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.

APAmiRank integrates alternative polyadenylation isoform usage into microRNA target prioritization

Deren Demirel Yavuz
Koç University
Irem Durmaz Sahin
irsahin@ku.edu.tr
Koç University

Research Article

Keywords: alternative polyadenylation, microRNA target prediction, 3' untranslated region, isoform usage, target prioritization, reporter assay, CLASH, CLEAR-CLIP

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11465332/v1

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

Additional Declarations: No competing interests reported.

APAmiRank integrates alternative polyadenylation isoform usage into microRNA target prioritization

Deren Demirel Yavuz[a,b], Irem Durmaz Sahin[a,b,c*]

[a] *Koc University Research Center for Translational Medicine (KUTTAM), Sariyer, 34450, Istanbul, Türkiye*

[b] *Koc University, Graduate School of Health Sciences, Sariyer, 34450, Istanbul, Türkiye*

[c] *Koc University, School of Medicine, Sariyer, 34450, Istanbul, Türkiye*

* Corresponding author. Koç University, School of Medicine, Sariyer, 34450, Istanbul, Turkey.

E-mail addresses: irsahin@ku.edu.tr (Irem Durmaz Şahin), ddemirel19@ku.edu.tr (Deren Demirel Yavuz)

ORCID: 0000-0001-5037-7883 (Irem Durmaz Sahin), 0009-0000-4705-9294 (Deren Demirel Yavuz)

Abstract

Background

Alternative polyadenylation (APA) changes the fraction of transcript isoforms that retain individual microRNA (miRNA)-binding sites, but many target-ranking workflows treat the 3′ untranslated region as static. We developed APAmiRank, a command-line framework that weights fixed canonical-site evidence by measured 3′UTR isoform usage and produces miRNA-specific gene rankings. We evaluated overall target prioritization and the incremental contribution of APA exposure using functional perturbation, direct-interaction, curated reporter, and construct-level evidence.

Results

In six primary cell-line/miRNA comparisons containing 776–2,968 genes, adding the exposure-weighted score to covariates and the matched unweighted score increased repeated cross-validated R^2 by 0.0055–0.0813. Among six miRNAs with at least five observable orthogonally supported reporter targets, APA-exposure positive–unlabeled AUCs ranged from 0.722 to 0.899 and all rank tests passed Benjamini–Hochberg $q<0.001$. The APA-weighted score improved AUC relative to the matched unweighted score for three of these six miRNAs after adjustment. CLEAR-CLIP data supported above-background prioritization for four of five miRNAs, whereas the corresponding CLASH tests did not pass $q<0.05$. In a rank-independent construct-level holdout, 12 of 21 eligible pairs were observable; their mean and median percentiles were 0.7847 and 0.8185, respectively, with four pairs in the top decile. The holdout supported prioritization conditional on 57.1% coverage but did not independently resolve an incremental APA advantage. In an exploratory method-native comparison of the 21 strict holdout pairs, APAmiRank ranked 12, miRDB v6.0 ranked 15, and nine were ranked by both; differing candidate universes precluded a superiority claim.

Conclusions

APAmiRank reproducibly prioritizes experimentally supported targets within its canonical 3′UTR-site universe. Functional and reporter analyses indicate that APA exposure can add predictive information, although the gain varies by miRNA and context. The current evidence supports use as an auditable candidate-prioritization framework; common-universe external predictor comparisons and prospective experiments with matched, replicated APA measurements are needed to establish comparative performance and transferability.

Keywords

alternative polyadenylation; microRNA target prediction; 3′ untranslated region; isoform usage; target prioritization; reporter assay; CLASH; CLEAR-CLIP

Background

MicroRNAs (miRNAs) guide Argonaute-containing silencing complexes to target transcripts and typically reduce gene expression through transcript destabilization and translational repression [1]. Target recognition is strongly influenced by pairing between the miRNA seed region and complementary sites in the target RNA [2]. However, canonical seed pairing alone does not determine regulatory efficacy: site type, local nucleotide composition, site position, and supplementary pairing contribute additional information [3–5]. Contemporary prediction models therefore rank candidate targets using combinations of sequence and contextual features rather than treating every canonical match as equally effective [4, 5].

Even with increasingly refined sequence models, target prioritization remains difficult. Prediction algorithms differ in their candidate universes, transcript annotations, feature sets, conservation requirements, and score calibration, and a canonical or conserved match does not by itself establish a functionally consequential interaction. Experimental analyses have shown that large predicted target sets can contain biologically weak or false-positive assignments [6]. In a standardized reporter-based comparison of 88 target predictors, performance also varied substantially among methods and confidence thresholds [7]. These observations support treating computational outputs as relative evidence for prioritization rather than as definitive binary labels, particularly when the transcript sequence presented in a biological context is uncertain.

A major source of such uncertainty is alternative cleavage and polyadenylation (APA). Selection among polyadenylation sites can generate transcript isoforms with different 3' ends while preserving the protein-coding sequence, thereby changing 3' untranslated region (3'UTR) length, regulatory-element content, localization, stability, and translational control [8]. Proliferating cells frequently express shortened 3'UTRs with fewer predicted miRNA sites [9], and widespread 3'UTR shortening can release specific cancer-associated transcripts from post-transcriptional regulation [10]. Large-scale polyadenylation atlases and tissue-resolved studies further demonstrate that APA is widespread and varies across biological contexts [11, 12]. Thus, the genomic presence

of a miRNA site does not guarantee that the site is retained to the same extent in the transcript isoforms used by a particular cell state.

The functional connection between cellular context, alternative 3'UTR isoform use, and miRNA-mediated repression has been demonstrated directly. Nam et al. showed that changes in 3′UTR isoform abundance across cellular contexts can alter apparent miRNA targeting responses [13]. This finding motivates a distinction between site evidence and site exposure: the sequence features of a candidate site may remain constant, whereas the fraction of expressed transcript isoforms containing that site may vary. Methods such as 3'READS measure cleavage and polyadenylation sites directly [14], whereas tools such as QAPA estimate relative annotated 3'UTR isoform usage from conventional RNA-sequencing data [15]. Other approaches, including DaPars, infer dynamic APA from RNA-sequencing coverage [16]. Comparative evaluations have shown that polyadenylation-site identification and quantification methods differ in their assumptions and performance, reinforcing the need for explicit, reproducible downstream use of APA estimates [17].

The relationship between APA and miRNA regulation is not an unrecognized problem. APADB, for example, combined experimentally supported polyadenylation sites with predicted changes in miRNA-binding-site content and enabled event-level exploration across tissues [18]. TargetScan also incorporated 3′UTR isoform profiles and affected isoform ratios into weighted context predictions [4, 13]. These precedents establish isoform-aware targeting as an existing methodological principle. An APA catalog or quantifier nevertheless addresses a different task from a configurable workflow that exposes the site-level contributions underlying a gene ranking. Here, we focus on an explicit implementation that combines fixed canonical-site evidence with supplied isoform-usage fractions and evaluates the exposure term against an otherwise identical unweighted score. The contribution is therefore the implementation and evaluation of this prioritization framework, rather than the first use of APA information in miRNA targeting.

This distinction matters because APA is quantitative rather than strictly binary. A distal site can be absent from the shortest isoform, partially exposed when proximal and distal isoforms coexist, or nearly constitutive when distal isoforms dominate. Genes can also contain multiple canonical sites distributed across common and extension-specific segments. A ranking framework should therefore preserve site identity and evidence while allowing each site's contribution to vary with the summed usage of the isoforms in which it is present. Comparing

that exposure-aware score with an otherwise identical unweighted architecture provides a direct way to ask whether APA measurements add predictive information, without attributing differences to a changed site universe or a different sequence model.

Evaluation of such a framework requires evidence sources with complementary strengths and limitations. CLASH and CLEAR-CLIP identify ligated miRNA-target chimeras and capture interactions extending beyond canonical seed rules [19, 20], but their coverage depends on assay protocol, ligation efficiency, annotation, expression, and cellular context. Curated resources such as miRTarBase integrate heterogeneous experimental evidence, including reporter assays, Western blotting, quantitative PCR, and high-throughput interaction data [21]. Reporter-based comparisons further show that experimentally supported target sets can be useful for benchmarking predictors while remaining dependent on construct design and evidence selection [7]. Absence from any one interaction map or curated collection cannot therefore be interpreted as a confirmed negative interaction. Positive-unlabeled evaluation, transparent reporting of candidate coverage, and independent construct-level validation are more appropriate than conventional binary classification against unobserved pairs.

Here, we present APAmiRank, a condition-aware framework that combines canonical miRNA-site evidence with the measured usage of 3'UTR isoforms containing each site. APAmiRank is intended to prioritize candidates within a prespecified canonical-site universe rather than to produce calibrated probabilities of targeting. We assessed whether APA exposure contributed predictive information beyond matched unweighted site architecture using miRNA-perturbation data, CLEAR-CLIP and CLASH interaction maps, a curated reporter-supported benchmark, and a rank-independent locked luciferase holdout. This design tests both overall prioritization and the incremental contribution of APA weighting while retaining non-replication, incomplete observability, and assay heterogeneity as explicit limitations.

Methods

APAmiRank candidate universe and ranking

Strand-correct 3′ untranslated region (3′UTR) isoforms were ordered from proximal to distal polyadenylation-site usage within each locus. Canonical 8mer, 7mer-m8, 7mer-A1, and 6mer motifs were generated from the supplied mature miRNA sequence and enumerated across eligible isoforms. Mature sequences and MIMAT

accessions used in the benchmark panels were checked against the official miRBase mature-sequence resource [22]. Genomically identical sites represented in multiple isoforms were collapsed while retaining the first and last polyadenylation-site ranks at which each site was available and its classification as common or extension-specific. For condition-specific analyses, site exposure was defined as the summed usage fraction of all isoforms containing that site. The fixed site-evidence score was multiplied by exposure, and exposure-weighted evidence was summed across sites to obtain a gene-level APAmiRank score. The matched unweighted comparator used the same eligible sites and fixed evidence scores without the exposure multiplier. Both scores were interpreted as relative prioritization statistics rather than calibrated probabilities of targeting.

