NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES

DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES "DATA SCIENCE AND INFORMATION TECHNOLOGIES"

MASTER THESIS

Abundance and regulatory roles of tRNA-derived fragments: a computational exploration using Next-Generation Sequencing data

Vasileios A. Maroulis

Supervisor: Theodore Dalamagas, Research Director, IMSI ATHENA Research Center

ATHENS

JULY 2023

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αφθονία και ρυθμιστικοί ρόλοι των παραγόμενων από tRNA θραυσμάτων: μια υπολογιστική εξερεύνηση με χρήση δεδομένων από Αλληλούχηση Επόμενης Γενιάς

Βασίλειος Α. Μαρούλης

Επιβλέπων καθηγητής: Θεόδωρος Δαλαμάγκας, Διευθυντής Ερευνών, ΙΠΣΥ Ερευνητικό Κέντρο "Αθηνά"

ΑΘΗΝΑ

ΙΟΥΛΙΟΣ 2023

MASTER THESIS

Abundance and regulatory roles of tRNA-derived fragments: a computational exploration using Next-Generation Sequencing data

Vasileios A. Maroulis

ID: DS2180011

SUPERVISOR:

Theodore Dalamagas, Research Director, IMSI, ATHENA Research Center

EXAMINATION COMMITTEE

Theodore Dalamagas, Research Director, IMSI, ATHENA Research Center

Artemis Hatzigeorgiou, Professor, University of Thessaly/ Hellenic Pasteur Institute

Spyrıdon Tastsoglou, Postdoctoral Researcher, University of Thessaly/ Hellenic Pasteur Institute

July 2023

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αφθονία και ρυθμιστικοί ρόλοι των παραγόμενων από tRNA θραυσμάτων: μια υπολογιστική εξερεύνηση με χρήση δεδομένων από Αλληλούχηση Επόμενης Γενιάς

Βασίλειος Α. Μαρούλης

ΑΜ: DS2180011

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Θεόδωρος Δαλαμάγκας, Διευθυντής Ερευνών, ΙΠΣΥ, Ερευνητικό Κέντρο "Αθηνά"

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Θεόδωρος Δαλαμάγκας, Διευθυντής Ερευνών, ΙΠΣΥ, Ερευνητικό Κέντρο "Αθηνά"

Άρτεμις Χατζηγεωργίου, Καθηγήτρια, Πανεπιστήμιο Θεσσαλίας/Ελληνικό Ινστιτούτο Παστέρ

Σπυρίδων Ταστσόγλου, Μεταδιδακτορικός Ερευνητής, Πανεπιστήμιο Θεσσαλίας/Ελληνικό Ινστιτούτο Παστέρ

Ιούλιος 2023

ABSTRACT

The primary and most well-studied function of transfer RNAs (tRNAs) is their involvement in protein translation, where they carry amino acid residues into the ribosomes and participate in the codon-based sequential elongation of the nascent peptide chain. Recently, tRNAs were shown to act as a pool for the biogenesis of numerous short RNAs of regulatory potential. tRNA-derived fragments (tRFs) vary in their size and originate from different parts of the tRNA cloverleaf-shaped molecule. The main tRF species are ~34nt 5' and 3' halves of tRNAs (tRHs) and the smaller (18-22nt) 5' and 3' tRFs. Increasing evidence indicates that small tRFs are loaded into AGO proteins and guide RISC-dependent post-transcriptional repression, in the same manner that microRNAs do. In this thesis, tRF species were quantified from human small RNA-Seq (sRNA-Seq) datasets using Manatee tool. Manatee combines information on uniquely aligned read clusters and known annotation of transcriptomic features to guide the efficient placement of multi-mapping reads. The highly abundant tRFs that were identified were subsequently utilized to guide the analysis of publicly available AGO-Cross-Linking Immunoprecipitation (AGO-CLIP) sequencing experiments in relevant cell-lines using microCLIP tool. Directly identified tRF-mRNA interactions were functionally interrogated using pathway enrichment statistics. The results demonstrate the use of Manatee for the quantification of tRFs for the first time and the limitations of such an approach, and the subsequent characterization of tRF implication in gene expression regulation.

SUBJECT AREA: Bioinformatics

KEYWORDS: tRNA-derived fragments, microRNAs, AGO protein, small RNA-Seq, AGO-Cross-Linking Immunoprecipitation**,** gene expression regulation

ΠΕΡΙΛΗΨΗ

Η πρωταρχική και πιο καλά μελετημένη λειτουργία των μεταφορικών RNA (tRNAs) είναι η συμμετοχή τους στην μετάφραση των πρωτεϊνών, όπου μεταφέρουν αμινοξέα στα ριβοσώματα και συμμετέχουν στη βασισμένη στα κωδικόνια διαδοχική επιμήκυνση της νεοσυντιθέμενης πολυπεπτιδικής αλυσίδας. Πρόσφατα δείχθηκε ότι τα tRNAs αποτελούν τη δεξαμενή για τη βιογένεση πολυάριθμων RNAs μικρού μήκους με ρυθμιστικές δυνατότητες. Τα παραγόμενα από tRNA θραύσματα (tRFs) ποικίλουν ως προς το μέγεθος τους και προέρχονται από διαφορετικά σημεία του μορίου του tRNA, το οποίο έχει σχήμα τριφυλλιού. Τα κύρια είδη των tRFs είναι τα «μισά» των tRNAs (tRHs) με μέγεθος ~34 νουκλεοτίδια και τα μικρότερα (18-22 νουκλεοτίδια) 3' και 5' tRFs. Ένας αυξανόμενος όγκος δεδομένων υποδεικνύει ότι μικρά tRFs φορτώνονται σε πρωτεΐνες της οικογένειας AGO και καθοδηγούν την εξαρτώμενη από το σύμπλοκο RISC μετα-μεταγραφική καταστολή της γονιδιακής έκφρασης, με τον ίδιο τρόπο με οποίο δρουν τα microRNAs. Σε αυτή τη διπλωματική εργασία έγινε ποσοτικοποίηση των διαφορετικών ειδών tRFs από ανθρώπινα δεδομένα αλληλούχησης μικρών RNA (sRNA-Seq) με τη χρήση του εργαλείου Manatee. Το Manatee συνδυάζει πληροφορίες από συστάδες μοναδικά χαρτογραφημένων διαβασμάτων και τον ήδη γνωστό σχολιασμό των μεταγραφικών χαρακτηριστικών για να καθοδηγήσει την αποτελεσματική τοποθέτηση των διαβασμάτων που χαρτογραφούνται σε πολλαπλές γενωμικές θέσεις. Τα tRFs με την μεγαλύτερη αφθονία που ταυτοποιήθηκαν χρησιμοποιήθηκαν στη συνέχεια για να καθοδηγήσουν την ανάλυση δημόσια διαθέσιμων πειραμάτων αλληλούχησης διασταυρούμενης σύνδεσης και ανοσοκαθίζησης πρωτεϊνών AGO (AGO-Cross-Linking Immunoprecipitation, AGO-CLIP) σε αντίστοιχες ανθρώπινες κυτταρικές σειρές, με τη χρήση του εργαλείου microCLIP. Απευθείας αλληλεπιδράσεις μεταξύ των tRFs και mRNAs που ταυτοποιήθηκαν, διερευνήθηκαν λειτουργικά με στατιστική ανάλυση εμπλουτισμού μονοπατιών (pathway enrichment analysis). Τα αποτελέσματα επιδεικνύουν τη χρήση για πρώτη φορά του εργαλείου Manatee στην ποσοτικοποίηση των tRFs, τους περιορισμούς μιας τέτοιας προσέγγισης και τον επακόλουθο χαρακτηρισμό της εμπλοκής των tRFs στη ρύθμιση της γονιδιακής έκφρασης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: θραύσματα παραγόμενα από tRNA**,** μικρόRNAs, πρωτεΐνες οικογένειας AGO**,** αλληλούχηση μικρών RNA**,** διασταυρούμενη σύνδεση και ανοσοκαθίζηση πρωτεϊνών AGO**,** ρύθμιση γονιδιακής έκφρασης

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the members of the examination committee, Professor Theodore Dalamagas, Professor Artemis Hatzigeorgiou and Postdoctoral Researcher Spyrıdon Tastsoglou for their support and guidance throughout the course of this thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The current MSc thesis was implemented in the context of the program of postgraduate studies Data Science and Information Technologies, specialization in Bioinformatics and Biomedical Data Science, of the Department of Informatics and Telecommunications of National and Kapodistrian University of Athens.

1.INTRODUCTION

1.1 Noncoding RNAs and their emergence as regulatory molecules

The Central Dogma of Molecular Biology traditionally placed RNA as an intermediate for the transfer of genetic information from the DNA sequence of a gene to the protein that it encodes. A growing number of exceptions to this rule has been reported over the last decades. The term noncoding RNA (ncRNA) refers to RNA molecules that do not encode for a protein. Until recently, most of the known ncRNAs fulfilled generic cellular functions, with important structural and catalytic roles for gene expression: Small nuclear RNAs (snRNAs) are involved in splicing of messenger RNAS (mRNAs); transfer RNAs (tRNAs) decode the mRNA sequence into protein; ribosomal RNAs (rRNAs) are components of the ribosomes which are ribonucleoprotein complexes macromolecular structures essential for translation; small nucleolar RNAs (snoRNAs) are involved in the modification of rRNAs [1]. Our understanding of ncRNAs completely changed with the discovery in the 1990s that small ncRNAs could act as regulatory molecules, mediating post-transcriptional gene silencing of complementary mRNAs [2,3].

