Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Position-resolved k-mer analysis of mature microRNAs reveal widespread CG and UA depletion across taxa

Chandhru Srinivasan
Bharathidasan University
Mano Chitra Karthikeyan
Bharathidasan University
Abirami Jayaprakash
Bharathidasan University
Indhirakumar Murugesan
Bharathidasan University
Vignesh Packiyaraj
Bharathidasan University
Antony Joseph Velanganni Arockiam
velanganni@bdu.ac.in

Bharathidasan University

Research Article

Keywords: microRNA, k-mer, dinucleotide, trinucleotide, positional bias, CG depletion, UA depletion, permutation test, Shannon entropy

Posted Date: September 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10645057/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Mature-miRNA function depends on nucleotide position, particularly the 5' nucleotide and seed, but adjacent-base usage across the complete sequence remains unresolved at full scale. We analysed 48,885 records from 271 organism entries with 30,721 distinct sequences. Overlapping dinucleotides and trinucleotides were counted at 0-based starts and normalized by the number of eligible sequences. Taxon analyses included 185 organism entries in ten mutually exclusive groups, with duplicates removed within each organism. Significance was assessed using 10,000 permutations under two nulls: nucleotide shuffling within each miRNA and position-wise permutation among sequences of the same length.

CG was the least frequent dinucleotide (3.335%), followed by UA (4.897%), whereas UG was the most frequent (9.165%). Relative to the position-wise, length-stratified null, CG and UA were depleted by 2.466% and 1.852%, and UG was enriched by 1.667% (Benjamini–Hochberg $q = 2.13 \times 10^{-4}$ for each). CG and UA were depleted under both nulls in all ten taxonomic groups. Every CG- or UA-containing trinucleotide was also depleted under both nulls. UA nevertheless had a raw-frequency peak at start (9.834%). At 0th position, UA was enriched relative to within-sequence shuffling but depleted relative to the position-wise null. Many k-mer profiles were also less uniform across positions than predicted by the nulls.

Mature miRNAs thus have a widespread, lineage-associated positional organization. Elevated CpG mutability is compatible with CG depletion and UG enrichment, but the present analysis does not test methylation, mutation or causation. The biological basis of UA depletion remains unresolved.

Introduction

MicroRNAs (miRNAs) are short regulatory RNAs that guide Argonaute proteins to complementary sites in target transcripts. In animals, target recognition is strongly influenced by the seed, usually defined as nucleotides 2–8 of the mature miRNA, while additional pairing outside the seed can modify site efficacy and specificity [1,2]. The 5' nucleotide also affects the accumulation and sorting of small RNAs. A 5' uridine can promote loading of animal miRNAs into Argonaute, and plant Argonaute proteins show distinct preferences for the terminal nucleotide of their associated small RNAs [3,4]. These findings establish that nucleotide position matters, but they do not describe how adjacent bases are distributed across the mature-miRNA sequences.

Short sequence motifs have been studied at several stages of miRNA biology. Primary-sequence elements in precursor hairpins contribute to Microprocessor recognition, and position-aware motif methods can recover enriched and depleted k-mers near miRNA cleavage sites [5–7]. K-mer composition has also been used as a feature for precursor-miRNA classification and species categorization [8]. Those applications were designed to identify processing signals or discriminate sequence classes. They did not ask whether every mature-sequence position has a characteristic dinucleotide or trinucleotide repertoire.

Some previous studies addressed CG in more specific settings. Halytskiy and Komisarenko (2011) mapped CG and CNG sites in 1,083 mature miRNAs from human, mouse and rat and reported local peaks in their positional distribution [9]. Their comparison was focused on CG/CNG localization and average genomic frequency. Dai et al. (2022) studied human miRNAs carrying CG in the seed and related this feature to non-canonical bulge targets [10]. These studies show that CG can be biologically informative, but neither compared CG and UA with the complete k-mer analysis under null models that preserve relevant components of mature-miRNA composition. The difference is important because a motif may form local peaks while remaining scarce overall, and a raw positional peak can arise from the mononucleotide composition at that position.

Here, we constructed a position-resolved atlas of all overlapping dinucleotides and trinucleotides in miRBase release 22.1 [11]. The analysis was designed around three questions. First, which k-mers are scarce or abundant in the globally unique mature-miRNA sequences? Second, which positional patterns remain after controlling either the nucleotide composition of each sequence or the position-specific nucleotide composition of sequences with the same length? Third, how similar are the resulting profiles across selected animal lineages and angiosperms? We placed particular emphasis on CG, UA and UG because the observed data identified these dinucleotides as prominent depleted or enriched features. CpG hypermutability was evaluated only as a possible explanation for CG depletion; No underlying mechanism could be proposed for the observed UA depletion.

Statement of significance

(1) Problem or Issue	**Mature-miRNA function depends on nucleotide position, but position-specific adjacent-base usage across mature-miRNA sequences and its consistency among broad taxonomic groups are not well defined.**
(2) What is Already Known	The 5' nucleotide and seed influence Argonaute loading and target recognition. Earlier work examined processing motifs, guide-strand features or selected CG-containing miRNAs, rather than complete positional k-mer profiles under composition-preserving null models.
(3) What this Paper Adds	Analysis of 48,885 miRBase records identifies widespread CG and UA depletion, UG enrichment and positional concentration. Two permutation nulls support CG and UA depletion in ten taxonomic groups and separate the raw 5' UA peak from an independent adjacency preference.
(4) Who would benefit from the new knowledge in this paper	Researchers in microRNA biology, RNA sequence analysis, comparative genomics and small-RNA evolution can use these data to investigate mechanisms underlying positional k-mer bias.

Results

Construction of the mature-miRNA sequence dataset

The miRBase release 22.1 mature FASTA contained 48,885 records assigned to 271 organism prefixes. The free-text headers contained 265 distinct displayed organism names, but organism prefixes were used as the identifiers because they preserve separate database entries. All sequences contained only A, C, G and U. Their lengths ranged from 15 to 34 nucleotides, with a median of 22 nucleotides and an interquartile range of 21–23 nucleotides.

Exact deduplication across the complete database produced 30,721 distinct mature sequences. This globally unique set was used to describe the universal sequence dataset, so that an identical mature sequence deposited for several organisms did not receive repeated weight. Taxon comparisons used a different rule. Exact duplicates were removed within each organism, but an identical sequence present in different organisms was retained once for every organism in which it occurred. The ten selected groups contained 185 organism entries, 43,564 raw mature records and 40,854 within-organism unique sequence occurrences. Across those groups, 26,315 distinct sequence strings were represented (Table 1).