Percentile ranks were calculated within the miRNA-specific canonical-site candidate universe. Genes without an enumerated canonical site in an eligible 3′UTR were therefore outside the modeled universe rather than classified as negative. Unless otherwise stated, higher percentiles indicated higher APAmiRank priority.

Software implementation and scoring modes

APAmiRank was implemented as a Python command-line package with architecture and thermodynamic scoring modes. All validation analyses reported in this study used architecture mode, which combined canonical seed class and local AU context; consequently, RNAhybrid and RNAup outputs did not contribute to the reported benchmark scores. The optional thermodynamic mode calls RNAhybrid v2.1.2 [23] using the human 3′UTR parameter set (-s 3utr_human) and one reported hit per target (-b 1), and RNAup distributed with ViennaRNA v2.7.2 [24, 25] with auxiliary output-file generation disabled (-o); interaction energies and their components are read from standard output. The equal-weight seven-feature thermodynamic score combines seed strength and local AU context with RNAhybrid coordinate concordance and minimum free energy and RNAup coordinate concordance, total interaction energy, and target-opening energy. The analysed release was APAmiRank v0.2.18, requiring Python ≥3.10, PyYAML ≥6.0, and pyfaidx ≥0.8; Snakemake support is optional. Architecture mode is the biologically validated mode in this study. Thermodynamic mode is implemented and software-tested but remains experimental and was not biologically validated here. The complete ranking workflow and the validation boundary are summarized in **Fig. 1.**

Functional perturbation benchmark

The functional benchmark combined processed human 3P-seq poly(A)-site profiles from GEO series GSE52527 with quantile-normalized RNA-sequencing data following miR-124-3p, miR-155-5p, or mock transfection from GSE52530, the two relevant subseries of GSE52531 [13]. Processed files were retrieved from the NCBI Gene Expression Omnibus (GEO) [26]. HeLa, HEK293, Huh7, and IMR90 contexts were processed on the hg19 genome assembly using the hg19 sequence and refGene transcript coordinates obtained from the UCSC Genome Browser download resource [27].One processed 3P-seq profile was available per cell type. The miRNA-transfection RNA-sequencing data comprised two biological replicates per treatment; the separate Huh7 mock pairs corresponding to miR-124-3p and miR-155-5p were retained rather than pooled. The primary expression threshold was a mean control RPKM of at least 1. The response variable was −log2(treatment/control), calculated without a pseudocount.

Four ordinary least-squares models were fitted on identical gene sets: covariates only (C), C plus the unweighted architecture score (C+U), C plus the exposure-weighted score (C+W), and C plus both scores (C+U+W). Covariates comprised log10 mean control RPKM, log1p maximum benchmark 3′UTR length, log1p canonical site count, and mean raw AU fraction in 30-nt site flanks. Predictive performance was summarized as the median pooled out-of-fold R^2 across 100 deterministic repetitions of 10-fold cross-validation. The 2.5th–97.5th percentile range across repetitions was treated as a stability interval, not a confidence interval, because genes were reused across repetitions. The difference between C+U+W and C+U quantified predictive information contributed by W conditional on U and was not interpreted as mechanistic independence. In-sample partial F tests were secondary analyses and were not used as substitutes for out-of-fold validation.

Direct-interaction benchmarks

Ligation-based interaction data were evaluated as positive–unlabeled benchmarks. The primary analysis used CLEAR-CLIP interactions generated in Huh-7.5 cells (GSE73057) [20], downloaded from NCBI GEO [26], restricted to annotated 3′UTRs and compared with Huh7 APA profiles; Huh-7.5 and Huh7 were considered near-matched rather than identical contexts. The independent analysis used CLASH interactions generated in HEK293 cells (GSE50452) [19], also obtained from NCBI GEO [26]. CLASH transcript coordinates were mapped to hg19

UCSC ensGene models downloaded from the UCSC Genome Browser resource [27] and assigned to 5′UTR, coding-sequence, or 3′UTR regions; only the 3′UTR subset was used for the primary comparison. The prespecified panel comprised miR-16-5p, miR-17-5p, miR-25-3p, miR-92a-3p, and miR-186-5p.

For each miRNA, direct-interaction genes were compared with all remaining genes in the same canonical-site universe. Endpoints were positive–unlabeled area under the receiver-operating-characteristic curve (PU-AUC), a one-sided Mann–Whitney rank test, and top-decile and top-quintile hypergeometric enrichment. Mann–Whitney P values were adjusted using the Benjamini–Hochberg procedure across the five-miRNA family within each dataset, genomic region, and rank-metric family. Absence from CLEAR-CLIP or CLASH was not interpreted as evidence of a confirmed negative interaction.

Curated reporter-assay benchmark

A human Reporter Assay export was obtained from the official miRTarBase portal on 8 September 2026 [21]. Because the download endpoint was not version tagged, the file was treated as a dated portal snapshot rather than an immutable miRTarBase version 10 release. Of 7,790 exported rows, 7,669 were exact human-miRNA–human-target records. The seven-miRNA panel contributed 513 unique harmonized pairs and comprised miR-16-5p, miR-17-5p, miR-25-3p, miR-92a-3p, miR-124-3p, miR-155-5p, and miR-186-5p.

The primary positive set required Reporter Assay evidence together with Western blot or quantitative PCR support. Because the underlying studies spanned heterogeneous cell types and the export did not provide a usable cell-line annotation for every record, interactions were not classified as cell matched. The primary consensus universe required each gene to be rankable in at least two of four APA contexts: HeLa, HEK293, Huh7, and IMR90. Available percentile ranks were averaged and the resulting consensus score was re-ranked. Requiring rankability in all four contexts and retaining all reporter records were evaluated as sensitivity analyses.

Primary endpoints were candidate coverage, PU-AUC, one-sided Mann–Whitney tests with Benjamini–Hochberg adjustment across the seven-miRNA family, and top-decile and top-quintile enrichment. To isolate the effect of the exposure term within an otherwise matched architecture, APA-exposure and unweighted-additive PU-AUCs were compared on identical genes using 5,000 paired stratified bootstrap samples of positive and unlabeled genes. Inferential results were suppressed when fewer than five positive genes were observable.

Construct-level reporter pilot

Candidate pairs were manually audited for a reporter containing the human target 3′UTR, repression of the wild-type construct, and loss or material weakening of repression after mutation of the cognate miRNA site. Four pairs met all criteria: miR-16-5p–BCL2, miR-17-5p–BMPR2, miR-25-3p–PTEN, and miR-124-3p–RAB27A. A separately labeled sensitivity scope added miR-155-5p–ETS1 when construct-level corroboration was accepted. Because this audit was conducted after inspection of the broad reporter benchmark, it was reported descriptively without a P value or confidence interval.

Rank-independent locked luciferase holdout

To reduce outcome-aware selection, a deterministic manifest of 33 previously unaudited miRNA–target pairs was generated without rank or score columns. Original articles were manually screened under a frozen construct-level protocol before the rank table was opened. Eligibility required the exact mature miRNA, a human target, a reporter containing the cognate human target 3′UTR sequence, repression of the wild-type construct, and loss or material weakening of repression after mutation of the cognate site. Fail-closed exclusions were retained as protocol or provenance failures and were not reclassified as biological negatives. The completed screening form was frozen with SHA-256 checksum 8f7ecc1f89a3819365fdfc5ccbf7799d6d13ac7042874c0023d93fe779160ae5 before eligible pairs were joined to the prespecified consensus rank table.

The prespecified evaluation was descriptive and included coverage, mean and median APA-exposure percentile, fractions in the top decile and top quintile, and the paired median difference between APA-exposure and unweighted percentiles. Missing pairs remained missing; no scores were imputed and the candidate universe was not modified.

After descriptive results were opened, a separately labeled exploratory uncertainty analysis was designed. A theoretical rank-uniform null preserved, for each of the six represented miRNAs, the number of observable holdout pairs and the size of the corresponding full consensus ranking universe. In each of 50,000 iterations, ordinal rank positions were sampled without replacement and converted to percentiles. One-sided Monte Carlo P

values used a plus-one correction, and the four correlated tests were adjusted using Holm's method. This null did not preserve empirical score-tie multiplicities and was conditional on observability.

For cluster-aware sensitivity analysis, miRNA-specific statistics were calculated first and the six miRNA clusters were sampled with replacement for 50,000 iterations with equal miRNA weight. The 2.5th and 97.5th percentiles were reported. For APA-minus-unweighted ranking, the cluster statistic was the within-miRNA median paired difference. Given the presence of only six clusters, these intervals were interpreted as exploratory sensitivity estimates.

Reproducibility and statistical interpretation

All input transformations, eligibility rules, model specifications, random seeds, multiple-testing families, rank tables, and pair-level joins were retained with the APAmiRank v0.2.18 benchmark code. Raw public resources were retrieved from the NCBI GEO, UCSC Genome Browser, miRBase, and miRTarBase locations recorded in the research package, and checksums were retained for the downloaded direct-interaction, annotation, mature-miRNA, and reporter files [21, 22, 26, 27]. Candidate coverage was reported throughout. Tests used the prespecified canonical-site universe and did not create negative labels from unlabeled observations. P values were nominal unless an adjustment method was explicitly stated. Secondary P values for the locked holdout were generated after unblinding and were therefore considered exploratory even after Holm adjustment. All automated tests included in the v0.2.18 research package passed at release. Detailed provenance, benchmark procedures, computational settings, and reproducible execution information are provided in Supplementary Methods S1–S8 (Additional file 1). Public resource, accession, version, retrieval, and checksum information is summarized in Supplementary Table S1 (Additional file 2).

