has been previously demonstrated for some lncRNAs. For example, unannotated and largely non-coding human accelerated regions (Pollard, Salama, Lambert, et al. 2006; Pollard, Salama, King, et al. 2006) are conserved genomic regions across mammals that accumulate disproportionately more mutations in humans, many of which function as enhancers in neurodevelopment (Doan et al. 2016; Girskis et al. 2021). Although signs of positive selection in local secondary structures of human *NEAT1* have been reported (Walter Costa et al. 2019), our data do not support the hypothesis of accelerated evolution of *NEAT1* in the human lineage. The rate of evolution highlights species or taxon-specific adaptations to their ecological niches. We speculate that this mechanism may also influence *NEAT1* biogenesis, as *NEAT1* can directly interact with diverse mRNAs and miRNAs, possibly via complementary interactions of primary sequences. This may explain the high evolutionary speed observed in certain taxa and across mammals in general.

MATERIAL AND METHODS

Identification of coordinates for *NEAT1* and *MALAT1* orthologs

Mammalian genomes were downloaded from GenBank (Clark et al. 2016, July 2023). Annotated *NEAT1* and *MALAT1* orthologs from *Homo sapiens* (NR_131012.1), *Mus musculus* (NR_131212.1, O'Leary et al. 2016), and *Monodelphis domestica* (KX036207.1, Cornelis et al. 2016) were used for similarity searches and the identification of orthologs in the downloaded genomes. We additionally retrieved promoter regions and triple helix motifs, followed by tRNA-like structure sequences, for these annotated orthologs using in-house scripts. These sequences were subjected to a blastn (Altschul et al. 1990) search against the downloaded mammalian genomes. Approximate gene coordinates were obtained from the homology search results and were complemented with some manual curation in cases where *NEAT1* and *MALAT1* orthologs were found on different contigs due to fragmentary assembly. Genes were retrieved with some sequence excess at both the 5′- and 3′-ends and subjected to multiple sequence alignment (MSA, MAFFT, v7.487, Katoh and Standley 2013), default parameters). Since *NEAT1* showed noticeably higher divergence compared to *MALAT1*, we divided the mammals into eight groups according to the phylogenetic tree (Ns et al. 2019).

Group1: Monotremata, Didelphimorphia, Microbiotheria, Diprotodontia, Dasyuromorphia

Group2: Eulipotyphla, Perissodactyla, Pholidota

Group3: Macroscelidea, Pilosa, Proboscidea, Afrosoricida, Cingulata, Sirenia, Tubulidentata, Hyracoidea

Group 4: Lagomorpha, Rodentia, Scandentia

Group 5: Primates, Dermoptera

Group 6: Artiodactyla

Group 7: Chiroptera

Group 8: Carnivora

We then added the most relevant, phylogenetically closest annotated *NEAT1* ortholog(s) to these groups and performed MSA. MSA was visualised using the online tool AlignmentViewer (https://alignmentviewer.org/). The coordinates of the genes' start and stop sites (TATA-box and end of the triple helix) within the MSA were identified and used to build the final set of orthologs and their structural elements. The same procedure was applied for the *MALAT1* ortholog search, but sequences were divided into two groups: the aforementioned Group 1 and the remaining sequences.

Subsequently, we manually curated the results and removed orthologs with excessive assembly gaps or spurious sequences lacking the correct start or end. Coordinates, contig accessions, genome assembly versions, and other results and metadata can be found in Supplementary Table 1.

We examined the strand and genomic distance between *NEAT1* and *MALAT1*. Out of 428 organisms in which both genes were predicted, 92% had these genes located on the same contig. Since not all species possess complete chromosome-level assemblies, some genes were found on different contigs. In such cases, it is not possible to determine the true genomic positions of the genes. Among those located on the same contig, only two species had *NEAT1* and *MALAT1* coded on opposite strands. Assemblies of both these species, *Rousettus madagascariensis* and *Oryctolagus cuniculus*, do not belong to the GenBank reference set. After manual inspection, we found another reference assembly for *Oryctolagus cuniculus* (Suppl. Fig. 1B) and checked the strand and location of the predicted orthologs of *NEAT1* and *MALAT1*. Although these orthologs showed high sequence similarity to the reference assembly, the directionality of the genes was different: they were coded on the same strand, consistent with the majority of other orthologs. However, we cannot assess the impact of assembly quality on the opposing directionality observed in *Rousettus madagascariensis*, as no reference assembly is currently available.

Comparison of *NEAT1* and *MALAT1* orthologs to the results of Yamada *et al* and Weghorst *et al*

Our collection included a newer version of the naked mole-rat genome assembly than the one used in the publication by Yamada *et al*. (Yamada et al. 2022). We downloaded the assembly used in that study and performed a blastn search of the *Neat1* sequence identified in our study against this assembly. Our start coordinate for *Neat1* was 20,972,753, which is 204 bp downstream of the start coordinate reported by Yamada *et al*. (JH602080:20,972,549, with both genes coded on the minus strand). The start coordinate for *Malat1* was 20,907,466, which is 60 bp downstream of the coordinate reported by Yamada *et al*. (JH602080:20,907,406). We assume that the 3′-ends of the genes in the naked mole-rat are identical to those we identified, as Yamada *et al.* also defined them computationally based on the similarity of triple helix and tRNA-like motifs.

The higher agreement we found for the koala *Neat1* (Weghorst et al. 2024), where the starting coordinate differed by 12 bp only and the 3′-end was the same. The coordinates for koala *Malat1* were identical.

Prediction of short isoform in *NEAT1* orthologs

We divided the orthologs into two similarity groups, with marsupials and Monotremata (Group 1) sequences placed separately. The remaining orthologs were subjected to MSA (MAFFT, default parameters). We identified the position in the MSA corresponding to the PAS of human *NEAT1Short* and searched for the predicted PAS in the vicinity of this position in the orthologs. Polyadenylation signals were predicted by searching for the canonical motif ‘AATAAA’. If a single signal was detected within 110 bp (in both directions) of the PAS position in human *NEAT1*, it was considered an active PAS for the *NEAT1Short* orthologs.

In Group 1, we searched for two PASs using the same logic, based on the predicted sites for *Monodelphis domestica* (Cornelis et al. 2016). In this prediction, there were three orthologs where we could not identify a single alternative polyadenylation signal, and these were omitted from the analysis.

Prediction of TEs in *NEAT1* and *MALAT1* orthologs

We used the DFAM database (downloaded in April 2022, Storer et al. 2021) of transposable elements and searched for similarities using blastn algorithm and applying the 80-80-80 rule (a minimum alignment length of 80 bp with 80% nucleotide identity over an alignment covering at least 80% of the TE). Only non-overlapping TE annotations were selected.

Using this method, we identified four large fragments of LINE elements and six complete SINEs in human *NEAT1*: four *Alu* elements and two FLAM-C elements. Two of the identified *Alu* elements, *AluSx3* (17,804–18,067 bp) and *AluJr* (17,532–17,678 bp), can form an IRAlu secondary stem-loop structure, which may attract ADARs for A-to-I modification of *NEAT1* (Vlachogiannis et al. 2021). In mouse *NEAT1*, we identified four SINE elements (*B1_Mus1*, *B3*, *B1_Mm*, and *B1_Mus2*), which are non-complementary to each other.

Prediction of sequence elements in *NEAT1* and *MALAT1* orthologs

We retrieved 1kb of promoter sequence for each ortholog of *NEAT1* and *MALAT1* and predicted transcription factors binding sites using FIMO tool (Grant et al. 2011, part of MEME package v5.0.5, Bailey et al. 2015) and JASPAR database (core part, version 2022, vertebrates, Rauluseviciute et al. 2024). Sites with p-value < 10^{-4} were considered. GO terms were downloaded in October 2021 (Ashburner et al. 2000; The Gene Ontology Consortium et al. 2023), each gene was associated with all connected to it terms.

G-quadruplexes were predicted using pqsfinder R package (v.2.2.0, Hon et al. 2017).

Kmers were counted using in house script.

Self-complementary regions were assessed from the blastn search against an ortholog itself, and only reverse complementary hits were counted.