Table 1
Dataset and taxon inventory

Analysis set	Organism entries (n)	Raw records (n)	Within-organism unique (n)	Distinct sequences (n)	Median length (IQR), nt
All miRBase mature records	271	48885	45862	30721	22 (21–23)
Selected taxa combined	185	43564	40854	26315	22 (21–23)
Angiosperms	75	9297	6845	5265	21 (21–22)
Nematodes	11	3121	3088	2884	22 (22–23)
Insects	31	5004	4983	3410	22 (21–23)
Fishes (non-tetrapod vertebrates)	17	3990	3935	2293	22 (22–22)
Amphibians	2	429	427	371	22 (21–23)
Non-avian reptiles	5	1877	1857	1315	22 (21–22)
Aves	4	1997	1991	1723	22 (21–23)
Other mammals	20	11250	11190	7391	22 (21–23)
Non-human primates	19	3943	3906	1985	22 (21–22)
Human	1	2656	2632	2632	22 (21–22)

Table 1 note. Within-organism unique counts each exact sequence once per organism; distinct strings collapse identical sequences across the indicated set. Organism entries are miRBase prefixes, and the ten taxon groups are mutually exclusive.

The principal positional window comprised k-mer start positions 0–19. At each position, the denominator was the number of sequences long enough to contain the complete dinucleotide or trinucleotide. This position-specific denominator prevented later positions, which are contributed only by longer miRNAs, from being divided by the full sequence count. Figure 1 summarizes the dataset construction, length distribution and observed universal k-mer profiles.

The miRNA dataset has a nonuniform dinucleotide landscape

When counts were aggregated across start positions 0–19, the 16 dinucleotides occupied a broad frequency range (Table 2). CG was the least frequent dinucleotide at 3.335%, followed by UA at 4.897%. UG was the most frequent at 9.165%. CU, GG and AG each accounted for approximately 7% of eligible windows, while the remaining dinucleotides were distributed between 5.03% and 7.79%.

Table 2
Universal dinucleotide abundance relative to two composition-preserving null models

2-mer	Observed (%)	Effect: within-sequence (pp)	q: within-sequence	Effect: position-wise (pp)	q: position-wise	Interpretation
AA	5.847	-0.105	2.29e-04	+ 0.104	2.13e-04	Null-dependent direction
AC	5.121	+ 0.006	0.769	-0.203	2.13e-04	Depleted only under position-wise null
AG	6.999	+ 0.773	2.29e-04	+ 0.610	2.13e-04	Enriched under both nulls
AU	6.055	-0.586	2.29e-04	-0.510	2.13e-04	Depleted under both nulls
CA	6.546	+ 1.431	2.29e-04	+ 1.310	2.13e-04	Enriched under both nulls
CC	5.033	-0.190	2.29e-04	+ 0.165	2.13e-04	Null-dependent direction
CG	3.335	-2.131	2.29e-04	-2.466	2.13e-04	Depleted under both nulls
CU	7.014	+ 0.807	2.29e-04	+ 0.991	2.13e-04	Enriched under both nulls
GA	6.586	+ 0.359	2.29e-04	+ 0.439	2.13e-04	Enriched under both nulls
GC	5.632	+ 0.167	2.29e-04	-0.089	2.13e-04	Null-dependent direction
GG	7.047	-0.040	0.115	+ 0.190	2.13e-04	Enriched only under position-wise null
GU	6.549	-0.596	2.29e-04	-0.540	2.13e-04	Depleted under both nulls
UA	4.897	-1.744	2.29e-04	-1.852	2.13e-04	Depleted under both nulls
UC	6.381	+ 0.174	2.29e-04	+ 0.126	2.13e-04	Enriched under both nulls
UG	9.165	+ 2.021	2.29e-04	+ 1.667	2.13e-04	Enriched under both nulls
UU	7.792	-0.348	2.29e-04	+ 0.059	0.021	Null-dependent direction

Table 2 note. Effects are observed minus null means (percentage points); q values are Benjamini–Hochberg-adjusted over 16 motifs per null (10,000 permutations).

The heatmap in Fig. 1a shows that these differences were not constant across positions. UA and UG were prominent at start position 0, whereas CG remained low through most of the displayed window. Other motifs showed narrower positional changes that would be obscured by a single overall frequency. The observed heatmaps are descriptive; inference was based on the two composition-preserving permutation models.

Two null models identify CG and UA depletion and UG enrichment

The within-sequence null preserved the length and exact mononucleotide composition of every mature miRNA while removing its nucleotide order. Under this null, CG was depleted by 2.131% points and UA by 1.744%, whereas UG was enriched by 2.021% points ($q = 2.29 \times 10^{-4}$ for each motif). The position-wise, length-stratified null preserved the mononucleotide composition of every position among sequences of the same length and therefore provided the more stringent test of adjacent-base association. Relative to this null, the CG, UA and UG effects were − 2.466, − 1.852 and + 1.667% points, respectively ($q = 2.13 \times 10^{-4}$ for each motif). Thus, their overall directions were not explained by either the base composition of individual miRNAs or the position-specific mononucleotide composition of the repertoire.

Several additional dinucleotides had concordant effects under both nulls. AG, CA, CU, GA, UC and UG were enriched, while AU, CG, GU and UA were depleted. AC was detectably depleted only under the position-wise null, and GG was enriched only under that null. AA, CC, GC and UU changed direction between the null models, showing that the biological interpretation of a frequency difference depends on the constraints preserved by the comparison (Table 2).

The depletion extended to the trinucleotide level. Each of the eight contexts containing CG (ACG, CCG, CGA, CGC, CGG, CGU, GCG and UCG) was depleted under both null models. The eight UA-containing contexts (AUA, CUA, GUA, UAA, UAC, UAG, UAU and UUA) were also depleted under both. Their observed frequencies ranged from 0.667% to 1.065% for CG contexts and from 1.015% to 1.380% for UA contexts. All 16 effects remained significant after false-discovery-rate correction ($q \leq 2.33 \times 10^{-4}$; Table S4).

UA is depleted overall but has a distinct 5' localization

The raw UA frequency at start position 0 was 9.834%, approximately twice its overall frequency across positions 0–19. The two nulls separated two components of this pattern. UA exceeded the within-sequence null at position 0 by 3.200% points ($q = 3.23 \times 10^{-4}$), indicating that UA was preferentially localized to the first dinucleotide compared with random positions within the same sequences. It was nevertheless 1.756 percentage points below the position-wise, length-stratified expectation at the same position ($q = 3.37 \times 10^{-4}$). The observed 5' peak therefore reflects the strong nucleotide composition at the start of mature miRNAs, including the known enrichment of 5' U, but does not show a composition-independent preference for U followed by A.

UG behaved differently at position 0. Its observed frequency was 13.890%, which exceeded the within-sequence and position-wise null means by 6.744 and 1.790 percentage points, respectively. CG remained depleted at the first start position under both nulls. These results distinguish a position-specific UG enrichment from the raw UA peak.

The canonical seed comprises mature-miRNA nucleotides 2–8 (1-based), corresponding to fully contained dinucleotide starts 1–6 in the present indexing. Under the position-wise null, CG, UA and AU were depleted at all six of these starts, whereas AG, CA, CU and UG were enriched at all six. At the trinucleotide level, ACG, CGA, CGC, CGU, GCG and UCG were depleted at all five fully seed-contained starts. ACA, CAC, CAG, CCU, CUG, GCA and UGG were enriched at all five. These are position-level results rather than a separate regional test, but they show that the focal depletion is part of a wider seed-contained sequence pattern.