Computational scaling benchmark

Practical scaling was evaluated with deterministic synthetic architecture-mode inputs containing 100, 1,000, or 5,000 genes and two ordered 3′UTR isoforms per gene. Each run used 250 bootstrap iterations and 1,000 opportunity iterations in APAmiRank v0.2.18 under Python 3.12.14. Wall time was measured with a monotonic process clock, and maximum resident set size was collected from the child-process resource record. A single

measured run per size was reported descriptively; the benchmark was intended to characterize operational scale, not to estimate a population parameter. Synthetic input and output sizes, commands, and the fixed seeds are provided in **Supplementary Methods and Table S6**.

Exploratory external-predictor comparison

A version-locked exploratory comparison used miRDB v6.0 predictions generated by MirTarget v4 against miRBase release 22 [28, 29]. Official miRDB result pages for the seven-miRNA panel were retrieved on 16 September 2026 and preserved with SHA-256 checksums. miRDB scores were converted to within-miRNA native percentiles without treating an absent prediction as a biological negative. Recovery was summarized for the dated reporter-positive sets and for the 21 strict locked-holdout inclusions. For the nine holdout pairs observable by both methods, native percentiles were compared descriptively and with a two-sided paired Wilcoxon test. Because APAmiRank and miRDB used different transcript annotations, site rules, and candidate universes, this analysis was not a common-universe accuracy benchmark and was not interpreted as evidence of superiority.

Computational environment and generative-AI assistance

Analyses were executed in Linux environments, including a high-performance-computing (HPC) system; MobaXterm was used only as an SSH terminal and file-transfer client for remote access. OpenAI Codex was used as an interactive coding and editorial assistant for drafting and refactoring analysis code, diagnostic review, literature-screening support, and language editing. It did not determine eligibility rules, interpret primary articles autonomously, or make final scientific decisions. The author reviewed the generated code, source extractions, numerical outputs, statistical interpretations, and manuscript text and retains full responsibility for the work.

Results

APA exposure contributed predictive information in the functional perturbation benchmark

In the GSE52527/GSE52530 functional benchmark [13, 26], at the mean-control RPKM threshold of at least 1, the six primary cell-line/miRNA combinations contained 776–2,968 genes. Adding the exposure-weighted score to the unweighted model increased cross-validated R^2 in all six comparisons. The C+U+W minus C+U

difference ranged from +0.0055 for Huh7–miR-155-5p to +0.0813 for HEK293–miR-155-5p. Direct comparison of C+W with C+U favored exposure weighting in five of six primary combinations; Huh7–miR-155-5p was the exception, with a difference of −0.0010. The largest improvements were observed for miR-155-5p in HeLa (+0.0720 for C+W versus C+U) and HEK293 (+0.0821). In the supporting IMR90–miR-124-3p comparison, C+W was inferior to C+U (−0.0124), whereas W retained a positive increment when added to U (+0.0121). Thus, in this dataset, the exposure-weighted score contributed predictive information beyond the prespecified covariates and unweighted score, but the magnitude and form of the contribution varied by biological context. The matched model results and APA-exposure increments are shown in **Fig. 2A–B.**

The condition number of the full standardized model ranged from 4.07 to 5.40, providing no indication of severe numerical instability. In-sample partial F tests for adding W to C+U were nominally significant in all six primary comparisons (largest P=0.0105), but these tests were not interpreted as evidence of independent generalization or a causal mechanism. Full model summaries are provided in **Supplementary Table S2**.

Direct-interaction prioritization replicated in CLEAR-CLIP but not CLASH

In the primary Huh-7.5 CLEAR-CLIP 3′UTR analysis (GSE73057) [20, 26], APA-exposure PU-AUC values were 0.671 for miR-16-5p, 0.652 for miR-17-5p, 0.667 for miR-25-3p, 0.551 for miR-92a-3p, and 0.648 for miR-186-5p (Fig. 3A). One-sided rank tests remained significant after Benjamini–Hochberg adjustment for miR-16-5p ($q=2.76\times10^{-4}$), miR-17-5p ($q=2.73\times10^{-7}$), miR-25-3p (q=0.0431), and miR-186-5p (q=0.00525), but not miR-92a-3p (q=0.159). Candidate coverage ranged from 15.6% to 30.3%, reflecting both the canonical-site restriction and the presence of interactions outside the modeled 3′UTR universe. Top-decile enrichment was greater than one for all five miRNAs and ranged from approximately 2.1-fold to 3.0-fold **(Fig. 3B)**. Complete summary statistics are reported in **Supplementary Table S3**.

The independent HEK293 CLASH 3′UTR analysis (GSE50452) [19, 26] was weaker and heterogeneous. APA-exposure PU-AUC values ranged from 0.504 to 0.601, and none of the five one-sided rank tests passed q<0.05 (Fig. 3A). Top-decile enrichment ranged from approximately 0.44-fold to 1.36-fold and exceeded one for only two miRNAs **(Fig. 3B)**. This result was treated as a non-replication of the stronger CLEAR-CLIP rank-distribution effect. Differences in assay protocol, annotation, cell state, and the prevalence of noncanonical

interactions remain plausible explanations, but none was used as a post hoc exclusion criterion. Exposure-weighted and unweighted rankings also lacked a uniform direction of difference across these direct-interaction datasets, precluding a general APA-increment claim from this evidence layer.

Reporter-supported targets were strongly prioritized with miRNA-specific APA increments

In the orthogonally supported miRTarBase Reporter Assay set [21] and the at-least-two-context consensus universe, APA-exposure PU-AUC ranged from 0.722 to 0.899 among the six miRNAs with at least five observable positives **(Fig. 4A)**. All six passed Benjamini–Hochberg $q<0.001$. The results were miR-16-5p, PU-AUC 0.830 (25 of 49 positives observable; $q=2.35\times10^{-8}$); miR-17-5p, 0.782 (24/45; $q=2.23\times10^{-6}$); miR-25-3p, 0.824 (9/25; $q=4.48\times10^{-4}$); miR-92a-3p, 0.899 (8/29; $q=8.76\times10^{-6}$); miR-124-3p, 0.802 (34/74; $q=4.82\times10^{-9}$); and miR-155-5p, 0.722 (23/72; $q=1.74\times10^{-4}$). miR-186-5p had only two observable positives and was not interpreted inferentially. Top-decile enrichment was greater than twofold for every panel member and ranged from approximately 2.2-fold to 6.7-fold **(Fig. 4B)**. Pairwise ranking and bootstrap summaries are provided in Supplementary Table S4.

APA-exposure PU-AUC exceeded unweighted-additive PU-AUC for every panel miRNA, but bootstrap support for the increment was not uniform **(Fig. 4C)**. Differences remained positive after Benjamini–Hochberg adjustment for miR-16-5p (+0.048, 95% bootstrap interval 0.005–0.087, q=0.0340), miR-25-3p (+0.070, 0.017–0.128, q=0.0150), and miR-124-3p (+0.060, 0.026–0.093, q=0.00120). Intervals crossed zero for miR-17-5p, miR-92a-3p, and miR-155-5p; miR-186-5p was not tested because fewer than five positives were observable. Sensitivity analyses retaining all reporter records preserved positive ranking performance, whereas requiring rankability in all four APA contexts substantially reduced coverage for several miRNAs.

Construct-level reporter examples occupied the upper tail of the rankings

All four full-text-verified wild-type/cognate-mutant reporter pairs selected through the miRTarBase-derived audit [21] were observable and fell within the top APA-exposure decile. Their mean percentile was 0.9673 and their median percentile was 0.9733. When construct-corroborated miR-155-5p–ETS1 was included in the separately labeled sensitivity scope, all five pairs remained in the top decile, with a mean percentile of 0.9668 and

a median of 0.9649. These examples demonstrate that selected construct-validated targets can occupy the extreme upper tail of APAmiRank rankings. Because the audit followed inspection of the broad reporter benchmark and included only four or five pairs, it was not treated as an unbiased performance estimate.

The locked holdout supported ranking ability conditional on 57.1% coverage

Of the 33 rank-free candidate pairs locked from the dated miRTarBase snapshot [21], 21 satisfied every construct-level inclusion criterion and 12 were excluded for protocol or provenance reasons. Twelve of the 21 strict inclusions were observable in the frozen consensus rank table, corresponding to 57.1% coverage; the remaining nine pairs were retained as missing. Among the observable pairs, the mean APA-exposure percentile was 0.7847 and the median was 0.8185 **(Fig. 5A)**. Four of 12 pairs fell in the top decile and seven of 12 fell in the top quintile. The paired median APA-exposure-minus-unweighted percentile difference was +0.0651. Pair-level evidence and ranks are listed in **Supplementary Table S5**.

