Systems-level control of microRNA regulation through transcriptomic plasticity

Dissertation
Zur Erlangung des Grades
Doktor der Naturwissenschaften

Am Fachbereich Biologie
Der Johannes Gutenberg-Universität Mainz

Mert Cihan

Geb. am 31.03.1994 in Bergisch Gladbach, Deutschland

Mainz, 2026

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 14.04.2026

Mert Cihan
Systems-level control of microRNA regulation through transcriptomic plasticity
Dissertation, February 2026
Reviewers:
Supervision:

Johannes Gutenberg University Mainz
AG – Computational Biology and Data Mining
Institute of Organismic and Molecular Evolution (iOME)
Faculty of Biology
Hanns-Dieter-Hüsch-Weg 15
55128 Mainz

Zusammenfassung

MicroRNAs (miRNAs) sind zentrale posttranskriptionelle Regulatoren der Genexpression, die eng in komplexe regulatorische Netzwerke eingebunden sind. Durch die koordinierte Regulation zahlreicher Zieltranskripte beeinflussen sie die Aktivität dieser Netzwerke und tragen zur Kontrolle zellulärer Identität bei. Trotz erheblicher Fortschritte durch Hochdurchsatz-Sequenzierung bleibt die rechnergestützte Analyse miRNA-vermittelter Regulation aufgrund ihrer hohen Komplexität und Kontextabhängigkeit anspruchsvoll.

Diese Dissertation widmet sich zentralen computergestützten Fragestellungen der miRNA-Forschung auf Basis transkriptomischer und genomischer Daten. Ein zentraler Ansatz dieser Arbeit ist die Analyse von miRNA-Funktionen im regulatorischen und strukturellen Kontext des Transkriptoms. Insbesondere die durch alternative Polyadenylierung (APA) vermittelte transkriptomische Plastizität verändert miRNA–Gen-Interaktionen grundlegend und stellt klassische Analyseansätze erheblich vor Herausforderungen.

Ein Schwerpunkt dieser Dissertation liegt auf der Verbesserung der Vorhersage funktioneller Zielgene. Durch die Integration vorhergesagter miRNA–mRNA-Interaktionen mit transkriptionsfaktorbasierten regulatorischen Netzwerken werden charakteristische Merkmale funktionell relevanter miRNA-Bindestellen identifiziert. Darauf aufbauend wird in einem netzwerkorientierten Ansatz untersucht, wie transkriptomische Plastizität die miRNA-vermittelte Regulation in Glioblastomen beeinflusst. Einzelzellanalysen zeigen, dass zelltypspezifische APA-Profile die Verfügbarkeit von miRNA-Bindestellen steuern und miRNA–Transkriptionsfaktor-Regulationsnetzwerke abhängig von Zellzustand und Tumormikromilieu reorganisieren. Eine weitere methodische Herausforderung ist die begrenzte Verfügbarkeit von miRNA-Sequenzierungsdaten in transkriptomischen Datensätzen. Mithilfe maschinellen Lernens wird gezeigt, dass sich miRNA-Expressionsmuster teilweise aus mRNA-Profilen ableiten lassen. Zudem deuten die Analysen darauf hin, dass vermeintliche Unterschiede in der miRNA-Expression häufig auf Veränderungen der Zieltranskripte oder der miRNA-Abbaudynamik zurückzuführen sind. Ergänzend werden genomische Variationen in der menschlichen Bevölkerung analysiert, um Zusammenhänge zwischen Sequenzkonservierung, Selektionsdruck und funktioneller Bedeutung von miRNAs zu untersuchen.

Zusammenfassend adressiert diese Dissertation zentrale Herausforderungen der miRNA-Forschung durch die methodische Integration transkriptomischer Plastizität. Dabei werden Annotation, Vorhersage der miRNA-Expression, Identifikation funktionell relevanter Zielmerkmale sowie die Charakterisierung der miRNA-Abbaudynamik und der durch alternative Polyadenylierung modulierten miRNA-Aktivität untersucht. Damit wird ein integrierter, kontextsensitiver und rechnergestützter Rahmen zur Analyse von miRNA-Funktionen in komplexen biologischen Systemen etabliert.

Abstract

MicroRNAs (miRNAs) are post-transcriptional regulators of gene expression and are deeply embedded within gene regulatory networks that coordinate cellular identity. By targeting large sets of transcripts and interacting with other regulatory layers, miRNAs contribute to adaptable control of gene expression programs. While high-throughput sequencing has enabled comprehensive profiling of miRNAs across biological contexts, the complexity and context dependence of miRNA-mediated regulation pose major challenges for computational analysis.

This dissertation addresses core computational challenges in miRNA research, ranging from miRNA annotation and target identification to the inference and interpretation of miRNA regulatory activity from transcriptomic and genomic data. A central premise of this work is that miRNA function must be analyzed within the regulatory and structural context of the transcriptome, as transcriptomic plasticity driven by alternative polyadenylation (APA) reshapes miRNA–target interactions and complicates conventional analytical approaches.

To address the high false-positive rates of miRNA target prediction, predicted miRNA–mRNA interactions are integrated with transcription factor regulatory networks. This network-based framework reveals sequence composition and positional binding-site features that distinguish functional targets and highlights the importance of regulatory context in miRNA targeting. Extending this network-based perspective to transcriptomic plasticity, single-cell analyses of glioblastoma show that cell-type-specific APA dynamically reshapes miRNA binding site availability and rewires miRNA–transcription factor co-regulatory networks in association with cellular states in the tumor microenvironment. To overcome the limited availability of small RNA sequencing data, interpretable machine learning models are applied to infer miRNA expression from matched mRNA profiles, enabling reconstruction of miRNA regulatory activity from widely available transcriptomic datasets. Consideration of transcript dynamics further reveals that apparent differential miRNA expression between conditions can arise from target-site dynamics rather than true changes in miRNA abundance, as miRNA decay mechanisms and APA-driven gain or loss of binding sites confound standard differential expression analyses. Complementing these transcriptomic findings, large-scale analyses of human genomic variation link sequence conservation and population-level constraint to miRNA functional relevance and annotation reliability.

Together, this work addresses key challenges in miRNA annotation, expression prediction, identification of targeting features, decay dynamics, and regulation of miRNA activity through alternative polyadenylation. By integrating regulatory network analysis, machine learning, modeling of transcriptomic dynamics, and population genetics, this dissertation provides an integrated, context-aware computational framework for studying miRNA function across complex biological systems.

Table of Contents

1 Introduction 1
1.1 Scope of this thesis 1
1.2 MicroRNA biology and gene regulation 2
1.2.1 MicroRNA biogenesis 2
1.2.2 Mechanisms of microRNA action 3
1.2.3 MicroRNA turnover and decay 5
1.3 Evolutionary and genetic architecture of microRNA regulation 6
1.3.1 Conservation of microRNAs 6
1.3.2 Genetic variation in microRNAs and targets 7
1.3.3 MicroRNA variation in human disease 8
1.4 MicroRNAs in cancer 8
1.4.1 The Cancer Genome Atlas (TCGA) 8
1.4.2 OncomiRs and tumor suppressors 9
1.4.3 Glioblastoma as a model of regulatory complexity 11
1.5 Alternative polyadenylation and microRNA-mediated regulation 12
1.5.1 Molecular mechanisms of APA 12
1.5.2 APA across cell types and states 15
1.6 Bioinformatics approaches to microRNA research 16
1.6.1 MicroRNA annotation 16
1.6.2 Target identification 17
1.6.3 MicroRNA networks and function 18
1.7 Machine learning in regulatory biology 19
1.7.1 Machine learning for gene regulatory network inference 19
1.7.2 Machine learning in microRNA research 20
1.7.3 Regularization methods 21
2 Research articles 23
2.1 Detection of features predictive of microRNA targets by integration of network data 23
2.1.1 Preamble 23
2.1.2 Abstract 23
2.1.3 Introduction 23
2.1.4 Results 25
2.1.5 Discussion 30
2.1.6 Materials and methods 33

2.2 Unveiling cell-type-specific microRNA networks through alternative polyadenylation in glioblastoma....36
2.2.1 Preamble....36
2.2.2 Abstract....36
2.2.3 Background....36
2.2.4 Results....38
2.2.5 Discussion....48
2.2.6 Conclusions....51
2.2.7 Methods....51
2.3 Evaluating Genetic Regulators of MicroRNAs Using Machine Learning Models....55
2.3.1 Preamble....55
2.3.2 Abstract....55
2.3.3 Introduction....55
2.3.4 Results....57
2.3.5 Discussion....64
2.3.6 Methods....67
2.3.7 Conclusions....69
2.4 Target-site Dynamics and Alternative Polyadenylation Explain a Large Share of Apparent MicroRNA Differential Expression....70
2.4.1 Preamble....70
2.4.2 Abstract....70
2.4.3 Introduction....70
2.4.4 Methods....72
2.4.5 Results....74
2.4.6 Discussion....84
2.5 Genomic variation of human microRNAs and its association with functional features....87
2.5.1 Preamble....87
2.5.2 Abstract....87
2.5.3 Introduction....87
2.5.4 Methods....88
2.5.5 Results....92
2.5.6 Discussion....102
2.6 Applied motif analysis in collaborations....105
3 General discussion....106
3.1 Systems-level integration of microRNA regulation....106

3.2 Functional and disease implications of regulatory dynamics.....107
3.3 Computational advances in microRNA analysis.....109
4 Conclusion and outlook.....110
5 Acknowledgements.....112
6 Supplementary Information.....113
7 Bibliography.....114

1 Introduction

1.1 Scope of this thesis

MicroRNA (miRNA) research has greatly benefited from bioinformatics approaches that enable systematic study of regulatory interactions and miRNA function. Despite extensive knowledge of miRNA biogenesis and targeting rules, the functional interpretation of miRNA regulation remains limited by transcriptomic plasticity, incomplete data modalities, and the lack of systems-level integration. This thesis addresses these limitations by treating miRNA regulation not as a static interaction problem but as a dynamic property of regulatory networks embedded in a changing transcriptome.

The first study (Chapter 2.1) tackles the fundamental challenge of false-positive miRNA target prediction by leveraging transcription factor (TF)–miRNA overlap to learn properties that characterize functionally relevant target genes. It addresses the limitation of sequence-based prediction methods by examining how 3’ untranslated region (3’UTR) structure, sequence composition, and binding site abundance define features of reliable miRNA targets.

Building on the interplay of regulatory layers, the second study (Chapter 2.2) extends the analysis to co-regulatory miRNA–TF networks in single-cell sequencing data. It addresses the challenge of missing miRNA data at single-cell resolution by inferring regulatory activity through co-regulatory network expression and cell-type-specific alternative polyadenylation (APA) patterns that dynamically alter miRNA binding site presence. This analysis is performed in the biological context of glioblastoma progression and the dynamics of the tumor microenvironment.

Remaining within the cancer framework, the third study (Chapter 2.3) addresses the limitation that transcriptomic studies typically offer extensive mRNA sequencing data, whereas small RNA sequencing is less consistently available. It investigates whether miRNA expression can be predicted from matched mRNA expression profiles in large-scale cancer cohorts, enabling reconstruction of miRNA activity from widely available transcriptomic data.

The fourth study (Chapter 2.4) extends this perspective by examining how target site dynamics influence the interpretation of differential miRNA expression. It identifies apparent expression changes between conditions and links them, for a subset of miRNAs, to structural remodeling of 3′UTRs that alters miRNA binding site availability rather than to changes in miRNA expression itself.

Using large-scale genomic variation data, the fifth study (Chapter 2.5) addresses the problem of potential misannotation of miRNAs by investigating how sequence variation across human populations relates to functional relevance and annotation accuracy.

The last chapter (Chapter 2.6) summarizes contributions to collaborative research, building on computational approaches introduced in Chapter 2.2 to study TF binding site annotation in close spatial proximity within the context of T cell differentiation.

In summary, this thesis addresses multiple core challenges in miRNA research through systematic analysis of transcriptomic plasticity: annotation of miRNA genes and targets, prediction of miRNA expression from transcriptomic data, interpretation of differential expression in the presence of APA-driven structural changes, and understanding miRNA regulation in the highly complex single-cell cancer environment.

1.2 MicroRNA biology and gene regulation

MicroRNAs (miRNAs) are short (~22 nucleotide) non-coding RNAs that regulate gene expression post-transcriptionally, primarily through binding to complementary sequences within target mRNAs [1]. The first microRNA, *lin-4*, was discovered in *Caenorhabditis elegans* in 1993 by Victor Ambros and Gary Ruvkun, who showed that it represses *lin-14* by binding to complementary sites in its 3′-UTR, thereby reducing LIN-14 protein production [2,3]. This discovery revealed a previously unrecognized layer of gene regulation and was later recognized by the awarding of the Nobel Prize in Physiology or Medicine to Ambros and Ruvkun in 2024 [4].

Since their discovery, microRNAs have become a central research topic in molecular biology, with extensive efforts directed toward understanding their biogenesis, mechanisms of action, and decay pathways [5].

1.2.1 MicroRNA biogenesis

The biogenesis of miRNAs begins with the transcription of miRNA genes by RNA polymerase II (RNA Pol II) to produce long primary miRNA (pri-miRNA) transcripts in the nucleus. These miRNA genes can be stand-alone transcriptional units, located within introns of protein-coding genes, or organized in polycistronic clusters that are transcribed together as a single primary transcript. pri-miRNA transcripts generated by RNA pol II are typically capped and polyadenylated and contain the sequence information for one or more miRNAs [1,5,6]. Within the human genome, there are approximately 1,900 annotated microRNA genes, though the exact number depends on database and annotation criteria [7,8].

Characteristically, the pri-miRNA folds into a hairpin consisting of a stem and a loop structure in which complementary sequences form a double-stranded stem with a terminal loop and are flanked by single-stranded 5′ and 3′ regions of variable length within the longer primary transcript. Within the stem, imperfect base pairing is common, including G–U wobble pairs and other mismatches [5,9].

In the canonical miRNA biogenesis pathway, the Microprocessor complex, composed of the RNase III enzyme Drosha and its RNA-binding partner DGCR8, binds the hairpin structure of the pri-miRNA and performs endonucleolytic cleavage of the double-stranded stem to generate a precursor-miRNA (pre-miRNA) with a characteristic two-nucleotide 3′ overhang [10,11].

After cleavage by the Microprocessor complex, the pre-miRNA is exported from the nucleus to the cytoplasm by Exportin-5 in a Ran-GTP–dependent manner. Exportin-5 specifically recognizes the structural features generated by Drosha processing, including the short 3′ overhang, thereby facilitating efficient nuclear export and protecting the pre-miRNA from degradation during transport [12].

In the cytoplasm, the pre-miRNA undergoes further processing by the RNase III endonuclease Dicer. In association with proteins, such as TRBP, Dicer binds and cleaves the stem–loop structure near the terminal loop, producing a short double-stranded RNA duplex of approximately 21–23 nucleotides [5,13]. This miRNA duplex consists of two partially complementary strands and retains characteristic two-nucleotide 3′ overhangs that are typical of RNase III–mediated cleavage.

Subsequently, the miRNA duplex is incorporated into an Argonaute (AGO) protein, forming the RNA-induced silencing complex (RISC). During this process, one strand of the duplex is preferentially stabilized as the guide strand, whereas the opposing passenger strand is eliminated. Strand selection is influenced by the relative thermodynamic stability of the duplex ends and intrinsic sequence properties of the miRNA (Figure 1.2.1) [1,5].

In contrast to the canonical pathway, non-canonical miRNA biogenesis routes exploit alternative precursor RNAs and distinct processing modules, including Drosha/DGCR8-independent pathways such as in mirtrons, which use splicing to generate pre-miRNA-like intron hairpins that are then exported and processed by Dicer, and Dicer-independent pathways directly cleaved by AGO2 [5,14,15].

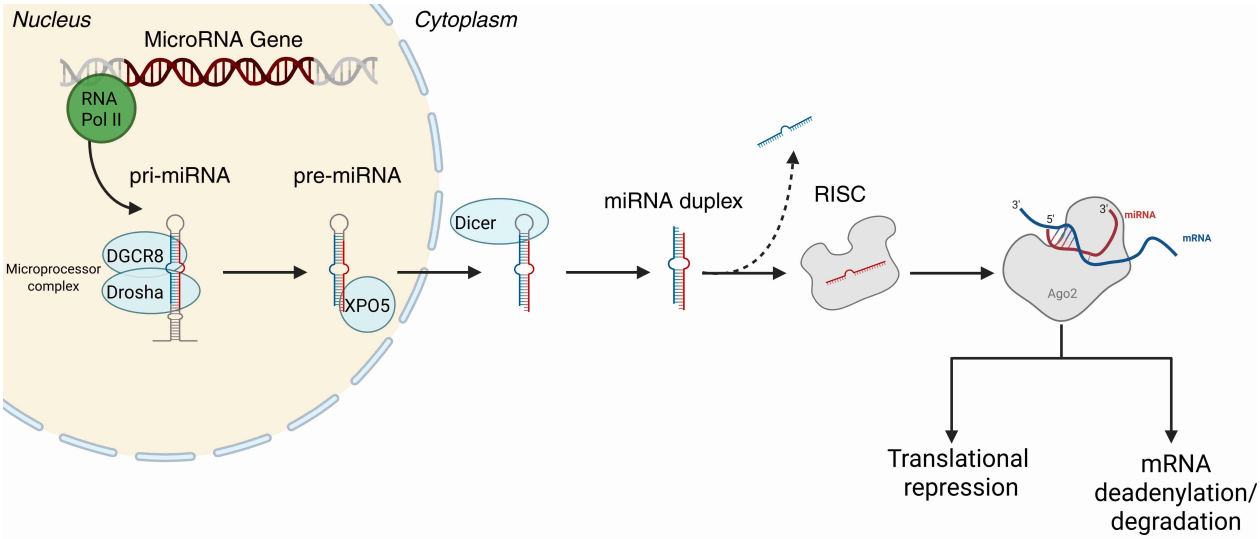


Figure 1.2.1: The canonical miRNA pathway. miRNA genes are transcribed by RNA Polymerase II (RNA Pol II) to produce pri-miRNA transcripts, which are processed by the Drosha–DGCR8 complex into pre-miRNAs in the nucleus. These precursors are exported by Exportin-5 (XPO5) to the cytoplasm, where Dicer generates miRNA duplexes, and the guide strand is subsequently loaded into the Argonaute (AGO)-containing RISC to repress or degrade target mRNAs. Created in BioRender.

1.2.2 Mechanisms of microRNA action

Mechanistically, miRNAs regulate gene expression by targeting complementary sequences within the 3′ untranslated region (3’UTR) of mRNAs. Nucleotides 2–8 from the 5′ end of the miRNA define the seed region, which is essential for target recognition, and AGO proteins guide the miRNA to its mRNA target [1,16]. Target binding begins with complementary pairing of nucleotides 2–5, followed by a conformational movement of the α7 helix of AGO that enables base pairing beyond the fifth nucleotide [5,17]. Mismatches or wobble base pairs can disrupt this helix

movement, leading to the release of the target RNA. When complementarity is extensive, direct AGO-mediated endonucleolytic cleavage can occur, although in humans this mechanism is rare and limited to only a few specific transcripts. [5,18,19].

The predominant outcome of miRNA targeting is mRNA deadenylation followed by degradation. The protein GW182, also known as TNRC6 in humans, functions as a central effector of miRNA-mediated silencing by interacting with AGO within the RISC and through interacting with the poly(A)-binding protein of the target mRNA. Through these interactions, GW182 recruits the CCR4–NOT and PAN2–PAN3 deadenylase complexes, which cooperatively shorten the poly(A) tail of the target mRNA. Poly(A) tail shortening destabilizes the transcript, weakens its association with poly(A)-binding proteins, and commits the mRNA to decapping and exonucleolytic decay. This pathway represents the dominant mechanism by which miRNAs reduce steady-state mRNA levels in animal cells. Deadenylation not only promotes degradation but also creates a translationally inactive mRNA prior to its removal. [20–23].

Translational repression often follows or accompanies deadenylation and further limits protein output. GW182-mediated recruitment of the CCR4–NOT complex interferes with translation initiation by disrupting the recruitment or activity of cap-binding initiation factors, particularly the eIF4F complex, thereby reducing ribosome loading onto the mRNA. This inhibition targets the rate-limiting step of translation, making repression rapid and highly effective [24,25]. By weakening the functional interaction between the 5′ cap and the poly(A) tail, miRNAs prevent efficient mRNA circularization. Some repressed mRNAs remain ribosome-associated, but show reduced initiation frequency rather than complete ribosome exclusion [21,26,27]. In certain contexts, miRNAs may also affect post-initiation steps such as elongation efficiency or ribosome stability [28]. Together, coordinated mRNA degradation and translational repression ensure robust and sustained suppression of gene expression by miRNAs.

The mechanistic role of miRNAs has been further described as activating translation under specific cellular conditions. Upon cell cycle arrest, AU-rich elements and microRNA target sites can shift from repressing to activating translation through the recruitment of microRNAs bound to associated regulatory proteins such as FXR1, thereby mediating a dynamic oscillation between translational repression and activation [29].

Nuclear and extracellular functions expand the mechanistic repertoire of miRNAs beyond their cytoplasmic roles. Mature miRNAs and AGO proteins are also present in the nucleus, where they regulate transcription. Nuclear miRNAs can guide AGO complexes to promoter-associated or nascent RNAs to modulate transcription, either repressing or activating gene expression depending on the chromatin context and changing epigenetic states [30,31].

Beyond the cell, miRNAs act as intercellular messengers. Selective packaging into extracellular vesicles, such as exosomes, protects miRNAs from degradation and enables their transfer to recipient cells, where they impact gene expression and cellular behavior [32].

1.2.3 MicroRNA turnover and decay

miRNA turnover is controlled by several coordinated mechanisms. During loading into AGO, the passenger strand of the miRNA duplex is rapidly displaced and degraded by exonucleases such as XRN1 and XRN2 [33]. In contrast, the guide strand is selectively retained and stabilized through its association with AGO, often resulting in relatively long half-lives of 12-24 hours [34,35].

Once incorporated into the AGO–miRNA complex, the miRNA is largely protected from nuclease attack. The 5′ end of the miRNA is anchored within the MID domain of AGO, while the 3′ end is secured by the PAZ domain. This configuration effectively shields both terminal miRNA ends, restricting access by exonucleases and contributing substantially to miRNA stability [36].

A major pathway regulating miRNA decay involves post-transcriptional modification at the 3′ end, a process known as tailing and trimming. Tailing consists of the addition of non-templated nucleotides to the 3′ end of mature miRNAs and is mediated by terminal nucleotidyl transferases. Among these, TUT4 and TUT7 predominantly add uridine residues, and 3′ adenylation is catalyzed by TUT2 and TUT3 and can promote miRNA stability in certain cellular contexts [37,38].

Following tailing, miRNAs can undergo progressive 3′ to 5′ trimming by exonucleases, which shortens the miRNA and can ultimately lead to complete degradation. For instance, the exonuclease DIS3L2 preferentially targets oligouridylated RNAs, providing a direct mechanistic link between uridylation and miRNA decay. Notably, miRNA degradation is not solely dependent on tailing. Tailing-independent 3′ trimming pathways have also been described, underscoring the existence of multiple, parallel mechanisms governing miRNA turnover [34,36,37].

Another mechanism that controls miRNA turnover is target-directed miRNA degradation (TDMD). In this process, a target RNA, referred to as a trigger, induces degradation of its bound miRNA, reversing the typical miRNA–target relationship. TDMD triggers differ from canonical targets by pairing extensively with both the seed and 3′ region of the miRNA, which alters the AGO–miRNA complex and promotes miRNA decay rather than target repression [39,40].

This extensive pairing displaces the miRNA 3′ end from AGO’s protective PAZ domain, exposing it to tailing and trimming enzymes. As a result, TDMD-affected miRNAs often accumulate as heterogeneous 3′-modified isoforms, including uridylated or shortened species that are more susceptible to exonuclease-mediated decay [37]. Central mismatches or bulges within the duplex prevent AGO-mediated slicing, allowing the complex to persist long enough to initiate degradation [5,34].

A defining feature of TDMD is the coupling of RNA and protein turnover. The conformationally altered AGO recruits the E3 ubiquitin ligase adaptor ZSWIM8 as part of a Cullin–RING complex, leading to AGO polyubiquitination and subsequent proteasomal degradation. Loss of AGO dismantles the RISC complex and leaves the miRNA unprotected, resulting in rapid decay. Notably, a single trigger RNA can sequentially eliminate multiple AGO–miRNA complexes, making TDMD highly efficient (Figure 1.2.2) [34,39,40].

TDMD is distinct from general miRNA decay pathways because it is actively driven by target RNAs and involves ubiquitin-mediated protein degradation. Although tailing and trimming frequently accompany TDMD and enhance its efficiency, AGO degradation alone can be sufficient. Biologically, TDMD is conserved across viruses and eukaryotes [41].

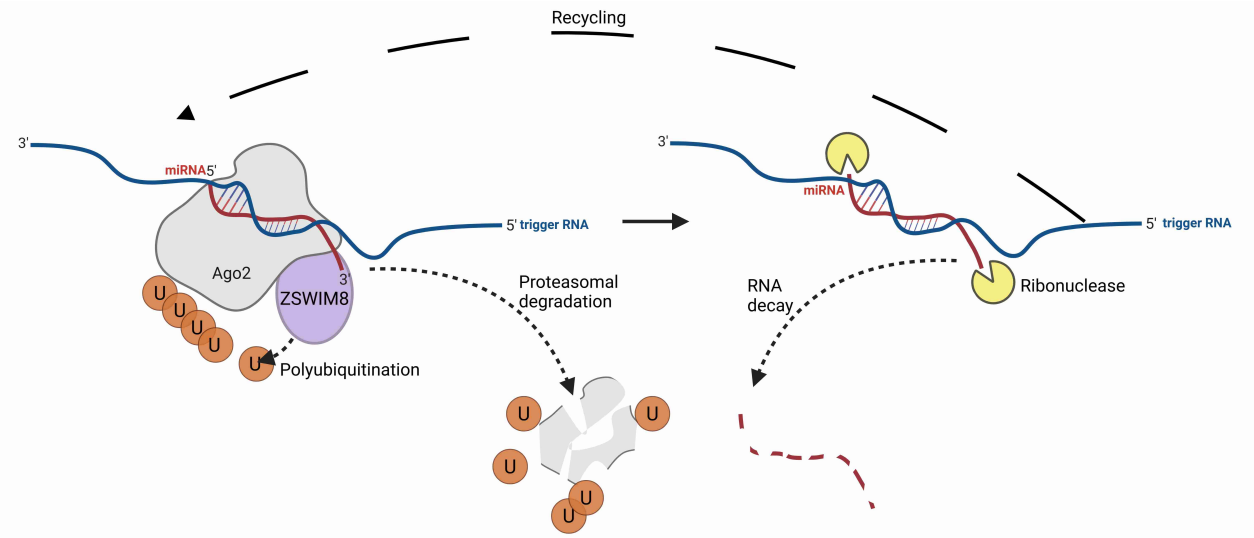


Figure 1.2.2: The current model of action of target-directed miRNA degradation (TDMD). Extensive pairing between a miRNA and its trigger RNA exposes the miRNA's 3′ end and recruits the ZSWIM8 E3 ligase, which polyubiquitinates Ago2 and targets it for proteasomal degradation. After Ago2 is removed, the miRNA is released and degraded by ribonucleases, while the trigger RNA is recovered and can engage in additional rounds of TDMD. Created in BioRender.

1.3 Evolutionary and genetic architecture of microRNA regulation

1.3.1 Conservation of microRNAs

miRNAs constitute a fundamental layer of post-transcriptional gene regulation whose evolutionary dynamics reflect a balance between regulatory robustness and adaptability [42]. Given that miRNAs regulate a substantial fraction of protein-coding genes post-transcriptionally, they exert wide-reaching effects on essential developmental, physiological, and homeostatic processes [1]. Individual miRNAs typically target numerous transcripts, and most mRNAs are co-regulated by multiple miRNAs, forming dense and partially redundant regulatory networks that buffer gene expression [42,43]. Because a single miRNA can influence entire gene networks, even subtle changes in miRNA activity can have multifactorial effects. This network-level robustness helps explain why many functionally important miRNAs are deeply conserved across metazoans [16,44].

However, conservation is not uniform across all miRNAs. Lineage-specific miRNAs are widespread and represent a major source of regulatory novelty and adaptive evolution [45,46]. They are frequently gained through multiple mechanisms, including duplication, *de novo* emergence from intronic or intergenic regions, transposons, or other non-coding RNAs [47]. Comparative genomics shows that a large proportion of miRNA families are restricted to specific evolutionary lineages, indicating rapid turnover alongside a conserved core set. Many recently emerged miRNAs are characterized by low expression levels and strong tissue or context specificity, which limits their immediate functional impact [46–48].

The loss of miRNAs is largely governed by functional relevance. miRNAs that fail to establish beneficial or stable regulatory interactions are prone to elimination through genetic drift or purifying selection [46,48]. In contrast, miRNAs that become incorporated into stable gene regulatory modules tend to increase and stabilize in expression over evolutionary time and are preferentially retained [46,47]. Consequently, more evolutionarily conserved miRNAs generally display higher expression levels, broader regulatory roles, and stronger selective constraints.

Finally, the regulatory impact of miRNAs is closely tied to the evolutionary conservation of their binding sites in target 3′UTRs. miRNAs with conserved target sites tend to exert stronger and more stable regulatory effects, whereas rapidly evolving miRNAs often interact with less conserved binding sites, promoting regulatory flexibility and turnover [45,49,50].

1.3.2 Genetic variation in microRNAs and targets

Genetic variants located within miRNA genes or in their immediate regulatory regions, including promoters, can alter the regulatory effects by changing their expression levels, maturation efficiency, and target recognition. Because miRNAs act as fine-tuners of gene expression, even subtle sequence changes can influence multiple downstream pathways [51,52].

Single-nucleotide polymorphisms (SNPs) within miRNA loci are relatively rare compared with other noncoding regions, reflecting strong evolutionary constraints. Large population-scale analyses show that miRNA variants in the seed region are less frequent than in non-seed regions [53,54]. Despite their rarity, miRNA SNPs can have disproportionate functional effects. Variants within the mature miRNA sequence may disrupt existing binding sites or generate novel ones, leading to potentially large shifts in target specificity and altered repression of entire gene sets [55].

Variants affecting miRNA biogenesis represent another important category. Changes within primary or precursor miRNA sequences can modify local RNA secondary structure and influence processing by Drosha in the nucleus or Dicer in the cytoplasm [56,57]. In such cases, the overall repertoire of target mRNAs typically remains the same, but the abundance of the mature miRNA is altered, resulting in globally increased or decreased repression of its targets. Experimental studies have shown that single-nucleotide changes in hairpin loops or stem regions can either enhance or impair processing efficiency, leading to measurable differences in mature miRNA levels [58,59].

Beyond miRNA genes themselves, variants within miRNA target sites play a major role in modulating regulation. While many such variants act by directly disrupting or creating seed-matching sequences, others exert their effects indirectly by altering the secondary structure of 3′UTRs. These structural changes can lead to inaccessible binding sites, effectively reducing repression without changing the binding motif itself [55,60].

1.3.3 MicroRNA variation in human disease

Genetic variation affecting miRNAs or their binding sites has been described in a variety of human disease onset and progression across many pathological contexts.

One well-studied example is a SNP in miR-146a, which alters miRNA maturation. This variant has been associated with increased cancer susceptibility and progression, as well as more severe coronary artery disease, through dysregulated control of inflammatory and DNA damage response pathways involving targets such as BRCA1, IRAK1, and TRAF6 [61].

In cancer, SNPs in miRNA binding sites are also highly relevant. A BRCA1 3′UTR variant disrupts miRNA-mediated repression and is linked to triple negative breast cancer, while the MDM4 3′UTR SNP creates novel miRNA binding sites and potentially modifies risk across several tumor types [61].

Cardiovascular and metabolic diseases are similarly influenced by such variants. SNPs in 3′UTRs of genes within the renin–angiotensin–aldosterone system, including AVPR1A and NR3C2, impair miRNA binding and contribute to hypertension and early myocardial infarction [62]. Additional variants in the 3′UTRs of INSR, ACSL1, FABP2, and APOL6 disrupt miRNA regulation and are associated with metabolic phenotypes [63].

Neurodegenerative diseases provide further examples. The Parkinson’s disease-associated SNP in the FGF20 3′UTR has been reported to weaken the binding of miR-433, thereby increasing FGF20 and α-synuclein expression [64]. Further, variants in the SNCA and LRRK2 3′UTRs that alter repression by miR-34b and miR-138-2-3p, respectively, were established in Parkinson’s disease [65,66].

Rare variants demonstrate strong effects as well. Seed region mutations in MIR96 cause autosomal dominant progressive hearing loss, and MIR184 mutations lead to familial corneal dystrophy and cataracts [67,68].

Together, these common and rare variants illustrate how disruption of miRNA networks can drive disease susceptibility and progression.

1.4 MicroRNAs in cancer

1.4.1 The Cancer Genome Atlas (TCGA)

The role of miRNAs in cancer has been significantly clarified through the large-scale, multi-institutional effort of The Cancer Genome Atlas (TCGA). As a pioneering cancer genomics project, TCGA systematically profiled over 20,000 primary cancer and matched normal samples spanning 33 cancer types. Using standardized protocols, including small RNA sequencing, it generated quantitative data on miRNA expression levels and isoforms. These datasets, covering a wide range of tissues and

disease states, enabled unprecedented comparisons between tumor and normal conditions [69].

A key advantage of TCGA lies in its comprehensive integration of diverse data types into a unified resource. In addition to miRNA expression, it includes matched genomic mutations, mRNA expression, DNA methylation, copy number variation, and extensive clinical information. This layered dataset offers a broad and interconnected view of tumor biology, allowing researchers to uncover how miRNAs shape gene regulation, reflect underlying genetic alterations, and influence clinical outcomes. As a result, the role of miRNAs has been elevated from secondary contributors to central regulators in cancer biology, with both mechanistic and translational significance.

A major achievement stemming from TCGA-based studies is the identification of miRNA biomarkers with diagnostic and stratification value. Through large-scale expression analyses, researchers have developed miRNA signatures that reliably distinguish tumor types and molecular subtypes. These signatures often surpass mRNA-based classifiers, owing to the high tissue specificity, stability, and regulatory precision of miRNAs, making them ideal candidates for diagnostic assays [70–72].

Another prominent application is the development of prognostic models. Survival analyses leveraging TCGA data have linked specific miRNAs and multi-miRNA signatures to patient outcomes across multiple cancer types. These risk scores provide tools for stratifying patients based on expected disease progression, independent of conventional clinical metrics [73,74].

Additionally, miRNA profiles have shown promise in predicting therapeutic response. Certain miRNAs have been associated with sensitivity or resistance to chemotherapy, radiotherapy, and targeted therapies. Mapping these expression patterns helps identify patients more likely to respond to specific treatments, offering a path toward more personalized and effective cancer care [75–77].

TCGA findings have also guided the development of miRNA-based therapeutic approaches. By spotlighting miRNAs that are consistently dysregulated and clinically significant, TCGA has enabled the prioritization of candidates for therapeutic targeting, whether through anti-miRNA agents to inhibit oncogenic miRNAs or replacement strategies to restore tumor-suppressive ones [78–80].

1.4.2 OncomiRs and tumor suppressors

Multiple molecular mechanisms interfere with microRNA regulation that contribute to cancer progression. As key post-transcriptional regulators, miRNAs are responsible for maintaining fine-tuned control over gene expression programs [16]. When this balance is lost, either through the loss of miRNAs that normally suppress oncogenic targets or the overexpression of miRNAs that inhibit tumor suppressor genes, the result can be uncontrolled proliferation, impaired differentiation, and other hallmarks of cancer. miRNAs that promote cancer by repressing tumor suppressor genes are referred to as oncogenic miRNAs, or oncomiRs, while miRNAs that act as tumor suppressors can contribute to malignancy when underexpressed or lost [81,82]. This

disruption in target regulation can be attributed to several upstream alterations (Figure 1.4.1).

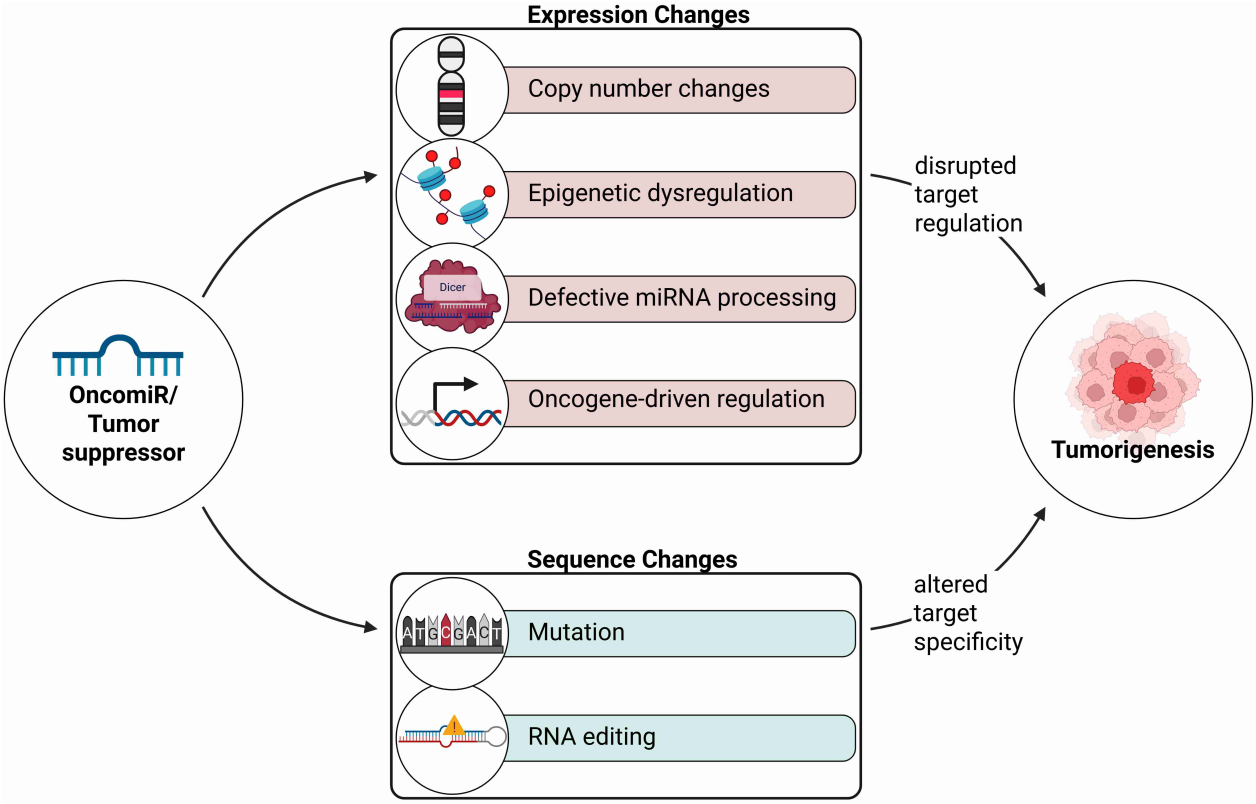


Figure 1.4.1: Overview of mechanisms driving oncomiR or tumor-suppressor miRNA dysregulation in cancer. Created in BioRender.

A major set of these mechanisms acts at the sequence level and affects the target repertoire of the miRNA. miRNAs can acquire tumor-promoting properties through alterations in their sequence. These include somatic mutations that occur within miRNA genes, especially in the seed region critical for target recognition [83]. Even a single nucleotide change can alter a miRNA's processing efficiency, strand selection, or target repertoire. Some mutations result in the gain of new targets or the loss of repression of essential tumor suppressors, effectively reprogramming the miRNA's function [61,84]. While these events are less common than expression-level disruptions, their impact can be substantial due to the broad downstream effects of mis-targeting.

RNA editing represents another layer of sequence-based regulation. Adenosine-to-inosine editing by ADAR enzymes can alter the seed sequence of a mature miRNA, redirecting its target specificity [85,86]. In some cancers, editing is reduced, which maintains a miRNA in its unedited, potentially oncogenic form. In other contexts, aberrant editing creates miRNAs with new pro-oncogenic properties [87,88]. This type of editing-driven shift can remodel post-transcriptional networks in ways that subtly or dramatically alter cellular behavior, often with consequences for metastasis, immune evasion, or resistance to therapy.

In contrast to these sequence-level mechanisms, defective miRNA biogenesis also contributes to widespread dysregulation, though through different means and often at the processing level.

Another important contributor is genomic copy number change. Amplification of miRNA loci can lead to abnormally high levels of certain miRNAs that downregulate genes that normally restrict growth, while deletions may eliminate miRNAs with tumor-suppressive functions [89]. These gains or losses are common across cancers, with oncomiRs often found in amplified regions [90].

A widespread mechanism involves epigenetic silencing. Many tumor-suppressive miRNAs are located near CpG-rich promoter regions that are susceptible to hypermethylation in cancer cells [91]. Methylation of these regions effectively shuts down transcription, leading to decreased levels of mature miRNA and consequently the loss of regulation over their target oncogenes [92]. Similarly, histone modifications that induce a closed chromatin state around miRNA loci can prevent transcriptional access [93]. These epigenetic mechanisms are especially significant in early tumor development, where they contribute to silencing of key regulatory miRNAs without requiring structural genetic alterations [94].

In addition to these changes, aberrant transcriptional regulation by oncogenic signaling pathways also reshapes miRNA expression. Oncogenes such as MYC can directly activate or repress miRNA transcription [95]. This commonly results in coordinated upregulation of miRNAs that support proliferation, metabolic adaptation, or evasion of apoptosis. The downstream effect is the reinforcement of oncogenic networks, as these miRNAs preferentially suppress tumor suppressor genes or modulate immune evasion pathways [95,96].

Altogether, disrupted biogenesis, copy number changes, epigenetic silencing, and oncogene-driven transcription contribute to global shifts in miRNA expression that impair normal gene regulation and favor malignant transformation.

1.4.3 Glioblastoma as a model of regulatory complexity

Glioblastoma multiforme (GBM) is the most prevalent and aggressive brain tumor in adults, characterized by heterogeneity, rapid progression, and resistance to current therapies [97]. Among the many molecular drivers of GBM, miRNAs have emerged as critical modulators of tumorigenesis. Their dysregulation in GBM has become a major focus of research, revealing their influence on core oncogenic processes like proliferation, apoptosis, invasion, and cellular plasticity [98–100].

Several miRNAs function as oncogenic drivers in GBM. miR-21 is among the most frequently upregulated miRNAs and acts by repressing multiple tumor suppressors, thereby enhancing proliferation [101–103]. Inhibition of miR-21 in experimental models leads to reduced tumor growth, while its overexpression in patient tumors is consistently linked to poor prognosis [101,104,105]. Similarly, miR-10b, largely absent in normal brain tissue, is highly expressed in gliomas, where it suppresses apoptosis and promotes cell cycle progression by targeting genes such as Bim and p21. High levels of miR-10b correlate with increased proliferative capacity and significantly reduced patient survival [106,107].

Conversely, several miRNAs enriched in the healthy brain are markedly downregulated in GBM and act as tumor suppressors. miR-7, which inhibits key components of the EGFR and MAPK pathways, is one such example. Loss of miR-7 enhances oncogenic signaling, while its restoration reduces tumor growth [108]. miR-128, another neural-specific miRNA, is also significantly reduced in glioma cells [109]. These tumor-suppressive miRNAs help maintain differentiation programs, and their downregulation contributes to the aggressive and stem-like phenotype that characterizes high-grade GBM.

miRNAs also influence the tumor microenvironment. For example, miR-21 has been implicated in modulating immune cell infiltration and promoting an immunosuppressive milieu, while other miRNAs have been linked to angiogenesis and extracellular matrix remodeling [99,110,111]. The interactions between GBM cells and their microenvironment are increasingly recognized as essential for tumor maintenance and progression [112,113].

Importantly, miRNA expression patterns carry clinical significance. Distinct miRNA profiles have been linked to GBM subtypes and patient outcomes. For instance, high expression of miR-181d is associated with reduced MGMT levels and improved response to temozolomide, the frontline chemotherapeutic agent for GBM [114]. In contrast, elevated levels of miR-21 or miR-10b are linked to treatment resistance and shorter survival [100,107].

1.5 Alternative polyadenylation and microRNA-mediated regulation

One of the key determinants of miRNA activity is the presence, composition, and accessibility of their binding sites within target transcripts. This regulatory landscape is highly dynamic, largely due to a mechanism known as alternative polyadenylation (APA) [115]. APA occurs in an estimated 70 percent of protein-coding genes and results in the generation of transcript isoforms with variable 3′UTR lengths [116]. Consequently, APA modifies the availability of microRNA binding sites, thereby influencing post-transcriptional gene regulation [117].

1.5.1 Molecular mechanisms of APA

During mRNA production, polyadenylation marks the final processing step before a transcript is fully matured and exported from the nucleus to the cytoplasm [115]. The poly(A) tail is a sequence of adenosine nucleotides added to the 3′ end of a precursor mRNA and protects the transcript from degradation, ensures efficient export, and enables translation. Without polyadenylation, mRNAs are unstable and often fail to be translated into protein [115,118].

The addition of the poly(A) tail depends on sequence motifs within the 3′UTR and the coordinated action of several protein complexes. The main players are the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), and cleavage factors I and II (CFIm and CFIIm) [115,118–120]. CPSF binds to the polyadenylation site (PA site), an upstream AAUAAA sequence motif or a close variant, while CstF interacts with a downstream GU-rich region. Once these sites are engaged, CFIm and CFIIm help align the machinery for cleavage. After the cut is

made, poly(A) polymerase (PAP) adds the tail, and poly(A)-binding proteins control the length and protect from degradation (Figure 1.5.1) [115,118].

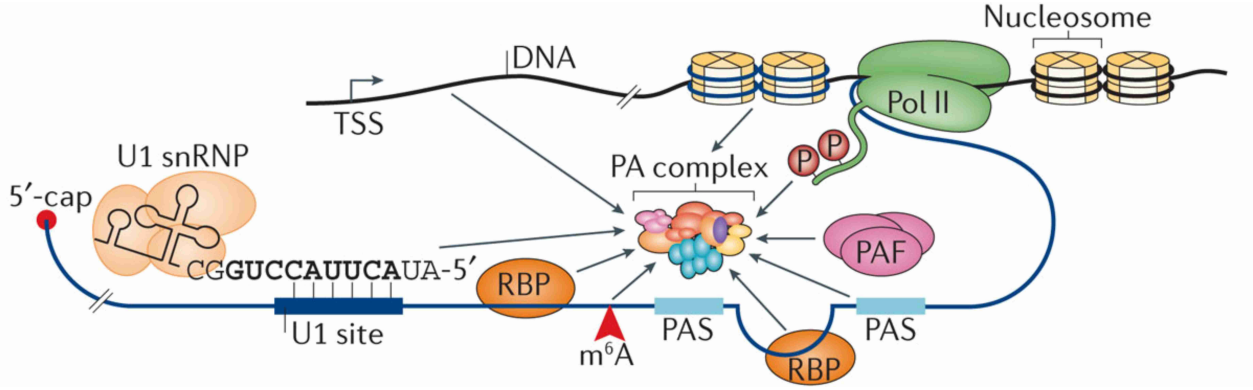


Figure 1.5.1: Factors influencing alternative polyadenylation. PAS choice can be affected by the gene promoter, polyadenylation-related proteins, local nucleosome density, RNA polymerase II elongation, RNA-binding proteins, m6A marks, and inhibition by U1 snRNP. Figure adapted from Tian & Manley (2017).

APA increases the diversity of mRNA isoforms produced from a single gene by cleavage and tail addition at different positions when multiple PA sites are located within the same gene.

Experimental studies have shown that APA is a regulated process that responds to changes in the abundance and activity of specific processing factors. For instance, reducing levels of the CFIm complex, particularly the CFIm25 subunit (NUDT21), causes widespread use of proximal PA sites and leads to shorter 3′UTRs [121–123]. These results show that PA site choice depends not only on the sequence motifs present but also on the availability and balance of core and accessory RNA-binding proteins that interact with the polyadenylation machinery. APA is therefore a dynamic and flexible mechanism for fine-tuning gene expression, not just background noise.

Several distinct types of APA have been described, characterized based on where alternative PA sites are located in the mRNA. The most common type is tandem APA, where two or more PA sites lie within the same terminal exon, resulting in 3′UTRs of varying lengths but identical protein-coding sequences. Intronic APA occurs when cleavage happens within an intron, often before splicing is complete, leading to truncated transcripts that may code for shorter proteins. In splicing APA, the inclusion or exclusion of certain exons alters the set of available PA sites. Internal APA refers to the use of less common upstream polyadenylation signals within coding exons or introns, often producing unstable or nonfunctional transcripts (Figure 1.5.2) [115,124–126].

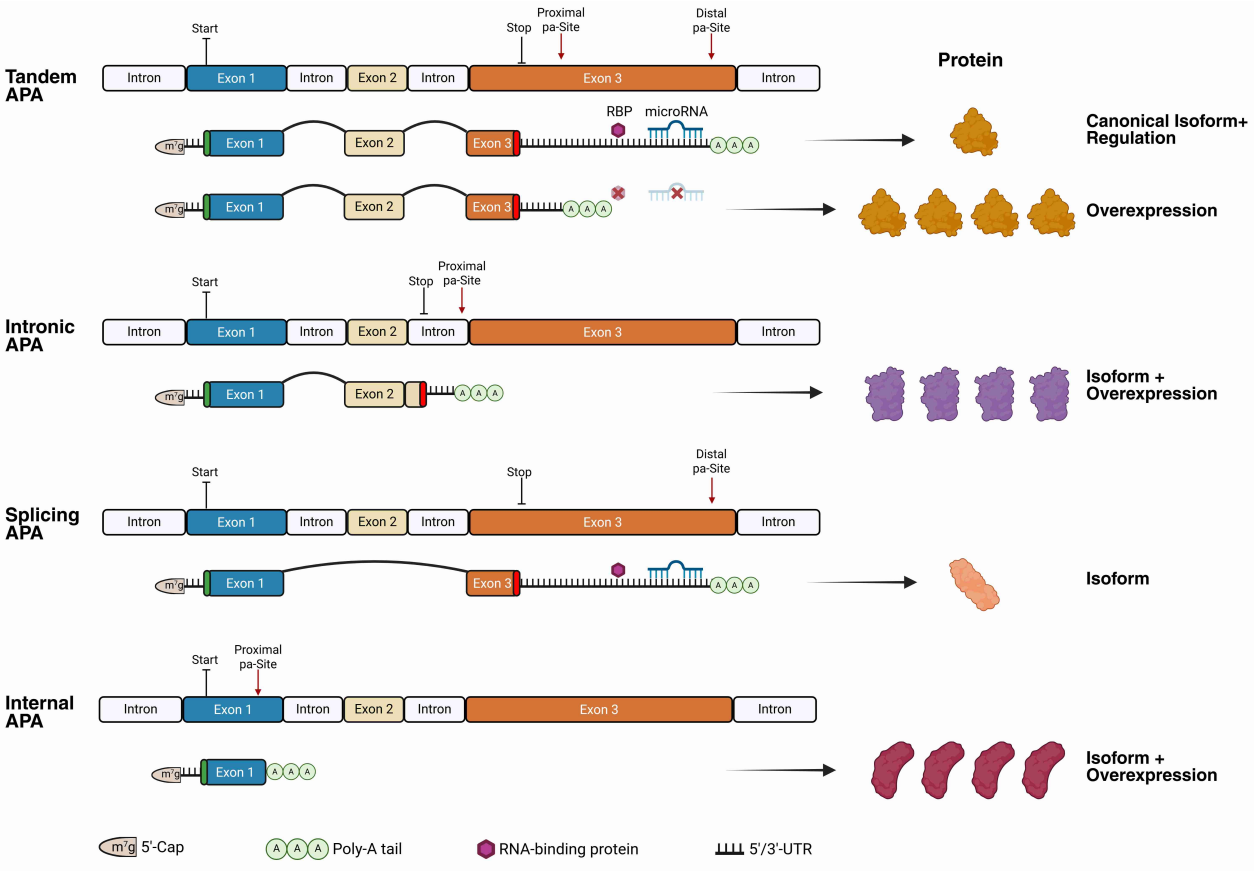


Figure 1.5.2: Alternative polyadenylation (APA). Tandem APA occurs when proximal or distal poly(A) sites (PA sites) are selected within the terminal exon, producing mRNAs with shorter or longer 3′UTRs. Use of a proximal site shortens the 3′UTR and can remove regulatory elements such as miRNA-binding sites, potentially increasing protein expression. Intronic APA generates truncated transcripts by selecting PA sites within introns, resulting in shorter isoforms that may be more highly expressed through loss of miRNA binding sites. Similarly, splicing-coupled APA links alternative exon choice with PA site selection, yielding distinct transcript variants with different regulatory potentials. Internal APA uses upstream PA sites to produce shortened mRNAs that lack portions of the 3′UTR and can therefore potenitally escape post-transcriptional repression. Created in BioRender.