CG and UA depletion recurs across the selected taxonomic groups

The ten mutually exclusive analytical groups comprised angiosperms, nematodes, insects, non-tetrapod vertebrates (Fishes), amphibians, non-avian reptiles, Aves, other mammals, non-human primates and human. Exact group definitions and organism inventories are given in Table S1. All ten groups showed overall CG and UA depletion under both permutation nulls (Fig. 3a). Under the position-wise, length-stratified null, relative CG depletion ranged from 15.0% in nematodes to 63.4% in human, while relative UA depletion ranged from 8.9% in amphibians to 40.6% in angiosperms. Each comparison remained significant after correction within its dataset and test family.

Recurrence of the overall direction did not imply identical positional profiles. Nematodes had a localized CG increase around start position 4, for example, while vertebrate groups generally showed lower and flatter CG profiles at internal positions. UA position 0 also varied. Relative to within-sequence shuffling, UA was enriched at position 0 in nine groups and showed no detectable deviation in angiosperms. Under the position-wise null, position-0 UA was depleted in Aves, human, non-human primates, other mammals and angiosperms, and showed no detectable deviation in the other five groups. No taxon showed significant position-0 UA enrichment under the position-wise null (Fig. 3d).

An organism-level principal-component analysis summarized the complete observed dinucleotide profiles without using the focal motifs as a separate input. The analysis included 103 organism entries with at least 100 within-organism unique mature sequences. PC1 and PC2 explained 25.4% and 10.7% of the total variance. Vertebrate, invertebrate-animal and angiosperm profiles occupied partly separated regions, with some overlap at their boundaries (Fig. 2d). Human was located within the vertebrate distribution rather than forming an isolated profile. The PCA is descriptive and does not test evolutionary distance or phylogenetic independence.

Positional concentration extends beyond CG and UA

For each motif, positional concentration was defined as one minus the normalized Shannon entropy of its frequency profile across start positions 0–19. A value near zero indicates an approximately even distribution, while a larger value indicates localization to a subset of positions. Under the within-sequence null, all 16 dinucleotides and all 64 trinucleotides in the globally unique repertoire were more positionally concentrated than expected. The position-wise null accounted for more of this structure: eight dinucleotides were more concentrated, four were more uniform and four showed no detectable difference. Among trinucleotides, 39 were more concentrated, two were more uniform and 23 showed no detectable difference.

The focal motifs illustrate why abundance and concentration are separate properties. Universal CG abundance was strongly depleted under the position-wise null, but its concentration did not differ detectably from that null (q = 0.602). CG was therefore scarce across much of the window rather than confined to a narrow set of positions. UA was both depleted overall and more concentrated than expected (concentration effect = 0.00417, $q = 4.00 \times 10^{-4}$), consistent with its 5' localization. UG was enriched overall but slightly more uniform than the position-wise expectation (effect = − 0.000603, $q = 4.00 \times 10^{-4}$).

Taxon-level concentration effects were widespread but varied in extent (Fig. 4). Depending on the group, 31.25–87.5% of dinucleotides and 37.5–93.75% of trinucleotides were significantly more concentrated than the position-wise null. CG was more concentrated in eight of ten groups and UA in nine of ten groups under this null. All 16 CG- or UA-containing trinucleotide contexts had positive median concentration effects across taxa, and each context was significant in seven to ten groups. These summaries show recurrent positional organization without treating the taxonomic groups as independent evolutionary replicates.

Discussion

The overall result of this study is that mature-miRNA sequences have a reproducible position-resolved k-mer composition that is not explained by sequence length, within-sequence nucleotide composition or position-specific mononucleotide composition alone. CG and UA were depleted in overall abundance under both null models in the globally unique miRNA sequences and in each of the ten selected taxonomic groups. UG was enriched in the universal miRNA sequences, and all CG- and UA-containing trinucleotide contexts were depleted. The positional analysis also resolved a feature that would be ambiguous in an overall count: UA was scarce across the full window but formed a raw-frequency peak at the first dinucleotide.

The CG result adds a different comparison to earlier work. Halytskiy and Komisarenko (2011) reported positional CG peaks in a smaller set of mammalian mature miRNAs and compared CG occurrence with average genomic frequency [9]. A local peak can coexist with overall depletion, and a genomic background answers a different question from a null that preserves miRNA nucleotide composition. Dai et al. (2022) examined the target relationships of miRNAs that already contained CG in the seed [10]. That analysis

addressed the possible functional consequences of a CG-bearing seed rather than whether CG was scarce relative to other adjacent-base combinations. The present dataset and null models therefore do not contradict those studies. They show that, across a larger and taxonomically broader miRNA sequences, the observed number of CG adjacencies is below the expectation generated from the bases available to form them.

Elevated mutability of methylated CpG sites is one plausible contributor to this pattern. In vertebrate DNA, methylated CpG sites are mutation hotspots because deamination of 5-methylcytosine produces thymine, and CpG loss has accumulated over evolutionary time [12,13]. If a CpG on the genomic strand corresponding to a mature miRNA underwent a C-to-T substitution, the transcribed sequence would contain UG. The combination of CG depletion and UG enrichment is compatible with this route. We did not locate each mature sequence in a genome, measure methylation, infer ancestral alleles or estimate substitution rates. The same pattern could also be influenced by precursor processing, Argonaute loading, target-mediated constraints, miRNA-family history or other sequence-dependent processes. Locus-level comparative analyses will be required to separate these explanations, especially because DNA methylation systems and sequence contexts differ among lineages.

UA requires a separate interpretation. The overall depletion of UA was consistent under both nulls and recurred in all ten groups. Its position-0 peak did not reverse that conclusion because the two summaries address different parts of the distribution. Recognition of a 5' U during Argonaute loading provides a biological basis for enrichment of U at the first nucleotide [3,4]. Once the observed position-specific mononucleotide composition was preserved, however, UA at position 0 was depleted in the universal dataset and was not enriched in any taxon. The data therefore support 5' localization of UA, but not a separate preference for A immediately after the first U. We do not have a supported explanation for UA depletion at internal positions. Structural effects, precursor processing and target-related selection can be tested, but assigning any of them now would convert an observation into an unmeasured mechanism.

The taxonomic analysis shows breadth and variation rather than a simple evolutionary progression. CG and UA depletion had the same overall direction in angiosperms, invertebrate groups and vertebrates, yet their positional profiles differed. The organism-level PCA recovered partially separated macrolineage distributions from the complete 320-feature dinucleotide profile. This pattern indicates that taxonomic composition is associated with the mature-miRNA k-mer landscape, but it does not show that plant, invertebrate and vertebrate miRNAs form discrete biological types. Human remained within the vertebrate distribution. The selected groups also differ in species number, annotation depth and miRNA-family composition. Phylogenetically controlled models are needed before individual profile changes can be assigned to particular evolutionary transitions.

The comparison of the two null models was central to interpretation. Within-sequence shuffling asks whether a motif's abundance or location follows from the bases present in each mature miRNA. The position-wise, length-stratified null asks whether adjacent-base associations remain after the observed

nucleotide composition of every position has been retained. The opposite directions for UA at position 0 show why one null is insufficient. The first null identified localization to the 5' end; the second showed that the localization was weaker than expected from the available first- and second-position nucleotides. Similar null-dependent directions for AA, CC, GC and UU indicate that some apparent dinucleotide effects are largely consequences of positional mononucleotide bias.