In the explicitly post-unblinding matched-null analysis, the pooled observed mean percentile exceeded the theoretical null mean (0.7847 versus 0.4992; Holm-adjusted P=0.00080; **Fig. 5B**). Holm-adjusted P values were 0.01164 for the median percentile, 0.02534 for the top-decile fraction, and 0.01164 for the top-quintile fraction. Equal-miRNA cluster analysis estimated a mean percentile of 0.7915 with a 95% bootstrap interval of 0.7030–0.8847. In contrast, the cluster-equal APA-minus-unweighted estimate was +0.0423 and its interval crossed zero (−0.0136 to 0.0899; **Fig. 5C**). The holdout therefore supported prioritization of construct-validated targets conditional on observability but did not independently establish an incremental advantage from APA weighting. The complete locked-candidate screening and observability flow is summarized in **Fig. 5D.**

Architecture-mode runtime scaled approximately linearly in the tested range

For deterministic synthetic inputs containing 100, 1,000, and 5,000 genes (200, 2,000, and 10,000 isoforms), measured wall times were 0.162, 1.092, and 5.443 s, respectively (**Supplementary Fig. S1, Additional file 3; Supplementary Table S6**). Recorded maximum resident set size was 230.9 MB for all three runs, indicating that interpreter and package overhead dominated memory use within this range. Input size increased from 0.167 to

8.315 MB and output size from 0.542 to 27.295 MB. These single-run synthetic measurements describe the tested environment and should not be treated as hardware-independent performance guarantees.

Exploratory miRDB comparison showed complementary coverage

Among the 21 strict locked-holdout inclusions, APAmiRank ranked 12 pairs, miRDB v6.0 ranked 15, and nine were observable by both methods (**Supplementary Fig. S2; Supplementary Tables S7–S8**). Within these nine common observable pairs, the median method-native percentile was 0.8220 for APAmiRank and 0.6636 for miRDB, with a median paired difference of +0.1584 (two-sided Wilcoxon P=0.00781). Across the broader reporter-positive sets, miRDB coverage varied by miRNA from 47.2% to 73.5%. These figures describe recovery under each method's native universe. They do not establish comparative accuracy because the universes were not harmonized, absent miRDB predictions were not imputed as negatives, and the same literature-derived evidence underlies both the reporter panel and holdout selection.

Cross-benchmark interpretation

Three findings converged on the primary conclusion. First, reporter-supported targets were strongly prioritized across the evaluable miRNA panel. Second, the rank-independent holdout placed strict construct-validated targets above the midpoint of their respective candidate universes. Third, exposure-aware scores contributed predictive information in the matched perturbation dataset. CLEAR-CLIP provided additional support for direct-interaction prioritization. However, CLASH did not reproduce the stronger rank-distribution effect, candidate coverage was incomplete, and the gain from APA exposure relative to the matched unweighted score varied across miRNAs and contexts. Collectively, these results support APAmiRank as a reproducible candidate-prioritization framework with context-dependent evidence for additional information from APA exposure; they do not establish universal improvement or calibrated target-prediction accuracy.

Discussion

Principal findings and scope of the evidence

APAmiRank prioritized experimentally supported miRNA targets within a defined canonical 3′UTR-site universe, while the additional value of APA exposure depended on the evaluation setting. Reporter-supported

targets were enriched near the top of the rankings, and the locked construct-level holdout supported prioritization among observable pairs. The exposure-weighted score also added predictive information to the unweighted functional model in all six primary cell-line/miRNA comparisons. However, neither the direct-interaction benchmarks nor the holdout established a uniform advantage from exposure weighting. These findings support a practical use for APAmiRank in selecting candidates for follow-up, with a narrower claim for its APA component than for its overall ranking performance.

Relationship to existing prediction and APA frameworks

Isoform-aware miRNA targeting has clear precedents. Nam et al. demonstrated that alternative 3′UTR usage can influence targeting across cellular contexts [13], and TargetScan incorporated affected isoform ratios into weighted context++ predictions [4]. Thus, weighting a site according to the isoforms that retain it is not itself a new biological principle. APAmiRank makes this contribution explicit at the site level and compares exposure-weighted and unweighted evidence on an identical candidate universe. Its proposed value lies in a transparent, configurable prioritization workflow and the accompanying validation design; the present results do not establish an advance over TargetScan's scoring model.

FilTar provides a closely related example of context-aware prediction: it uses RNA-sequencing evidence to revise sample-specific 3′UTR boundaries and filters predicted interactions according to transcript expression [30]. APAmiRank instead represents site retention continuously through the summed usage of isoforms containing each site. These approaches address related but nonidentical questions: whether a candidate transcript and its annotated target region are present, and how strongly a retained site should contribute when multiple isoforms coexist. PRIMATA-APA examines another level of the same biology by quantifying collective APA-associated changes in miRNA target-site availability across the transcriptome [31]. Its miRNA-level landscape analysis differs from APAmiRank's gene prioritization for a specified miRNA. QAPA and DaPars address upstream estimation of APA [15, 16], whereas APAmiRank consumes isoform-usage information rather than estimating it from sequencing reads.

Sequence-based resources also differ in scope. DIANA-microT 2023 provides predictions for both coding sequences and 3′UTRs together with expression and other contextual annotations [32], whereas the current

APAmiRank benchmarks are restricted to canonical 3′UTR sites. Thermodynamic accessibility is a further, distinct property: PITA evaluated the energetic cost of exposing a site within RNA structure [33]. In APAmiRank, exposure denotes the fraction of isoforms retaining a site, not the probability that its nucleotides are structurally unpaired. RNAhybrid and RNAup supply optional thermodynamic features [23–25], but all biological benchmarks here used architecture mode. Their performance therefore cannot be attributed to those external programs or interpreted as validation of the thermodynamic extension.

Interpreting the contribution of APA exposure

The functional comparison most directly tested the exposure term because weighted and unweighted scores used identical sites and fixed evidence, with expression, 3′UTR length, site count, and AU context included as covariates. The observed improvement is consistent with the relevance of isoform usage reported by Nam et al. [13]. However, we reanalyzed data from that study; agreement is not an independent replication of its biological finding. Moreover, cross-validation across genes within each condition estimates within-dataset prediction, not transfer to an unseen tissue or study. The positive increment from adding W to C+U, despite occasional underperformance of C+W relative to C+U, suggests that weighted and unweighted summaries can carry complementary information. It does not show that exposure weighting should replace the unweighted score in every setting.

The reporter benchmark similarly separates overall prioritization from incremental benefit. PU-AUC values of 0.722–0.899 indicate that the six adequately represented positive sets tended to rank above unlabeled candidates, but only three miRNAs had supported APA-minus-unweighted increments after multiple-testing adjustment. In the locked holdout, the equal-miRNA increment was +0.0423 with an interval spanning zero. Consequently, the holdout supports recovery of validated targets without independently resolving the APA advantage. A mechanistic expectation is that exposure weighting will matter most when informative sites lie in variably retained extensions; little rank change is expected when most informative sites are constitutively retained. This interpretation remains a hypothesis because the present analyses did not prospectively test a prespecified interaction between site location, APA variation, and performance gain.

Complementary assays and the limits of convergence

The benchmarks measure different aspects of targeting. Chimeric CLIP data support physical miRNA–RNA associations [19, 20], whereas wild-type/cognate-mutant reporters test sequence-dependent repression in a constructed assay. Perturbation RNA-sequencing measures transcript responses, which can also include indirect effects. Transcript abundance is a biologically relevant endpoint: Eichhorn et al. found that mRNA destabilization accounts for much of mammalian miRNA-mediated repression once substantial repression develops [34]. Nevertheless, our RNA-sequencing analysis cannot resolve all translational effects or demonstrate direct regulation for each gene. Evidence across these assays is complementary rather than interchangeable, and it does not constitute multiple fully independent replications because APA profiles, miRNA panels, and reporter provenance are partly shared.

CLEAR-CLIP supported above-background prioritization more consistently than CLASH, where none of the five primary rank tests passed the adjusted significance threshold. Both original studies identified interactions beyond canonical seed matching [19, 20], but the evaluated APAmiRank universe excludes such sites. Differences in protocol, cell state, annotation, or interaction composition could contribute to the discrepancy; the current comparison cannot identify the cause. Huh-7.5 interaction data were also paired with Huh7 APA measurements, not an exact cellular match. The weaker CLASH result therefore remains a constraint on generalization rather than a dataset to discard. Likewise, high reporter recovery should not be read as more accurate measurement of endogenous binding: reporter selection and construct design answer a different experimental question.

Limitations

The principal comparative limitation is that the executed miRDB v6.0 analysis was method-native rather than a harmonized head-to-head benchmark. It adds version-locked external coverage and rank context, but APAmiRank and miRDB differ in transcript annotations, site rules, and candidate universes. The planned TargetScan bulk comparison could not be completed because the required version-matched bulk files were not available during retrieval. The matched unweighted baseline isolates the exposure term within APAmiRank, but neither it nor the exploratory miRDB comparison establishes superiority to TargetScan, FilTar, miRDB, or another

method. A common transcript and gene universe, fixed eligibility criteria, and prospectively specified metrics remain necessary before making comparative-accuracy claims.

Coverage and label uncertainty further restrict interpretation. The canonical-site and annotation requirements exclude noncanonical interactions, coding-region sites, and otherwise unobservable genes. Only 12 of 21 strict holdout inclusions were rankable, so its favorable rank summaries apply to 57.1% of eligible pairs; missingness may be systematic. The remaining benchmark genes are unlabeled, not verified negatives. PU-AUC therefore measures separation from an incompletely annotated background and does not estimate specificity, positive predictive value, or calibrated targeting probability. Published reporter collections may also favor well-studied genes and candidates originally selected through seed-based prediction. Construct-level screening improves evidence specificity but does not remove that ascertainment bias. Gene-level recovery additionally does not demonstrate that the exact experimentally assayed site drove the score.