In addition to sequence elements and regulatory proteins, the speed of transcription also influences APA. Faster RNA Polymerase II elongation reduces distal PA site recognition, favoring upstream cleavage and shorter isoforms. Slower elongation favors downstream signals to be processed. Chromatin structure and specific histone modifications can also affect how quickly polymerase moves along the gene, and thus indirectly influence APA outcomes [127,128].

This ability to generate different mRNA isoforms from the same gene has significant implications for how transcripts are regulated after transcription. In particular, changes in 3′UTR length directly affect the presence or absence of regulatory elements such as microRNA binding sites [125,129]. Thus, APA provides a molecular switch that controls how susceptible an mRNA is to post-transcriptional regulation, even when its overall expression level remains unchanged.

1.5.2 APA across cell types and states

APA usage is highly cell type–specific, and also varies with developmental stage, differentiation status, and disease state. This specificity is largely governed by the availability and activity of core 3′ end processing factors, which can shift PA site usage globally. Different cell types express varying levels and sometimes specialized paralogs of these factors, allowing them to shape distinct APA profiles that contribute to their identity and function [115,130,131].

Tissues such as muscle, brain, and immune cells express RBPs that promote or repress polyadenylation at certain sites, tuning the output of APA [130,132]. For example, the neuron-specific RNA binding protein (RBP) Nova1 and the immune cell–expressed CstF-64 each influence site usage in ways that align with tissue-specific gene expression programs [133]. In general, APA provides a mechanism for the same gene to be regulated differently across cell types by modifying transcript stability, localization, and translation, and often in coordination with miRNA networks.

During embryonic development, APA undergoes dynamic reprogramming. Early embryonic cells, which are highly proliferative, tend to use proximal PA sites. As cells exit proliferation and begin lineage specification, there is a progressive shift toward longer 3′ UTR isoforms enriched in regulatory elements [134–136]. This lengthening coincides with increased engagement of post-transcriptional regulatory mechanisms, including miRNAs and RBPs, and supports the emergence of stable, tissue-specific gene expression programs [135].

The nervous system is another striking example of APA specialization. Neurons exhibit widespread usage of long 3′-UTR isoforms, a phenomenon conserved across mammals [137]. In neurons, APA is tightly linked to function: long 3′ UTRs facilitate mRNA transport to dendrites or axons and support activity-dependent translation at synapses [138]. For example, the BDNF gene uses a long isoform that targets the mRNA to dendrites and responds to synaptic signals, while a shorter isoform remains confined to the soma [139,140]. Moreover, many neural transcripts are co-regulated by shared miRNAs through inclusion of conserved binding sites supporting coordinated control over functionally related genes [141].

Similarly, the immune system uses APA to diversify its transcriptome and control gene expression during activation and differentiation. One example is immunoglobulin heavy-chain transcripts in B cells, which switch APA sites to produce either membrane-bound or secreted antibody isoforms [133]. This switch is driven in part by elevated expression of CstF-64 during plasma cell differentiation [133].

In T cells, APA remodeling also occurs during activation and effector differentiation. Upon antigen stimulation, many genes involved in T cell signaling, trafficking, and cytokine production undergo 3′-UTR shortening or isoform switching [142,143].

In contrast to the lengthening seen in differentiated cells like neurons, cancer cells often display global 3′-UTR shortening [117]. This tendency towards shorter isoforms mirrors APA profiles seen in undifferentiated cells [131]. Shortened 3′ UTRs in tumors may also disrupt cross-talk among transcripts competing for shared miRNAs, altering the balance of gene expression across entire regulatory networks [144]. Moreover,

loss of APA regulation in cancer is not limited to oncogenes and can affect tumor suppressors and signaling molecules [144,145].

1.6 Bioinformatics approaches to microRNA research

Bioinformatics has been fundamental in miRNA research, providing the frameworks required to analyse and understand miRNA-mediated gene regulation. Computational approaches underpin all major stages of miRNA research, including miRNA gene annotation, target identification, expression profiling, and regulatory network reconstruction. By leveraging sequence analysis, evolutionary conservation, and large-scale high-throughput data, bioinformatics has allowed miRNAs to be studied systematically at the genome and transcriptome level, establishing them as integral components of gene regulatory systems across diverse biological and pathological contexts [146].

1.6.1 MicroRNA annotation

Initially, miRNAs were discovered through classical genetics and low-throughput molecular biology, most notably in *C. elegans* [2,3]. These studies established defining miRNA features, including short length, derivation from hairpin precursors, precise Drosha/Dicer processing, and incorporation into AGO complexes [147]. Early bioinformatic approaches exploited these properties by identifying hairpin-forming loci, characteristic small RNA sequencing read patterns, and evolutionary conservation, enabling scalable, genome-wide miRNA discovery [148,149].

As discovery accelerated, several databases emerged to catalog annotated miRNAs, with miRBase becoming the most widely used resource [7,150]. miRBase aggregates published miRNA sequences and standardizes nomenclature, but its annotation strategy has historically relied on author-submitted evidence. High-confidence annotations were initially supported by deep sequencing and required features such as perfectly matching reads mapping to both mature arms of the precursor, reads with 3′ overhangs, and prediction of a stable hairpin structure [7,150]. Despite these guidelines, large-scale analyses revealed substantial false-positive rates in miRBase, particularly among species-specific and lowly expressed miRNAs [151].

In contrast, MirGeneDB was introduced as a manually curated, evolution-aware alternative that applies strict and explicit criteria emphasizing canonical biogenesis, precise processing, and evolutionary conservation [8,152]. Consequently, MirGeneDB is far more conservative, recognizing only a subset of miRBase annotations as high-confidence miRNA genes [152]. This discrepancy highlights the trade-off between annotation breadth and reliability and underscores the importance of rigorous curation for downstream analyses such as target prediction and regulatory network modeling.

High-confidence annotations have enabled more advanced computational discovery strategies. Curated datasets such as MirGeneDB have been used to train covariance models that integrate sequence, secondary structure, and evolutionary constraints, allowing highly specific, genome-wide identification of novel miRNA loci [153].

Despite these advances, miRNA annotation remains challenging, particularly for lowly expressed, species-specific, or condition-dependent miRNAs, where distinguishing functional molecules from background noise continues to be difficult [154].

1.6.2 Target identification

A defining challenge in miRNA research is the identification of mRNA targets. Because miRNAs regulate gene expression through short and often imperfect base pairing, experimental identification of all miRNA–mRNA interactions is not scalable [155]. Bioinformatics algorithms have therefore played a central role in defining miRNA targets computationally, enabling large-scale functional interpretation of miRNA activity.

A principle that could be exploited computationally, and that forms the foundation of most target prediction algorithms, is the search for complementary sequences to the miRNA seed region within target mRNAs [16,42]. However, this approach alone is insufficient to explain targeting specificity, as large transcriptomes contain millions of potential seed matches. To address this challenge, bioinformatic approaches incorporated additional contextual features, including evolutionary conservation, local sequence composition, site accessibility, thermodynamic stability, and the presence of multiple binding sites, to distinguish functional targets from background noise [16,42,156,157].

One major advance in miRNA target prediction was the introduction of evolutionary conservation as a filtering principle [42]. The underlying assumption is that functional miRNA binding sites are often preserved across related species due to selective pressure. This concept is central to TargetScan, which identifies conserved seed matches of potentially extended length across vertebrate or metazoan genomes and ranks predicted targets based on both site type and conservation metrics [42]. Subsequent versions of TargetScan incorporated context-based scoring metrics that account for features such as AU-rich flanking sequences, binding site position within the UTR, and UTR length [16,158]. Conservation-based methods substantially increased prediction specificity, although they are inherently biased against species-specific and recently evolved miRNAs whose functional targets may lack detectable conservation [159].

Alternative algorithms emphasized the biophysical properties of miRNA–mRNA interactions, focusing on sequence complementarity beyond the seed region and thermodynamic considerations. These methods enabled the identification of non-canonical and non-conserved binding sites but typically generated larger candidate target sets, often with increased false-positive rates. Incorporation of RNA secondary structure and local accessibility further refined these predictions, underscoring the importance of transcript context in miRNA-mediated regulation [160–162].

Algorithms such as DIANA-microT combine multiple features, including seed pairing, conservation, thermodynamics, and site context, into unified probabilistic scores [157]. In parallel, the integration of expression-based evidence became an

important complementary strategy, whereby miRNA–mRNA pairs exhibiting anticorrelated expression patterns were analyzed [163]. While anticorrelation alone is not sufficient to establish direct targeting, its incorporation alongside sequence-based predictions improves biological relevance.

More recently, machine learning and deep learning approaches have been applied to miRNA target prediction. These models are trained on experimentally validated miRNA–mRNA interactions and can automatically learn complex, non-linear feature combinations, including non-canonical interactions [164–166].

To mitigate uncertainty inherent in computational prediction, bioinformatics has increasingly incorporated high-throughput experimental evidence. Crosslinking and immunoprecipitation sequencing (CLIP-seq) technologies identify transcript regions physically bound by AGO proteins, providing maps of RISC-occupied regions enriched for miRNA target sites. Several databases integrate CLIP-seq data with computational predictions, substantially increasing confidence in inferred miRNA–mRNA interactions [156,167,168].

1.6.3 MicroRNA networks and function

The functional annotation of miRNAs primarily relies on systematic analysis of their mRNA targets. Since a single miRNA can modulate the expression of hundreds of genes, understanding its role requires large-scale approaches that go beyond isolated interactions [42]. One widely used strategy involves identifying enrichment and over-representation of biological processes or pathways among predicted or experimentally validated targets [169]. These analyses, conceptually similar to those applied in mRNA expression studies, draw on gene annotation resources such as Gene Ontology (GO) and KEGG, and are implemented through tools like DAVID, g:Profiler, or miRNA-focused platforms such as DIANA-miRPath and miEAA [169–172]. This provides a first approximation of a miRNA's potential biological roles.

Beyond enrichment, bioinformatics also supports analysis of tissue and context-specific miRNA expression [173,174]. Integrating target enrichment data with spatiotemporal expression profiles, such as those available in resources like miRNATissueAtlas, helps refine functional interpretations by identifying in which biological settings a given miRNA is likely to be active [173,174].

miRNAs function within broader regulatory networks, where their activity intersects with that of TFs, RBPs, and other miRNAs. For instance, multiple miRNAs may co-target the same mRNA, or a single miRNA might regulate a TF that subsequently controls the expression of additional miRNAs. Modeling these coregulatory networks enables a shift from static lists of targets to dynamic interpretations of miRNA function within specific expression contexts [175,176].

A commonly studied network motif is the feed-forward loop (FFL), where a TF regulates a miRNA or vice versa, and both jointly regulate a shared target gene (Figure 1.6.1) [176]. This creates a three-component regulatory circuit that can result in aligned or opposing regulatory effects. FFLs are widespread regulatory networks, and their identification requires the integration of binding site data,

transcriptional control information, and expression dynamics [176]. FFLs can buffer transcriptional noise, and facilitate rapid, transient responses. Additionally, they have been observed in systems ranging from stem cell differentiation to tumor progression [177–180]. A well-known example is the c-Myc, E2F, miR-17-92 circuit, where c-Myc activates both E2F and the miR-17-92 cluster, the latter acting to repress E2F, forming a regulated proliferative module [181]. Another example involves NF-κB, which activates pro-inflammatory genes as well as miRNAs like miR-146a that, in turn, suppress key signaling intermediates, thereby modulating the immune response [182,183].

However, FFLs represent only the core units of more complex regulatory architectures. In real biological contexts, these loops are shaped by additional regulators and mechanisms that may alter or disrupt their function. The incorporation of such post-transcriptional layers, including APA, adds substantial context dependency to miRNA-centered regulatory networks (Figure 1.6.1). What begins as a relatively simple loop may, in reality, behave as a dynamic, multi-input module that is sensitive to changes in transcriptome composition, cellular state, or subcellular localization [184,185].

Bioinformatics remains central to resolving this complexity. Through the development of integrative models that account for multiple regulatory layers and contextual dependencies, computational approaches are moving the field beyond static catalogs of miRNA–mRNA pairs toward a deeper understanding of miRNAs as adaptive components of larger gene regulatory-systems [176,186,187].

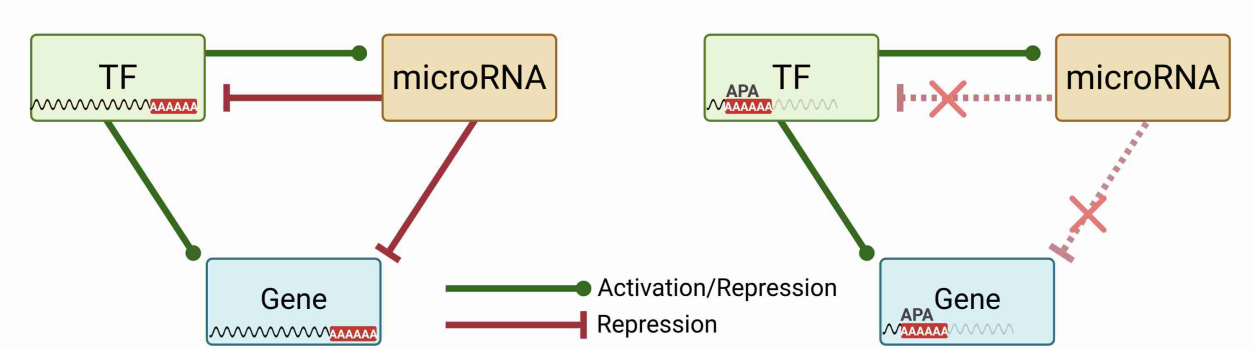


Figure 1.6.1: Transcription factor (TF)–microRNA (miRNA)-gene feed-forward loops (FFLs). The left panel shows a coregulatory FFL, where the TF and the miRNA each regulate the target gene, and at least one of them also regulates the other component, allowing for mutual regulation as a possible configuration. The right panel illustrates disruption of this FFL when APA-mediated 3′UTR shortening removes miRNA-binding sites targeting the TF and target-gene, preventing miRNA repression and breaking the regulatory loop. Created in BioRender.

1.7 Machine learning in regulatory biology

1.7.1 Machine learning for gene regulatory network inference

The availability of high-throughput data, such as RNA sequencing, single-cell omics, and genome-wide profiling technologies, has produced rich datasets that are high-dimensional, complex, and noisy. Traditional statistical methods regularly fall short in obtaining valuable insights from such data. Consequently, machine learning (ML) has become a key tool for understanding regulatory biology, as it enables the

identification of patterns without specific prior assumptions, handles large feature spaces, and supports predictive modeling from complex molecular datasets [188,189].

At the core of regulatory biology is the gene regulatory network (GRN), which represents how transcription factors, non-coding RNAs, and other regulators, such as chromatin modifiers, coordinate gene expression. These networks consist of numerous interacting components whose relationships are frequently nonlinear, context-dependent, and combinatorial [189,190]. Machine learning approaches, including Bayesian networks, random forests, support vector machines (SVMs), and regression-based models, were among the first computational strategies applied to GRN inference, allowing more accurate prediction of regulatory interactions than earlier methods based solely on unsupervised clustering or pairwise correlation analyses. More recently, deep learning and advanced structured ML models, such as graph neural networks, have additionally improved the field's capacity to capture nonlinear interactions and latent regulatory patterns across diverse omics layers, including transcriptomic and epigenomic data [189,191,192].

A key distinction in ML modeling lies in the choice between linear and nonlinear techniques. Linear models offer interpretability and robustness, particularly in settings with limited sample sizes, but they may oversimplify biological systems that involve feedback loops, combinatorial regulation, and threshold-dependent effects. In contrast, nonlinear methods, including tree-based ensemble models and deep neural architectures, can learn complex dependencies and interaction effects that are common in gene regulatory processes. However, these models typically require careful regularization and hyperparameter tuning, along with stringent validation, to reduce overfitting, especially in biological datasets in which the number of measured features far exceeds the number of samples. As a result, studies often evaluate multiple ML approaches to balance predictive performance with biological interpretability across experimental conditions [189,192–195].

The application of machine learning in regulatory biology goes beyond network inference. ML methods are widely used for multi-omics data integration, classification of cell types and cellular states, discovery of regulatory modules, and prediction of phenotypic outcomes from molecular profiles [190]. Through these approaches, researchers can move beyond simple lists of differentially expressed genes toward mechanistic models of regulation, allowing the discovery of key regulatory hubs and pathways through analyses such as feature importance and model-based inference.

1.7.2 Machine learning in microRNA research

ML contributes across most tasks in miRNA research. First, ML is widely used for miRNA discovery and classification, identifying novel miRNA precursors and mature sequences from genomic data by learning sequence and structural features that differentiate true miRNAs from other non-coding elements. It also improves the accuracy of miRNA target prediction by combining various features derived from sequence complementarity, structural accessibility, and expression patterns. ML models outperform heuristic, rule-based target predictors by learning complex patterns associated with functional miRNA–mRNA interactions [196,197].

Beyond sequence analysis, ML models were applied in modeling miRNA expression and assessing regulatory influence. For example, ML methods have been used to predict miRNA expression levels from gene expression profiles, revealing the multifactorial regulatory relationships underlying miRNA control [198,199].

Machine learning also enables regulatory module detection and disease association studies by identifying groups of miRNAs and target genes whose coordinated expression patterns discriminate between biological states [200–202]. In clinical research, ML classifiers trained on miRNA profiles have identified candidate miRNA signatures that distinguish disease from control samples with high accuracy. For example, in prostate cancer diagnostics, random forest models applied to miRNA expression data yielded strong classification metrics and revealed miRNA expression ratios with superior diagnostic performance compared to conventional biomarkers [203].

Furthermore, ML frameworks support the construction of miRNA regulatory networks that integrate miRNA–mRNA, miRNA–lncRNA, and circ-RNA interactions, enabling integrated views of post-transcriptional regulation. In these network contexts, ML assists not only in prediction but also in identifying key regulatory hubs, potential therapeutic targets, and pathways that are enriched in specific disease contexts [201,203,204].

Deep learning (DL) has recently extended traditional machine learning approaches in miRNA research by using neural network architectures to automatically learn informative representations from raw sequence, structural, and expression data. DL models have shown improved performance in miRNA discovery and target prediction by capturing complex, nonlinear patterns underlying miRNA biogenesis and miRNA–mRNA interactions without relying on selection of features [164,205].

1.7.3 Regularization methods

A challenge in regulatory biology and miRNA research is the high-dimensional nature of data. Here, the number of predictors often exceeds the number of samples. In such settings, standard regression models are prone to overfitting [206]. They capture noise rather than biologically relevant patterns and perform poorly on unseen data. Regularization addresses this problem by introducing constraints that penalize complex models.

In linear machine learning models, regularization is typically implemented by adding a penalty term to the loss function, which shrinks large coefficient values [207]. Regularization is especially important in regulatory biology, where many features may be weakly informative or highly correlated [188,208].

Three regularized linear models are especially prominent in biological research: ridge regression, LASSO, and elastic net [194,207,209].

Ridge regression applies an L2 penalty that shrinks all coefficients toward zero but does not eliminate any predictors. This makes ridge well-suited in situations where many features contribute small, cumulative effects. This is often the case in gene expression regulation. Ridge regression also handles multicollinearity effectively by distributing weights across correlated predictors [194].

LASSO introduces an L1 penalty that forces some coefficients to exactly zero, producing sparse models that perform automatic feature selection. However, LASSO can behave unstably when predictors are highly correlated, arbitrarily selecting one feature while discarding others that may be biologically relevant [207].

Elastic net combines L1 and L2 penalties, inheriting strengths from both LASSO and ridge regression. It performs feature selection while also retaining groups of correlated predictors. This makes it particularly interesting for transcriptomic and miRNA data, where regulatory signals often occur in coordinated sets. Although elastic net requires tuning multiple parameters, it often yields more stable and biologically plausible models than either method alone [209].

From a biological perspective, regularized linear models are especially valuable because they combine interpretability and predictive power. Specifically, each coefficient directly links a molecular feature to an outcome, which supports mechanistic hypothesis generation [189,190].

2 Research articles

In the following, the research articles contributing to this thesis are presented. The final subsection provides an overview of additional publications in which the author participated as a collaborator.

2.1 Detection of features predictive of microRNA targets by integration of network data

2.1.1 Preamble

This chapter is published in the journal PLOS ONE:

Cihan, M., & Andrade-Navarro, M. A. (2022). Detection of features predictive of microRNA targets by integration of network data. PLOS ONE, 17(6), e0269731. https://doi.org/10.1371/journal.pone.0269731

The supplementary files associated with this publication are available on the publisher's website via the article's DOI.

2.1.2 Abstract

Gene activity is controlled by multiple molecular mechanisms, for instance through transcription factors or by microRNAs (miRNAs), among others. Established bioinformatics tools for the prediction of miRNA target genes face the challenge of ensuring accuracy, due to high false positive rates. Further, these tools present poor overlap. However, we demonstrated that it is possible to filter good predictions of miRNA targets from the bulk of all predictions by using information from the gene regulatory network. Here, we take advantage of this strategy that selects a large subset of predicted microRNA binding sites as more likely to possess less false-positives because of their over-representation in RE1 silencing transcription factor (REST)-regulated genes from the background of TargetScanHuman 7.2 predictions to identify useful features for the prediction of microRNA targets. These enriched miRNA families would have silencing activity for neural transcripts overlapping the repressive activity on neural genes of REST. We analyze properties of associated microRNA binding sites and contrast the outcome to the background. We found that the selected subset presents significant differences respect to the background: (i) lower GC-content in the vicinity of the predicted miRNA binding site, (ii) more target genes with multiple identical microRNA binding sites and (iii) a higher density of predicted microRNA binding sites close to the 3' terminal end of the 3'-UTR. These results suggest that network selection of miRNA-mRNA pairs could provide useful features to improve microRNA target prediction.

2.1.3 Introduction

Post-transcriptional repression of mRNAs by microRNAs (miRNAs) is one of multiple layers of regulation of gene expression [1]. Since the discovery of the first miRNA, lin-4 in Caenorhabditis elegans in 1993 [2], more than 2,300 human miRNAs with numerous regulatory functions have been identified [151]. Particularly, the malfunctioning of miRNA regulation has been described as promoting neurological diseases [210] and various types of cancer [211], among other illnesses [212].

Estimates suggest that approximately 60% of protein-coding genes in the human genome may be regulated by miRNAs [213]. However, miRNA functional characterization is experimentally difficult due to their regulatory mechanisms, which are more subtle and less specific than transcription factors and epigenetic modifications [214]. This has fueled the development of many bioinformatics tools for the prediction of miRNA-mRNA interactions [166].

A commonly applied algorithm for detecting miRNA-target genes is based on finding 3’-UTR sequences with conserved sites that are complementary to the seed region of broadly conserved miRNAs, following the rules of Watson-Crick base pairing. Along with the integration of further criteria, such as the conservation of the 3’-UTR across mammalian species, the presence of complementary sequences around the matching seed and the assignment of binding free energy, many tools perform ranking of predicted miRNA-mRNA interactions to determine the probability of conserved targeting [42,150,162,215].

Regardless of these efforts, and although the regulatory mechanisms and biogenesis of miRNAs are well studied, the computational prediction of target genes and binding sites faces the challenge of ensuring accuracy due to false positive rates reaching 70% [159] and established predictors and databases such as TargetScan, miRanda and miRBase demonstrate poor agreement [216]. The lack of large collections of validated miRNA-mRNA interactions hampers the improvement of methods to predict these interactions.

Although transcription factors perform activity on the pre-transcriptional level and miRNAs on the post-transcriptional level, their systematics and effects exhibit a strong resemblance. Transcription factors and miRNAs are crucial components of the gene regulatory network which operate as trans-acting factors by interaction with cis-regulatory elements in the target gene [217]. The coordinated action of cell- and tissue-specific sets of transcription factors with multiple cis-regulatory elements controls development and often determines cell identity. Furthermore, many miRNAs are described as being exclusively present in specific cell types and having related functions [217]. Moreover, the 3’-UTRs of target genes are capable of possessing multiple cis-regulatory elements for distant miRNAs, indicating cluster-wise regulation and coordinated gene repression [217,218]. Notably, coding genes for transcription factors and miRNAs regulate each other in feedback and feedforward loops, hinting at their interaction in a gene regulatory network [179,217], and pairs of transcription factors and miRNAs coregulating common targets have been noted [219].

Previously, we analyzed the overlap between targets of transcription factors and targets of miRNAs for the purpose of identifying redundancy in the global regulatory network and to add support to large subsets of predicted miRNA-mRNA interactions [141]. Potential target genes for RE1 silencing transcription factor (REST) were identified by the analysis of multiple ChIP-seq datasets for diverse human cell types [141]. The selected transcription factor REST has been found to exert biological activity by regulating genes associated with abundant neuronal but also non-neuronal functions [220]. From the background of all miRNA-mRNA interactions predicted by TargetScanHuman 6.2 [215], we found 20 broadly conserved miRNA

families (REST miRNAs) whose targets were over-represented in genes potentially regulated by REST. Several of these REST miRNAs had been previously described as contributing to neural cell differentiation and tumor suppression in glioblastoma. One of the REST miRNA-mRNA interactions with the highest support, miRNA-448 with the oncogene PI3KR1, was experimentally validated in our original work [141], and recent work found reported further effects of this miRNA in the regulation of the PI3K/AKT signaling pathway through targeting of ROCK1, inhibiting the progression of retinoblastoma [221].

Under the assumption that the predicted interactions between REST miRNAs and mRNAs are more accurate than other predictions, we propose that the study of the differences between these targets and the background of all predicted miRNA-mRNA interactions will point to features characterizing real interactions. Our aim is to discover target features that could be used to improve miRNA target prediction. We particularly focus on the analysis of the properties of miRNA binding sites in the 3'-UTR, thus attempting to reveal new features for factor-associated miRNA predictions. For this purpose, we study the position of miRNA binding sites, the presence of multiple targets, and the GC-content around the seed matching sequence in the 3'-UTR. Considering that the genetic locus as well as the combinatorial activity [218] and the accessibility of cis-regulatory elements [222], play a crucial role in transcriptional regulation, we assume that these features could also be important for miRNA regulation and could contribute to the improvement of the prediction of conserved miRNA targeting.

In our study, we compute the over-representation of REST miRNAs for miRNA-mRNA interactions predicted with the most recent version of TargetScanHuman (version 7.2), which extends the previous prediction model by considering several additional features when scoring predicted interactions, including the structural accessibility of the miRNA binding site, global and local nucleotide composition and 3'-UTR length [158].

2.1.4 Results

We analyzed properties of predicted miRNA binding sites for miRNA families targeting sets of genes enriched in genes that are potentially bound by the transcription factor REST (S2 Table), in terms of their position and nucleic environment. For simplicity, hereafter we name these families as REST miRNAs, the genes predicted to be bound by REST (or their 3'-UTRs) as REST genes (or 3'-UTRs), and the binding sites pairing REST miRNAs and their targets as REST pairs. We then contrasted the outcome to predicted miRNA-mRNA interactions for all human genes, as annotated by TargetScanHuman 7.2 (see Materials and Methods section for details). Conversely, we name the set of miRNAs as TargetScanHuman (TSH) miRNAs, the 3'-UTRs/genes predicted to be bound by TSH miRNAs as TSH 3'-UTRs/genes, and the binding sites pairing TSH miRNAs and their targets as TSH pairs. Table 2.1.1 presents an overview of descriptive statistics and calculated p-values for each analysis.

Table 2.1.1: Descriptive statistics and p-values for parsed properties of REST-associated miRNA-target gene pairs (REST) and TargetScanHuman miRNA-target gene pairs (TSH) or TargetScanHuman miRNA-target gene pairs in REST 3'-UTRs (TSH-REST) (see Materials and Methods for details).

Analysis	Dataset	N	Mean	Std. dev.	p-value
3'-UTR length	REST	2791	4635 nt	3377 nt	<0.001
	TSH	12989	2482 nt	2098 nt	
Distance from 3'-UTR start to miRNA binding site	REST	17325	1979 nt	2480 nt	<0.001
	TSH	103467	1648 nt	1926 nt	
Distance miRNA binding site to 3'-UTR end	REST	17325	2650 nt	2713 nt	<0.001
	TSH	103467	2226 nt	2261 nt	
Position of miRNA binding site (relative)	REST	17325	0.437	0.325	0.012
	TSH	103467	0.444	0.323	
GC-content of 3'-UTRs	REST	2781	0.39	0.069	<0.001
	TSH	12989	0.441	0.095	
Distance between multiple miRNA binding sites	REST	2689	1886 nt	2978 nt	<0.001
	TSH (REST 3'-UTRs)	25946	1932 nt	2594 nt	<0.001
	TSH	11127	1451 nt	2014 nt	
GC-content between multiple miRNA binding sites	REST	2689	0.363	0.087	<0.001
	TSH (REST 3'-UTRs)	25946	0.391	0.094	<0.001
	TSH	11127	0.402	0.108	

2.1.4.1 3'-UTR length

Predictions for REST pairs cover 2,781 target genes and are compared with 12,989 target genes for TSH pairs. The mean length for the 3'-UTR of REST-bound genes is 4,635 nt, which is 1.87-fold greater than for TargetScanHuman genes with an average length of 2,482 nt. The calculated p-value of <0.001 indicates the statistical significance of the difference in these means (Table 2.1.1). Both sets of genes present noticeably higher mean than median values, since they exhibit numerous outliers (Figure 2.1.1A). The histograms for REST-bound genes particularly display a higher density for 3'-UTRs longer than 3,000 nt and a lower density for 3'-UTRs shorter than 2,200 nt (Figure 2.1.1AB). The plot of the distribution supports the statistical assessment that the subset of REST-bound genes with predicted miRNA binding sites have longer 3'-UTRs than the remaining annotated TargetScanHuman genes. We take this difference in consideration for the interpretation of the results of our further analyses.

2.1.4.2 Position of miRNA binding site

We measured the distance from the 3' and 5' terminal end of the 3'-UTR to the predicted miRNA binding site, as well as the relative miRNA binding site position in the 3'-UTR, for TSH pairs and REST pairs. We found that the mean of the absolute distance from the 5' terminal end to the predicted miRNA binding site is significantly higher for REST pairs, with a p-value of <0.001 (Table 2.1.1; mean distances 1,979 nt and 1,648 nt, respectively) consistently with the significantly longer length of the 3'-UTRs of REST-bound genes. The distributions, however, display a strong resemblance (Figure 2.1.1C;D).

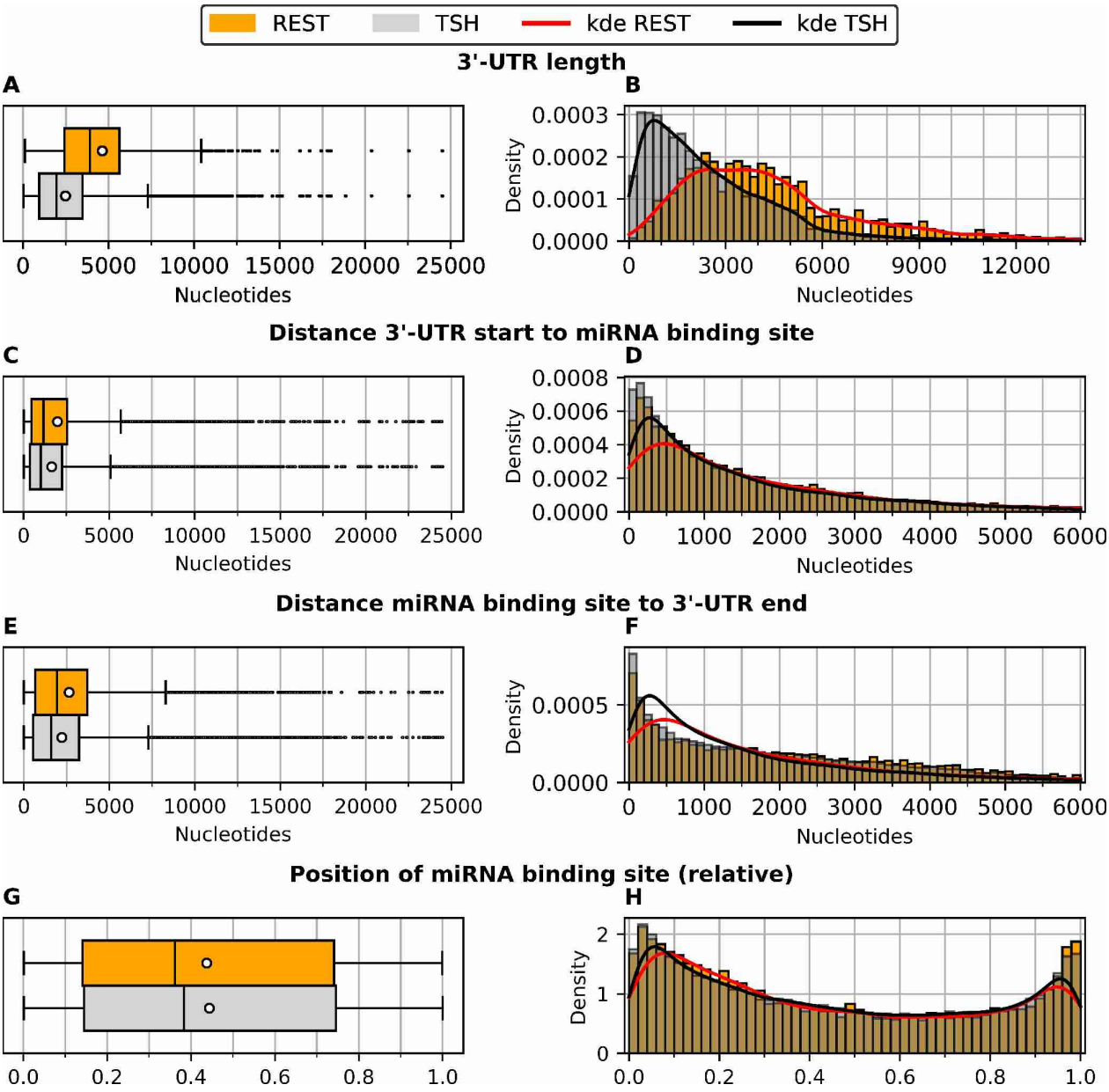


Figure 2.1.1: Properties of REST-associated miRNA-target gene pairs (REST) and TargetScanHuman miRNA-target gene pairs (TSH) or TargetScanHuman miRNA-target gene pairs in REST 3'-UTRs (TSH-REST). (A, B) Distribution of 3'-UTR length. (C, D) Distance from 3'-UTR start to miRNA binding site. (E, F) Distance from miRNA binding site to 3'-UTR end. (G, H) Relative position of miRNA binding site to the 3'-UTR length. Left side: the box plots indicate median, second and third quartile, mean (white dot) and standard deviation (whiskers). Right side: kde = kernel density of the corresponding distribution.

Similarly, the mean of the absolute distance from the predicted miRNA binding site to the 3' terminal end of the 3'-UTR is significantly higher for REST pairs, with a p-value of <0.001 (Table 2.1.1; mean distances 2,650 nt and 2,226 nt, respectively) consistently with the significantly longer length of the 3'-UTRs of REST-bound genes. Again, however, the distributions present strong similarities (Figure 2.1.1AE;F).

To obtain results independent of the length of the 3'-UTR, we calculated the position of the miRNA binding site relative to the length of the associated target 3'-UTR. The

result shows that miRNA binding sites in both REST pairs and TSH pairs are located on average closer to the 5’ terminal end of the 3’-UTR. The mean position of miRNA binding sites for REST pairs is 0.437 and for TSH pairs 0.444. The statistical test yielded a p-value of 0.012, indicating that there were no significant differences for these means (Table 2.1.1). This result is consistent with the graphical representations of the distributions, which present strong similarities (Figure 2.1.1AG;H).

All in all, miRNA binding sites for REST pairs are observed to have a greater absolute distance to the 5’ and 3’ terminal end of their 3’-UTRs, consistent with the longer length of observed REST 3’-UTRs, as well as a very similar position, relative to the 3’-UTR length.

2.1.4.3 GC-content around predicted miRNA binding sites

The GC-content around predicted miRNA binding sites was calculated for 50 nt bins in the range 500 nt before and after the site (Figure 2.1.2A). This analysis revealed a notable decrease in GC-content towards the predicted miRNA binding site for all subsets, which was more marked in the REST miRNA-target pairs than in TSH miRNA-pairs (Figure 2.1.2A; orange and gray bars, respectively). However, the differences in the leftmost and rightmost values suggested that REST 3’-UTRs have a lower GC background content and, additionally, that the GC background content has a decreasing gradient that must be appreciable in a 1000 nt region. To test this hypothesis, we computed these backgrounds using 1000 nt regions taken at random positions from all REST 3’-UTRs and from all TSH 3’-UTRs, respectively. We obtained the expected results (lower GC content in REST 3’-UTRs and decreasing values from 5’ to 3’; Figure 2.1.2B).

To test that the differences in GC content variation surrounding REST miRNA and TSH miRNA pairs are not just due to differences in 3’-UTR properties, we examined separately the GC-content surrounding TSH miRNA pairs in REST 3’-UTRs (red bars in Figure 2.1.2A) and confirmed that their decrease in GC contents is also less pronounced than that for REST miRNA pairs.

Interestingly, we observed the largest drop in GC-content in the 50 nt bin right after the miRNA binding site, which could be a property used to improve miRNA predictions.

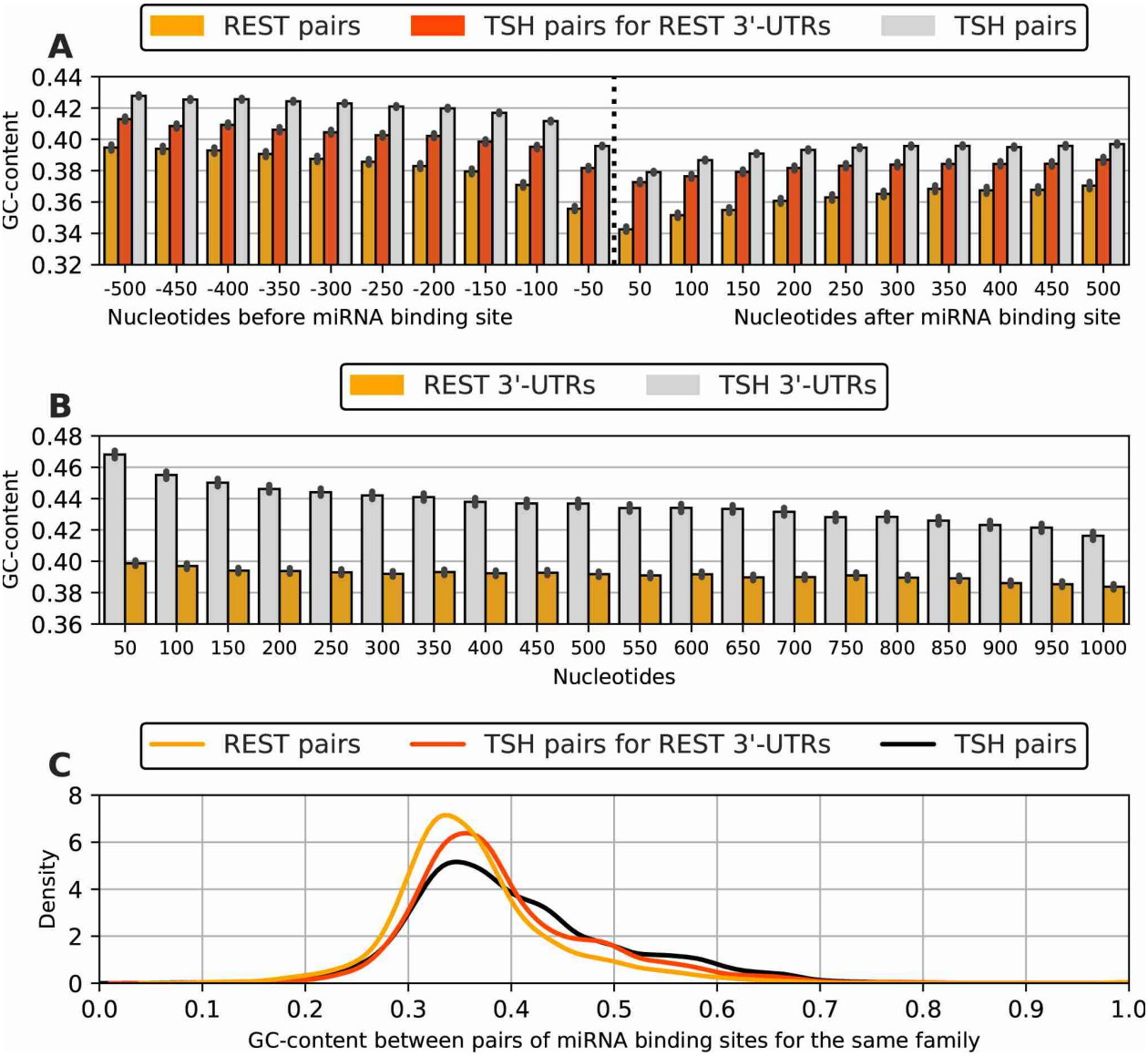


Figure 2.1.2: GC-content in the vicinity of predicted miRNA binding for subsets of miRNA-mRNA pairs. (A) GC-content relative to the distance around the miRNA binding site for REST pairs, TSH pairs and TSH pairs in the 3'-UTRs of REST regulated genes (TSH pairs in REST 3'-UTRs). (B) Background GC content for REST 3'-UTRs and TSH 3'-UTRs (see Methods for details). The error bars indicate the 99% confidence interval. (C) Kernel density estimation for the distribution of GC-content values between pairs of miRNA binding sites for the same miRNA family. Values obtained from TSH pairs, TSH pairs in REST 3'-UTRs and REST pairs.

2.1.4.4 Presence of multiple predicted binding sites of a given miRNA family

Many 3'-UTRs possess multiple miRNA binding sites for the same miRNA family. For the set of TSH pairs, we identified 7,871 miRNA-target gene combinations with multiple miRNA binding sites within the target 3'-UTR. Furthermore, we observed 1,855 out of 15,009 miRNA-target gene pairs with more than one predicted binding site of the same miRNA family in the set of predicted REST pairs. In fact, all 67 miRNA families that are enriched in the subset of REST regulated genes have multiple binding sites in the 3'-UTRs of the target genes.

p-values were calculated using the Fisher's exact test, to test for statistically significant differences in the proportions of target genes with at least one predicted binding site and target genes with multiple binding sites, between REST pairs and TSH pairs. The analysis returned a p-value of <0.001, concluding statistical significance (S4 Table). We also computed p-values for the proportion of multiple binding sites for each miRNA family enriched for REST-bound genes, compared to the observation of TSH pairs. A total of 36 out of 67 enriched miRNA families present a statistically significant difference (p-value <0.05) for the proportion of target genes with multiple miRNA binding sites for the same family, compared to background predictions for all considered miRNA-mRNA interactions. Remarkably, the 21 REST miRNA families with the largest number of associated miRNA-target gene pairs (277 or more) show significance (S4 Table). This suggests that the presence of multiple miRNA binding sites for the same family is a good predictor of miRNA binding sites.

2.1.4.5 Distance and GC-content between multiple predicted binding sites

We further examined properties from previously identified multiple binding sites for a given miRNA family in 3'-UTRs, in terms of distance and GC-content. The mean of the distance between multiple miRNA binding sites for the subset of REST pairs is 1,886 nt, whereas the mean for TSH predictions is 1,451 nt. However, this difference could be due to the longer length of REST 3'-UTRs (Figure 2.1.1A). In fact, the mean of the distance between multiple miRNA binding sites for TSH predictions in REST 3'-UTRs is 1,932 nt, which is very close to the value observed for REST pairs (Table 2.1.1).

We then computed the GC-content between multiple binding sites for the same family. The distributions of values are shown in Figure 2.1.2C. The average value is slightly lower for REST pairs than for TSH pairs (0.363 and 0.402, respectively; Table 2.1.1) and this is reflected in the distributions of values (orange and black curves, respectively; Figure 2.1.2C). This difference is not just due to the differences in GC content of the 3'-UTRs, since TSH pairs in REST 3'-UTR do have a higher average GC content of 0.391 (Table 2.1.1) and a distribution shifted to higher values (red curve; Figure 2.1.2C) compared to that of REST pairs. This confirms that lower GC content is a good predictor of miRNA binding sites, particularly between pairs of miRNA binding sites for the same family.

2.1.5 Discussion

Predictions of miRNA binding sites are not very accurate and are likely to include many false positives. Our hypothesis is that, given a subset of predictions assumed to be enriched in true positives, one could use it to compare its properties with those of the background of predictions to learn discriminant properties. For this strategy to have any chance of success, the selected subset needs to be significantly large. The largest databases that annotate experimentally supported target genes for human miRNAs, DIANA-TarBase v.8 and miRTarBase, overlap only about 10% [223]. Moreover, these databases contain indirect interactions that could originate from outside the 3' UTR, with less strong evidence from high-throughput experiments, making these databases unsuitable candidates for our study.

In this work, we take advantage of an integrative approach that selects a large subset of predictions as more likely to be true because of their enrichment in transcription factor-regulated genes. This assumption is supported by the finding that the majority of microRNA families with validated high confidence interactions from miRTarBase are enriched in their targets with at least one specific transcription factor [224]. The redundancy in transcription factors and microRNA-associated targets could then be used to select for true interactions in biologically significant pathways [224], which highlights the potential of integrating network data for the selection of predictions with less false-positives.

Here we focused our analysis on miRNA families with targets enriched for REST targets. The transcription factor REST is a repressor that could be expected to have activity correlated to some miRNA families. Indeed, activity of miR-203 and REST co-regulate gene expression related to neuronal activity [225]. Additionally, miR-26 and miR-132 have been reported to target REST and its complex by participating in networks by negative feedback loops in neural tissue and controlling neurogenesis [226,227].

Our strategy found 67 miRNA families with targets enriched in REST-regulated genes (S2 Table; see Methods and Materials for details). The targets of these 67 miRNA families in REST-regulated genes (S3 Table) constituted therefore our subset (REST pairs) to be compared to a background consisting of all other target predictions from TargetScanHuman 7.2 (TSH pairs).

Interestingly, REST pairs have more experimental support than TSH pairs (17.9% to 14.3%) according to DIANA-TarBase v.8 [156]; this difference increases when considering only the miRNAs reported in DIANA-TarBase v.8 (25.2% to 20.5%).

Since the 3’-UTR is the genetic region of the mRNA that contains cis-regulatory elements for miRNA regulation, we compared its length in the subset of putative REST-regulated genes with its length in all genes that have predicted miRNA binding sites, according to TargetScanHuman 7.2. We validated the prior observation that the 3’-UTRs of the set of REST-bound genes are significantly longer than those of the background [141] (Table 2.1.1; Figure 2.1.1A;B).

Position-specific analysis of miRNA binding sites in the 3’-UTRs of REST genes and TSH genes indicated that the predicted sites are located close to the 3’ and 5’ terminal ends, relative to the 3’-UTR length, more often than in the middle. Predictions for REST genes demonstrated a slightly higher density for the relative position of 0.8 and downstream (Figure 2.1.1H). Since for long 3’-UTR’s (>1300 nt), regions near the 3’-UTR terminals have been reported to carry more frequently conserved targeting sites [228], this finding is consistent with our assumption of a good selection of binding sites.

Moreover, we demonstrated that the subset of REST pairs has a continuous lower GC-content in the area surrounding the predicted target site, even when contrasting REST pairs with TSH pairs situated in REST 3’-UTRs. GC-poor, respectively AU-rich, 3’-UTR regions in close vicinity to miRNA binding sites have been described as correlating with target efficiency in multiple ways, such as by destabilizing mRNA and impeding the formation of stable and functional secondary structures, thus,

providing accessible miRNA binding sites [228–230]. We conclude that the even lower GC content of REST pairs is consistent with our expectation that their predictions are more accurate than those of all TSH pairs.

Our results also revealed that REST pairs include significantly more target genes with multiple miRNA binding sites for a particular miRNA family than TSH pairs, as a proportion of the sum of target genes, suggesting that this feature is predictive of miRNA target sites. Multiple binding sites for the same miRNA might provide resistance against changes in the environment and accessibility, thus ensuring regulatory efficiency. The observation that 21 miRNA families, with the most predicted miRNA-mRNA interactions for REST pairs, display statistically significantly more target genes with multiple miRNA binding sites than the background hints at redundancy of the gene regulatory network and supports our assumption that our selection of transcription factor associated miRNA families and related predictions possess less frequent false positive predictions.

The analysis of GC-content between multiple miRNA binding sites provides another feature that separates REST pairs. We found lower GC-content between multiple miRNA binding sites for REST pairs than for TSH pairs and TSH predictions in REST 3’-UTRs (Figure 2.1.2C). Lower GC content might enable RNA-protein interaction by preventing stable secondary structure and this property can be taken as indicating good target predictions.

Our approach has revealed properties for miRNA families that are enriched in a subset of genes bound by the transcription factor REST, which indicate regulatory network interaction and clustered gene repression on the post-transcriptional level by miRNAs. To the best of our knowledge, our work represents the first attempt that selected large subsets of miRNA targets of different quality, based on the integration of miRNA-target relations with data from the network of transcriptional regulation, to collect features for the prediction of miRNA targets.

Our study has a number of limitations, including potential bias in the predictions used, and that our exploration used only ChIP-seq data regarding REST targets. Considering possible expansions of our approach, it is worth noting that we were able to provide a statistical assessment of significance given the relatively large number of genes targeted by REST. Using our approach with ChIP-seq or any other type of DNA-binding data for other transcriptional regulators will possibly only work for factors that regulate as many genes as REST does; these are not abundant. This means that extending the type of integrating method proposed here will need to add complexity, for example by pooling data for multiple factors and/or considering other indirect regulatory connections.

The reward of testing further network-based selections of miRNA targets is that our results could receive further support if the above-mentioned characteristics of GC-content and miRNA binding sites were detected for further transcription factors or network contexts. Network-based selection of miRNA-mRNA pairs can potentially provide further features to improve the algorithms used in miRNA prediction tools to ensure identification of conserved miRNA targeting.

2.1.6 Materials and methods

2.1.6.1 Datasets

2.1.6.1.1 Human 3’-UTR sequences

TargetScanHuman 7.2 provides sequences for representative human 3’-UTRs based on GENCODE annotations with most 3P-seq tags. A total of 12,989 human 3’-UTRs were considered in the analysis.

2.1.6.1.2 miRNA binding site predictions

miRNA binding site predictions for annotated human 3’-UTRs were obtained from TargetScanHuman 7.2. The predictions were based on finding complementary conserved mRNA sequences to the seed region of miRNAs (2–8 nt) and were ranked by the integration of further criteria [158]. To minimize biases in the predictions of TargetScan that could affect our analyses, we considered only predictions for broadly conserved miRNA families (conserved across most vertebrates) with miRNA targets conserved between human and mouse. The outcome comprises 219 broadly conserved miRNA families and 109,249 unique miRNA-target gene pairs for 120,702 predicted miRNA binding sites in human 3’-UTRs. TargetScanHuman 7.2 predictions and 3’ UTR sequences are publicly available and can be downloaded from http://www.targetscan.org/vert_72.

2.1.6.1.3 REST target genes

We previously assigned target genes to the repressor REST by analyzing ChIP-seq datasets of 15 different cell types, including both neural and non-neural. In total 12,344 genes that are potentially regulated by REST were identified [141].

2.1.6.1.4 Over-represented miRNA families for REST target genes

To determine the overlap between targets of miRNAs and transcription factors, we calculated the over-representation of broadly conserved miRNA families for the subset of REST-bound genes, from the background of all TargetScanHuman genes with predicted miRNA binding sites as previously described [141]. Briefly, for one ChIPseq dataset, given n REST-bound genes and m of them predicted to be target of a particular miRNA A, we randomly take n genes from the set of all genes with predicted TargetScan miRNA targets 10,000 times and count for the number of targets for miRNA A (z). To correct for the fact that REST-bound genes could have a higher tendency to have miRNA targets (e.g. due to longer 3’UTRs) we compute a factor (r) to correct z, which is the ratio between the number of all miRNA targets found in the n REST-bound genes and the number of all miRNA targets found in the random set of n genes. Then we multiply z by r to obtain the corrected value z*. This is repeated 10,000 times and we count how many times z* is smaller than m. The number of positive tests divided by the number of tests (10,000) is then taken as p-value of enrichment of miRNA A targets in the REST-bound genes. Computed p-values were corrected for multiple testing using the Benjamini and Hochberg method (S1 Table). The significance level for adjusted p-values was set to 0.05. The analysis resulted in 67 miRNA families with a number of predicted miRNA binding sites in the subset of potentially REST-regulated genes that was significantly higher than the background (S2 Table).

2.1.6.1.5 Sets of miRNA-target gene pairs

For the purpose of studying the properties of miRNA binding sites in factor-bound genes, two sets of miRNA-target gene pairs were analyzed further. The first set comprised over-represented miRNA families for potentially REST-regulated genes that contain predicted miRNA binding sites according to TargetScanHuman 7.2; this includes 15,009 unique REST-associated miRNA-target gene pairs (REST pairs) (S3 Table). The second set covered 94,240 unique predictions for TargetScanHuman miRNA-target gene pairs (TargetScanHuman pairs), after excluding predictions of the first set from the background of all considered miRNA predictions listed in TargetScanHuman 7.2.

