Generation of *de novo* miRNAs from template switching during DNA replication

Heli A. M. Mönttinen[1], Mikko J. Frilander[1] and Ari Löytynoja[1,*]

[1] Institute of Biotechnology, HiLIFE, University of Helsinki, Finland
*email: ari.loytynoja@helsinki.fi

Abstract

Micro-RNA (miRNA) genes represent one of the most constrained examples of genetic information found in metazoan genomes. All primary transcripts generated from these elements fold into a short RNA stem-loop structure that is further processed to a ∼22 bp dsRNA duplex. Incorporation of one of the strands into the RNA Induced Silencing complex (RISC) can then facilitate negative gene expression regulation. While a substantial number of miRNA genes are ancient and highly conserved, many are subject to rapid evolutionary change. Entirely novel miRNA genes have been shown to emerge in a lineage-specific manner in primates and plants. However, the specific mechanisms that generate these genetic elements have not been determined. Here, we show that up to 100 lineage-specific miRNA genes have emerged in primates via mutations caused by template switching during DNA replication. Template-switching mutations (TSMs) are complex mutations that can introduce in a single replication cycle inverted DNA repeats capable of forming perfect hairpin structures. Our findings demonstrate that TSMs are actively generating complex biological sequence variants which in suitable circumstances create miRNA genes from previously non-functional genomic sequences. This mechanism is orders of magnitude faster than others proposed for the *de novo* creation of genes, enabling near-instant rewiring of genetic information and rapid adaptation to changing environments.

The emergence of novel genes from already pre-existing ones during evolution is a rather common and well-characterised process involving a variety of mechanisms such as gene fusions and fissions, horizontal gene transfers, gene duplication and insertions of mobile DNA elements (Long et al., 2003). In contrast, the *de novo* emergence of new genes has been deemed extremely rare, particularly with protein-coding genes (Weisman, 2022) that need to satisfy the requirement for both transcription and translation processes and post-transcriptional RNA processing events such as pre-mRNA splicing and polyadenylation. Non-coding RNA (ncRNA) genes have no translational requirement and are thus more relaxed to emerge and evolve. Consistent with this, deep sequencing analyses have documented a large number of lineage-specific non-coding RNA transcripts, in particular the long noncoding RNAs (lncRNAs) subclass. While a subset of these has been experimentally shown to have a biological function, the bulk of lncRNAs are poorly characterised and subject to an ongoing debate as to whether they represent functional genes or rather the result of pervasive transcriptional activity without any obvious biological role (Nojima & Proudfoot, 2022)