APA estimates introduce uncertainty that was not propagated into ranking intervals. One processed 3P-seq profile per cell type precluded estimation of biological variation in isoform usage, and the historical hg19 annotations and conservative mapping filters can affect site retention and coverage. Benchmarks of APA quantification methods show why dependence on upstream estimation deserves explicit evaluation [17]. The four-context reporter consensus accommodates heterogeneous source studies but does not recreate their endogenous isoform mixtures or prove cell specificity. Extrapolation to primary tissues, patient samples, other species, or different sequencing protocols remains untested.

The locked screening design reduced rank-aware selection, but it was not a prospective experimental validation or a preregistered trial. The small post hoc pilot is illustrative only, and the holdout uncertainty analyses were designed after unblinding. Its theoretical rank-uniform null does not preserve empirical score ties or model the missingness process, while bootstrap inference is fragile with only six miRNA clusters. Repeated gene-level cross-validation also does not create independent biological replicates, and significance adjustments within specified test families do not remove broader analytical-selection uncertainty. Future evaluation should freeze endpoints and exclusion rules before examining independent data.

Finally, architecture mode is a deliberately limited model of targeting. Additive aggregation does not explicitly represent site cooperativity, saturation, Argonaute occupancy, or competition among transcripts, and its

fixed evidence terms do not capture all determinants considered in richer targeting models [3–5]. Computational exposure is not a measurement of endogenous binding or repression. The optional thermodynamic mode requires separate biological and computational benchmarking; passing software tests alone is insufficient. No clinical performance, therapeutic efficacy, or prospective experimental hit rate can be inferred from the current analyses.

Implications for subsequent validation

The immediate priorities are a common-universe external predictor comparison and validation with replicated APA measurements matched to the interaction or perturbation experiment. A prospective reporter panel should span the ranking distribution, include common and extension-specific sites, and retain failures and unrankable candidates. Testing short and long 3′UTR constructs alongside cognate-site mutants, followed where feasible by endogenous site or polyadenylation-site perturbation, would separate sequence recognition from isoform-dependent exposure. Study-level or cell-type-held-out evaluation and uncertainty propagation from APA quantification would then address transferability. Until these tests are available, APAmiRank is best used to generate auditable candidate lists whose site architecture and exposure contributions can guide experimental selection.

Conclusions

APAmiRank provides a reproducible framework for ranking miRNA targets while accounting for the fraction of expressed 3′UTR isoforms that retain each canonical site. Experimentally supported targets were preferentially recovered across reporter and locked construct-level evaluations, and exposure weighting added predictive information in the functional perturbation benchmark. The size and statistical support of this increment varied among miRNAs and evidence sources, and the independent CLASH analysis did not reproduce the stronger CLEAR-CLIP result. These findings support APAmiRank as an auditable candidate-prioritization tool within its defined canonical-site universe. They do not establish calibrated target probabilities or superiority to established predictors. Common-universe external comparisons replicated and sample-matched APA measurements, and prospective tests of common versus extension-specific sites will be required to determine comparative accuracy and the conditions in which APA weighting provides the greatest benefit.

List of abbreviations

APA: alternative polyadenylation; 3′UTR: 3′ untranslated region; miRNA: microRNA; PAS: polyadenylation site; PU-AUC: positive–unlabeled area under the receiver-operating-characteristic curve; CLASH: crosslinking, ligation, and sequencing of hybrids; CLEAR-CLIP: covalent ligation of endogenous Argonaute-bound RNAs followed by crosslinking immunoprecipitation; GEO: Gene Expression Omnibus; CV: cross-validation; R^2: coefficient of determination; BH: Benjamini–Hochberg; HPC: high-performance computing.

Declarations

Ethics approval and consent to participate

Not applicable. This study reanalysed publicly available, de-identified datasets and did not involve new experiments with human participants, human tissue, or animals.

Consent for publication

Not applicable.

Availability of data and materials

The public datasets analysed in this study are available from NCBI GEO under accessions GSE52527, GSE52530, GSE73057, and GSE50452. Genome and transcript annotations were obtained from the UCSC Genome Browser, mature miRNA sequences from miRBase, and reporter-assay evidence from a dated miRTarBase portal snapshot. Processed matrices, complete rank tables, the locked-holdout screening manifest, benchmark code, exact commands, environment files, checksums, and the sanitized release-verification record are archived with APAmiRank v0.2.18 in Zenodo [https://doi.org/10.5281/zenodo.23079713]. The development repository is available at https://github.com/deren0demirel-a11y/APAmiRank.

Software availability

Project name: APAmiRank. Version analysed: v0.2.18. Project home: https://github.com/deren0demirel-a11y/APAmiRank. Archived version: [https://doi.org/10.5281/zenodo.23079713]. Verified release tag: v0.2.18. Verified commit: 87defcb3970686d604746d6f4cfec92e93506c8b. Operating system: Linux tested; other platforms were not formally evaluated. Programming language: Python ≥3.10. Core requirements: PyYAML ≥6.0 and pyfaidx ≥0.8. Optional workflow support uses Snakemake. Optional thermodynamic mode requires

RNAhybrid v2.1.2 and RNAup from ViennaRNA v2.7.2. License: BSD 3-Clause. Restrictions on use by non-academics: none.

Competing interests

The authors declare no competing interests.

Funding

The authors received no specific funding for this work.

Authors' contributions

DDY: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review and editing, and Project administration.

IDS: Methodology, Investigation, Data curation, Writing – original draft, Writing – review and editing, and Project administration. All authors read and approved of the final manuscript.

Acknowledgements

Not applicable.

Additional files

Additional file 1: APAmiRank Supplementary Methods S1–S8 (.docx). Detailed provenance, benchmark procedures, computational settings, and reproducible execution information.

Additional file 2: APAmiRank Supplementary Tables S1–S8 (.xlsx). Public-resource provenance, benchmark-level metrics, pair-level holdout screening, runtime measurements, and the exploratory miRDB comparison.

Additional file 3: APAmiRank Supplementary Figures S1–S2 (.pptx). Architecture-mode runtime scaling and the exploratory method-native comparison with miRDB v6.0.

References

1. Bartel DP. Metazoan microRNAs. Cell. 2018;173:20-51. https://doi.org/10.1016/j.cell.2018.03.006.

2. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15-20. https://doi.org/10.1016/j.cell.2004.12.035.

3. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell. 2007;27:91-105. https://doi.org/10.1016/j.molcel.2007.06.017.

4. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015;4:e05005. https://doi.org/10.7554/eLife.05005.

5. McGeary SE, Lin KS, Shi CY, et al. The biochemical basis of microRNA targeting efficacy. Science. 2019;366:eaav1741. https://doi.org/10.1126/science.aav1741.

6. Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, et al. microRNA target prediction programs predict many false positives. Genome Res. 2017;27:234-245. https://doi.org/10.1101/gr.205146.116.

7. Kern F, Krammes L, Danz K, Diener C, Kehl T, Küchler O, et al. Validation of human microRNA target pathways enables evaluation of target prediction tools. Nucleic Acids Res. 2021;49:127-144. https://doi.org/10.1093/nar/gkaa1161.

8. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol. 2017;18:18-30. https://doi.org/10.1038/nrm.2016.116.

9. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science. 2008;320:1643-1647. https://doi.org/10.1126/science.1155390.

10. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009;138:673-684. https://doi.org/10.1016/j.cell.2009.06.016.

11. Derti A, Garrett-Engele P, MacIsaac KD, et al. A quantitative atlas of polyadenylation in five mammals. Genome Res. 2012;22:1173-1183. https://doi.org/10.1101/gr.132563.111.

12. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 2013;27:2380-2396. https://doi.org/10.1101/gad.229328.113.

13. Nam JW, Rissland OS, Koppstein D, et al. Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell. 2014;53:1031-1043. https://doi.org/10.1016/j.molcel.2014.02.013.

14. Hoque M, Ji Z, Zheng D, et al. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat Methods. 2013;10:133-139. https://doi.org/10.1038/nmeth.2288.

15. Ha KCH, Blencowe BJ, Morris Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. Genome Biol. 2018;19:45. https://doi.org/10.1186/s13059-018-1414-4.

16. Xia Z, Donehower LA, Cooper TA, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nat Commun. 2014;5:5274. https://doi.org/10.1038/ncomms6274.

17. Bryce-Smith S, Burri D, Gazzara MR, et al. Extensible benchmarking of methods that identify and quantify polyadenylation sites from RNA-seq data. RNA. 2023;29:1839-1855. https://doi.org/10.1261/rna.079849.123.

18. Müller S, Rycak L, Afonso-Grunz F, Winter P, Zawada AM, Damrath E, et al. APADB: a database for alternative polyadenylation and microRNA regulation events. Database (Oxford). 2014;2014:bau076. https://doi.org/10.1093/database/bau076.

19. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013;153:654-665. https://doi.org/10.1016/j.cell.2013.03.043.

20. Moore MJ, Scheel TKH, Luna JM, et al. miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. Nat Commun. 2015;6:8864. https://doi.org/10.1038/ncomms9864.

21. Cui S, Yu S, Huang HY, et al. miRTarBase 2025: updates to the collection of experimentally validated microRNA-target interactions. Nucleic Acids Res. 2025;53:D147-D156. https://doi.org/10.1093/nar/gkae1072.

22. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47:D155-D162. https://doi.org/10.1093/nar/gky1141.

23. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004;10:1507-1517. https://doi.org/10.1261/rna.5248604.

24. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. Bioinformatics. 2006;22:1177-1182. https://doi.org/10.1093/bioinformatics/btl024.

25. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.

26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets - update. Nucleic Acids Res. 2013;41:D991-D995. https://doi.org/10.1093/nar/gks1193.

27. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996-1006. https://doi.org/10.1101/gr.229102.

28. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020;48(D1):D127-D131. https://doi.org/10.1093/nar/gkz757.