Average nucleotide identity calculation

ANI between two sequences was calculated by using all blastn hits longer than 100bp and following the formula:

$$ANI = \left(\frac{\sum(blastn\ hits\ Gene1)}{Length\ Gene1} + \frac{\sum(blastn\ hits\ Gene2)}{Length\ Gene2}\right)/2 * 100$$

where

$$blastn\ hit = \frac{blast\ pident}{100} * Length\ blast\ HSP$$

Analysis of CDS and UTR regions of protein-coding gene orthologs in mammals

CDS and UTR regions of protein-coding gene orthologs in mammals were retrieved from GenBank (Clark et al. 2016) in September 2024 using the NCBI Datasets tool (Na et al. 2024, command: *datasets download gene symbol "$GENE" --ortholog mammals --include gene,cds,3p-utr,product-report*; for genes with at least 150 orthologs from different genera of our collection). In cases where multiple transcripts were available, the longest single transcript per ortholog was selected. ANI, G-quadruplexes, and nucleotide usage were predicted in the same manner as for *NEAT1* and *MALAT1* orthologs; values were averaged across all orthologs per gene before being used in distribution plots. A total of 15,461 protein-coding genes were included in the CDS analysis, and 13,847 genes were used in the UTR analysis.

Phylogenetic tree

The phylogenetic tree of Ns *et al.* (Ns et al. 2019) was used. Species for which *NEAT1* and *MALAT1* orthologs were identified but absent in the phylogenetic tree were associated with their closest relatives. A full list of these connections can be found in the

Part3_PhylogeneticTree python notebook (https://github.com/kseniaarkhipova/NEAT1-MALAT1). Visualisation and graphical adjustments of the phylogenetic tree were made using the iTOL web server (Letunic and Bork 2024). Tree parsing, pruning, and the retrieval of time information were performed using the ete3 Python package (v.3.1.2, Huerta-Cepas et al. 2016).

Other used resources and software

RNAfold web-server (Lorenz et al. 2011) was used to predict and visualise folding of structural elements. LocRNA software (v. 2.0.0, http://rna.informatik.uni-freiburg.de, Will et al. 2012) was used to analyse coevolutionary patterns of tRNA-like structures. Sequence logos were generated using WebLogo web-server (Crooks et al. 2004). Taxonomic tree of NCBI (downloaded on April 2022, Schoch et al. 2020) was used to classify the studied genomes. Most of analysis was performed with customs scripts, which were written in Python 3.7.0 and used the following packages: scipy (v.1.7.1, Virtanen et al. 2020), numpy (v. 1.18.5, Harris et al. 2020), pandas (v 1.1.5, https://zenodo.org/records/10957263), matplotlib (v.3.4.3, Hunter 2007), seaborn (v. 0.11.2, Waskom 2021) and Jupyter notebook (v.4.8.1, Kluyver et al. 2016). Code and orthologs sequences are available on GitHub (https://github.com/kseniaarkhipova/NEAT1-MALAT1, DOI:0.5281/zenodo.15147921).

ACKNOWLEDGEMENTS

We would like to thank the John Templeton Foundation for their generous support through Grant #62572.

AUTHOR CONTRIBUTIONS

Ksenia Arkhipova: performed the research, writing—original draft and editing. Micha Drukker: conceived the study, writing—review.

CONFLICT OF INTEREST

The authors declare no competing interests or financial conflicts related to this research.

FUNDING

The John Templeton Foundation [grant number #62572].

DATA AVAILABILITY

Code and orthologs sequences are available on GitHub (https://github.com/kseniaarkhipova/NEAT1-MALAT1, DOI:0.5281/zenodo.15147921).

FIGURE LEGENDS

Figure 1. Identification of *NEAT1* and *MALAT1* orthologs.

A. The organisation of *NEAT1* and *MALAT1* genes and the logic of orthologs' coordinates identification. Promoter areas, TATA-boxes, tRNA-like structures, and triple helices are highlighted, with colours used uniformly throughout the scheme. Genomic regions lacking sequence similarity to confirmed orthologs are depicted in black.

B. Confirmation of *NEAT1* and *MALAT1* ortholog predictions in *Tachyglossus aculeatus* (Monotremata order, short-beaked echidna)—the phylogenetically oldest species in our collection. Predicted coordinates were overlaid (shaded areas) on mapped transcriptomic read profiles in the Genome Browser (*http://genome.ucsc.edu*, Raney et al. 2024). In the 'Genes' section of the Genome Browser, the automatically predicted genes identified in the region are shown. *Neat1* and *Malat1* are coded on the minus strand, with transcription direction indicated by arrows.

C. Two predicted PASs in *Tachyglossus aculeatus*. Zoom-in view on the transcription profiles of *Neat1* in *Tachyglossus aculeatus* near the 3′-end of the *Neat1*Short isoform. The coordinates of the main and alternative PASs are overlaid. The primary PAS corresponds more closely to the drop in transcriptomic reads.

D. Location and taxonomic distribution of alternative PASs in mammals. Most species possess an alternative PAS within 600 bp up- or downstream of the main PAS.

Figure 2. Conservation of 3′-end motifs of *NEAT1* and *MALAT1* orthologs.

A. Secondary structure and sequence diversity of triple helices and tRNA-like structures in *NEAT1* and *MALAT1* orthologs. Secondary structures of the human triple helix and tRNA-like structure are shown at the top of the figure, with individual structural elements highlighted. Colours are used consistently throughout the figure. An example of the multiple sequence alignment of 3′-end structures of both *NEAT1* and *MALAT1* orthologs from the listed randomly selected species is depicted. The summary of sequence diversity across all orthologs is presented as a coloured sequence conservation logo. The variance in length (mean ± std) of hairpins I and II of triple helices in *MALAT1* and *NEAT1* orthologs is specified. Highly conserved triple-helix-forming sequence regions are highlighted in both the alignment and logo figures.

B. Co-evolving patterns of tRNA-like structures across all *NEAT1* and *MALAT1* orthologs in Eutherians. The most conserved base pairs are shown in dark red. High-intensity yellow and green indicate perfectly matching alternative base pairs (coevolving) in the MSA. The co-evolving patterns of the tRNA-like structure of *MALAT1* exhibit a much higher level of conservation in the whole secondary structure, while the tRNA-like structure of *NEAT1* mainly involves hairpin III with a highly variable hairpin II.

Figure 3. Transcriptional regulation of *NEAT1* and *MALAT1* orthologs and their length distribution.

A. Conservation of TATA-boxes and promoter areas in *NEAT1* and *MALAT1* orthologs (sequence logo). TATA-boxes are highlighted with grey boxes, and the transcription start site is marked by an arrow.

B. The most frequent GO terms associated with transcription factors, the binding sites of which were identified in at least 65% of orthologs of *NEAT1* and *MALAT1*. Shared biological processes associated with the same TF are depicted as overlaps.

C. Average ortholog length and its variation across mammals. Marsupials exhibit the longest average length for *Neat1* and *Malat1*, while the *NEAT1Short* isoform shows much smaller length variation compared to *NEAT1Long*. Only non-gapped ortholog assemblies are taken into account.

D. Length distribution of *NEAT1* and *MALAT1* orthologs in mammalian orders, arranged along a time-scaled phylogenetic tree. Only orthologs with non-gapped gene assemblies were used. The number of orthologs used for the assessment is indicated on the bars.

Figure 4. Primary sequence diversity of *NEAT1* and *MALAT1* orthologs in mammals.

A. Schematic representation of how average nucleotide identity (ANI) was calculated between pairs of genes and visualised as heatmaps. Patches of similarity from pairwise blastn alignments were normalised to the length of individual genes, averaged for both, and expressed as a percentage (see Methods). The obtained percentages were assigned corresponding colours and plotted in an all-to-all heatmap.

B. Heatmap of ortholog similarity in pairwise comparisons (all-to-all). Orthologs are arranged along a phylogenetic tree, with the colour bar on the left indicating the mammalian orders of individual orthologs; colours are explained in the legend. For visual clarity, the phylogenetic tree was simplified, and phylogenetically estimated divergence times were omitted. Red clusters represent groups of highly similar orthologs, which align well with mammalian orders (yellow arrows), while dark blue areas indicate a lack of similarity. The three largest similarity clusters are framed, and the low sequence similarity between them is highlighted with double-sided arrows and ANI values. On the right side of the heatmap, the positions of archetypes are marked. The colour code indicates the availability of RNA-seq data for the predicted gene regions in Figure 1B and Supplementary Figure 1B: red – data available, blue – data not available, yellow – human and mouse genes.