Regulation of gene expression is a fundamental process, important for the development, homeostasis, and adaptation of all living cells [4]. It requires high fidelity and precise control and can occur at several levels by multiple mechanisms. Transcription is considered the primary regulatory point in gene expression and has received the most attention, however, post-transcriptional regulation adds extra levels of control and complexity, and there seems to be considerable coordination and interdependence between those two control points. Over the past few decades ncRNAs emerged as crucial gene expression regulators. A large body of studies revealed the abundance of ncRNAs and the existence of numerous distinct mechanisms of regulation in all three domains of life (archaea, bacteria, and eukaryotes), and it is likely that their diversity, functions, and underlying mechanisms are still underestimated [5,6].

1.2 Classes of regulatory noncoding RNAs

1.2.1 Long noncoding RNAs

Long noncoding RNAs (lncRNAs) are a diverse group of regulatory noncoding RNAs arbitrarily considered to have a minimum size of 200nt [7]. lncRNAs represent a significant portion of the mammalian transcriptome, with an estimated abundance of more than 5900 transcripts in humans [8]. Initially lncRNAs were thought to be transcriptional noise and non-functional, despite the evidence suggesting that many of them exhibited developmentally regulated expression and specific subcellular localization. The assertion that lncRNAs are not functional seemed to be supported by their low sequence conservation, although many examples of conserved lncRNAs exist and, in contrast to protein-coding genes, lncRNAs require to maintain high conservation over short regions of their length in order to preserve their function. Even though only a small fraction of lncRNAs have been mechanistically characterized to date, it is evident that they fulfil a broad range of functional roles, including the maintenance of nuclear architecture and the regulation of gene expression.

Regarding their functions, lncRNAs can mediate epigenetic changes by acting as guide molecules for the recruitment of chromatin remodelling complexes to target genomic loci, resulting to transcriptional silencing. A number of lncRNAs associate with enhancers and promoters, modulating the binding of transcription factors or pre-initiation complex to these elements, thus inducing or repressing the transcription of target genes. Long ncRNAs can also affect post-transcriptional processing of mRNAs, either indirectly via targeting protein complexes to complementary mRNAs that promote stabilization or degradation, or directly, interacting with mRNAs to modulate their splicing and translation. Several lncRNAs can act as competing endogenous RNAs (ceRNAs) or miRNA sponges, containing binding sites for one or multiple miRNAs and sequestering them away from their canonical targets, thus fine-tuning gene expression. Many lncRNAs are not directly connected to gene regulation, but rather act as scaffolds, forming structures together with proteins that maintain and modulate nuclear architecture [9,10].

1.2.2 PIWI-interacting RNAs

PIWI-interacting RNAs (piRNAs) are small ncRNAs of 21-35 nucleotides in length, with extremely diverse and rarely conserved sequences. They are transcribed as long single-stranded precursor transcripts from one or both DNA strands on genomic loci called piRNA clusters via canonical or non-canonical transcription. These precursors are processed in a Dicer-independent manner that involves endonucleolytic cleavage, trimming and 2'-O-methylation of the 3' end, to produce multiple and diverse mature piRNAs from each precursor. Mature piRNAs possess 5´ monophosphate and 2´-O-methyl 3´ ends and are loaded to the PIWI-clade of Argonaute proteins to guide the silencing of complementary transcripts.

Regarding the functions of piRNAs, the ancestral and main known function is the silencing of transposable and repetitive elements in the germline towards maintaining genomic stability. In some invertebrates, viral RNAs can enter the piRNA pathway and produce piRNAs that can be used to fight viral infection. There is also evidence that piRNAs may have a role in regulating gene expression, by guiding the PIWI-dependent cleavage of target mRNAs during meiosis and spermatogenesis [11].

1.2.3 Small interfering RNAs (siRNAs)

Small interfering RNAs (siRNAs) are double-stranded RNA molecules of 21-25 nucleotides in length that are derived from longer double-stranded RNA precursors or from precursors harboring long hairpins that may be produced endogenously or may be supplied exogenously. siRNAs are produced in a similar way and have similar mechanisms of action as microRNAs (miRNAs). They are processed from their precursors by the RNase III endonuclease Dicer (but do not require Drosha) as duplexes with a 2 nucleotide 3' overhang on each strand. These duplexes are loaded to the RISC ribonucleoprotein complex that contains a member of the Argonaute family of proteins. Usually only one strand of the duplex is retained into the RISC complex and guides the recognition and incorporation of complementary mRNAs into the RISC complex that leads to translational repression by a yet unknown mechanism in case of an imperfect match, or to cleavage of target mRNAs in case of a near perfect match.

siRNAs mainly appear to act as an antiviral defence in plants and flies, where during infection viral dsRNAs produce siRNAs that target the complementary viral mRNAs, and a similar mechanism has been suggested in mammals. They also seem to have a role in silencing transposable elements in the mammalian female germline, contributing to the maintenance of genomic stability [12].

1.2.4 microRNAS (miRNAs)

1.2.4.1 Introduction

microRNAs (miRNAs) are an abundant class of small RNAs in most somatic tissues, found in plants, animals and some viruses. They are endogenous, evolutionarily conserved RNAs, of approximately 22 nucleotides in length and regulate gene expression at the post-transcriptional level by guiding translational repression or degradation of target mRNAs. The first miRNA (lin-4) was discovered in *C. elegans* in 1993 and the first human miRNA (let-7) was discovered in 2000. Currently, the latest version of the reference miRNA database miRBase (release 22.1) contains entries for 38,589 hairpin precursors and 48,860 mature microRNAs from 271 organisms, including 2,654 human mature miRNAs [13,14,15].

miRNA genes are one of the most abundant gene families, with multiple loci with similar sequences that arose from gene duplication existing in many species. The majority of miRNAs are located within introns or exons of noncoding genes, or within introns of coding genes. Intergenic miRNAs also exist, and often several miRNA loci in proximity are transcribed as a single polycistronic unit. Intergenic miRNAs have their individual promoters, while other miRNAs share the promoters of their host genes. Most miRNA genes are transcribed by RNA Pol II as long primary miRNAs (pri-miRNAs) that possess a 5'-cap but not always a 3' end polyadenylation signal, and typically contain one or more stem-loop structures and single stranded 5' and 3' sides.

Inside the nucleus, pri-miRNAs are processed by a complex called Microprocessor, consisting of the RNase III Drosha and the RNA binding protein DGCR8, that cleaves the stem-loop structure to release hairpin-shaped RNAs of ~70 nucleotides in length called pre-miRNAs. pre-miRNAs are subsequently exported into the cytoplasm by the transport complex formed by exportin 5 (XPO5) and the GTP-binding protein RAN (14). In the cytoplasm, pre-miRNAs are cleaved near the loop by the RNase III Dicer which in humans is associated with the RNA binding protein TRBP, to produce small RNA duplexes with 2-3 nucleotide overhangs at their 3' ends. This maturation procedure mediated by Drosha and Dicer represents the canonical biogenesis pathway of miRNAs, however alternative non-canonical pathways have been described that can be Drosha-independent or Dicer-independent and generate miRNAs and miRNA-like small RNAs. A well-recognized example of non-canonical miRNA biogenesis are miRNAs located within introns of coding genes, where after splicing of the host gene mRNA, the lariat structure of the intron is debranched and forms a hairpin structure resembling a pre-miRNA which is subsequently exported into the cytoplasm and continues into the canonical biogenesis pathway to be processed by Dicer. In a similar, Drosha-independent manner some miRNA-like small RNAs may be produced from other noncoding RNAs, such as tRNAs, tRNA-like precursors and snoRNAs [16]. An overview schematic of miRNA biogenesis is provided in **Figure 1**.