29. miRDB. Data download and version information. https://mirdb.org/download.html. Accessed 16 Sep 2026.

30. Bradley T, Moxon S. FilTar: using RNA-Seq data to improve microRNA target prediction accuracy in animals. Bioinformatics. 2020;36:2410-2416. https://doi.org/10.1093/bioinformatics/btaa007.

31. Kim S, Bai Y, Fan Z, Diergaarde B, Tseng GC, Park HJ. The microRNA target site landscape is a novel molecular feature associating alternative polyadenylation with immune evasion activity in breast cancer. Brief Bioinform. 2021;22:bbaa191. https://doi.org/10.1093/bib/bbaa191.

32. Tastsoglou S, Alexiou A, Karagkouni D, Skoufos G, Zacharopoulou E, Hatzigeorgiou AG. DIANA-microT 2023: including predicted targets of virally encoded miRNAs. Nucleic Acids Res. 2023;51:W148-W153. https://doi.org/10.1093/nar/gkad283.

33. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39:1278-1284. https://doi.org/10.1038/ng2135.

34. Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, Baek D, et al. mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Mol Cell. 2014;56:104-115. https://doi.org/10.1016/j.molcel.2014.08.028.

Table Legends

Table S1. Public datasets and external resources. Accession, retrieval, exact-file, version, checksum, and persistent-location information for public and external inputs.

Table S2. Functional perturbation models. Primary comparisons at mean-control RPKM ≥1. Stability intervals are repetition percentiles, not confidence intervals.

Table S3. Direct-interaction metrics. APA-exposure rank performance in the 3′UTR-restricted positive–unlabelled CLEAR-CLIP and CLASH benchmarks.

Table S4. Reporter benchmark metrics. Primary orthogonally supported Reporter Assay set in the at-least-two-of-four-context consensus universe. Bootstrap results were suppressed when fewer than five positives were observable.

Table S5. Locked-holdout screening and ranks. All 33 locked candidates are shown with inclusion/exclusion status, observability, reasons for missingness, and ranks where available. No ranks were imputed.

Table S6. Runtime and resource scaling. Single deterministic synthetic run per size; 250 bootstrap and 1,000 opportunity iterations. Values are descriptive and hardware-specific.

Table S7. miRDB reporter-positive recovery. Recovery of reporter-positive genes in official miRDB v6.0 method-native result pages. Missing predictions were not treated as biological negatives.

Table S8. miRDB locked-holdout comparison. Pair-level exploratory method-native comparison for the nine targets observable by both methods, followed by a summary row; the candidate universes differ.

Figure Legends

Figure 1. APAmiRank workflow and validation boundary. Mature-miRNA sequence, ordered strand-correct 3′UTR isoforms, and optional condition-specific isoform-usage fractions are combined to enumerate canonical sites and quantify their retention. Architecture mode produces matched unweighted and APA-exposure scores and is the mode validated in this study. The optional RNAhybrid/RNAup thermodynamic extension is implemented but was not biologically validated here. Outputs include auditable site- and gene-level rankings and run manifests.

Figure 2. Functional perturbation benchmark. (A) Median repeated cross-validated R^2 for covariates only (C), covariates plus the matched unweighted score (C+U), covariates plus the exposure-weighted score (C+W), and the full model (C+U+W) across the six primary comparisons at mean-control RPKM ≥1. **(B)** Median repeated-cross-validation increment from adding the exposure-weighted score to C+U; horizontal intervals show the 2.5th–97.5th percentile range across 100 deterministic repetitions and are stability intervals rather than confidence intervals.

Figure 3. Direct-interaction benchmarks. (A) Positive–unlabeled AUC for APA-exposure rankings in the Huh-7.5 CLEAR-CLIP and HEK293 CLASH 3′UTR benchmarks. The dashed line indicates random ranking (PU-AUC=0.5). **(B)** Fold enrichment of observed direct-interaction genes in the top 10% of each miRNA-specific canonical-site universe. The dashed line indicates no enrichment. CLEAR-CLIP supported above-background prioritization for four of five miRNAs after multiple-testing correction, whereas the corresponding CLASH rank-distribution tests did not pass $q<0.05$. Absence from either interaction map was treated as unlabeled rather than as a confirmed negative.

Figure 4. Curated reporter-assay benchmark. (A) Positive–unlabeled AUC for APA-exposure consensus rankings. Labels indicate the number of observable orthogonally supported targets; miR-186-5p was retained descriptively but not tested inferentially because $n<5$. **(B)** Fold enrichment of supported targets in the top 10% of each miRNA-specific canonical-site universe; the dashed line indicates no enrichment. **(C)** Paired difference

between APA-exposure and unweighted-additive PU-AUC on identical genes. Points indicate observed differences and error bars indicate 95% paired stratified bootstrap intervals from 5,000 resamples. The horizontal dashed line denotes no difference. Positive increments were supported after Benjamini–Hochberg adjustment for miR-16-5p, miR-25-3p, and miR-124-3p, but not uniformly across the panel.

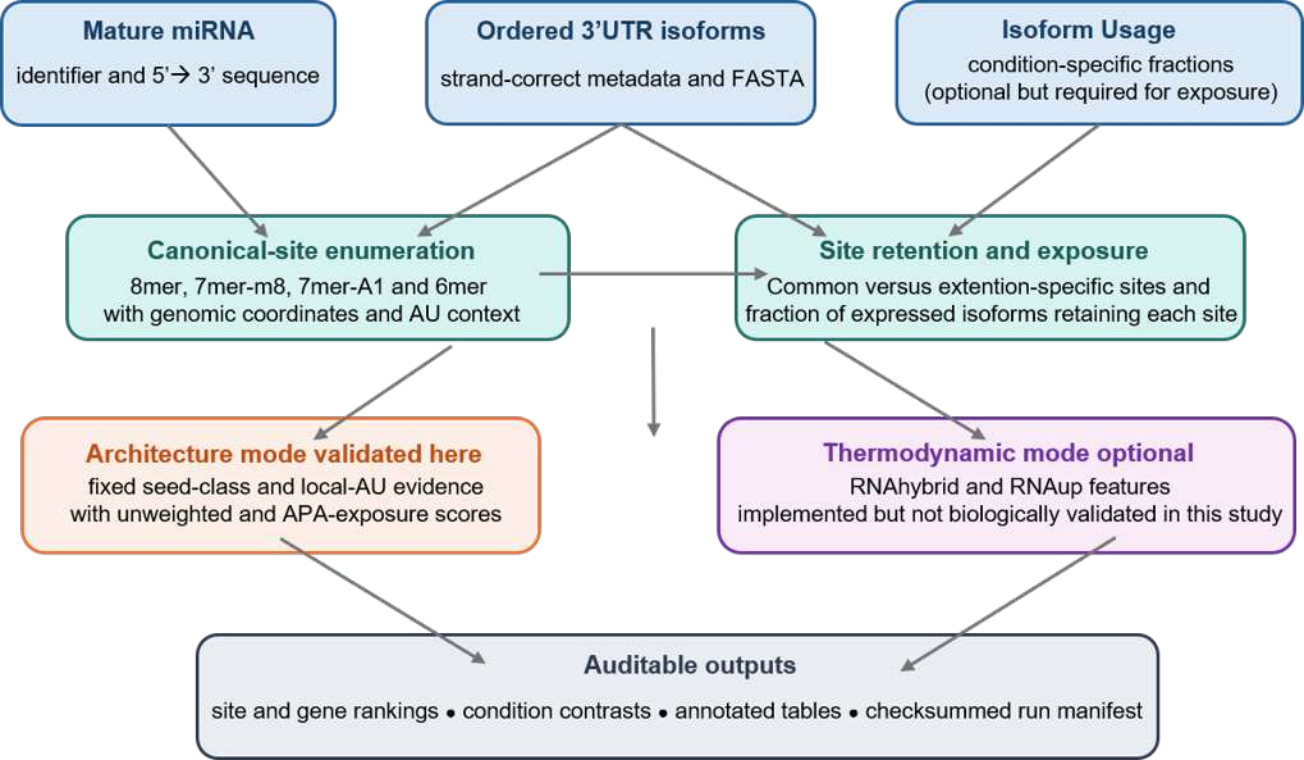
Figure 5. Rank-independent locked luciferase holdout and exploratory uncertainty analysis. (A) APA-exposure percentile ranks of the 12 observable strict inclusions. Reference lines indicate the 50th, 80th, and 90th percentiles. **(B)** Null distribution of the pooled mean percentile from 50,000 theoretical miRNA-matched rank-uniform simulations. The orange line marks the observed mean of 0.7847 and the dashed line marks the null mean; the Holm-adjusted exploratory P value was 0.00080. **(C)** Paired comparison of APA-exposure and unweighted-additive percentiles for observable pairs. The diagonal denotes equal ranks; the median paired difference was +0.0651. **(D)** Screening and rank coverage. Of 33 rank-free candidates, 21 met the frozen human 3′UTR wild-type/cognate-mutant reporter criteria and 12 were observable in the consensus rank table. Protocol or provenance exclusions were not biological negatives, and unobservable strict inclusions were not imputed. All rank summaries are conditional on 57.1% observability; panel B and the associated uncertainty analyses were designed after unblinding and are exploratory.

Figure S1. Architecture-mode computational scaling.