C. Bar plot of averaged ANI for specified groups of orthologs or genes, estimated in pairwise all-to-all comparisons. In addition to the two *NEAT1* isoforms, we also present *NEAT1*_3.5kb+—a part of *NEAT1Long* excluding the 5′-end of the gene, which is shared with *NEAT1Short*. Archetypes refers to a subset of 16 of the most diverged *NEAT1* orthologs. For this species subset, we estimated the average ANI of *MALAT1* orthologs and of protein-coding genes (see Methods). The averaged ANI of two structural parts of transcripts of protein-coding genes were included for comparison—CDS regions (15,461 orthologous genes were used) and 3′-UTR regions (n=13,847).

D. Heatmap of primary sequence similarity among *NEAT1* archetypes. Orthologs are arranged along a phylogenetic tree, and the colour bar on the left side corresponds to the mammalian orders of individual orthologs; colours are explained in the legend. The phylogenetic tree is time-scaled.

Figure 5. Transposable elements contribute to *NEAT1* sequence diversity.

A. Bar plot of the average number of TEs and their type per ortholog across mammalian orders. The phylogenetic tree is not time-scaled.

B. Distribution of TEs within *NEAT1* orthologs. The length of individual orthologs was binned into 5% segments, and the number of annotated TEs within each bin was summed for all orthologs per mammalian order. Two segments (30-40% and 70-80% bins), which most frequently contain TEs, are highlighted in green.

C. Bar plot showing the number of self-complementary regions per ortholog, averaged per mammalian order. The number of orthologs used in the assessment is indicated on the left.

D. Graphical representation of the distribution of self-complementary regions in four selected *NEAT1* orthologs. Each ortholog is aligned to itself, and reverse complementary regions are connected with lines, forming a visual 'cross' shape. The second 'cross' (around 20kb) in human *NEAT1* corresponds to an IRAlu-formed hairpin. Self-complementary regions often coincide with TEs and can occur over short distances, resembling the IRAlu of human *NEAT1*, or over much longer distances. Additional plots are in Suppl. Fig. 3C.

E. Averaged nucleotide usage (nucleotide composition of a molecule estimated as a percentage) of *NEAT1* and *MALAT1* orthologs compared to the averaged nucleotide usage of CDS and 3′-UTRs of protein-coding genes. The nucleotide usage of *NEAT1* and *MALAT1* is overlaid on coloured bars representing the nucleotide usage of CDSs (grey bars) and 3′-UTRs (orange bars). The standard deviation whisker is shifted for visual clarity.

F. Plot of nucleotide usage of human *NEAT1* along the sequence. The length of the ortholog was binned into 5% segments, and the nucleotide usage of each bin was estimated. The average nucleotide usage of human *NEAT1* is depicted with dashed lines. Additional plots for *NEAT1* are in Suppl. Fig. 6, and for *MALAT1* in Suppl. Fig. 7.

Figure 6. Simple and complex repeats in *NEAT1* and *MALAT1* orthologs.

A. Distribution of G-quadruplexes and two groups of hexamers in *NEAT1* archetypes. Each ortholog's length was divided into 10% bins, and the number of detected elements was summed per bin. In the bottom part of the panel the distribution of the studied elements is summarised for all the archetypes (mean ± std).

B. Frequency of G-quadruplex detection in CDSs and 3′-UTRs of protein-coding transcripts. The number of detected G-quadruplexes per kb in orthologs was averaged per gene and used in the plot.

C. Summary of identified conserved features of *NEAT1*, potentially contributing to the function and stabilisation of paraspeckles. I. Distribution and interactions of G-quadruplexes with NONO and potentially other proteins of the DSHS family. The structure of a G-quadruplex is depicted in the zoom-in insert. II. The predominant distribution pattern of two universal hexamer sequence motifs potentially recognised by TDP-43 and other RNA-binding protein(s). III. Bimodal pattern of TEs integration, frequently associated with self-complementary interactions in close proximity, potentially forming IRAlu-like structures possibly recognised by DSHS family proteins, and also at long-range distances, possibly facilitating *NEAT1* conformation and paraspeckle stabilisation.

Figure 7. Uneven speed of *NEAT1* evolution.

A. Heatmap of primary sequence similarity between *NEAT1* orthologs of the Rodentia order, in pairwise all-to-all comparison. Clusters of red represent groups of highly similar orthologs which aligns well to family borders (yellow arrows), while dark blue areas indicate a lack of similarity between them. The colour bar on the left represents individual families, as seen in the legend. Six selected genera for the alignment in part B are highlighted with yellow frames. The red dashed frame highlights the similarity cluster of 13 Rodentia families mentioned in the Results section.

B. Graphical representation of the pairwise alignment of *NEAT1* orthologs from six Rodentia genera. Orthologs are aligned according to the time-scaled phylogenetic tree on the left. Individual genera are highlighted in grey for visual clarity. *Mus* and *Acomys* genera exhibit a higher evolutionary rate compared to the others. An example of an excised SINE element is highlighted with a dashed frame.

REFERENCES

Adriaens C, Standaert L, Barra J, Latil M, Verfaillie A, Kalev P, Boeckx B, Wijnhoven PWG, Radaelli E, Vermi W, et al. 2016. p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nat. Med.* 22:861–868.

Alfeghaly C, Sanchez A, Rouget R, Thuillier Q, Igel-Bourguignon V, Marchand V, Branlant C, Motorin Y, Behm-Ansmant I, Maenner S. 2021. Implication of repeat insertion domains in the trans-activity of the long non-coding RNA ANRIL. *Nucleic Acids Res.* 49:4954–4970.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

An H, Williams NG, Shelkovnikova TA. 2018. NEAT1 and paraspeckles in neurodegenerative diseases: A missing lnc found? *Non-Coding RNA Res.* 3:243–252.

Asamitsu S, Shioda N. 2021. Potential roles of G-quadruplex structures in RNA granules for physiological and pathological phase separation. *J. Biochem. (Tokyo)* 169:527–533.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25:25–29.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Res.* 43:W39–W49.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3:370–379.

Binder M, Lasho TL, Ismail WM, Ben-Crentsil NA, Fernandez JA, Kim M, Geyer SM, Mazzone A, Finke CM, Mangaonkar AA, et al. 2023. Abstract 806: Enhancer deregulation inTET2-mutant clonal hematopoiesis is associated with increased COVID-19 severity and mortality. *Cancer Res.* 83:806.

Bourdon S, Herviou P, Dumas L, Destefanis E, Zen A, Cammas A, Millevoi S, Dassi E. 2023. QUADRatlas: the RNA G-quadruplex and RG4-binding proteins database. *Nucleic Acids Res.* 51:D240–D247.

Brown JA, Valenstein ML, Yario TA, Tycowski KT, Steitz JA. 2012. Formation of triple-helical structures by the 3′-end sequences of MALAT1 and MENβ noncoding RNAs. *Proc. Natl. Acad. Sci.* 109:19202–19207.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.

Che F, Ye X, Wang Y, Ma S, Wang X. 2021. Lnc NEAT1/miR-29b-3p/Sp1 form a positive feedback loop and modulate bortezomib resistance in human multiple myeloma cells. *Eur. J. Pharmacol.* 891:173752.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. *Nucleic Acids Res.* 44:D67-72.

Cornelis G, Souquere S, Vernochet C, Heidmann T, Pierron G. 2016. Functional conservation of the lncRNA NEAT1 in the ancestrally diverged marsupial lineage: Evidence for NEAT1 expression and associated paraspeckle assembly during late gestation in the opossum Monodelphis domestica. *RNA Biol.* 13:826–836.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.

Daniels GR, Deininger PL. 1985. Integration site preferences of the Alu family and similar repetitive DNA sequences. *Nucleic Acids Res.* 13:8939–8954.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* 22:1775–1789.

De Silva NIU, Lehman N, Fargason T, Paul T, Zhang Z, Zhang J. 2024. Unearthing a novel function of SRSF1 in binding and unfolding of RNA G-quadruplexes. *Nucleic Acids Res.* 52:4676–4690.