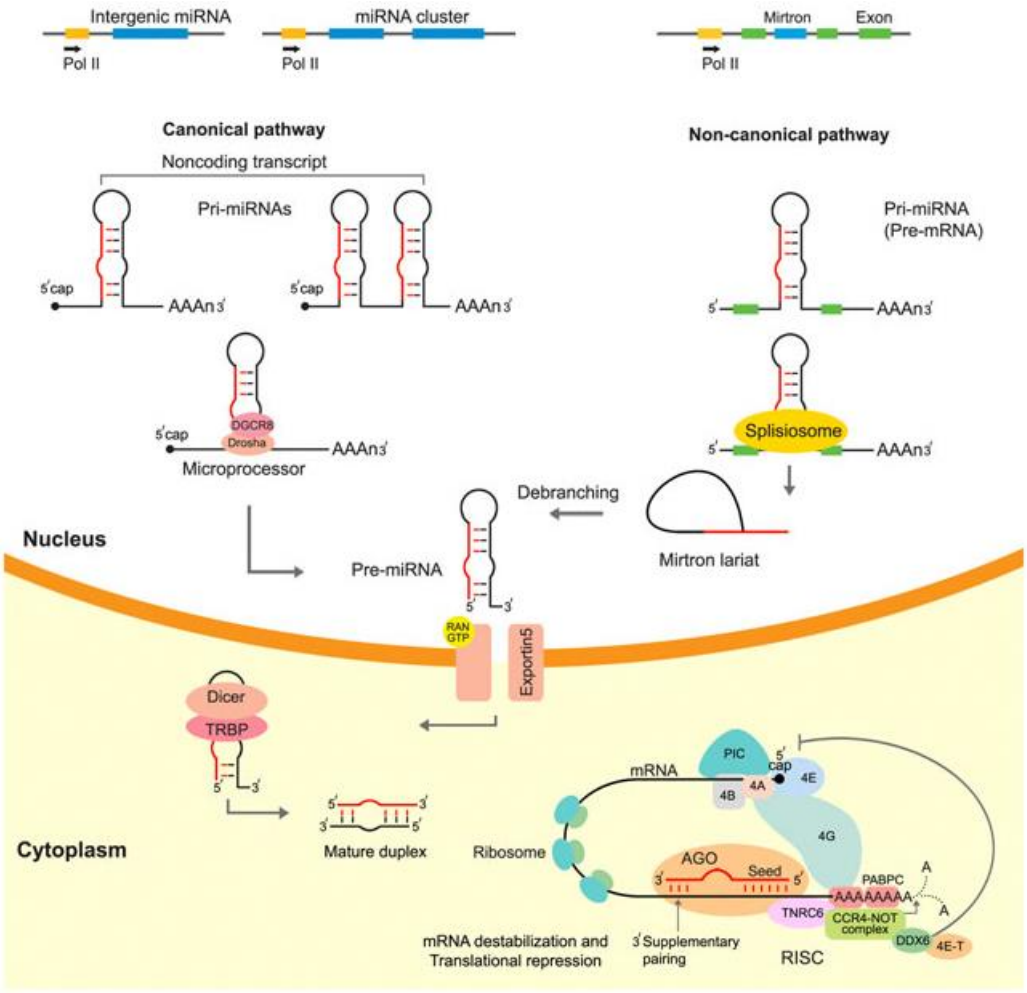

Figure 1. Overview of miRNA biogenesis.In the canonical pathway the pri-miRNAs transcribed from miRNA genes are cleaved into pre-miRNAs by the Microprocessor complex consisting of Drosha and DGCR8 in the nucleus. Subsequently they are exported to the cytoplasm by Exportin 5 and RAN, where they are cleaved by Dicer and form the RISC complex that mediates the silencing of target mRNAs. In the non-canonical pathway, the lariat structure of the intron forms a structure resembling a pre-miRNA which is subsequently exported into the cytoplasm and continues into the canonical pathway. (Figure reprinted by Saliminejad et al., 2019).

1.2.4.2 miRNA loading, RISC assembly and target mRNA recognition

The small RNA duplexes produced by Dicer are subsequently loaded into a member of the AGO protein family to form the ribonucleoprotein complex known as RNA-induced silencing complex (RISC) and guide the silencing of their complementary target mRNAs through cleavage, translational repression, and/or degradation [17]. Loading of the RNA duplex is mediated by chaperones and one of the two strands is selected as the passenger strand and is ejected from the AGO protein, while the other strand becomes the guide strand and is retained, a selection that depends on the thermodynamic stability of the 5'-end as well as the identity of the 5'-end nucleotide. The guide strand is divided into four functional domains, **(i)** the seed, **(ii)** central, **(iii)** supplementary and **(iv)** tail regions (**Figure 2A**). The seed region, which includes nucleotides g2-g8 from the 5'-end, is critical for target mRNA recognition through base pairing. In target mRNAs, miRNA-binding sites with perfect complementarity to nucleotides g2-g8 as well as an adenine at residue t1 (**Figure 2B**), collectively termed the 8-mer site, present the highest affinity for RISC. Binding sites with complementarity to g2–g8 (7-mer-m8), g2–g7 plus t1A (7- mer-A1), g2–g7 (6-mer), and g3–g8 (Offset 6-mer) are also considered canonical binding sites, despite their lower silencing efficacy. Non-canonical sites contain mismatches in the seed region, require additional base pairing in the supplementary region and have

much lower affinity. Most commonly, miRNA-binding sites are located in the 3'-UTR of target mRNAs, but binding sites in the 5'-UTR or open reading frame (ORF) also exist, although with lower efficiency [18].

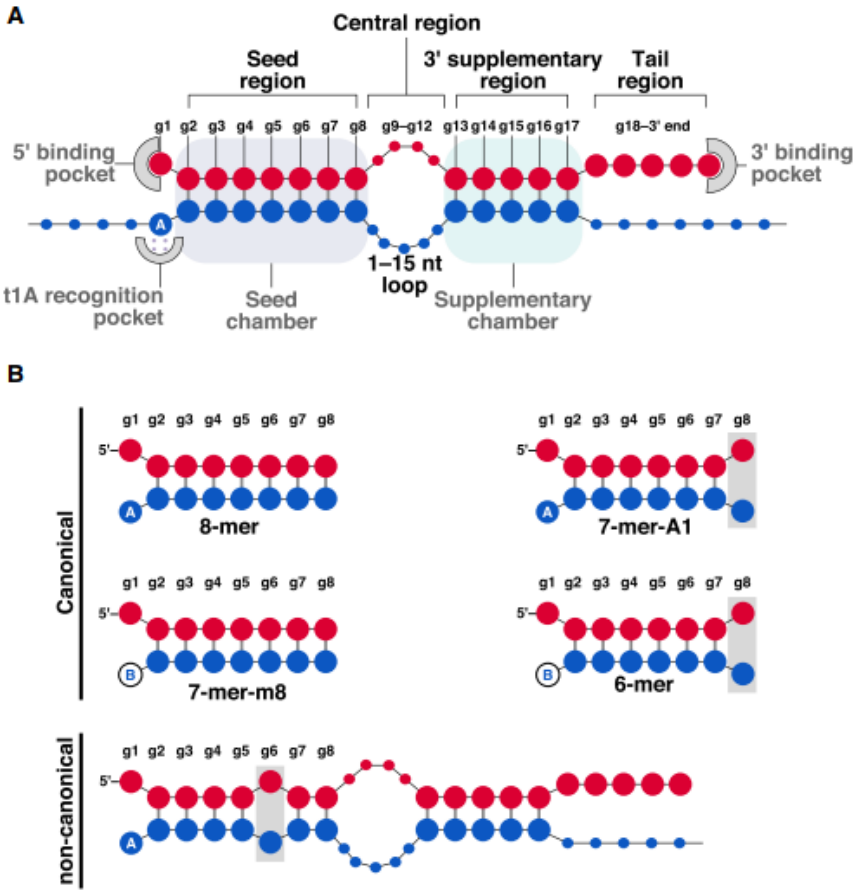

Figure 2. Target mRNA recognition by RISC in animals.(A) Base pairing between the guide strand of miRNA (red) and the target mRNA (blue). The guide strand is divided into 4 regions, the seed (g2–g8), central (g9–g12), 3' supplementary (g13–g17), and tail (g18–3'end). The seed region is important for target recognition, the central region is important for cleavage of target mRNA, the 3' supplementary region stabilizes target mRNA binding and the tail region regulates RISC function. (B) Canonical and non-canonical target sites in animals. The non-canonical binding sites have mismatches in the seed region and require additional base pairing in the 3' supplementary region. (Figure reprinted by Iwakawa1 and Tomari, 2022)

1.2.4.3 Target mRNA cleavage, translational repression, and/or degradation by RISC

Target mRNA cleavage requires extensive complementarity, with base pairing in the central region in addition to the seed region of miRNAs, and is catalysed by the PIWI domain of the AGO protein, although not all AGO proteins have this activity. Cleavage is the main mechanism of action for plant miRNAs, but not for animal miRNAs, where it is required for the silencing of only a few mRNAs.

The main mechanism of action for animal miRNAs is translational repression and mRNA degradation, although the molecular details and the order in which these events co-operate remain unclear and controversial [18]. The current consensus is that translational repression occurs first, followed by mRNA degradation, although there is evidence that translational repression may be independent from degradation and have an important role by itself [19]. The GW182 protein (TNRC6 in mammals) plays a key role in mRNA

degradation that includes deadenylation, decapping, and exonucleolytic degradation **(Figure 3A)**. This protein has glycine-tryptophan (GW) repeats that interact with the PIWI domain of AGOs and acts as a scaffold, promoting the removal of the poly(A)-binding protein (PABP) from the poly(A) tail of the mRNA and the recruitment of the deadenylation complexes CCR4-NOT and PAN2-PAN3. The CCR4-NOT complex in turn acts as a scaffold for the recruitment of decapping factors and activators (DCP1, DCP2, DDX6). Finally, the deadenylated and decapped mRNA is degraded by the exoribonuclease XRN1. The mechanisms of translational repression are less understood (**Figure 3B**). PABP interacts with eIF4G to form a closed-loop structure that facilitates translation initiation, and it has been proposed that PABP displacement by RISC abolishes this structure and inhibits translation initiation. Another proposed mechanism is the recruitment of translational inhibitors by RISC, that target eIF4E/4G, like DDX6. There is also evidence suggesting that translational repression may occur through removal of eIF4F components, and more specifically of eIF4A, a process that can be GW182-independent [18].