2.1.6.2 Statistics

In order to test statistical significance regarding the difference between the proportion of target genes with one and multiple predicted binding sites between TargetScanHuman pairs and REST pairs, p-values were calculated using the Fisher's exact test [231].

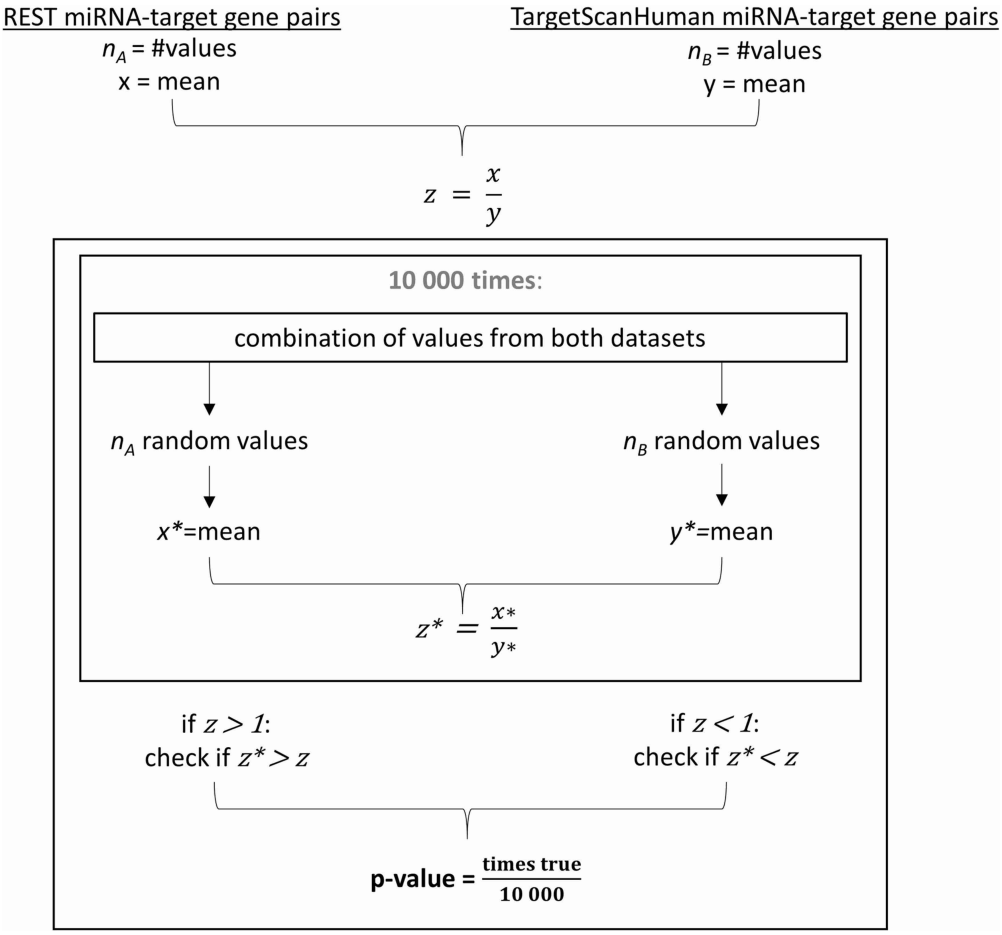


Figure 2.1.3: Illustration of p-value calculation by conducting 10,000 random tests for the statistical comparison of two sample means.

We also evaluated the significance of differences in several properties between REST miRNA-target gene pairs (REST pairs in REST-bound genes) and TargetScanHuman miRNA-target gene pairs (REST pairs in all genes considered in TargetScanHuman). These properties were distance from 3'UTR-start to miRNA binding site, distance from 3'UTR-end to miRNA binding site and relative position of the miRNA binding site in the 3'UTR. Statistical significance in terms of p-values was computed as described in the next paragraph (illustrated in Figure 2.1.3).

For each group of miRNA-target gene pairs to be compared (REST miRNA-target gene pairs and TargetScanHuman miRNA-target gene pairs), we obtained nA and nB values (for measured distances or GC-content), respectively. Means of the values (x and y, respectively) were calculated to produce the ratio value z = x / y. To produce random test ratios, we picked nA and nB values at random 10,000 times from the corresponding complete datasets of targets (all miRNA-targets in REST genes and all miRNA-targets in TargetScanHuman genes), without replacement. Means of the sampling values (x* and y*, respectively) were calculated to produce the random test ratio z* = x* / y*. Next, we examined whether the deviation from one was greater for z* than for z. For $z > 1$ we checked whether $z^* > z$ and took the number of successful tests divided by 10,000 as the p-value. In the case of $z < 1$ we defined a successful test as $z^* < z$ and calculated the p-value identically. The significance level was set to 0.05.

2.2 Unveiling cell-type-specific microRNA networks through alternative polyadenylation in glioblastoma

2.2.1 Preamble

This chapter is published in the journal BMC Biology:

Cihan, M., Schmauck, G., Sprang, M., & Andrade-Navarro, M. A. (2025). Unveiling cell-type-specific microRNA networks through alternative polyadenylation in glioblastoma. BMC biology, 23(1), 15. https://doi.org/10.1186/s12915-024-02104-8

The supplementary files associated with this publication are available on the publisher's website via the article's DOI.

2.2.2 Abstract

2.2.2.1 Background

Glioblastoma multiforme (GBM) is characterized by its cellular complexity, with a microenvironment consisting of diverse cell types, including oligodendrocyte precursor cells (OPCs) and neoplastic CD133 + radial glia-like cells. This study focuses on exploring the distinct cellular transitions in GBM, emphasizing the role of alternative polyadenylation (APA) in modulating microRNA-binding and post-transcriptional regulation.

2.2.2.2 Results

Our research identified unique APA profiles that signify the transitional phases between neoplastic cells and OPCs, underscoring the importance of APA in cellular identity and transformation in GBM. A significant finding was the disconnection between differential APA events and gene expression alterations, indicating that APA operates as an independent regulatory mechanism. We also highlighted the specific genes in neoplastic cells and OPCs that lose microRNA-binding sites due to APA, which are crucial for maintaining stem cell characteristics and DNA repair, respectively. The constructed networks of microRNA-transcription factor-target genes provide insights into the cellular mechanisms influencing cancer cell survival and therapeutic resistance.

2.2.2.3 Conclusions

This study elucidates the APA-driven regulatory framework within GBM, spotlighting its influence on cell state transitions and microRNA network dynamics. Our comprehensive analysis using single-cell RNA sequencing data to investigate the microRNA-binding sites altered by APA profiles offers a robust foundation for future research, presenting a novel approach to understanding and potentially targeting the complex molecular interplay in GBM.

2.2.3 Background

Glioblastoma multiforme (GBM) is recognized as the most aggressive and prevalent primary brain tumor in adults, posing significant challenges in neuro-oncology [97]. A key aspect in understanding the pathogenesis of GBM is identifying its cell of origin.

The observed heterogeneity in GBM points to various theories, including the prominence of glial progenitor cells and the stem cell hypothesis [232–234]. The latter suggests a subpopulation of stem cells within tumors that are characterized by

their ability to propagate tumors, self-renew, differentiate into multiple lineages, express specific markers, exhibit low frequency, and resist drugs [235–238].

Recent research has revealed the presence of a subset of CD133 + radial glia-like cells in adult human glioblastomas, which exhibit characteristics similar to normal human fetal radial glia. These cells exist in various states, ranging from dormancy to active cycling, highlighting their significant role in the dynamics of tumor growth and maintenance [239].

Additionally, CD133 + cells, known for expressing genes associated with radial glial and neural crest cell development, have been implicated in the seeding of recurrent GBM tumors. These recurrent tumors display a diverse array of properties, including both neural and mesenchymal traits [240,241].

The inherent capacity of radial glia-like cells to transdifferentiate demonstrates the plasticity and adaptability of tumor cells to microenvironmental cues, which complicates the dynamics of GBM progression and therapy resistance.

Experimental evidence has further elucidated the potential of these radial glia-like cells to differentiate into oligodendrocyte precursor cells (OPCs). Specifically, radial glia-like cells derived from human pluripotent stem cells have shown an enhanced capacity for oligodendrocyte lineage commitment, especially under the influence of specific growth factors, highlighting their adaptability and importance in neurogenesis [242]. Additionally, it has been found that pre-OPCs express neurogenic outer radial glia cell markers, indicating a lineage relationship and suggesting a complex interplay in the development of neural cell types [243,244].

Within the dynamic landscape of the GBM tumor microenvironment, especially at the tumor borders, OPCs assume a critical role. These cells significantly contribute to the cellular architecture and intricacies of the tumor periphery, thereby influencing the neoplastic trajectory. OPCs, in concert with macrophages, construct a specialized microenvironmental niche that perpetuates the stemness of GBM cells and their resilience against therapeutic interventions [112,113].

The cellular and molecular heterogeneity of GBM underscores the importance of gene regulation in its progression. Notably, microRNAs play a significant role in oligodendrocyte differentiation [245] and within the tumor microenvironment [246].

MicroRNAs like miR-219-5p, miR-219–2-3p, and miR-338-3p show heightened expression at the tumor fringes, indicating an abundant presence of oligodendrocyte lineage cells in these regions, a strong contrast to their sparse distribution within the central tumor mass [98,112]. The role of these microRNAs is multifaceted and extends into various biological processes; they actively modulate the immune response, reshape the epigenetic environment, and alter the dynamics of GBM subpopulations. This broad spectrum of activity affects everything from cell proliferation to programmed cell death, significantly impacting tumor behavior and patient outcomes [100,247–249].

An additional critical aspect that must be considered to understand the mechanisms by which microRNAs regulate cellular transitions in GBM is the role of microRNA-binding through alternative polyadenylation (APA) modification of the 3' untranslated

regions (UTRs) of target mRNAs. APA can alter microRNA-binding sites by shortening or extending the UTRs [250]. In cancer cells, the shortening of 3' UTRs compared to normal tissue may enhance mRNA stability and subsequently affect microRNA-mediated gene regulation [117].

However, a comprehensive understanding of regulatory mechanisms controlling microRNA binding cannot be explained by APA alone and needs to consider the interplay with other regulatory elements. This requirement for precision often leads to functional synergy with transcription factors, orchestrating a layered regulatory network [141,251]. This complex interplay can be studied as the body of feed-forward loops (FFLs), sophisticated networks composed of microRNAs, transcription factors (TFs), and their target genes. FFLs play a crucial role in cellular proliferation, tumor formation, inhibition of cellular aging processes, and are used for classification of subtypes in oncological research [177,252–255].

In this study, we investigate the complex regulatory framework of GBM, emphasizing the intricate dynamics between microRNAs, APA, and FFLs. We seek to illuminate the role of microRNAs in influencing the fate of neoplastic radial-glia-like cells and their potential transdifferentiation to OPCs, employing single-cell RNA sequencing data. By examining APA dynamics and the shifts in microRNA-binding sites, we aim to understand how these molecular alterations impact cell fate decisions and transitions in GBM. We particularly focus on building cell-type-specific FFL networks, which will provide a detailed view of the co-regulatory interactions involving microRNAs and their contribution to the complex cellular ecosystem of GBM. Our analysis includes constructing pseudotime trajectories to model the progression paths of individual cell clusters within the tumor. This approach aims to unravel the subtleties of APA dynamics, cellular shifts, and the broader implications for cell fate and tumor heterogeneity in GBM.

2.2.4 Results

2.2.4.1 APA reveals an additional regulatory layer in GBM cellular heterogeneity

In order to study the role of APA in gene regulatory networks, including its dysregulation in a cancer setup, we chose to investigate single-cell RNA samples from GBM. Batch effect correction was critically applied to single-cell RNA sequencing data from three distinct GBM samples: GBM27, GBM28, and GBM29 (see “ Methods” for details). As delineated in Figure 2.2.1a, this methodological step was imperative for normalizing across disparate samples, thereby ensuring a robust integration and comparative analysis of cell populations. Cell clustering based on gene expression profiles coupled with literature-derived cell type annotations resulted in the identification of distinct cellular contingents within the GBM milieu, specifically neoplastic cells, OPCs, endothelial cells, macrophages, and oligodendrocytes, as depicted in Figure 2.2.1b (Additional file 1: Table S1). Notably, a substantial proportion of the cells were categorized as OPCs or neoplastic cells. To quantitatively assess the proximity of neoplastic cells to OPCs, we calculated the mean Euclidean distance from the neoplastic centroid to each cell in target clusters in PCA space. The results showed that the neoplastic-OPC distance (mean: 23.96, standard deviation [SD]: 7.12) was notably shorter than distances to other clusters: macrophages (mean: 50.15, SD: 10.9), endothelial cells (mean: 49.99, SD: 11.0), and oligodendrocytes (mean: 42.0, SD: 16.5). This marked reduction in distance

highlights a distinct spatial alignment between neoplastic cells and OPCs, reinforcing a potential transitional relationship or shared lineage characteristics between these cell types.

In an innovative twist to traditional clustering approaches, cells were also clustered based on their APA profiles and colored in the gene expression-based UMAP. This analysis revealed an APA cluster 1, intriguingly interspersed between the OPC and neoplastic cells, as shown in Figure 2.2.1c and further supported in Figure 2.2.1e. This interposition suggests an overlap in APA patterns between these cells, despite their distinct placements when clustered by gene expression. While the majority of genes in the dataset exhibited a singular APA site, a total of 16.6% of genes featured multiple APA sites (Figure 2.2.1d). To validate the APA events identified, we compared genes with multiple APA sites to PolyASite v2.0 [256] and PolyA_DB 3 [257]. Our results showed high overlap: 89.22% of genes with multiple APA sites in our dataset were also present in PolyASite, and 91.49% were found in PolyA_DB. Notably, only 142 genes were not captured by either database. These findings validate the robustness of our analysis and highlight the near-comprehensive coverage provided by these databases. This revelation hints at a layer of post-transcriptional regulation that warrants further investigation.

Upon examining the relationship between differential gene expression and APA site variability, it became evident that most genes that are differentially expressed are not subject to differential APA, and conversely, genes with differential APA are not predominantly differentially expressed (Figure 2.2.1f). This apparent lack of overlap underlines a critical nuance; APA does not directly mirror gene expression changes but rather provides an additional layer of information, possibly affecting gene expression stability, localization, and protein translation efficiency. This realization emphasizes APA as a distinct, regulatory axis that complements traditional gene expression analysis.

Focusing on differential APA events among neoplastic, OPC, and oligodendrocyte clusters, it was observed that significant differential APA events do not necessarily correlate with large changes in gene expression in the clusters of OPC vs neoplastic, as exemplified in Figure 2.2.1g (log2 fold change). This observation is pivotal, suggesting that while APA might not always dramatically shift gene expression levels, it could still be critically modulating gene function and cell state in a more subtle or context-dependent manner. Differential APA events between OPC and neoplastic cells were particularly observed for RPS3, DVL3, DEF8, EGFR, OLFM1, and GRB2.

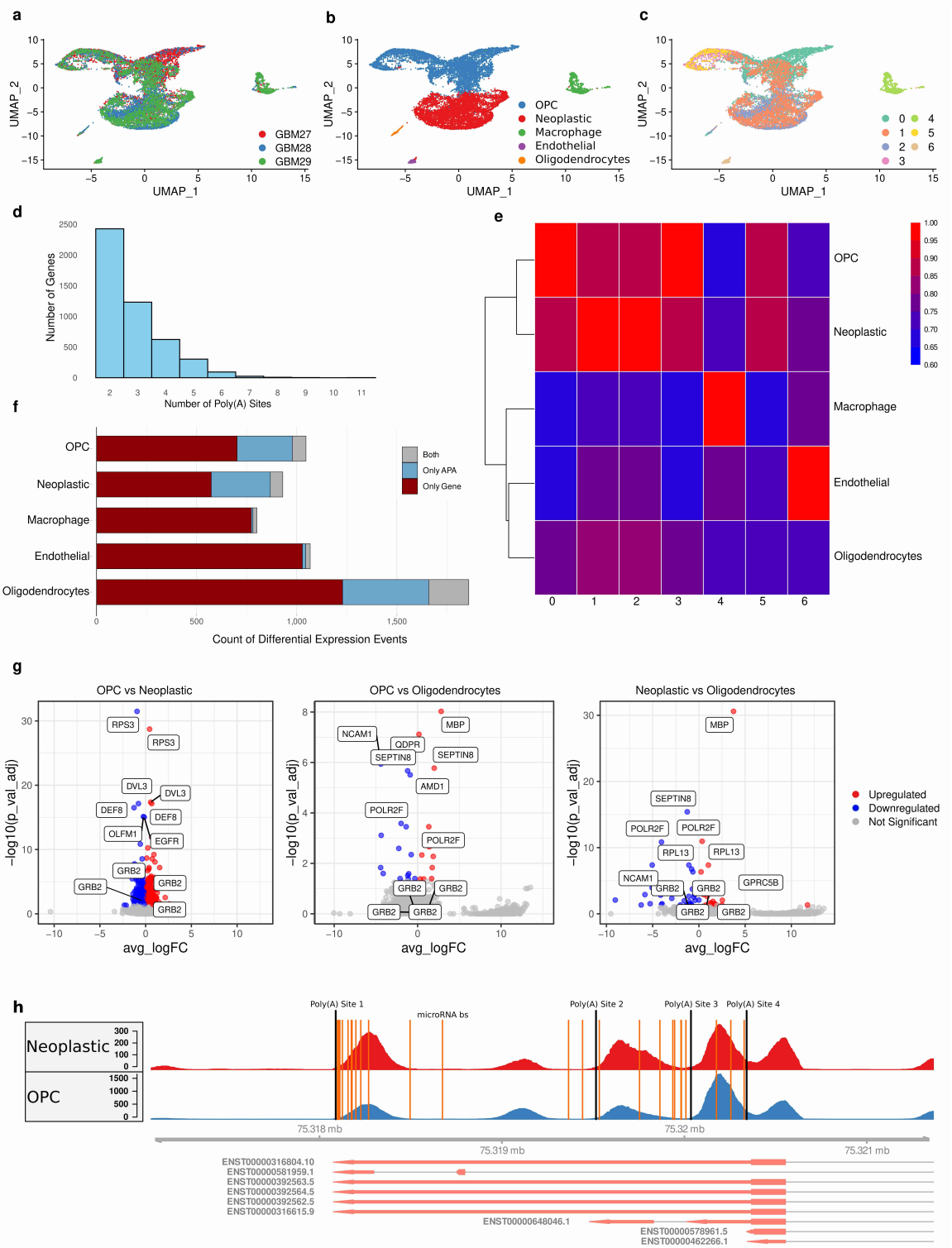


Figure 2.2.1: Single-cell clustering, APA events, and differential expression analysis. a UMAP plot showing single-cell clustering post batch effect correction, colored by sample origin. b UMAP plot depicting cell clustering results, colored by annotated cell types. c UMAP visualization of cell clusters and colored based on clustering results applied on APA matrix. d Histogram indicating the number of APA sites per gene. e Heatmap showing the overlap between APA and gene expression clusters. f Bar chart comparing the total differential expression event counts for each cell cluster vs. all other cell clusters. g Volcano plots showing differential gene expression differences between cell types. h Visualization of APA site variation for gene GBR2 across OPCs and neoplastic cells generated from BAM files.

The differential usage of the APA site 1 in GRB2, more frequently used in OPC than in neoplastic cells as evidenced in Figure 2.2.1h, highlights APA's potential impact on microRNA regulation. This trend might suggest a mechanism where shorter transcripts in OPCs resulting from preferential APA site usage reduce microRNA-binding opportunities, reflecting the interaction between APA and post-transcriptional regulation within GBM's cellular diversity. To further explore the role of APA-regulatory genes in GBM cellular heterogeneity, we analyzed genes annotated under the Gene Ontology terms GO:0031124 (mRNA 3' processing) and GO:0110104 (mRNA alternative polyadenylation) to determine whether these genes exhibited significant differential APA events. Among the 29 identified APA-regulatory genes, 26 significant differential APA events were detected across 10 genes. These events were distributed across cell clusters, with 9 observed in the OPC cluster, 4 in the neoplastic cluster, 12 in the macrophage cluster, and 1 in the endothelial cluster (Additional file 2: Table S1-S6). Interestingly, only two of the APA regulatory genes, CPSF4 and ZC3H3, were among the highly variable genes used in the original PCA for clustering. However, none of these genes exhibited significant differential APA site usage, suggesting that their inclusion in the PCA was driven solely by expression variability rather than APA dynamics. This lack of representation among highly variable genes highlights a limited contribution of APA regulatory genes to the overall clustering process.

2.2.4.2 Coordinated microRNA-binding site avoidance by APA highlights cell-specific regulatory strategies in GBM

Investigation into the loss of microRNA-binding sites due to APA uncovers a sophisticated landscape of evasion strategies within the GBM tumor microenvironment. Analysis of cell-specific microRNA avoidance reveals that OPCs and neoplastic cells selectively evade distinct sets of microRNA families (Figure 2.2.2a). For OPCs, out of the 912 microRNA families analyzed, 334 were avoided significantly by APA, whereas neoplastic cells show a significant avoidance in 192 microRNA families. This targeted avoidance is exclusive to each cell type, with no microRNA family found to be commonly avoided across both OPCs and neoplastic cells. However, both OPCs and neoplastic cells share significantly avoided microRNA families with other cell types (Figure 2.2.2a). Such specificity in the regulatory landscape suggests that it may contribute to a complex adaptive mechanism of post-transcriptional regulation orchestrated by APA, which may underpin the cellular heterogeneity observed in GBM. This avoidance is not merely a binary event; several genes are involved in avoiding microRNA-binding sites across multiple clusters. Specifically, 773 genes are involved in avoiding microRNA-binding sites across both OPC and neoplastic clusters exclusively, suggesting a subset of regulatory processes that are critical to both cell types (Figure 2.2.2b). In contrast, genes exclusively avoiding significant microRNA families in OPCs number 1404, whereas neoplastic cells have a unique set of 196 genes engaged in such avoidance.

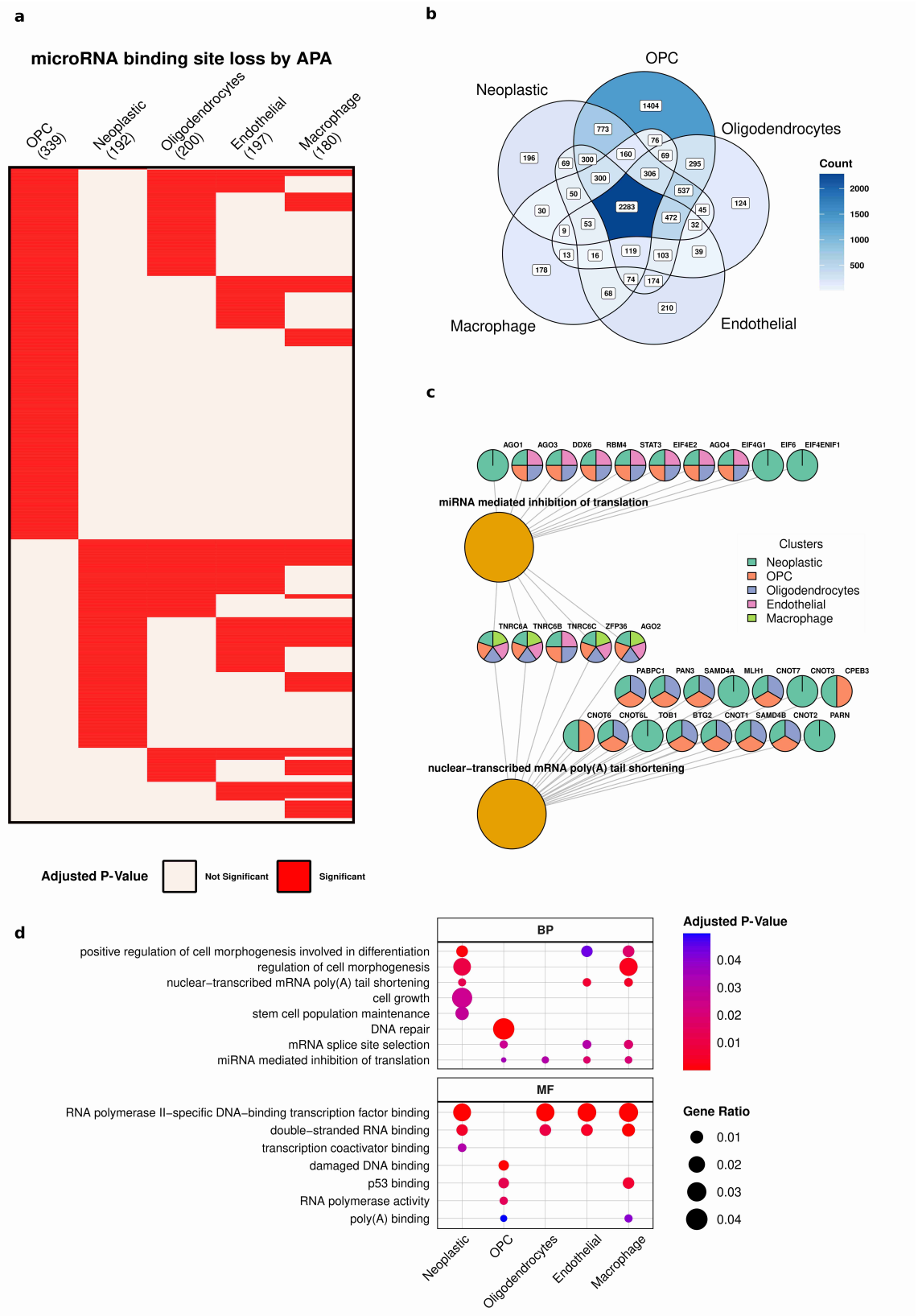


Figure 2.2.2: MicroRNA-binding site loss and GO term enrichment of APA events. a Heatmap displaying microRNA families with statistically significant loss of binding sites due to APA across cell clusters.

Figure 2.2.2 (continued): b Venn diagram presenting the overlap of genes losing microRNA-binding sites in each cell cluster. c Network of GO terms and respective genes illustrated for different cell clusters. d Manually curated GO term enrichment results highlighting significantly enriched biological processes and molecular functions in different clusters

The GO term analysis for these genes reveals a profound involvement in key processes and functions. In neoplastic cells, genes associated with “positive regulation of cell morphogenesis involved in differentiation” and “regulation of cell morphogenesis” present significant APA events. This suggests an APA-mediated emphasis on maintaining differentiation potential within neoplastic cells. Furthermore, “nuclear-transcribed mRNA poly(A) tail shortening” is a notable process, with 18 genes involved, which may indicate APA’s significant role in the regulation of mRNA stability and turnover, essential for the dynamic cellular states in neoplastic cells (Figure 2.2.2c;d). Among the genes associated with this GO term, MLH1, CNOT3, TOB1, and PARN are exclusively associated with neoplastic cells.

Furthermore, in neoplastic cells, the significant enrichment of genes in processes such as “cell growth” and “stem cell population maintenance” points to a strategic use of APA in supporting the malignant phenotype (Figure 2.2.2d). These genes are crucial in sustaining proliferative capacity and adaptability, with APA potentially acting as a regulatory buffer to mitigate the effects of oncogenic stress on these cells.

The molecular function analysis aligns with these biological processes. For neoplastic clusters, the significant evasion of microRNA regulation by APA is linked to “RNA polymerase activity,” “double-stranded RNA binding,” and “transcription coactivator binding.” These significant ratios indicate APA’s potential influence on transcriptional regulation, with possible implications for cellular identity and response to environmental cues. Moreover, functions like “damaged DNA binding” and “p53 binding” for OPC suggest a role for APA in modulating the DNA damage response and interactions with tumor suppressor networks (Figure 2.2.2d).

2.2.4.3 APA-mediated microRNA regulation in cell-type-specific networks and feed-forward loops

To further investigate the influence of microRNA regulation and APA on GBM, we delved into the intricacies of regulatory networks, focusing on simple feed-forward loops (SFFLs) and module feed-forward loops (MFFLs). SFFLs represent regulatory motifs that consist of three-node interactions among microRNA, TF, and gene, where the gene is regulated by both the TF and the microRNA, with additional interactions involving TF regulating microRNA, microRNA regulating TF, or both. MFFLs extend this concept by incorporating multiple such SFFLs into a larger, interconnected network, enabling a more complex and robust regulatory framework (see “ Methods” for details). This exploration is particularly pertinent given the observed enrichment of “transcription coactivator binding” in the GO terms, suggesting a complex interplay of transcriptional regulation within the tumor's cellular milieu.

In the explored GBM dataset, our analysis revealed a total of 3558 significant SFFLs. The neoplastic cluster accounted for the highest number with 382 SFFLs, followed by the OPC cluster with 277. Within these significant networks, 83 SFFLs in the OPC and 30 in the neoplastic cluster were associated with microRNA-binding site loss due to APA. The other cell types—oligodendrocytes, endothelial, and macrophages—

presented 62, 37, and 198 SFFLs, respectively, reflecting their unique contributions to the complex regulatory milieu of GBM.

Further dissecting the specificity of SFFLs within GBM, the upset plot analysis identified cell-type-specific loops: neoplastic cells exhibited 168 unique SFFLs, OPCs had 86, and macrophages showed 87. Additionally, an overlap was observed with 82 SFFLs shared between neoplastic and OPC cells, indicative of potential regulatory intersections between these cell types (Figure 2.2.3a).

The analysis of MFFLs revealed 91 networks in the neoplastic cluster, the highest number across all cell types, and 77 in the OPC cluster. Neoplastic cells had a slightly lower count of central TF node MFFLs at 15, compared to 16 in the OPC cluster. In contrast, the OPC cluster showed a significant occurrence of microRNA-binding site loss due to APA in 41 microRNA-centric MFFLs. The counts for oligodendrocytes, endothelial, and macrophages were 19, 9, and 57 MFFLs, respectively, each with their own set of MFFLs affected by APA, illustrating the cell-specific regulatory strategies within the GBM tumor microenvironment (Figure 2.2.3b).

Pathway profiling of OPC and neoplastic cell clusters, derived from MFFL analysis, reveals distinct biological pathways relevant to each cluster's role in GBM. OPCs are associated with neurodegenerative diseases and glioblastoma signaling pathways, highlighting their versatile nature. The BDNF signaling pathway's presence in OPCs may be pivotal for neuronal-like functions within the tumor milieu.

Neoplastic cells are involved in pathways like nervous system development and axon guidance, aligning with their invasive characteristics. The presence of the defective intrinsic pathway for apoptosis in neoplastic cells aligns with cancer's typical evasion of programmed cell death, aiding in tumor survival (Figure 2.2.3c).

To explore the potential prognostic relevance of MFFLs across different cell clusters, Kaplan–Meier survival analysis was conducted. Cox regression analysis was performed to account for potential confounding factors, including age, tumor immune dysfunction, and exclusion, with these variables included as covariates. This comprehensive analysis revealed seven MFFLs with p-values less than 0.05, underscoring their potential prognostic relevance in GBM. Among these, five belonged to the OPC cluster, one to the neoplastic cluster, and one to the macrophage cluster (Additional file 1: Table S4). Notably, the MFFL involving MIR499A, YY1, CBX5, and DDX39B in the OPC cluster not only demonstrated statistical significance with the lowest p-value of 0.017 (Figure 2.2.3d) but also further distinguished by the loss of binding sites for the respective microRNA family due to APA in the OPC cluster exclusively, as previously identified in our cluster-specific analysis.

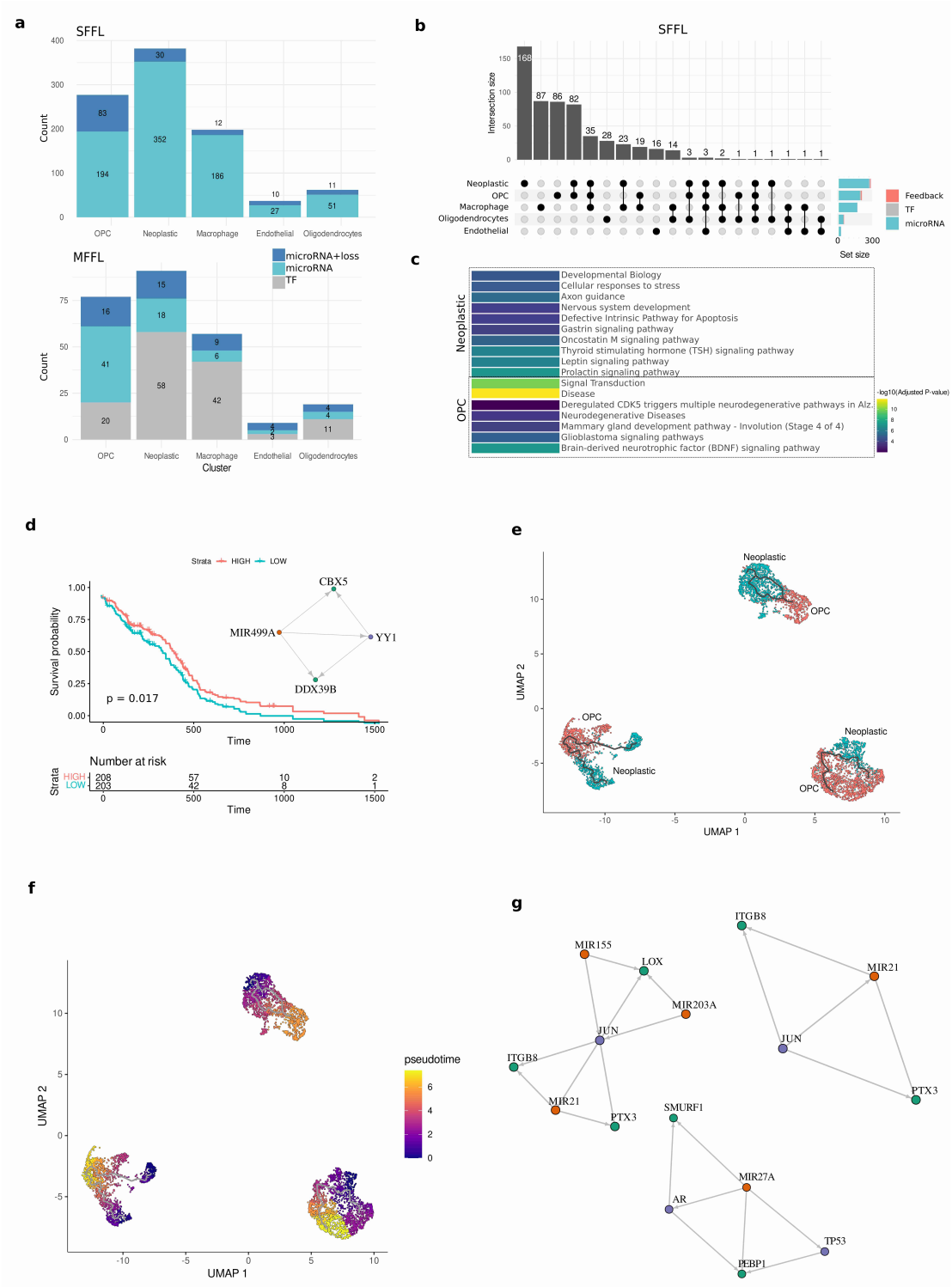


Figure 2.2.3: Network analysis and visualization of pseudotime trajectory. a Simple feed-forward loops (SFFLs) of microRNAs, transcription factors and target genes, as well as merged networks (MFFLs) for the same center node and their distribution among cell clusters. Number of microRNA families with significant loss of binding sites for respective cell cluster is indicated.

Figure 2.2.3 (continued): b Upset plot showing the overlap of significant SFFLs among cell clusters. c Manually curated list of enriched pathways for genes and transcription factors involved in MFFLs in OPC and neoplastic clusters, computed with Reactome and Wiki pathways. d Kaplan–Meier plot based on GBM-TCGA data. e UMAP of filtered neoplastic cells and OPCs f Pseudotime trajectory analysis for neoplastic cell to OPC differentiation. g MFFL networks of microRNA, TF, and target gene containing genes involved in differential APA events and differential gene expression derived from pseudotime analysis and significant for neoplastic cells

2.2.4.4 Impact of APA and co-regulatory networks on neoplastic cell transdifferentiation to OPC in pseudotime analysis

To decipher the role of APA in the transdifferentiation of neoplastic cells to OPC and to understand the impact of co-regulatory networks, SFFLs and MFFLs, we embarked on a detailed modeling of the pseudotime trajectory. This trajectory encompasses cells positioned intermediately between neoplastic and OPC cell clusters, representing a dynamic spectrum of transdifferentiation.

The pseudotime trajectory analysis incorporated 5429 cells, with 2461 belonging to the neoplastic and 2968 to OPC cluster. UMAP analysis, based on gene expression profiles, delineated these into 3 distinct clusters consisting of 2051, 1716, and 1662 cells, respectively, each embodying a blend of neoplastic and OPC cells (Figure 2.2.3e;f). This expression-based clustering was chosen to maintain consistency with the initial identification of neoplastic and OPC clusters and to track changes in gene expression patterns during the transitional phase. Root nodes, representing the origination points of differentiation to OPCs, were selected among the possible nodes provided by Monocle 3 as those most distant from OPC clusters (Figure 2.2.3f).

A comprehensive differential expression analysis over pseudotime yielded 3974 significant differential expression events (q-value < 0.05, average log expression > 0.1). Among the genes identified as differentially expressed, 189 genes showcased significant differential APA events (adjusted p-value < 0.05, average log expression > 0.1) when juxtaposing OPC versus neoplastic cells. Differential APA events can alter regulatory elements, such as microRNA-binding sites, potentially driving or modulating the observed differential expression events. This underscores the intricate interplay between APA and gene expression, highlighting APA's pivotal role in shaping regulatory landscapes during cell state transitions. These overlapping genes, characterized by differential alternative polyadenylation and expression, play crucial roles in significant SFFLs. Specifically, 23 SFFLs reported as significant in the neoplastic cluster and those reported for the OPC cluster are involved. Among these SFFLs, eight have been identified as subject to coordinated microRNA family regulation in the OPC cluster and four in the neoplastic cluster, underscoring the nuanced regulatory landscape (Table 2.2.1).

Additionally, these pivotal genes are components of two significant MFFLs in the OPC cluster (one TF-centric and one microRNA-centric) and seven significant MFFLs (four TF-centric and two microRNA-centric) in the neoplastic cluster (Table 2.2.2). This highlights the complex and integrated regulatory networks at play, driving the nuanced transitions and cellular dynamics inherent in GBM progression.

Table 2.2.1: Significant SFFLs comprising genes with differential expression in the pseudotime trajectory between neoplastic and OPC clusters and concurrent differential APA events.

Cluster Type	microRNA Family	Transcription Factor	Target Gene
Neoplastic	MIR153-1/MIR153-2	SP3	IRS2
Neoplastic	MIR155	JUN	LOX
Neoplastic	MIR155	STAT1	IL6ST
Neoplastic	MIR32	SP3	IRS2
OPC	MIR101-1/MIR101-2	CREB1	PEB1
OPC	MIR143	AR	PEB1
OPC	MIR214	TP53	PEB1
OPC	MIR27A	AR	PEB1
OPC	MIR27A	TP53	PEB1
OPC	MIR194-2	TP53	PEB1
OPC	MIR214	JUN	ITGB8
OPC	MIR214	JUN	PTX3

Table 2.2.2: Significant MFFLs comprising genes with differential expression in the pseudotime trajectory between neoplastic and OPC clusters and concurrent differential APA events.

Cluster	MicroRNA(s)	Transcription factors /Target genes	Center node
Neoplastic	MIR155, MIR203A, MIR21	JUN, LOX, ITGB8, PTX3	TF
Neoplastic	MIR128-2, MIR129-1, MIR129-2, MIR140, MIR150, MIR155, MIR194-2, MIR223, MIR338, MIR494	STAT1, FAM172A, IL6ST	TF
Neoplastic	MIR101-1	BACH1, KLF6, PEBP1, TGFBR1	MicroRNA
Neoplastic	MIR101-2	BACH1, FOS, KLF6, PEBP1, WEE1, TGFBR1	MicroRNA
Neoplastic	MIR27A	AR, TP53,PEBP1	MicroRNA
Neoplastic	MIR143, MIR16-2, MIR27A, MIR21	AR, PEBP1, F2R	TF
Neoplastic	MIR214, MIR27A, MIR377	TP53, PEBP1, YWHAG	TF
OPC	MIR27A	AR, TP53,PEBP1, SMURF1	MicroRNA
OPC	MIR21	JUN,ITGB8, PTX3	TF

2.2.5 Discussion

In exploring the landscape of GBM, our study delves into the diverse cellular environments of neoplastic CD133 + radial glia-like cells and OPCs. Central to this investigation is the role of APA in modulating the landscape of microRNA binding, studied at single-cell resolution. This mechanism unveils a hidden layer of post-transcriptional regulation, pivotal in contributing to the cellular heterogeneity characteristic of GBM. This approach has been instrumental in elucidating the complex interplay between APA and gene expression, highlighting APA's distinct influence over cellular behaviors and identities within the GBM matrix.

A significant finding from our study is the shared APA profile in cells transitioning from neoplastic to OPC states. The closer spatial alignment of neoplastic cells with OPCs in PCA space, as evidenced by significantly shorter Euclidean distances, aligns with existing hypotheses regarding a possible neoplastic-OPC transition. This observation suggests a regulatory mechanism, potentially orchestrated by APA, that prepares these cells for a shift in function or identity [115,131]. The similarity in APA patterns, despite the differences in gene expression profiles, indicates a potential functional state or latent capacity that transcends the cells immediate phenotypic presentations.

The study further reveals a minimal overlap between differential APA events and gene expression changes, underscoring APA as an independent layer of regulatory information. We identified crucial genes that are differentially polyadenylated when contrasting neoplastic cells and OPCs, such as RPS3, DVL3, DEF8, EGFR, OLFM1, and GRB2, indicating their potential importance as biomarkers and for shifts in cellular identity. Specifically, the involvement of EGFR and GRB2 in the EGFR signaling pathway [258–260], the participation of DVL3 in the Wnt signaling pathway [261,262], and the association of RPS3 with chemotherapy resistance [263] have been established, underscoring their potential roles in the GBM microenvironment concerning the differentiation of neoplastic cells into OPCs. The mechanisms regulating APA are multifaceted and involve interactions among cleavage and polyadenylation factors, splicing machinery, and transcription elongation processes. For instance, core components such as CPSF4 and ZC3H3 are known to influence 3' end processing efficiency and site selection, which may indirectly affect downstream gene expression and post-transcriptional regulation [115]. However, in our dataset, the lack of significant differential APA site usage in these genes indicates that their regulatory activity may not be cell-type-specific, at least in the context of glioblastoma heterogeneity. Furthermore, APA regulatory genes were notably absent from SFFLs/MFFLs in our analysis, suggesting that while these genes play crucial roles in global mRNA processing, they do not appear to participate directly in the co-regulatory networks shaping cell identity in glioblastoma.

The identification of specific microRNA families selectively evaded by different cell types emphasizes the need for precise control over microRNA interactions. In neoplastic cells, APA-regulated genes are predominantly associated with cell growth and stemness, signifying APA’s role in driving the proliferative and adaptive nature of these tumor cells. In OPCs, APA produces mRNA variants of DNA repair genes that

evade microRNA repression, potentially strengthening their genomic stability capabilities. This enhanced ability to repair damage could confer a higher level of resistance to chemotherapeutic agents, contributing to the tumor's resilience [264,265]. Therefore, APA may represent a critical factor in maintaining the integrity and robustness of the tumor, particularly considering OPCs roles in the tumor microenvironment and chemo-resistance [112,113].

The overlap in APA patterns observed in the UMAP clustering (Figure 2.2.1c) reflects global transcriptomic similarities in polyadenylation usage, particularly during the transitional phase between neoplastic cells and OPCs. This clustering emphasizes shared regulatory mechanisms and gene expression profiles that may contribute to a common functional state. In contrast, the microRNA avoidance analysis (Figure 2.2.2a) focuses on specific APA events that alter microRNA-binding site availability, shedding light on cell-type-specific regulatory consequences. While the shared APA patterns observed in UMAP reflect a broader regulatory foundation, the microRNA avoidance analysis highlights how certain APA events drive fine-tuned adaptations in each cell type's regulatory landscape. This distinction underscores the complementary nature of these analyses in understanding both the global and specific impacts of APA.

The critical impact of alternative polyadenylation in altering microRNA interactions and contributing to increased glioma cell migration by evading the binding sites of the miR-124 family has been previously demonstrated [121]. In our research, this family is notably identified for its substantial loss of binding sites in OPCs and is instrumental in the formation of MFFLs. In the broader landscape of glioblastoma research, our findings resonate with studies highlighting the complex interplay of APA, microRNA, and their impacts on cellular dynamics [123,266,267].

To delve deeper into the mechanistic aspects of microRNA regulation, we conducted a comprehensive network analysis which allowed us to further investigate the complex relationship with TFs, target genes, and APA. The computation of FFLs as a basis for exploring the interaction of microRNAs and transcription factors with target genes to fine-tune gene regulation has been a widely studied topic in the literature. Numerous studies [175,176,180,255] have shown the importance of these regulatory loops in understanding how transcription factors and microRNAs interact to regulate gene expression. The integration of SFFLs into MFFLs within our study has elucidated a more comprehensive understanding of the regulatory dynamics in GBM. This approach has revealed the prominence of cell-type specificity in these networks, notably in the context of genes associated with neoplastic cells, which are found to be enriched in critical pathways such as the defective intrinsic pathway for apoptosis. This observation not only highlights the role of these networks in sustaining the malignant phenotype of GBM cells but also their contribution to resistance against apoptotic mechanisms [268].

In contrast, the genes implicated in OPC-specific MFFLs exhibit a distinct association with GBM signaling pathways. This divergence underscores the unique functional roles and identities of different cell types within the tumor microenvironment, reflecting the complex cellular architecture and the multifaceted nature of GBM.

Further scrutiny of our data reveals a significant role for key microRNAs, including miR-21, renowned for its oncogenic potential and involvement in chemoresistance [99,101,269,270]. The emergence of miR-21 as a central figure in OPC-specific MFFLs not only validates our methodological framework but also expands the understanding of its role within the GBM microenvironment. Additionally, the presence of miR-138–2, associated with high expression levels in oligodendrocyte differentiation [242], in significant OPC MFFLs, further substantiates the importance of our network analysis. The observed significant loss of binding sites for these microRNAs in OPCs, driven by APA, underscores a complex regulatory adjustment, highlighting the dynamic interplay between APA and microRNA regulation.

The utility of MFFL computation is further underscored not only by the significant loss of binding sites for microRNA families within neoplastic cells, attributable to APA, but also by their potential for utilization in survival analysis, enhancing our understanding of prognostic indicators in GBM and providing a foundation for hypothesis generation that warrants further investigation. Our utilization of pseudotime trajectory analysis has been pivotal in revealing that genes differentially expressed during the transition from neoplastic cells to OPCs are actively involved in MFFLs. MicroRNAs such as miR-21 reemerge as crucial elements in these networks, underscoring their influential role in GBM pathology across various regulatory scenarios. Given that the roles of transcription factors can vary depending on the cellular context, our resource of cell-type-specific SFFLs and MFFLs provides a flexible framework for exploring these interactions in a context-dependent manner, without assuming a universally fixed role for transcription factors as activators. Instead, the direction and impact of these interactions are determined by the specific expression patterns and regulatory relationships in each cell type and context. Our MFFL computation has elucidated key aspects of regulatory networks, considering the role of microRNA networks in oncogenesis and tumor suppression [177,271], pan-cancer relevance [178], potential as drug targets [178,253,272], and their demonstrated ability to uncover glioblastoma heterogeneity and cell state transitions.

One limitation in studying SFFLs and MFFLs is the reliance on interaction databases, which can be error-prone or incomplete. To mitigate this, we focused on experimentally validated interactions and conserved interactions across species. However, the networks may still be incomplete, and further refinement may be necessary to enhance their accuracy and comprehensiveness. Additionally, the absence of comprehensive microRNA expression data, particularly in single-cell RNA sequencing, may limit the interpretation of certain interactions.

Our study’s primary focus on neoplastic CD133 + cells and OPCs may limit the broader understanding of GBM's cellular complexity. Additionally, the static nature of our research overlooks the dynamic changes that GBM undergoes over time, emphasizing the need for longitudinal studies, for example, APA profile changes in response to therapy. While we emphasized APA and microRNA regulation, other post-transcriptional mechanisms remain unexplored. Experimental validation and the incorporation of emerging single-cell sequencing technologies are necessary to strengthen the biological relevance of our findings.

In summary, our research provides an in-depth analysis of the cellular dynamics within GBM, focusing on the interplay between cell-type-specific APA, microRNAs, TFs, and target genes. We have successfully constructed and analyzed networks that capture the complex regulatory interactions within GBM, particularly highlighting the unique APA profiles of OPCs and neoplastic CD133 + radial glia-like cells. Our study unravels the subtle yet impactful ways in which APA contributes to cell identity and transformation in GBM, emphasizing the role of microRNA and TF in these processes. These insights provide a comprehensive view of the molecular intricacies that drive the heterogeneity and progression of GBM, offering a valuable resource for future research and potential clinical applications.

2.2.6 Conclusions

In conclusion, our study has effectively mapped the APA landscape in GBM, highlighting its role in the transition between neoplastic cells and OPCs. Utilizing single-cell RNA sequencing data, we identified APA profiles crucial for this cellular transition, providing insights into the post-transcriptional regulation involving microRNAs, transcription factors, and gene networks.

Our workflow pinpointed genes undergoing significant APA changes and subsequent microRNA-binding site loss linked to stem cell maintenance and DNA repair, essential for understanding GBM's adaptability and resilience. The elucidation of APA's influence on microRNA-transcription factor-gene networks has revealed new dimensions of GBM's molecular complexity, revealing both potential therapeutic targets and marker genes for survival analysis.

This research not only sheds light on the regulatory dynamics within GBM but also sets the stage for future studies aiming to exploit APA mechanisms for therapeutic innovation, aligning with the goals of precision medicine in oncology.

2.2.7 Methods

2.2.7.1 Dataset and single-cell clustering

We procured single-cell RNA sequencing data from three glioblastoma samples (GBM27, GBM28, and GBM29), obtained from GSE139448 [273]. The single-cell datasets were processed using CellRanger [274], employing the "mkfastq" function for generating FASTQ files, the "count" function for aligning reads and quantifying gene expression, and the "aggr" function for merging data across samples. Initial quality control steps involved filtering cells based on mitochondrial reads (< 10%) and selecting for RNA counts in the range of 2000 to 10,000. This resulted in 13,600 cells eligible for further analysis using Seurat v3.2.3 package [275]. Normalization and variable feature identification were conducted using Seurat package. Principal component analysis was utilized for dimensionality reduction, with the number of significant components determined via elbow plot methodology, resulting in 17 principal components. We computed Euclidean distances from the neoplastic cell centroid to each cell in other clusters, thus quantifying their spatial proximity in PCA space. To correct for batch effects across samples, we employed the "FindIntegrationAnchors" function, to integrate data from the three distinct samples. Unsupervised clustering was then executed using "FindNeighbors" and "FindClusters" functions. Cell-type annotations were assigned by both genetic expression profiles and known markers (Additional file 1: Table S1). Differential gene

expression was computed using the “FindAllMarkers” function, enabling the identification of unique molecular signatures across different cell clusters.

2.2.7.2 MicroRNA interactions

In assembling our high-quality microRNA-target gene interaction dataset, we included reported interactions from TRANSFAC [276], enriched with experimentally supported interactions from DIANA-TarBase v8 [156] and miRTarBase [168]. We then cross-referenced these with predictive binding sites from TargetScanHuman 8.0 (TSH) [16], supplemented by predictions from broadly conserved microRNAs and their binding sites in TSH. This resulted in a dataset comprising 2,908,279 predicted binding sites for 912 microRNAs, targeting 12,612 genes. To enhance compatibility with current genomic studies, we used UCSC’s liftOver [277] to update these binding sites from hg19 to hg38. This comprehensive dataset stands as a critical resource for studying microRNA-mediated gene regulation.

2.2.7.3 Alternative polyadenylation events

In our study, the APA matrix was computed using the SCAPE program with its default parameters [278]. The APA matrix represents expression values quantified for each polyadenylation site within a gene, providing a finer resolution of polyadenylation usage across samples rather than summarizing expression at the gene level. Additionally, the “FindDE” function was employed to identify differential APA events. This function applies DEXSeq-based analysis, comparing APA usage for each cell cluster against all other clusters to identify cell type-specific APA patterns. Statistical significance was determined using adjusted p-values < 0.05. APA events meeting this threshold were considered statistically significant and prioritized for downstream analyses, ensuring a rigorous and reliable identification of cell type-specific APA usage (Additional file 2: Table S1-S5). We focused our analysis on genes expressed in at least 10% of the cells in each cluster. This threshold ensured that our analysis was based on genes with a significant presence in each cellular subset, providing a more accurate reflection of the APA dynamics within the different cell clusters.