Doan RN, Bae B-I, Cubelos B, Chang C, Hossain AA, Al-Saad S, Mukaddes NM, Oner O, Al-Saffar M, Balkhy S, et al. 2016. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* 167:341-354.e12.

Dumas L, Herviou P, Dassi E, Cammas A, Millevoi S. 2021. G-Quadruplexes in RNA Biology: Recent Advances and Future Directions. *Trends Biochem. Sci.* 46:270–283.

Elbarbary RA, Maquat LE. 2015. CARMing down the SINEs of anarchy: two paths to freedom from paraspeckle detention. *Genes Dev.* 29:687–689.

Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, Zakian SM. 2008. A Dual Origin of the Xist Gene from a Protein-Coding Gene and a Set of Transposable Elements. *PLOS ONE* 3:e2521.

Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC, et al. 2012. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* 486:353–360.

Fox AH, Nakagawa S, Hirose T, Bond CS. 2018. Paraspeckles: Where Long Noncoding RNA Meets Phase Separation. *Trends Biochem. Sci.* 43:124–135.

Garcia-Jove Navarro M, Kashida S, Chouaib R, Souquere S, Pierron G, Weil D, Gueroui Z. 2019. RNA is a critical element for the sizing and the composition of phase-separated RNA–protein condensates. *Nat. Commun.* 10:3230.

Girskis KM, Stergachis AB, DeGennaro EM, Doan RN, Qian X, Johnson MB, Wang PP, Sejourne GM, Nagy MA, Pollina EA, et al. 2021. Rewiring of human neurodevelopmental gene regulatory programs by human accelerated regions. *Neuron* 109:3239-3251.e7.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27:1017–1018.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy. *Nature* 585:357–362.

He S, Gu W, Li Y, Zhu H. 2013. ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. *BMC Evol. Biol.* 13:247.

Hennig S, Kong G, Mannen T, Sadowska A, Kobelke S, Blythe A, Knott GJ, Iyer KS, Ho D, Newcombe EA, et al. 2015. Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles. *J. Cell Biol.* 210:529–539.

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep.* 11:1110–1122.

Hirose T, Yamazaki T, Nakagawa S. 2019. Molecular anatomy of the architectural NEAT1 noncoding RNA: The domains, interactors, and biogenesis pathway required to build phase-separated nuclear paraspeckles. *WIREs RNA* 10:e1545.

Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, Finstermeier K, Stahringer A, Wilfert W, Beutner F, et al. 2013. Alu Elements in ANRIL Non-Coding RNA at Chromosome 9p21 Modulate Atherogenic Cell Functions through Trans-Regulation of Gene Networks. *PLOS Genet.* 9:e1003588.

Hon J, Martínek T, Zendulka J, Lexa M. 2017. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* 33:3373–3379.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* 33:1635–1638.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9:90–95.

Jacq A, Becquet D, Guillen S, Boyer B, Bello-Goutierrez M-M, Franc J-L, François-Bellan A-M. 2021. Direct RNA–RNA interaction between Neat1 and RNA targets, as a mechanism for RNAs paraspeckle retention. *RNA Biol.* 18:2016–2027.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30:772–780.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* 13:R107.

Khong A, Matheny T, Jain S, Mitchell SF, Wheeler JR, Parker R. 2017. The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Mol. Cell* 68:808-820.e5.

Kilgas S, Syed A, Toolan-Kerr P, Swift ML, Roychoudhury S, Sarkar A, Wilkins S, Quigley M, Poetsch AR, Botuyan MV, et al. 2024. NEAT1 modulates the TIRR/53BP1 complex to maintain genome integrity. *Nat. Commun.* 15:8438.

Kluyver T, Ragan-Kelley B, P&#233, Rez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, et al. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press. p. 87–90. Available from: https://ebooks.iospress.nl/doi/10.3233/978-1-61499-649-1-87

Knott GJ, Bond CS, Fox AH. 2016. The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. *Nucleic Acids Res.* 44:3989–4004.

Kowalska E, Ripperger JA, Muheim C, Maier B, Kurihara Y, Fox AH, Kramer A, Brown SA. 2012. Distinct Roles of DBHS Family Members in the Circadian Transcriptional Feedback Loop. *Mol. Cell. Biol.* 32:4585–4594.

Kumar S, Mishra S. 2022. MALAT1 as master regulator of biomarkers predictive of pan-cancer multi-drug resistance in the context of recalcitrant NRAS signaling pathway identified using systems-oriented approach. *Sci. Rep.* 12:7540.

Lagemaat LN van de, Landry J-R, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19:530–536.

Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* 52:W78–W82.

Li R, Harvey AR, Hodgetts SI, Fox AH. 2017. Functional dissection of NEAT1 using genome editing reveals substantial localization of the NEAT1_1 isoform outside paraspeckles. *RNA* 23:872–881.

Li S, Wang Q, Qiang Q, Shan H, Shi M, Chen B, Zhao S, Yuan L. 2015. Sp1-mediated transcriptional regulation of MALAT1 plays a critical role in tumor. *J. Cancer Res. Clin. Oncol.* 141:1909–1920.

Lin Y, Schmidt BF, Bruchez MP, McManus CJ. 2018. Structural analyses of NEAT1 lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture. *Nucleic Acids Res.* 46:3742–3752.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6:26.

Lu X, Huang J, Wu S, Zheng Q, Liu P, Feng H, Su X, Fu H, Xi Q, Wang G. 2020. The tRNA-like small noncoding RNA mascRNA promotes global protein translation. *EMBO Rep.* 21:e49684.

Mao YS, Sunwoo H, Zhang B, Spector DL. 2011. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat. Cell Biol.* 13:95–101.

Marz M, Wehner S, Stadler PF. 2014. Homology Search for Small Structured Non-coding RNAs. In: Handbook of RNA Biochemistry. John Wiley & Sons, Ltd. p. 619–632. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527647064.ch29

McCown PJ, Wang MC, Jaeger L, Brown JA. 2019. Secondary Structural Model of Human MALAT1 Reveals Multiple Structure–Function Relationships. *Int. J. Mol. Sci.* 20:5610.

Miyagawa R, Tano K, Mizuno R, Nakamura Y, Ijiri K, Rakwal R, Shibato J, Masuo Y, Mayeda A, Hirose T, et al. 2012. Identification of cis- and trans-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. *RNA* 18:738–751.

Modic M, Grosch M, Rot G, Schirge S, Lepko T, Yamazaki T, Lee FCY, Rusha E, Shaposhnikov D, Palo M, et al. 2019. Cross-Regulation between TDP-43 and Paraspeckles Promotes Pluripotency-Differentiation Transition. *Mol. Cell* 74:951-965.e13.

Monroy-Eklund A, Taylor C, Weidmann CA, Burch C, Laederach A. 2023. Structural analysis of MALAT1 long noncoding RNA in cells and in evolution. *RNA* 29:691–704.

Mou X, Liew SW, Kwok CK. 2022. Identification and targeting of G-quadruplex structures in MALAT1 long non-coding RNA. *Nucleic Acids Res.* 50:397–410.

Na O, E C, Jb H, Wr A, R F, V H, Mtn T, Gd S, X Z, J Torcivia, et al. 2024. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* [Internet] 11. Available from: https://pubmed.ncbi.nlm.nih.gov/38969627/

Naganuma T, Nakagawa S, Tanigawa A, Sasaki YF, Goshima N, Hirose T. 2012. Alternative 3′-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J.* 31:4020–4034.

Naveed A, Cooper JA, Li R, Hubbard A, Chen J, Liu T, Wilton SD, Fletcher S, Fox AH. 2021. NEAT1 polyA-modulating antisense oligonucleotides reveal opposing functions for both long non-coding RNA isoforms in neuroblastoma. *Cell. Mol. Life Sci.* 78:2213–2230.

Ns U, Ja E, W J. 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* [Internet] 17. Available from: https://pubmed.ncbi.nlm.nih.gov/31800571/

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44:D733-745.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22:1–5.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006. Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLOS Genet.* 2:e168.

Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172.

Raney BJ, Barber GP, Benet-Pagès A, Casper J, Clawson H, Cline MS, Diekhans M, Fischer C, Navarro Gonzalez J, Hickey G, et al. 2024. The UCSC Genome Browser database: 2024 update. *Nucleic Acids Res.* 52:D1082–D1088.

Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D, et al. 2024. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 52:D174–D182.

Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2015. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol. Spectr.* 3:10.1128/microbiolspec.mdna3-0061–2014.

Romero-Barrios N, Legascue MF, Benhamed M, Ariel F, Crespi M. 2018. Splicing regulation by long noncoding RNAs. *Nucleic Acids Res.* 46:2169–2184.

Rot G, Wang Z, Huppertz I, Modic M, Lenče T, Hallegger M, Haberman N, Curk T, von Mering C, Ule J. 2017. High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. *Cell Rep.* 19:1056–1067.

Sasaki YTF, Ideue T, Sano M, Mituyama T, Hirose T. 2009. MENε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc. Natl. Acad. Sci.* 106:2525–2530.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, et al. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database J. Biol. Databases Curation* 2020:baaa062.

Senft AD, Macfarlan TS. 2021. Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.*:1–21.

Simko EAJ, Liu H, Zhang T, Velasquez A, Teli S, Haeusler AR, Wang J. 2020. G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. *Nucleic Acids Res.* 48:7421–7438.

Souquere S, Beauclair G, Harper F, Fox A, Pierron G. 2010. Highly Ordered Spatial Organization of the Structural Long Noncoding NEAT1 RNAs within Paraspeckle Nuclear Bodies. *Mol. Biol. Cell* 21:4020–4027.

Stadler PF. 2010. Evolution of the Long Non-coding RNAs MALAT1 and MENβ/ε. In: Ferreira CE, Miyano S, Stadler PF, editors. Advances in Bioinformatics and Computational Biology. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. p. 1–12.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* 12:2.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. 2009. MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* 19:347–359.

The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, et al. 2023. The Gene Ontology knowledgebase in 2023. *Genetics* 224:iyad031.

Tian F, Zhang Y, Li J, Chu Z, Zhang J, Han H, Jia L. 2023. Effects of the SPI/lncRNA NEAT1 Axis on Functions of Trophoblast and Decidual Cells in Patients with Recurrent Miscarriage. *Crit. Rev. Eukaryot. Gene Expr.* [Internet] 33. Available from: https://www.dl.begellhouse.com/journals/6dbf508d3b17c437,6c9018655d68e94a,3ab570e41315ac65.html

Torres M, Becquet D, Blanchard M-P, Guillen S, Boyer B, Moreno M, Franc J-L, François-Bellan A-M. 2016. Circadian RNA expression elicited by 3'-UTR IRAlu-paraspeckle associated elements.Singer RH, editor. *eLife* 5:e14837.

Torres M, Becquet D, Blanchard M-P, Guillen S, Boyer B, Moreno M, Franc J-L, François-Bellan A-M. 2017. Paraspeckles as rhythmic nuclear mRNA anchorages responsible for circadian gene expression. *Nucleus* 8:249–254.

Van Treeck B, Parker R. 2018. Emerging Roles for Intermolecular RNA-RNA Interactions in RNP Assemblies. *Cell* 174:791–802.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17:261–272.

Vlachogiannis NI, Sachse M, Georgiopoulos G, Zormpas E, Bampatsias D, Delialis D, Bonini F, Galyfos G, Sigala F, Stamatelopoulos K, et al. 2021. Adenosine-to-inosine Alu RNA editing controls the stability of the pro-inflammatory long noncoding RNA NEAT1 in atherosclerotic cardiovascular disease. *J. Mol. Cell. Cardiol.* 160:111–120.

Walter Costa MB, Höner zu Siederdissen C, Dunjić M, Stadler PF, Nowick K. 2019. SSS-test: a novel test for detecting positive selection on RNA secondary structure. *BMC Bioinformatics* 20:151.

Waskom ML. 2021. seaborn: statistical data visualization. *J. Open Source Softw.* 6:3021.

Weghorst F, Torres Marcén M, Faridi G, Lee YCG, Cramer KS. 2024. Deep Conservation and Unexpected Evolutionary History of Neighboring lncRNAs MALAT1 and NEAT1. *J. Mol. Evol.* 92:30–41.

West JA, Mito M, Kurosaka S, Takumi T, Tanegashima C, Chujo T, Yanaka K, Kingston RE, Hirose T, Bond C, et al. 2016. Structural, super-resolution microscopy analysis of paraspeckle nuclear body organization. *J. Cell Biol.* 214:817–830.

Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. 2012. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* 18:900–914.

Wilusz JE, Freier SM, Spector DL. 2008. 3′ End Processing of a Long Nuclear-Retained Noncoding RNA Yields a tRNA-like Cytoplasmic RNA. *Cell* 135:919–932.

Wilusz JE, Whipple JM, Phizicky EM, Sharp PA. 2011. tRNAs Marked with CCACCA Are Targeted for Degradation. *Science* 334:817–821.

Yamada A, Toya H, Tanahashi M, Kurihara M, Mito M, Iwasaki S, Kurosaka S, Takumi T, Fox A, Kawamura Y, et al. 2022. Species-specific formation of paraspeckles in intestinal epithelium revealed by characterization of NEAT1 in naked mole-rat. *RNA* 28:1128–1143.

Zhang B, Mao YS, Diermeier SD, Novikova IV, Nawrocki EP, Jones TA, Lazar Z, Tung C-S, Luo W, Eddy SR, et al. 2017. Identification and Characterization of a Class of MALAT1-like Genomic Loci. *Cell Rep.* 19:1723–1738.

Zhou X, Li X, Yu L, Wang R, Hua D, Shi C, Sun C, Luo W, Rao C, Jiang Z, et al. 2019. The RNA-binding protein SRSF1 is a key cell cycle regulator via stabilizing NEAT1 in glioma. *Int. J. Biochem. Cell Biol.* 113:75–86.