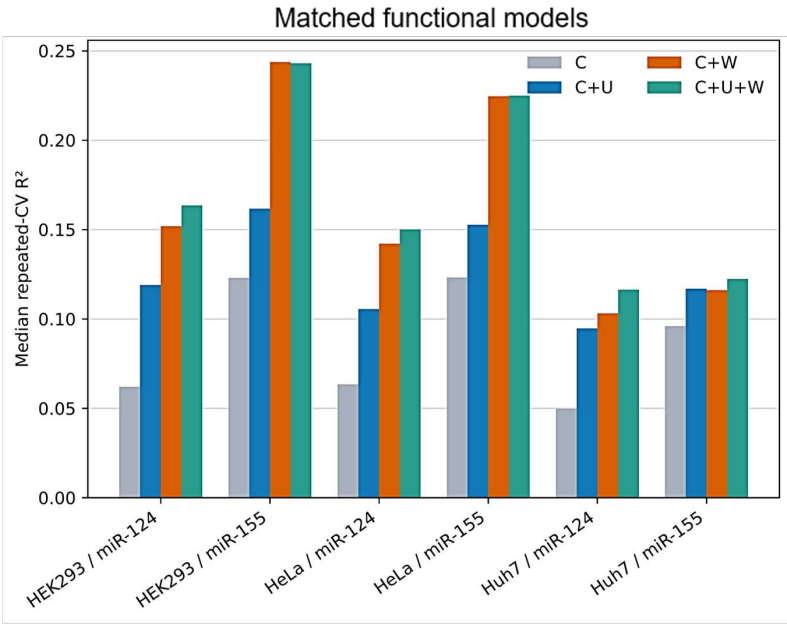
Wall time increased from 0.162 s for 100 genes to 5.443 s for 5,000 genes. Recorded maximum resident set size was 230.9 MB at all three sizes, consistent with fixed interpreter/package overhead dominating memory use over the tested range. One deterministic run was performed per size; values are descriptive and hardware-specific.

Figure S2. Exploratory comparison with miRDB v6.0.

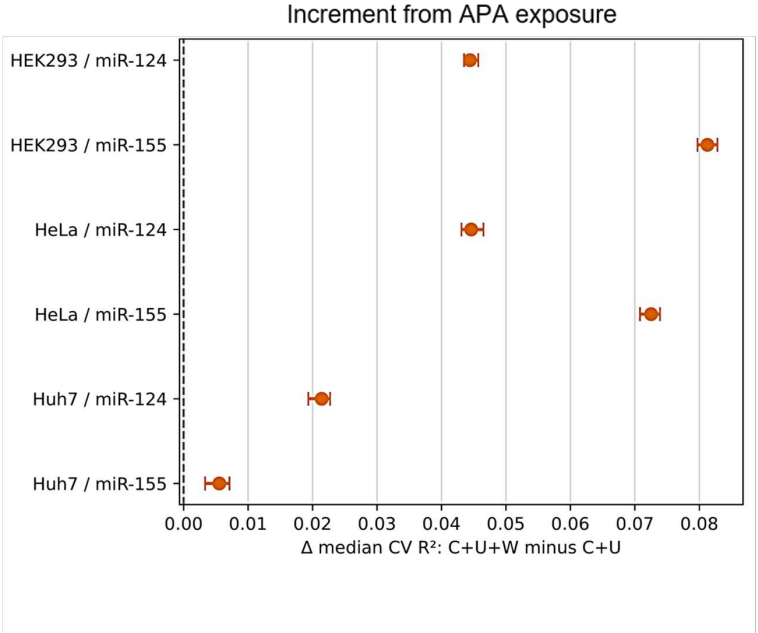
(A) Coverage of the 21 strict locked-holdout inclusions in the APAmiRank consensus universe, official miRDB result pages, and both. **(B)** Method-native percentiles for the nine pairs observable by both methods. Lines join the same target. The comparison does not use a harmonized candidate universe and does not establish superiority.


Mature miRNA
identifier and 5'→ 3' sequence
Ordered 3'UTR isoforms
strand-correct metadata and FASTA
Isoform Usage
condition-specific fractions
(optional but required for exposure)
Canonical-site enumeration
8mer, 7mer-m8, 7mer-A1 and 6mer
with genomic coordinates and AU context
Site retention and exposure
Common versus extention-specific sites and
fraction of expressed isoforms retaining each site
Architecture mode validated here
fixed seed-class and local-AU evidence
with unweighted and APA-exposure scores
Thermodynamic mode optional
RNAhybrid and RNAup features
implemented but not biologically validated in this study
Auditable outputs
site and gene rankings • condition contrasts • annotated tables • checksummed run manifest

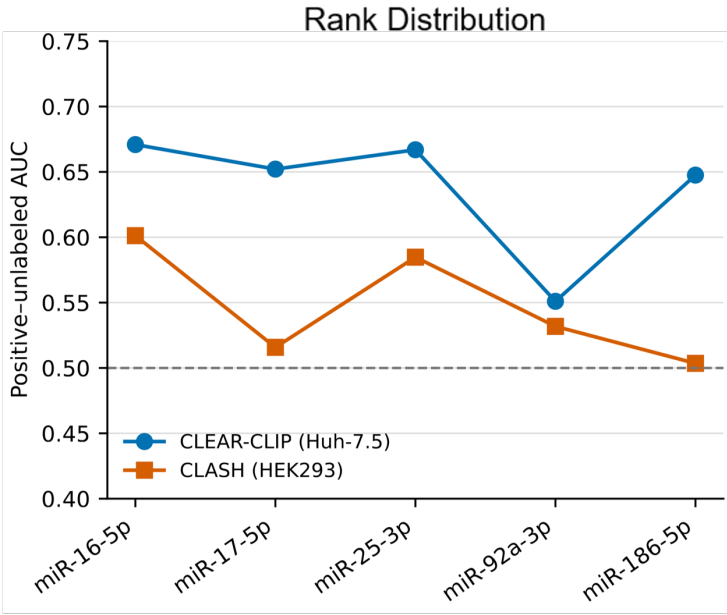
A


Matched functional models
C
C+W
C+U
C+U+W
Median repeated-CV R²
0.25
0.20
0.15
0.10
0.05
0.00
HEK293 / miR-124
HEK293 / miR-155
HeLa / miR-124
HeLa / miR-155
Huh7 / miR-124
Huh7 / miR-155

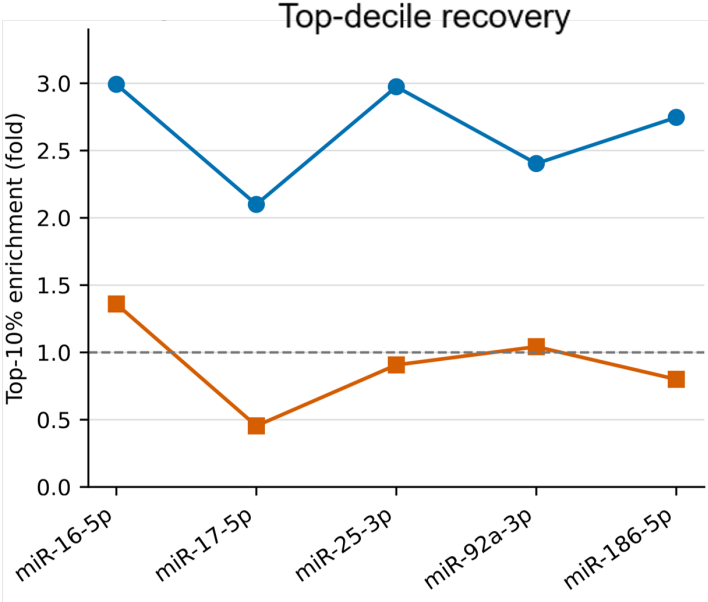
B


Increment from APA exposure
HEK293 / miR-124
HEK293 / miR-155
HeLa / miR-124
HeLa / miR-155
Huh7 / miR-124
Huh7 / miR-155
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Δ median CV R²: C+U+W minus C+U

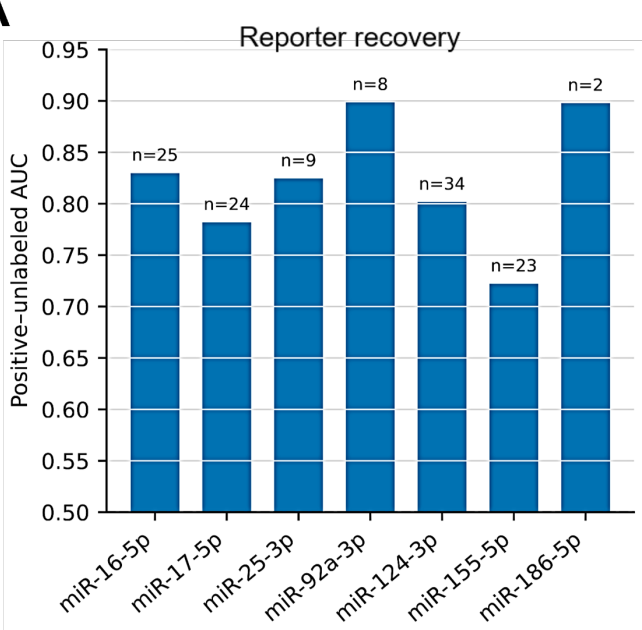
A


Rank Distribution
Positive-unlabeled AUC
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
CLEAR-CLIP (Huh-7.5)
CLASH (HEK293)
miR-16-5p
miR-17-5p
miR-25-3p
miR-92a-3p
miR-186-5p