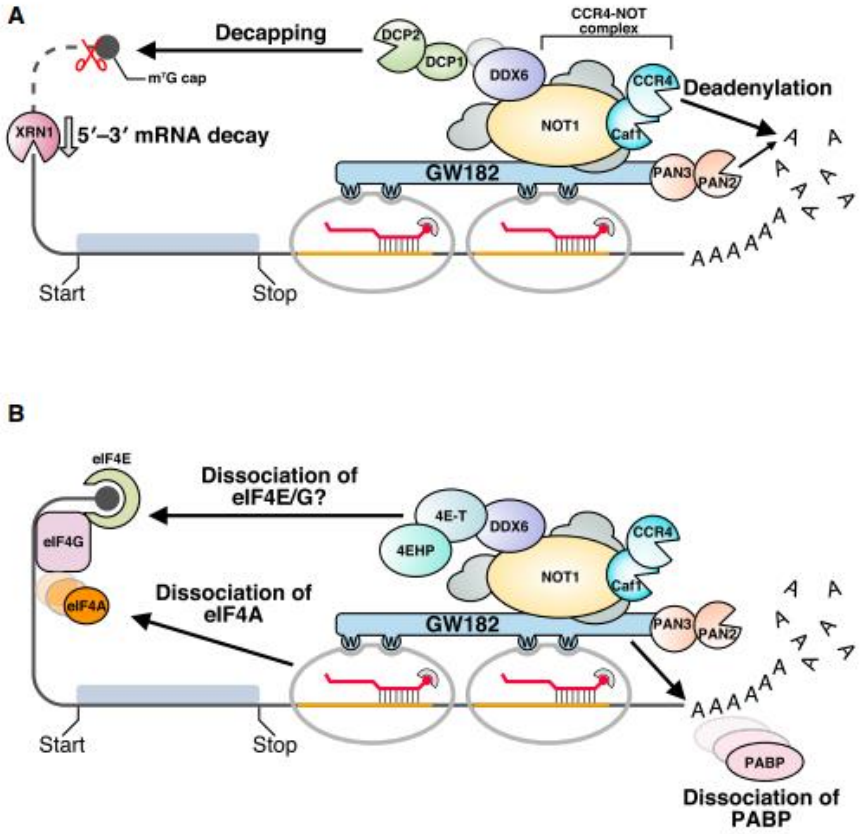

Figure 3. Action of RISC complex in animalsGW182 has glycine-tryptophan (GW) repeats that interact with tryptophan (W) binding pockets of AGOs (A) GW182 destabilizes target mRNAs by recruiting the deadenylation complexes CCR4-NOT and PAN2-PAN3 and the decapping factor DCP2. The decapped and deadenylated mRNA is then degraded by XRN1. (B) The mechanisms of translational repression are less understood, and it can either occur by recruitment of inhibitors that target eIF4E/ 4G, such as DDX6, or by displacement of PABP and eIF4A from target mRNA. (Figure reprinted by Iwakawa and Tomari, 2022)

1.2.4.4 Functions of miRNAs and association with disease

Functional studies, as well as computational approaches have revealed the importance of miRNAs in a wide diversity of biological processes and in a multitude of organisms.

Plant miRNAs have the propensity to target Transcription Factor gene families that are implicated in developmental processes such as patterning and cell differentiation. This inclination is much less pronounced in animals with a small portion of animal miRNAs being involved in developmental processes, while the rest are implicated in a wide variety of biological processes. miRNAs have been involved in cell proliferation, apoptosis and fat metabolism in flies, neuronal development in nematodes and hematopoietic lineage differentiation, neuronal development and cell fate decisions in mammals (13). Changes in the miRNA expression profile have been described in many human diseases, including cancer, neurological disorders, cardiovascular disorders, diabetes, metabolic disorders and viral infection. These dysregulations in miRNA expression can occur either through epigenetic changes (such as promoter hypermethylation and histone modifications) and genetic alterations in miRNA loci, that can affect the production, processing and interactions of a miRNA with the target mRNAs, or by mutations that disrupt components of the miRNA processing machinery and RISC complex. The association of miRNAs with disease makes them useful as potential diagnostic, prognostic and predictive biomarkers, and many relevant studies have highlighted many prominent candidate miRNAs with biomarker capacity, although their utilization has yet to reach maturity [20].

1.2.5 Small RNAs derived from housekeeping noncoding RNAs

The recent advances in high-throughput RNA sequencing (RNA-Seq) and bioinformatics analysis have led to the discovery of a large number of novel noncoding RNAs, among them, small RNAs derived from abundant “housekeeping” noncoding RNAs (rRNA, tRNA, snRNA, snoRNA, etc.). These noncoding RNAs have a well-established role in generic functions of the cell, such as splicing and translation, and small RNAs derived from these noncoding RNA were initially considered as random degradation products. Yet evidence suggests that these small RNAs are conserved and precisely processed. In particular, small RNAs processed from tRNAs have been identified in a wide range of organisms, and there is evidence suggesting that they may exert regulatory functions [6,24].

1.2.5.1 Transfer RNAs (tRNAs)

Transfer RNAs (tRNAs) are highly conserved and abundant RNAs, 70 to 90 bases in length, that have an important role in protein synthesis. There are >500 tRNA genes encoded in the human genome, as well as numerous genetic loci with sequences resembling nuclear and mitochondrial tRNAs termed “tRNA-lookalikes”. The tRNA genes are transcribed by RNA Pol III as precursor tRNAs (pre-tRNAs), which contain additional bases at the 5’ and 3’-ends called leader and trailer sequences that are subsequently trimmed by RNase P and RNase Z respectively. Also, some eukaryotic tRNAs contain introns that are spliced out, followed by folding and post-transcriptional modification. After proper folding a tRNA has a cloverleaf structure with four distinct arms, namely the D arm, anticodon loop, TψC arm and variable loop (**Figure 4**). A CCA trinucleotide is then added to the 3’ end by the enzyme tRNA nucleotidyl-transferase and the now mature tRNAs are aminoacylated by their respective aminoacyl-tRNA synthetase and exported to the cytoplasm, a procedure mediated by a nuclear export receptor [21].

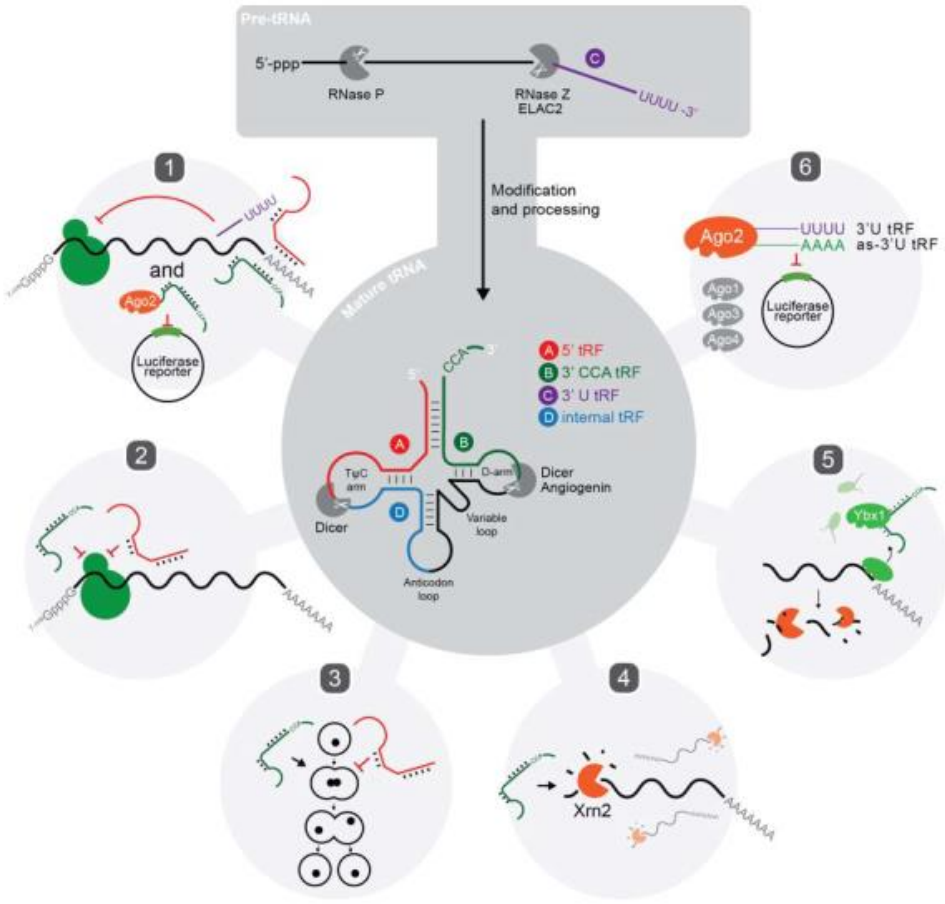

Figure 4. Biogenesis of tRNAs, types and functions of tRFs.Pre-tRNAs are transcribed from tRNA genes by RNA pol III and undergo nuclease cleavage by RNase P and RNase Z, modification including the addition of a CCA trinucleotide and folding to form a cloverleaf structure. The 4 recognized types of tRFs (A, B, C, D) are produced by cleavage from different nucleases at different points of the mature tRNAs. Functions of the tRFs include silencing of complementary target mRNAs (1), translational repression (2), regulation of cell proliferation (3), modulation of mRNA stability (4, 5). (Figure reprinted by Keam and Hutvagner, 2015)

1.2.5.2 Biogenesis of tRNA-derived short RNAs

Short RNAs derived from tRNAs are generally classified into two classes depending on their biogenesis. tRNA-halves are produced under stress conditions (e.g., oxidative stress, heat shock, ultraviolet irradiation, starvation) by cleavage of mature tRNAs in the anticodon loop by the ribonuclease Angiogenin, are 31-40 bases long, and can be further distinguished into 5'-halves and 3'-halves [22]. In contrast, tRNA-derived fragments (tRFs) are shorter in length (13-32nt) and are produced through endonucleolytic cleavage of mature and precursor tRNAs near the D or the TψC arm. tRFs can be further classified into four main types based on the region of the pre-tRNA or mature tRNA they originate from (**Figure 4**). The 5'-tRFs are produced through a cleavage in the TψC-arm, 3'-tRFs are produced through a cleavage in the D arm and include the added CCA trinucleotide, while simultaneous cleavage in the anticodon loop and either D arm or TψC arm produces the internal tRFs (itRFs). Finally, 3'-U-tRFs are produced by the 3'-end of pre-tRNAs through cleavage by RNase Z and include characteristic poly-U residues at the 3'-end [23, 24]. In general little is known regarding the biogenesis of tRFs and it has been proposed that they are produced in a manner similar to the canonical miRNA pathway. Limited evidence suggests that some tRFs are generated in a Dicer-dependent manner in humans, mice and flies and that Dicer is able to generate tRFs *in vitro*, but recent evidence suggests that tRFs can also be produced independently of the canonical miRNA

machinery. RNase Z, required for maturation of tRNAs is indispensable for the generation of 3'-U-tRFs, and another endonuclease, Elac2/RNaseZL has been shown to be required for 3'-U-tRF generation. Angiogenin can also produce tRFs *in vitro* and *in vivo* under non-stress conditions [24].