2.2.7.4 MicroRNA avoidance analysis

We quantified the avoidance of microRNA regulation through APA for each cell cluster by using the computed APA expression matrix and annotated microRNA-binding sites. The analysis involved calculating the ratio of total microRNA-binding sites lost versus retained for each microRNA family within each cell cluster. To assess statistical significance, we executed 10,000 permutations, where the cell assignments were randomly shuffled across all clusters while maintaining the original cell counts. In each permutation, the ratio was recalculated, and the incidence of ratios higher than the observed ratio was tracked. The resulting p-value, derived from the frequency of higher ratios divided by the total permutations, indicates the significance of microRNA regulation avoidance in each cell cluster. To account for multiple testing, p-values were adjusted using the Benjamini–Hochberg [279] method, with a significance threshold set at 0.05 (Additional file 1: Table S2).

2.2.7.5 Construction of feed-forward loops

To construct the FFL networks, our methodology integrated four types of directed regulatory interactions: microRNA to TF, microRNA to gene, TF to gene, and TF to microRNA. The microRNA to TF and microRNA to gene interactions were derived

from our compiled microRNA-target gene interaction dataset. For TF to gene and TF to microRNA interactions, we utilized TRANSFAC [276] and TransmiR v2.0 databases [280].

We defined a list of transcription factors as defined in [281] and extracted validated and conserved microRNA to gene/TF interactions. We merged these with TF to microRNA interactions from TransmiR [280] and TF to gene interactions from TRANSFAC [276].

From this background graph, we extracted all possible three-node interactions between microRNA, TF, and gene, categorizing them as simple feed-forward loops (SFFLs). We differentiated these SFFLs based on their central nodes: microRNA-centric (microRNA to gene and microRNA to TF, followed by TF to gene), TF-centric (microRNA to gene and TF to microRNA, followed by TF to gene), and feedback loops (microRNA to gene, TF to gene, microRNA to TF, and TF to microRNA). Subsequently, we identified module feed-forward loops (MFFLs) by merging all significant SFFLs that originated from the same central node such as a specific microRNA. Further, we set the constraint that nodes of these SFFLs must be expressed in at least 10% of the cells belonging to the cluster. This approach allowed us to discern broader regulatory patterns and understand the influence of individual entities like a specific microRNA on multiple pathways in GBM. While missing small RNA expression data, especially for microRNAs, is a known limitation in single-cell RNA sequencing, our use of MFFLs helps mitigate this challenge by integrating multiple SFFLs into broader networks. This approach enables the capture of biologically meaningful interactions even when individual microRNA data are incomplete, allowing us to interpret regulatory dynamics driven by APA and transcription factors with greater robustness.

2.2.7.6 Significance testing of SFFLs and MFFLs

To ascertain the significance of SFFLs in diverse cell types, we scored networks comprising microRNA, TF, and gene nodes. In order to score computed FFLs, we applied the methodology prior demonstrated by [178]. Node scores, based on adjusted p-values from differential expression analysis, were computed by comparing the expression of a gene in a specific cell cluster versus in all other cells. These scores underwent inverse normal cumulative distribution transformation. Edge scores were computed using Fisher transformation of Pearson's correlation z-scores between nodes. The average of node and edge scores constituted the overall SFFL score. Due to limitations in capturing small RNAs in single-cell RNA data, if microRNA scores were absent, we calculated the score using the remaining nodes and edges. However, to address the sparsity of microRNA expression data, we enhanced the gene expression matrix by employing the PPMS software [282], which profiles primary microRNAs with cell-type specificity, concentrating on the inclusion of those microRNAs that were absent in the initial gene expression dataset.

To assess the significance of the SFFLs, we permuted the network 10,000 times by selecting three random nodes—one microRNA, one TF, and one target gene—from the same cell cluster as the original SFFL. For each permutation, we computed the overall SFFL score again by calculating the node scores and edge scores for the randomly selected nodes. The randomized overall scores were then compared to the original SFFL score. The p-value for each SFFL was determined by calculating the

ratio of random scores greater than or equal to the observed score. Finally, the Benjamini–Hochberg FDR correction [279] was applied to account for multiple testing, with a significance threshold of 0.05 (Additional file 1: Table S3). All significant SFFLs within a specific cluster were merged together based on the center node, which makes up the MFFLs for that cluster, providing a consolidated view of the regulatory networks for that cell type (Additional file 1: Table S4).

2.2.7.7 GO term enrichment for genes avoiding microRNA regulation by APA

For the Gene Ontology (GO) term analysis, we conducted an examination specifically on genes identified as avoiding microRNA regulation by significant microRNA families. This analysis was restricted to genes pertinent to each individual cell cluster. Utilizing the "enrichGO" functions from ClusterProfiler [171], we focused on biological processes (BP) (Additional file 1: Table S6) and molecular functions (MF) (Additional file 1: Table S7) to discern the biological and molecular underpinnings influenced by this evasion. This approach allowed us to gain insights into the unique biological processes and molecular functions affected by the avoidance of microRNA regulation in each specific cell cluster context.

2.2.7.8 Pseudotime trajectory

Utilizing Monocle 3 [283], we computed pseudotime trajectories, specifically targeting cells from the neoplastic cell to OPC clusters to model a potential transdifferentiation (Additional file 1: Table S5). To analyze differential expression over time, we conducted graph-autocorrelation analysis using the "graph_test" function. This was preceded by the application of the "estimate_size_factors" function, adhering to its default settings.

2.2.7.9 Kaplan–Meier survival analysis

Kaplan–Meier survival analysis was performed using TCGA-GBM clinical and gene expression data, processed with TCGAbiolinks [284], survival [285], and survminer packages in R. Clinical data pertinent to survival was extracted, and gene expression data were normalized using DESeq2's [286] variance stabilizing transformation. Cox regression analysis was conducted using the "coxph" function [285], incorporating gene expression values, age at diagnosis and TIDE-derived exclusion and dysfunction scores [287] as covariates, as they reflect the functional state of immune cells and the extent of immune evasion in the tumor microenvironment.

The coefficients for the respective genes obtained from the Cox regression model were combined with their expression values to stratify patients into high-risk and low-risk groups. Kaplan–Meier survival curves were then generated for these stratified groups, and statistical significance was assessed with a threshold of $p < 0.05$ (Additional file 1: Table S4). This approach, as demonstrated by, e.g., [288,289], allowed us to explore the prognostic significance of MFFLs in the context of multiple clinical and molecular factors.

2.3 Evaluating Genetic Regulators of MicroRNAs Using Machine Learning Models

2.3.1 Preamble

This chapter is published in the journal International Journal of Molecular Sciences:

Cihan, M., Anyaegbunam, U. A., Albrecht, S., Andrade-Navarro, M. A., & Sprang, M. (2025). Evaluating Genetic Regulators of MicroRNAs Using Machine Learning Models. International Journal of Molecular Sciences, 26(12), 5757. https://doi.org/10.3390/ijms26125757

The supplementary files associated with this publication are available on the publisher's website via the article's DOI.

2.3.2 Abstract

This study explores the genetic regulators of microRNAs (miRNAs) using a set of machine learning models to predict miRNA expression levels from gene expression data. Employing machine learning, we accurately predicted the expression of 353 human miRNAs ($R2 > 0.5$), revealing robust miRNA–gene regulatory relationships. By analyzing the coefficients of these predictive models, we identified genetic regulators for each miRNA and highlighted the multifactorial nature of miRNA regulation. Further network analysis uncovered that miRNAs with higher predictive accuracy are more densely connected to their top predictive genes, reflecting strong regulatory control within miRNA–gene networks. To refine these insights, we filtered the miRNA–gene interaction networks to identify miRNAs specifically associated with enriched pathways, such as synaptic function and cardiovascular processes. From this pathway-centric analysis, we present a curated list of miRNAs and their genetic regulators, pinpointing their activity within distinct biological contexts. Additionally, our study provides a comprehensive set of metrics and coefficients for the genes most predictive of miRNA expression, along with a filtered subnetwork of miRNAs linked to specific pathways and phenotypes. By integrating miRNA expression predictors with network analysis and pathway enrichment, this work advances our understanding of miRNA regulatory mechanisms and their roles across distinct biological systems. Our approach enables researchers to train custom models using TCGA data and predict miRNA expression from gene expression inputs.

2.3.3 Introduction

MicroRNAs (miRNAs) play a critical role in the regulation of gene expression by binding to target messenger RNAs (mRNAs) and either promoting their degradation or inhibiting their translation [290]. These small, non-coding RNAs are involved in a wide array of cellular processes, including development, differentiation, and apoptosis, making them essential for maintaining cellular homeostasis [7,291].

Accurate profiling of miRNA expression is crucial for understanding miRNA functions. To predict miRNA targets, among other methods, the negative correlation between miRNA and mRNA expression is used to identify potential novel miRNA target binding sites on genes [199,292,293]. By mapping these interactions, researchers can elucidate how miRNAs influence various cellular pathways and processes, highlighting their potential as therapeutic targets and biomarkers [294,295]. This characterization not only relies on direct binding site identification but also on

integration with annotation databases, high-throughput experimental validation, evolutionary conservation studies, and network-based analysis [7,16,251,296,297].

Quantifying miRNA expression remains challenging due to biases in current experimental methods. Small RNA-seq often relies on adapter ligation and PCR amplification, introducing representation biases that affect accuracy and reliability [298–300]. Issues like false high fold changes from low expression values and alignment errors also arise, especially when compared to the higher precision of qRT-PCR [301]. Moreover, integrating data across platforms is complicated by differing biases and error profiles. A major limitation is the lack of validated reference controls for normalization, leading to variability and poor cross-study comparability [298,302–304].

To address the limitations in miRNA expression quantification, a range of machine learning and computational methods have been developed. For instance, a constrained least squares approach has been reported for imputing missing miRNA expression values, improving data completeness in partially observed miRNA matrices [305]. Moreover, MMpred employs regression to predict miRNA expression from microarray data, facilitating the inference of miRNA–mRNA interactions [199]. miREACT utilizes motif enrichment analysis to estimate miRNA activity from single-cell RNA-seq data, providing insights into miRNA regulation at the single-cell level [306]. Similarly, miRSCAPE leverages tree-based machine learning to infer miRNA expression from single-cell RNA-seq data, enabling the study of miRNA activity in contexts where direct measurement is challenging [198]. Other frameworks aim to infer miRNA activity or regulatory influence, such as the enrichment-based method for estimating miRNA repression from gene expression profiles [307] and the causal inference approach for detecting miRNA–mRNA regulatory relationships directly from expression data [308]. Collectively, these methods underscore the versatility of machine learning in tackling both expression-level and functional characterization challenges in miRNA biology.

While mRNA data is used to infer miRNA expression, the genetic regulators of miRNAs remain poorly understood as even intronic miRNAs often show weak correlation with their host genes [309,310]. Moreover, no existing tool offers pretrained models that allow direct inference of miRNA expression from bulk RNA-seq input alone, limiting broader applicability.

In this study, we use gene expression data from RNA sequencing to predict miRNA expression levels, offering an approach that leverages the correlations between gene and miRNA expressions to build predictive machine learning models, providing a more accessible and accurate computational alternative to direct miRNA measurement. By doing so, it allows us to infer miRNA activity and its regulatory impact on genes, facilitating deeper insights into both cellular mechanisms and disease pathways. To this end, we applied ridge regression [194], a regularization technique suited for handling multicollinearity and high-dimensional data, to predict miRNA levels from RNA-seq data obtained from The Cancer Genome Atlas (TCGA) [69] from both normal and cancer tissues across thousands of samples. By analyzing the regression coefficients, we identified predictive genes for each miRNA considered, revealing key regulatory elements within the gene–miRNA network.

Subsequent network analysis, incorporating miRNA binding data, enabled us to map out intricate interactions and pathways to characterize the functional relevance and biological implications of these predictive genes. Our approach is the first to provide pretrained, reproducible models that directly infer miRNA expression from bulk RNA-seq data while simultaneously uncovering gene-level regulators—offering a framework to explore miRNA control across diverse tissues and disease contexts.

2.3.4 Results

In this study, we applied ridge regression to develop a set of models to predict miRNA expression levels from RNA sequencing data. By leveraging the correlations between gene and miRNA expressions, our approach provides a computational alternative to direct miRNA measurement. Additionally, we constructed a network of miRNAs and their target genes, integrating experimentally validated interactions and predicted conserved interactions to understand the functional relevance and biological implications of these regulatory relationships better. We then used the feature coefficients from these models to identify key predictive genes, allowing us to explore the regulatory elements within the gene–miRNA network. This analysis offers a deeper insight into miRNA–gene interactions and their roles in cellular mechanisms and disease pathways.

2.3.4.1 Model Development and Performance Assessment

We thoroughly evaluated the performance of the ridge regression models that we utilized to predict miRNA expression levels from RNA sequencing data, using various statistical metrics. The distribution of R2 values across all miRNAs (Figure 2.3.1A) reveals a wide range of predictive accuracies. The cumulative distribution function (CDF) overlay shows that 353 out of the 1300 miRNAs analyzed achieve R2 values greater than 0.5, demonstrating strong predictive capabilities for these specific targets. This may stem from their inherently high expression levels, as reflected by the median TPM values extracted from respective TCGA samples ($R2 \leq 0.5$: 0.14; $R2 > 0.5$: 39; see Section 4 for details).

The comparison between observed and predicted miRNA levels (Figure 2.3.1B) shows a strong linear correlation, as evidenced by a Pearson correlation coefficient of 0.99, indicating the model's proficiency in accurately capturing the mean expression levels for the majority of miRNAs, reinforcing the validity of using gene expression data as a reliable surrogate for direct miRNA measurement. In addition to R2-based evaluation, we calculated the Spearman correlation between observed and predicted miRNA expression values across all miRNAs. We obtained a mean Spearman correlation of 0.55, indicating a strong monotonic relationship between observed and predicted values.

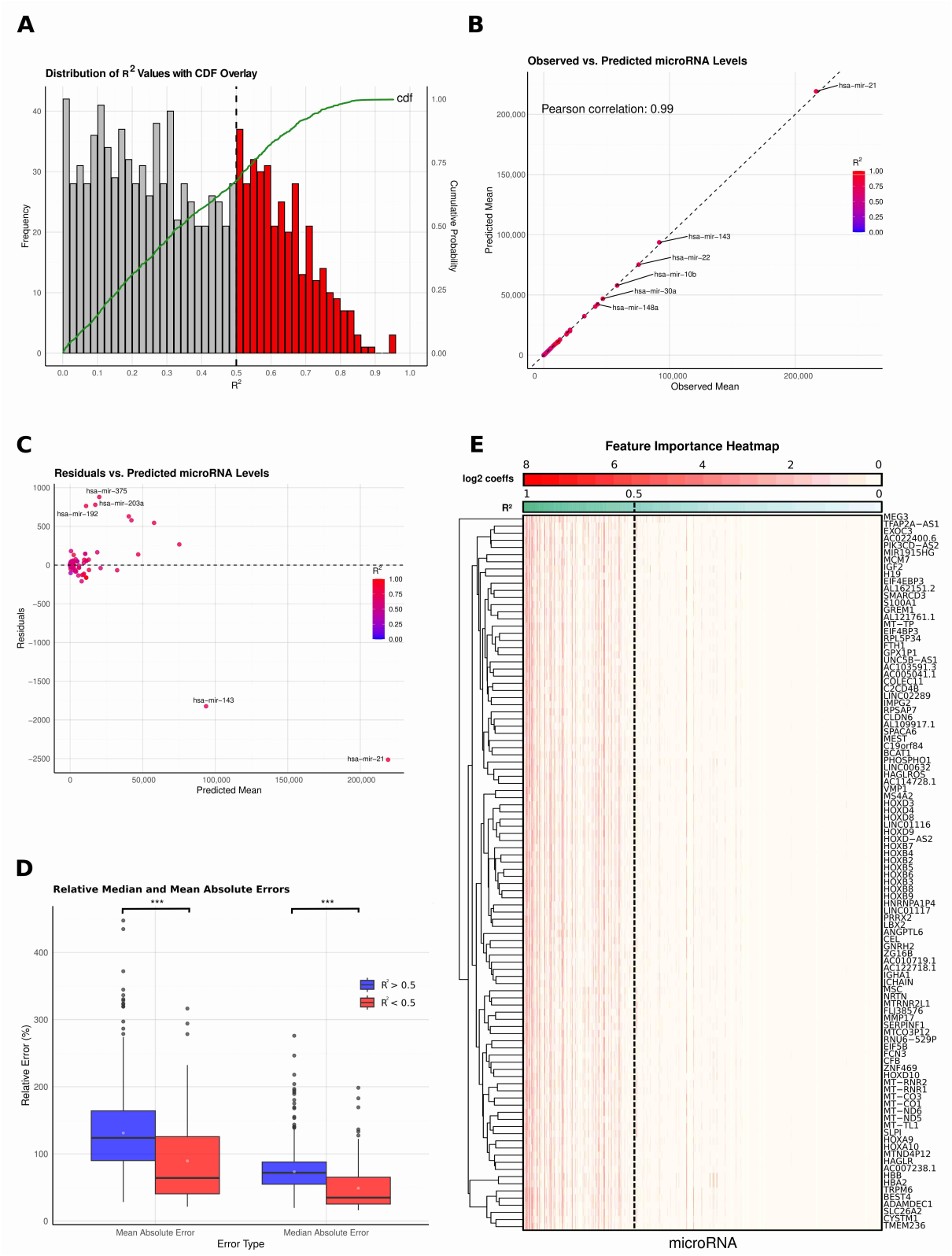


Figure 2.3.1: miRNA Prediction Performance and Feature Importance Analysis. (A) Distribution of R2 values across predicted miRNAs, with a histogram illustrating the range of prediction accuracies and cumulative probability overlayed in green.

*Figure 2.3.1 (continued): (B) Comparison of observed versus predicted mean miRNA expression levels, shown in a scatterplot with data points colored according to R2 values, highlighting prediction accuracy across samples. The strong correlation indicates the model's effectiveness in estimating overall expression patterns. (C) Residual analysis displaying the differences between observed and predicted mean values plotted against predicted mean miRNA levels, with outliers highlighted. This indicates expression ranges where the model performs less consistently. *** denotes p < 0.01. (D) Boxplots comparing relative errors (median and mean absolute errors) for miRNAs grouped by predicted R2 values (<0.5 and >0.5), providing insight into prediction reliability across different accuracy levels. Lower errors in the high-R2 group emphasize the model's robustness for better-predicted miRNAs. (E) Clustered heatmap of the top 100 genes with the highest absolute coefficients, showing feature importance by miRNA. Genes are sorted in descending order of R2 values, visualizing the predictive contributions across miRNAs, and log2-transformed absolute coefficients are visualized to highlight the relative contribution of each gene across different miRNA models. This panel illustrates the diverse gene contributions underlying miRNA expression and supports the model's reliance on multiple features.*

To evaluate the model's consistency, we examined the correlation between the coefficients of variation (CV) for observed and predicted miRNA levels. Overall, this correlation was moderate ($r = 0.32$), indicating some alignment between observed and predicted variability. For miRNAs with R2 values above 0.5, the correlation was much stronger ($r = 0.98$), demonstrating high stability and consistency in predictions. Conversely, for miRNAs with R2 values < 0.5, the correlation dropped to 0.20, highlighting greater challenges in accurately predicting these miRNAs. These results reinforce the model's robustness, especially for miRNAs with higher predictive accuracy.

Residual analysis (Figure 2.3.1C) provides additional insights into the model's robustness and areas for improvement. While the residuals generally cluster around zero, indicating unbiased predictions across most expression levels, certain miRNAs, such as hsa-mir-21, exhibit significant deviations from the trend. These deviations are primarily associated with miRNAs that have very high expression levels, suggesting that outlier values or extreme expression levels may introduce some noise or variability into the predictive model.

The relative errors, both mean absolute error (MAE) and median absolute error (MedAE), were significantly lower for miRNAs with R2 > 0.5 compared to those with R2 < 0.5, as shown in Figure 2.3.1D, with a p-value less than 0.01. This highlights the improved predictive accuracy for miRNAs with higher R2 values, pointing at the model being more robust for targets with generally higher expression values (Figure 2.3.1D).

For the top 100 genes with the highest variability, the feature importance heatmap (Figure 2.3.1E) illustrates the absolute log2 values of the coefficients for miRNAs with R2 values greater than 0.5, revealing that many features have high coefficients. This indicates that the prediction of miRNA expression is not driven by a single gene but rather by a complex interaction among multiple genes. The presence of several genes with substantial coefficients underscores the multifactorial nature of miRNA regulation, validating the model's strategy of using a diverse set of gene expression data to enhance predictive accuracy. Conversely, for miRNAs with R2 values < 0.5, there are no significantly high coefficients, suggesting a lack of strong predictive features and highlighting the challenges in predicting these miRNAs accurately (Figure 2.3.1E).

To assess whether alternative modeling approaches could improve predictive accuracy, we also implemented Lasso regression as a linear model and Random Forest regression as a non-linear model. Both approaches underperformed relative to ridge regression, with 196 miRNAs (Lasso) and 239 miRNAs (Random Forest) achieving R2 values > 0.5. A full comparison of model performance metrics is provided in Supplementary Tables S1–S3.

Overall, the results demonstrate that our ridge regression models provide a robust framework for predicting miRNA expression from RNA-seq data, particularly for miRNAs with clear expression patterns.

2.3.4.2 MiRNA–Gene Network Connectivity and Centrality Analysis

We analyzed the connectivity of the top 3% (632) of predictive genes, determined by the highest absolute coefficients for each miRNA, within the gene–miRNA network (Figure 2.3.2A). For miRNAs with R2 > 0.5, a higher proportion of predictive genes were found to be directly interacting with the miRNA (1-node distance), averaging 125 genes, compared to 102 genes for miRNAs with R2 < 0.5. Additionally, when examining the 3-node distance (3 degrees of separation), the difference between the two groups becomes more pronounced, with miRNAs that are better predicted (R2 > 0.5) showing an average of 401 connected genes, compared to 358 for those with R2 < 0.5. This suggests that miRNAs with higher predictive power tend to form stronger direct regulatory relationships with their target genes, highlighting the connection between prediction accuracy and regulatory interactions (Figure 2.3.2A).

We also examined the distribution of lncRNAs and protein-coding genes among the top predictive genes. Both groups, miRNAs with R2 > 0.5 and R2 < 0.5, had a similarly small proportion of lncRNAs among the top predictive genes. However, the proportion of protein-coding genes was higher for miRNAs with R2 > 0.5 (Figure 2.3.2B).

Analysis of the network's communities (groups of densely connected nodes, see Methods for details) shows variability in the proportion of well-predicted miRNAs (R2 > 0.5) across different communities, with some communities having a higher concentration of accurately predicted miRNAs. While this observation suggests differences in the predictive relationships within these communities, no consistent pattern was observed regarding community size (nodes/edges) and prediction quality (see Supplementary Table S4).

We analyzed the relationship between miRNA expression variability and their connectivity within the network by focusing on the 55 miRNAs with a high coefficient of variation (CV > 10). Correlating their R2 values with different network centrality measures revealed notable relationships: a Pearson correlation of 0.49 for both degree centrality and betweenness centrality and 0.47 for eigenvector centrality. These positive correlations suggest that miRNAs with higher variability in expression tend to be predicted better when they occupy more central and influential positions in the network.

This finding implies that miRNAs with significant network connectivity—either by having numerous direct interactions (degree centrality), being central to communication pathways (betweenness centrality), or influencing other highly

connected nodes (eigenvector centrality)—are more likely to exhibit predictable expression patterns. This could indicate that miRNAs deeply embedded in the regulatory network play crucial roles in maintaining network stability, which could explain why their expression is better captured by predictive models.

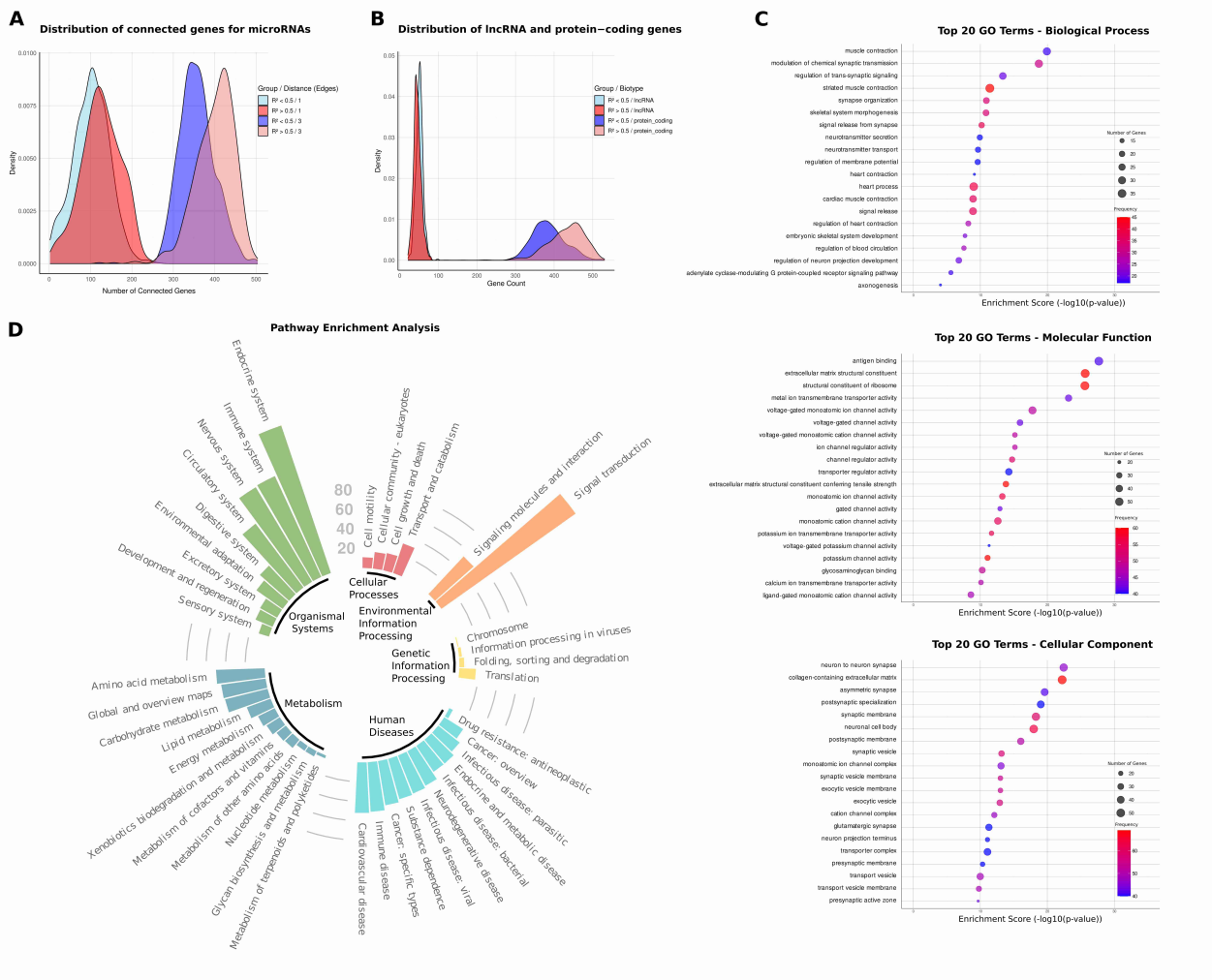


Figure 2.3.2: Functional Characterization of Predictive Genes in miRNA Networks. (A) Distribution of direct and 3-node distance gene interactions among the top 632 predictive genes for each miRNA, divided into two groups: R2 > 0.5 and R2 < 0.5. Distribution plot highlights differences in connectivity for high and low-accuracy miRNAs. Higher connectivity in the well-predicted group reflects stronger and more direct regulatory relationships in these miRNA models. (B) Density distribution of gene biotypes, with long non-coding RNAs and protein-coding genes represented among the top predictive genes for each miRNA, split by R2 category (R2 > 0.5 and R2 < 0.5). Protein-coding genes dominate predictive gene sets, particularly for miRNAs with high prediction accuracy. (C) Gene Ontology (GO) term analysis results for the top 20 enriched terms in biological process (BP), cellular component (CC), and molecular function (MF) categories. Enrichment is based on the frequency of significance across miRNAs, using the top 632 predictive genes per miRNA (see Supplementary Table S5). (D) KEGG-pathway enrichment analysis for the top predictive genes across miRNAs, illustrated in a bar plot where the height of each bar represents the number of miRNAs significantly enriched in each pathway. This visualization highlights the pathways most frequently associated with genes that predict miRNA expression.

2.3.4.3 Biological Signatures of Predictive miRNA Regulators

In the GO term analysis of the top predictive genes for miRNAs with R2 > 0.5, we found that many terms in the biological process category are related to synaptic function and cardiovascular processes. Notably, terms such as the modulation of chemical synaptic transmission, synapse organization, neurotransmitter secretion, and the regulation of neuron projection development had the highest number of predictive genes associated with them. In addition, cardiovascular-related terms like cardiac muscle contraction, heart process, and the regulation of blood circulation were also highly enriched, underscoring the involvement of these genes in critical physiological pathways (Figure 2.3.2C).

In the molecular function category, many of the enriched terms pertained to ion channel activity, particularly those involved in synaptic signaling. Higher-level categories like voltage-gated ion channel activity, monoatomic cation channel activity, and potassium channel activity dominated, with a large number of genes contributing to these functions. This suggests that most well-predicted miRNA families have predictive genes involved in regulating ion transport and signaling, further emphasizing their role in synaptic function and neuronal regulation (Figure 2.3.2C).

The cellular component category also reflected a strong focus on synaptic structures, with terms such as synaptic vesicle membrane, postsynaptic membrane, and neuronal cell body being the most enriched. These terms highlight the cellular environments where the predictive genes are most active, particularly in synapse-related functions. The enrichment in these synaptic components suggests that the genes associated with better-predicted miRNAs are often localized to critical regions involved in neural communication (Figure 2.3.2C).

These findings illustrate that the majority of well-predicted miRNA families have predictive genes that are heavily involved in synaptic and cardiovascular processes, as reflected by their enrichment in both functional and structural terms across the GO categories.

2.3.4.4 miRNA-Linked Pathway Enrichments

We subsequently performed pathway enrichment analysis using the KEGG database for the same set of predictive genes. Pathways significantly enriched across the majority of miRNAs include signal transduction, which involved 170 out of the 353 miRNAs considered (48%), the endocrine system with 160 miRNAs (45%), the nervous system with 113 miRNAs (32%), and cardiovascular diseases with 52 miRNAs (15%). These findings align with the GO term enrichment results, emphasizing synaptic and cardiovascular processes (Figure 2.3.2D).

We then filtered the miRNA–gene network to focus specifically on the genes associated with the pathw ays identified in the previous enrichment analysis, retaining only miRNAs directly connected to these genes. Additionally, we incorporated specific pathways corresponding to the enriched terms. For the nervous system, we presented this filtered network in Figure 2.3.3A, where key miRNAs such as miR-137 and miR-488 emerged as highly connected nodes within the network. This strategy resulted in the selection of 11 miRNAs, revealing a clear concentration of regulatory interactions within neural-associated pathways.

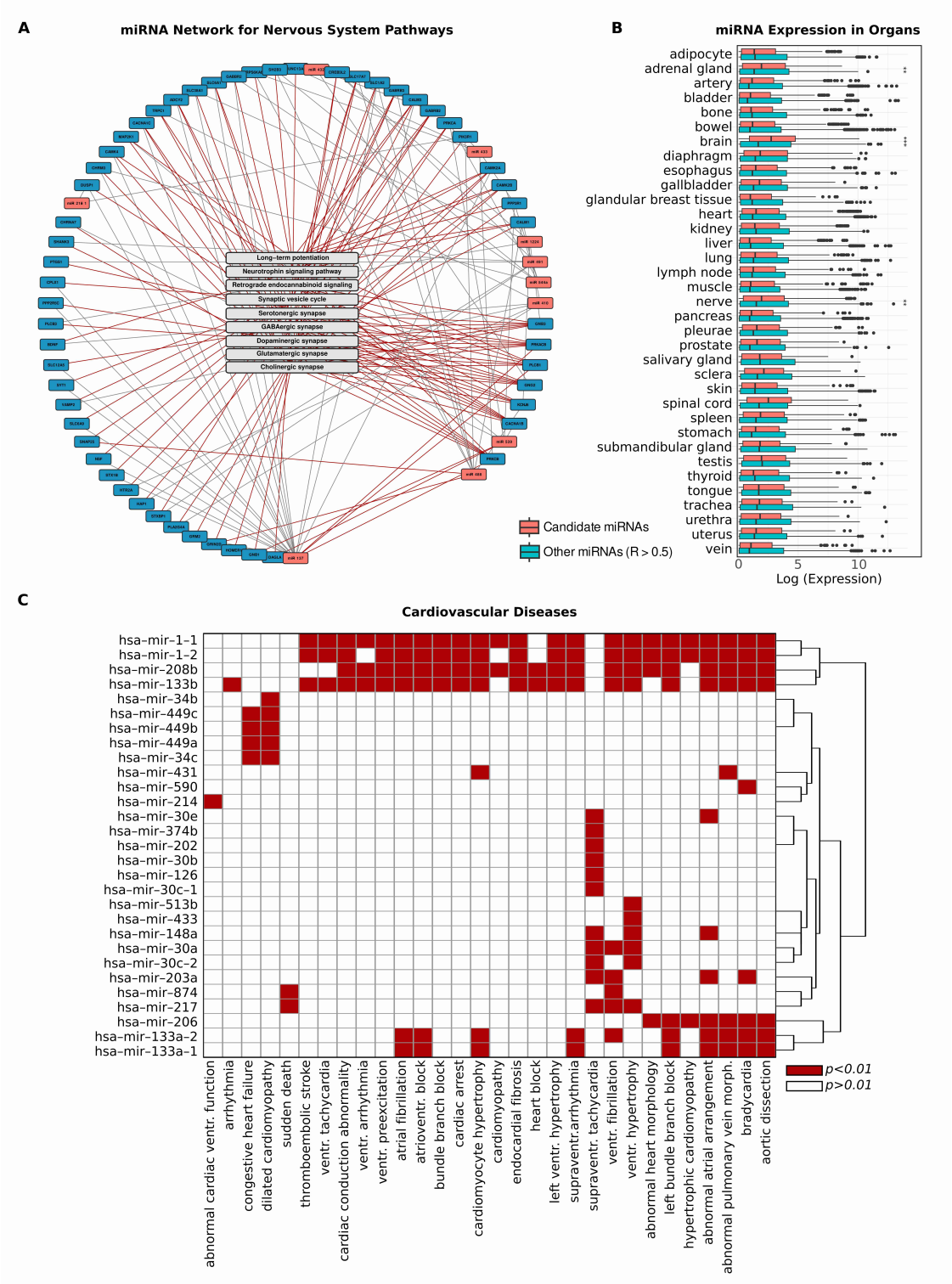


Figure 2.3.3: Predictive miRNAs in Neural and Cardiovascular Pathways. (A) Filtered network of miRNA-gene interactions, focused on pathways related to the nervous system. Network visualizes specific interactions in nervous system-associated pathways, emphasizing miRNAs like miR-137 that show high connectivity. Red edges indicate interactions involving genetic predictors and direct pathway associations, while grey lines represent additional connections with non-predictive regulators and indirect pathway associations.

*Figure 2.3.3 (continued): (B) Bar plot displaying tissue-specific expression levels of miRNAs that are filtered for signal transduction pathways and for the background of all other miRNAs predicted with R2 > 0.5. Elevated expression in neural tissues supports the functional relevance of the filtered miRNA set; *** denotes p < 0.01, ** denotes p between 0.01 and 0.05. (C) Heatmap showing miRNAs significantly enriched in cardiovascular disease pathways (p < 0.01), linking predictive genes to disease-relevant regulatory modules.*

We applied the same filtering strategy to isolate pathways associated with signal transduction, which also led to the selection of 43 miRNAs. For these miRNAs, we further analyzed their tissue-specific expression patterns using the TAM 2.0 database [169]. We contrasted their expression levels with the expression profiles of all other miRNAs with R2 > 0.5. Through this comparison, we identified significantly higher normalized expression levels of these signal-transduction-associated miRNAs in several key tissues, particularly the brain, nerve, and adrenal gland (Figure 2.3.3B).

These findings are consistent with the biological roles of signal transduction and nervous system pathways and provide additional validation of our network filtering methodology. The enriched expression in neural and endocrine-related tissues supports the functional relevance of the extracted miRNAs and highlights their potential regulatory impact within these critical systems. This underscores the biological coherence of our approach, linking the predictive genes and pathways to specific tissue contexts, thus reinforcing the importance of these miRNAs in neural and signal-transduction-related processes.

2.3.4.5 Cardiovascular Disease Associations in Predictive Gene Networks

We next conducted disease enrichment analysis for the predictive genes of miRNAs with R2 > 0.5, focusing on terms related to cardiovascular diseases, which were among the most prevalent in the pathway enrichment results (Figure 2.3.3C). Notably, the terms arrhythmia, abnormal cardiac ventricular function, and cardiac conduction abnormality were among the most frequent. Specifically, we observed significant enrichment of cardiovascular-disease-related terms for the predictive genes of miR-1-1, miR-1-2, miR-208b, and miR-133b, indicating their strong association with various cardiovascular conditions. miR-208b and miR-133b also appeared consistently when constructing the subnetwork for cardiovascular diseases, which included a total of 16 miRNAs. This reinforces the role of these specific miRNAs in cardiovascular regulation and highlights their potential importance in disease-associated regulatory networks.

2.3.5 Discussion

In this study, we employed ridge regression to predict miRNA expression from RNA sequencing data, leveraging its strengths in handling high-dimensional data and capturing multicollinearity. Ridge regression has been widely used in genetic studies to address the challenges posed by complex datasets, particularly those involving gene regulatory networks [311–314]. Its ability to manage large numbers of correlated features while maintaining robust predictions makes it ideal for exploring the regulatory interactions between miRNAs and their target genes, which are often characterized by overlapping regulatory roles and multicollinear gene expression profiles [156,168,251]. Our models performed well for over 353 miRNAs (R2 > 0.5), likely due to their higher expression levels facilitating stronger signal detection.

Among the top 100 miRNAs with the highest mean expression, only eight were predicted with R2 values below 0.5, reinforcing the strength of the model in capturing the regulatory dynamics of highly expressed miRNAs.

In contrast, miRNAs with lower expression levels posed a greater challenge. This is likely due to their lower signal-to-noise ratios, making it difficult for the model to distinguish true signals from background noise. Random Forest regression did not improve predictions, suggesting limited predictability may stem from biological variability or sparse input features. Lastly, regularization in ridge regression tends to shrink the coefficients of low-expression miRNAs, reducing their predictive accuracy. However, this trade-off is crucial for preventing overfitting as the model must balance capturing meaningful patterns without allowing noise to dominate the predictions [195,315,316]. A critical factor in the model's success was carefully selecting the regularization parameter (alpha), which we set to 11,000. This relatively high regularization was essential due to the inclusion of a large number of gene features.

While existing methods predominantly address miRNA expression imputation through techniques such as constrained least squares and GO-based similarity measures, our approach broadens the application to both healthy and tumor tissues, enhancing predictive performance without the need for imputation strategies [305,317]. Although no current tool provides pretrained models for direct miRNA prediction from gene expression, we benchmark our approach against miRSCAPE [198], which infers miRNA levels from scRNA-seq and reports a mean Spearman correlation of 0.45 across 10 TCGA cancer cohorts, focusing on miRNAs expressed in over 50% of samples. miRSCAPE also reports improved performance over miREACT based on bulk RNA-seq data [306]. In comparison, our model achieves a higher mean Spearman correlation of 0.55 across 1300 miRNAs, despite using a less stringent filtering threshold. We attribute this performance gain to two factors: the use of a linear ridge regression model, which generalizes well in high-dimensional settings, and the approximately two-fold increase in training data as we aggregated samples across all cancer types rather than limiting to individual cohorts. By predicting miRNA expression directly from RNA-seq-derived expression matrices, our set of models offers broader applicability without the need for pre-existing miRNA profiles. Our study is further strengthened by the availability of scripts that enable researchers to train their own models using TCGA or similar datasets, providing flexibility in adapting the approach to diverse research questions. By integrating these computational tools, we aim to facilitate reproducibility and extend the practical applications of miRNA prediction. These resources empower users to not only predict miRNA expression but also explore novel regulatory relationships tailored to specific datasets and biological systems.

Our analysis of connectivity within the miRNA–gene network reveals significant relationships between prediction accuracy ($R2 > 0.5$) and the degree of direct gene interactions. Positive correlations with degree centrality and betweenness centrality suggest that miRNAs with greater regulatory influence tend to exhibit more stable and predictable expression patterns. This highlights the importance of considering miRNAs not in isolation but within the context of their broader network interactions.

miRNAs with high centrality likely act as regulatory hubs, influencing a wide range of target genes across critical biological pathways.

We observed a significant enrichment of biological processes related to synaptic function and cardiovascular systems. Terms such as “synapse organization” and “cardiac muscle contraction” consistently appeared in the GO analysis for miRNAs with high R2 values, indicating a role in crucial physiological pathways. This is further validated by the pathway enrichment results, where signal transduction and cardiovascular pathways dominated. Filtered miRNAs, such as miR-1-1, miR-208b, and miR-133b, which showed strong associations with cardiovascular-disease-related pathways, have been extensively documented as key players in their role in cardiovascular disease progression and biomarkers in the literature [318–321], providing further validation of our findings and highlighting their critical role in cardiovascular regulation. miRNAs may play an essential role in the heart’s adaptability to varying physiological stimuli, allowing for rapid regulatory responses critical in maintaining cardiac rhythm and contractility [322,323].

In neurons, miRNAs may function as highly localized regulatory elements, helping to control mRNA pools in distant cellular locations like dendrites. This setup supports rapid, synapse-specific protein synthesis and is well suited to the nervous system’s dynamic demands, where miRNAs act as localized regulators that inhibit protein production from stored mRNAs [324]. This regulatory mechanism is consistent with the presence of multiple polyadenylation sites in neural transcripts [325,326], allowing for flexible transcript pools finely regulated by miRNAs.

By subsetting these miRNA–gene networks based on genes predictive for miRNAs and enriched in signal transduction pathways, we identified miRNAs specifically linked to these pathways. When cross-referenced with independent databases, these selected miRNAs also showed higher expression in their respective tissues that are nerve, adrenal gland, and brain.

This organ-specific expression additionally validates our methodological approach, indicating that the miRNAs identified as key players especially in synaptic and cardiovascular pathways are biologically relevant. This finding is consistent with existing research that demonstrates the tissue-specific roles of miRNAs in both the heart and brain [327,328]. Our analysis highlights several miRNAs—miR-1, miR-133b, miR-208b, miR-137, and miR-124—that not only show high predictive accuracy but are also strongly linked to known biological functions, reinforcing the validity of our models. miR-1 and miR-133b are well-established regulators of cardiac function, implicated in muscle differentiation, arrhythmias, and heart failure [329,330]. miR-208b, intronic to MYH7, is known to regulate cardiac hypertrophy [331]. In the nervous system, miR-137 plays a central role in regulating synaptic vesicle trafficking, neurotransmitter release, and presynaptic plasticity—functions essential for proper synaptic activity and neuronal signaling [332]. miR-124, one of the most abundant brain miRNAs, has been shown to fine-tune the timing and amount of adult neurogenesis [333]. The strong model performance and pathway enrichment of these miRNAs support the biological coherence of our predictions and functional connection to respective genetic regulators.

One limitation of this study is the model's reduced accuracy in predicting miRNAs with low expression levels, likely due to weaker signal-to-noise ratios and sparse data for these miRNAs. Additionally, the miRNA–gene network used in our analysis is based on current datasets, which are continuously evolving as new experimental data becomes available. As a result, some regulatory interactions may be missing, and the network may not fully capture all relevant biological relationships, potentially limiting the scope and applicability of our findings.

This study illustrates the effectiveness of ridge regression for predicting miRNA expression levels from gene expression data, offering a computational alternative to direct miRNA measurement. By analyzing the predictive gene–miRNA relationships, we revealed key insights into the functional roles of miRNAs, particularly in synaptic and cardiovascular processes. Our findings provide a foundation for further exploration of miRNA regulation in disease contexts, and the framework developed here has the potential for broader applications in miRNA-related research. In future work, integrating deep-learning-based approaches may offer improved adaptability and performance, particularly for harmonizing and modeling data originating from multiple experimental platforms [334].

2.3.6 Methods

2.3.6.1 Data Collection

We sourced expression data from TCGA [69], selecting all samples across all cancer types that included both miRNA sequencing and RNA-seq data from the same tissue condition (either tumor or normal). Our dataset comprised 10,464 matched expression profiles, with 9828 derived from tumor tissues and 636 from normal tissues. To ensure uniformity and comparability across all samples, we used normalized expression values using Transcripts Per Million (TPMs). This normalization method accounts for differences in sequencing depth and gene length, providing a standardized framework for subsequent analysis.

2.3.6.2 Data Preparation

To model miRNA expression levels, we utilized gene expression values as predictive features. The initial data preparation phase involved several key steps: loading the RNA-seq and miRNA-seq datasets and filtering the feature matrix by removing rows with more than 1050 zeros and 1050 NaNs, a threshold set to 10% of the total samples. This threshold was selected to eliminate features with excessive missing or non-expressed values while retaining a broad set of informative predictors. For the label matrix, we applied a less stringent threshold of 10,000 zeros and 10,000 NaNs. Given that miRNA expression is often tissue-specific and generally lower in abundance, this threshold allows us to retain miRNAs present in at least 325 samples, capturing approximately 70% of miRNAs, even those expressed at lower levels. This filtering strategy enabled elimination of non-informative features and labels, enhancing the robustness of subsequent analyses. We also removed features with zero variance to focus on informative predictors.

Next, we applied a Z-transformation using the StandardScaler function from the scikit-learn library [192] to standardize the features so that the values of different features are on the same scale to ensure that the regularization applies equally to all coefficients, which is important for the performance of the ridge regression.

2.3.6.3 miRNA Expression Modeling

For each miRNA, the dataset was split into training (80%) and test (20%) subsets using random sampling, and ridge regression was applied using the Ridge function from scikit-learn. This regularization technique is effective in managing multicollinearity and high-dimensional datasets, chosen for its ability to prevent overfitting and enhance prediction stability. The regularization parameter (alpha) was set at 11,000 after initial tuning after an initial grid search with 5-fold cross-validation on the same training set over a range of values (1, 10, 100, 1000, 10,000, 11,000, and 15,000), selecting 11,000 as it yielded the highest number of miRNAs with R2 values greater than 0.5 while maintaining the lowest mean absolute error. This methodology ensures that the test set remained completely independent throughout the training and model optimization process to provide an unbiased evaluation of model performance. Interestingly, running the same model with the top 3% of features with the highest absolute coefficients yielded similar results (Supplementary Tables S1 and S2). While we focused on these top features for downstream analysis, we retained all features for modeling, as computing efficiency was not significantly impacted, and the overall results improved. Given the set of TPM-normalized gene expression data, our method can be applied to predict miRNA expression using the provided predict_microRNA_expression.py script.

2.3.6.4 Model Performance Evaluation

The models were trained on the training set and evaluated on the test set using metrics such as R2, median absolute error, and mean absolute error (Supplementary Tables S1 and S2). These metrics provided a comprehensive view of model performance across different miRNAs. The final dataset consisted of 21,044 features (mRNA expression data) and 1300 miRNAs as targets. All features were retained in the final models as feature exclusion did not improve model efficiency and retaining them enhanced predictive accuracy. The top 3% genes with the highest absolute coefficients for miRNAs predicted with $R2 > 0.5$ are provided in Supplementary Table S6. The entire modeling process was conducted using Python version 3.10 and R version 4.4. For comparison, we also trained models using the Lasso and RandomForestRegressor functions from the scikit-learn library [192], using the same training and test splits as for ridge regression (Supplementary Table S3).

2.3.6.5 Network Construction and Subsetting

Following the predictive modeling, we constructed a network of miRNAs and their target genes. This network was built by combining experimentally validated interactions from the miRTarBase [168], DIANA-TarBase [156], and TRANSFAC databases [276]. Additionally, we complemented these interactions with predicted conserved interactions from the TargetScanHuman 8.0 [16] with conservation scores higher than 0.5 to ensure a comprehensive representation of miRNA–gene interactions. In this network, each node represents a gene or miRNA, and the edges represent the interactions between them.

2.3.6.6 Connectivity Metrics

To analyze miRNA connectivity within the network, we used the biomaRt package [335] to classify gene biotypes as protein-coding or long non-coding RNAs (lncRNAs). Key connectivity metrics were computed using the igraph package [336]: degree assessed the number of direct interactions, betweenness measured the role of

miRNAs as bridges in the network, and evcent evaluated their influence based on connections. Communities were computed with cluster_louvain. Pearson correlation coefficients between R2 values and centrality measures were calculated using the cor function. Network metrics for miRNA–gene regulatory interactions, including connectivity measures, centrality scores, and community structure, are provided in Supplementary Tables S4 and S7.

2.3.6.7 Gene Ontology (GO) Term Analysis

For the Gene Ontology (GO) term analysis, we focused on the top 3% of the most predictive genes for each miRNA, selected based on R2 values greater than 0.5. This analysis concentrated on significant findings with adjusted p-values smaller than 0.05. Utilizing the enrichGO function from clusterProfiler [171], we examined biological processes (BP), molecular functions (MF), and cellular components (CC) to uncover the biological and molecular underpinnings influenced by these highly predictive genes. Detailed GO term enrichment results are provided in Supplementary Table S5.

2.3.6.8 Pathway Analysis

For the pathway analysis, we again focused on the top 3% most predictive genes for each miRNA with R2 values greater than 0.5. We performed enrichment analysis using the KEGG database [172], applying a threshold of adjusted p-values smaller than 0.05 to identify significant pathways (see Supplementary Tables S8 and S9). This involved utilizing the enrichKEGG function from the clusterProfiler [171] package to map the predictive genes to their associated biological pathways, providing insights into the regulatory frameworks governing miRNA-mediated gene expression.

2.3.6.9 Disease Enrichment

The disease enrichment analysis was conducted using the top 3% of predictive genes for each miRNA, identified based on R2 values greater than 0.5. These genes were cross-referenced with the Human Phenotype Ontology database [337] through the g:Profiler web platform [338], with a p-value cutoff of 0.01 (see Supplementary Table S10)

2.3.6.10 Organ-Specific miRNA Expression

For the comparison of organ-specific miRNA expression levels between miRNAs associated with signal transduction pathways and all other miRNAs with R2 values greater than 0.5, we utilized the TAM 2.0 database [169] to extract miRNA levels for each organ. The expression levels of the pathway-specific miRNAs were contrasted with the remaining miRNAs, and the significance of the differences was assessed using a t-test using t test function of the R stats library.

2.3.7 Conclusions

We present a robust ridge regression framework for predicting miRNA expression from RNA-seq data, identifying over 350 miRNAs with high predictive accuracy. Our models uncover key gene–miRNA regulatory relationships, particularly in synaptic and cardiovascular pathways, and highlight tissue-specific expression patterns linked to biological function and disease. The accompanying pretrained models and scripts offer a reproducible tool for miRNA prediction, with broad applicability in both research and clinical settings.

2.4 Target-site Dynamics and Alternative Polyadenylation Explain a Large Share of Apparent MicroRNA Differential Expression

2.4.1 Preamble

This chapter is a preprint on bioRxiv:

Cihan, M., More, P., Sprang, M., Marini, F., & Andrade, M. (2025). Target-site Dynamics and Alternative Polyadenylation Explain Large Share of Apparent MicroRNA Differential Expression. bioRxiv, 2025-09. https://doi.org/10.1101/2025.09.29.679194

The supplementary files associated with this publication are available via the article's DOI.

2.4.2 Abstract

MicroRNA (miRNA) abundance reflects a dynamic balance between biogenesis, target engagement, and decay, yet differential expression analyses typically ignore changes in target-site availability driven by alternative polyadenylation (APA). We introduce MIRNAPEX, an expression-stratification-based machine learning framework that quantifies miRNA regulatory effect sizes from RNA-seq data by integrating target-gene expression with 3′UTR isoform usage to infer effective binding-site dosage. Using pan-cancer training sets, we train models that learn relationships between transcriptomic features and miRNA log-fold changes, with APA patterns providing predictive information beyond gene expression alone. When applied to knockdowns of core APA regulators, MIRNAPEX captured widespread 3′UTR shortening and accurately anticipated miRNA-specific shifts whose direction and magnitude mirrored APA-driven changes in binding-site availability. Analysis of target-directed miRNA degradation interactions further showed that loss of distal decay-trigger sites coincides with increased miRNA abundance, consistent with reduced degradation. Together, these findings demonstrate that apparent miRNA differential expression can arise from dynamic target-site landscapes rather than altered miRNA transcription, and that neglecting this dimension can lead to misestimation of regulatory effect sizes.