Shannon concentration provided a complementary view of the same profiles. It did not identify whether a motif was abundant or depleted; it measured how unevenly that motif was distributed across positions. CG was strongly depleted in the combined miRNA sequence data without being more concentrated than the position-wise null. UA was depleted and concentrated, reflecting its uneven profile. The significant concentration of many other dimers and trimers shows that the positional architecture extends beyond the motifs selected for detailed discussion. These results support use of the complete atlas as a discovery resource rather than a two-motif catalogue.

Limitations

miRBase combines annotations supported by different amounts and types of evidence, and annotation depth differs markedly among organisms [11,14]. The analysis treated deposited mature sequences as a repertoire and did not weight them by expression, tissue, developmental stage or confidence category. Global deduplication was appropriate for the universal sequence-repertoire question, while taxon-level within-organism deduplication retained conserved sequences across organisms. Neither choice removes miRNA-family dependence or unequal taxonomic sampling. The fixed start-position window allowed consistent comparison from the 5' end but should not be interpreted as a fixed 3'-terminal region because mature-miRNA lengths vary. The analysis also lacked precursor structures, arm-of-origin models, genomic loci, methylation data and experimental measurements. Finally, both nulls preserve selected constraints and disrupt others. They identify sequence features not explained by those preserved constraints, but they are not generative models of miRNA biogenesis or evolution.

These limitations lead to specific next tests. Mapping the mature sequences to orthologous precursor loci would permit substitution reconstruction and direct evaluation of the CpG-mutability hypothesis. Controlled mutation of CG and UA contexts in precursor hairpins could test effects on processing, strand loading and mature-miRNA stability while holding the remaining sequence constant. Phylogenetic models and miRNA-family-aware resampling could determine which taxon-level differences persist after shared ancestry and annotation depth are considered. The position-resolved profiles may also provide features for future models of mature-miRNA candidate recognition or mutation tolerance, but any predictive use will require purpose-built training data, held-out species and prospective validation.

Conclusion

Mature miRNAs have a widespread position-dependent dinucleotide and trinucleotide organization. CG and UA were depleted relative to two composition-preserving null models in the globally unique miRNA

sequence data and in every selected taxonomic group. UA nevertheless formed a 5' raw-frequency peak, which was attributable to terminal nucleotide composition rather than an independent UA adjacency preference. UG was enriched, and many other motifs showed position-specific abundance or concentration effects. CpG hypermutability is a plausible explanation for part of the CG pattern, but it was not tested directly. The cause of UA depletion remains open.

Materials and methods

Data source and sequence validation

The overall work is given as a schematic diagram. All mature-miRNA sequences were obtained from the miRBase release 22.1, downloaded in December 2025 (Kozomara et al. 2019). FASTA records were parsed in the 5'-to-3' orientation supplied by miRBase. Sequences were converted to uppercase and validated against the alphabet A, C, G and U. All 48,885 records passed this check. Both 5p- and 3p-annotated mature products were retained. No expression or confidence filter was applied because the study addressed the complete deposited mature-sequence dataset.

The miRBase organism prefix was used as the primary organism identifier. The prefix avoids ambiguities in free-text display names and is consistent with the miRBase taxonomic hierarchy. Sequence lengths, raw record counts and duplicate counts were recorded before analysis.

Deduplication and analytical datasets

Two deduplication schemes were used because the universal and taxonomic analyses asked different questions. For the universal miRNA sequence analysis, identical mature sequences were collapsed globally, irrespective of organism or miRNA name. This produced 30,721 globally unique sequences.

For taxon-level analyses, exact duplicates were removed separately within each organism prefix. The same sequence was retained in different organisms when it occurred in more than one organism. This prevented repeated database entries within an organism from receiving extra weight while preserving the taxonomic occurrence of conserved mature sequences. Taxon-level statistics were calculated from the pooled within-organism unique occurrences in each group.

Taxonomic grouping

Ten mutually exclusive groups were selected for the main comparison: angiosperms; Nematoda; Insecta; non-tetrapod vertebrates; Amphibia; non-avian reptiles; Aves; mammals excluding primates; non-human primates; and Homo sapiens. Non-tetrapod vertebrates comprised the miRBase Teleostei entries and *Petromyzon marinus*. Non-avian reptiles comprised Crocodylia, Testudines and Squamata with birds excluded. Human, non-human primates and other mammals were separated so that nested mammalian categories were not counted as independent groups in the same comparison. Angiosperms included eudicots, monocots and *Amborella trichopoda*.

Every selected organism prefix was assigned to one group only. The taxon analysis included 185 organism entries. Viruses, other eukaryotes, non-angiosperm plants and metazoan groups that did not match the prespecified categories were retained in the dataset but were not pooled into the main taxon comparison. The full organism manifest, NCBI Taxonomy identifiers and inclusion rationale are supplied in Table S1.

Overlapping k-mer counting and normalization

All overlapping dinucleotides (k = 2) and trinucleotides (k = 3) were counted from the 5' end. A sequence of length L contributed L − k + 1 overlapping k-mers. Start positions were indexed from 0.

For motif m at start position p, the observed frequency was

$$f\left(m,p\right)=100\times\ c\left(m,p\right)/n\left(p\right)$$

Where $c\left(m,p\right)$ is the number of eligible sequences containing motif m at p and $n\left(p\right)$ is the number of sequences with length at least $p+k$. Overall abundance across the displayed window was calculated as the sum of motif counts across positions divided by the sum of eligible sequence windows across those positions. Frequencies were therefore not divided by a fixed sequence count at later positions.

Permutation null models

Two nucleotide-shuffling null models were applied. In the within-sequence null, each mature-miRNA sequence was independently shuffled by a Fisher–Yates procedure. This preserved each sequence's length and exact mononucleotide composition while disrupting nucleotide order, adjacent-base associations and positional composition.

In the position-wise, length-stratified null, sequences were first separated by exact length. At every nucleotide position within a length stratum, the observed nucleotides were independently permuted among sequences. This preserved the number of A, C, G and U nucleotides at every position for every sequence length while disrupting associations between adjacent positions. Shuffling was performed within each analytical dataset, so taxon-level nulls did not exchange nucleotides among taxa.

Each dataset was analysed with 10,000 independent permutations per null model. Overlapping dinucleotides and trinucleotides were recounted after every shuffle. The deterministic base seed was set; dataset-, null- and replicate-specific seed sequences were derived from it.

Effect sizes, empirical probabilities and multiple testing

For abundance and position-specific analyses, the effect size was the observed frequency minus the mean frequency across the corresponding permutation distribution, expressed in percentage points. Relative depletion in the taxon summary was calculated as

$$100 \times \frac{(null\ mean - observed)}{null\ mean}$$

Two-sided empirical P values were calculated as twice the smaller of the upper- and lower-tail permutation probabilities. A value of one was added to each tail count and to the number of permutations, preventing zero P values [15]. Values greater than one after doubling were set to one. The Benjamini–Hochberg procedure controlled the false discovery rate [16]. Correction was performed separately for each dataset, k and null model, and separately for the overall-abundance, position-specific and positional-concentration test families. Results with q ≤ 0.05 were called enriched or depleted according to the sign of the effect.