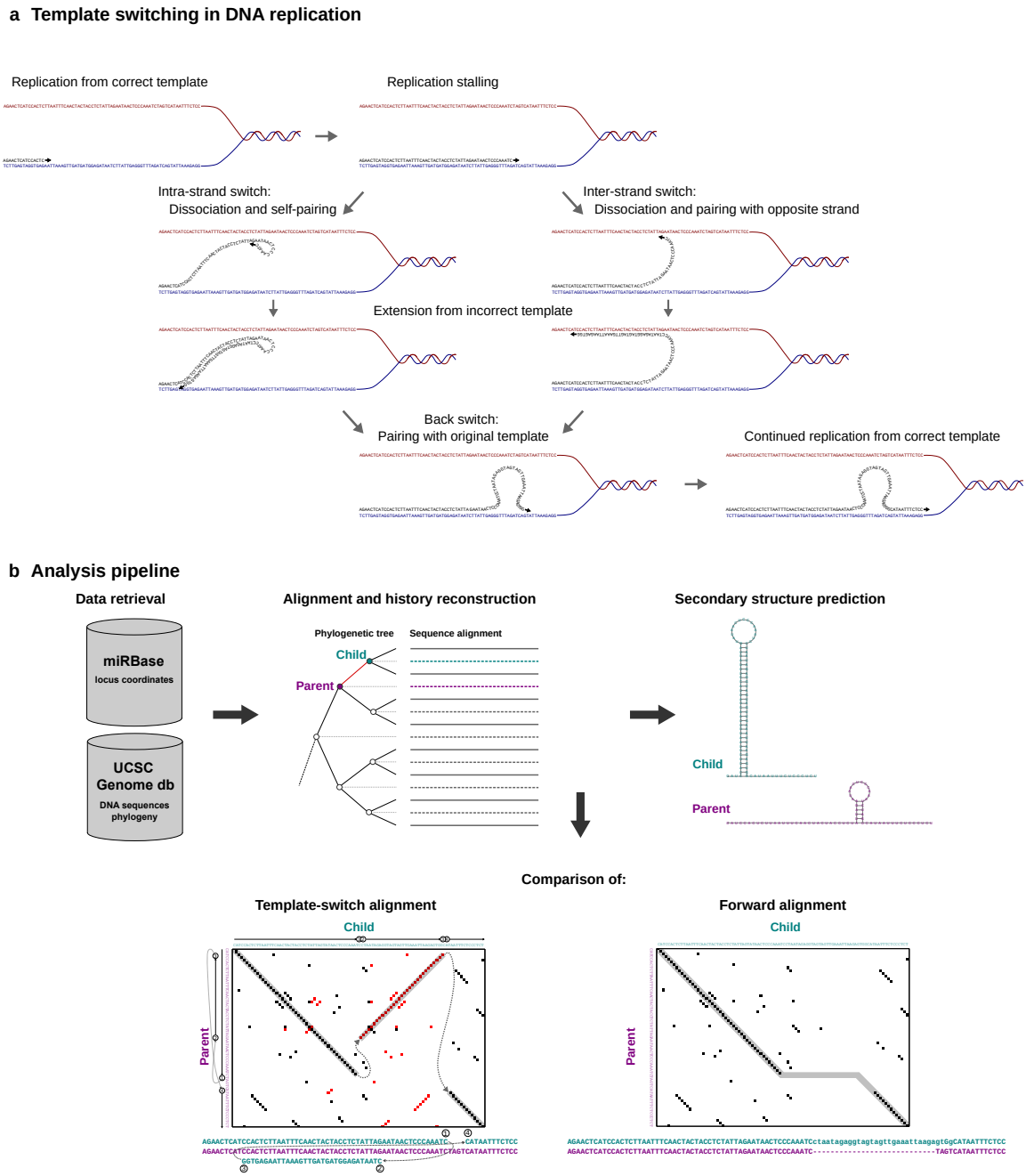

Fig. 1 The mechanism for template-switching mutations and workflow for their analysis. **a** TSMs are proposed to result from DNA replication machinery bypassing DNA lesions with a jump to the other DNA strand. After stalling of the replication fork, the nascent strand may either fold over and undergo intramolecular pairing (Intra-strand switch) or pair with the opposite strand (Inter-strand switch); the nascent strand is then briefly extended from the incorrect template before it realigns with the original template for the replication to continue normally. In addition to a complex sequence change (shown as a loop), this may create a reverse-complement repeat with the potential to form a perfect hairpin. **b** For each miRNA in miRBase, the surrounding genome sequences were fetched for a set of mammals from the UCSC Genome database, and multiple alignment and sequence history were inferred following a consensus phylogeny. The resulting tree-sequence structure was traversed and each parent-child sequence pair was searched for complex mutations. Mutation loci where the template-switch alignment was superior to the traditional forward alignment were recorded and secondary structures before and after the inferred TSM event were predicted. The example is based on hsa-mir-5690.

Amongst the ncRNAs, genes encoding microRNAs (miRNAs) are unique in gene structure and function. The primary transcript from a miRNA locus folds to an RNA stem-loop structure with a strong preference for 35 bp stem length and 10 nt or longer unstructured apical loop (Fang & Bartel, 2015). This primary transcript undergoes both nuclear and cytoplasmic processing events that result in the final ∼22-nt miRNA gene product loaded on the RNA-induced silencing complex (Treiber et al., 2019; Fang & Bartel, 2015). While the stem-loop structure alone is both necessary and in many cases sufficient for any RNA to enter the canonical miRNA processing pathway, processing can be significantly stimulated by additional sequence elements in the stem or the loop (Fang & Bartel, 2015). Ultimately, the functionality of a miRNA is determined by the potential for base-pairing interactions with an mRNA. In animals, miRNAs associate predominantly with the 3'UTR sequence of an mRNA through partial base-pairing interactions. This provides significant flexibility in mRNA recognition so that a single miRNA can target multiple mRNAs.

The *de novo* formation of a miRNA gene requires a mutational mechanism that can introduce inverted DNA sequences to a transcribed region of the genome. Recent studies on mutation clusters within primate genomes have revealed the emergence of short inverted DNA sequences through a DNA replication-related mutation mechanism (Löytynoja & Goldman, 2017; Walker et al., 2021; Löytynoja, 2022). This mechanism was initially proposed to explain the conversion of quasi-palindromic sequences into perfect palindromes by template strand switching during replication (Ripley, 1982). The outcomes from this event are a loopback hairpin formation in the growing strand (intra-strand switch) or a brief jump to the opposing strand (inter-strand switch) followed by continuation of the DNA replication on the original strand (see Fig. 1; Löytynoja & Goldman, 2017). In both cases the outcome is a one-step formation of a complex replacement mutation, creating either an inverted repeat, an inversion, or a combination of inverted and direct repeats (Fig. S1). Hereafter, we refer to this as a template-switching mutation (TSM). Transcription of an inverted repeat will give rise to an RNA hairpin structure (Fig. 1). In contrast to the inverted repeats generated by various mobile DNA elements (e.g., Alus or LINEs) inserted in opposing orientations, the capability of TSM is not limited to a particular DNA sequence. Rather, TSMs can emerge in any DNA sequence context and the inverted repeats do not show length preferences.

Template switch mutations create and modify miRNA stem structures

We studied the emergence of miRNA genes by reconstructing evolutionary histories for the gene loci across mammals, focusing on the human lineage and the densely sampled primates. Applying the approach described in Mönttinen & Löytynoja (2022), we reconstructed the sequence history of all 1,821 unique human miRNA loci present in miRBase (v. 22.1; Kozomara et al., 2019). Given the phylogenetic tree-sequence alignment structure (Fig. 1b), we systematically traversed the tree and found 1,633 of the miRNA loci to have arisen in the primate lineage. We tested whether TSM could provide a more parsimonious explanation for the complex differences observed between the sequences at a parental and child node rather than traditional point substitutions and insertion-deletions (indels). As ancestral reconstruction loses accuracy with high sequence divergence (Vialle