2.4.3 Introduction

MicroRNAs (miRNAs) are short (~22-nt) non-coding RNAs that post-transcriptionally regulate gene expression by binding to partially complementary sites in the 3′ untranslated regions (3'UTRs) of target genes [290]. By doing so, miRNAs fine-tune developmental programs, buffer cellular stress responses, and, if dysregulated, contribute to diverse disease phenotypes, including metabolic disorders, cancer, and neurogenerative diseases [11,339–341]. Considering the essential role of miRNAs in maintaining normal physiology and their potential as predictive biomarkers [342], accurate quantification of their expression level is crucial. In comparative transcriptomic analyses, differences in miRNA expression between conditions are commonly interpreted as indicators of altered post-transcriptional regulation and a reflection of broader regulatory state changes [343,344]. However, the steady-state level of a mature miRNA reflects a moving balance of three broad processes that determine the cellular abundance of each miRNA species. First, biogenesis, which includes transcription of the primary transcript, Microprocessor cleavage, nuclear

export, Dicer processing and loading of Argonaute (AGO) proteins to form RNA-induced silencing complexes (RISCs), sets the potential pool of mature miRNAs [5,11]. Second, target engagement redistributes miRNA–RISC complexes across the transcriptome and determines how strongly a given miRNA can repress its targets in a particular cellular state [16,24,339]. Third, decay pathways remove mature miRNAs, with general turnover mechanisms such as tailing and trimming followed by exonucleolytic decay controlling their half-life [345,346].

In addition, target-directed miRNA degradation (TDMD) is a process in which binding to a highly complementary target transcript actively triggers destabilization and decay of the miRNA itself, thereby accelerating its decay. For instance, systematic AGO-CLASH analyses have revealed numerous endogenous transcripts that act as TDMD triggers, indicating that target-directed decay of miRNAs is more prevalent than previously recognized [34,41,347]. Because biogenesis, target engagement, and decay, including TDMD, act simultaneously and dynamically, observed changes in miRNA abundance may not directly report transcriptional output but can reflect shifts in target availability and turnover.

Accordingly, several studies have estimated miRNA activity from properties of their targets, showing that target gene abundance and site affinity predict miRNA levels, AGO binding, and competition effects [307,348–350]. Intuitively, target expression sets the demand placed on a miRNA, such that abundant, site-rich targets increase demand, whereas depletion of those targets reduces it [349,351,352]. However, most target-centric approaches treat each gene as if it had a single, fixed 3’UTR, overlooking the widespread phenomenon of alternative polyadenylation (APA) [115]. APA leads to the generation of transcript isoforms with distinct 3′UTR lengths and can therefore add or remove canonical miRNA binding sites as well as highly complementary TDMD trigger sites, dynamically altering the effective binding-site dosage available to each miRNA. Indeed, more than half of human genes utilize APA to generate alternative 3′UTR isoforms, meaning that dynamic 3′UTR length changes broadly modulate available miRNA binding sites and thus influence post-transcriptional regulation [115,125,326,353]. Since APA is a widespread mechanism that controls 3′UTR length and miRNA-site availability, perturbation experiments of core APA factors (such as CFIm25, CFIm68 and CPSF6) have shown that knocking them down remodels thousands of 3′UTRs across the transcriptome [122,123]. Such APA-driven alterations in miRNA targeting have been shown to impact gene expression programs stem cell function and differentiation, and oncogenic transformation, among others, highlighting the crucial interplay between APA and miRNA regulation in fine-tuning cellular phenotypes [354,355].

Despite this, most DE analyses of miRNAs ignore dynamic changes in effective target-site dosage caused by APA and their impact on TDMD, creating a gap in how observed miRNA shifts are interpreted. We therefore hypothesize that APA-driven 3′UTR remodeling, by altering binding sites, systematically shapes mature miRNA levels such that effective target-site availability correlates with the effect size of miRNA regulation, typically quantified as log fold-change (logFC), between states and samples. Consequently, apparent miRNA DE often reflects target-site dynamics rather than altered miRNA transcription.

Motivated by this gap, we introduce MIRNAPEX, an expression-stratification-based interpretable machine learning (ML) framework that predicts miRNA logFC from RNA-seq by integrating target-gene expression with 3′UTR isoform usage to estimate effective binding-site dosage. Beyond prediction, we quantify the relative impact of APA variation on each miRNA and apply MIRNAPEX to APA-factor perturbation datasets to test whether global 3′UTR shortening produces predictable shifts in miRNA levels. We further examine curated TDMD trigger–miRNA pairs to see if loss of distal TDMD sites coincides with expected increased miRNA abundance [34]. Altogether, this approach shows how transcriptomic variation, including 3′UTR remodeling, shapes miRNA abundance, underscoring that miRNA DE and its estimated effect size should be interpreted in the context of dynamic target-site landscapes.

2.4.4 Methods

2.4.4.1 Data Collection

To train the ML models for predicting miRNA logFCs based on RNA sequencing data, we assembled datasets from The Cancer Genome Atlas (TCGA) [69]. We downloaded all available mRNA and miRNA quantification data from TCGA and cross-referenced these samples with the TC3A database [356], a resource that applies the DaPars algorithm to TCGA RNA-seq data to quantify APA patterns [117]. APA is represented by percentage of distal usage index (PDUI) values, that serve as a measure for distinguishing long and short 3′UTRs. PDUI values range between 0 and 1. For miRNA we obtained isoform-level quantification files and mapped them to mature miRNA entries using miRBase annotations [7].

The final dataset comprised 8460 samples with matched mRNA, miRNA, and APA profiles. Specifically, it includes TPM values for gene expression, mean RPM values for 2,000 mature miRNAs, and PDUI values for between 1058 and 11,266 genes per cancer type. Because APA usage is influenced by gene expression and biological context, the number of genes with valid PDUI values varies across cancer types. In total, the dataset spans 32 distinct TCGA cancer types and forms the basis for training the miRNA-specific ML models.

2.4.4.2 Feature Engineering and Sample Definition

For each miRNA, putative target genes were obtained from the microT database [157] and ranked according to their gene-level microT interaction scores. To systematically evaluate the impact of feature set size, we constructed multiple input variants per miRNA by selecting the top 25, 50, 75, 100, 250, 500, 750, 1000, and 2000 highest-scoring target genes that are reported to have APA measurements. To generate training examples for each miRNA, we first randomly split all available samples into training (80%) and test (20%) sets. For model evaluation, the training set was further divided into five folds for cross-validation (CV). Within each fold, samples were stratified into high and low expression groups based on the expression level of the respective miRNA. Each sample from a specific fold was then randomly paired with a sample from the opposite expression group within the same fold. This strategy enabled the creation of diverse sample pairs representing varying expression differences while preventing data leakage between training and validation subsets. For each generated sample pair, we computed gene-level differential features for all target genes. Specifically, we calculated the logFC in

mRNA expression between the two samples and the corresponding difference in PDUI values (ΔPDUI) for APA usage. To avoid undefined values due to zero expression, a correction of +1 was applied prior to logFC calculation. ΔPDUI values range between −1 and 1, reflecting relative changes in distal polyadenylation site usage. Features are computed in the same way for the test and full training sets (Figure 2.4.1).

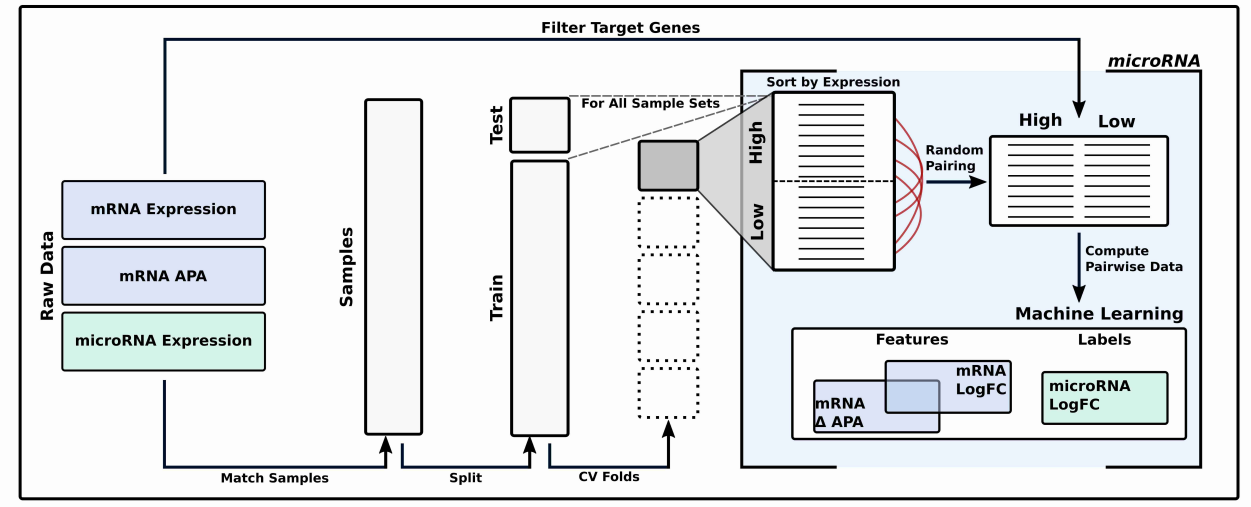


Figure 2.4.1: Workflow for training ML models to predict miRNA logFC. Matched TCGA samples containing mRNA expression, APA profiles, and miRNA expression are split into random training (80%) and test (20%) sets, with the training set further divided into cross-validation folds. Within each fold, samples are stratified by miRNA expression levels and randomly paired in both directions to prevent data leakage during training. For each miRNA, annotated target genes with valid APA measurements are selected, and differential features (mRNA logFC, ΔPDUI) are computed for each sample pair. The observed miRNA logFC serves as the prediction label, and the same feature computation is applied across folds, the test set, and final training set after hyperparameter tuning. A separate ML model is trained for each miRNA to capture the relationship between transcriptomic changes and miRNA expression dynamics.

2.4.4.3 Training ML models to predict miRNA expression changes

We trained a range of ML algorithms to predict logFC values of mature miRNAs using the transcriptomic feature sets described above. To evaluate model performance under varying input conditions, we generated multiple training datasets based on different random pairings of samples and varying numbers of miRNA target genes (ranging from 25 to 2000). For each configuration, we trained both linear and non-linear regression models implemented in the scikit-learn library [192], including ordinary least squares (OLS), Lasso (LA), Ridge (RI), Elastic Net (EN), histogram-based gradient boosting regressor (HB), random forest regressor (RF) and multilayer perceptron (MLP). Hyperparameters for each model were optimized using CV within the training folds. For each miRNA, the model and hyperparameter combination that achieved the highest mean R2 across CVs was selected and retrained on the full training set to produce the final predictive model.

2.4.4.4 The MIRNAPEX workflow

The resulting miRNA-specific ML models form the core of MIRNAPEX, enabling prediction of miRNA logFC values between two groups of RNA-seq samples. The MIRNAPEX pipeline automates the full process, starting from the raw FASTQ files. It integrates the GDC mRNA quantification pipeline [357] and DaPars-based APA

analysis [117,356,358] to compute gene-level logFC and ΔPDUI values for predefined miRNA target genes. These features are then passed to pretrained miRNA-specific regression models to predict logFC values for 1165 miRNAs across any two user-defined sample groups (Supplementary Figure 6.1).

2.4.4.5 APA perturbation

To test whether APA-driven changes in binding-site availability translate into shifts in mature miRNA levels, four RNA-seq comparisons of APA-regulator knockdowns with matched controls were analyzed. These perturbations remodel 3′UTRs, altering binding-site dosage. MIRNAPEX was then applied to predict miRNA log-fold changes between perturbed and control samples, and concordance with APA-driven target-site changes was evaluated. The datasets involve CFIm25 knockdown in HCT116 (GSE158591) as CFIm25-KD-1 [359]; CPSF6 knockdown in HEP3B (GSE229281) as CPSF6-KD [360]; and the HEK293 experiments comprising an independent CFIm25 knockdown replicate and a CFIm68 knockdown (GSE179630) as CFIm25-KD-2 and CFIm68-KD [122], respectively. We validated mature miRNA expression levels in respective cell lines using DIANA-miTED [173] and annotated miRNA binding sites on target transcripts with predictions from the DIANA-microT [157].

To approximate transcriptional contributions to miRNA abundance, we derived intronic transcriptional proxy measurements for a subset of TDMD pairs in which the miRNA is encoded within an intron of the trigger or host transcript. These measurements capture changes in transcription at the pri-miRNA locus and were used to evaluate whether mature miRNA logFC could be explained by altered transcription.

2.4.5 Results

2.4.5.1 Transcriptomic Prediction of miRNA expression changes

miRNA logFC values between sample groups were predicted using features derived from their putative target genes. Two types of features were considered: gene-level logFC values, reflecting differential mRNA expression, and ΔPDUI values, capturing changes in APA patterns. Together, these measures serve as proxies for the relative abundance of miRNA binding sites within their target genes. To avoid reliance on a single modeling assumption, we compared several commonly used regression algorithms that differ in their treatment of high-dimensional feature spaces. These include linear models with L1 or L2 regularization, which emphasize feature selection or coefficient shrinkage, as well as non-linear models that capture complex relationships through ensemble or kernel-based approaches. To assess predictive performance, ML algorithms were built using feature sets of varying size, defined by ranked microT interaction scores (see Methods for details).

Linear models substantially outperformed non-linear approaches in predicting miRNA logFC values across feature set sizes (Figure 2.4.2A). EN achieved the highest mean R2, followed closely by LA and RI, while non-linear models such as RF, HB, and MLP performed worse. As a baseline algorithm, OLS exhibited a marked decline in performance once the feature set exceeded ~250 genes, highlighting the importance of regularization in high-dimensional settings. These findings are

consistent with prior observations that linear models are well suited for modeling miRNA expression dynamics [350].

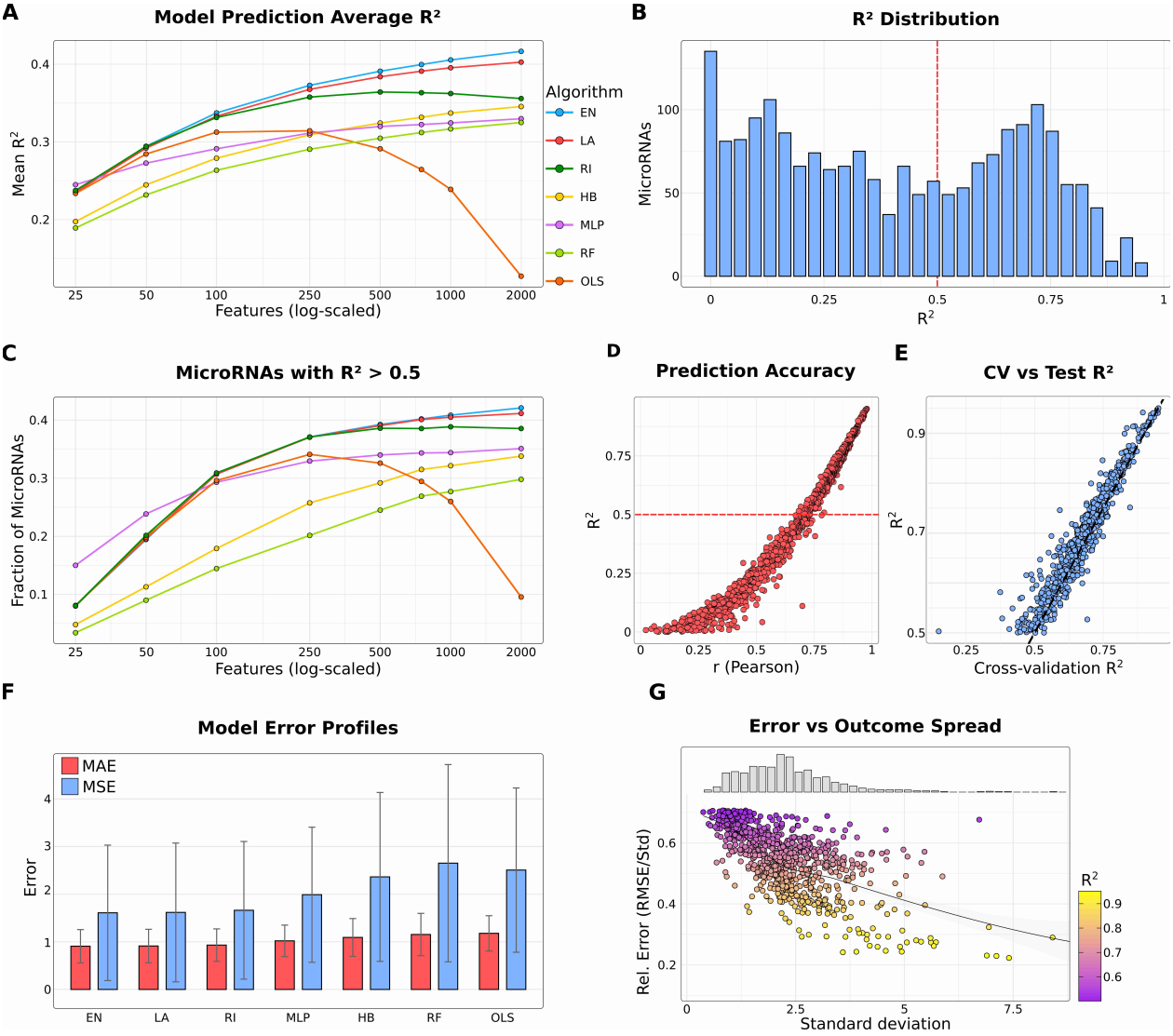


Figure 2.4.2: Prediction performance of ML models for miRNA logFC. (A) Line plot of mean R2 values across different ML algorithms as a function of the number of input features (tick marks represent logarithmically spaced values). Algorithms are abbreviated as follows: EN, elastic net; LA, lasso; RI, ridge regression; HB, HistGradientBoost ; MLP, multilayer perceptron; RF, random forest; OLS, ordinary least squares. (B) Distribution of R2 values from EN trained with 1000 features across all evaluated miRNAs. (C) Fraction of miRNAs achieving R2>0.5 as a function of the number of input features (tick marks represent logarithmically spaced values). (D) Dot plot of Pearson correlation between predicted and observed logFC values versus miRNAspecific R2 . (E) Comparison of cross-validation R2 against test set R2 across miRNAs. (F) Mean absolute error (MAE) and mean squared error (MSE) across algorithms (for 1000 features). (G) Relationship between the standard deviation of predicted logFC values and relative error (RMSE/standard deviation). The distribution of standard deviation values is shown as a histogram.

A feature set size of 1000 was selected as the optimal balance between predictive accuracy and interpretability (Supplementary Table 1). At this scale, EN achieved a mean R2 of 0.41 across all miRNAs, and the distribution of prediction accuracies showed that 817 miRNAs (41%) surpassed the R2 > 0.5 threshold, with a mean R2

of 0.69 for this subset (Figure 2.4.2B,C). These highly predictable miRNAs (HP-miRNAs) were prioritized for downstream analyses. Across all miRNAs, the average Pearson correlation between predicted and observed logFC values was 0.62, while the correlation increased to 0.83 for HP-miRNAs (Figure 2.4.2D). Robustness of the models was further supported by the strong concordance between cross-validation and test set performance, with a Pearson correlation of 0.98 (Figure 2.4.2E), indicating minimal over- or underfitting.

To further benchmark model accuracy, prediction errors were compared across algorithms at the 1000-feature setting (Figure 2.4.2F). EN achieved the lowest mean absolute error (0.91) and mean squared error (1.60), further highlighting its robustness relative to the other methods. Moreover, analysis of relative error revealed that prediction error scaled proportionally to the variance of the observed logFC values and remained small relative to the standard deviation, particularly for HP-miRNAs (Figure 2.4.2G).

Together, these analyses demonstrate that effect size estimate of the miRNA expression regulation can be predicted with high accuracy and robustness from transcriptomic features.

2.4.5.2 Expression- and APA-driven signals jointly shape miRNA prediction accuracy

The prediction of miRNA activity from transcriptomic data has traditionally been based on mRNA expression levels measured by RNA-seq or microarray platforms [198,306]. To investigate the added predictive value of 3′UTR patterns, we evaluated the role of APA. Specifically, we trained ML models for each miRNA using three different feature sets: expression-only, APA-only, and combined expression plus APA features.

Expression-only models achieved a mean R2 of 0.39 across all miRNAs, while APA-only models performed slightly lower with a mean R2 of 0.36. Importantly, the combined models improved performance to a mean R2 of 0.41, demonstrating that APA contributes complementary predictive information beyond gene expression alone (Figure 2.4.3A). Among the HP-miRNAs, there were 802 miRNAs for the expression-only and 617 for the APA-only models scoring with $R^2 > 0.5$. Although the gain in overall prediction performance when using expression and APA features together is modest compared with using either feature set alone, this analysis demonstrates two important points. First, APA-only models perform comparably to expression-only models, indicating that a measure independent of gene expression quantification, namely 3′UTR length patterns, can predict differential miRNA behavior. Second, combining expression and APA features provides a unified framework to assess their relative and context-dependent contributions to miRNA regulation. To assess model performance on high-confidence miRNAs, we evaluated MIRNAPEX predictions for MirGeneDB-supported miRNA genes [8]. Across all MirGeneDB miRNA entries, the mean predictive accuracy was R^2 of 0.61. At the gene level, allowing either mature arm to contribute, 364 of 506 MirGeneDB miRNA genes showed strong predictability ($R^2 > 0.5$). Together, these results indicate that MIRNAPEX performance is strongest for high-confidence miRNA annotations and is not driven by low-confidence miRBase entries.

To further investigate the predictive signal, we examined feature contributions from both the miRNA and target gene perspectives.

From the miRNA perspective, analysis of average coefficients confirmed that both expression- and APA-derived features contributed substantially to prediction accuracy, with no miRNA relying exclusively on a single modality (Figure 2.4.3B). Expression features were moderately more influential overall, with 77% of miRNAs showing higher weights for expression than for APA. This bias, however, was rather modest than extreme, and no outliers exhibited complete dependence on one feature type, consistent with previous observations.

From the gene perspective, we assessed whether target genes contributed systematically through expression or APA features. Among 8260 target genes across all HP-miRNAs, 70% exhibited a bias toward expression-derived contributions. Specific examples included CITED1, SLC52A2, and ACTG2, which were primarily expression-driven, whereas IFITM1 and PRDX6 were dominated by APA. Nonetheless, most genes featured in many miRNA models (>200) showed no strong preference, again highlighting the balanced contributions of both modalities (Figure 2.4.3C).

To further dissect modality-specific contributions, we stratified genes into APA- and expression-dominant groups based on the 10th and 90th percentile cutoffs of their dominance fraction. This classification yielded 891 APA-dominant and 768 expression-dominant genes across all HP-miRNAs. Notably, genes with high prevalence across multiple miRNAs typically exhibited only moderate dominance (Figure 2.4.3D).

When comparing variability across modalities, expression-dominant genes showed higher variance in both expression and APA relative to APA-dominant genes. For expression values, median coefficients of variation were 0.716 versus 0.650, and for APA, 0.225 versus 0.221. A Wilcoxon rank-sum test confirmed significantly greater variability in expression ($p < 0.001$) and APA ($p < 0.01$) for expression-dominant genes. Importantly, APA-dominant genes did not exhibit elevated APA variability across miRNAs, indicating that their predictive contribution reflects systematic APA regulation rather than noise (Figure 2.4.3E). Similarly, the top 100 recurrently used genes were rarely exclusive to APA or expression, but instead reflected mixed contributions (Figure 2.4.3F).

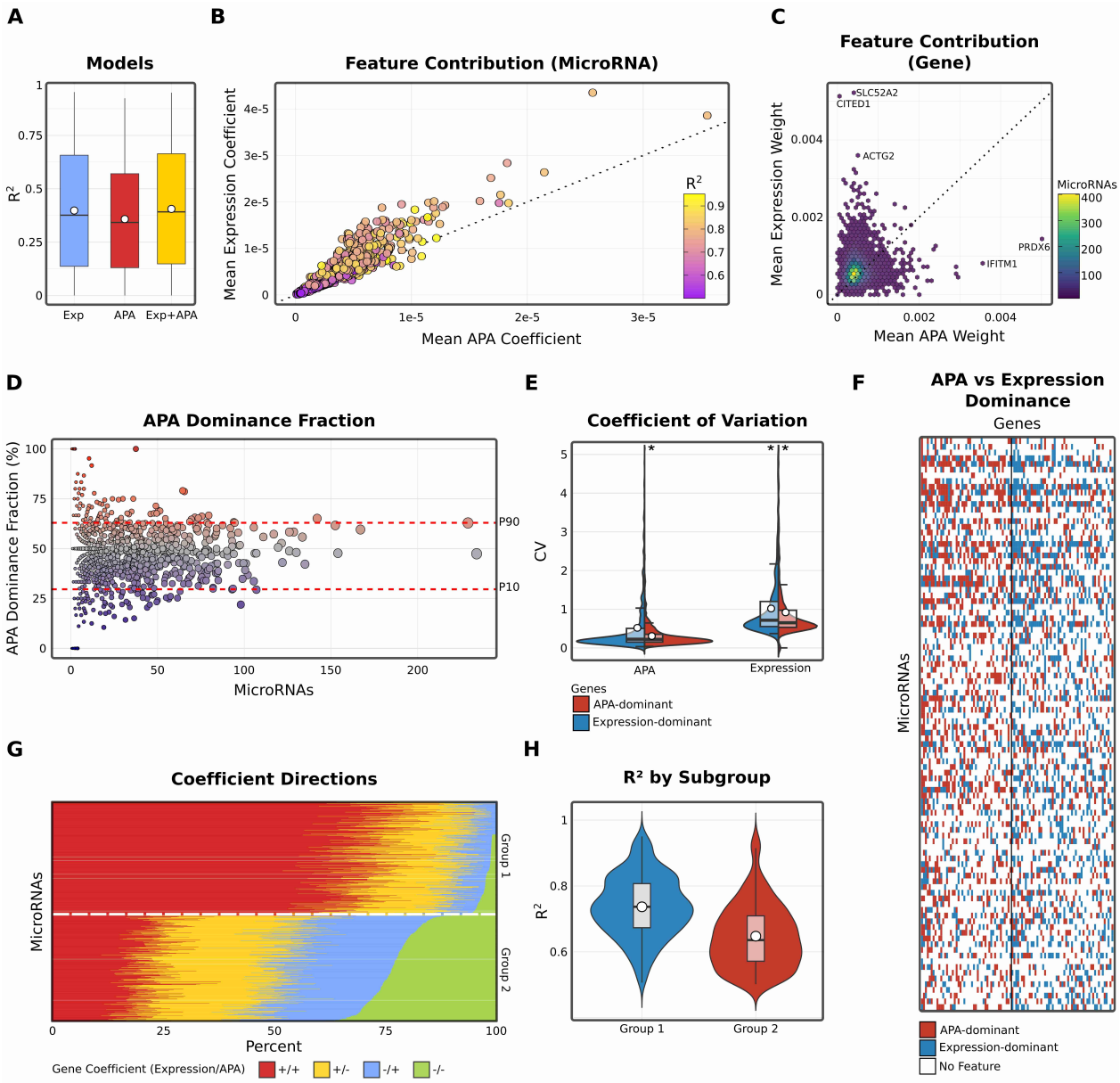


Figure 2.4.3: Contribution of APA and expression features to miRNA logFC prediction. (A) Comparison of predictive performance for models trained with APA-only, expression-only, or combined features. Boxplots show the distribution of R^2 scores across miRNAs. (B) Scatter plot of average normalized absolute coefficients for APA versus expression features across miRNAs with R2>0.3. Each point represents one miRNA, colored by predictive performance (R^2). (C) Hexbin plot of gene-level contributions, showing mean percentage weight of APA versus expression features across highly predictable miRNAs ($R^2 \geq 0.5$). Color scale denotes the number of miRNAs in which a given gene contributes. (D) Scatter plot of APA dominance fraction versus gene prevalence across all genes contributing to miRNAs with $R^2 \geq 0.5$. Each point represents one gene, with dashed lines marking the 10th and 90th percentile cutoffs used to define expression-dominant versus APA-dominant gene sets. (E) Coefficient of variation for APA versus expression features within the extreme APA-dominated and expression-dominated gene deciles. These measurements represent the raw variability from which APA and expression features are derived for model training. Stars denote outliers above the plotted range. (F) Heatmap of categorical dominance (APA versus expression) across all miRNAs with $R^2 \geq 0.5$ and the 100 most prevalent genes. Rows are ordered by decreasing miRNA R^2, and columns by decreasing gene prevalence. White indicates that the gene was not a selected feature for that miRNA, blue indicates higher absolute expression coefficient, and red indicates higher absolute APA coefficient.

Figure 2.4.3 (continued): (G) Stacked barplots of sign concordance between APA and expression coefficients across all genes for miRNAs with $R^2 \geq 0.5$. Colors denote the four possible sign combinations: both positive (++), APA positive with expression negative (+−), APA negative with expression positive (−+), and both negative (−−). miRNAs are stratified into two groups based on their composition, using a threshold of 10% (−−). (H) Comparison of R^2 values between the two stratified groups.

2.4.5.3 Bimodal sign patterns reveal distinct Expression–APA relationships

Beyond their relative magnitudes, the coefficient signs for expression- and APA-derived features reveal how these two modalities tend to co-vary within our models. In general, positive coefficients for both expression and APA of target genes indicate that higher expression together with more distal 3′UTR usage is statistically associated with higher predicted miRNA levels, whereas negative coefficients for both modalities indicate the opposite—lower expression combined with more proximal site usage is statstically associated with lower predicted miRNA levels. These associations reflect predictive relationships and do not imply a specific direction of causality.

To assess whether individual miRNAs exhibit systematic patterns in how expression- and APA-derived contributions relate across their target genes, we summarized the distribution of coefficient sign combinations separately for each miRNA. miRNAs were then stratified according to whether concordant (++ and −−) or discordant (+− and −+) sign patterns predominated among their targets. This grouping was introduced to distinguish miRNAs for which expression and 3′UTR architecture tend to act in the same direction from those in which the two modalities contribute in opposing directions.

Applying this stratification revealed two dominant groups of miRNAs. About 423 HP-miRNAs (52 %) were dominated by the concordant sign patterns (++ and −−), in which expression and APA coefficients share the same sign. The remaining 390 HP-miRNAs (48 %) were dominated by the discordant sign patterns (+− and −+), in which coefficients have opposite signs (Figure 2.4.3G).

This bimodality highlights two prevalent modes by which expression and APA features relate to miRNA levels. Interestingly, these two groups also differed in predictive performance, with mean R2 values of 0.74 and 0.65, respectively (Wilcoxon rank-sum test, $p < 0.01$; Figure 2.4.3H). While the coefficients come from regularised models and cannot be interpreted as direct effect sizes, the systematic separation is consistent with opposing mechanisms such as compensatory biogenesis versus target-directed decay. However, because miRNA–target interactions are intrinsically bidirectional, increased target expression and 3′UTR lengthening may coincide with either higher or lower mature miRNA abundance, and the present framework cannot distinguish whether observed associations reflect dominant target-mediated sequestration, miRNA-driven repression, or a combination of both. Importantly, this does not diminish the biological relevance of the observed patterns, as the reproducible contribution of APA and expression features demonstrates that dynamic changes in target-site availability systematically shape steady-state miRNA levels.

2.4.5.4 Global APA regulation as a determinant of miRNA logFC

To examine how APA modulates miRNA expression dynamics, we analyzed four perturbation experiments in which key APA-regulatory proteins were knocked down and compared with matched controls. These datasets included knockdowns of CFIm25 (two independent experiments), CFIm68, and CPSF6, factors that shape 3′UTR processing and thereby influence miRNA binding-site availability [115]. For each dataset we applied the MIRNAPEX pipeline to predict miRNA log-fold changes based solely on the observed gene-expression changes and APA shifts of their target genes.

Across all four perturbation experiments (CFIm25-KD-1, CFIm25-KD-2, CFIm68-KD, CPSF6-KD) we observed predicted miRNA logFC in both directions, with many exceeding an absolute value of 1 (Figure 2.4.4A). In CFIm25-KD-1 (4 miRNAs up-regulated and 15 down-regulated), CFIm25-KD-2 (6 up and 11 down), CFIm68-KD (20 up and 25 down) and CPSF6-KD (6 up and 14 down), the MIRNAPEX predictions indicated a range of miRNA logFC rather than a uniform shift. Notably, hsa-miR-182-5p showed consistent down-regulation in three of the four experiments.

Since the direction of individual miRNA changes correlated with the expression and APA shifts of their target genes, we investigated how many genes with APA changes also display corresponding differences in gene expression, and how the direction of 3′UTR change differences relates to the gene logFC to reveal the directionality of these effects.

Across the four perturbation experiments, we observed widespread changes in 3′UTR usage, with a clear predominance of shortening events and varying degrees of buffering, expression changes in the opposite direction to the APA effect, likely reflecting compensatory mechanisms such as altered miRNA activity as consequence of binding site modulation.

In CFIm25-KD-1, 6721 genes displayed altered 3′UTR usage (defined as ΔPDUIs ≥ 0.05 for genes with |logFC| ≥ 0.1), with 5325 (79 %) showing shortening; about 1698 (25 %) of these APA-changed genes exhibited opposite (buffering) expression shifts (Figure 2.4.4B). In CFIm25-KD-2, 385 genes showed altered 3′UTR usage, with 286 (74 %) showing shortening and 175 (46 %) displaying opposite expression changes (Figure 2.4.4D). In CFIm68-KD, 6902 genes had altered 3′UTR usage, with 5742 (83 %) showing shortening and roughly 1767 (26 %) exhibiting opposite expression changes (Figure 2.4.4F). In CPSF6-KD, 361 genes displayed altered 3′UTR usage, with 240 (67 %) showing shortening and about 105 (29 %) showing opposite expression changes (Figure 2.4.4H).

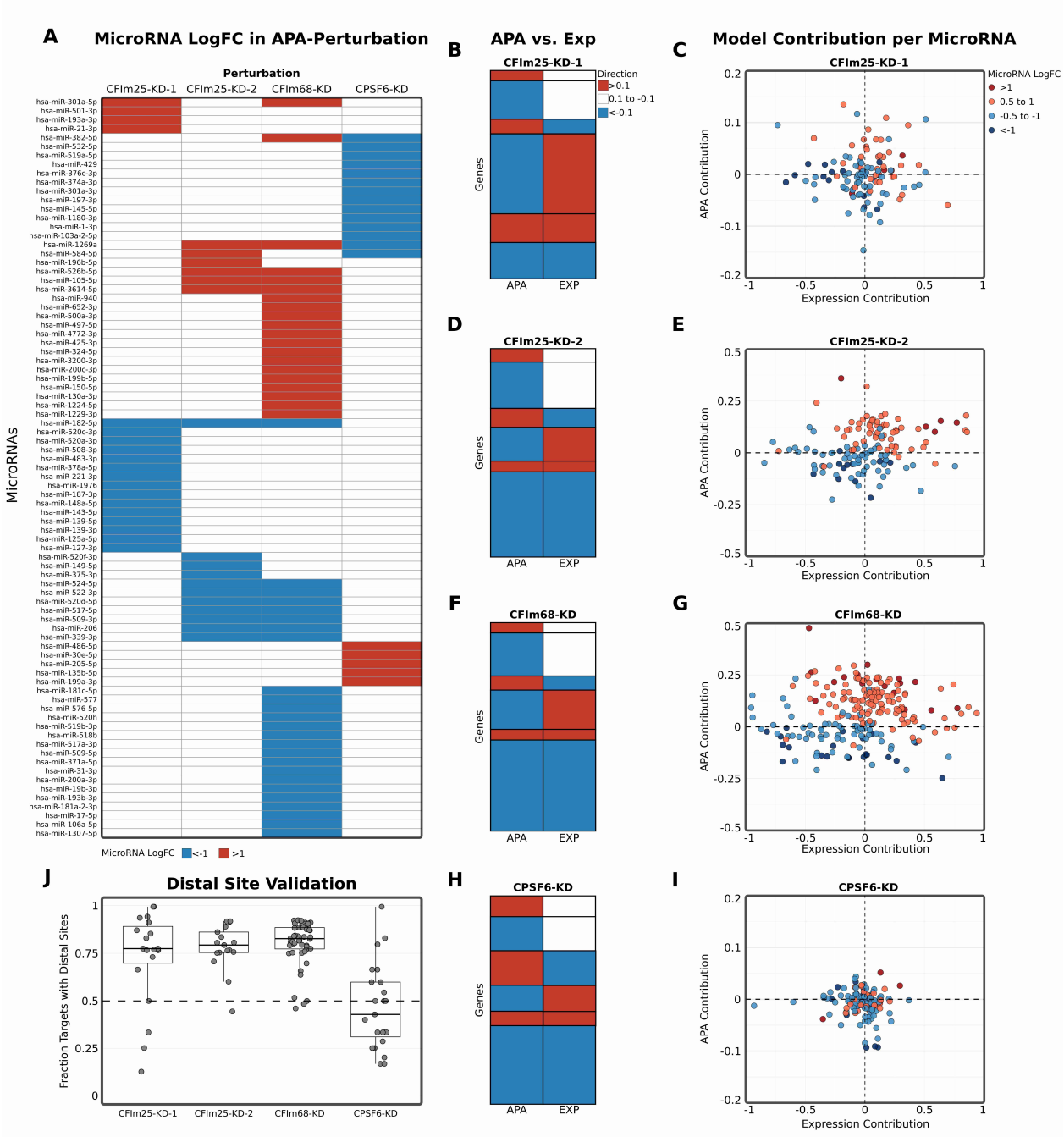


Figure 2.4.4: MiRNA behaviour in APA perturbation experiments. (A) Heatmap of MIRNAPEX-computed miRNA logFC across the four knockdown (KD) experiments of APA-regulatory factors (CFIm25-KD-1, CFIm25-KD-2, CFIm68-KD, CPSF6-KD). Only miRNAs with |logFC|>1 are displayed. (B, D, F, H) Heatmaps display, for each gene with an APA change of |ΔPDUI| > 0.1, the direction of 3′UTR change (APA column) together with the corresponding gene logFC (EXP column). Lengthening of 3′UTRs or positive gene logFC values are indicated in red, shortening or negative gene logFC values in blue, and no expression change in white. This representation highlights the directionality of APA changes relative to gene expression, showing whether 3′UTR lengthening or shortening coincides with increases or decreases in gene logFC across APA perturbation experiments. (C, E, G, I) Scatter plots show, for each miRNA across the perturbation experiments, the combined (unscaled) contribution of gene expression changes versus APA changes to the predicted miRNA logFC as computed by MIRNAPEX. Points are shaded according to the predicted miRNA logFC values across the defined thresholds. (J) Boxplots showing the proportion of APA-changed target genes that harbour at least one distal binding site for the same miRNA (|logFC|>1) in each perturbation experiment.

We found that miRNAs with larger predicted changes cluster in quadrants where the APA component change and the observed miRNA change point in the same direction. In CFIm25-KD-1, CFIm25-KD-2, and CFIm68-KD this concordance is significant (one-sided Fisher's exact test, $p < 0.01$ for miRNAs with $|logFC| > 0.5$), indicating that a stronger APA contribution is associated with larger miRNA shifts and vice versa (Figure 2.4.4C;E;G). In CPSF6-KD no enrichment is observed ($p = 0.53$), consistent with this perturbation showing the lowest fraction of 3′UTR-shortened genes among the datasets considered Figure 2.4.4I). This pattern further emphasizes that extensive gene shortening in APA perturbations coincides with the largest shifts in miRNA levels and that, where shortening is limited, the APA component contributes less strongly to miRNA log fold-changes.

For each miRNA with a predicted change, we assessed whether the 3′UTR shortening or lengthening of its target genes in the same experiment is associated with altered availability of binding sites for that specific miRNA. Using the microT predictions, APA-changed genes were screened for the presence of at least one binding site for the same miRNA in the region between the proximal and distal polyadenylation sites. This analysis showed that in CFIm25-KD-1, CFIm25-KD-2 and CFIm68-KD a substantial fraction of shortened targets indeed contained a distal binding site for the same miRNA, with median values of 77.8%, 79.7% and 83.1% of APA-changed genes, respectively. In CPSF6-KD the median proportion was much lower (42.9%), consistent with the weaker shortening seen in this dataset (Figure 2.4.4J). These results validate that the predicted miRNA shifts reflect real changes in binding-site dosage caused by APA remodeling. To further validate PDUI as a proxy for miRNA binding-site availability, we extended this analysis to all expressed miRNAs in each perturbation dataset, independent of their predicted logFC. Across datasets, APA-regulated target genes frequently harbored distal binding sites for the corresponding miRNA, with median fractions of 72.7% in CFIm25 KD—HCT116, 70.7% in CFIm25 KD—HEK293, 70.7% in CFIm68 KD—HEK293, and 56.5% in CPSF6 KD—HEP3B. These genome-wide results support PDUI as a meaningful measure of miRNA binding-site dosage in the present analyses.

In summary, our findings show a strong statistical concordance between APA-driven target shortening and miRNA logFC. This suggests that at least part of the apparent miRNA DE we observe may reflect changes in binding-site availability rather than direct changes in miRNA transcription.

2.4.5.5 APA-driven loss of TDMD trigger sites coincides with miRNA abundance shifts

As APA-perturbation experiments globally shorten 3′UTRs, we asked whether this remodeling also affects established TDMD interactions. We therefore investigated eight curated trigger–miRNA pairs, as well as one negative control, in which the trigger transcript harbours highly complementary sites known to direct miRNA decay and for which we observed APA changes of the trigger gene. These included CYRANO with hsa-miR-7-5p, SDC2 with hsa-miR-15a-5p, SERTAD3 with hsa-miR-92a-3p, SSR1 with hsa-miR-218-5p, TRIM9 with hsa-miR-218-5p, TDP1 with hsa-miR-320a-3p, NREP with hsa-miR-29b-3p, and BCL2L11 with hsa-miR-221-3p, alongside BCL2L11 with hsa-miR-221-5p as a negative control [41,361,362].

For each perturbation dataset we extracted the trigger genes, quantified their ΔPDUI to confirm 3′UTR shortening, and compared these changes with MIRNAPEX-predicted log fold-changes of the respective miRNAs (Figure 2.4.5A-I). This analysis directly tests whether loss of distal 3′UTR regions containing TDMD sites under APA perturbation translates into increases in the abundance of the targeted miRNAs.

Across the majority of TDMD pairs, negative ΔPDUI values of the trigger gene, indicative of 3′UTR shortening, coincided with increased mature miRNA abundance. This trend was particularly evident for pairs involving CYRANO with hsa-miR-7-5p (Figure 2.4.5A), SDC2 with hsa-miR-15a-5p (Figure 2.4.5B), SSR1 with hsa-miR-218-5p (Figure 2.4.5D), TRIM9 with hsa-miR-218-5p (Figure 2.4.5E), TDP1 with hsa-miR-320a-3p (Figure 2.4.5G), and BCL2L11 with hsa-miR-221-3p (Figure 2.4.5H). In these cases, the strongest miRNA up-regulation was observed in datasets exhibiting pronounced trigger 3′UTR shortening, consistent with reduced TDMD-mediated degradation following loss of distal trigger regions [34]. By contrast, SERTAD3 with hsa-miR-92a-3p (Figure 2.4.5C) did not exhibit consistent 3′UTR shortening across perturbations and accordingly showed decreased miRNA abundance in conditions associated with 3′UTR lengthening, supporting the directional relationship between trigger 3′UTR architecture and miRNA stability. NREP with hsa-miR-29b-3p (Figure 2.4.5F) deviated from the general trend, displaying divergent miRNA responses across perturbations despite trigger shortening, suggesting that additional regulatory inputs or context-dependent effects may modulate TDMD efficiency for this pair.

Importantly, the negative control pair BCL2L11 with hsa-miR-221-5p (Figure 2.4.5I) did not show systematic miRNA up-regulation despite APA changes of the trigger gene, indicating that TDMD sensitivity is arm-specific and reinforcing that the observed effects are not a general consequence of trigger gene expression changes.

To distinguish post-transcriptional effects from altered miRNA production, we additionally compared mature miRNA logFC values with transcriptional proxy measurements derived from host-gene or pri-miRNA-associated expression estimates (hollow circles in Figure 2.4.5A–F). For all TDMD pairs showing increased mature miRNA abundance under trigger shortening, the transcriptional proxy logFC was lower or unchanged, indicating that the observed miRNA up-regulation cannot be explained by increased transcription. In contrast, NREP with hsa-miR-29b-3p showed no consistent separation between transcriptional proxy and mature miRNA changes, in line with its context-dependent behaviour.

Taken together, these results indicate that MIRNAPEX captures TDMD-linked miRNA behaviour in the majority of curated cases, and that APA-driven loss of distal trigger regions is frequently associated with increased mature miRNA abundance independent of transcriptional changes. This supports the view that a substantial fraction of miRNA expression changes observed under APA perturbation reflects altered TDMD site availability rather than solely changes in miRNA transcription.

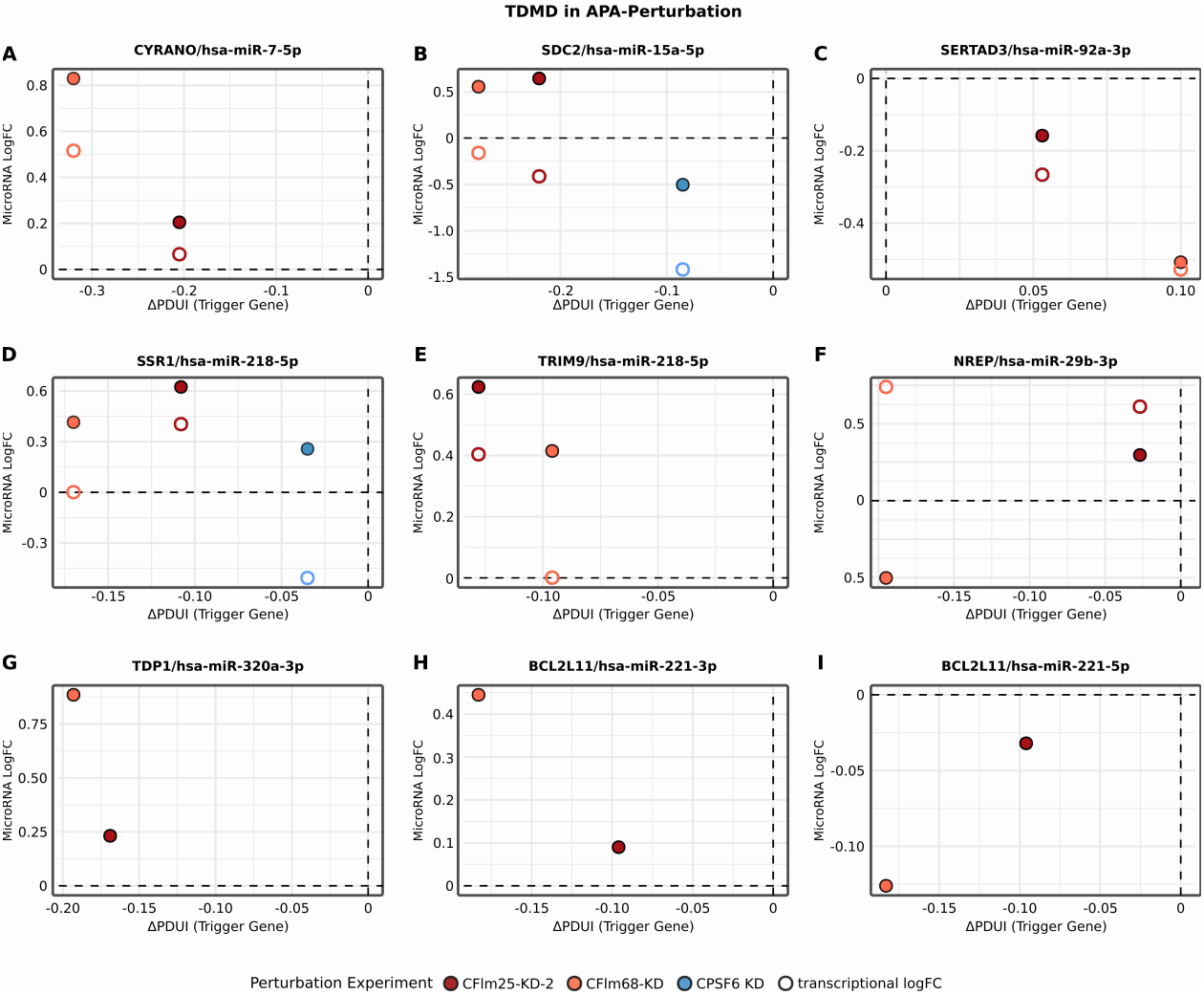


Figure 2.4.5: APA-driven 3′UTR shortening of TDMD trigger genes and miRNA abundance changes. Scatter plots relating APA changes of TDMD trigger genes to the log fold-change of their paired mature miRNAs across APA-perturbation datasets. Each panel corresponds to one trigger–miRNA pair: (A) CYRANO with hsa-miR-7-5p, (B) SDC2 with hsa-miR-15a-5p, (C) SERTAD3 with hsa-miR-92a-3p, (D) SSR1 with hsa-miR-218-5p, (E) TRIM9 with hsa-miR-218-5p, (F) NREP with hsa-miR-29b-3p, (G) TDP1 with hsa-miR-320a-3p, (H) BCL2L11 with hsa-miR-221-3p, and (I) BCL2L11 with hsa-miR-221-5p (negative control). Filled symbols indicate mature miRNA logFC values predicted by MIRNAPEX, while open symbols represent transcriptional proxy logFC estimates. Points are coloured by perturbation dataset. The x-axis shows ΔPDUI values (KD − control), with negative values indicating 3′UTR shortening, and the y-axis shows log fold-changes.

2.4.6 Discussion

Steady-state miRNA levels arise from a dynamic balance of biogenesis, target engagement and decay, including TDMD, yet DE of miRNAs is often interpreted as evidence of altered post-transcriptional regulation [16,34,345]. Within this balance, APA reshapes 3′UTR isoform usage and therefore affects effective dosage of canonical binding sites and highly complementary decay triggers [115,363]. To test whether such transcriptome remodeling predicts miRNA logFCs between conditions, we developed MIRNAPEX, which reads out target-centric features derived jointly from gene expression changes and APA.

Across miRNAs, features derived solely from APA were predictive on their own and, when combined with mRNA expression, consistently improved performance, showing that 3′UTR remodeling contributes information absent from total transcript levels. Mechanistically this is consistent with the idea that site dosage and binding strength together determine AGO occupancy and repression efficacy [18,364,365].

From a gene-centric perspective, many targets contributed to prediction mainly through expression features, reflecting changes in total abundance and baseline miRNA binding site load, while others were dominated by APA, consistent with isoform switches that add or remove distal sites or decay triggers without large changes in total transcript levels [115].

Both target-gene expression and APA influence predicted miRNA logFC: expression dominates overall, but APA leads for many targets. Interestingly, miRNAs fall into two behavior classes: in one, expression and APA effects align, more transcript and longer 3′UTRs go with higher miRNA levels, and in the other, they oppose, so increased site availability is associated with lower miRNA. This bimodal pattern suggests distinct regulatory modes for different miRNAs, not just variation in target abundance; such modes are consistent with competition and sequestration effects and observed differences in miRNA-mRNA network behavior in recent studies [129].

Applying MIRNAPEX to experimental data where core APA regulators were knocked down revealed broad 3′UTR shortening, consistent with the known global impact of APA perturbations, with variable buffering at the expression level and predicted miRNA changes in both directions [119,120,123]. Decomposing predictions showed that the largest miRNA shifts clustered where the APA contribution and miRNA direction agreed, especially in perturbations that induce extensive shortening, whereas when shortening was limited, the contribution was weaker. Screening shortened targets confirmed that a substantial majority harbored distal sites for the same miRNA in these datasets, validating that MIRNAPEX effects stemming from site dosage rather than spurious correlations.

For TDMD trigger–miRNA pairs such as BCL2L11–miR-221-3p, NREP–miR-29b-3p, SSR1–miR-218-5p, TDP1–miR-320a-3p and TRIM9–miR-218-5p, 3’UTR shortening generally coincided with higher predicted miRNA abundance, consistent with relief from decay [34,41]. TDMD triggers located in the 3′UTR have been shown to degrade miRNAs more effectively than identical triggers placed in coding sequences [361], which supports our focus on 3′UTR site loss in interpreting miRNA changes. A recent study found an endogenous TDMD trigger with minimal non-canonical 3′-end base-pairing that is nevertheless sufficient to induce degradation of the miR-279 family [366]. This suggests that many APA-associated miRNA expression changes could reflect widespread but previously uncharacterized TDMD triggers captured by MIRNAPEX. These findings further indicate that MIRNAPEX is sensitive to mechanistically defined decay events embedded within broader APA remodeling.

MIRNAPEX has some important limitations. It relies on bulk RNA-seq–derived APA metrics and selection of features based on predicted target sets, which are imperfect proxies for the true binding landscape. Bulk data also mix cell types and states, so shifts in composition could confound expression and 3′UTR usage. Furthermore, coefficients from regularized models help interpretation but don’t

explain causal effect sizes and cannot fully separate biogenesis from decay. Finally, curated TDMD interactions are incomplete and context-dependent, limiting ground-truth validation.