Shannon entropy and positional concentration

For each motif, its frequency profile across positions was normalized to sum to one. Shannon entropy was calculated as $H = -\sum q(p) \ln q(p)$ where $q(p)$is the normalized share of the motif profile at position p [17]. Positional concentration was defined as

$$C = \frac{1 - H}{\ln(20)}$$

A concentration of zero corresponds to a perfectly even profile across the 20 positions; larger values indicate increasing localization. The concentration effect was the observed value minus the mean value from the corresponding null distribution. Empirical probabilities and false-discovery-rate correction were calculated as described above.

Organism-level principal-component analysis

The descriptive principal-component analysis included miRBase organism entries containing at least 100 exact unique mature sequences after within-organism deduplication. Each of the 103 eligible organisms was represented by 320 observed-frequency features: 16 dinucleotides at 20 start positions. Features were mean-centred but not variance-scaled. Principal components were obtained by singular-value decomposition. Covariance ellipses in Fig. 2 summarize broad macrolineage distributions and were not used for hypothesis testing.

Software and reproducibility

Sequence processing and statistical analyses were performed in Python 3.12 using NumPy and pandas. Figures were generated with Python and exported as editable SVG files and annotated with Inkscape (v1.4.4). The exact FASTA input, taxon manifest, observed counts, eligible denominators, complete permutation results and plotting tables accompany the manuscript as supplementary data.

Declarations

Data availability

The mature-miRNA sequences analysed in this study are publicly available from miRBase release 22.1 at https://www.mirbase.org/download/. The exact mature.fa input used here is supplied as Supplementary Data S1. The taxonomic manifest, processed count and frequency tables, complete permutation results and Shannon-concentration results are supplied as Supplementary Data S2–S7. Analysis and figure-generation code will be available at https://github.com/chandhruyuva006/Kmer_Positional_Bias_in_Mature_miRNA_Sequences .

Ethics approval

Not applicable. The study used publicly available sequence records and did not involve human participants, identifiable personal data or animal experimentation.

Author contributions

Chandhru Srinivasan: Conceptualization, methodology, formal analysis, data curation, visualization, and writing—original draft. Mano Chitra Karthikeyan: Writing—review and editing. Abirami Jayaprakash: Writing—review and editing. Indhirakumar Murugesan: Formal analysis and data curation. Vignesh Packiyaraj: Formal analysis and data curation. Antony Joseph Velanganni Arockiam: Supervision and writing—review and editing.

Acknowledgements

The authors acknowledge the Council of Scientific and Industrial Research for the fellowship awarded to Chandhru Srinivasan (File No. 09/0475(11339)/2021-EMR-I), DST-INSPIRE for support to Mano Chitra Karthikeyan (IF200178), and RUSA 2.0 Biological Sciences, ICMR for support to Abirami Jayaprakash. The authors also acknowledge DST-FIST, Government of India, for infrastructure support to the department.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used ChatGPT for language and grammar editing. The author(s) reviewed and edited the output as needed and take full responsibility for the content of the published article.

References

1. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215–233. https://pubmed.ncbi.nlm.nih.gov/19167326. https://doi.org/10.1016/j.cell.2009.01.002
2. Wang X. Composition of seed sequence is a major determinant of microRNA targeting patterns. Bioinformatics. 2014;30(10):1377–1383. https://academic.oup.com/bioinformatics/article/30/10/1377/267311. https://doi.org/10.1093/bioinformatics/btu045
3. Mi S et al. Sorting of Small RNAs into Arabidopsis Argonaute Complexes Is Directed by the 5' Terminal Nucleotide. Cell. 2008;133(1):116–127. http://dx.doi.org/10.1016/j.cell.2008.02.034. https://doi.org/10.1016/j.cell.2008.02.034
4. Seitz H, Tushir JS, Zamore PD. A 5'-uridine amplifies miRNA/miRNA* asymmetry in Drosophila by promoting RNA-induced silencing complex formation. Silence. 2011;2(1):4. http://dx.doi.org/10.1186/1758-907x-2-4. https://doi.org/10.1186/1758-907x-2-4
5. Auyeung VC, Ulitsky I, McGeary SE, Bartel DP. Beyond Secondary Structure: Primary-Sequence Determinants License Pri-miRNA Hairpins for Processing. Cell. 2013;152(4):844–858. https://linkinghub.elsevier.com/retrieve/pii/S0092867413000834. https://doi.org/10.1016/j.cell.2013.01.031
6. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. Cell. 2009;136(2):215–233. https://linkinghub.elsevier.com/retrieve/pii/S0092867409000087. https://doi.org/10.1016/j.cell.2009.01.002
7. Wu X, Bartel DP. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Research. 2017;45(W1):W534–W538. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx323. https://doi.org/10.1093/nar/gkx323
8. Yousef M, Khalifa W, Acar İE, Allmer J. MicroRNA categorization using sequence motifs and k-mers. BMC Bioinformatics. 2017;18(1):170. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1584-1. https://doi.org/10.1186/s12859-017-1584-1
9. Halytskiy VA, Komisarenko S V. 5'-CG-3', 5'-CNG-3', 5'-GC-3' and 5'-GNC-3' site localization in microRNA sequences. Biopolymers and Cell. 2011;27(6):499–505. http://dx.doi.org/10.7124/bc.00011c. https://doi.org/10.7124/bc.00011c
10. Dai W et al. An Alternative Class of Targets for microRNAs Containing CG Dinucleotide. Biology. 2022;11(3):478. http://dx.doi.org/10.3390/biology11030478. https://doi.org/10.3390/biology11030478
11. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2018;47. https://doi.org/10.1093/nar/gky1141. https://doi.org/10.1093/nar/gky1141
12. Cooper DN, Krawczak M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Human Genetics. 1989;83(2):181–188. http://dx.doi.org/10.1007/bf00286715.

https://doi.org/10.1007/bf00286715

13. Pfeifer GP. Mutagenesis at Methylated CpG Sequences. In: DNA Methylation: Basic Mechanisms. Springer-Verlag; p 259–281. http://dx.doi.org/10.1007/3-540-31390-7_10. https://doi.org/10.1007/3-540-31390-7_10
14. Fromm B et al. MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Research. 2019;48(D1):D132–D141. http://dx.doi.org/10.1093/nar/gkz885. https://doi.org/10.1093/nar/gkz885
15. Phipson B, Smyth GK. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology. 2010;9(1). http://dx.doi.org/10.2202/1544-6115.1585. https://doi.org/10.2202/1544-6115.1585
16. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1995;57(1):289–300. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
17. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948;27(4):623–656. http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Figures