1.2.5.3 Functionality of tRFs

tRFs are deeply conserved and are present in almost every branch of life, including archaea, bacteria, algae, protozoa, plants, flatworms, flies and mammals. Despite their universality, a major occurring concern is that tRFs are just degradation products generated by endonuclease activity, considering the universality, deep conservation and abundance of their precursor molecules. As a response to that, there is a large amount of evidence from independent groups supporting that tRFs are functional molecules. Several studies suggest that tRFs have a role in translational repression, as they can bind to the ribosomes and reduce translational efficiency, most likely by inhibiting elongation [25, 26]. Another putative role for tRFs is regulation of cell proliferation. A 3'-tRF has been reported to inhibit proliferation in mature human B cells, while a 3'-U-tRF has been reported to promote cell proliferation [27, 28]. 3'-tRFs have also been shown to modulate RNA stability, either by activating exonucleases or by binding and sequestering RNA-binding proteins that stabilize transcripts [29, 30]. Probably the most important function tRFs could have would be the ability to silence the expression of complementary target mRNAs in a manner similar to miRNAs and siRNAs [31].

1.2.5.4 Binding of tRFs to AGO proteins and repression of target mRNAs

The similar size of tRFs and miRNAs makes it reasonable to assume that tRFs may be bound to AGO proteins. Many different studies in human cell lines have identified through AGO protein immunoprecipitation (AGO-IP) that different types of tRFs (5'-, 3'-, 3'-U-tRFs) are indeed bound to AGO proteins and that, for many of these instances, the binding is selective for specific members of AGO proteins. Furthermore, data from Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) verified the selective binding of tRFs to AGO proteins and also suggested that tRFs are loaded onto AGO proteins in a similar manner as miRNAs, as indicated by the positioning of the crosslink-induced mutations. Binding of tRFs to AGO proteins has also been reported for other organisms such as mice, flies and plants, as well as binding of tRFs to the PIWI-clade of Argonaute proteins [32].
Evidence of AGO-bound tRFs functioning as miRNAs was first shown on viruses. It has been shown that an abundantly expressed 3'-tRF in HIV-1 infected cells could be loaded onto AGO2 and silence the complementary viral sequence and a complementary reporter gene [33]. Similarly in RSV-infected cells, it was shown that an accumulated 5'-tRF could silence complementary reporter genes and also was involved in regulating viral replication [34]. These results, along with the high complementarity that 3'-tRFs have to human endogenous retroviral sequences, suggest that tRFs may also have a role in silencing those endogenous sequences. In another study it was reported that AGO-bound 3'- and 3'-U-tRFs could silence complementary target RNAs [35]. Finally, in another example it was reported that overexpression of a particular tRNA led to up-regulation of the respective 3'-tRFs and that these tRFs could in turn repress reporter genes with 3'-UTR complementarity in a Dicer-independent but AGO-dependent manner [31]. In addition to translational repression, tRFs may also exhibit a role in other AGO protein-related functions, such as modulation of histone methylation, mRNA splicing, and DNA damage repair [32].

1.2.5.5 tRFs as potential disease biomarkers

The potential involvement of tRFs in infection and disease makes them useful as biomarkers. As mentioned above, evidence suggests that tRFs have a role in viral infection and can either repress the viral sequence and replication or promote viral replication by modulating host genes' expression [33, 34]. The strong association of tRFs with cell proliferation [26, 27, 28] suggests that tRFs may be used to manipulate highly proliferative cancer cells. There is also evidence that tRFs are differentially expressed in tumors and their expression can distinguish tumor from normal tissue, while in another study it was shown that tRFs can regulate and suppress oncogenic transcripts [23, 30, 36].

2.MATERIALS AND METHODS

2.1 Datasets

sRNA-Seq datasets for two prostate cancer cell lines (i.e., 22Rv1 and DU145) and 10 prostate cancer samples from the TCGA-PRAD project were used for the quantification of tRFs. AGO-PAR-CLIP-Seq datasets for 22Rv1 and DU145 cell lines were used for the identification of tRF-gene interactions. The utilized datasets are described in **Table 1**.

Table 1. Description of sRNA-Seq and AGO-PAR-CLIP-Seq datasets analyzed for the quantification of tRF expression and the identification of tRF-gene interactions.

Accession/id	Repository	Cell type/ Tissue	Dataset type
SRR6082010	Sequence Read Archive	22Rv1	sRNA-Seq FASTQ
SRR6082021	Sequence Read Archive	22Rv1	sRNA-Seq FASTQ
SRR5689199	Sequence Read Archive	DU145	sRNA-Seq FASTQ
TCGA-HC-7080-01A-11R-1964-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-KK-A7AU-01A-11R-A360-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-KK-A8IK-01A-11R-A36B-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-M7-A722-01A-12R-A36B-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-VN-A88N-01A-11R-A36B-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-XK-AAIW-01A-11R-A41R-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-XQ-A8TA-01A-11R-A36B-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-Y6-A9XI-01A-11R-A41R-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-YL-A8HL-01A-11R-A36B-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
TCGA-YL-A8SA-01A-21R-A37H-13	TCGA-PRAD	Prostate	sRNA-Seq FASTQ
SRR3502967	Sequence Read Archive	22Rv1	AGO-PAR-CLIP-Seq BAM
SRR3502969	Sequence Read Archive	22Rv1	AGO-PAR-CLIP-Seq BAM
SRR3502970	Sequence Read Archive	22Rv1	AGO-PAR-CLIP-Seq BAM
SRR3502975	Sequence Read Archive	DU145	AGO-PAR-CLIP-Seq BAM

2.2 Extraction of tRF information from MINTbase v2.0

MINTbase v2.0 [37] is a repository containing information about tRFs found in a variety of human tissues. The tRFs included in the database were identified by the analysis of

11,719 human datasets, including 11,198 small RNA-Seq (sRNA-Seq) datasets from The Cancer Genome Atlas (TCGA), using the MINTmap algorithm [38] and by retaining only tRFs exhibiting a relative abundance of ≥ 1 Reads-Per-Million (RPM). MINTbase consists of five 'vistas', each focusing into a different aspect of tRFs: genomic loci, RNA molecule, tRNA alignment, expression, and summary. The 'genomic loci' vista is the most detailed and focuses on the genomic information of each tRF, 'RNA molecule' vista contains basic information about each tRF molecule, 'tRNA alignment' vista focuses on the mature tRNA molecule and the alignment with the tRFs that map to it, 'expression' vista provides information on the tissues and datasets containing a given tRF and 'summary' vista provides access to all the available information for a given tRF. Due to the finite number of human tRNAs, it is feasible to enumerate all possible tRFs with a specific range of lengths, and MINTbase contains a summary record for each possible tRF with a length of 16-50 nt. MINTbase incorporates two different labeling schemes for tRFs. The genome-centric labels contain information about the genomic coordinates of the tRNA gene, as well as the start and end position relative to the mature tRNA and the length of the tRF. The tRF license plates depend solely on the tRF sequence and are independent from the genome-assembly.

The 'genomic loci' vista of all the 5'-, 3'- and i-tRFs with a relative abundance of ≥ 1 RPM was downloaded from MINTbase. Since the tRF genomic coordinates included in MINTbase correspond to the GRCh37/hg19 genome assembly, they were remapped to the GRCh38 genome assembly using the Remap tool from NCBI (https://www.ncbi.nlm.nih.gov/genome/tools/remap). The information of the genomic coordinates (chromosome, start/stop position and strand) together with the genome-centric labels, the license plates and the tRF type for each tRF was incorporated into the existing noncoding RNA annotation GTF file of Manatee (https://github.com/jehandzlik/Manatee/blob/annotation/ncRNA_hg38.gtf).

2.3 Quantification of tRFs from sRNA-Seq data with Manatee

sRNA-Seq is the gold standard high-throughput technique for the quantification of the expression of small RNA (sRNA) species, as well as for the identification of novel sRNAs and sRNA species. The analysis of sRNA-Seq data is less mature and requires additional caution compared to RNA-Seq, due to technical hindrances that arise from the small length of reads and transcripts. Short sequences tend to map to multiple genetic locations, and this problem of multi-mapping is exacerbated by the fact that many small RNAs originate from repeat genetic loci, and also undergo post-transcriptional modifications. Current algorithms try to address the multi-mapping problem using various approaches. One approach is alignment against known sRNA annotations, but in this case the methods are limited to quantifying only known sRNAs, a limitation that is even greater if the algorithm is dedicated to quantifying a single biotype of sRNA. Another approach is to assign multi-mapping reads to all their mapping positions, or to split them between the mapping positions equally or in a weighted way, but still there can be loss of biological information regarding the expression of different sRNA biotypes.

The sMAll rNa dATa analysis pipElinE (MANATEE) is an algorithm for sRNA quantification that attempts to address the multimap issue by combining information from existing annotation and density of uniquely aligned reads without prioritizing any specific sRNA biotype. The algorithm also attempts to salvage highly multimapping and unaligned reads by aligning them against the transcriptome based on the provided annotation, while gradually increasing the number of allowed mismatches. Manatee also enables the detection of expressed unannotated genomic loci. The algorithm requires pre-processed (barcode and adapter removed) FASTQ/FASTA sRNA-Seq files and ncRNAs annotation

in GTF format as input and generates three tab-separated files containing the quantified transcripts, isomiR sequences and putative novel expressed loci. The workflow of Manatee is shown in **Figure 5** [39].