Integrating direct AGO-binding data with isoform-resolved APA profiles and more detailed predictors of site efficacy will sharpen our understanding of how target landscapes influence miRNA levels. Applying such analyses at single-cell or time-course resolution could help separate cell-state effects from true regulatory changes, while controlled perturbations with matched mRNA, APA and small-RNA measurements would provide stricter benchmarks for testing mechanistic hypotheses.

Our findings suggest that a more comprehensive view of miRNA regulation can be obtained when dynamic changes in target-site availability and decay processes are explicitly taken into account. Incorporating these dimensions has the potential to improve the quality of commonly applied analysis workflows, strengthen functional interpretations and increase the reliability of biomarker discovery.

2.5 Genomic variation of human microRNAs and its association with functional features

2.5.1 Preamble

This chapter is published in the journal Cellular and Molecular Life Sciences:

Cihan, M., Andrade-Navarro, M.A. & Morett, E. Genomic variation of human microRNAs and its association with functional features. Cell. Mol. Life Sci. 82, 411 (2025). https://doi.org/10.1007/s00018-025-05936-x

The supplementary files associated with this publication are available on the publisher's website via the article's DOI.

2.5.2 Abstract

Current methods for annotating microRNAs (miRNAs) often rely on phylogenetic conservation or expression data, with less attention paid to the impact of human genetic variation. This limits our understanding of how variation shapes miRNA function and regulatory dynamics across populations. In this study, we systematically annotated genomic variants within human miRNAs and investigated their relationship to functional features and evolutionary constraint. To facilitate this, we developed a population-based conservation metric that integrates allele frequency and positional coverage across miRNA loci. We show that miRNA conservation effectively links to functional roles, as evidenced by associations with higher expression levels, broader target gene regulation, and enrichment in essential biological pathways. Conserved miRNAs also preferentially target genes with fewer alternative polyadenylation sites, indicating more stable and consistent regulatory interactions. This trend is also reflected in miRNAs with both 5p and 3p arms, where the more conserved arm typically regulates more targets, especially when conservation differences between arms are pronounced. Moreover, we find that miRNA genomic variants display population-specific patterns, often co-occurring with target site variants to form compensatory pairs that preserve base-pairing. These events suggest co-evolution and several involve pathogenic variants, indicating that some deleterious regulatory disruptions in target genes may be mitigated through compensatory changes in miRNAs that restore binding.

2.5.3 Introduction

Genomic variability, driven by changes in both protein-coding and non-coding regions, governs key biological processes that influence cellular functions and shape phenotypic diversity across populations [367]. Specifically, variability in non-coding regions significantly influences gene expression by affecting chromatin accessibility, altering mRNA stability, affecting transcription and translation, and modulating key post-transcriptional mechanisms [368,369]. Among these mechanisms, microRNAs (miRNAs)—small non-coding RNAs—emerge as a critical layer of regulation by binding to complementary sequences on target mRNAs, thereby directing their degradation or repressing their translation [290].

While the initial detection and annotation of miRNAs were primarily guided by criteria such as phylogenetic conservation of sequence and hairpin structure [370], more recent efforts have integrated large-scale small RNA sequencing data to address the complexity of miRNAs. These advances have introduced criteria such as

precise read mapping profiles, consistency of mature miRNA 5′ ends, and duplex characteristics, which allow for higher precision in the identification of species-specific miRNAs alongside conserved ones [7]. Challenges in miRNA identification persist due to biases from RNA sequencing quality and depth, leading to misannotation [371]. Additionally, inconsistencies arise from the lack of standardized identification criteria across species, as thresholds for conservation and sequencing read counts vary, complicating accurate identification and comparisons of lineage-specific miRNAs [372]. Moreover, species-specific miRNAs with low conservation and low read profiles are particularly prone to be identified as false positives, necessitating rigorous experimental confirmation to validate computational predictions [7].

While accurate annotation of miRNAs is critical to avoid false positives, the outcome of miRNA function are further shaped by sequence variability in the population. These genetic variants can occur both within the miRNA itself or their target sites, altering seed specificity and therefore target recognition, by disrupting or creating new binding regions in mRNA 3′UTRs [53,373]. Consequently, variants can lead to changes in gene expression patterns, contributing to phenotypic diversity and disease, with roles in oncogenesis, metabolic disorders, and inflammatory conditions through the modulation of regulatory networks and protein levels [61].

The genomic variability of miRNAs within human populations highlights a clear distinction between high-confidence miRNAs, which show minimal genetic variation likely due to their essential regulatory roles, and species-specific miRNAs, which exhibit greater variability as they adapt to lineage-specific functions [54].

With the availability of extensive human genomic variation data from gnomAD v4.1 [374], we aim to investigate miRNA conservation within human populations by developing a scoring method that evaluates the authenticity and functional relevance of miRNAs. This scoring approach reflects their biological relevance by assessing variant frequency and distribution in relation to their location within the mature miRNA. We further expanded these analyses by examining multiple factors, including target interaction dynamics, and genomic variation characteristics. These include aspects such as annotating target genes, functional enrichment, alternative polyadenylation patterns, co-evolution with targets, and variant properties such as transition-transversion and InDels ratios. Our method aims to enhance our understanding of miRNA function in a context of genomic variant distribution and points towards potential misannotations of miRNAs.

2.5.4 Methods

2.5.4.1 Data collection and processing

We obtained genomic variation data from gnomAD v4.0 [374], which includes 76,215 Whole Genome Sequencing (WGS) and 730,947 Whole Exome Sequencing (WES) datasets. We filtered these datasets to retain only genomic variants corresponding to alleles located within the genomic regions of 2,883 mature human miRNAs and 1,918 precursor miRNAs, as defined by miRBase [7], considering only alleles that passed quality control for genomes. The seed region of each miRNA was defined as positions 2–7 within the mature miRNA sequence. Joint allele frequencies (AF) from gnomAD [374], selected for presence in WGS data but optionally in WES

variant data without discrepant AF, were used for downstream analyses. For 156 mature miRNAs there are no reported genomic variants that pass the filters. We consider these miRNAs as having a conservation score of 1 but excluded them from downstream analyses, as they do not provide any population-specific information.

2.5.4.2 Conservation score computation

To quantify conservation for each miRNA, we computed an Overall Conservation Score (OCS). This score integrates parameters that capture conservation within the seed region, outside the seed region, the number of positions with variants, and the total number of variants across the miRNA. The Seed Conservation Score (SCS) is defined as

$$SCS = 1 - \frac{\sum (f_s)}{n_s}$$

where fs represents the AF of a seed variant and ns is the total number of seed variants. Similarly, the Non-Seed Conservation Score (NSCS) quantifies conservation outside the seed region, using the same formula but applied to non-seed variants, defined as

$$NSCS = 1 - \frac{\sum (f_{ns})}{n_{ns}}$$

where fns represents the AF of a non-seed variant and nns is the total number of non-seed variants. The Positional Coverage Score (PCS) evaluates the proportion of variant-free positions within the miRNA and is defined as

$$PCS = 1 - \frac{p_v}{L}$$

where pv is the number of unique positions with variants and L is the total length of the miRNA. The Total Variants Score (TVS) assesses the overall burden of variation and is defined as

$$TVS = 1 - \frac{v_t}{L}$$

where vt is the total number of variants across the miRNA and L is the total length of the miRNA. The OCS combines these scores with weights assigned to SCS, NSCS, PCS, and TVS, respectively:

$$OCS = (\alpha * SCS) + (\beta * NSCS) + (\gamma * PCS) + (\delta * TVS)$$

These parameters capture conservation in the seed region, outside the seed region, positional coverage, and the overall burden of genomic variation, respectively.

To determine the optimal weights, we used the expression-conservation overlap as a robust estimation of the functional relevance of miRNAs. Specifically, we aim to maximize the overlap between the most expressed and most conserved miRNAs for various weight combinations. The tested weight combinations are constrained to

sum up to 1, with the additional conditions that the weight for the seed region (α) must be higher than the weight for the non-seed region (β) and that each weight must be greater than 0, varying in increments of 0.1.

To ensure a robust evaluation, we consider two independent datasets reporting miRNA expression: the miRNATissueAtlas2 [174] and the Adult Genotype-Tissue Expression (GTEx) database [375]. For each weight combination, we tested thresholds of 5%, 10%, 15%, 20%, and 25% to capture the computed overlaps of the high conserved and top-expressed miRNAs. Although certain miRNAs may be conditionally expressed, those consistently detected across tissues are likely to be genuine and functionally relevant. We therefore used these subsets to find the weight parameters. At the same time, it does not exclude conditionally expressed miRNAs from being identified as highly conserved and such miRNAs can still receive high OCS values based on their variant profiles, regardless of their expression levels.

For each weight combination and threshold, we calculate the average F1-score across the two datasets, integrating both precision and recall to evaluate conservation-expression overlap. To identify the most robust weight combination, we computed the z-score of the average F1-score for each weight combination relative to all others at each threshold. The z-scores quantify how much a weight combination deviates from the mean performance at a given threshold. Finally, we calculated the average absolute z-score across all thresholds for each weight combination. The weight combination with the highest average absolute z-score is selected as the optimal choice, as it consistently outperforms others relative to the mean, highlighting its robust performance across diverse biological contexts. Sequences identical across multiple genomic loci were collapsed under a single mature miRNA name, and the final OCS was computed at the level of mature miRNA sequences.

2.5.4.3 miRNA expression

We obtained a comprehensive dataset of 42,494 RPM-normalized miRNA expression values across multiple tissues and conditions from isomiRdb [376], and compared OCS values to the mean expression across the entire dataset. Additionally, tissue average expression values for 2,656 miRNAs were obtained from miRNATissueAtlas 2 [174]. The small RNA-sequencing data from the GTEx Portal used for the analyses described comprises 2,564 miRNAs and were obtained on December 16 2024 [377] . While GTEx and miRNATissueAtlas databases were utilized during the weight optimization process, isomiRdb expression data served as the primary dataset for downstream analyses.

2.5.4.4 miRNA targets

Human miRNA targets based on the hg38 human reference genome were downloaded from the microT database [157]. Only binding sites within the 3'UTR regions of genes were considered, and interactions were restricted to those with a default gene-miRNA interaction score of at least 0.7 and a miRNA recognition element (mre) score of 0.01. This resulted in 11,421,667 annotated miRNA binding sites for 183,257 unique miRNA-gene pairs. In addition, we downloaded all conserved miRNA binding sites from TargetScanHuman 8.0 [16] and cross-validated

our findings by comparing the number of targeted genes across high- and low-conservation miRNA groups with respect to the presence of APA sites.

2.5.4.5 Polyadenylation sites

Polyadenylation sites mapped on hg38 were obtained from PolyASite 2.0 [256]. Only polyadenylation sites located in exon regions of protein-coding genes were considered in the analysis. The representative alternative polyadenylation (APA) site for each gene was defined in PolyASite 2.0 as the position with the highest read support among all APA sites.

2.5.4.6 Computation of compensatory variant pairs

Genomic variation in miRNA seed regions or in their binding sites in target genes can disrupt regulatory interactions. We aim to investigate whether variations in miRNA seed regions and their corresponding binding positions exhibit potential co-evolution patterns, ensuring that despite sequence variation, the binding pair remains preserved.

To do this, we obtained 1,851,543 unique miRNA-gene pairs with experimental support from miRTarBase [168] and TarBase-v9 [378]. We then retrieved 3' UTR sequences from the hg38 reference genome using BioMart [379], selecting the longest 3' UTR as the representative sequence. Next, we aligned miRNA seed sequences from miRBase [7] to these 3' UTR sequences, identifying all matching 6mer, 7mer-A1, 7mer-m8 and 8mer miRNA binding sites [228].

By definition, this alignment guarantees at minimum a perfect match between positions 2–7 of the miRNA seed and the target site (6mer). A genomic variant occurring in either the miRNA or the target gene alone would always introduce a mismatch, potentially disrupting the binding. However, when simultaneous variants occur at aligned positions in both the miRNA seed and its 3' UTR binding site, their effect depends on whether they maintain or disrupt base-pairing. If the variants preserve complementary base-pairing, they are classified as compensatory variant pairs, suggesting potential co-evolution. In contrast, if complementary base-pairing is not preserved, they are considered disruptive and likely interfere with miRNA-target regulation. We identified 66,240 compensatory variant pairs, indicating a potential evolutionary relationship between miRNAs and their targets (Supplementary Table 2).

To assess whether these compensatory variant pairs are population-specific, we analyzed AF data from gnomAD, which reports the AF for each variant and each population. A compensatory variant pair was classified as population-specific if both variants reported their highest AF within the same population rather than in any other population. Using this criterion, we identified 14,869 population-specific compensatory variant pairs, while the remaining 51,371 compensatory variant pairs did not exhibit this population-specific pattern.

2.5.4.7 Functional enrichment and disease association of miRNAs

To assess the functional relevance of miRNAs, we performed enrichment analysis using MiEAA 2.0 [380], accessed via rbioapi [381], on the top 5% highest conserved (HC) and 5% least conserved (LC) miRNAs based on OCS; each group corresponding to 124 mature miRNAs. The enrichment profiles of over- and under-represented

terms were compared to identify patterns of functional divergence and overlap. Categories analyzed included GO terms, pathways, and disease associations and cover only significant terms, based on adjusted p-values. Clinical significance was assessed by cross-referencing variant positions with the ClinVar database [382].

2.5.5 Results

In this study, we investigated genomic variants within miRNAs and their target sites obtained from gnomAD v4 across different ancestry groups in the human population. To quantify miRNA conservation, we applied a scoring method that integrates AF and positional impact of variants within seed and non-seed regions. We further evaluated this conservation score by comparing HC- and LC-miRNAs (highest and least conserved miRNAs, respectively) in terms of their expression profiles, targeting dynamics, and functional roles. By analyzing allele types and frequencies across ancestry groups, we identified population-specific variants and assessed their potential functional impact. Rare and common variants were characterized to explore their relevance in regulatory processes and potential links to disease. This comprehensive analysis reveals how genomic variation shapes the regulatory roles of miRNAs across human populations, providing insights into their functional significance.

2.5.5.1 Characterization of miRNA variants

We identified 55,796 miRNA variants, of which 32,713 mapped in precursor miRNAs and 23,083 in mature miRNA regions. Among the mature miRNA variants, 16,707 were located in the non-seed region, while 6,376 were in the seed region. These variants were derived from WGS and WES data. Because many miRNAs are intronic and therefore not consistently captured by exome sequencing, discrepancies between WGS- and WES-derived variant counts are expected [5]. After filtering for variants detected in WGS, the total number of variants was reduced to 27,261, of which 10,677 mapped to mature miRNAs: 7,765 in the non-seed region and 2,912 in the seed region (Table 2.5.1). On average, one genomic variant occurs every 2.73 base pairs (WGS: 1 SNV every 5.55 bp), with no significant differences between miRNA regions (precursor: 2.83 bp, non-seed: 2.58 bp, seed: 2.59 bp). We categorized variants by their AF as common (AF>0.05), single nucleotide polymorphisms (SNPs; 0.01<AF<=0.05) or rare (AF<=0.01). Among mature miRNAs, we found 10,511 rare variants, 98 SNPs, and 68 common variants. Regarding SNPs, precursor miRNAs exhibited a significantly higher SNP density than mature miRNAs (Fisher's Exact Test, p = 0.006). and notably, no SNPs were detected in the HC-miRNAs (Figure 2.5.1a). Within precursor miRNA regions, we identified 32,378 rare variants, 127 SNPs, and 208 common variants from WGS and WES data. Region-specific mapping revealed 7,701 variants in the 5′ lower stem (1 variant every 2.82 bp), 6,313 in the 3′ lower stem (1 variant every 2.99 bp), and 14,152 in the terminal loop (1 variant every 2.48 bp). For hairpins with only one mature strand annotated, we inferred the opposite arm using the canonical 2-nt 3′ overhang to define the loop and lower-stem boundaries [383]. In the flanking regions of precursor miRNAs, we annotated 10,479 variants upstream (1 variant every 3.11 bp) and 10,834 variants downstream (1 variant every 3.01 bp) of the precursor on the primary transcript.

A total of 1,710 variants were identified as insertions or deletions (InDels) (Figure 2.5.1b), with 1,093 occurring in precursor miRNAs, where the InDels density is

significantly higher compared to mature miRNA regions (Fisher's Exact Test, $p = 0.007$). In the mature miRNA regions, 447 InDels were found in non-seed regions, while 170 were located in the seed region. Within mature miRNAs, the distribution of variants remained similar between seed and non-seed regions (Figure 2.5.1c). However, when comparing HC-miRNAs with LC-miRNAs, consistently fewer variants are observed in the highly conserved group. Moreover, within highly conserved miRNAs, positions 7 and 16 showed a notably lower number of variants after accounting for miRNA coverage (Figure 2.5.1d).

Among the annotated variants, we computed the transition-to-transversion (Ti/Tv) ratio across allele frequency thresholds in miRNAs. Seed regions exhibit a higher Ti/Tv ratio than non-seed regions for very rare variants ($AF \leq 0.002$), but this trend shifts at higher allele frequencies. Furthermore, precursor regions consistently show lower Ti/Tv ratios than the combined mature miRNA regions. (Figure 2.5.1e). We next examined RNA-level mismatches introduced by transitions. Among mature miRNAs, we found 3,971 pyrimidine transitions (C:U), 3,856 purine transitions (A:G), and 1,163 G:U wobble mismatches (Figure 2.5.1g). Across allele frequency thresholds, A:G transitions were more common in non-seed regions, whereas C:U transitions were enriched in seed regions. G:U wobbles predominated in non-seed regions at low allele frequencies but converged between seed and non-seed regions at higher frequencies (Figure 2.5.1f).

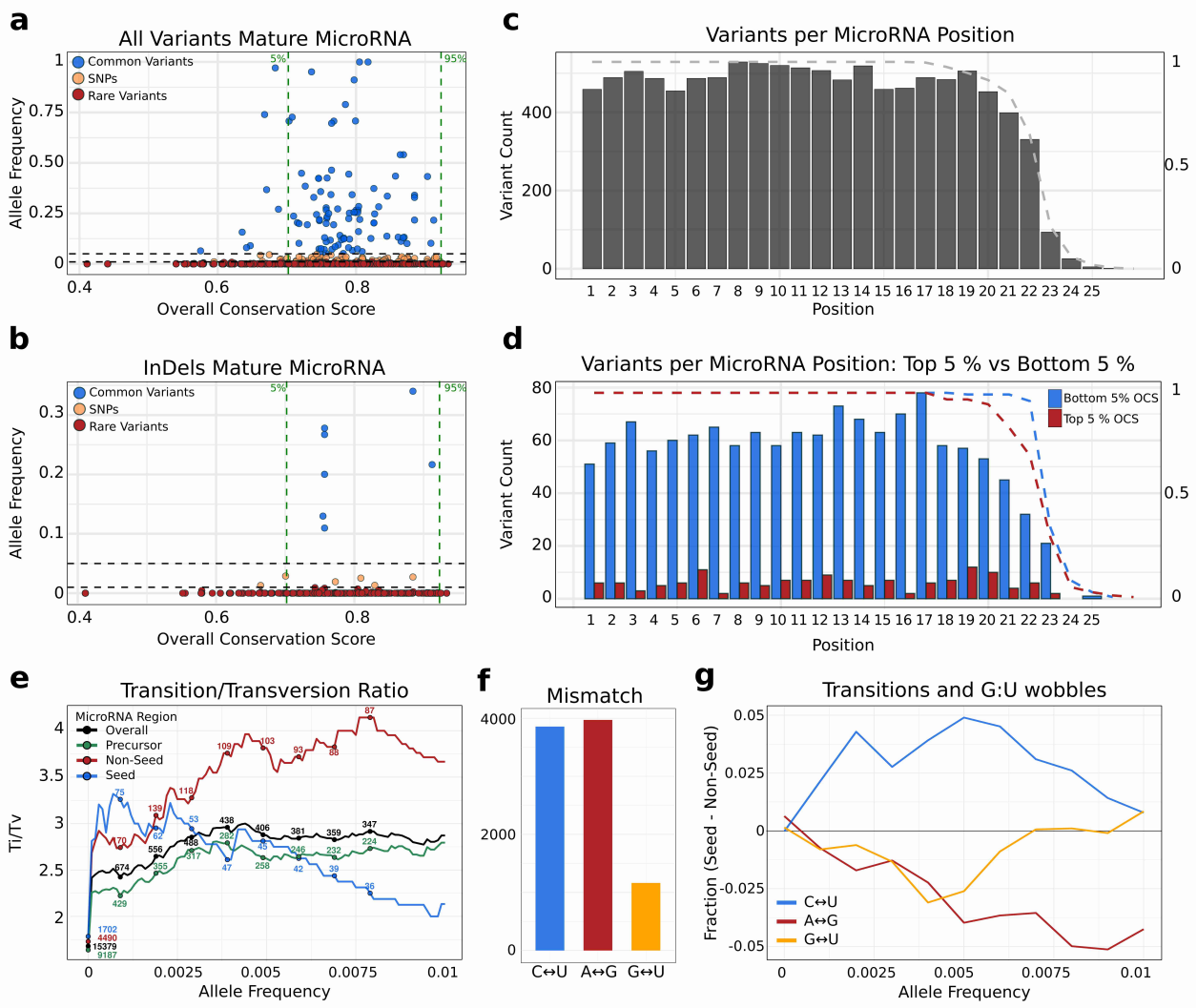


Figure 2.5.1: MiRNA Variant Distribution and Conservation in Human Populations a) Distribution of AF for all genetic variants in mature miRNA regions. Variants are categorized based on AF thresholds: common variants (AF>0.05), SNPs (0.01<AF<=0.05), and rare variants (AF<=0.01). The 5% and 95% quantiles for miRNA OCS are indicated. b) Distribution of insertions and deletions (InDels) in mature miRNA regions, shown with Overall Conservation Scores. Variants are colored according to AF as in a). c) Variant count per position in mature miRNA regions. The dotted line indicates the fraction of mature miRNAs covering a given position length. d) Same plot as c but for the top 5% most conserved and least conserved mature miRNAs (124 in each category). e) Transition to transversion ratios of AF across different regions: all variants, precursor regions (excluding mature miRNAs), non-seed regions and seed regions of mature miRNAs. The number of variants analyzed in each category is labeled. f) Barplot showing the total number of transition types and G:U wobbles in mature miRNAs. g) Seed–Non-Seed differences in mismatch type frequencies across allele-frequency thresholds. Positive values indicate higher fractions in seed regions and negative values indicate higher fractions in non-seed regions.

Table 2.5.1: Genetic Variants in miRNAs. An overview of 55,796 genetic variants classified into distinct regions—precursor, mature, and seed—based on their location within miRNAs, further stratified by sequencing method (genome or exome) and human population groups as annotated by gnomAD database. Variants from genome-sequencing and exome-sequencing can be overlapping.

	Joint	WGS	WES	Joint	WGS	WES	Joint	WGS	WES
African/African American	7,791	6,711	3,171	3,688	3,122	1,496	1,450	1,197	638
Admixed American	5,697	3,556	3,632	2,705	1,580	1,831	1,019	614	690
Ashkenazi Jewish	1,253	731	953	506	301	387	189	115	152
East Asian	3,823	1,989	2,603	1,856	837	1,364	746	349	554
European (Finnish)	2,093	1,074	1,652	970	406	804	354	171	293
European (non-Finnish)	19,309	8,395	15,056	9,935	3,878	8,158	3,837	1,457	3,182
Middle Eastern	1,718	591	1,566	817	227	754	290	81	277
South Asian	6,900	2,196	5,967	3,414	926	3,020	1,350	335	1,230
Remaining	5,102	1,678	4,409	2,547	682	2,276	986	259	897
Total	32,713	16,584	24,149	16,707	7,765	13,066	6,376	2,912	4,986

2.5.5.2 Conservation scoring and miRNA confidence analysis

To quantify miRNA conservation from human population variant data, a scoring approach was developed that integrates four key parameters: allele frequency-based conservation in the seed and non-seed regions of mature miRNAs (SCS and NSCS, respectively), positional coverage (PCS), and total number of distinct variants. These four parameters were weighed and combined into an Overall Conservation Score (OCS). We evaluated different weight combinations (34 combinations; see Methods for details) by the overlap between highly scored and highly expressed miRNAs at various thresholds. The hypothesis here was that miRNAs with demonstrated and clear expression are very likely real; conversely, miRNAs that show low or no expression in many samples have high probability of not being real. Importantly, this approach does not exclude conditionally or cell-type–specific miRNAs from being identified as highly conserved, as such miRNAs can still score highly based solely on their variant profiles, independent of their expression levels. The most effective weight combination assigned the highest importance to PCS (0.6), followed by SCS (0.2), NSCS (0.1), and TVS (0.1). This weight combination produced the highest average overlap between highly conserved and highly expressed miRNAs across all thresholds, emphasizing the critical role of positional variant burden in defining conservation. In particular, this weight combination captured the HC-miRNAs better at the threshold of 5%. OCS computation relying mainly on AF performed suboptimally, indicating that it is insufficient for accurately predicting conservation (Figure 2.5.2a).

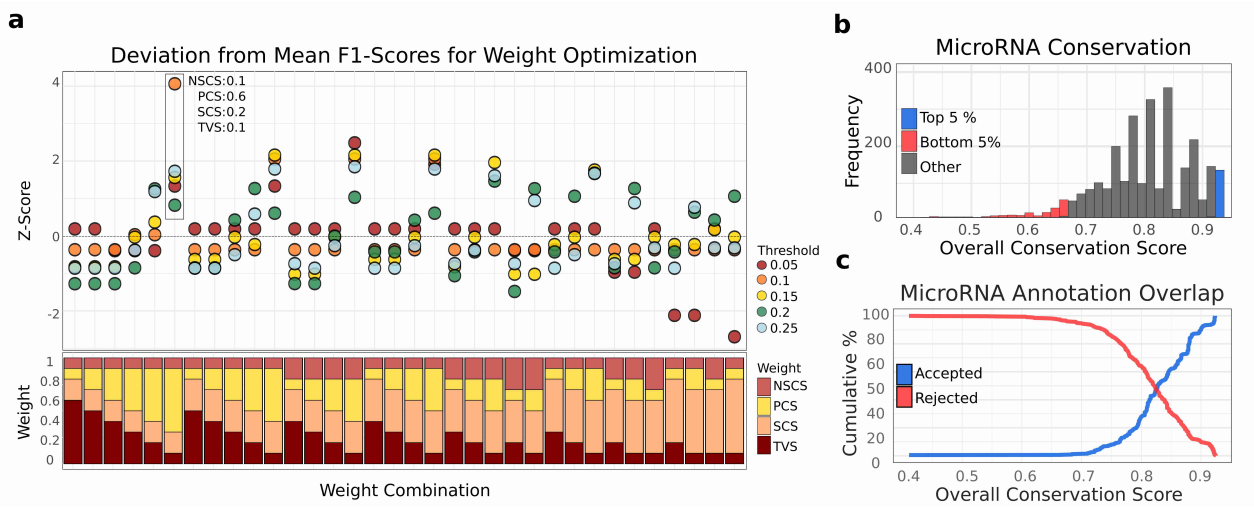


Figure 2.5.2: Weight Optimization and Overall Conservation Score a) Weight Optimization for Seed Conservation Score (SCS), Non-Seed Conservation Score (NSCS), Positional Coverage Score (PCS), and Total Variants Score (TVS). The stacked bar plot at the bottom visualizes the weight combinations for these scores, ensuring the weights sum to 1. For each weight combination, we computed the overlap of conserved miRNAs (based on the computed OCS) with the top-expressed miRNAs derived from the miRNA Tissue Atlas and GTEx expression data, evaluated at different thresholds on the OCS. At each threshold, F1 scores for the overlaps were calculated, and a z-score was computed to indicate the deviation from the average F1 score across thresholds, identifying the optimal weight configuration (see Methods for details). b) Distribution of OCS across all miRNAs. The top 5% most conserved miRNAs, as defined by OCS values, are highlighted in blue, while the bottom 5% are highlighted in red. c) Cumulative Coverage of miRNA Confidence by OCS. Line plots illustrate the cumulative percentage of miRNAs from mirGeneDB, categorized by their confidence (accepted; rejected), as covered by OCS.

The second-best weight combination (PCS = 0.4, SCS = 0.3, NSCS = 0.2, TVS = 0.1) outperformed those assigning an even higher weight to PCS (e.g., PCS = 0.5) at higher thresholds. This suggests that while positional conservation is a key factor, increasing its weight beyond a certain point does not necessarily improve performance. Instead, incorporating allele frequency-based conservation parameters, particularly non-seed conservation, enhances predictive accuracy. This effect becomes more pronounced at the tested thresholds, where AF considerations play an increasingly significant role in distinguishing conserved miRNAs (Figure 2.5.2a).

The distribution of conservation scores among mature miRNAs indicates that a subset of miRNAs exhibits particularly low conservation, distinguishing them from the majority. The average conservation score is 0.82, with a standard deviation of 0.07 (Figure 2.5.2b). To further validate the OCS, we compared conservation scores with all curated annotations from MirGeneDB, which defines sets of accepted and rejected human miRNAs [8,372]. The cumulative distribution analysis reveals a clear trend where miRNAs annotated as high-confidence are predominantly found among the most highly conserved miRNAs according to OCS. Conversely, miRNAs with the lowest conservation scores are more frequently associated with low-confidence annotations in mirGeneDB (Figure 2.5.2c). This pattern supports the agreement between conservation scores and annotation confidence, and therefore reinforces the OCS as an effective score to distinguish functionally relevant miRNAs. In addition, we examined whether paralogous copies of miRNAs influence

conservation levels. Among 2,496 mature miRNAs with variant data, 116 are encoded by multiple genomic loci while 2,380 occur as single copies. Multi-copy miRNAs show significantly higher conservation than single-copy miRNAs (Wilcoxon test, $p < 0.05$). A comprehensive list of conservation scores for each mature microRNA, along with their corresponding parameters, is provided in Supplementary Table 1.

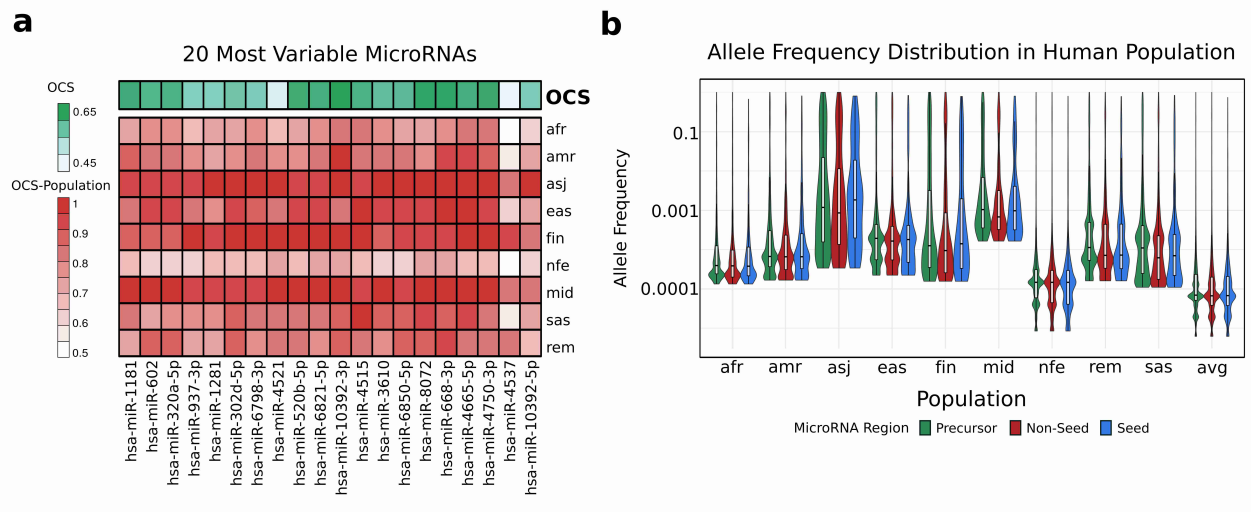


Figure 2.5.3: MiRNA Variants in Human Populations a) Heatmap of population-specific OCS values, displayed for the 20 most variable miRNAs. The populations considered include African/African American (afr), Admixed American (amr), Ashkenazi Jewish (asj), East Asian (eas), Finnish (fin), Non-Finnish European (nfe), Middle Eastern (mid), South Asian (sas), and Remaining (rem). b) Allele distribution across human populations shown for precursor miRNA regions (excluding mature miRNAs), non-seed regions, and seed regions of mature miRNAs. The joint AF is presented as an average (avg).

2.5.5.3 Population-specific miRNA variability

The OCS can be computed using subsets of variants corresponding to populations (as annotated in gnomAD). Population-specific analysis of mature miRNA conservation revealed variability across different groups. For example, hsa-miR-4537, despite having low overall conservation, remains relatively preserved in Finnish, Middle Eastern, and Ashkenazi Jewish populations. Similarly, hsa-miR-10392-5p is notably conserved only in Ashkenazi Jews, suggesting a potential role in population-specific adaptation or genetic drift (Figure 2.5.3a). The population-specific analysis indicated that the conservation trends described above are consistent among populations. For example, we observed that the distribution of allele frequencies across all miRNA regions is significantly different between precursor and mature miRNA regions across all populations, with precursors harboring more alleles at higher frequencies (Mann-Whitney U, $p < 0.05$). No significant differences were observed between seed and non-seed regions, except for the East Asian population (Mann-Whitney U, $p < 0.05$; Figure 2.5.3b).

2.5.5.4 Functional impact of miRNA conservation on gene regulation

Since our score of miRNA conservation was optimized using expression, it is important to examine how expression levels correspond to this measure to see if we induced any undesired bias. Mean expression levels across diverse biological conditions and tissues from isomiRdb were analyzed in relation to miRNA OCS values. A pattern emerged in which, as expected, lower conservation was

consistently linked to reduced expression, but higher conservation did not necessarily imply high expression (Figure 2.5.4a). This indicates that the score can still highlight biologically relevant miRNAs that may otherwise go unnoticed in conventional profiling.

Measuring miRNA expression is inherently linked to its precursor form, where both the 5p and 3p arms are transcribed together and subsequently measured as a single unit. While expression levels of both arms may be correlated, differences in biological activity and functional stability could lead to distinct conservation patterns. To investigate this, we analyzed 784 precursor miRNAs with annotated 5p and 3p arms, comparing their respective OCS values. We found that in 380 cases, the 3p arm is more conserved, while in 404 cases, the 5p arm shows higher conservation. The distribution of conservation differences between the two arms is balanced (Figure 2.5.4b). Permutation analysis of OCS values reveals significantly lower conservation differences between miRNA arms than expected by chance ($p < 0.01$), suggesting a strong tendency for 3p and 5p arms to maintain similar conservation levels. This indicates potential functional or evolutionary constraints preserving balanced conservation between both strands. To assess whether conservation differences between miRNA arms influence their functional relevance, miRNAs were binned based on their OCS difference, and the number of target interactions (as defined in the microT database [157]) for each arm was analyzed. In the group with the smallest conservation differences, the trend remains subtle, yet the more conserved arm tends to have a higher number of target interactions. However, in the group with the largest OCS differences, this pattern becomes more pronounced, with 124 cases compared to 74 cases where the more conserved arm also exhibits a greater number of target genes, suggesting a potential link between conservation and regulatory significance (Figure 2.5.4c). To test if this association exceeds random expectation, a permutation analysis confirmed a significant difference from chance ($p < 0.01$), supporting a link between conservation and miRNA targeting preferences.

Alternative polyadenylation (APA), leading to different length of the 3' UTRs, influences miRNA binding site availability [115]. We tested whether genes lacking APA sites, which maintain stable miRNA binding, are more frequently targeted by highly conserved miRNAs compared to the least conserved ones, integrating OCS scores to assess this relationship. We observe that among 2,784 genes lacking APA sites, the HC-miRNAs consistently target more genes than the LC-miRNAs across multiple thresholds of miRNA binding site scores. For instance, without restriction on binding site scores, this difference is 2,449 vs. 2,061 genes, while considering only binding sites with an mre score >= 0.1, the difference becomes 70 vs. 28 genes. Similarly, among 15,955 genes with multiple APA sites, this trend persists. Without restriction on binding site scores, the HC-miRNAs target 10,784 genes compared to 10,009 for the LC-miRNAs. When considering only binding sites with an MRE score >= 0.1, this difference increases to 760 vs. 166 genes (Figure 2.5.4d).

Further, for each miRNA, we investigated how frequently its miRNA binding sites are located upstream of a representative APA site (on genes with multiple APA sites). We found that this frequency is consistently higher for the most conserved miRNAs (0.48) compared to the least conserved miRNAs (0.39). This difference becomes

more pronounced at higher binding site stringency thresholds (Figure 2.5.4e). To validate these results independently, we repeated the analysis using conserved binding sites from TargetScanHuman [16] and confirmed that highly conserved miRNAs preferentially target genes lacking APA sites (911 of 7,339 targets, 12.4%) compared to the least conserved miRNAs (50 of 622 targets, 8.0%; Fisher's exact test $p < 0.01$). This independent dataset thus supports our conclusion that conserved miRNAs are preferentially associated with stable miRNA regulation.

As an additional measure of OCS validity, we assessed the functional implications of miRNAs and their target genes by annotating and contrasting the Gene Ontology (GO) term enrichment and pathway associations of genes regulated by the HC-miRNAs and LC-miRNAs. This analysis evaluates whether higher conservation corresponds to greater regulatory involvement in essential biological processes and key molecular pathways. We found 2,341 significant terms for HC-miRNAs compared to 257 significant terms for LC-miRNAs. In particular, no significant GO terms were identified for LC-miRNAs, whereas HC-miRNAs were significantly associated with 950 GO biological process terms. Additionally, we observed that significant terms for LC-miRNAs are particularly associated with under-representation in KEGG-pathways and GO terms derived from miRTarBase. The large discrepancy in enriched terms annotated between HC-miRNAs and LC-miRNAs supports the notion that conservation is strongly linked to functional relevance, with HC-miRNAs engaging in more biologically significant and well-annotated regulatory roles (Figure 2.5.4f).

2.5.5.5 Evolutionary and clinical insights into compensatory miRNA-target variant pairs

To investigate potential co-evolution patterns between miRNAs and their targets, we analyzed variants occurring in both the miRNA seed region and the corresponding binding site in the 3' UTR of the target gene, while preserving complementary base-pairing.

Interestingly, we identified 66,240 compensatory variant pairs in which sequence variation in both the miRNA and its target maintained base-pairing. Among these, 14,869 pairs exhibited their highest AF in the same population and were classified as population-specific, comprising 14,464 rare compensatory variant pairs and 405 cases in which one or both sites contained a SNP (Figure 2.5.4h; see Methods for details). These compensatory pairs were located in 5,792 6mer, 2,576 7mer-A1, 4,519 7mer-m8, and 1,982 8mer miRNA binding sites. It is noteworthy that compensatory variants can only arise in the context of an initially deleterious mutation that disrupts efficient binding of a miRNA to its target, whether through a change in the miRNA itself or in its binding site. This means that deleterious mutations can sometimes survive in the population and that there is a strong selection pressure to reestablish the regulatory interaction by a compensatory mutation.

The strongest signals of population-specific compensatory variant pairs were observed in the Non-Finnish European population (6,525 rare variants; 203 with one SNP) and the African population (5,702 rare variants; 159 with one SNP; 4 with both SNPs) (Figure 2.5.4g, h). We interpret these rare variants as those that arose recently, in contrast to the frequent ones that have arisen a long time ago, such that

they are present in a large proportion of the population. Moreover, the African population harbors a significantly greater proportion of compensatory to disruptive binding pairs compared to all other populations, suggesting stronger evolutionary constraints or selective pressures shaping miRNA-target interactions (Fisher's Exact Test, $p < 0.01$). Significance is retained when the analysis is restricted to binding sites stronger than 6mers (Fisher's Exact Test, $p < 0.01$). This pattern may also reflect the higher genetic diversity found in African populations [384].

Co-evolution appears to be more frequent in HC-miRNAs than in LC-miRNAs, as HC-miRNAs exhibit a significantly higher ratio of compensatory to disruptive binding pairs at binding positions (Fisher's Exact Test, $p < 0.01$; consistent for sites stronger than 6mers). This suggests that evolutionary constraints are stronger in functionally essential miRNAs, where compensatory variants may act to preserve regulatory interactions despite sequence variation.

To evaluate the clinical significance of these population-specific compensatory variant pairs, we cross-referenced them with ClinVar and identified 108 matches. Of these, one variant is classified as pathogenic: MMACHC, co-occurring with a compensatory variant in hsa-miR-6516-5p. This compensatory variant pair is associated with methylmalonic aciduria and homocystinuria (7mer-m8 binding site), in the Non-Finnish European population. Moreover, 8 variants have conflicting pathogenicity classifications, and 109 are of uncertain significance (Figure 2.5.4i). While these variants are identified as pathogenic or likely pathogenic, we propose that their effects are through disruption of a miRNA target site and that they can be compensated by the corresponding variation in the interacting miRNA.

To better understand the impact of variants in targeting-dynamics, we also analyzed the remaining 51,371 compensatory variants that did not meet the criterion of population-specificity. Removing this restriction accounts for unequal population representation in the data and natural AF fluctuations that could obscure evolutionary patterns. Additionally, it helps identify functional compensatory mechanisms that persist across multiple populations rather than being confined to a single group. We identified 359 compensatory variant pairs with at least one variant annotated in ClinVar, including 354 in target genes and 5 in miRNAs. Notably, one variant in the COMT target gene, co-occurring with hsa-miR-3907, is classified as a drug response variant associated with Tramadol response. In addition, we observed 8 cases with conflicting pathogenicity classifications and 202 cases of uncertain significance (Supplementary Table 3).

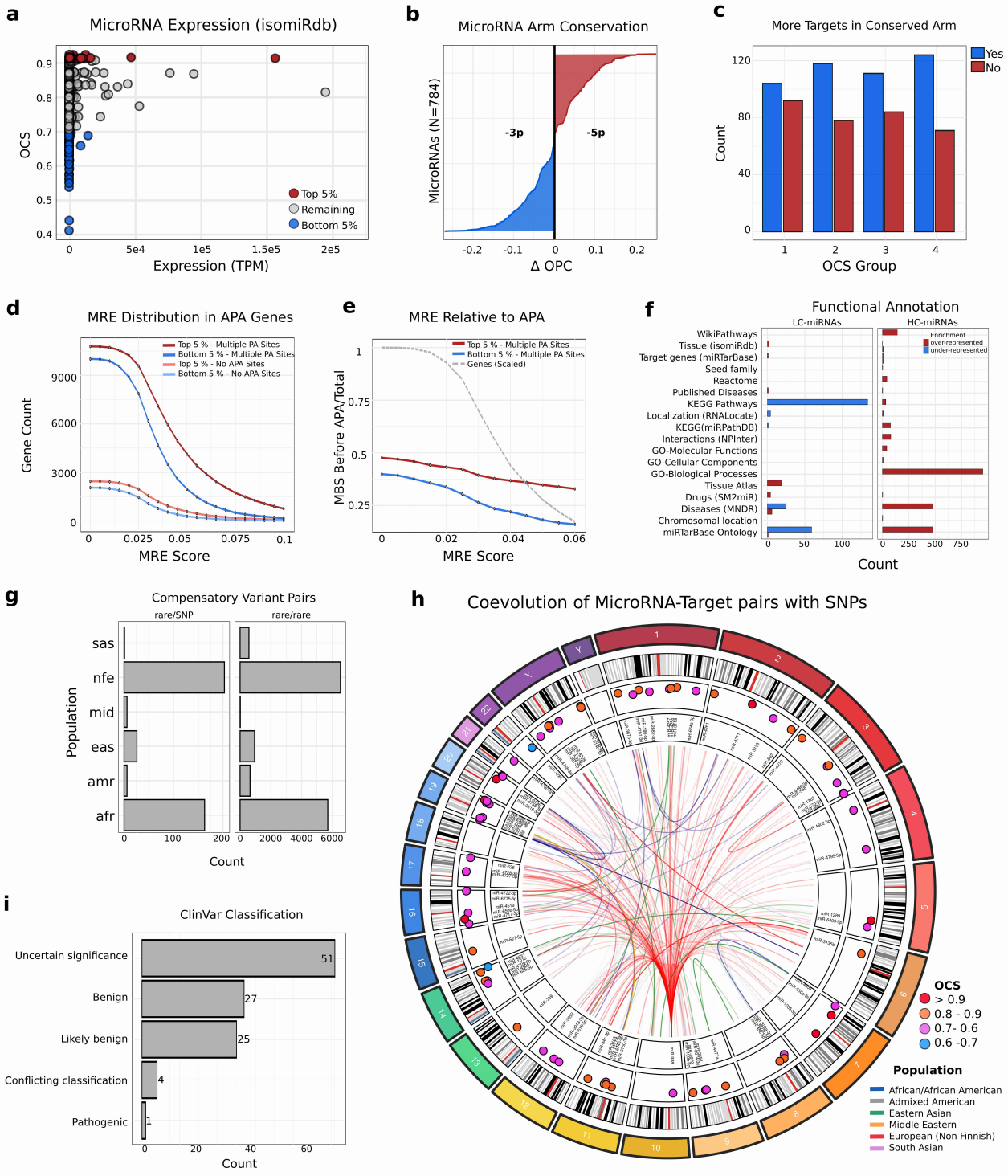


Figure 2.5.4: MiRNA Conservation: Expression, Targeting, and APA Dynamics a) Normalized mean expression levels of miRNAs from isomiRdb and OCS values. b) Sorted differences in OCS between the -3p and -5p arms of mature miRNAs. Blue represents cases where the -3p arm is more conserved, while red indicates higher conservation for the -5p arm derived from the same precursor miRNA. c) miRNAs with two arms were grouped into four bins based on their OCS differences, and the number of cases where the more conserved arm had more target genes was counted. d) Number of genes lacking APA regulation for top 5% and bottom 5% miRNAs, as well as for genes with APA, across different thresholds of miRNA binding site scores (MRE).

Figure 2.5.4 (continued): e) Ratio of binding sites before a representative APA site per gene for top 5% conserved and bottom 5% conserved miRNAs. The number of genes considered for each threshold of miRNA binding is scaled and shown in grey. f) Overview of categories with significantly enriched terms for of HC-miRNAs and LC-miRNAs identified using miEAA (see Methods for details). g) Counts of co-occurring compensatory variants in miRNA binding sites, with the highest frequency observed in the same population. h) Circos plot illustrating co-evolved miRNA–target gene variants that are compensatory both in the miRNA seed and at its binding site within the human genome. Dots represent the OCS value of the corresponding miRNA. Connections are drawn when both variants share the highest AF for the same population, with increased opacity representing higher AF. At least one of the two variants must be a SNP (0.01<AF<=0.05). i) ClinVar classifications of population-specific compensatory variants.

2.5.6 Discussion

This study investigates the population-specific conservation patterns of miRNAs, emphasizing their variability, functional roles, and evolutionary significance across diverse human populations. By leveraging the expanded gnomAD v4 dataset [374], which provides substantially greater coverage than gnomAD v3 (a 5-fold increase in WGS and a 6-fold increase in WES datasets), we identified 55,796 variants in precursor and mature miRNAs. This represents an approximately 2.3-fold increase in variant data within genomic miRNA coordinates as compared to prior characterizations [54], highlighting the enhanced resolution and expanded analytical scope provided by the latest genomic resources.

The computation of OCS emphasizes positional conservation as the most impactful determinant of miRNA functionality. While the seed region's role in determining binding specificity and regulatory activity is well established [290,385], the non-seed region also plays a critical role by stabilizing 3′-end interactions [16,386], enabling the recognition of non-canonical target sites, the creation of functionally relevant isomiRs through altered enzymatic cleavage patterns that shift seed location [345] and allowing RNA editing in non-seed regions, which impacts loading efficiency in the RISC complex [348]. This dual importance supports the rationale for scoring variant-free positions equally in both seed and non-seed regions to comprehensively capture the full spectrum of miRNA functionality. The optimization of the computation of the OCS based on expression-conservation overlap implies read coverage as a key metric for miRNA annotation, linking it to greater annotation confidence [7,387]. This approach is further validated by demonstrating a link between higher expression levels and functional significance in our analyses, reinforcing the role of miRNA expression across multiple biological conditions and tissues in inferring function [380,388]. Additionally, we validated our findings through comparison with the accepted and rejected miRNAs from mirGeneDB [8,372].

To further validate OCS computation, comparisons between HC-miRNAs and LC-miRNAs reveal consistent alignment with biological expectations. LC-miRNAs, while detectable, present lower expression levels compared to HC-miRNAs, suggesting their functional roles may be more context-dependent or specialized. Further, these miRNAs with low expression levels have been linked to less connected target networks [350]. In contrast, HC-miRNAs show balanced conservation between -5p and -3p arms, and in cases of imbalance, the less conserved arm consistently targets fewer genes, reinforcing the link between conservation and regulatory relevance. HC-miRNAs also preferentially target sites upstream of representative

APA sites. Cell-type-specific 3′UTR shortening correlates with the loss of binding sites for functional microRNAs, implying selective evasion of miRNA regulation [326]. Their higher frequency of enriched pathways and Gene Ontology terms further emphasizes their functional significance, whereas LC-miRNAs show sparse associations. Moreover, we find that multi-copy miRNAs are more conserved than single-copy counterparts, consistent with the expectation that redundancy across loci enhances robustness. This finding is counterintuitive since one could have expected that multiple copies of the same miRNA could be more robust to mutation than single copy ones. Patterns of genetic variation also support these findings. The reported density of 1 SNV per 5.55 bp within precursor miRNAs is lower than the genome-wide average of 1 SNV per 4.9 bp [374]. Although no significant differences in SNV density were observed between seed and non-seed regions across all mature miRNAs, a higher density was observed in non-seed regions compared to seed regions when contrasting HC-miRNAs with LC-miRNAs, aligning with the idea that non-seed regions in LC-miRNAs are under reduced evolutionary pressure, allowing for greater variability.

Additionally, precursor miRNAs exhibit a significantly higher density of common variants across all populations, again reflecting their greater tolerance for variation likely due to reduced functional constraints compared to mature miRNA regions [51]. Regional mapping showed slightly higher variant densities in terminal loops compared to lower stems, consistent with the lower stem's essential role in Drosha processing [389]. While overall Ti/Tv ratios of miRNAs align with previous findings [390], we observed higher Ti/Tv ratios in seed regions for rare alleles, preserving binding stability, while higher-frequency alleles favor non-seed regions, reflecting relaxed selective pressure. Given the absence of SNPs in HC-miRNAs, this pattern underscores the critical role of seed regions in maintaining stable miRNA-mRNA interactions, whereas non-seed regions in LC-miRNAs tolerate greater variability due to reduced constraints. The higher transition rate observed for pyrimidines in seed regions aligns with expectations from RNA structural studies, which show that pyrimidine–pyrimidine mismatches are structurally more compatible with the A-form helices adopted by miRNA precursors and miRNA–mRNA duplexes than bulkier purine–purine mismatches [390,391].

Collectively, these findings validate the biological relevance of OCS, demonstrating its capacity to distinguish miRNAs with critical regulatory roles from those with limited impact. Moreover, OCS effectively identifies confidently annotated miRNAs and distinguish from the ones that may have lost functionality through evolutionary divergence.

Our analyses identified population-specific differences in conservation of miRNAs, such as in the case of hsa-miR-10392-5p, which is more conserved in Ashkenazi Jews than in other populations. These findings suggest that certain miRNAs may have evolved lineage-specific regulatory roles or adaptations influenced by population-specific selective pressures or genetic drift. Additionally, we observed that certain miRNA variants often co-occur with their corresponding binding site variants in a population-specific manner. The fact that we found many compensatory mutations highlights a very interesting phenomenon of surviving deleterious mutations in a population until they are rescued by a second mutation that restores the base-

pairing. Our results show that these mutations occur rather frequently in the African population, which harbors the highest genetic diversity among all populations studied [392]. Among these co-occurring population-specific variants, three cases were identified as likely pathogenic, highlighting their potential clinical relevance. Specifically, variants in target genes MMACHC and COMT that were identified as pathogenic or responding to drugs, are accompanied by compensatory miRNA variants in the respective binding position. These findings highlight the interplay between population-specific genetic diversity and miRNA-mediated regulatory mechanisms, underscoring the potential clinical relevance of such co-evolutionary patterns. With the expanding knowledge of the links between genomic variants and disease, our annotation of potentially co-evolved compensatory variant pairs of miRNAs and target genes offers a valuable resource for investigating these relationships.

We are aware that our study has some limitations: These include the bias introduced by the over-representation of European samples within the gnomAD dataset [374], the discrepancies in AF annotation resulting from differences between WES and WGS data [393], and the lack of sample-specific annotations for co-occurring variants. This last limitation hinders our ability to fully conclude the biological function of co-occurring variants in miRNA seed regions and their corresponding target sites.

In conclusion, our study highlights miRNA conservation patterns across human populations using a novel scoring system that integrates genomic variability, further validated with correlations with targeting dynamics and functional enrichment. By distinguishing highly conserved from less conserved miRNAs, we reveal their regulatory significance, population-specific adaptations, and compensatory co-evolutionary patterns.