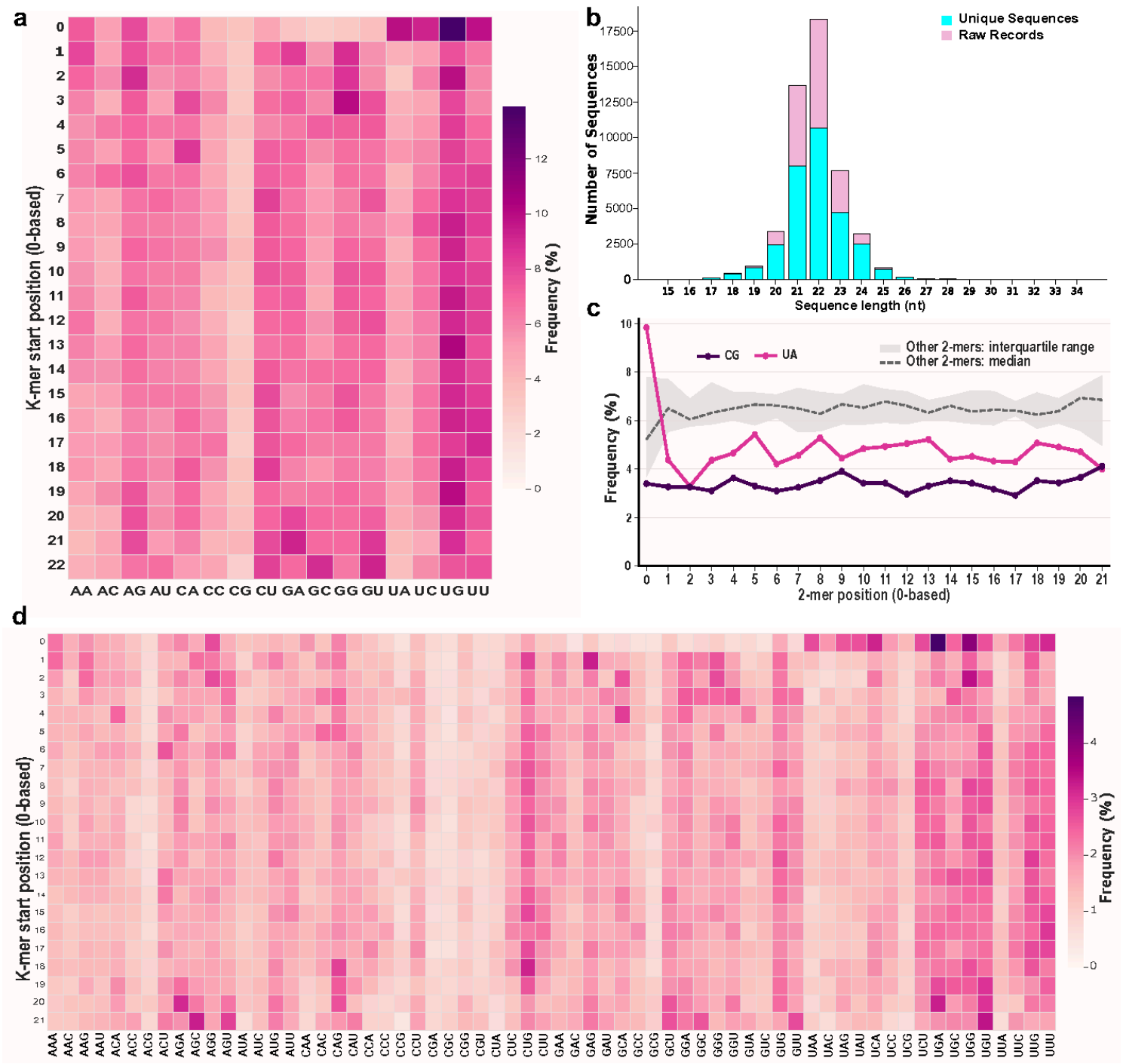


Figure 1

Global mature-miRNA sequence repertoire and observed positional k-mer frequencies

a, Observed positional frequencies of all 16 overlapping dinucleotides in the globally unique mature-miRNA sequences. Exact deduplication of 48,885 miRBase release 22.1 records yielded 30,721 distinct sequence strings. For motif m at zero-based start position p, frequency was calculated as 100×c(m,p)/n(p), where c(m,p) is the number of globally unique sequences containing the motif at that position and n(p) is the number long enough to contribute the complete motif. The displayed

dinucleotide starts, 0-22, each retained at least 10% of the globally unique miRNA dataset as eligible sequences. b, Mature-miRNA length distribution from 15 to 34 nucleotides. The cyan lower segment gives the number of globally unique sequences at each length; the full stacked height gives the corresponding number of raw records, with the pink segment representing records removed by global exact deduplication. c, Observed positional-frequency profiles of CG and UA across dinucleotide starts 0-21. The dashed grey line and shaded band show, respectively, the median and interquartile range of the other 14 dinucleotides at each position. d, Observed positional frequencies of all 64 overlapping trinucleotides. And each cell was normalized by the number of sequences eligible to contribute a complete trinucleotide at that position. Colour scales in a and d show frequency (%).