The barcode- and adapter-cleaned FASTQ sRNA-Seq files were processed with Manatee using the default parameter settings and the modified GTF annotation file that incorporates the tRF information extracted from MINTbase.

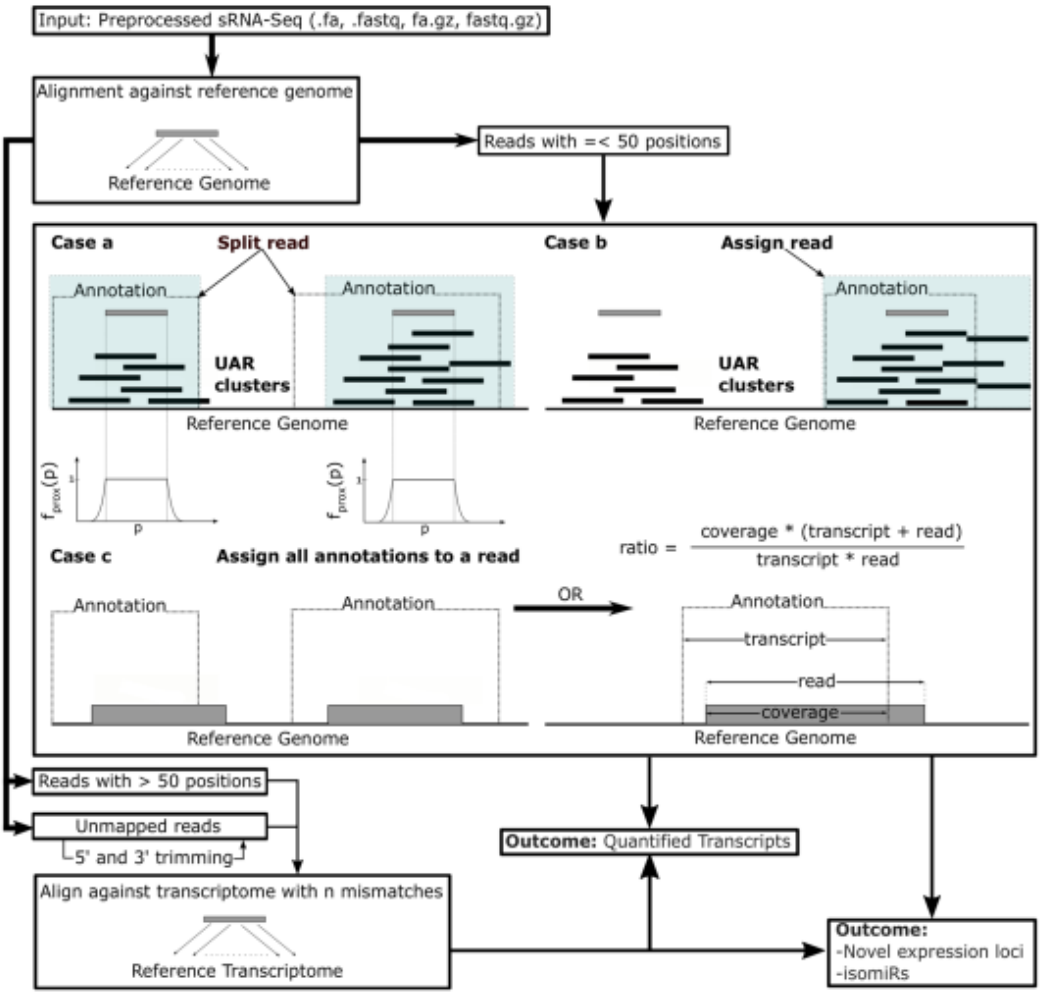

Figure 5. The Manatee workflow.Multi-mapping reads are either (a) split between annotated and uniquely aligned read (UAR) containing loci, (b) assigned to annotated and UAR-containing loci, or (c) assigned to annotated loci. Highly multi-mapped and unmapped reads are aligned against the transcriptome with a gradual increase in the number of allowed mismatches (Figure reprinted by Handzlik et al., 2020)

2.4 Analysis of tRF-gene interactions from AGO-PAR-CLIP data with microCLIP

RNA transcripts in eukaryotes are subject to post-transcriptional control by RNA-binding proteins (RBPs) and ribonucleoprotein complexes (RNPs) that modulate their expression. More specifically, a large number of miRNAs bound to members of the Argonaute (AGO) protein family mediate translational repression and/or degradation of complementary target mRNAs, and to explore and map those interactions different experimental methodologies can be used. Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) is a method that includes the incorporation of 4-thiouridine (4SU) into transcripts of cultured cells, followed by UV-triggered RNA-protein

crosslinking and immunoprecipitation, partial digestion of protein bound RNA with RNase treatment, separation of the protein-RNA complexes with SDS-PAGE, recovery of RNA molecules and conversion to cDNA, and finally deep sequencing (**Figure 6**). The crosslinking of 4SU to the protein amino acid side chains increases the frequency of T-to-C transitions in the sequenced cDNA and these transitions reveal the crosslinked sites [40]

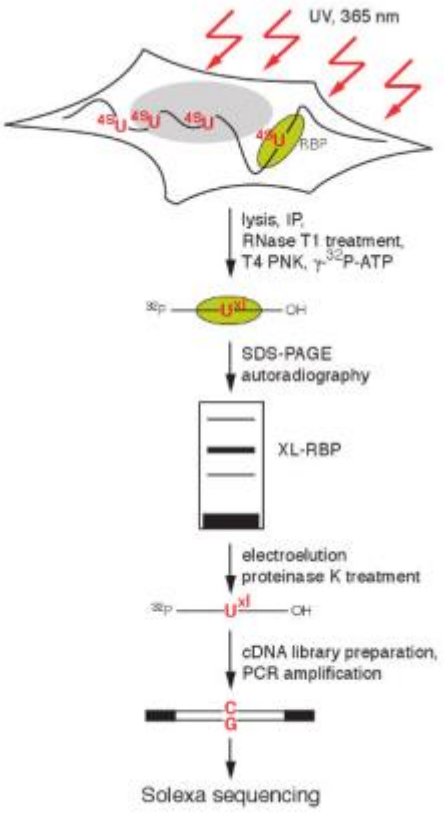

Figure 6. Overview of PAR-CLIP.The incorporation of 4-thiouridine into transcripts during cell culture enhances the RNA-protein crosslinking by UV irradiation. The RNA-protein complexes are then RNase-treated, immunoprecipitated and size-fractionated. The RNA is then recovered, converted to cDNA and deep sequenced. The T-to-C transitions reveal the crosslinked sites. (Figure reprinted by Hafner et al., 2010)

PAR-CLIP of AGO proteins is thus considered among the most powerful methods for miRNA target characterization on a transcriptome wide scale. Computational methods developed for the complex analysis of AGO-PAR-CLIP data utilize different mathematical models but fail to identify a large portion of true miRNA-gene interactions mainly due to their dependence on the T-to-C conversions, introduced during PAR-CLIP, to identify miRNA-binding sites. microCLIP is an *in silico* framework for the identification of miRNA-target interactions from AGO-PAR-CLIP-Seq data, that utilizes an extensive collection of experimental data, including AGO-PAR-CLIP data, miRNA-binding events from highly specific high/low throughput techniques and high throughput miRNA perturbation data. microCLIP is based on a super learning scheme that combines deep learning, random forest and gradient boosting classifiers and has been shown to perform better than a single model. Positive/negative miRNA-target pairs extracted from the experimental data combined with signal from AGO-PAR-CLIP were used for the training and validation of the algorithm based on a set of 131 descriptors related to CLIP-Seq, MRE (miRNA recognition elements) and miRNA/MRE hybrid derived characteristics. microCLIP utilizes

a multi-layer super learner classification scheme, with nine base classifiers specialized for subsets of features constituting the first layer and a gradient boosting meta-classifier aggregating their outcomes in the second layer (**Figure 7**). microCLIP is the only available implementation able to initiate the AGO-PAR-CLIP data analysis from SAM/BAM files and requires a SAM/BAM AGO-PAR-CLIP alignment file and a list of miRNAs as minimum input. It proceeds to scan the AGO-enriched read clusters for candidate MREs, including a wide range of binding types (canonical and non-canonical), scoring them through a super learner ensemble scheme, the first implementation to utilize such a scheme. Due to indications that MREs supported by AGO-enriched clusters without T-to-C conversions may have functional importance, microCLIP, unlike other implementations, processes and scores all AGO-enriched clusters. The inclusion of MREs supported by non-T-to-C clusters results in an increased number of identified interactions and an enhancement of downstream analyses, such as pathway enrichment analysis. The utilization of an extensive collection of experimental data coupled with a super learning scheme, and the inclusion of previously omitted non-T-to-C clusters, result in an increased accuracy of microCLIP in the identification of miRNA-target interactions, when compared with other similar implementations [41].

The AGO-PAR-CLIP BAM files, together with a list of the 100 top-expressed tRFs from the matched sRNA-Seq files, were processed with microCLIP using the default parameter settings.

2.5 Enrichment analysis of the tRF-interacting genes

clusterProfiler is a R package that provides an interface that enables functional annotation of genes as well as access, manipulation and visualization of enrichment results. The enrichGO() function performs enrichment analysis of a gene set for GO (Gene Ontology) terms. The enrichment analysis of the tRF-interacting genes for Biological Processes GO terms. was performed using the Benjamini–Hochberg method and a FDR < 0.01.