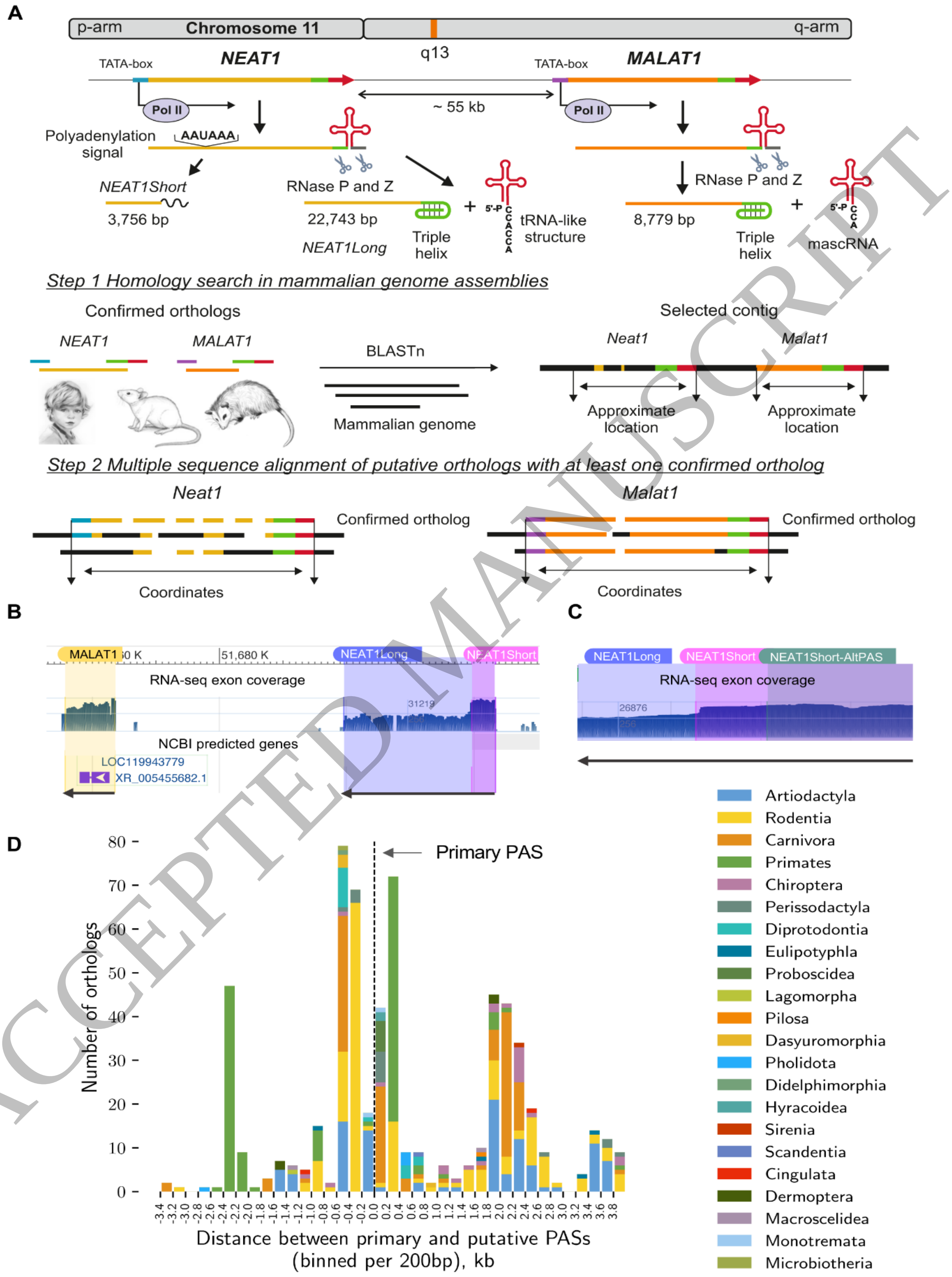

Figure 1
190x275 mm (x DPI)

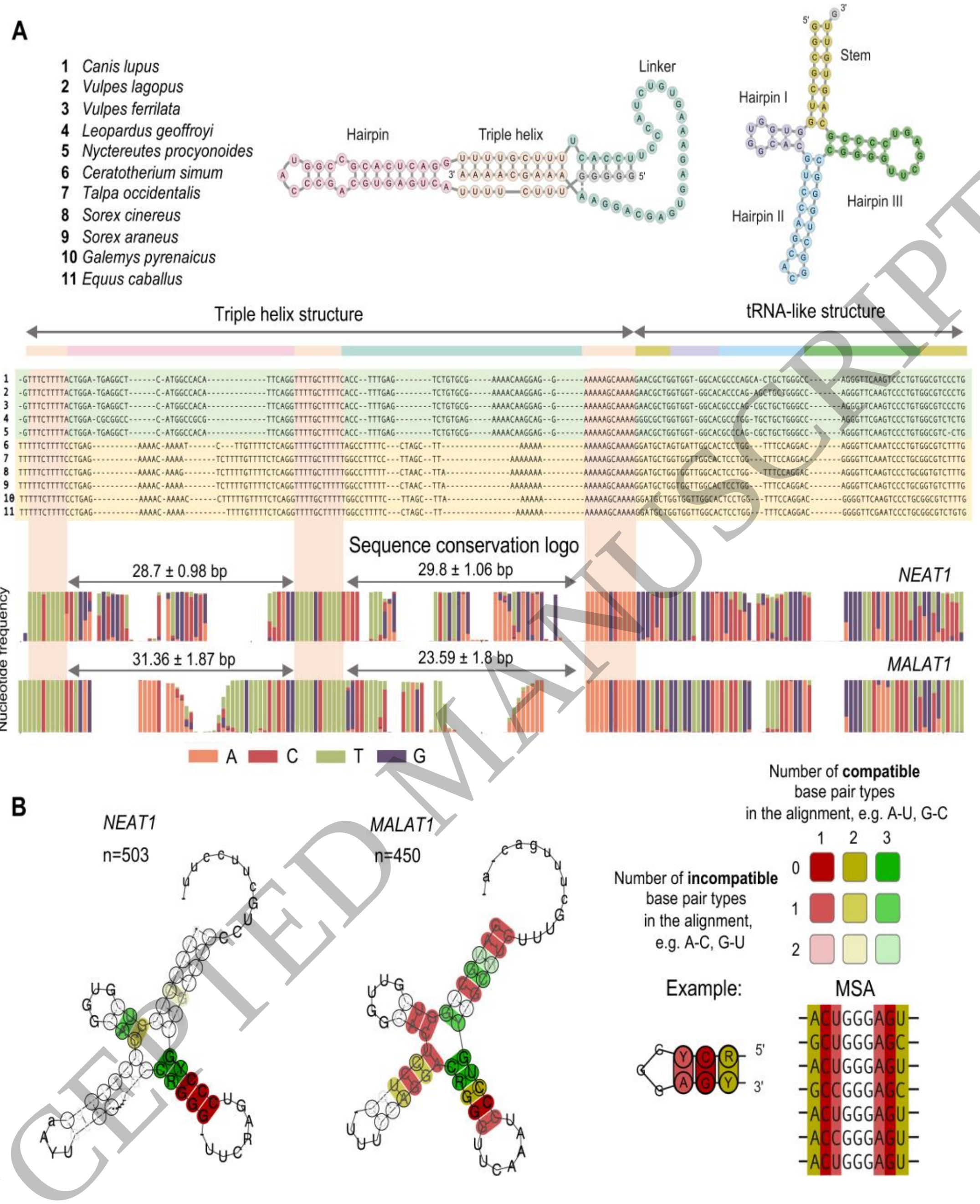

Figure 2
190x187 mm (x DPI)

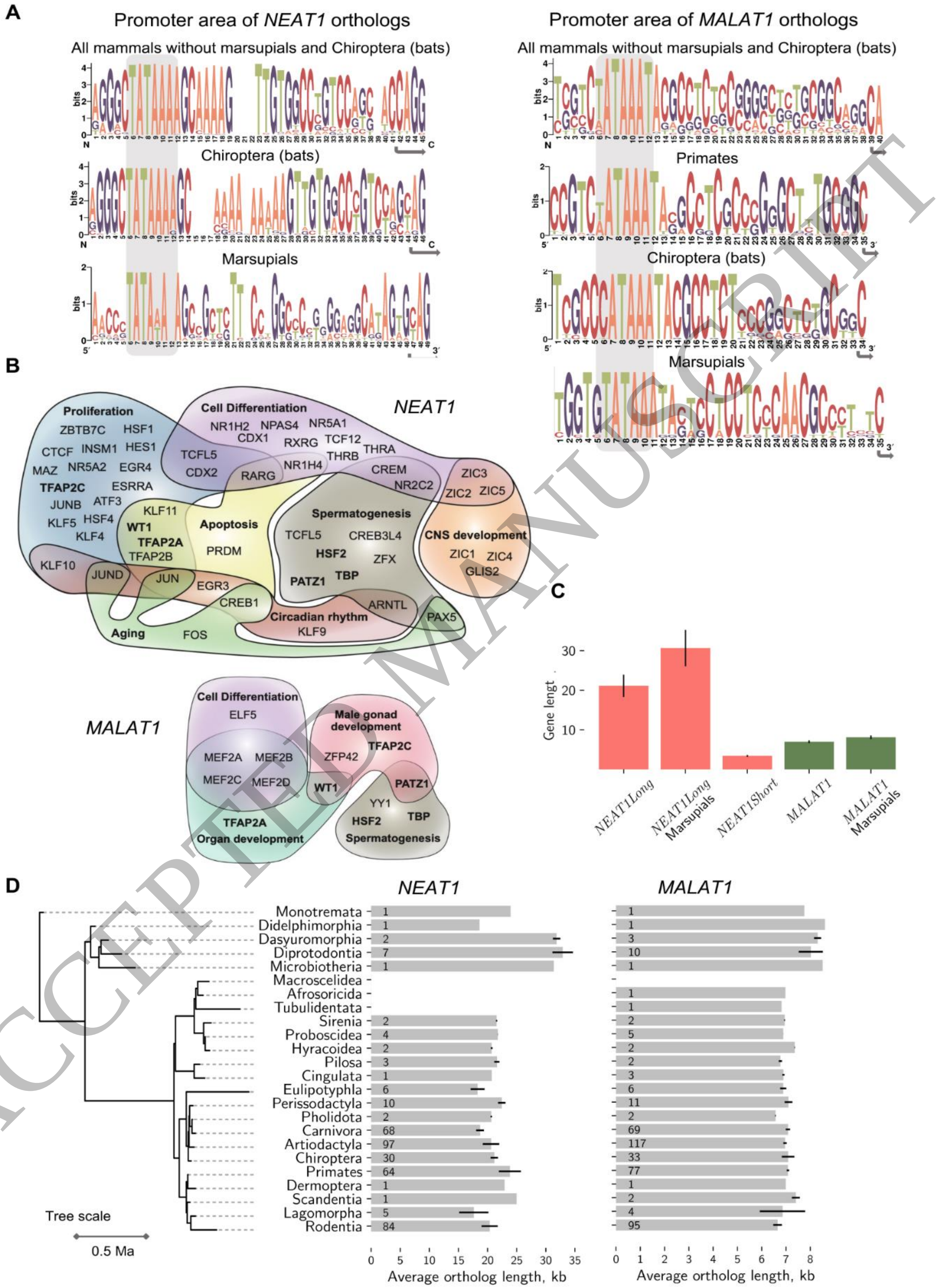

Figure 3

190x275 mm (x DPI)