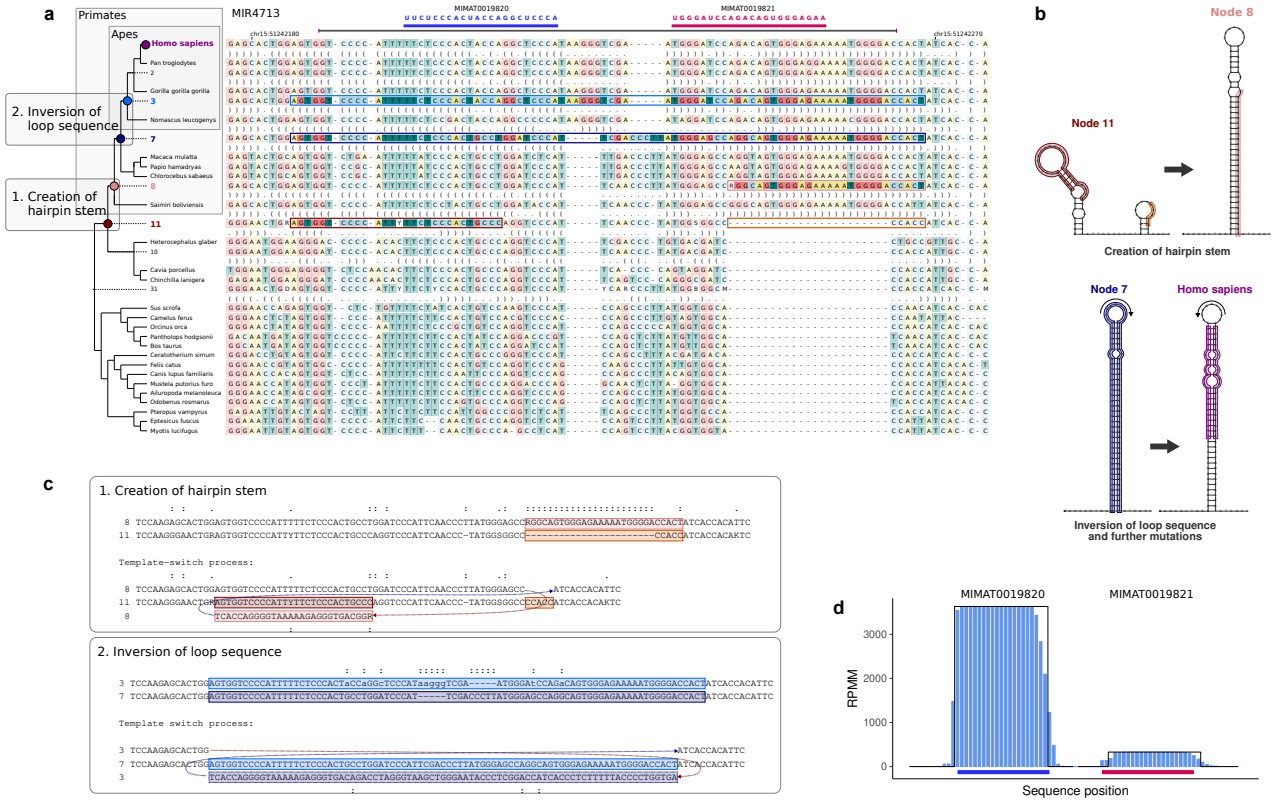

Fig. 2 The emergence and evolution of the hsa-mir-4713 gene. **a** Two TSM events are inferred to have taken place in the locus among the sampled mammals. The first event in the ancestor of primates creates a new inverted repeat with the capacity of forming a hairpin. The second event at the root of apes modifies the existing hairpin by inverting the hairpin loop sequence. **b** Predicted secondary structures for the sequences before and after the inferred TSM events. Mismatches in the predicted stems are caused by inaccuracies in the sequence reconstruction and independent point mutations in-between the TSM event and the next internal node in the tree. **c** The inferred template switch solutions for the two complex mutations. **d** Expression support for the hsa-mir-4713 gene in the miRCarta database. Light blue bars show the read counts (y-axis) for the two annotated transcripts (x-axis; blue and red bars, see panel a). Black lines indicate the ideal stacks with uniform counts (either zero or the maximum count) used for scoring the locus.

et al., 2018), we applied strict criteria for the implied sequence homology and conservation.

Among the cases passing our criteria, we found 79 TSMs that fully and 30 TSMs that partially explain the *de novo* creation of miRNA stem structure at a given locus. Additionally, we found four cases where TSM had modified an existing miRNA hairpin by inverting its loop sequence (Table S1, Table S2, Supplementary Data). Hsa-mir-4713 is an example that combines both types of events: the initial emergence of miRNA stem in the ancestor of primates followed by an inversion of the loop sequence in the ancestor of apes (Fig. 2a,b). The first event is explained by a 28-base-long TSM (Fig. 2c), which may have occurred either intra- or intermolecularly (see Fig. 1a). In contrast, the second event is consistent with an intermolecular TSM of up to 83 bases (Fig. 2c).

Stable miRNA stem structures emerge from unstructured sequences

Although template switching was originally proposed as a mechanism for converting nearly perfect reverse-complement repeats into perfect ones (Ripley, 1982; Dutra & Lovett, 2006; Ladoukakis & Eyre-Walker, 2008; Lavi et al., 2018), it has been shown to take place also at loci with no pre-existing similarity (Löytynoja & Goldman, 2017; Walker et al., 2021; Löytynoja, 2022). To investigate the requirement for pre-existing repeat sequences, we assessed the genomic context and the impact of the inferred TSMs by predicting secondary structures for the loci before and after the inferred events (Fig. 2b). Of the 1,366 primate-specific miRNA loci, we focused on the 79 where template switching fully explains the emergence of the stem structure (Table S1, Table S2) and identified homologous sequences in a set of mammalian genomes. For these, we predicted the RNA secondary structures and determined the corresponding lowest free energy values.