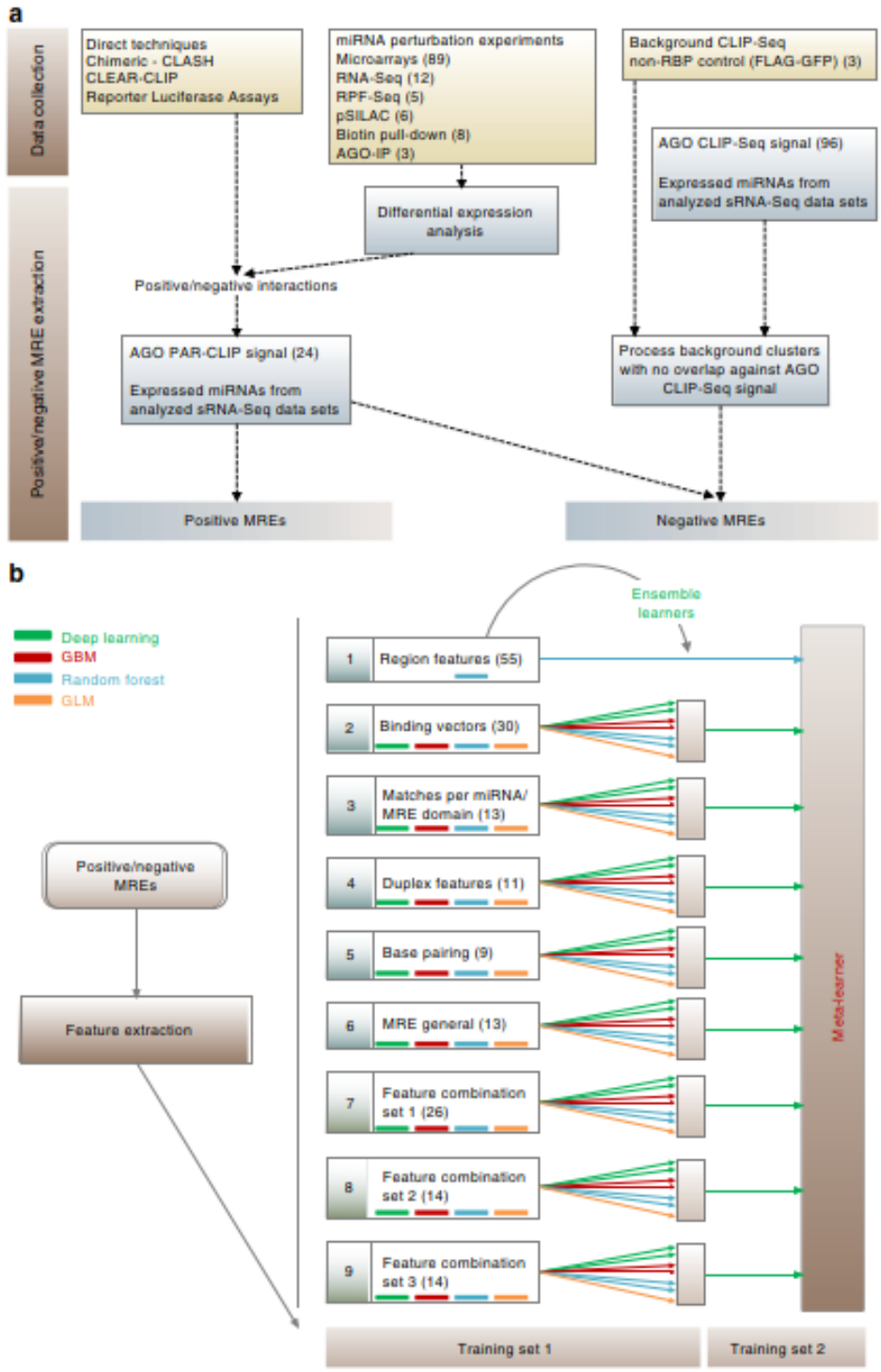

Figure 7. Overview of microCLIP framework.(a) Positive/negative miRNA-target pairs extracted from high/low throughput techniques, signal from AGO-PAR-CLIP libraries and background CLIP-Seq formed the training/test set of microCLIP. (b) The algorithm includes nine base classifiers in the first layer, specialized for feature subsets, with 8 of them utilizing a super learning scheme. A gradient boosting meta-classifier in the second layer aggregates the output from the first layer. (Figure reprinted by Paraskevopoulou et al., 2018)

3.RESULTS

3.1 Workflow

The overview of the applied workflow is shown in **Figure 8**.

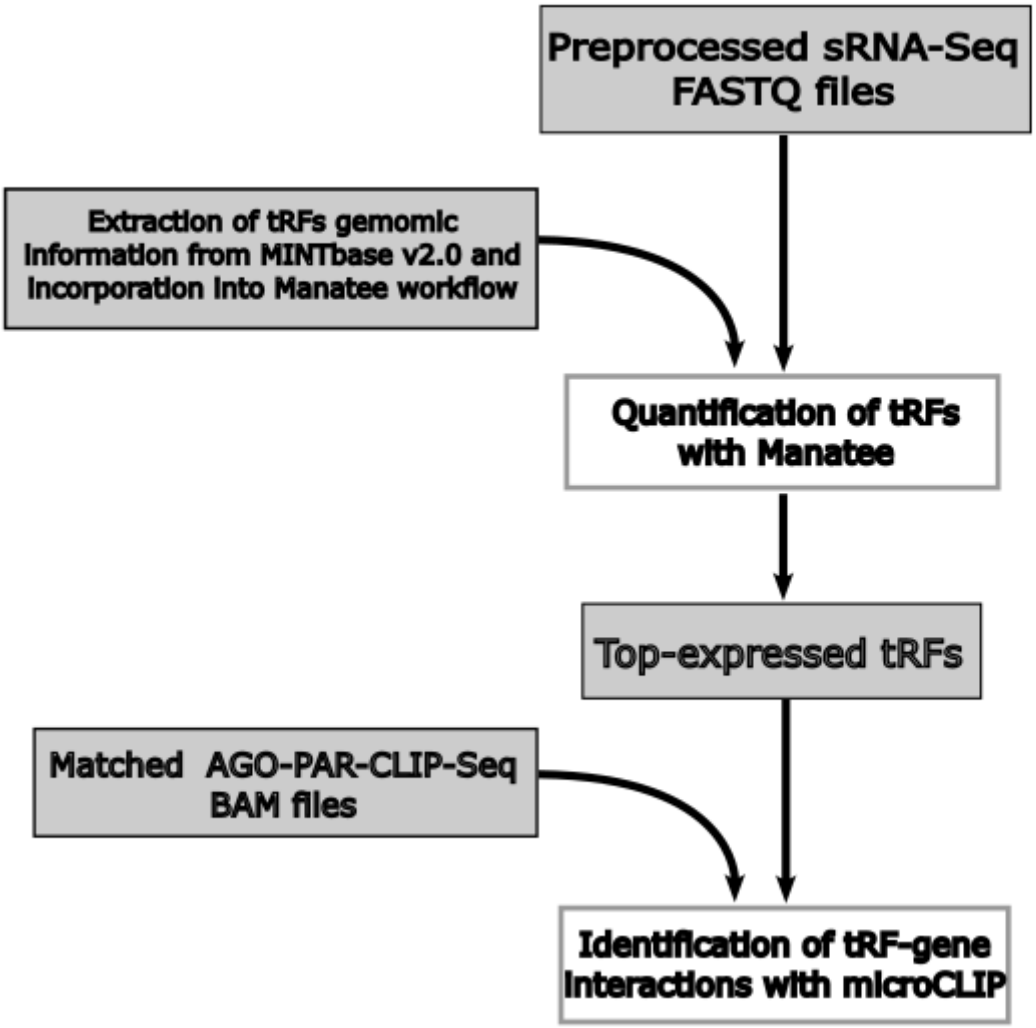

Figure 8. Overview of the applied workflow.The adapter trimmed sRNA-Seq FASTQ files were processed with a modified Manatee workflow that incorporates tRF information from MINTbase v2.0, to quantify tRF expression. The top-expressed tRFs were subsequently used as input for the identification of tRF-gene interactions from AGO-PAR-CLIP-Seq BAM files with microCLIP.

3.2 Quantification of tRF expression with Manatee

The output count files from the processing of the 10 TCGA and the 3 cell line sRNA-Seq datasets with Manatee were merged into a single count matrix containing counts for 16,498 tRFs. The datasets were labeled as “TCGA” and “cell line” accordingly and the count matrix was filtered (with the filterByExpr() function of the edgeR R package) to keep only tRFs that have sufficiently large read counts in a significant number of datasets (i.e., at least the size of the smallest group, in this case 3 datasets for the cell line group). The filtered count matrix contained read counts for 741 tRFs and was normalized and the values converted to counts per million (CPM) on a log2 scale (with functions calcNormFactors(), and cpm() of edgeR R package). Hierarchical clustering was applied

to the normalized and log_2-scaled values to produce the heatmap (with pheatmap R package) depicted in **Figure 9**.