2.6 Applied motif analysis in collaborations

The author of this thesis applied the transcription factor–related regulatory knowledge and TRANSFAC database [276] expertise gained in Chapter 2.2 to collaborative projects. In the first study, the author contributed computational motif analysis to investigate transcription factor binding logic in immune cell differentiation. This work was published as

Gabele, A., Sprang, M., **Cihan, M**., … Andrade-Navarro, M. A., Tenzer, S., Luck, K., Bopp, T., Distler, U. (2025). Unveiling IRF4-steered regulation of context-dependent effector programs in $CD4^+$ T cells under Th17- and Treg-skewing conditions. Cell Reports, 44(3) [394]

and focused on how IRF4 regulates gene expression in $CD4^+$ T cells under Th17- and iTreg-skewing conditions. The author's contribution centered on the analysis of IRF4 binding sites, including motif enrichment, motif co-occurrence, and the identification of composite motifs, defined as IRF4 motifs located in close proximity to motifs of potential co-binding transcription factors. In addition, the author contributed to the cross-validation of selected composite motifs through comparison with public ChIP-seq datasets.

By integrating motif analysis with Bio-ChIP-seq profiles, proteomics, and interactome datasets, this work helped to uncover substantial differences in IRF4's DNA-binding architecture between Th17 and iTreg cells. Th17 cells showed more IRF4 binding sites overall and a higher proportion of peaks containing composite motifs, indicating a richer and more diverse cofactor environment. Several of these composite elements were cell-type specific, reflecting distinct IRF4 partner transcription factors in each differentiation state. Notably, the analysis highlighted IRF4–FLI1 composite motifs, which were enriched near key Th17-associated genes and supported by interaction data and functional experiments demonstrating the requirement of FLI1 for Th17 differentiation.

The second publication,

Gabele, A., **Cihan, M**., … Andrade-Navarro, M. A., Bopp, T., Distler, U. (2025). Protocol for mapping murine transcription factor interactomes and composite motifs combining affinity purification mass spectrometry and ChIP-seq. STAR Protocols, 6(4), 104184 [395],

presents a generalized and standardized version of the analytical workflow used in the first study. This protocol outlines how AP–MS interactome data, Bio-ChIP-seq peak sets, and TRANSFAC-based transcription factor annotations can be integrated into a coherent analysis pipeline. The author of this thesis contributed to structuring and formulating the sections that detail the bioinformatics analysis, including data preprocessing, peak extraction, database-supported TF annotation, and the systematic detection of composite motifs. By organizing these analytical steps into a reproducible and modular format, the protocol offers a broadly applicable framework for investigating the interactomes and DNA-binding architectures of any biotinylated transcription factor.

3 General discussion

3.1 Systems-level integration of microRNA regulation

miRNA regulation operates within complex regulatory environments shaped by transcriptional dynamics and evolutionary constraint. Rather than arising from isolated binding events, miRNA function emerges through interactions among multiple layers [5,219,351]. Throughout this thesis, integrative analyses demonstrate how miRNA regulation adapts to cellular state while remaining constrained by higher-order regulatory organization. This view reframes miRNA activity from a collection of pairwise interactions to a systems-level property of coordinated regulatory architectures.

Chapter 2.1 establishes this perspective by addressing limitations of sequence-based miRNA target prediction. By analyzing miRNA–mRNA interactions within REST-regulated gene sets, targeting patterns are evaluated in the context of transcriptional co-regulation. When regulatory context is incorporated, miRNAs preferentially associate with targets displaying features of functional regulation, including multiple binding sites, favorable local sequence composition, and enrichment toward distal regions of 3′UTRs. These properties differ significantly from background predictions, indicating that effective targeting reflects coordinated regulatory programs rather than intrinsic sequence features alone.

At a broader scale, these findings suggest that miRNAs primarily act on gene groups that are already transcriptionally coordinated [141]. Instead of introducing independent regulation, miRNAs tend to modulate existing programs by stabilizing, buffering, or fine-tuning established expression patterns [290]. This principle recurs throughout subsequent chapters, where network context consistently shapes the apparent impact of miRNA regulation.

Chapter 2.2 extends this framework to cell-type-specific regulation in glioblastoma. Here, TF–miRNA–gene feed-forward loops are examined as regulatory motifs whose function depends on transcript structure. Cell-type-specific APA alters 3′UTR length, systematically reshaping miRNA binding site availability across tumor states. Although many genes maintain similar transcriptional regulation, differences in transcript structure lead to distinct post-transcriptional connectivity. The limited overlap between differential gene expression and differential APA highlights transcript structure as a source of regulatory variation that is not captured by expression-based analyses alone. Consequently, specific tumor cell populations can attenuate miRNA-mediated repression without altering core transcriptional programs.

These observations underscore transcript structure as an additional organizational layer in regulatory systems [115,117]. In glioblastoma, APA-driven remodeling contributes to differential control of pathways linked to stemness, survival, and DNA repair, providing a mechanistic basis for regulatory heterogeneity that remains invisible at the expression level.

The role of transcript structure in miRNA regulation is examined more systematically in Chapters 2.3 and 2.4. Chapter 2.3 demonstrates that miRNA expression can be inferred from mRNA expression profiles, indicating that miRNA abundance is closely coupled to global regulatory states rather than independently regulated. Chapter 2.4 builds on this insight by showing that APA-derived changes in miRNA binding site availability are informative for predicting apparent miRNA expression differences across conditions. While transcript structure alone provides a detectable signal, integrating it with gene expression substantially improves predictive performance. Together, these results suggest that miRNA steady-state levels are influenced not only by transcription and processing, but also by the structure of the target landscape [41]. In particular, APA-mediated gain or loss of binding sites modulates target engagement and miRNA turnover, consistent with models of TDMD [34].

Chapter 2.5 adds an evolutionary dimension by examining population-scale variation in human miRNAs. Highly conserved miRNAs are associated with stable regulatory architectures, target genes with limited APA variability, and occupy central network positions. In contrast, less conserved miRNAs display greater sequence diversity and preferentially interact with structurally plastic transcripts. This uneven distribution of regulatory stability and flexibility reflects distinct evolutionary constraints across the miRNA–target network. The identification of compensatory miRNA–target variants further demonstrates that some interactions are buffered against system-level genetic perturbations.

Collectively, these chapters show that systematic miRNA targeting arises from the dynamics of reguatory layers, network structure, and evolutionary constraint.

3.2 Functional and disease implications of regulatory dynamics

The analyses presented in this thesis indicate that miRNA regulation contributes to functional diversity and disease phenotypes primarily through reconfiguration of regulatory interactions rather than uniform changes in expression. Functional consequences consistently emerge when miRNA-mediated regulation is reshaped by cellular context, highlighting context-dependent regulatory reorganization as a central determinant of miRNA function in disease [118,212].

In Chapter 2.2, single-cell analyses of glioblastoma illustrate how transcriptomic plasticity reshapes miRNA regulation in a cell-state-dependent manner. APA accompanies transitions from neoplastic radial glia–like cells toward oligodendrocyte precursor–like states, with shared APA profiles marking transitional populations along pseudotime trajectories. These changes are spatially associated with cell types of tumor border niches, linking transcriptomic remodeling to the interplay of neoplastic cells with the tumor microenvironment.

Functional annotation reveals that APA-mediated loss of miRNA regulation affects distinct biological modules across tumor states. In neoplastic populations, affected pathways are enriched for stemness, proliferation, morphogenesis, and impaired apoptotic control, whereas OPC-like cells preferentially evade miRNA regulation of DNA damage response and glioblastoma-associated signaling. Rather than globally

escaping post-transcriptional control, different tumor states selectively modulate repression of specific functional programs. Coordinated APA and expression changes in genes such as EGFR, GRB2, and DVL3 further connect post-transcriptional remodeling to functional adaptation.

At the regulatory motif level, disease relevance is concentrated within miRNA-centered FFLs. Many significant motifs involve genes undergoing concurrent changes in expression and APA along pseudotime, indicating dynamic reconfiguration during cellular transitions. Recurrently affected miRNAs include neural and lineage-associated families as well as oncogenic hubs. Survival analyses reinforce the functional significance of this reorganization, as miRNA-centered FFLs associated with OPC-like tumor populations correlate with patient survival outcome.

Broader relevance is supported by Chapter 2.3, which shows that miRNA expression reflects coordinated pathway-level gene activity rather than regulation by individual upstream factors. Predictive gene networks are densely connected and enriched for shared biological functions, consistent with miRNAs acting as integrative regulatory nodes. Neuronal and synaptic processes dominate these signatures, indicating tight coupling between miRNA abundance and physiological signaling programs rather than cell identity alone.

Within this framework, miR-137 emerges as a central hub embedded in networks related to synaptic vesicle cycling and neuronal signaling, suggesting that its expression reflects synaptic functional state. A second regulatory axis involves cardiac and muscle-associated miRNAs linked to pathways governing ion transport, contractility, and excitation–contraction coupling. Disease enrichment analyses connect these modules to clinically relevant phenotypes, including arrhythmias and cardiomyopathies, with tissue-specific expression patterns providing additional validation [169].

Chapter 2.5 further refines these interpretations by linking evolutionary conservation to regulatory importance. Highly conserved miRNAs are more deeply embedded in regulatory networks, control larger target sets, and are associated with essential biological processes, whereas less conserved miRNAs occupy more context-dependent positions. Selective constraint extends beyond the seed region across the mature miRNA, and differences between 5p and 3p arms further reflect asymmetric regulatory responsibility.

Conservation also correlates with target transcript architecture with conserved miRNAs preferentially regulating genes with limited APA variability, whereas less conserved miRNAs acting within structurally dynamic environments. Disease relevance is further supported by evidence that miRNA–target interactions often co-evolve, with compensatory variants preserving regulatory pairing despite genetic change. Such events are enriched among conserved miRNAs and overlap with disease- and clinical variants.

Overall, these analyses demonstrate that miRNA regulation contributes to pathology by selectively reshaping regulatory motifs and functional gene modules in a tissue-

and state-dependent manner. Transcriptomic plasticity, particularly through APA, enables disease states such as cancer to attenuate post-transcriptional control of specific pathways while preserving core transcriptional programs.

3.3 Computational advances in microRNA analysis

A central objective of this thesis is to address computational limitations that constrain the analysis of miRNA regulation in complex transcriptomic datasets. These challenges include incomplete or absent miRNA expression measurements, uncertainty in miRNA–target annotation, static representations of regulatory interactions, and limited methods for integrating miRNA regulation across datasets and biological scales [155,298]

Chapter 2.1 mitigates the high false-positive rate of sequence-based target prediction by embedding miRNA–mRNA interactions within transcriptionally defined regulatory programs. By reframing target prediction as a context-aware network problem, this approach identifies functionally coherent miRNA–target associations across diverse regulatory settings. Importantly, the framework is extensible and can be adapted to incorporate additional regulatory layers, such as epigenetic regulation or RBPs that intersect with miRNA-mediated control.

Chapter 2.2 addresses the absence of miRNA measurements in single-cell transcriptomics by inferring miRNA regulatory activity from APA-derived binding site landscapes and regulatory network structure. Emphasizing changes in target accessibility rather than miRNA abundance enables the identification of biologically meaningful regulators, supported by concordance with prior literature and associations with patient survival [117,125,129].

In bulk transcriptomic datasets, missing or incomplete miRNA profiling similarly limits large-scale and comparative analyses. Chapter 2.3 introduces a machine-learning framework that predicts miRNA expression directly from mRNA expression profiles by modeling multivariate transcriptional structure. Regularization enhances robustness and generalizability across cancer types, enabling scalable reconstruction of miRNA expression with improved performance relative to existing approaches [198].

Building on this framework, Chapter 2.4 presents MIRNAPEX, which predicts differential miRNA expression by integrating gene expression with transcript-structure-derived features directly from raw sequencing data. By explicitly modeling APA-driven changes in miRNA binding site availability, MIRNAPEX avoids the error propagation inherent in two-step prediction strategies and provides a principled approach for identifying miRNAs potentially regulated through TDMD.

Finally, Chapter 2.5 introduces a population-based miRNA conservation score that captures selective constraints within human populations, complementing traditional phylogenetic conservation metrics. The identification of compensatory miRNA–target variants further enables assessment of regulatory robustness and co-evolution at the system level.

Together, these computational contributions support systems-level analysis of miRNA regulation under data availability limitations. Methods that are presented enable context-aware target annotation, inference of miRNA regulation in the absence of direct measurements, modeling of transcriptomic plasticity, and robust prioritization of functionally relevant miRNAs and regulatory interactions.

4 Conclusion and outlook

miRNA-mediated gene regulation poses substantial conceptual and computational challenges, largely due to its strong dependence on transcriptomic context, regulatory integration, and limitations of current large-scale datasets [159,339,349]. By integrating network-based modeling, analyses of transcriptomic plasticity, machine learning approaches, and population-scale genomic data, this thesis demonstrates that miRNA function cannot be understood as an isolated property of individual interactions but instead emerges from the dynamic interplay between miRNAs, their targets, and the structural organization of the transcriptome.

Across the presented studies, a recurring theme is that apparent miRNA activity often reflects changes in target availability rather than changes in miRNA abundance itself. APA was shown to be a major driver of this effect, reshaping miRNA binding landscapes in a cell-type-specific manner. This finding challenges common interpretations of differential miRNA expression and highlights the need to explicitly account for transcript isoform usage when inferring regulatory activity. In parallel, integration of transcription factor networks revealed that functional miRNA targeting is constrained by regulatory context, allowing a shift from purely sequence-driven target prediction toward biologically informed models with improved specificity.

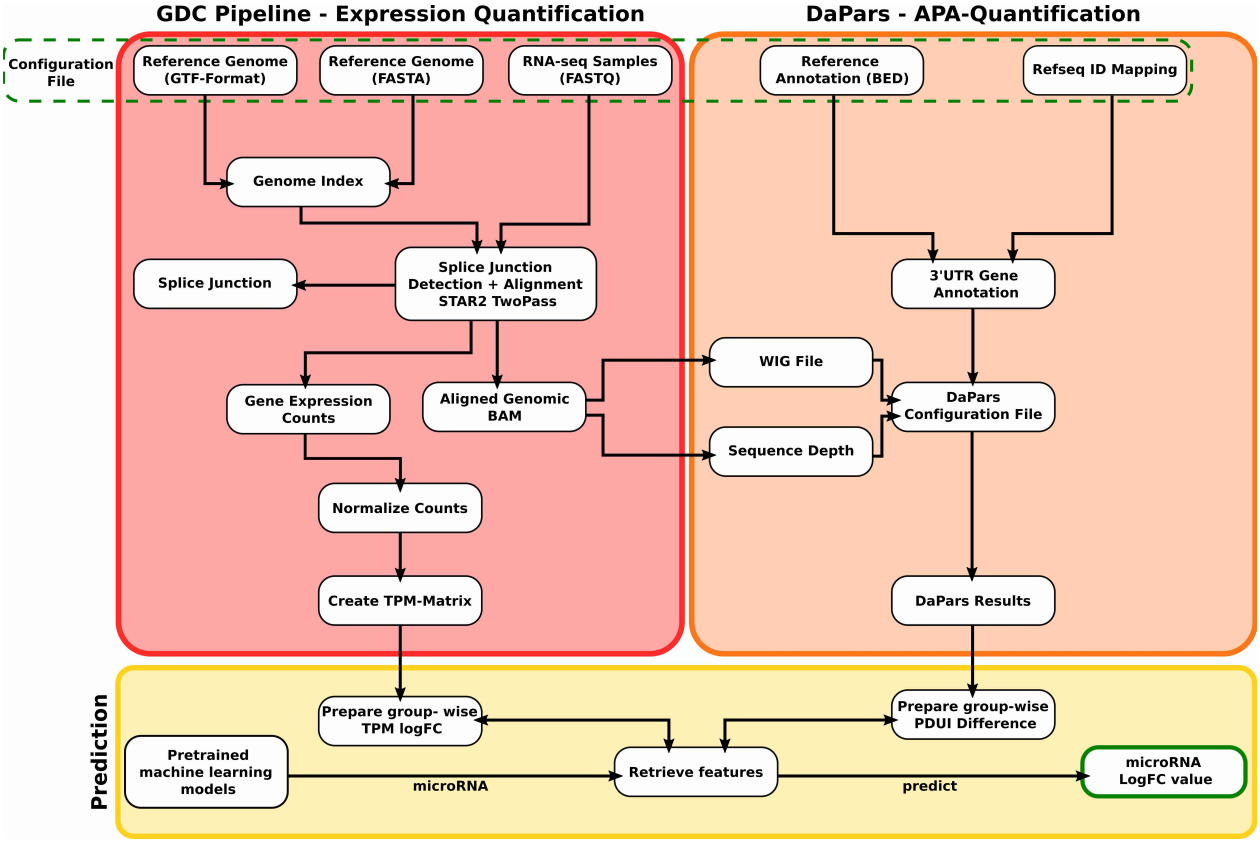
While miRNA biogenesis and processing are now well characterized and structurally resolved at high precision, the downstream rules governing target recognition, repression strength, and regulatory outcome are highly context-dependent [5]. Non-canonical targeting mechanisms, including seed-independent interactions, miRNA-mediated activation of translation, nuclear functions, and TDMD, further expand the regulatory repertoire beyond classical repression model [5,29,34]. Systematic investigation of these non-canonical modes remains limited, yet they are likely to play key roles in shaping miRNA turnover, target selectivity, and regulatory feedback loops. In this context, improved resolution of TDMD mechanisms holds particular promise for identifying robust biomarkers, as TDMD links miRNA stability directly to specific target transcripts and cellular states [34,39,366].

Looking forward, a major bottleneck in the field remains the limited availability of large-scale miRNA data at single-cell resolution. The integration of high-quality single-cell miRNA measurements with matched mRNA and isoform-level data will be essential to disentangle cell-type-specific regulatory programs and to accurately map miRNA activity in complex tissues such as tumors [396]. As data generation shifts from quantity toward higher quality and resolution, future miRNA target prediction efforts will increasingly benefit from fewer but more informative interactions that are supported by experimental data.

Artificial intelligence and machine learning approaches will play a central role in this transition [165,397]. Rather than merely expanding prediction catalogs, future models are expected to integrate multi-layered information, including APA profiles, miRNA decay dynamics, network topology, and single-cell expression patterns. Interpretable models that prioritize mechanistic insight over raw predictive performance will be particularly valuable for identifying functionally relevant interactions.

5 Acknowledgements

6 Supplementary Information



Supplementary Figure 6.1: Schematic of the MIRNAPEX pipeline, showing how RNA-seq reads are aligned and quantified, 3′UTR usage is estimated with DaPars, and expression and APA features are combined with reference annotations to feed the pre-trained models that predict miRNA log-fold changes.

7 Bibliography

[1] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004;116:281–97. https://doi.org/10.1016/s0092-8674(04)00045-5.

[2] Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993;75:843–54. https://doi.org/10.1016/0092-8674(93)90529-y.

[3] Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 1993;75:855–62. https://doi.org/10.1016/0092-8674(93)90530-4.

[4] Press release. NobelPrize.org. Nobel Prize Outreach 2025. Mon. 8 Dec 2025. <https://www.nobelprize.org/prizes/medicine/2024/press-release/> n.d.

[5] Kim H, Lee Y-Y, Kim VN. The biogenesis and regulation of animal microRNAs. Nat Rev Mol Cell Biol 2025;26:276–96. https://doi.org/10.1038/s41580-024-00805-0.

[6] Lee Y, Kim M, Han J, Yeom K, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J 2004;23:4051–60. https://doi.org/10.1038/sj.emboj.7600385.

[7] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Research 2019;47:D155–62. https://doi.org/10.1093/nar/gky1141.

[8] Clarke AW, Høye E, Hembrom AA, Paynter VM, Vinther J, Wyrożemski Ł, et al. MirGeneDB 3.0: improved taxonomic sampling, uniform nomenclature of novel conserved microRNA families and updated covariance models. Nucleic Acids Res 2025;53:D116–28. https://doi.org/10.1093/nar/gkae1094.

[9] Starega-Roslan J, Koscianska E, Kozlowski P, Krzyzosiak WJ. The role of the precursor structure in the biogenesis of microRNA. Cell Mol Life Sci 2011;68:2859–71. https://doi.org/10.1007/s00018-011-0726-2.

[10] Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ. Processing of primary microRNAs by the Microprocessor complex. Nature 2004;432:231–5. https://doi.org/10.1038/nature03049.

[11] O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne) 2018;9:402. https://doi.org/10.3389/fendo.2018.00402.

[12] Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev 2003;17:3011–6. https://doi.org/10.1101/gad.1158803.

[13] Wilson RC, Tambe A, Kidwell MA, Noland CL, Schneider CP, Doudna JA. Dicer–TRBP complex formation ensures accurate mammalian microRNA biogenesis. Mol Cell 2015;57:397–407. https://doi.org/10.1016/j.molcel.2014.11.030.

[14] Hansen TB, Venø MT, Jensen TI, Schaefer A, Damgaard CK, Kjems J. Argonaute-associated short introns are a novel class of gene regulators. Nat Commun 2016;7:11538. https://doi.org/10.1038/ncomms11538.

[15] Kakumani PK, Ko Y, Ramakrishna S, Christopher G, Dodgson M, Shrinet J, et al. CSDE1 promotes miR-451 biogenesis. Nucleic Acids Res 2023;51:9385–96. https://doi.org/10.1093/nar/gkad619.

[16] McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science 2019;366:eaav1741. https://doi.org/10.1126/science.aav1741.

[17] Klum SM, Chandradoss SD, Schirle NT, Joo C, MacRae IJ. Helix-7 in Argonaute2 shapes the microRNA seed region for rapid target recognition. EMBO J 2018;37:75–88. https://doi.org/10.15252/embj.201796474.
[18] Elkayam E, Kuhn C-D, Tocilj A, Haase AD, Greene EM, Hannon GJ, et al. The Structure of Human Argonaute-2 in Complex with miR-20a. Cell 2012;150:100–10. https://doi.org/10.1016/j.cell.2012.05.017.
[19] Schirle NT, MacRae IJ. The Crystal Structure of Human Argonaute2. Science 2012;336:1037–40. https://doi.org/10.1126/science.1221551.
[20] Wilczynska A, Bushell M. The complexity of miRNA-mediated repression. Cell Death Differ 2015;22:22–33. https://doi.org/10.1038/cdd.2014.112.
[21] Naeli P, Winter T, Hackett AP, Alboushi L, Jafarnejad SM. The intricate balance between microRNA-induced mRNA decay and translational repression. The FEBS Journal 2023;290:2508–24. https://doi.org/10.1111/febs.16422.
[22] Briskin D, Wang PY, Bartel DP. The biochemical basis for the cooperative action of microRNAs. Proceedings of the National Academy of Sciences 2020;117:17764–74. https://doi.org/10.1073/pnas.1920404117.
[23] Fabian MR, Cieplak MK, Frank F, Morita M, Green J, Srikumar T, et al. miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4–NOT. Nat Struct Mol Biol 2011;18:1211–7. https://doi.org/10.1038/nsmb.2149.
[24] Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet 2015;16:421–33. https://doi.org/10.1038/nrg3965.
[25] Meijer HA, Kong YW, Lu WT, Wilczynska A, Spriggs RV, Robinson SW, et al. Translational repression and eIF4A2 activity are critical for microRNA-mediated gene regulation. Science 2013;340:82–5. https://doi.org/10.1126/science.1231197.
[26] Iwakawa H, Tomari Y. The Functions of MicroRNAs: mRNA Decay and Translational Repression. Trends in Cell Biology 2015;25:651–65. https://doi.org/10.1016/j.tcb.2015.07.011.
[27] Fukaya T, Iwakawa H-O, Tomari Y. MicroRNAs block assembly of eIF4F translation initiation complex in Drosophila. Mol Cell 2014;56:67–78. https://doi.org/10.1016/j.molcel.2014.09.004.
[28] Valinezhad Orang A, Safaralizadeh R, Kazemzadeh-Bavili M. Mechanisms of miRNA-Mediated Gene Regulation from Common Downregulation to mRNA-Specific Upregulation. Int J Genomics 2014;2014:970607. https://doi.org/10.1155/2014/970607.
[29] Vasudevan S, Tong Y, Steitz JA. Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. Science 2007;318:1931–4. https://doi.org/10.1126/science.1149460.
[30] Hu X, Yin G, Zhang Y, Zhu L, Huang H, Lv K. Recent advances in the functional explorations of nuclear microRNAs. Front Immunol 2023;14:1097491. https://doi.org/10.3389/fimmu.2023.1097491.
[31] Liu H, Lei C, He Q, Pan Z, Xiao D, Tao Y. Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. Mol Cancer 2018;17:64. https://doi.org/10.1186/s12943-018-0765-5.
[32] Mori MA, Ludwig RG, Garcia-Martin R, Brandão BB, Kahn CR. Extracellular miRNAs: From Biomarkers to Mediators of Physiology and Disease. Cell Metab 2019;30:656–73. https://doi.org/10.1016/j.cmet.2019.07.011.

[33] Chatterjee S, Fasler M, Büssing I, Großhans H. Target-Mediated Protection of Endogenous MicroRNAs in *C. elegans*. Developmental Cell 2011;20:388–96. https://doi.org/10.1016/j.devcel.2011.02.008.
[34] Buhagiar AF, Kleaveland B. To kill a microRNA: emerging concepts in target-directed microRNA degradation. Nucleic Acids Res 2024;52:1558–74. https://doi.org/10.1093/nar/gkae003.
[35] Guo Y, Liu J, Elfenbein SJ, Ma Y, Zhong M, Qiu C, et al. Characterization of the mammalian miRNA turnover landscape. Nucleic Acids Res 2015;43:2326–41. https://doi.org/10.1093/nar/gkv057.
[36] Han J, Mendell JT. MicroRNA turnover: a tale of tailing, trimming, and targets. Trends in Biochemical Sciences 2023;48:26–39. https://doi.org/10.1016/j.tibs.2022.06.005.
[37] Yang A, Shao T-J, Bofill-De Ros X, Lian C, Villanueva P, Dai L, et al. AGO-bound mature miRNAs are oligouridylated by TUTs and subsequently degraded by DIS3L2. Nat Commun 2020;11:2765. https://doi.org/10.1038/s41467-020-16533-w.
[38] Sanei M, Chen X. Mechanisms of microRNA turnover. Curr Opin Plant Biol 2015;27:199–206. https://doi.org/10.1016/j.pbi.2015.07.008.
[39] Shi CY, Kingston ER, Kleaveland B, Lin DH, Stubna MW, Bartel DP. The ZSWIM8 ubiquitin ligase mediates target-directed microRNA degradation. Science 2020;370:eabc9359. https://doi.org/10.1126/science.abc9359.
[40] Han J, LaVigne CA, Jones BT, Zhang H, Gillett F, Mendell JT. A ubiquitin ligase mediates target-directed microRNA decay independently of tailing and trimming. Science 2020;370:eabc9546. https://doi.org/10.1126/science.abc9546.
[41] Li L, Sheng P, Li T, Fields CJ, Hiers NM, Wang Y, et al. Widespread microRNA degradation elements in target mRNAs can assist the encoded proteins. Genes Dev 2021;35:1595–609. https://doi.org/10.1101/gad.348874.121.
[42] Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009;19:92–105. https://doi.org/10.1101/gr.082701.108.
[43] Herranz H, Cohen SM. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes Dev 2010;24:1339–44. https://doi.org/10.1101/gad.1937010.
[44] Dai A, Lan W, Lyu Y, Zhou X, Mi X, Tang T, et al. MicroRNA-mediated network redundancy is constrained by purifying selection and contributes to expression robustness in Drosophila melanogaster. Commun Biol 2024;7:1431. https://doi.org/10.1038/s42003-024-07162-w.
[45] Wang Y, Tang X, Lu J. Convergent and divergent evolution of microRNA-mediated regulation in metazoans. Biological Reviews 2024;99:525–45. https://doi.org/10.1111/brv.13033.
[46] Meunier J, Lemoine F, Soumillon M, Liechti A, Weier M, Guschanski K, et al. Birth and expression evolution of mammalian microRNA genes. Genome Res 2013;23:34–45. https://doi.org/10.1101/gr.140269.112.
[47] França GS, Vibranovski MD, Galante PAF. Host gene constraints and genomic context impact the expression and evolution of human microRNAs. Nat Commun 2016;7:11438. https://doi.org/10.1038/ncomms11438.
[48] Liu N, Okamura K, Tyler DM, Phillips MD, Chung W-J, Lai EC. The evolution and functional diversification of animal microRNA genes. Cell Res 2008;18:985–96. https://doi.org/10.1038/cr.2008.278.

[49] Ebert MS, Sharp PA. Roles for MicroRNAs in Conferring Robustness to Biological Processes. Cell 2012;149:515–24. https://doi.org/10.1016/j.cell.2012.04.005.
[50] Simkin A, Geissler R, McIntyre ABR, Grimson A. Evolutionary dynamics of microRNA target sites across vertebrate evolution. PLoS Genet 2020;16:e1008285. https://doi.org/10.1371/journal.pgen.1008285.
[51] Cammaerts S, Strazisar M, De Rijk P, Del Favero J. Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. Front Genet 2015;6. https://doi.org/10.3389/fgene.2015.00186.
[52] Machowska M, Galka-Marciniak P, Kozlowski P. Consequences of genetic variants in miRNA genes. Computational and Structural Biotechnology Journal 2022;20:6443–57. https://doi.org/10.1016/j.csbj.2022.11.036.
[53] Gong J, Tong Y, Zhang H-M, Wang K, Hu T, Shan G, et al. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Human Mutation 2012;33:254–63. https://doi.org/10.1002/humu.21641.
[54] Oak N, Ghosh R, Huang K, Wheeler DA, Ding L, Plon SE. Framework for microRNA variant annotation and prioritization using human population and disease datasets. Hum Mutat 2019;40:73–89. https://doi.org/10.1002/humu.23668.
[55] Cao W, He J, Feng J, Wu X, Wu T, Wang D, et al. miRNASNP-v4: a comprehensive database for miRNA-related SNPs across 17 species. Nucleic Acids Res 2025;53:D1066–74. https://doi.org/10.1093/nar/gkae888.
[56] Lin X, Steinberg S, Kandasamy SK, Afzal J, Mbiyangandu B, Liao SE, et al. Common miR-590 Variant rs6971711 Present Only in African Americans Reduces miR-590 Biogenesis. PLOS ONE 2016;11:e0156065. https://doi.org/10.1371/journal.pone.0156065.
[57] Fernandez N, Cordiner RA, Young RS, Hug N, Macias S, Cáceres JF. Genetic variation and RNA structure regulate microRNA biogenesis. Nat Commun 2017;8:15114. https://doi.org/10.1038/ncomms15114.
[58] Xiong X, Kang X, Zheng Y, Yue S, Zhu S. Identification of Loop Nucleotide Polymorphisms Affecting MicroRNA Processing and Function. Mol Cells 2013;36:518–26. https://doi.org/10.1007/s10059-013-0171-1.
[59] Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, et al. SNPs in human miRNA genes affect biogenesis and function. RNA 2009;15:1640–51. https://doi.org/10.1261/rna.1560209.
[60] Haas U, Sczakiel G, Laufer SD. MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3′-UTR via altered RNA structure. RNA Biol 2012;9:924–37. https://doi.org/10.4161/rna.20497.
[61] Chhichholiya Y, Suryan AK, Suman P, Munshi A, Singh S. SNPs in miRNAs and Target Sequences: Role in Cancer and Diabetes. Front Genet 2021;12. https://doi.org/10.3389/fgene.2021.793523.
[62] Quiat D, Olson EN. MicroRNAs in cardiovascular disease: from pathogenesis to prevention and treatment. J Clin Invest 2013;123:11–8. https://doi.org/10.1172/JCI62876.
[63] Gottmann P, Ouni M, Zellner L, Jähnert M, Rittig K, Walther D, et al. Polymorphisms in miRNA binding sites involved in metabolic diseases in mice and humans. Sci Rep 2020;10:7202. https://doi.org/10.1038/s41598-020-64326-4.
[64] Wang G, van der Walt JM, Mayhew G, Li Y-J, Züchner S, Scott WK, et al. Variation in the miRNA-433 Binding Site of FGF20 Confers Risk for Parkinson

Disease by Overexpression of α-Synuclein. Am J Hum Genet 2008;82:283–9. https://doi.org/10.1016/j.ajhg.2007.09.021.
[65] Cardo LF, Coto E, Ribacoba R, Mata IF, Moris G, Menéndez M, et al. The screening of the 3′UTR sequence of LRRK2 identified an association between the rs66737902 polymorphism and Parkinson's disease. J Hum Genet 2014;59:346–8. https://doi.org/10.1038/jhg.2014.26.
[66] Tagliafierro L, Chiba-Falek O. Up-regulation of SNCA Gene Expression: Implications to Synucleinopathies. Neurogenetics 2016;17:145–57. https://doi.org/10.1007/s10048-016-0478-0.
[67] Soldà G, Robusto M, Primignani P, Castorina P, Benzoni E, Cesarani A, et al. A novel mutation within the MIR96 gene causes non-syndromic inherited hearing loss in an Italian family by altering pre-miRNA processing. Hum Mol Genet 2012;21:577–85. https://doi.org/10.1093/hmg/ddr493.
[68] Abreu Costa M, Sadan AN, Bhattacharyya N, Chai N, Zarouchlioti C, Liu S, et al. Rare variants in MIR184 are a novel genetic cause of Fuchs endothelial corneal dystrophy. Genet Med 2025;27:101562. https://doi.org/10.1016/j.gim.2025.101562.
[69] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20. https://doi.org/10.1038/ng.2764.
[70] Cheerla N, Gevaert O. MicroRNA based Pan-Cancer Diagnosis and Treatment Recommendation. BMC Bioinformatics 2017;18:32. https://doi.org/10.1186/s12859-016-1421-y.
[71] Kim T-M, Huang W, Park R, Park PJ, Johnson MD. A developmental taxonomy of glioblastoma defined and maintained by MicroRNAs. Cancer Res 2011;71:3387–99. https://doi.org/10.1158/0008-5472.CAN-10-4117.
[72] Zadran S, Remacle F, Levine RD. miRNA and mRNA cancer signatures determined by analysis of expression levels in large cohorts of patients. Proceedings of the National Academy of Sciences 2013;110:19160–5. https://doi.org/10.1073/pnas.1316991110.
[73] Shi X-H, Li X, Zhang H, He R-Z, Zhao Y, Zhou M, et al. A Five-microRNA Signature for Survival Prognosis in Pancreatic Adenocarcinoma based on TCGA Data. Sci Rep 2018;8:7638. https://doi.org/10.1038/s41598-018-22493-5.
[74] Tian B, Hou M, Zhou K, Qiu X, Du Y, Gu Y, et al. A Novel TCGA-Validated, MiRNA-Based Signature for Prediction of Breast Cancer Prognosis and Survival. Front Cell Dev Biol 2021;9:717462. https://doi.org/10.3389/fcell.2021.717462.
[75] Gasparini P, Lovat F, Fassan M, Casadei L, Cascione L, Jacob NK, et al. Protective role of miR-155 in breast cancer through RAD51 targeting impairs homologous recombination after irradiation. Proceedings of the National Academy of Sciences 2014;111:4536–41. https://doi.org/10.1073/pnas.1402604111.
[76] Lin S, Zhou J, Xiao Y, Neary B, Teng Y, Qiu P. Integrative analysis of TCGA data identifies miRNAs as drug-specific survival biomarkers. Sci Rep 2022;12:6785. https://doi.org/10.1038/s41598-022-10662-6.
[77] Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics 2016;32:2891–5. https://doi.org/10.1093/bioinformatics/btw344.
[78] Teplyuk NM, Uhlmann EJ, Gabriely G, Volfovsky N, Wang Y, Teng J, et al. Therapeutic potential of targeting microRNA-10b in established intracranial glioblastoma: first steps toward the clinic. EMBO Mol Med 2016;8:268–87. https://doi.org/10.15252/emmm.201505495.

[79] Bassot A, Dragic H, Haddad SA, Moindrot L, Odouard S, Corlazzoli F, et al. Identification of a miRNA multi-targeting therapeutic strategy in glioblastoma. Cell Death Dis 2023;14:630. https://doi.org/10.1038/s41419-023-06117-z.
[80] Liu S-H, Hsu K-W, Lai Y-L, Lin Y-F, Chen F-H, Peng P-H, et al. Systematic identification of clinically relevant miRNAs for potential miRNA-based therapy in lung adenocarcinoma. Molecular Therapy - Nucleic Acids 2021;25:1–10. https://doi.org/10.1016/j.omtn.2021.04.020.
[81] Svoronos AA, Engelman DM, Slack FJ. OncomiR or Tumor Suppressor? The Duplicity of MicroRNAs in Cancer. Cancer Res 2016;76:3666–70. https://doi.org/10.1158/0008-5472.CAN-16-0359.
[82] Dhawan A, Scott JG, Harris AL, Buffa FM. Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. Nat Commun 2018;9:5228. https://doi.org/10.1038/s41467-018-07657-1.
[83] Urbanek-Trzeciak MO, Galka-Marciniak P, Nawrocka PM, Kowal E, Szwec S, Giefing M, et al. Pan-cancer analysis of somatic mutations in miRNA genes. eBioMedicine 2020;61. https://doi.org/10.1016/j.ebiom.2020.103051.
[84] Hamilton MP, Rajapakshe K, Hartig SM, Reva B, McLellan MD, Kandoth C, et al. Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. Nat Commun 2013;4:2730. https://doi.org/10.1038/ncomms3730.
[85] Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, et al. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat Struct Mol Biol 2006;13:13–21. https://doi.org/10.1038/nsmb1041.
[86] Paul D, Sinha AN, Ray A, Lal M, Nayak S, Sharma A, et al. A-to-I editing in human miRNAs is enriched in seed sequence, influenced by sequence contexts and significantly hypoedited in glioblastoma multiforme. Sci Rep 2017;7:2466. https://doi.org/10.1038/s41598-017-02397-6.
[87] Wang H, Chen S, Wei J, Song G, Zhao Y. A-to-I RNA Editing in Cancer: From Evaluating the Editing Level to Exploring the Editing Effects. Front Oncol 2021;10. https://doi.org/10.3389/fonc.2020.632187.
[88] Ramírez-Moya J, Baker AR, Slack FJ, Santisteban P. ADAR1-mediated RNA editing is a novel oncogenic process in thyroid cancer and regulates miR-200 activity. Oncogene 2020;39:3738–53. https://doi.org/10.1038/s41388-020-1248-x.
[89] Czubak K, Lewandowska MA, Klonowska K, Roszkowski K, Kowalewski J, Figlerowicz M, et al. High copy number variation of cancer-related microRNA genes and frequent amplification of DICER1 and DROSHA in lung cancer. Oncotarget 2015;6:23399–416. https://doi.org/10.18632/oncotarget.4351.
[90] Zhang L, Huang J, Yang N, Greshock J, Megraw MS, Giannakakis A, et al. microRNAs exhibit high frequency genomic alterations in human cancer. Proceedings of the National Academy of Sciences 2006;103:9136–41. https://doi.org/10.1073/pnas.0508889103.
[91] Shen Z, Zhou C, Li J, Ye D, Li Q, Wang J, et al. Promoter hypermethylation of *miR-34a* contributes to the risk, progression, metastasis and poor survival of laryngeal squamous cell carcinoma. Gene 2016;593:272–6. https://doi.org/10.1016/j.gene.2016.07.047.
[92] Chuang JC, Jones PA. Epigenetics and MicroRNAs. Pediatr Res 2007;61:24–9. https://doi.org/10.1203/pdr.0b013e3180457684.

[93] Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. Hum Mol Genet 2007;16:R50–9. https://doi.org/10.1093/hmg/ddm018.
[94] Suzuki H, Maruyama R, Yamamoto E, Kai M. Epigenetic alteration and microRNA dysregulation in cancer. Front Genet 2013;4:258. https://doi.org/10.3389/fgene.2013.00258.
[95] Izreig S, Samborska B, Johnson RM, Sergushichev A, Ma EH, Lussier C, et al. The miR-17~92 microRNA Cluster Is a Global Regulator of Tumor Metabolism. Cell Reports 2016;16:1915–28. https://doi.org/10.1016/j.celrep.2016.07.036.
[96] Zhou K, Liu M, Cao Y. New Insight into microRNA Functions in Cancer: Oncogene–microRNA–Tumor Suppressor Gene Network. Front Mol Biosci 2017;4. https://doi.org/10.3389/fmolb.2017.00046.
[97] Stoyanov GS, Dzhenkov D, Ghenev P, Iliev B, Enchev Y, Tonchev AB. Cell biology of glioblastoma multiforme: from basic science to diagnosis and treatment. Med Oncol 2018;35:27. https://doi.org/10.1007/s12032-018-1083-x.
[98] Buruiană A, Florian Ștefan I, Florian AI, Timiș T-L, Mihu CM, Miclăuș M, et al. The Roles of miRNA in Glioblastoma Tumor Cell Communication: Diplomatic and Aggressive Negotiations. Int J Mol Sci 2020;21:1950. https://doi.org/10.3390/ijms21061950.
[99] Makowska M, Smolarz B, Romanowicz H. microRNAs (miRNAs) in Glioblastoma Multiforme (GBM)-Recent Literature Review. Int J Mol Sci 2023;24:3521. https://doi.org/10.3390/ijms24043521.
[100] Møller HG, Rasmussen AP, Andersen HH, Johnsen KB, Henriksen M, Duroux M. A Systematic Review of MicroRNA in Glioblastoma Multiforme: Micro-modulators in the Mesenchymal Mode of Migration and Invasion. Mol Neurobiol 2013;47:131–44. https://doi.org/10.1007/s12035-012-8349-7.
[101] Aloizou A-M, Pateraki G, Siokas V, Mentis A-FA, Liampas I, Lazopoulos G, et al. The role of MiRNA-21 in gliomas: Hope for a novel therapeutic intervention? Toxicol Rep 2020;7:1514–30. https://doi.org/10.1016/j.toxrep.2020.11.001.
[102] Nieland L, Solinge TS van, Cheah PS, Morsett LM, Khoury JE, Rissman JI, et al. CRISPR-Cas knockout of miR21 reduces glioma growth. Molecular Therapy - Oncolytics 2022;25:121–36. https://doi.org/10.1016/j.omto.2022.04.001.
[103] Yang CH, Yue J, Pfeffer SR, Fan M, Paulus E, Hosni-Ahmed A, et al. MicroRNA-21 Promotes Glioblastoma Tumorigenesis by Down-regulating Insulin-like Growth Factor-binding Protein-3 (IGFBP3). J Biol Chem 2014;289:25079–87. https://doi.org/10.1074/jbc.M114.593863.
[104] Chan JA, Krichevsky AM, Kosik KS. MicroRNA-21 Is an Antiapoptotic Factor in Human Glioblastoma Cells. Cancer Res 2005;65:6029–33. https://doi.org/10.1158/0008-5472.CAN-05-0137.
[105] Jiang G, Mu J, Liu X, Peng X, Zhong F, Yuan W, et al. Prognostic value of miR-21 in gliomas: comprehensive study based on meta-analysis and TCGA dataset validation. Sci Rep 2020;10:4220. https://doi.org/10.1038/s41598-020-61155-3.
[106] Zhang Y, Rabinovsky R, Wei Z, Fatimy RE, Deforzh E, Luan B, et al. Secreted PGK1 and IGFBP2 contribute to the bystander effect of miR-10b gene editing in glioma. Molecular Therapy Nucleic Acids 2023;31:265–75. https://doi.org/10.1016/j.omtn.2022.12.018.
[107] Gabriely G, Yi M, Narayan RS, Niers JM, Wurdinger T, Imitola J, et al. Human glioma growth is controlled by microRNA-10b. Cancer Research 2011;71:3563–72. https://doi.org/10.1158/0008-5472.CAN-10-3568.

[108] Bhere D, Arghiani N, Lechtich ER, Yao Y, Alsaab S, Bei F, et al. Simultaneous downregulation of miR-21 and upregulation of miR-7 has anti-tumor efficacy. Sci Rep 2020;10:1779. https://doi.org/10.1038/s41598-020-58072-w.
[109] Shang C, Hong Y, Guo Y, Liu Y-H, Xue Y-X. miR-128 regulates the apoptosis and proliferation of glioma cells by targeting RhoE. Oncology Letters 2016;11:904–8. https://doi.org/10.3892/ol.2015.3927.
[110] Abels ER, Maas SLN, Nieland L, Wei Z, Cheah PS, Tai E, et al. Glioblastoma-Associated Microglia Reprogramming Is Mediated by Functional Transfer of Extracellular miR-21. Cell Rep 2019;28:3105-3119.e7. https://doi.org/10.1016/j.celrep.2019.08.036.
[111] Grigore IA, Rajagopal A, Chow JT-S, Stone TJ, Salmena L. Discovery of miRNA–mRNA regulatory networks in glioblastoma reveals novel insights into tumor microenvironment remodeling. Sci Rep 2024;14:27493. https://doi.org/10.1038/s41598-024-78337-y.
[112] Hide T, Shibahara I, Kumabe T. Novel concept of the border niche: glioblastoma cells use oligodendrocytes progenitor cells (GAOs) and microglia to acquire stem cell-like features. Brain Tumor Pathol 2019;36:63–73. https://doi.org/10.1007/s10014-019-00341-2.
[113] Hide T, Komohara Y, Miyasato Y, Nakamura H, Makino K, Takeya M, et al. Oligodendrocyte Progenitor Cells and Macrophages/Microglia Produce Glioma Stem Cell Niches at the Tumor Border. eBioMedicine 2018;30:94–104. https://doi.org/10.1016/j.ebiom.2018.02.024.
[114] Zhang W, Zhang J, Hoadley K, Kushwaha D, Ramakrishnan V, Li S, et al. miR-181d: a predictive glioblastoma biomarker that downregulates MGMT expression. Neuro Oncol 2012;14:712–9. https://doi.org/10.1093/neuonc/nos089.
[115] 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.
[116] Yalamanchili HK, Alcott CE, Ji P, Wagner EJ, Zoghbi HY, Liu Z. PolyA-miner: accurate assessment of differential alternative poly-adenylation from 3′Seq data using vector projections and non-negative matrix factorization. Nucleic Acids Res 2020;48:e69. https://doi.org/10.1093/nar/gkaa398.
[117] Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, 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.
[118] Danckwardt S, Hentze MW, Kulozik AE. 3′ end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J 2008;27:482–98. https://doi.org/10.1038/sj.emboj.7601932.
[119] Zhu Y, Wang X, Forouzmand E, Jeong J, Qiao F, Sowd GA, et al. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. Mol Cell 2018;69:62-74.e4. https://doi.org/10.1016/j.molcel.2017.11.031.
[120] Liu S, Wu R, Chen L, Deng K, Ou X, Lu X, et al. CPSF6 regulates alternative polyadenylation and proliferation of cancer cells through phase separation. Cell Rep 2023;42:113197. https://doi.org/10.1016/j.celrep.2023.113197.
[121] Jonnakuti VS, Ji P, Gao Y, Lin A, Chu Y, Elrod N, et al. NUDT21 alters glioma migration through differential alternative polyadenylation of LAMC1. J Neurooncol 2023;163:623–34. https://doi.org/10.1007/s11060-023-04370-y.
[122] Ghosh S, Ataman M, Bak M, Börsch A, Schmidt A, Buczak K, et al. CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA

biogenesis. Nucleic Acids Res 2022;50:3096–114. https://doi.org/10.1093/nar/gkac114.
[123] Masamha CP, Xia Z, Yang J, Albrecht TR, Li M, Shyu A-B, et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature 2014;510:412–6. https://doi.org/10.1038/nature13261.
[124] Fahmi NA, Saha S, Song Q, Lou Q, Yong J, Zhang W. Computational methods for alternative polyadenylation and splicing in post-transcriptional gene regulation. Exp Mol Med 2025;57:1631–40. https://doi.org/10.1038/s12276-025-01496-z.
[125] Fu Y, Chen L, Chen C, Ge Y, Kang M, Song Z, et al. Crosstalk between alternative polyadenylation and miRNAs in the regulation of protein translational efficiency. Genome Res 2018;28:1656–63. https://doi.org/10.1101/gr.231506.117.
[126] Gruber AR, Martin G, Keller W, Zavolan M. Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors. WIREs RNA 2014;5:183–96. https://doi.org/10.1002/wrna.1206.
[127] Geisberg JV, Moqtaderi Z, Struhl K. Chromatin regulates alternative polyadenylation via the RNA polymerase II elongation rate. Proceedings of the National Academy of Sciences 2024;121:e2405827121. https://doi.org/10.1073/pnas.2405827121.
[128] Geisberg JV, Moqtaderi Z, Struhl K. The transcriptional elongation rate regulates alternative polyadenylation in yeast. eLife 2020;9:e59810. https://doi.org/10.7554/eLife.59810.
[129] Mao Z, Zhao H, Qin Y, Wei J, Sun J, Zhang W, et al. Post-Transcriptional Dysregulation of microRNA and Alternative Polyadenylation in Colorectal Cancer. Front Genet 2020;11. https://doi.org/10.3389/fgene.2020.00064.
[130] Mitschka S, Mayr C. Context-specific regulation and function of mRNA alternative polyadenylation. Nat Rev Mol Cell Biol 2022;23:779–96. https://doi.org/10.1038/s41580-022-00507-5.
[131] Sommerkamp P, Cabezas-Wallscheid N, Trumpp A. Alternative Polyadenylation in Stem Cell Self-Renewal and Differentiation. Trends in Molecular Medicine 2021;27:660–72. https://doi.org/10.1016/j.molmed.2021.04.006.
[132] Gallicchio L, Olivares GH, Berry CW, Fuller MT. Regulation and function of alternative polyadenylation in development and differentiation. RNA Biology 2023;20:908–25. https://doi.org/10.1080/15476286.2023.2275109.
[133] MacDonald CC. Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond (2018 update). Wiley Interdiscip Rev RNA 2019;10:e1526. https://doi.org/10.1002/wrna.1526.
[134] Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proceedings of the National Academy of Sciences 2009;106:7028–33. https://doi.org/10.1073/pnas.0900028106.
[135] Agarwal V, Lopez-Darwin S, Kelley DR, Shendure J. The landscape of alternative polyadenylation in single cells of the developing mouse embryo. Nat Commun 2021;12:5101. https://doi.org/10.1038/s41467-021-25388-8.
[136] Hoffman Y, Bublik DR, Ugalde AP, Elkon R, Biniashvili T, Agami R, et al. 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. PLOS Genetics 2016;12:e1005879. https://doi.org/10.1371/journal.pgen.1005879.
[137] Zhang H, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. Genome Biol 2005;6:R100. https://doi.org/10.1186/gb-2005-6-12-r100.