The inferred TSM events are associated with a sharp drop in free energy, indicating a sudden appearance of complex secondary structures (Fig. 3). All of the inferred events occurred within the primate subtree: 57/79 took place in the ancestor of all primates; 11 and 5 in the ancestor of Old World monkeys and apes, respectively; and the last 6 stem structures are only present in a subset of apes (Fig. 3, Table S2). However, this does not imply a primate-specific mechanism but rather reflects the limitations of the underlying data. Ancestral sequence reconstruction requires dense species sampling and inevitably loses accuracy in deeper nodes (Vialle et al., 2018). With the current species set, we could theoretically reach only one branch deeper in the phylogenetic tree, to the ancestor of primates and rodents. Rodent sequences are known to evolve exceptionally fast (Fabre et al., 2012) and difficulties in identifying homologous sequences (cf. Fig. 3, grey cells) likely explain the absence of inferred TSM events in that relatively short branch.

Existing promoters induce transcription of novel stem structures

Our analyses confirm that perfect stem structures can be created by single mutation events and such nascent reverse-complement repeats can further evolve into putative miRNA genes. However, the potential to form a hairpin structure is alone not sufficient for turning a TSM locus into a gene. Transcription of the hairpin sequence and its subsequent processing in the miRNA maturation pathway are both necessary for a TSM-derived sequence to function as a RISC-associated miRNA.

Transcriptional activation of the miRNA genes is associated with their genomic context. Intragenic miRNAs within protein-coding or lncRNA genes are often dependent on the host gene transcriptional activity and are subsequently released from excised introns or mRNA sequences by the Drosha exonuclease (Treiber et al., 2019). A subset of intragenic miRNA genes and most intergenic miRNA genes have their own promoter sequences, but intergenic miRNA may also utilise the bidirectional activity of nearby promoters or readthrough transcription of an upstream protein-coding or lncRNA gene (Rodriguez et al., 2004; Liu et al., 2019; Ha & Kim, 2014). To differentiate between these possibilities we determined the genomic context of the TSM-derived miRNA genes (intergenic vs intragenic) and searched for genomic elements capable of providing transcription for the miRNAs.

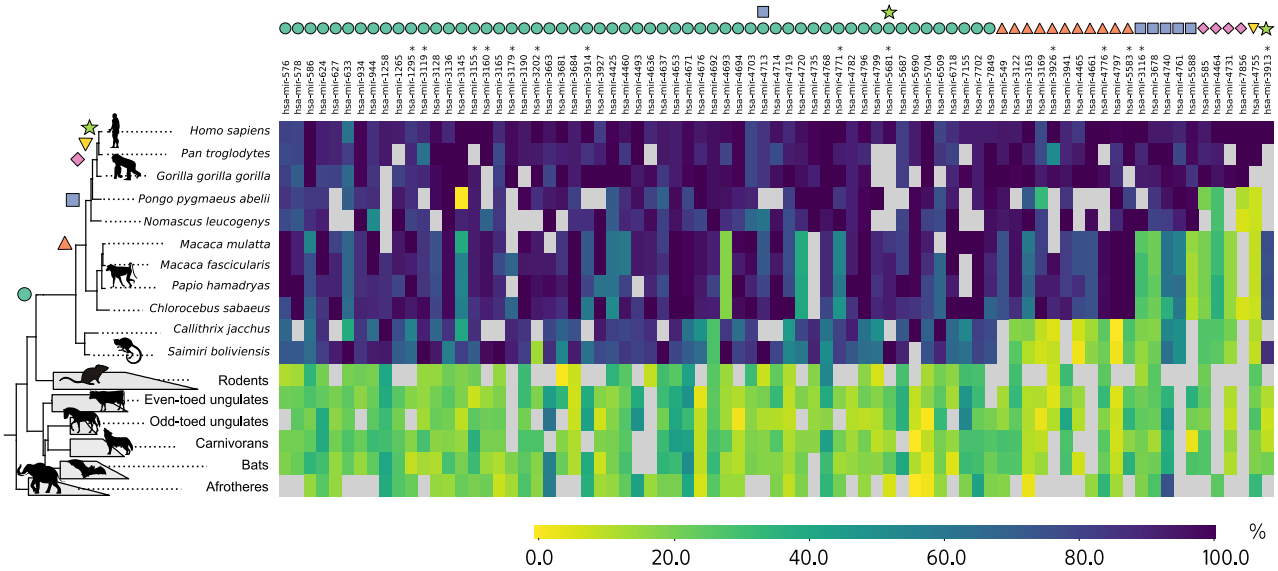

Fig. 3 Normalised free-energy change at human miRNA loci. Matrix cells show the relative stability (free-energy change) of the predicted secondary structures for sequences (species; y-axis) homologous to the human miRNA loci (x-axis, top) in comparison to the sequence with the greatest free-energy change. Darker colours indicate greater changes and thermodynamically more stable secondary structures. Taxa with no homologous sequence are shown in grey. Outside of primates, species within the order are merged (grey trapezoids) and, within the merged group, the species with the greatest free-energy change is shown. Only taxa with data for at least 10 loci are included. Coloured symbols on the top indicate the tree branch where the TSM event was inferred to have taken place, in two cases the mutation hitting the same locus twice. Asterisks after the gene names indicate loci with multiple annotated miRNA transcripts. See Fig. S2 for the full matrix.