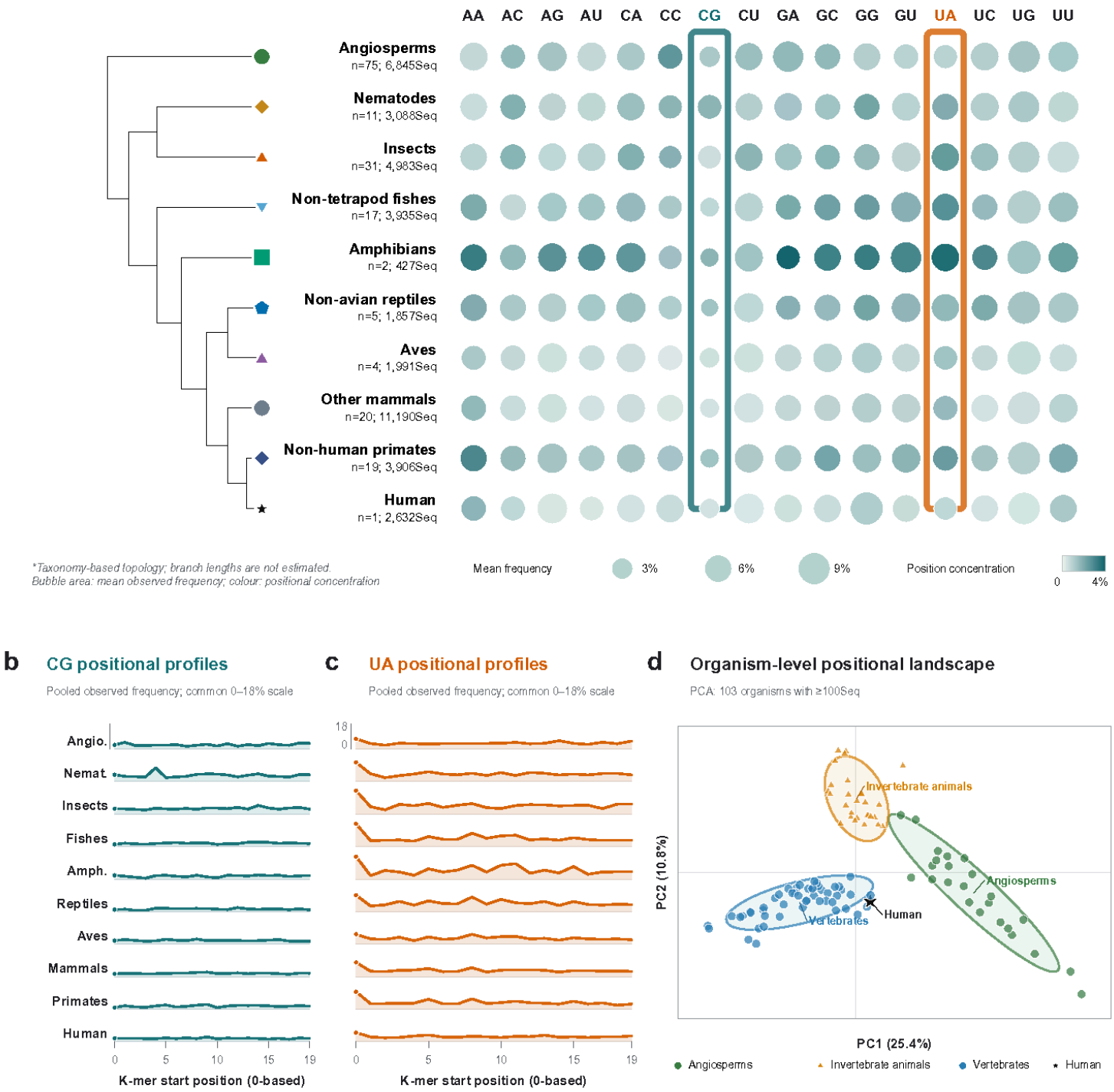


Figure 2

Taxonomic breadth and lineage-associated variation in observed mature-miRNA dinucleotide profiles

a, Taxonomy-aligned plot of the observed dinucleotide variations in ten mutually exclusive analytical groups. Branch lengths are not estimated and do not represent evolutionary distance. Exact duplicate sequences were removed separately within each organism before taxon-level pooling, whereas the same

sequence occurring in different organisms was retained once per organism. Beside each taxon, n is the number of miRBase organism entries and Seq is the summed number of within-organism unique sequence occurrences. Bubble area is proportional to the mean pooled observed frequency of each dinucleotide across zero-based starts 0-19. Bubble colour represents positional concentration, 1-H/ln(20), after the 20-position profile was normalized to sum to one; zero indicates an even positional distribution and larger values indicate stronger localization to a subset of positions. CG and UA are outlined as the focal motifs. b,c, Pooled observed positional-frequency profiles for CG (b) and UA (c) across miRNA sequences. Taxa follow the order in a, and both panels use the same 0-18% frequency scale. d, Principal-component analysis of the complete observed dinucleotide profiles of 103 organism entries containing at least 100 within-organism unique mature-miRNA sequences. Each organism was represented by 320 features (16 dinucleotides x 20 start positions). Features were mean-centred without variance scaling, and components were obtained by singular-value decomposition. PC1 and PC2 explain 25.4% and 10.7% of total variance, respectively. Ellipses are descriptive covariance summaries for the three broad macrolineages, and human is marked separately by a black star. The PCA is descriptive and is not a phylogenetic or evolutionary-distance test. No null-model expectation or significance test is shown in this figure.

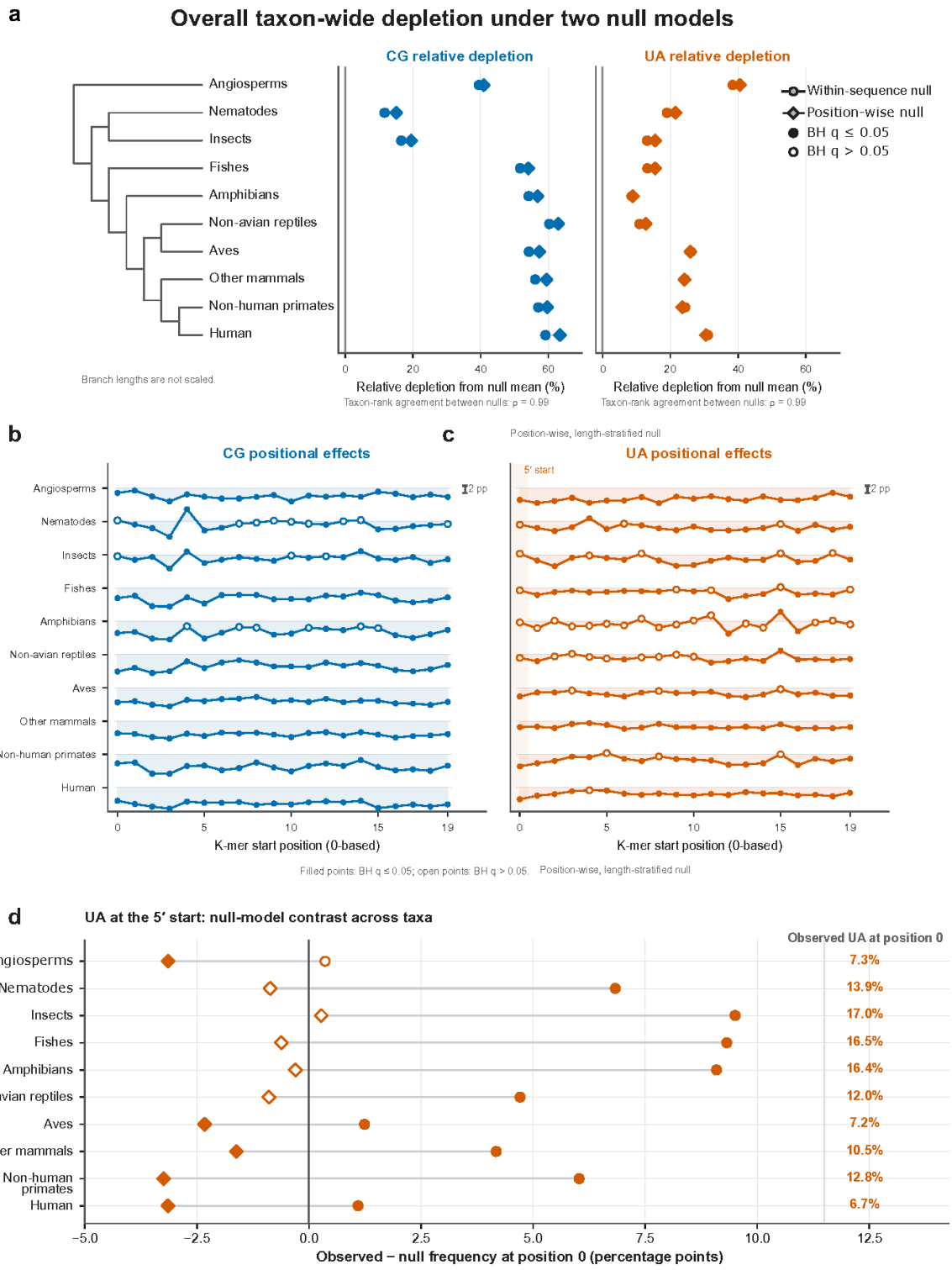


Figure 3

Taxon-wide permutation evidence for CG and UA abundance and positional effects

a, Overall CG and UA depletion in ten mutually exclusive taxonomic groups under two permutation null models. Relative depletion was calculated as 100 × (null mean-observed)/null mean. Circles denote the within-sequence null and diamonds the position-wise, length-stratified null; paired symbols refer to the