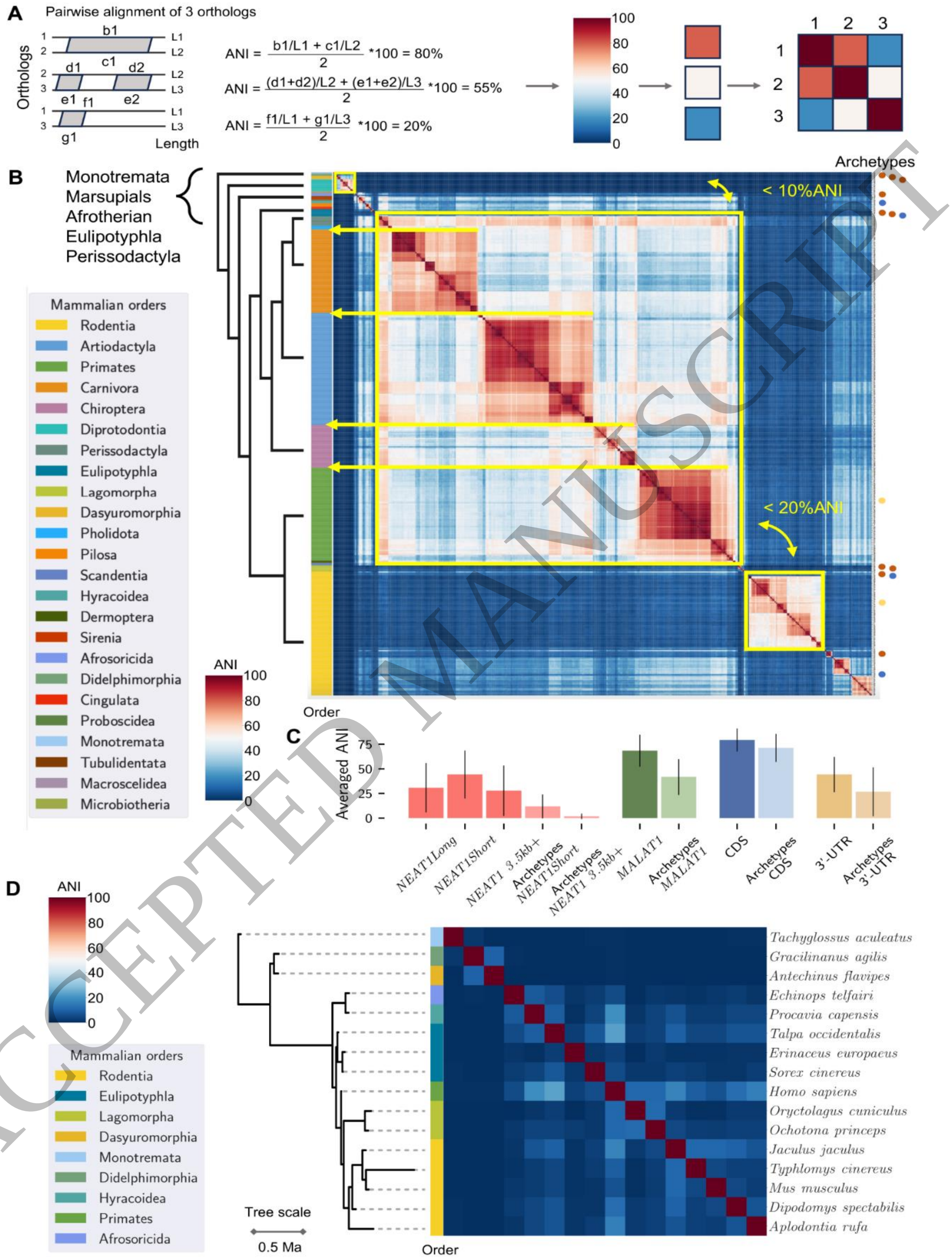

Figure 4

190x275 mm (x DPI)

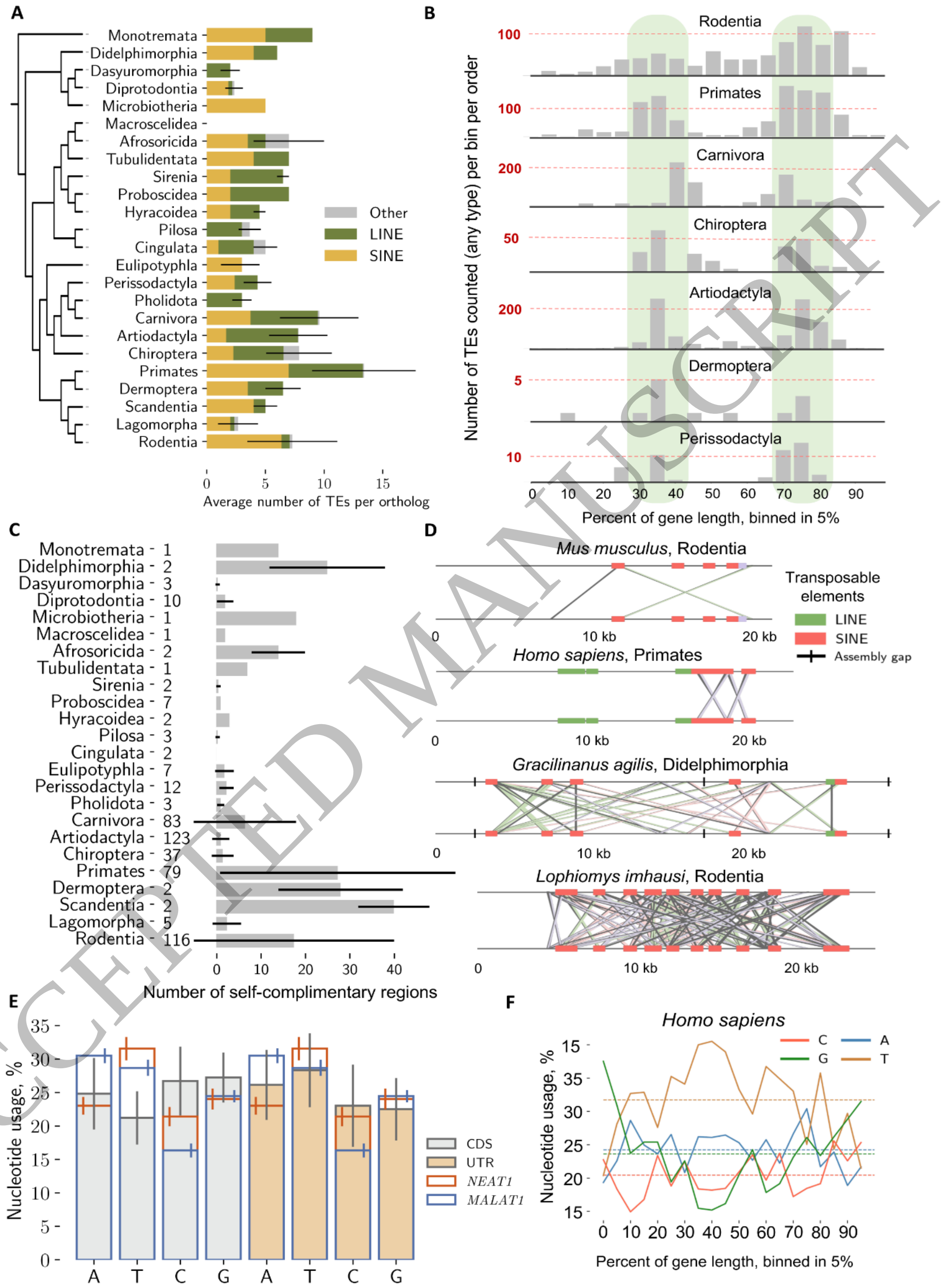

Figure 5
190x275 mm (x DPI)