We particularly looked for SINE elements that can provide a sense promoter and LTR and LINE1 elements that can provide both sense and antisense promoters for the nearby miRNA gene.

Consistent with earlier observations of introns as hotspots for young miRNAs (Meunier et al., 2013), 89.2% (99/111; Table S2) of our TSM-associated miRNA loci are located in intragenic regions. A great majority of these (82/99) are in the sense strand of a host gene intron and can be expressed as a part of the transcriptional unit with a predominantly neutral effect on the host gene. Six intragenic miRNAs (6/99) partly overlap with exonic regions of host genes, and four intergenic miRNAs (4/12) are within 1 kb of a gene promoter which could support their expression by bidirectional activity (Trinklein et al., 2004; Neil et al., 2009). Finally, ten intronic miRNAs (10/99) in antisense orientation to the host gene and six intergenic miRNAs (6/12) are located within 1 kb of an LTR, SINE, or LINE1 promoter. In total, the genomic context provides a potential expression strategy for 108/111 of the TSM-associated miRNA genes.

The miRCarta (Backes et al., 2018) expression profiles (Fig. 2d) summarise the present RNAseq evidence for miRNA genes annotated in miRBase. Our analysis of the miRCarta profiles confirmed the expression of 110/111 of the TSM-derived miRNA genes. A closer evaluation of the read stacks revealed that at least 95 of the miRNA genes produce clearly defined miRNA-sized stacks (Table S1, Supplementary Data) indicating that their primary transcripts are successfully processed by the miRNA processing pathway. Fang & Bartel (2015) showed that wobble pairs and a mismatched motif in the basal stem region enhance the processing of the pri-miRNAs. Consistent with this,

several sequences inferred for nodes after the TSM event contain mismatches, suggesting base change mutations after the creation of a perfect inverted repeat. Further, a majority of the loci (105/111) have gained additional mutations after the initial TSM event.

A promoter is a limiting factor for *de novo* miRNA genes

Our findings are in line with the suggestions of transposons being involved in the emergence of new miRNA genes (Piriyapongsa et al., 2007; Piriyapongsa & Jordan, 2008) and the reports of young miRNAs being enriched in introns of protein-coding genes (Meunier et al., 2013). This raises the question of whether the enrichment reflects the requirement for a pre-existing promoter or whether it results from a corresponding bias in the distribution of TSM events. To distinguish between these options, we reanalysed the candidate TSM loci found in human polymorphism data (Löytynoja, 2022), shuffling the coordinates of the 3,705 TSM-explained mutation clusters across the complex parts of the genome. We found that the TSM events are strongly depleted in protein-coding genes ($p = 2.85 \times 10^{-6}$), including the introns ($p = 0.012$). Furthermore, the distances between TSM loci and the closest transposable elements do not differ from random (Fig. S3). Together, the results strongly support the first option that the enrichment of novel miRNAs in specific genomic regions reflects the need for transcriptional expression machinery. One should note, though, that our phylogenetic approach for identifying TSM events may have favoured miRNA loci within gene regions due to their greater conservation and that we could not distinguish alternative mutation mechanisms acting on non-complex sequences. Because of the latter, we excluded all potential TSMs that overlap with repeat elements from our polymorphism data, which prevented closer analyses of the transposon loci (see Fig. S4).

To estimate the rate at which novel hairpin structures created by TSM events get converted into miRNA genes, we identified all TSMs located within genic regions in the human evolutionary lineage. We split the gene regions within primate genomes into 1 kbp blocks and scanned for TSM-like mutation patterns. In total, we identified 53,666 putative TSMs of which 31,007 formed a new stem structure at the RNA level and 6,054 were within the size range of a typical miRNA gene (Table S2). Our results indicate that of the human-lineage TSM loci that encode suitable hairpin structures only 0.9% (57/6,054) show sufficient evidence to be annotated as miRNA loci in humans.

Discussion

Our reconstructions of the primate lineage sequence histories demonstrate that the evolutionary emergence of a considerable proportion (Table S2) of the annotated human miRNA genes can be traced with high confidence to individual TSM events during DNA replication. Consistent with earlier estimates (Löytynoja & Goldman, 2017; Walker et al., 2021; Löytynoja, 2022), we identified abundant (>50,000) TSMs within the genic regions of the reconstructed human evolutionary lineage. Of these, more than 6,000 TSMs have the potential to produce pri-miRNA-sized hairpin structures following transcription. This suggests a potent mechanism to produce the evolutionary raw material

for miRNA biogenesis. Our analysis of the annotated human miRNA genes generated by TSMs indicates that the majority (105/111) have gained additional mutations which may enhance their processing or functional interactions with the target mRNAs. Most TSM-associated miRNAs show consistent transcriptional evidence in annotation databases (e.g., miRBase and miRCarta) indicating a successful passage through the miRNA processing pathway (Fig. 2d, Supplementary Data). However, we do note that the majority of these lack detailed functional annotations (see Table S1), such as validated high-confidence target mRNAs or related phenotypic evidence to fulfil the most stringent definitions of miRNA genes (Fromm et al., 2015). Therefore, following the analogy with protein-coding protogenes described in yeast (Carvunis et al., 2012), it is possible that a subset of the annotated TSM-associated miRNA genes are proto-miRNAs and still evolving towards functionality. These young miRNAs may have acquired expression and processing characteristics but not necessarily such regulatory targets that provide a positive selection advantage as proposed by Svoboda & Di Cara (2006). On the other end of the spectrum are TSM-associated miRNAs with clearly documented functional significance. A noteworthy example is hsa-mir-576 (Table S1) which has validated mRNA targets and is linked to the antiviral response threshold in primates (Yarbrough et al., 2014).