[138] Tushev G, Glock C, Heumüller M, Biever A, Jovanovic M, Schuman EM. Alternative 3′ UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron 2018;98:495-511.e6. https://doi.org/10.1016/j.neuron.2018.03.030.
[139] Vicario A, Colliva A, Ratti A, Davidovic L, Baj G, Gricman Ł, et al. Dendritic targeting of short and long 3′ UTR BDNF mRNA is regulated by BDNF or NT-3 and distinct sets of RNA-binding proteins. Front Mol Neurosci 2015;8:62. https://doi.org/10.3389/fnmol.2015.00062.
[140] Lau AG, Irier HA, Gu J, Tian D, Ku L, Liu G, et al. Distinct 3′UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). Proceedings of the National Academy of Sciences 2010;107:15945–50. https://doi.org/10.1073/pnas.1002929107.
[141] Gebhardt ML, Reuter S, Mrowka R, Andrade-Navarro MA. Similarity in targets with REST points to neural and glioblastoma related miRNAs. Nucleic Acids Res 2014;42:5436–46. https://doi.org/10.1093/nar/gku231.
[142] Blake D, Gazzara MR, Breuer I, Ferretti M, Lynch KW. Alternative 3′UTR expression induced by T cell activation is regulated in a temporal and signal dependent manner. Sci Rep 2024;14:10987. https://doi.org/10.1038/s41598-024-61951-1.
[143] Seyres D, Gorka O, Schmidt R, Marone R, Zavolan M, Jeker LT. T helper cells exhibit a dynamic and reversible 3′-UTR landscape. RNA 2024;30:418–34. https://doi.org/10.1261/rna.079897.123.
[144] Li L, Wang D, Xue M, Mi X, Liang Y, Wang P. 3′UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network. Sci Rep 2014;4:5406. https://doi.org/10.1038/srep05406.
[145] Fan Z, Kim S, Bai Y, Diergaarde B, Park HJ. 3′-UTR Shortening Contributes to Subtype-Specific Cancer Growth by Breaking Stable ceRNA Crosstalk of Housekeeping Genes. Front Bioeng Biotechnol 2020;8. https://doi.org/10.3389/fbioe.2020.00334.
[146] Chen L, Heikkinen L, Wang C, Yang Y, Sun H, Wong G. Trends in the development of miRNA bioinformatics tools. Brief Bioinform 2019;20:1836–52. https://doi.org/10.1093/bib/bby054.
[147] Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell 2009;136:642–55. https://doi.org/10.1016/j.cell.2009.01.035.
[148] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 2005;37:766–70. https://doi.org/10.1038/ng1590.
[149] 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.
[150] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006;34:D140-144. https://doi.org/10.1093/nar/gkj112.
[151] Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, et al. An estimate of the total number of true human miRNAs. Nucleic Acids Res 2019;47:3353–64. https://doi.org/10.1093/nar/gkz097.
[152] Fromm B, Domanska D, Høye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, et al. MirGeneDB 2.0: the metazoan microRNA complement. Nucleic Acids Res 2020;48:D132–41. https://doi.org/10.1093/nar/gkz885.
[153] Umu SU, Paynter VM, Trondsen H, Buschmann T, Rounge TB, Peterson KJ, et al. Accurate microRNA annotation of animal genomes using trained covariance

models of curated microRNA complements in MirMachine. Cell Genomics 2023;3:100348. https://doi.org/10.1016/j.xgen.2023.100348.
[154] Hackenberg M, Kalogeropoulos P, Peterson KJ, Friedländer MR, Fromm B. Knowing is not enough; we must apply: the case for rigorous microRNA annotation standards. Nucleic Acids Res 2025;53:gkaf1049. https://doi.org/10.1093/nar/gkaf1049.
[155] Barbato C, Arisi I, Frizzo ME, Brandi R, Da Sacco L, Masotti A. Computational Challenges in miRNA Target Predictions: To Be or Not to Be a True Target? BioMed Research International 2009;2009:803069. https://doi.org/10.1155/2009/803069.
[156] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic Acids Res 2018;46:D239–45. https://doi.org/10.1093/nar/gkx1141.
[157] Tastsoglou S, Alexiou A, Karagkouni D, Skoufos G, Zacharopoulou E, Hatzigeorgiou AG. DIANA-microT 2023: including predicted targets of virally encoded miRNAs. Nucleic Acids Research 2023;51:W148–53. https://doi.org/10.1093/nar/gkad283.
[158] Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015;4:e05005. https://doi.org/10.7554/eLife.05005.
[159] 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;27:234–45. https://doi.org/10.1101/gr.205146.116.
[160] Reyes~Herrera PH, Ficarra E. One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms. Genom Proteom Bioinform 2012;10:254–63. https://doi.org/10.1016/j.gpb.2012.10.001.
[161] Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB. Common features of microRNA target prediction tools. Front Genet 2014;5. https://doi.org/10.3389/fgene.2014.00023.
[162] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol 2004;2:e363. https://doi.org/10.1371/journal.pbio.0020363.
[163] Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, et al. MicroRNA target prediction by expression analysis of host genes. Genome Res 2009;19:481–90. https://doi.org/10.1101/gr.084129.108.
[164] Pla A, Zhong X, Rayner S. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. PLoS Comput Biol 2018;14:e1006185. https://doi.org/10.1371/journal.pcbi.1006185.
[165] Bi Y, Li F, Wang C, Pan T, Davidovich C, Webb GI, et al. Advancing microRNA target site prediction with transformer and base-pairing patterns. Nucleic Acids Res 2024;52:11455–65. https://doi.org/10.1093/nar/gkae782.
[166] Saçar Demirci MD, Yousef M, Allmer J. Computational Prediction of Functional MicroRNA-mRNA Interactions. Methods Mol Biol 2019;1912:175–96. https://doi.org/10.1007/978-1-4939-8982-9_7.
[167] Cui S, Yu S, Huang H-Y, Lin Y-C-D, Huang Y, Zhang B, et al. miRTarBase 2025: updates to the collection of experimentally validated microRNA–target interactions. Nucleic Acids Research 2025;53:D147–56. https://doi.org/10.1093/nar/gkae1072.
[168] Huang H-Y, Lin Y-C-D, Cui S, Huang Y, Tang Y, Xu J, et al. miRTarBase update 2022: an informative resource for experimentally validated miRNA-target

interactions. Nucleic Acids Res 2022;50:D222–30. https://doi.org/10.1093/nar/gkab1079.
[169] Li J, Han X, Wan Y, Zhang S, Zhao Y, Fan R, et al. TAM 2.0: tool for MicroRNA set analysis. Nucleic Acids Res 2018;46:W180–5. https://doi.org/10.1093/nar/gky509.
[170] Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic Acids Res 2015;43:W460-466. https://doi.org/10.1093/nar/gkv403.
[171] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.
[172] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457-462. https://doi.org/10.1093/nar/gkv1070.
[173] Kavakiotis I, Alexiou A, Tastsoglou S, Vlachos IS, Hatzigeorgiou AG. DIANA-miTED: a microRNA tissue expression database. Nucleic Acids Research 2022;50:D1055–61. https://doi.org/10.1093/nar/gkab733.
[174] Keller A, Gröger L, Tschernig T, Solomon J, Laham O, Schaum N, et al. miRNATissueAtlas2: an update to the human miRNA tissue atlas. Nucleic Acids Research 2022;50:D211–21. https://doi.org/10.1093/nar/gkab808.
[175] Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Research 2020;48:W244–51. https://doi.org/10.1093/nar/gkaa467.
[176] Xie G-Y, Xia M, Miao Y-R, Luo M, Zhang Q, Guo A-Y. FFLtool: a web server for transcription factor and miRNA feed forward loop analysis in human. Bioinformatics 2020;36:2605–7. https://doi.org/10.1093/bioinformatics/btz929.
[177] WU Q, QIN H, ZHAO Q, HE X-X. Emerging role of transcription factor-microRNA-target gene feed-forward loops in cancer. Biomed Rep 2015;3:611–6. https://doi.org/10.3892/br.2015.477.
[178] Jiang W, Mitra R, Lin C-C, Wang Q, Cheng F, Zhao Z. Systematic dissection of dysregulated transcription factor-miRNA feed-forward loops across tumor types. Brief Bioinform 2016;17:996–1008. https://doi.org/10.1093/bib/bbv107.
[179] Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell 2007;26:753–67. https://doi.org/10.1016/j.molcel.2007.05.018.
[180] Baroudi ME, Corà D, Bosia C, Osella M, Caselle M. A Curated Database of miRNA Mediated Feed-Forward Loops Involving MYC as Master Regulator. PLOS ONE 2011;6:e14742. https://doi.org/10.1371/journal.pone.0014742.
[181] Martinez NJ, Walhout AJM. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. BioEssays 2009;31:435–45. https://doi.org/10.1002/bies.200800212.
[182] Taganov KD, Boldin MP, Chang K-J, Baltimore D. NF-κB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. Proceedings of the National Academy of Sciences 2006;103:12481–6. https://doi.org/10.1073/pnas.0605298103.
[183] Han R, Gao J, Wang L, Hao P, Chen X, Wang Y, et al. MicroRNA-146a negatively regulates inflammation via the IRAK1/TRAF6/NF-κB signaling pathway in dry eye. Sci Rep 2023;13:11192. https://doi.org/10.1038/s41598-023-38367-4.
[184] Gosline SJC, Gurtan AM, JnBaptiste CK, Bosson A, Milani P, Dalin S, et al. Elucidating microRNA regulatory networks using transcriptional, post-

transcriptional and histone modification measurements. Cell Rep 2016;14:310–9. https://doi.org/10.1016/j.celrep.2015.12.031.
[185] Erhard F, Haas J, Lieber D, Malterer G, Jaskiewicz L, Zavolan M, et al. Widespread context dependency of microRNA-mediated regulation. Genome Res 2014;24:906–19. https://doi.org/10.1101/gr.166702.113.
[186] Sibilio P, De Smaele E, Paci P, Conte F. Integrating multi-omics data: Methods and applications in human complex diseases. Biotechnology Reports 2025;48:e00938. https://doi.org/10.1016/j.btre.2025.e00938.
[187] Li B, Wang C, Wang Y, Li P, Liu Z-P. RegNetwork 2025: an integrative data repository for gene regulatory networks in human and mouse. Nucleic Acids Res 2026;54:D1234–41. https://doi.org/10.1093/nar/gkaf779.
[188] Pezoulas VC, Hazapis O, Lagopati N, Exarchos TP, Goules AV, Tzioufas AG, et al. Machine Learning Approaches on High Throughput NGS Data to Unveil Mechanisms of Function in Biology and Disease. Cancer Genomics Proteomics 2021;18:605–26. https://doi.org/10.21873/cgp.20284.
[189] Hegde A, Nguyen T, Cheng J. Machine learning methods for gene regulatory network inference. Brief Bioinform 2025;26:bbaf470. https://doi.org/10.1093/bib/bbaf470.
[190] Erbe R, Gore J, Gemmill K, Gaykalova DA, Fertig EJ. The use of machine learning to discover regulatory networks controlling biological systems. Molecular Cell 2022;82:260–73. https://doi.org/10.1016/j.molcel.2021.12.011.
[191] Razaghi-Moghadam Z, Nikoloski Z. Supervised learning of gene-regulatory networks based on graph distance profiles of transcriptomics data. Npj Syst Biol Appl 2020;6:21. https://doi.org/10.1038/s41540-020-0140-1.
[192] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.
[193] Heil BJ, Crawford J, Greene CS. The effect of non-linear signal in classification problems using gene expression. PLoS Comput Biol 2023;19:e1010984. https://doi.org/10.1371/journal.pcbi.1010984.
[194] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970;12:55–67. https://doi.org/10.1080/00401706.1970.10488634.
[195] McDonald GC. Ridge regression. WIREs Computational Statistics 2009;1:93–100. https://doi.org/10.1002/wics.14.
[196] Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Brief Bioinform 2015;16:780–94. https://doi.org/10.1093/bib/bbu044.
[197] Hadad E, Rokach L, Veksler-Lublinsky I. Empowering prediction of miRNA-mRNA interactions in species with limited training data through transfer learning. Heliyon 2024;10:e28000. https://doi.org/10.1016/j.heliyon.2024.e28000.
[198] Olgun G, Gopalan V, Hannenhalli S. miRSCAPE - inferring miRNA expression from scRNA-seq data. iScience 2022;25. https://doi.org/10.1016/j.isci.2022.104962.
[199] Stempor PA, Cauchi M, Wilson P. MMpred: functional miRNA – mRNA interaction analyses by miRNA expression prediction. BMC Genomics 2012;13:620. https://doi.org/10.1186/1471-2164-13-620.
[200] Mezlini AM, Wang B, Deshwar A, Morris Q, Goldenberg A. Identifying Cancer Specific Functionally Relevant miRNAs from Gene Expression and miRNA-to-Gene Networks Using Regularized Regression. PLOS ONE 2013;8:e73168. https://doi.org/10.1371/journal.pone.0073168.

[201] Gonzalo-Calvo D de, Karaduzovic-Hadziabdic K, Dalgaard LT, Dieterich C, Perez-Pons M, Hatzigeorgiou A, et al. Machine learning for catalysing the integration of noncoding RNA in research and clinical practice. eBioMedicine 2024;106. https://doi.org/10.1016/j.ebiom.2024.105247.
[202] Wu P, Li D, Zhang C, Dai B, Tang X, Liu J, et al. A unique circulating microRNA pairs signature serves as a superior tool for early diagnosis of pan-cancer. Cancer Letters 2024;588:216655. https://doi.org/10.1016/j.canlet.2024.216655.
[203] Singh S, Pathak AK, Kural S, Kumar L, Bhardwaj MG, Yadav M, et al. Integrating miRNA profiling and machine learning for improved prostate cancer diagnosis. Sci Rep 2025;15:30477. https://doi.org/10.1038/s41598-025-99754-7.
[204] Hamidi F, Gilani N, Kazemnejad A, Aftabi Y, Shirforoush-Sattari M, Jahanimoghadam A. Decision tree-based machine learning methods for identifying colorectal cancer-associated microRNA signatures and their regulatory networks. Sci Rep 2025;15:34700. https://doi.org/10.1038/s41598-025-17037-7.
[205] Zacharopoulou E, Paraskevopoulou MD, Tastsoglou S, Alexiou A, Karavangeli A, Pierros V, et al. microT-CNN: an avant-garde deep convolutional neural network unravels functional miRNA targets beyond canonical sites. Brief Bioinform 2025;26:bbae678. https://doi.org/10.1093/bib/bbae678.
[206] Wang Y, Miller DJ, Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. Br J Cancer 2008;98:1023–8. https://doi.org/10.1038/sj.bjc.6604207.
[207] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Royal Statistical Society Journal Series B: Methodological 1996;58:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[208] De Landtsheer S, Lucarelli P, Sauter T. Using Regularization to Infer Cell Line Specificity in Logical Network Models of Signaling Pathways. Front Physiol 2018;9:550. https://doi.org/10.3389/fphys.2018.00550.
[209] Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Ser B Stat Methodol 2005;67:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
[210] Haramati S, Chapnik E, Sztainberg Y, Eilam R, Zwang R, Gershoni N, et al. miRNA malfunction causes spinal motor neuron disease. Proc Natl Acad Sci U S A 2010;107:13111–6. https://doi.org/10.1073/pnas.1006151107.
[211] Adams BD, Kasinski AL, Slack FJ. Aberrant regulation and function of microRNAs in cancer. Curr Biol 2014;24:R762-776. https://doi.org/10.1016/j.cub.2014.06.043.
[212] Chen X, Xie D, Zhao Q, You Z-H. MicroRNAs and complex diseases: from experimental results to computational models. Brief Bioinform 2019;20:515–39. https://doi.org/10.1093/bib/bbx130.
[213] Dweep H, Gretz N, Sticht C. miRWalk Database for miRNA–Target Interactions. In: Alvarez ML, Nourbakhsh M, editors. RNA Mapping: Methods and Protocols, New York, NY: Springer; 2014, p. 289–305. https://doi.org/10.1007/978-1-4939-1062-5_25.
[214] Miska EA, Alvarez-Saavedra E, Abbott AL, Lau NC, Hellman AB, McGonagle SM, et al. Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. PLoS Genet 2007;3:e215. https://doi.org/10.1371/journal.pgen.0030215.

[215] Lewis BP, Shih I -hung, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of Mammalian MicroRNA Targets. Cell 2003;115:787–98. https://doi.org/10.1016/S0092-8674(03)01018-3.
[216] Li H, Xie S, Liu X, Wu H, Lin X, Gu J, et al. Matrine alters microRNA expression profiles in SGC-7901 human gastric cancer cells. Oncol Rep 2014;32:2118–26. https://doi.org/10.3892/or.2014.3447.
[217] Hobert O. Common logic of transcription factor and microRNA action. Trends Biochem Sci 2004;29:462–8. https://doi.org/10.1016/j.tibs.2004.07.001.
[218] Hobert O. Gene regulation by transcription factors and microRNAs. Science 2008;319:1785–6. https://doi.org/10.1126/science.1151651.
[219] Shalgi R, Lieber D, Oren M, Pilpel Y. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. PLoS Comput Biol 2007;3:e131. https://doi.org/10.1371/journal.pcbi.0030131.
[220] Coulson JM. Transcriptional regulation: cancer, neurons and the REST. Curr Biol 2005;15:R665-668. https://doi.org/10.1016/j.cub.2005.08.032.
[221] Wu S, Ai N, Liu Q, Zhang J. MicroRNA-448 inhibits the progression of retinoblastoma by directly targeting ROCK1 and regulating PI3K/AKT signalling pathway. Oncol Rep 2018;39:2402–12. https://doi.org/10.3892/or.2018.6302.
[222] Chen H, Li H, Liu F, Zheng X, Wang S, Bo X, et al. An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. Sci Rep 2015;5:8465. https://doi.org/10.1038/srep08465.
[223] Nazarov PV, Kreis S. Integrative approaches for analysis of mRNA and microRNA high-throughput data. Comput Struct Biotechnol J 2021;19:1154–62. https://doi.org/10.1016/j.csbj.2021.01.029.
[224] Prompsy PB, Toubia J, Gearing LJ, Knight RL, Forster SC, Bracken CP, et al. Making use of transcription factor enrichment to identify functional microRNA-regulons. Comput Struct Biotechnol J 2021;19:4896–903. https://doi.org/10.1016/j.csbj.2021.08.032.
[225] Bersten DC, Wright JA, McCarthy PJ, Whitelaw ML. Regulation of the neuronal transcription factor NPAS4 by REST and microRNAs. Biochim Biophys Acta 2014;1839:13–24. https://doi.org/10.1016/j.bbagrm.2013.11.004.
[226] Sauer M, Was N, Ziegenhals T, Wang X, Hafner M, Becker M, et al. The miR-26 family regulates neural differentiation-associated microRNAs and mRNAs by directly targeting REST. J Cell Sci 2021;134:jcs257535. https://doi.org/10.1242/jcs.257535.
[227] Wu J, Xie X. Comparative sequence analysis reveals an intricate network among REST, CREB and miRNA in mediating neuronal gene expression. Genome Biol 2006;7:R85. https://doi.org/10.1186/gb-2006-7-9-r85.
[228] Grimson A, Farh KK-H, 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.
[229] Plass M, Rasmussen SH, Krogh A. Highly accessible AU-rich regions in 3' untranslated regions are hotspots for binding of regulatory factors. PLoS Comput Biol 2017;13:e1005460. https://doi.org/10.1371/journal.pcbi.1005460.
[230] Chan CY, Carmack CS, Long DD, Maliyekkel A, Shao Y, Roninson IB, et al. A structural interpretation of the effect of GC-content on efficiency of RNA interference. BMC Bioinformatics 2009;10 Suppl 1:S33. https://doi.org/10.1186/1471-2105-10-S1-S33.
[231] Fisher R. Statistical methods for research workers 1934.

[232] Kondo T. Glioblastoma-initiating cell heterogeneity generated by the cell-of-origin, genetic/epigenetic mutation and microenvironment. Seminars in Cancer Biology 2022;82:176–83. https://doi.org/10.1016/j.semcancer.2020.12.003.
[233] Yao M, Li S, Wu X, Diao S, Zhang G, He H, et al. Cellular origin of glioblastoma and its implication in precision therapy. Cell Mol Immunol 2018;15:737–9. https://doi.org/10.1038/cmi.2017.159.
[234] Zong H, Parada LF, Baker SJ. Cell of Origin for Malignant Gliomas and Its Implication in Therapeutic Development. Cold Spring Harb Perspect Biol 2015;7:a020610. https://doi.org/10.1101/cshperspect.a020610.
[235] Kawamura Y, Takouda J, Yoshimoto K, Nakashima K. New aspects of glioblastoma multiforme revealed by similarities between neural and glioblastoma stem cells. Cell Biol Toxicol 2018;34:425–40. https://doi.org/10.1007/s10565-017-9420-y.
[236] Lathia JD, Mack SC, Mulkearns-Hubert EE, Valentim CLL, Rich JN. Cancer stem cells in glioblastoma. Genes Dev 2015;29:1203–17. https://doi.org/10.1101/gad.261982.115.
[237] Singh SK, Clarke ID, Terasaki M, Bonn VE, Hawkins C, Squire J, et al. Identification of a cancer stem cell in human brain tumors. Cancer Res 2003;63:5821–8.
[238] Singh SK, Hawkins C, Clarke ID, Squire JA, Bayani J, Hide T, et al. Identification of human brain tumour initiating cells. Nature 2004;432:396–401. https://doi.org/10.1038/nature03128.
[239] Wang Y, Yang J, Zheng H, Tomasek GJ, Zhang P, McKeever PE, et al. Expression of mutant p53 proteins implicates a lineage relationship between neural stem cells and malignant astrocytic glioma in a murine model. Cancer Cell 2009;15:514–26. https://doi.org/10.1016/j.ccr.2009.04.001.
[240] Liu Q, Nguyen DH, Dong Q, Shitaku P, Chung K, Liu OY, et al. Molecular properties of CD133+ glioblastoma stem cells derived from treatment-refractory recurrent brain tumors. J Neurooncol 2009;94:1–19. https://doi.org/10.1007/s11060-009-9919-z.
[241] Pavon LF, Sibov TT, de Oliveira DM, Marti LC, Cabral FR, de Souza JG, et al. Mesenchymal stem cell-like properties of CD133+ glioblastoma initiating cells. Oncotarget 2016;7:40546–57. https://doi.org/10.18632/oncotarget.9658.
[242] Gorris R, Fischer J, Erwes KL, Kesavan J, Peterson DA, Alexander M, et al. Pluripotent stem cell-derived radial glia-like cells as stable intermediate for efficient generation of human oligodendrocytes. Glia 2015;63:2152–67. https://doi.org/10.1002/glia.22882.
[243] Huang W, Bhaduri A, Velmeshev D, Wang S, Wang L, Rottkamp CA, et al. Origins and Proliferative States of Human Oligodendrocyte Precursor Cells. Cell 2020;182:594-608.e11. https://doi.org/10.1016/j.cell.2020.06.027.
[244] Pollen AA, Nowakowski TJ, Chen J, Retallack H, Sandoval-Espinosa C, Nicholas CR, et al. Molecular Identity of Human Outer Radial Glia during Cortical Development. Cell 2015;163:55–67. https://doi.org/10.1016/j.cell.2015.09.004.
[245] Zhao X, He X, Han X, Yu Y, Ye F, Chen Y, et al. MicroRNA-Mediated Control of Oligodendrocyte Differentiation. Neuron 2010;65:612–26. https://doi.org/10.1016/j.neuron.2010.02.018.
[246] Shea A, Harish V, Afzal Z, Chijioke J, Kedir H, Dusmatova S, et al. MicroRNAs in glioblastoma multiforme pathogenesis and therapeutics. Cancer Med 2016;5:1917–46. https://doi.org/10.1002/cam4.775.

[247] Banelli B, Forlani A, Allemanni G, Morabito A, Pistillo MP, Romani M. MicroRNA in Glioblastoma: An Overview. Int J Genomics 2017;2017:7639084. https://doi.org/10.1155/2017/7639084.
[248] Hasan H, Afzal M, Castresana JS, Shahi MH. A Comprehensive Review of miRNAs and Their Epigenetic Effects in Glioblastoma. Cells 2023;12:1578. https://doi.org/10.3390/cells12121578.
[249] Kohlhapp FJ, Mitra AK, Lengyel E, Peter ME. MicroRNAs as mediators and communicators between cancer cells and the tumor microenvironment. Oncogene 2015;34:5857–68. https://doi.org/10.1038/onc.2015.89.
[250] An J, Zhu X, Wang H, Jin X. A dynamic interplay between alternative polyadenylation and microRNA regulation: implications for cancer (Review). Int J Oncol 2013;43:995–1001. https://doi.org/10.3892/ijo.2013.2047.
[251] Cihan M, Andrade-Navarro MA. Detection of features predictive of microRNA targets by integration of network data. PLoS One 2022;17:e0269731. https://doi.org/10.1371/journal.pone.0269731.
[252] Arora S, Rana R, Chhabra A, Jaiswal A, Rani V. miRNA-transcription factor interactions: a combinatorial regulation of gene expression. Mol Genet Genomics 2013;288:77–87. https://doi.org/10.1007/s00438-013-0734-z.
[253] Bo C, Zhang H, Cao Y, Lu X, Zhang C, Li S, et al. Construction of a TF–miRNA–gene feed-forward loop network predicts biomarkers and potential drugs for myasthenia gravis. Sci Rep 2021;11:2416. https://doi.org/10.1038/s41598-021-81962-6.
[254] Chiarella E, Aloisio A, Scicchitano S, Bond HM, Mesuraca M. Regulatory Role of microRNAs Targeting the Transcription Co-Factor ZNF521 in Normal Tissues and Cancers. Int J Mol Sci 2021;22:8461. https://doi.org/10.3390/ijms22168461.
[255] Qin G, Mallik S, Mitra R, Li A, Jia P, Eischen CM, et al. MicroRNA and transcription factor co-regulatory networks and subtype classification of seminoma and non-seminoma in testicular germ cell tumors. Sci Rep 2020;10:852. https://doi.org/10.1038/s41598-020-57834-w.
[256] Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3’ end sequencing. Nucleic Acids Res 2020;48:D174–9. https://doi.org/10.1093/nar/gkz918.
[257] Zhang H, Hu J, Recce M, Tian B. PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res 2005;33:D116-120. https://doi.org/10.1093/nar/gki055.
[258] Azuaje F, Tiemann K, Niclou SP. Therapeutic control and resistance of the EGFR-driven signaling network in glioblastoma. Cell Commun Signal 2015;13:23. https://doi.org/10.1186/s12964-015-0098-6.
[259] Pan PC, Magge RS. Mechanisms of EGFR Resistance in Glioblastoma. International Journal of Molecular Sciences 2020;21:8471. https://doi.org/10.3390/ijms21228471.
[260] Rodriguez SMB, Kamel A, Ciubotaru GV, Onose G, Sevastre A-S, Sfredel V, et al. An Overview of EGFR Mechanisms and Their Implications in Targeted Therapies for Glioblastoma. International Journal of Molecular Sciences 2023;24:11110. https://doi.org/10.3390/ijms241311110.
[261] Lee Y, Lee J-K, Ahn SH, Lee J, Nam D-H. WNT signaling in glioblastoma and therapeutic opportunities. Laboratory Investigation 2016;96:137–50. https://doi.org/10.1038/labinvest.2015.140.
[262] Lu Y, Tian M, Liu J, Wang K. LINC00511 facilitates Temozolomide resistance of glioblastoma cells via sponging miR-126-5p and activating Wnt/β-catenin

signaling. Journal of Biochemical and Molecular Toxicology 2021;35:e22848. https://doi.org/10.1002/jbt.22848.
[263] Kim W, Youn H, Lee S, Kim E, Kim D, Sub Lee J, et al. RNF138-mediated ubiquitination of rpS3 is required for resistance of glioblastoma cells to radiation-induced apoptosis. Exp Mol Med 2018;50:e434–e434. https://doi.org/10.1038/emm.2017.247.
[264] Annovazzi L, Mellai M, Schiffer D. Chemotherapeutic Drugs: DNA Damage and Repair in Glioblastoma. Cancers 2017;9:57. https://doi.org/10.3390/cancers9060057.
[265] Erasimus H, Gobin M, Niclou S, Van Dyck E. DNA repair mechanisms and their clinical impact in glioblastoma. Mutation Research/Reviews in Mutation Research 2016;769:19–35. https://doi.org/10.1016/j.mrrev.2016.05.005.
[266] Chu Y, Elrod N, Wang C, Li L, Chen T, Routh A, et al. Nudt21 regulates the alternative polyadenylation of Pak1 and is predictive in the prognosis of glioblastoma patients. Oncogene 2019;38:4154–68. https://doi.org/10.1038/s41388-019-0714-9.
[267] Han T, Kim JK. Driving glioblastoma growth by alternative polyadenylation. Cell Res 2014;24:1023–4. https://doi.org/10.1038/cr.2014.88.
[268] Krakstad C, Chekenya M. Survival signalling and apoptosis resistance in glioblastomas: opportunities for targeted therapeutics. Mol Cancer 2010;9:135. https://doi.org/10.1186/1476-4598-9-135.
[269] Chen M, Medarova Z, Moore A. Role of microRNAs in glioblastoma. Oncotarget 2021;12:1707–23. https://doi.org/10.18632/oncotarget.28039.
[270] Masoudi MS, Mehrabian E, Mirzaei H. MiR-21: A key player in glioblastoma pathogenesis. Journal of Cellular Biochemistry 2018;119:1285–90. https://doi.org/10.1002/jcb.26300.
[271] Mitra R, Edmonds MD, Sun J, Zhao M, Yu H, Eischen CM, et al. Reproducible combinatorial regulatory networks elucidate novel oncogenic microRNAs in non-small cell lung cancer. RNA 2014;20:1356–68. https://doi.org/10.1261/rna.042754.113.
[272] Liu Z, Borlak J, Tong W. Deciphering miRNA transcription factor feed-forward loops to identify drug repurposing candidates for cystic fibrosis. Genome Med 2014;6:94. https://doi.org/10.1186/s13073-014-0094-2.
[273] Wang R, Sharma R, Shen X, Laughney AM, Funato K, Clark PJ, et al. Adult Human Glioblastomas Harbor Radial Glia-like Cells. Stem Cell Reports 2020;14:338–50. https://doi.org/10.1016/j.stemcr.2020.01.007.
[274] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049. https://doi.org/10.1038/ncomms14049.
[275] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell 2019;177:1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
[276] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 2006;34:D108-110. https://doi.org/10.1093/nar/gkj143.
[277] 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.
[278] Zhou R, Xiao X, He P, Zhao Y, Xu M, Zheng X, et al. SCAPE: a mixture model revealing single-cell polyadenylation diversity and cellular dynamics during cell

differentiation and reprogramming. Nucleic Acids Res 2022;50:e66. https://doi.org/10.1093/nar/gkac167.
[279] 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 (Methodological) 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
[280] Wang J, Lu M, Qiu C, Cui Q. TransmiR: a transcription factor–microRNA regulation database. Nucleic Acids Res 2010;38:D119–22. https://doi.org/10.1093/nar/gkp803.
[281] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell 2018;172:650–65. https://doi.org/10.1016/j.cell.2018.01.029.
[282] Ji J, Anwar M, Petretto E, Emanueli C, Srivastava PK. PPMS: A framework to Profile Primary MicroRNAs from Single-cell RNA-sequencing datasets. Brief Bioinform 2022;23:bbac419. https://doi.org/10.1093/bib/bbac419.
[283] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32:381–6. https://doi.org/10.1038/nbt.2859.
[284] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016;44:e71. https://doi.org/10.1093/nar/gkv1507.
[285] Borgan Ø. Modeling Survival Data: Extending the Cox Model. Terry M. Therneau and Patricia M. Grambsch, Springer-Verlag, New York, 2000. No. of pages: xiii + 350. Price: $69.95. ISBN 0-387-98784-3. Statistics in Medicine 2001;20:2053–4. https://doi.org/10.1002/sim.956.
[286] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.
[287] Fu J, Li K, Zhang W, Wan C, Zhang J, Jiang P, et al. Large-scale public data reuse to model immunotherapy response and resistance. Genome Med 2020;12:21. https://doi.org/10.1186/s13073-020-0721-z.
[288] Ning S, Gao Y, Wang P, Li X, Zhi H, Zhang Y, et al. Construction of a lncRNA-mediated feed-forward loop network reveals global topological features and prognostic motifs in human cancers. Oncotarget 2016;7:45937–47. https://doi.org/10.18632/oncotarget.10004.
[289] Jiang L, Yu X, Ma X, Liu H, Zhou S, Zhou X, et al. Identification of transcription factor-miRNA-lncRNA feed-forward loops in breast cancer subtypes. Comput Biol Chem 2019;78:1–7. https://doi.org/10.1016/j.compbiolchem.2018.11.008.
[290] Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. Cell 2009;136:215–33. https://doi.org/10.1016/j.cell.2009.01.002.
[291] Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med 2014;20:460–9. https://doi.org/10.1016/j.molmed.2014.06.005.
[292] Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling data to identify human microRNA targets. Nat Methods 2007;4:1045–9. https://doi.org/10.1038/nmeth1130.
[293] Ruike Y, Ichimura A, Tsuchiya S, Shimizu K, Kunimoto R, Okuno Y, et al. Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines. J Hum Genet 2008;53:515–23. https://doi.org/10.1007/s10038-008-0279-x.

[294] Alevizos I, Illei GG. MicroRNAs as biomarkers in rheumatic diseases. Nat Rev Rheumatol 2010;6:391–8. https://doi.org/10.1038/nrrheum.2010.81.
[295] Reda El Sayed S, Cristante J, Guyon L, Denis J, Chabre O, Cherradi N. MicroRNA Therapeutics in Cancer: Current Advances and Challenges. Cancers 2021;13:2680. https://doi.org/10.3390/cancers13112680.
[296] Jin S, Zeng X, Fang J, Lin J, Chan SY, Erzurum SC, et al. A network-based approach to uncover microRNA-mediated disease comorbidities and potential pathobiological implications. Npj Syst Biol Appl 2019;5:1–11. https://doi.org/10.1038/s41540-019-0115-2.
[297] van Iterson M, Bervoets S, de Meijer EJ, Buermans HP, ’t Hoen PAC, Menezes RX, et al. Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. Nucleic Acids Research 2013;41:e146. https://doi.org/10.1093/nar/gkt525.
[298] Xuan J, Shi L, Guo L. microRNA Profiling: Strategies and Challenges. microRNAs in Toxicology and Medicine, John Wiley & Sons, Ltd; 2013, p. 437–55. https://doi.org/10.1002/9781118695999.ch25.
[299] Wright C, Rajpurohit A, Burke EE, Williams C, Collado-Torres L, Kimos M, et al. Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. BMC Genomics 2019;20:513. https://doi.org/10.1186/s12864-019-5870-3.
[300] Benesova S, Kubista M, Valihrach L. Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. Diagnostics 2021;11:964. https://doi.org/10.3390/diagnostics11060964.
[301] Matullo G, Naccarati A, Pardini B. MicroRNA expression profiling in bladder cancer: the challenge of next-generation sequencing in tissues and biofluids. International Journal of Cancer 2016;138:2334–45. https://doi.org/10.1002/ijc.29895.
[302] Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, et al. Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies. Anal Chem 2016;88:2088–95. https://doi.org/10.1021/acs.analchem.5b03376.
[303] Madadi S, Schwarzenbach H, Lorenzen J, Soleimani M. MicroRNA expression studies: challenge of selecting reliable reference controls for data normalization. Cell Mol Life Sci 2019;76:3497–514. https://doi.org/10.1007/s00018-019-03136-y.
[304] Schwarzenbach H, da Silva AM, Calin G, Pantel K. Data Normalization Strategies for MicroRNA Quantification. Clinical Chemistry 2015;61:1333–42. https://doi.org/10.1373/clinchem.2015.239459.
[305] Webber JW, Elias KM. Fast and robust imputation for miRNA expression data using constrained least squares. BMC Bioinformatics 2022;23:145. https://doi.org/10.1186/s12859-022-04656-4.
[306] Nielsen MM, Pedersen JS. miRNA activity inferred from single cell mRNA expression. Sci Rep 2021;11:9170. https://doi.org/10.1038/s41598-021-88480-5.
[307] Cheng C, Li LM. Inferring MicroRNA Activities by Combining Gene Expression with MicroRNA Target Prediction. PLOS ONE 2008;3:e1989. https://doi.org/10.1371/journal.pone.0001989.
[308] Le TD, Liu L, Tsykin A, Goodall GJ, Liu B, Sun B-Y, et al. Inferring microRNA–mRNA causal regulatory relationships from expression data. Bioinformatics 2013;29:765–71. https://doi.org/10.1093/bioinformatics/btt048.

[309] Tan H, Huang S, Zhang Z, Qian X, Sun P, Zhou X. Pan-cancer analysis on microRNA-associated gene activation. EBioMedicine 2019;43:82–97. https://doi.org/10.1016/j.ebiom.2019.03.082.
[310] Monteys AM, Spengler RM, Wan J, Tecedor L, Lennox KA, Xing Y, et al. Structure and activity of putative intronic miRNA promoters. RNA 2010;16:495–505. https://doi.org/10.1261/rna.1731910.
[311] Frouin A, Dandine-Roulland C, Pierre-Jean M, Deleuze J-F, Ambroise C, Le Floch E. Exploring the Link Between Additive Heritability and Prediction Accuracy From a Ridge Regression Perspective. Front Genet 2020;11. https://doi.org/10.3389/fgene.2020.581594.
[312] Novianti PW, Snoek BC, Wilting SM, van de Wiel MA. Better diagnostic signatures from RNAseq data through use of auxiliary co-data. Bioinformatics 2017;33:1572–4. https://doi.org/10.1093/bioinformatics/btw837.
[313] Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet 2021;53:1097–103. https://doi.org/10.1038/s41588-021-00870-7.
[314] Liu C, Wei D, Xiang J, Ren F, Huang L, Lang J, et al. An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. Molecular Therapy - Nucleic Acids 2020;21:676–86. https://doi.org/10.1016/j.omtn.2020.07.003.
[315] Cule E, De Iorio M. Ridge Regression in Prediction Problems: Automatic Choice of the Ridge Parameter. Genet Epidemiol 2013;37:704–14. https://doi.org/10.1002/gepi.21750.
[316] Zhang R, McDonald GC. Characterization of Ridge Trace Behavior. Communications in Statistics - Theory and Methods 2005;34:1487–501. https://doi.org/10.1081/STA-200063266.
[317] Yang Y, Xu Z, Song D. Missing value imputation for microRNA expression data by using a GO-based similarity measure. BMC Bioinformatics 2016;17:10. https://doi.org/10.1186/s12859-015-0853-0.
[318] Chistiakov DA, Orekhov AN, Bobryshev YV. Cardiac-specific miRNA in cardiogenesis, heart function, and cardiac pathology (with focus on myocardial infarction). Journal of Molecular and Cellular Cardiology 2016;94:107–21. https://doi.org/10.1016/j.yjmcc.2016.03.015.
[319] Navickas R, Gal D, Laucevičius A, Taparauskaitė A, Zdanytė M, Holvoet P. Identifying circulating microRNAs as biomarkers of cardiovascular disease: a systematic review. Cardiovascular Research 2016;111:322–37. https://doi.org/10.1093/cvr/cvw174.
[320] Kaur A, Mackin ST, Schlosser K, Wong FL, Elharram M, Delles C, et al. Systematic review of microRNA biomarkers in acute coronary syndrome and stable coronary artery disease. Cardiovascular Research 2020;116:1113–24. https://doi.org/10.1093/cvr/cvz302.
[321] Widera C, Gupta SK, Lorenzen JM, Bang C, Bauersachs J, Bethmann K, et al. Diagnostic and prognostic impact of six circulating microRNAs in acute coronary syndrome. Journal of Molecular and Cellular Cardiology 2011;51:872–5. https://doi.org/10.1016/j.yjmcc.2011.07.011.
[322] Small EM, Olson EN. Pervasive roles of microRNAs in cardiovascular biology. Nature 2011;469:336–42. https://doi.org/10.1038/nature09783.
[323] Kalozoumi G, Yacoub M, Sanoudou D. MicroRNAs in heart failure: Small molecules with major impact. Glob Cardiol Sci Pract 2014;2014:79–102. https://doi.org/10.5339/gcsp.2014.30.

[324] Sambandan S, Akbalik G, Kochen L, Rinne J, Kahlstatt J, Glock C, et al. Activity-dependent spatially localized miRNA maturation in neuronal dendrites. Science 2017;355:634–7. https://doi.org/10.1126/science.aaf8995.
[325] Liu X, Xie H, Liu W, Zuo J, Li S, Tian Y, et al. Dynamic regulation of alternative polyadenylation by PQBP1 during neurogenesis. Cell Reports 2024;43. https://doi.org/10.1016/j.celrep.2024.114525.
[326] Cihan M, Schmauck G, Sprang M, Andrade-Navarro MA. Unveiling cell-type-specific microRNA networks through alternative polyadenylation in glioblastoma. BMC Biology 2025;23:15. https://doi.org/10.1186/s12915-024-02104-8.
[327] Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T. Identification of Tissue-Specific MicroRNAs from Mouse. Current Biology 2002;12:735–9. https://doi.org/10.1016/S0960-9822(02)00809-6.
[328] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. Nucleic Acids Research 2016;44:3865–77. https://doi.org/10.1093/nar/gkw116.
[329] Zhao Y, Ransom JF, Li A, Vedantham V, Drehle M von, Muth AN, et al. Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2. Cell 2007;129:303–17. https://doi.org/10.1016/j.cell.2007.03.030.
[330] Yang B, Lin H, Xiao J, Lu Y, Luo X, Li B, et al. The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2. Nat Med 2007;13:486–91. https://doi.org/10.1038/nm1569.
[331] van Rooij E, Sutherland LB, Qi X, Richardson JA, Hill J, Olson EN. Control of Stress-Dependent Cardiac Growth and Gene Expression by a MicroRNA. Science 2007;316:575–9. https://doi.org/10.1126/science.1139089.
[332] Mahmoudi E, Cairns MJ. MiR-137: an important player in neural development and neoplastic transformation. Mol Psychiatry 2017;22:44–55. https://doi.org/10.1038/mp.2016.150.
[333] Sun J, Sun J, Ming G, Song H. Epigenetic regulation of neurogenesis in the adult mammalian brain. European Journal of Neuroscience 2011;33:1087–93. https://doi.org/10.1111/j.1460-9568.2011.07607.x.
[334] Chen F, Wang Y-C, Wang B, Kuo C-CJ. Graph representation learning: a survey. APSIPA Transactions on Signal and Information Processing 2020;9:e15. https://doi.org/10.1017/ATSIP.2020.13.
[335] Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Res 2024;52:D891–9. https://doi.org/10.1093/nar/gkad1049.
[336] Csárdi G, Nepusz T. The igraph software package for complex network research, 2006.
[337] Gargano MA, Matentzoglu N, Coleman B, Addo-Lartey EB, Anagnostopoulos AV, Anderton J, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. Nucleic Acids Res 2024;52:D1333–46. https://doi.org/10.1093/nar/gkad1005.
[338] Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Research 2023;51:W207–12. https://doi.org/10.1093/nar/gkad347.
[339] Bartel DP. Metazoan MicroRNAs. Cell 2018;173:20–51. https://doi.org/10.1016/j.cell.2018.03.006.

[340] Vaghf A, Khansarinejad B, Ghaznavi-Rad E, Mondanizadeh M. The role of microRNAs in diseases and related signaling pathways. Mol Biol Rep 2022;49:6789–801. https://doi.org/10.1007/s11033-021-06725-y.
[341] Kapplingattu SV, Bhattacharya S, Adlakha YK. MiRNAs as major players in brain health and disease: current knowledge and future perspectives. Cell Death Discov 2025;11:7. https://doi.org/10.1038/s41420-024-02283-x.
[342] Kimura M, Kothari S, Gohir W, Camargo JF, Husain S. MicroRNAs in infectious diseases: potential diagnostic biomarkers and therapeutic targets. Clinical Microbiology Reviews 2023;36:e00015-23. https://doi.org/10.1128/cmr.00015-23.
[343] Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer 2006;6:857–66. https://doi.org/10.1038/nrc1997.
[344] Lu J, Clark AG. Impact of microRNA regulation on variation in human gene expression. Genome Res 2012;22:1243–54. https://doi.org/10.1101/gr.132514.111.
[345] Bofill-De Ros X, Vang Ørom UA. Recent progress in miRNA biogenesis and decay. RNA Biol n.d.;21:1–8. https://doi.org/10.1080/15476286.2023.2288741.
[346] 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;56:104–15. https://doi.org/10.1016/j.molcel.2014.08.028.
[347] Ameres SL, Horwich MD, Hung J-H, Xu J, Ghildiyal M, Weng Z, et al. Target RNA-directed trimming and tailing of small silencing RNAs. Science 2010;328:1534–9. https://doi.org/10.1126/science.1187058.
[348] Diener C, Keller A, Meese E. The miRNA–target interactions: An underestimated intricacy. Nucleic Acids Research 2024;52:1544–57. https://doi.org/10.1093/nar/gkad1142.
[349] Bosson AD, Zamudio JR, Sharp PA. Endogenous miRNA and Target Concentrations Determine Susceptibility to Potential ceRNA Competition. Mol Cell 2014;56:347–59. https://doi.org/10.1016/j.molcel.2014.09.018.
[350] Cihan M, Anyaegbunam UA, Albrecht S, Andrade-Navarro MA, Sprang M. Evaluating Genetic Regulators of MicroRNAs Using Machine Learning Models. International Journal of Molecular Sciences 2025;26:5757. https://doi.org/10.3390/ijms26125757.
[351] Denzler R, McGeary SE, Title AC, Agarwal V, Bartel DP, Stoffel M. Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. Mol Cell 2016;64:565–79. https://doi.org/10.1016/j.molcel.2016.09.027.
[352] Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Mol Cell 2014;54:766–76. https://doi.org/10.1016/j.molcel.2014.03.045.
[353] Marini F, Scherzinger D, Danckwardt S. TREND-DB—a transcriptome-wide atlas of the dynamic landscape of alternative polyadenylation. Nucleic Acids Res 2021;49:D243–53. https://doi.org/10.1093/nar/gkaa722.
[354] Boutet SC, Cheung TH, Quach NL, Liu L, Prescott SL, Edalati A, et al. Alternative polyadenylation mediates microRNA regulation of muscle stem cell function. Cell Stem Cell 2012;10:327–36. https://doi.org/10.1016/j.stem.2012.01.017.
[355] 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;320:1643–7. https://doi.org/10.1126/science.1155390.

[356] Feng X, Li L, Wagner EJ, Li W. TC3A: The Cancer 3’ UTR Atlas. Nucleic Acids Res 2018;46:D1027–30. https://doi.org/10.1093/nar/gkx892.
[357] Bioinformatics Pipeline: mRNA Analysis - GDC Docs n.d. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/ (accessed August 6, 2025).
[358] Li L, Huang K-L, Gao Y, Cui Y, Wang G, Elrod ND, et al. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. Nat Genet 2021;53:994–1005. https://doi.org/10.1038/s41588-021-00864-5.
[359] Scarborough AM, Flaherty JN, Hunter OV, Liu K, Kumar A, Xing C, et al. SAM homeostasis is regulated by CFIm-mediated splicing of MAT2A. Elife 2021;10:e64930. https://doi.org/10.7554/eLife.64930.
[360] Sim DY, Lee H-J, Ahn C-H, Park J, Park S-Y, Kil B-J, et al. Negative Regulation of CPSF6 Suppresses the Warburg Effect and Angiogenesis Leading to Tumor Progression Via c-Myc Signaling Network: Potential Therapeutic Target for Liver Cancer Therapy. International Journal of Biological Sciences 2024;20:3442–60. https://doi.org/10.7150/ijbs.93462.
[361] Li T, Li L, Hiers NM, Sheng P, Wang Y, Traugot CM, et al. Translation suppresses exogenous target RNA-mediated microRNA decay. Nat Commun 2025;16:5257. https://doi.org/10.1038/s41467-025-60374-4.
[362] Kleaveland B, Shi CY, Stefano J, Bartel DP. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. Cell 2018;174:350-362.e17. https://doi.org/10.1016/j.cell.2018.05.022.
[363] Mayr C. Regulation by 3’-Untranslated Regions. Annu Rev Genet 2017;51:171–94. https://doi.org/10.1146/annurev-genet-120116-024704.
[364] Smibert P, Yang J-S, Azzam G, Liu J-L, Lai EC. Homeostatic control of Argonaute stability by microRNA availability. Nat Struct Mol Biol 2013;20:789–95. https://doi.org/10.1038/nsmb.2606.
[365] Broderick JA, Salomon WE, Ryder SP, Aronin N, Zamore PD. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. RNA 2011;17:1858–69. https://doi.org/10.1261/rna.2778911.
[366] Hiers NM, Li L, Li T, Sheng P, Wang Y, Traugot CM, et al. An endogenous cluster of target-directed microRNA degradation sites induces decay of distinct microRNA families. Cell Reports 2025;44:116162. https://doi.org/10.1016/j.celrep.2025.116162.
[367] Engreitz JM, Lawson HA, Singh H, Starita LM, Hon GC, Carter H, et al. Deciphering the impact of genomic variation on function. Nature 2024;633:47–57. https://doi.org/10.1038/s41586-024-07510-0.
[368] Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature 2017;550:204–13. https://doi.org/10.1038/nature24277.
[369] Peña-Martínez EG, Rodríguez-Martínez JA. Decoding Non-coding Variants: Recent Approaches to Studying Their Role in Gene Regulation and Human Diseases. Front Biosci (Schol Ed) 2024;16:4. https://doi.org/10.31083/j.fbs1601004.
[370] Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. RNA 2003;9:277–9. https://doi.org/10.1261/rna.2183803.
[371] Ludwig N, Becker M, Schumann T, Speer T, Fehlmann T, Keller A, et al. Bias in recent miRBase annotations potentially associated with RNA quality issues. Sci Rep 2017;7:5162. https://doi.org/10.1038/s41598-017-05070-0.

[372] Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, et al. A Uniform System For The Annotation Of Human microRNA Genes And The Evolution Of The Human microRNAome. Annu Rev Genet 2015;49:213–42. https://doi.org/10.1146/annurev-genet-120213-092023.
[373] Liu C, Rennie WA, Carmack CS, Kanoria S, Cheng J, Lu J, et al. Effects of genetic variations on microRNA: target interactions. Nucleic Acids Research 2014;42:9543–52. https://doi.org/10.1093/nar/gku675.
[374] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature 2024;625:92–100. https://doi.org/10.1038/s41586-023-06045-0.
[375] dbGaP Study n.d. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v2.p1 (accessed December 16, 2024).
[376] Aparicio-Puerta E, Hirsch P, Schmartz GP, Fehlmann T, Keller V, Engel A, et al. isomiRdb: microRNA expression at isoform resolution. Nucleic Acids Res 2023;51:D179–85. https://doi.org/10.1093/nar/gkac884.
[377] The Genotype-Tissue Expression (GTEx) project | Nature Genetics n.d. https://www.nature.com/articles/ng.2653 (accessed August 26, 2025).
[378] Skoufos G, Kakoulidis P, Tastsoglou S, Zacharopoulou E, Kotsira V, Miliotis M, et al. TarBase-v9.0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs. Nucleic Acids Res 2024;52:D304–10. https://doi.org/10.1093/nar/gkad1071.
[379] Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acids Research 2025;53:D948–57. https://doi.org/10.1093/nar/gkae1071.
[380] Kern F, Fehlmann T, Solomon J, Schwed L, Grammes N, Backes C, et al. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. Nucleic Acids Research 2020;48:W521–8. https://doi.org/10.1093/nar/gkaa309.
[381] Rezwani M, Pourfathollah AA, Noorbakhsh F. rbioapi: user-friendly R interface to biologic web services' API. Bioinformatics 2022;38:2952–3. https://doi.org/10.1093/bioinformatics/btac172.
[382] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:D980-985. https://doi.org/10.1093/nar/gkt1113.
[383] Winter J, Jung S, Keller S, Gregory RI, Diederichs S. Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol 2009;11:228–34. https://doi.org/10.1038/ncb0309-228.
[384] An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65. https://doi.org/10.1038/nature11632.
[385] Hill CG, Jabbari N, Matyunina LV, McDonald JF. Functional and evolutionary significance of human microRNA seed region mutations. PLoS One 2014;9:e115241. https://doi.org/10.1371/journal.pone.0115241.
[386] Chipman LB, Pasquinelli AE. MiRNA Targeting – Growing Beyond the Seed. Trends Genet 2019;35:215–22. https://doi.org/10.1016/j.tig.2018.12.005.
[387] Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. Proceedings of the National Academy of Sciences 2015;112:E1106–15. https://doi.org/10.1073/pnas.1420955112.
[388] Liu B, Li J, Cairns MJ. Identifying miRNAs, targets and functions. Briefings in Bioinformatics 2014;15:1–19. https://doi.org/10.1093/bib/bbs075.

[389] Han J, Lee Y, Yeom K-H, Nam J-W, Heo I, Rhee J-K, et al. Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. Cell 2006;125:887–901. https://doi.org/10.1016/j.cell.2006.03.043.
[390] Omariba G, Xu F, Wang M, Li K, Zhou Y, Xiao J. Genome-Wide Analysis of MicroRNA-related Single Nucleotide Polymorphisms (SNPs) in Mouse Genome. Sci Rep 2020;10:5789. https://doi.org/10.1038/s41598-020-62588-6.
[391] Quarles KA, Sahu D, Havens MA, Forsyth ER, Wostenberg C, Hastings ML, et al. Ensemble Analysis of Primary miRNA Structure Reveals an Extensive Capacity to Deform near the Drosha Cleavage Site. Biochemistry 2013;52:795–807. https://doi.org/10.1021/bi301452a.
[392] Campbell MC, Tishkoff SA. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu Rev Genomics Hum Genet 2008;9:403–33. https://doi.org/10.1146/annurev.genom.9.081307.164258.
[393] Atkinson EG, Artomov M, Loboda AA, Rehm HL, MacArthur DG, Karczewski KJ, et al. Discordant calls across genotype discovery approaches elucidate variants with systematic errors. Genome Res 2023;33:999–1005. https://doi.org/10.1101/gr.277908.123.
[394] Gabele A, Sprang M, Cihan M, Welzel M, Nurbekova A, Romaniuk K, et al. Unveiling IRF4-steered regulation of context-dependent effector programs in CD4+ T cells under Th17- and Treg-skewing conditions. Cell Reports 2025;44. https://doi.org/10.1016/j.celrep.2025.115407.
[395] Gabele A, Cihan M, Sprang M, Klein M, Nurbekova A, Romaniuk K, et al. Protocol for mapping murine transcription factor interactomes and composite motifs combining affinity purification mass spectrometry and ChIP-seq. STAR Protoc 2025;6:104184. https://doi.org/10.1016/j.xpro.2025.104184.
[396] Maji RK, Leisegang MS, Boon RA, Schulz MH. Revealing microRNA regulation in single cells. Trends in Genetics 2025;41:522–36. https://doi.org/10.1016/j.tig.2024.12.009.
[397] Yang T, Wang Y, He Y. TEC-miTarget: enhancing microRNA target prediction based on deep learning of ribonucleic acid sequences. BMC Bioinformatics 2024;25:159. https://doi.org/10.1186/s12859-024-05780-z.