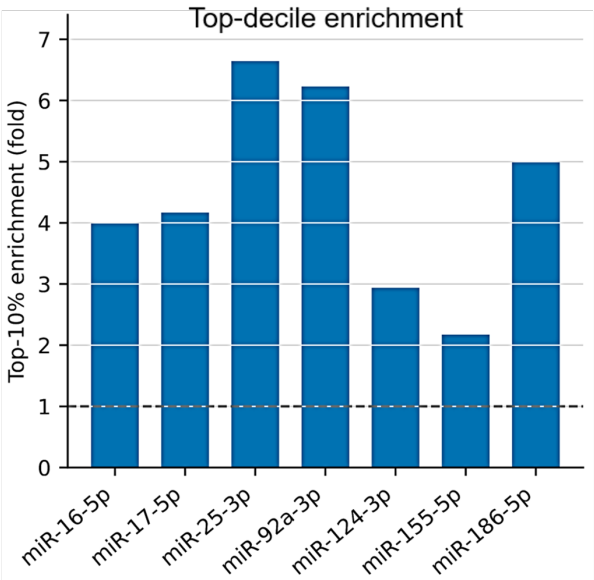
B


Top-decile recovery
Top-10% enrichment (fold)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
miR-16-5p
miR-17-5p
miR-25-3p
miR-92a-3p
miR-186-5p

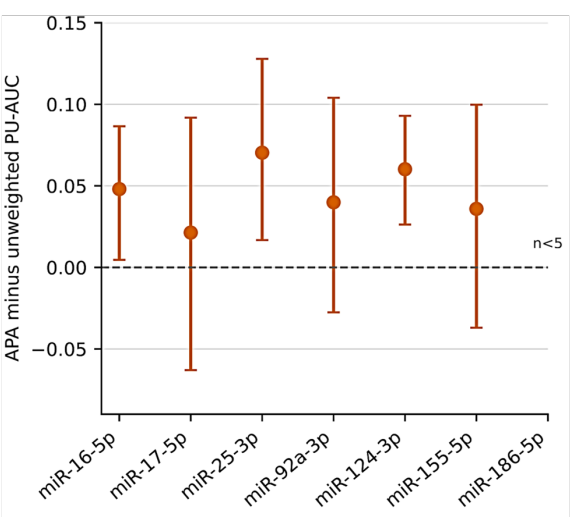
A

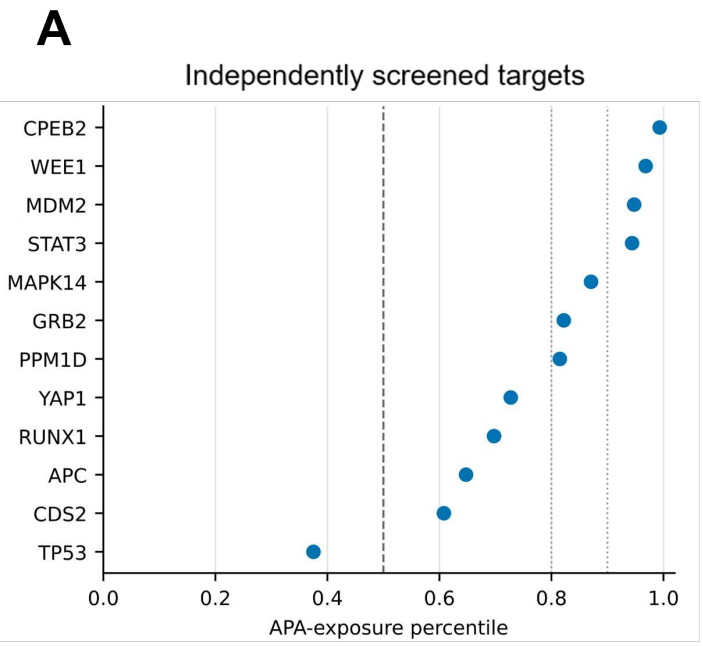

Reporter recovery
Positive-unlabeled AUC
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
n=25
n=24
n=9
n=8
n=34
n=23
n=2
miR-16-5p
miR-17-5p
miR-25-3p
miR-92a-3p
miR-124-3p
miR-155-5p
miR-186-5p

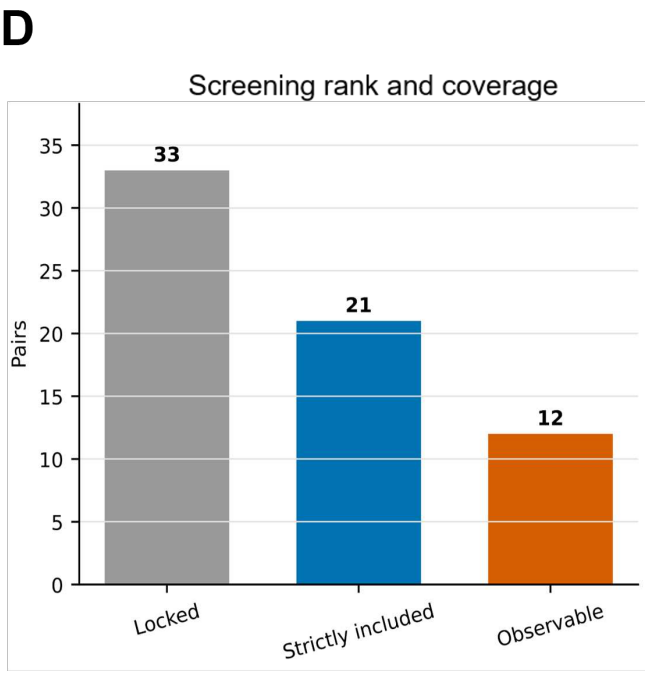
B

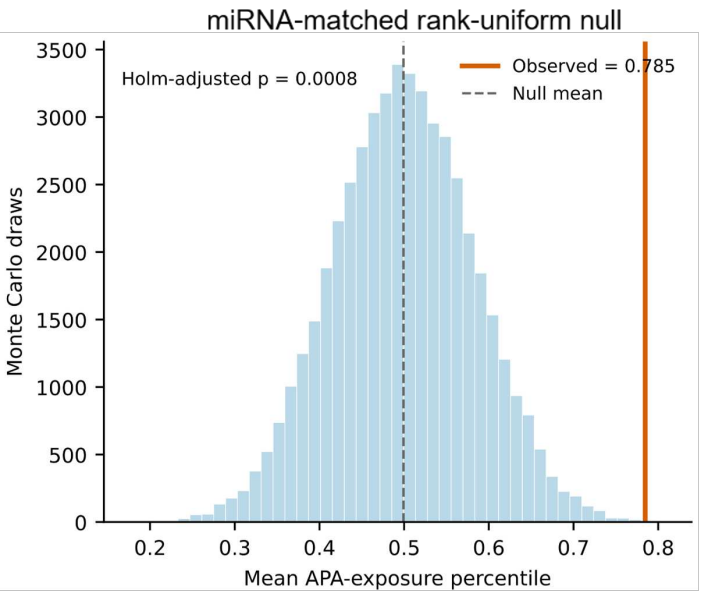

Top-decile enrichment
Top-10% enrichment (fold)
7
6
5
4
3
2
1
0
miR-16-5p
miR-17-5p
miR-25-3p
miR-92a-3p
miR-124-3p
miR-155-5p
miR-186-5p

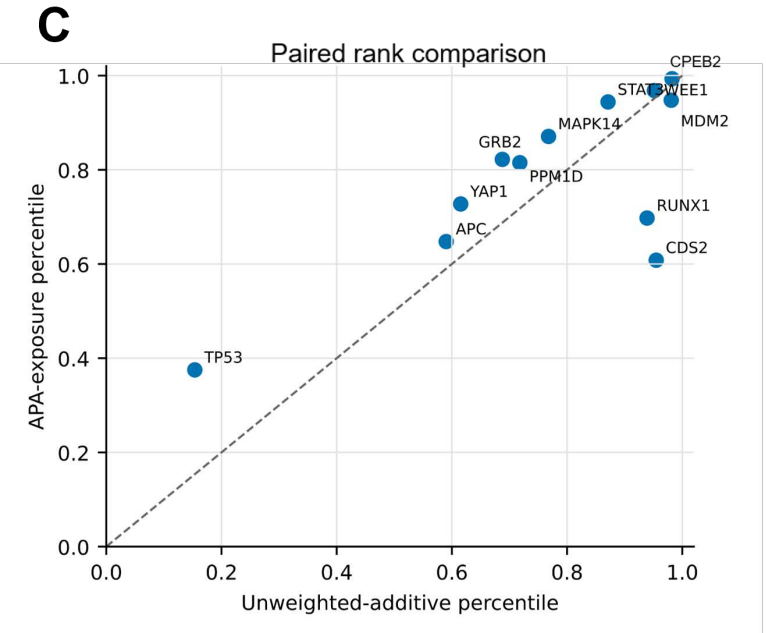
C


Increment from APA exposure
APA minus unweighted PU-AUC
0.15
0.10
0.05
0.00
−0.05
n<5
miR-16-5p
miR-17-5p
miR-25-3p
miR-92a-3p
miR-124-3p
miR-155-5p
miR-186-5p


A
Independently screened targets
CPEB2
WEE1
MDM2
STAT3
MAPK14
GRB2
PPM1D
YAP1
RUNX1
APC
CDS2
TP53
0.0
0.2
0.4
0.6
0.8
1.0
APA-exposure percentile


D
Screening rank and coverage
Pairs
33
21
12
Locked
Strictly included
Observable


B
miRNA-matched rank-uniform null
Holm-adjusted p = 0.0008
Observed = 0.785
Null mean
Monte Carlo draws
Mean APA-exposure percentile


C
Paired rank comparison
CPEB2
STAT3
WEE1
MDM2
MAPK14
GRB2
PPM1D
YAP1
RUNX1
APC
CDS2
TP53
APA-exposure percentile
Unweighted-additive percentile

Supplementary Files

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

- APAmiRankSupplementaryTablesS1S8revised.xlsx
- APAmiRankSupplementaryMethodsS1S8revised.docx
- APAmiRankSupplementaryFiguresS1S2BMCBioinformatics.pptx