Although earlier studies have documented fast evolutionary flux for miRNA genes in several species, the mechanistic basis of *de novo* emergence of the miRNA loci has remained elusive (Sun et al., 2012; Lu et al., 2008; Berezikov et al., 2010; Meunier et al., 2013). Our findings provide a robust model for the emergence of novel miRNA loci and their precursors. Unlike the gradual mutation process by individual single nucleotide changes (Eigen, 1971), TSMs can generate perfect base pairing for an entire RNA hairpin in a single event. In the right context, such as in introns that are known to be enriched for evolutionarily young miRNA loci (Meunier et al., 2013), they would also instantly gain the transcriptional requirement. Altogether, this makes the TSM-driven process of miRNA genesis orders of magnitude faster than the *de novo* evolution of functional proteins, potentially enabling rapid rewiring of existing genetic information and quick adaptation to changing environments. Our analyses focused on past events, but many TSM variants capable of becoming miRNA genes segregate among the human populations (Fig. S5) indicating that the TSM process is active and shaping our genomes right now.

Methods

Data collection and preprocessing

The genetic coordinates of the human miRNA were retrieved from the mirBase database version 22.1 (Kozomara et al., 2019) and all unique loci were retained. The genome assemblies of 62 mammals were downloaded from the UCSC database (https://hgdownload.soe.ucsc.edu/goldenPath/archive) and Entrez (see Table S3), and their alignments were downloaded from the UCSC database (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/maf/; Last modified 6th of May 2015). As maf-alignments consist of discontinuous homology blocks, the alignments were only used to transfer the coordinates of the human miRNA loci to other species. Of other species, we discarded ones whose homologous sequence (implied by the maf-alignment) was less than 30 bp long, showed

less than 80% identity to any other sequence, or had incongruent orientation around the start and end coordinates. For the remaining species, the sequence homologous to the human miRNA locus with an additional 50 bp of flanking sequences was extracted from the species' genome sequence. Three miRNAs (hsa-mir-605, hsa-mir-3688, hsa-mir-7856) are located next to a long human-specific insertion, and long enough sequences were extracted for these to ensure homology beyond the insertion gap.

The human gene coordinates were collected from the Ensemble (v. 105; Yates et al., 2020) genome annotation for the human reference genome GRCh38 (http://ftp.ensembl.org/pub/release-105/gff3/homo_sapiens/Homo_sapiens.GRCh38.105.gff3.gz). The overlapping genes were merged and the remaining gene regions were split into 1000 bp long chunks with 25 bp overlap for neighbouring chunks. Terminal chunks shorter than 1000 bp were merged with a previous chunk. The homologous sequences were extracted similarly to miRNAs loci with the exception that the species were limited to primates, dog, cat, pig, cow and sheep, and the minimum length of sequence implied by the maf-alignment was 100 bp.

Ancestral sequence reconstruction and inference of TSMs

The reference phylogenetic tree for the study species was downloaded from the UCSC database (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/hg38.100way.nh). The sequence alignments were done in two steps. First, the sequences for a miRNA locus were aligned with PAGAN2 (Löytynoja et al., 2012) according to the reference phylogeny. Based on this alignment, sequences were left out if 1) they were more than 1.3 times longer than the human sequence, 2) they contained ten or more Ns in a row, or 3) their sequence identity to human was <50% or <60% when using, respectively, the human and the target sequence as the reference. For the remaining sequences, the alignment and ancestral sequence inference were performed with PAGAN2 according to the same guide tree. Phylogenetic trees were traversed with a Python script using the ete3 library package v.3.1.1 (Huerta-Cepas et al., 2016). Sequences at each non-root parental node and child node were compared using the FPA tool (Löytynoja & Goldman, 2017). All results showing better inverted than forward alignment and >80% sequence identity were recorded. The secondary structures for sequences at the query node as well as its parental and children nodes were predicted using RNAfold v.2.4.14 (Lorenz et al., 2011) with default settings.

The identification of TSMs within gene regions was performed similarly except for quality thresholds. After the first alignment the sequence was discarded if 1) it contained more than 50 N's in a row, 2) either end of the sequence had a gap longer than 30% of the length of the human sequence, or 3) the sequence identity to human was <30% when using the target sequence as the reference.