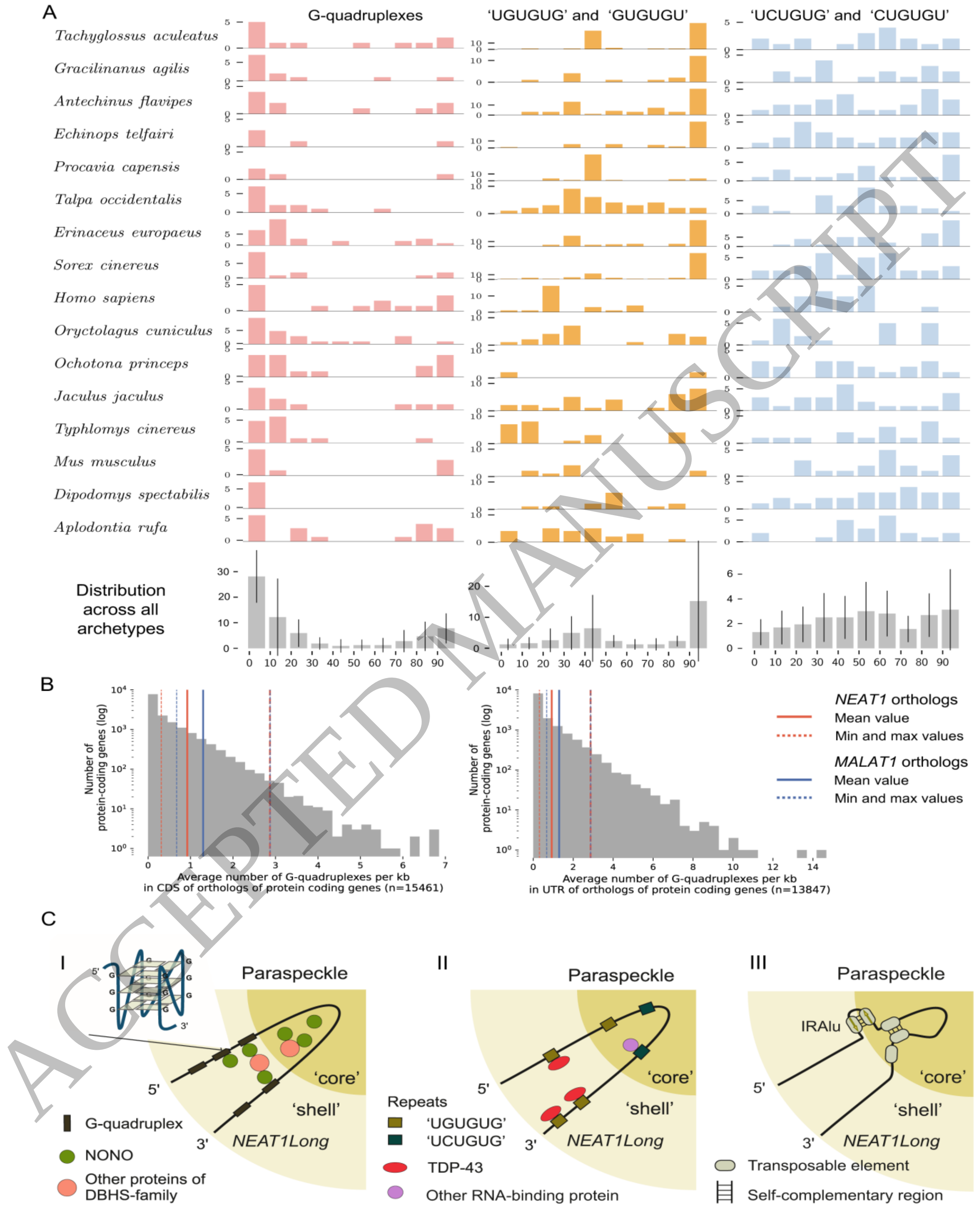

Figure 6

190x275 mm (x DPI)

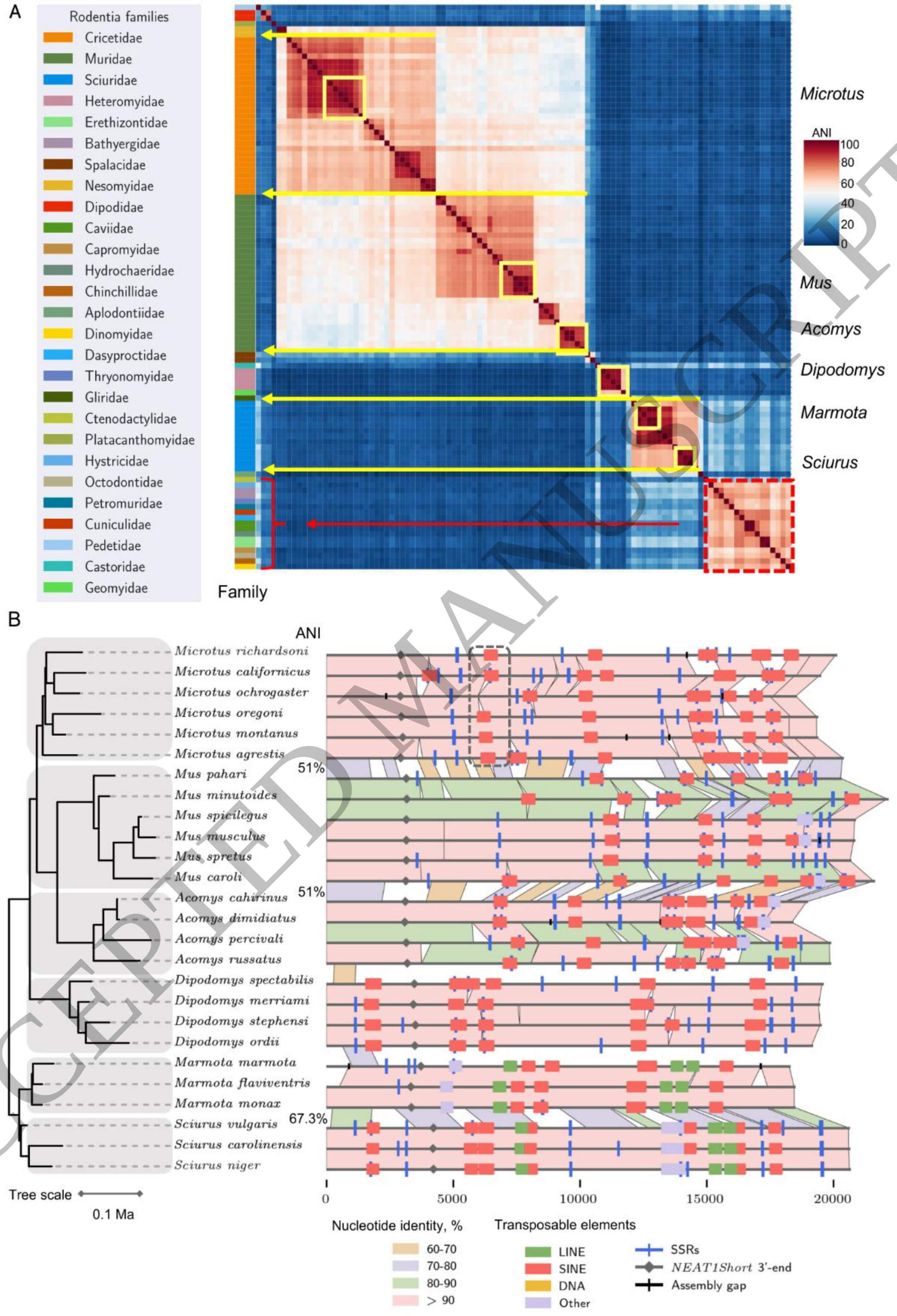

Figure 7
190x275 mm (x DPI)