same observed taxon profile. The branch lengths are not scaled. The value of rho (ρ) printed below each motif is the Spearman correlation across taxa between the relative-depletion estimates from the two null models. b,c, Position-specific observed-minus-null frequency effects for CG (b) and UA (c) under the position-wise, length-stratified null. Effects are expressed in percentage points across zero-based starts. Rows use a common vertical scale, indicated by the 2-percentage-point scale bar. The shaded strip at start position 0 in c marks the 5' UA start. d, Contrast between the two null models for UA at start position 0. The horizontal axis gives observed minus null mean frequency in percentage points; values printed at the right are the observed UA frequencies at that position. Circles and diamonds retain the null-model definitions used in a. In all panels, filled symbols indicate Benjamini-Hochberg-adjusted q <= 0.05, and open symbols indicate q>0.05, within the corresponding dataset, k-mer length, null model and test family. The within-sequence null independently shuffles the nucleotides of each mature miRNA, preserving its length and exact mononucleotide composition. The position-wise, length-stratified null independently permutes each nucleotide column among sequences of the same exact length, preserving position-specific mononucleotide composition within each length stratum while disrupting adjacent-base associations. Each null distribution contains 10,000 permutations.

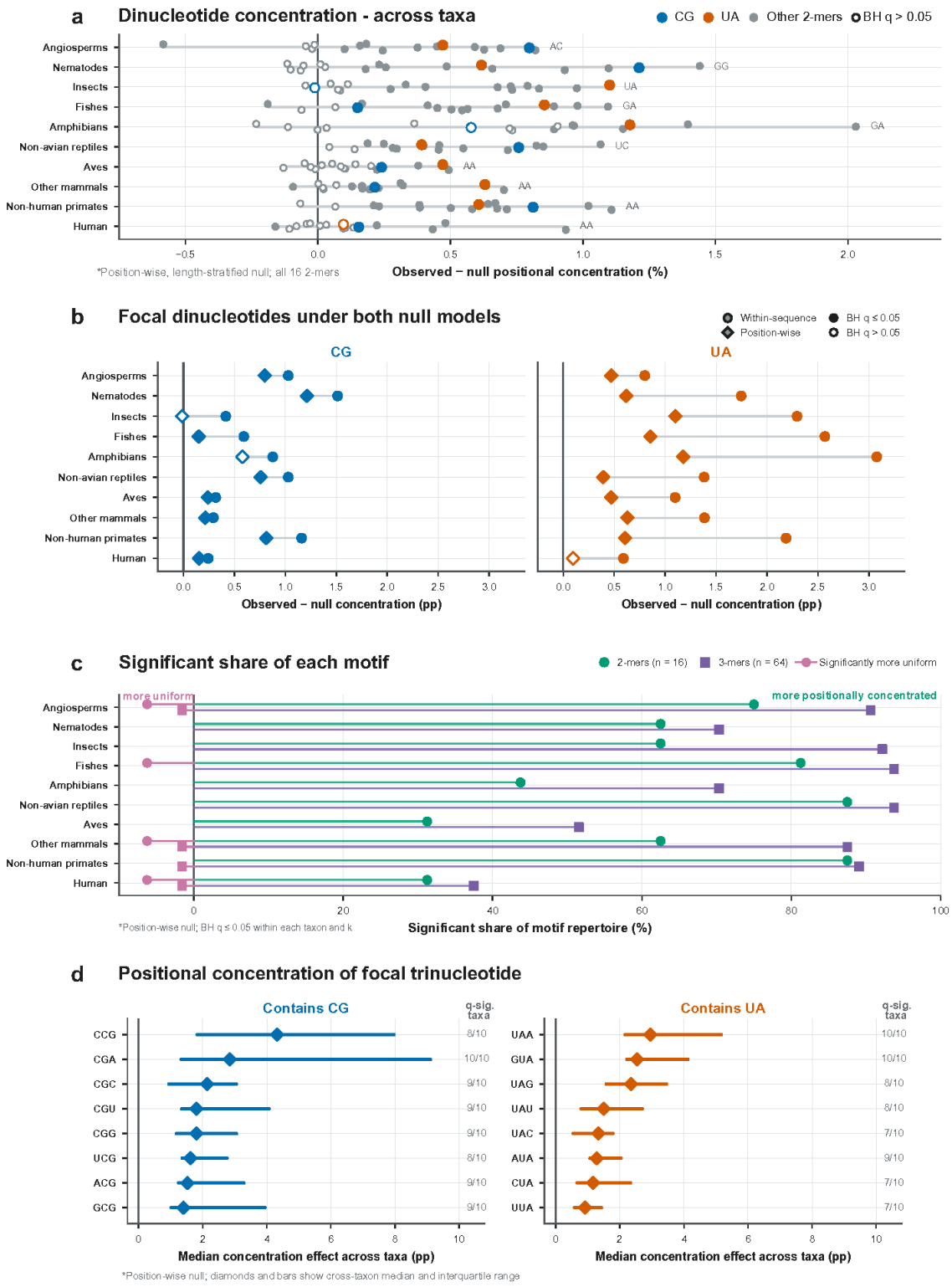


Figure 4

Taxon-wide positional concentration of mature-miRNA k-mer profiles

Positional concentration was defined as 1-H/ln(20), where H is the Shannon entropy of a motif's observed frequency profile across zero-based starts after that profile was normalized to sum to one. Larger values indicate localization to fewer positions. Concentration effects are observed concentration

minus the mean concentration under the specified null and are expressed in percentage points. a, Concentration effects for all 16 dinucleotides in each taxon under the position-wise, length-stratified null. CG and UA are highlighted; other dinucleotides are grey. The label at the right of each row identifies the largest positive dinucleotide effect in that taxon. Filled symbols indicate Benjamini-Hochberg-adjusted q <= 0.05, and open symbols indicate q>0.05. b, CG and UA concentration effects under the within-sequence null (circles) and position-wise, length-stratified null (diamonds). Symbols are connected because the two comparisons use the same observed positional profile. Filled and open symbols retain the significance definitions used in a. c, Percentage of the complete dinucleotide analysis (16 motifs; circles) and trinucleotide analysis data (64 motifs; squares) showing a significant concentration effect under the position-wise null. Values to the right of zero are the shares with significantly greater positional concentration; values mirrored to the left are the shares with significantly greater positional uniformity and are shown in magenta. Significance was defined by Benjamini-Hochberg-adjusted q <= 0.05 within each taxon and k-mer length. d, Cross-taxon median concentration effect and interquartile range for each of the eight CG-containing and eight UA-containing trinucleotides under the position-wise null. Diamonds denote medians and horizontal bars the interquartile ranges; the values at right give the number of taxa with q <= 0.05 out of the ten analysed taxa. The two null models preserve the sequence features described in Figure 3, and each distribution contains 10,000 permutations. Cross-taxon medians and interquartile ranges are descriptive summaries rather than a phylogenetic meta-analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterialzipped.zip
- graphicalabstracteditedfinal.pdf