Quality control for identified TSMs

As a quality control for the identified TSM cases, we compared the aligned sequences at the mutated node (child), its ancestral form (parent) and at their sister and parent nodes; we separately assessed the reliability of the ancestral sequence reconstruction at the TSM source region (where the sequence

is copied from) and at the TSM target region (where the sequence is copied to). For miRNA loci, the TSM source sites were compared between 1) the mutated node and its parent; 2) its parent and sister; 3) its parent and grant-parent; 4) its grant-parent and parent's sister; and 5) the mutated node and its children (one had to pass). The TSM target sites were compared between 1) the mutated node's parent and sister; 2) its parent and grant-parent and 3) its grant-parent and parent's sister. Minimum sequence identity of 80% was required between the mutated node and its parent/children, and between its parent and sister; and identity of 70% was required between its parent and grant-parent, and between its parent and sister of its parent (see Fig. S6). Uncertain bases were marked with IUPAC characters and, when possible, resolved using parsimony. To confirm the inheritance of the TSM target sequence, the TSM target sequence was compared between the mutated node and its child nodes, and sequence identity of 80% was required. Finally, the TSM locus had to cover at least 95% of the human miRNA locus.

The quality control for the inferred TSMs in the gene region data was similar with the exception that the root node was allowed to be the parent for the mutation and the comparisons of the grant-parent node were left out (Fig. S6). As repeats may sometimes look like a TSM just by accident, TSMs overlapping more than 50% of their length with simple repeat elements were excluded. Simple repeat regions were identified with dustmasker v.1.0 (Cock et al., 2015). The criteria for classifying a TSM as a new miRNA-like sequence were: 1) no equally-long hairpin was present at the same position in the parent sequence; 2) the length of the new hairpin stem was 30-55 bases; and 3) the loop of the new hairpin stem was located within the central third of the region defined by the TSM start and end sites.

Genomic enrichment of TSM events in humans

The coordinates of TSM loci segregating in human variation data with the complex genomic regions (not masked as repeats or simple sequences) were obtained from a previous study (Löytynoja, 2022). Using the Ensembl (v.105; Yates et al., 2020) genome annotation (gff3) for human, the coordinates of coding genes, non-coding RNA genes, pseudogenes, coding exons, non-coding exons, and UTR3/UTR5 regions were extracted and, within each subclass, overlapping regions were merged using bedtools v.2.26.0 (Quinlan & Hall, 2010). The genomic context of the inferred TSM loci was analysed using a custom script. The enrichment of the events was analysed using the Python package statistics v.3.11 by comparing the counts of inferred events in unmasked genomic regions to those of similar intervals shuffled randomly across the same genomic regions.

The coordinates of transposable elements were downloaded from the UCSC database (https://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigZips/hg38.fa.out.gz; Last modified 15th of January 2014). The distance between the TSM target site and the closest transposable element was computed using bedtools. Boxplots were drawn from the results using the R package ggplot2 (Wickham, 2016). The enrichment of TSMs in specific genomic regions was tested by comparing the original coordinates to random sets of coordinates. Shuffled sequence positions were drawn from unmasked sequence regions using the bedtools. In total, 100 replicates were generated.

Quality of microRNA expression patterns

To assess the expression of TSM-associated miRNAs, we downloaded their expression profiles, representing the number of RNAseq reads mapped on each position with zero or with one mismatch, from the miRCarta database v.1.1 (https://mircarta.cs.uni-saarland.de; 12th of December 2022; Backes et al., 2018). The expression pattern was visualized using the R package ggplot2 and its deviation from an optimal stack was computed with the function:

$$d = \sum \min(e_i, |m - e_i|) / \sum e_i$$

where e_i is the expression at position i and $m = max(e_i)$. We used the threshold $d < 0.2$ to classify the primary transcripts to be successfully processed.

Visualization of pri-microRNA hairpin free-energies

We considered the emergence of the stem structure to be credibly explained by template switching if the solution left less than nine bases unexplained. For these cases, sequences homologous to the human pri-miRNA loci were extracted from other species and their secondary structures were predicted using RNAfold v.2.4.13 (Lorenz et al., 2011) with default settings. As the free-energy values vary across loci, they were normalised for visualisation by dividing the inferred values by that of the greatest free energy. For the compact matrix, taxa within rodents, odd-toed ungulates, even-toed ungulates, carnivores, and afrotheres were merged and the weakest normalised value was retained. Taxa with data for fewer than ten microRNA loci were discarded. The heatmaps were generated using Python packages matplotlib v.3.5.1 and seaborn v.0.11.2.

Identification of primate-specific miRNA loci and post-TSM mutations

The mammalian sequences homologous to the human miRNA loci were compared pairwise to the human sequence, and their identity was calculated by dividing the count of identical bases with the length of the pairwise alignment. If sequence identity higher than 90% was not found outside primates, the miRNA locus was considered primate-specific. Post-TSM changes were similarly identified by pairwise comparison of the human sequence and the inferred sequence at the mutation node.

Data and code availability

The Python and R scripts for data processing, analyses and visualization will be available on GitHub.

Acknowledgements

This work was enabled by the Academy of Finland grant #322681 to AL and #341477 to MJF. We thank Brendan Battersby and Gunter Meister for the comments on an earlier version of the manuscript and CSC – IT Center for Science for the computational resources.

Supplementary Data

Supplementary Tables, Figures and Data will be available online.

References

Backes, C., Fehlmann, T., Kern, F., Kehl, T., Lenhof, H.-P., Meese, E., & Keller, A. (2018). miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, 46(D1), D160–D167.

Berezikov, E., Liu, N., Flynt, A. S., Hodges, E., Rooks, M., Hannon, G. J., & Lai, E. C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nat. Genet.*, 42(1), 6–9.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, 487(7407), 370–374.

Cock, P. J. A., Chilton, J. M., Grüning, B., Johnson, J. E., & Soranzo, N. (2015). NCBI BLAST+ integrated into Galaxy. *Gigascience*, 4, s13742–015–0080–7.

Dutra, B. E. & Lovett, S. T. (2006). Cis and trans-acting effects on a mutational hotspot involving a replication template switch. *J. Mol. Biol.*, 356(2), 300–311.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10), 465–523.

Fabre, P.-H., Hautier, L., Dimitrov, D., & Douzery, E. J. P. (2012). A glimpse on the pattern of rodent diversification: a phylogenetic approach. *BMC Evol. Biol.*, 12, 88.

Fang, W. & Bartel, D. P. (2015). The menu of features that define primary MicroRNAs and enable de novo design of MicroRNA genes. *Mol. Cell*, 60(1), 131–145.

Fromm, B., Billipp, T., Peck, L. E., Johansen, M., Tarver, J. E., King, B. L., Newcomb, J. M., Sempere, L. F., Flatmark, K., Hovig, E., & Peterson, K. J. (2015). A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, 49, 213–242.

Ha, M. & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, 15(8), 509–524.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, 33(6), 1635–1638.

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, 47(D1), D155–D162.

Ladoukakis, E. D. & Eyre-Walker, A. (2008). The excess of small inverted repeats in prokaryotes. *J. Mol. Evol.*, 67, 291–300.

Lavi, B., Levy Karin, E., Pupko, T., & Hazkani-Covo, E. (2018). The prevalence and evolutionary conservation of inverted repeats in proteobacteria. *Genome Biol. Evol.*, 10(3), 918–927.

Liu, B., Shyr, Y., Cai, J., & Liu, Q. (2019). Interplay between miRNAs and host genes and their role in cancer. *Brief. Funct. Genomics*, 18(4), 255–266.

Long, M., Betrán, E., Thornton, K., & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.*, 4(11), 865–875.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms Mol. Biol.*, 6, 26.

Löytynoja, A. (2022). Thousands of human mutation clusters are explained by short-range template switching. *Genome Res.*, 32(8), 1437–1447.

Löytynoja, A. & Goldman, N. (2017). Short template switch events explain mutation clusters in the human genome. *Genome Res.*, 27(6), 1039–1049.

Löytynoja, A., Vilella, A. J., & Goldman, N. (2012). Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, 28(13), 1684–1691.

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R. W., Wang, S. M., & Wu, C.-I. (2008). The birth and death of microRNA genes in Drosophila. *Nat. Genet.*, 40(3), 351–355.

Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., Hu, H., Khaitovich, P., & Kaessmann, H. (2013). Birth and expression evolution of mammalian microRNA genes. *Genome Res.*, 23(1), 34–45.

Mönttinen, H. A. M. & Löytynoja, A. (2022). Template switching in DNA replication can create and maintain RNA hairpins. *Proc. Natl. Acad. Sci. U. S. A.*, 119(4), e2107005119.

Neil, H., Malabat, C., d'Aubenton Carafa, Y., Xu, Z., Steinmetz, L. M., & Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457(7232), 1038–1042.

Nojima, T. & Proudfoot, N. J. (2022). Mechanisms of lncRNA biogenesis as revealed by nascent transcriptomics. *Nat. Rev. Mol. Cell Biol.*, 23(6), 389–406.

Piriyapongsa, J. & Jordan, I. K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA*, 14(5), 814–821.

Piriyapongsa, J., Mariño-Ramírez, L., & Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics*, 176(2), 1323–1337.

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842.

Ripley, L. S. (1982). Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc. Natl. Acad. Sci. U. S. A.*, 79, 4128–4132.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, 14(10A), 1902–1910.

Sun, J., Zhou, M., Mao, Z., & Li, C. (2012). Characterization and evolution of microRNA genes derived from repetitive elements and duplication events in plants. *PLoS One*, 7(4), e34092.

Svoboda, P. & Di Cara, A. (2006). Hairpin RNA: a secondary structure of primary importance. *Cell. Mol. Life Sci.*, 63(7-8), 901–908.

Treiber, T., Treiber, N., & Meister, G. (2019). Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nat. Rev. Mol. Cell Biol.*, 20(1), 5–20.

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. *Genome Res.*, 14(1), 62–66.

Vialle, R. A., Tamuri, A. U., & Goldman, N. (2018). Alignment modulates ancestral sequence reconstruction accuracy. *Mol. Biol. Evol.*, 35(7), 1783–1797.

Walker, C. R., Scally, A., De Maio, N., & Goldman, N. (2021). Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLoS Genet.*, 17(3), e1009221.

Weisman, C. M. (2022). The origins and functions of de novo genes: Against all odds? *J. Mol. Evol.*, 90(3-4), 244–257.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Use R! Springer International Publishing.

Yarbrough, M. L., Zhang, K., Sakthivel, R., Forst, C. V., Posner, B. A., Barber, G. N., White, M. A., & Fontoura, B. M. A. (2014). Primate-specific mir-576-3p sets host defense signalling threshold. *Nat. Commun.*, 5, 4963.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B.,

Nuhn, M., Oheh, D. N., Parker, A., Parton, A., Patricio, M., Sakthivel, M. P., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S. E., IIsley, G., Kostadima, M., Langridge, N., Loveland, J. E., Martin, F. J., Morales, J., Mudge, J. M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S. J., Cunningham, F., Howe, K. L., Zerbino, D. R., & Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Res.*, 48(D1), D682–D688.