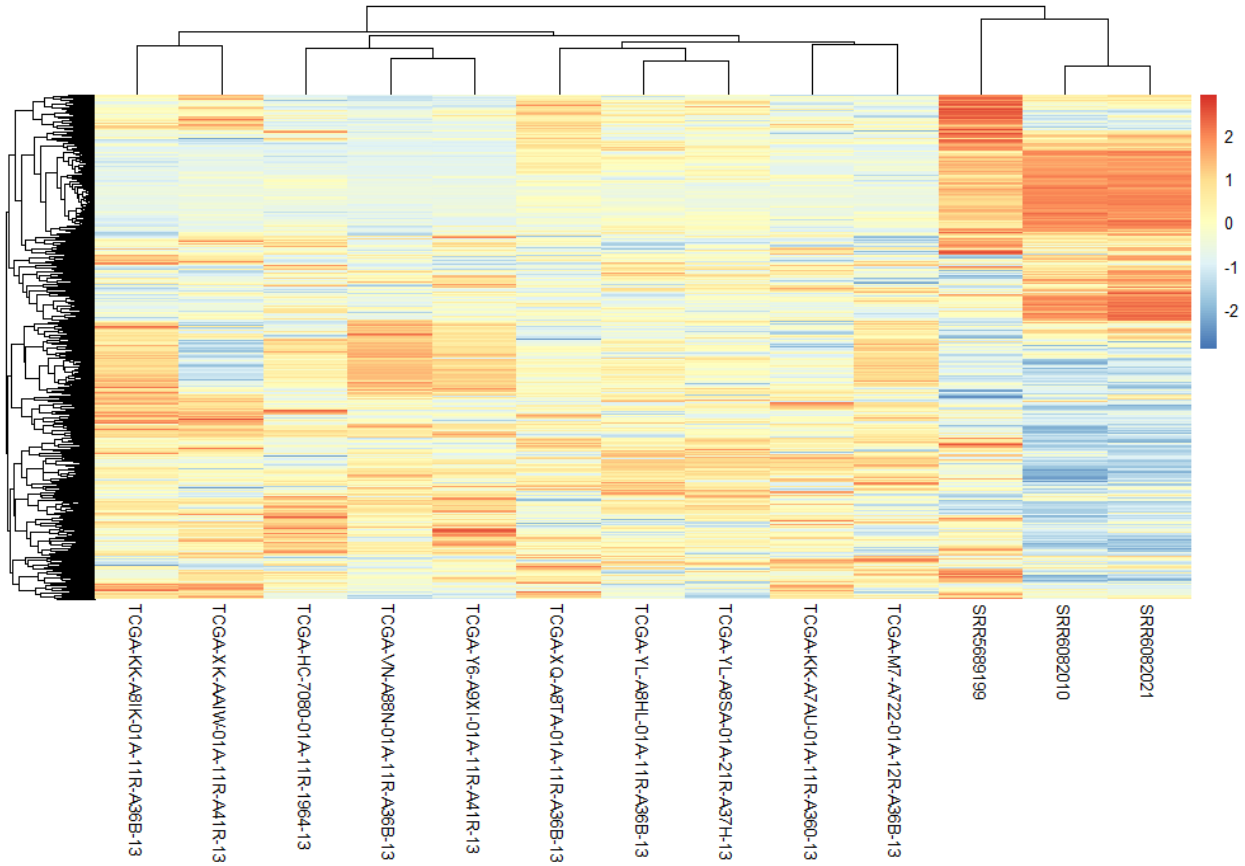

Figure 9. Heatmap showing hierarchical clustering of the normalized and log2-scaled expression values of 741 tRFs that have sufficient large expression in a significant number of the datasets (at least 3).

From the tRF expression heatmap it is evident that the datasets from the 22Rv1 and DU145 cell lines (SRR6082010, SRR6082021, SRR5689199) cluster together and separately from the TCGA datasets, with the two datasets from the 22Rv1 cell line (SRR6082010 and SRR6082021) showing high similarity. The TCGA datasets exhibit high heterogeneity of tRF expression, something that can be attributed to the fact that they originate from human tissue samples and probably there are biological and technical factors that may contribute to this heterogeneity.

The 100 top-expressed tRFs with a maximum length of 28nt for each cell line (for 22Rv1 the mean expression of the two datasets was used) were selected to be used as input for the analysis of the cell line-matched AGO-PAR-CLIP-Seq datasets.

3.3 Analysis of tRF-gene interactions from AGO-PAR-CLIP-Seq data with microCLIP

Four AGO-PAR-CLIP-Seq datasets (three for 22Rv1 cells and one for DU145 cells) were processed with microCLIP to identify tRF-gene interactions for the 100 top-expressed tRFs of each cell line that were identified in the previous step. The MREs (miRNA recognition elements) in the output file of microCLIP were filtered in order to keep only those MREs that overlap with the 3' UTR gene regions in the GENCODE v.41 gene annotation (downloaded from https://genome.ucsc.edu/). The total reads, number of

MREs identified, and number of genes these MREs are located per dataset are shown in **Table 2**. A significant positive correlation (Pearson's coef. 0.99, $p<0.05$) exists between the number of MREs and genes identified and the number of reads each dataset contains.

Table 2. Number of reads, MREs and genes per PAR-CLIP dataset.

Dataset	Cell type	Number of reads	Number of MREs	Number of genes
SRR3502967	22Rv1	4,478,609	319,859	5,891
SRR3502969	22Rv1	7,565,726	607,264	8,258
SRR3502970	22Rv1	432,603	8,535	1,009
SRR3502975	DU145	3,097,933	158,175	4,252

Additional metrics for each dataset regarding the number of genes each tRF interacts with, the gene types, the number of MREs each gene harbors and the distribution of the score that microCLIP calculates for each MRE are shown in **Figures 10-13**.

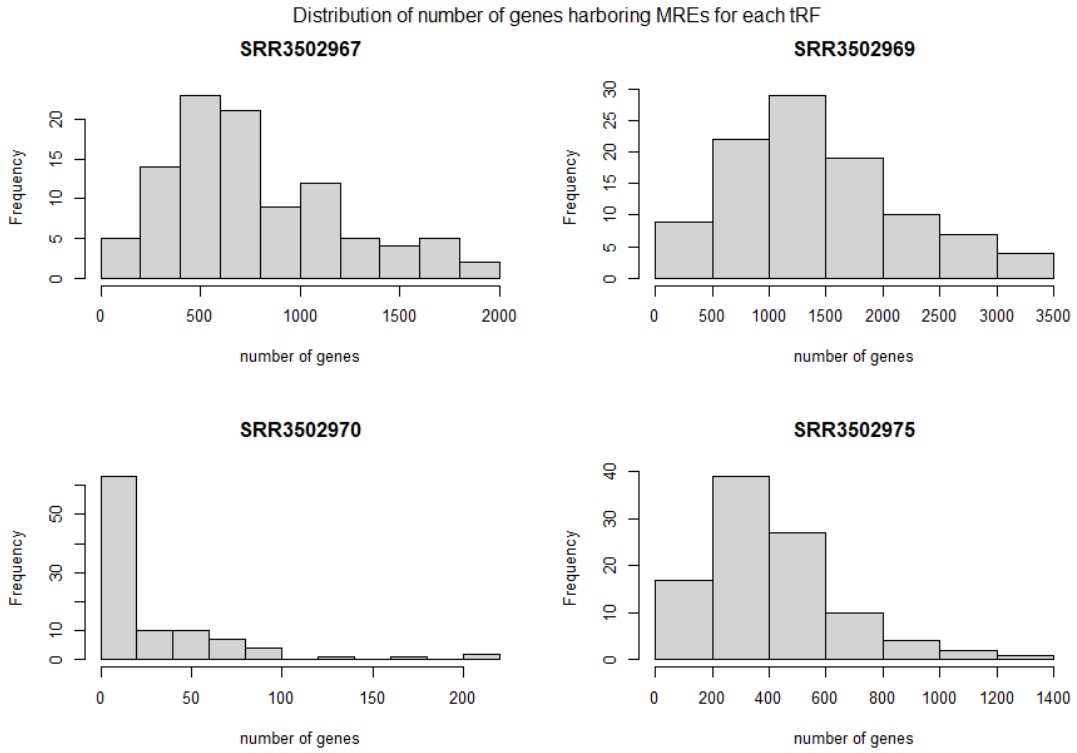

Figure 10. Distribution of the number of genes each of the top-expressed tRFs interacts with.

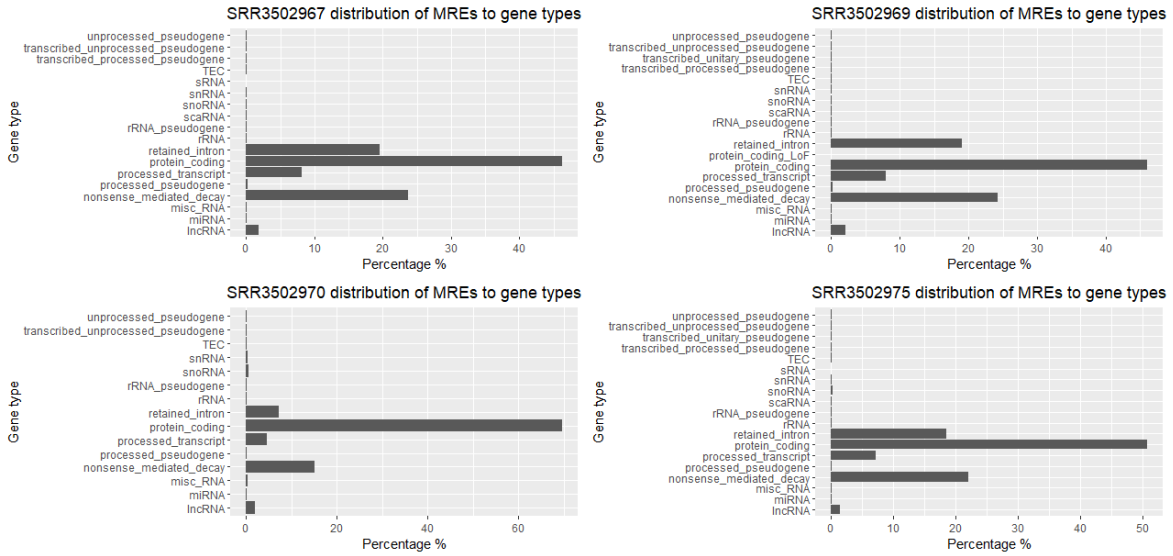

Figure 11. Distribution of gene types for the genes with identified MREs.

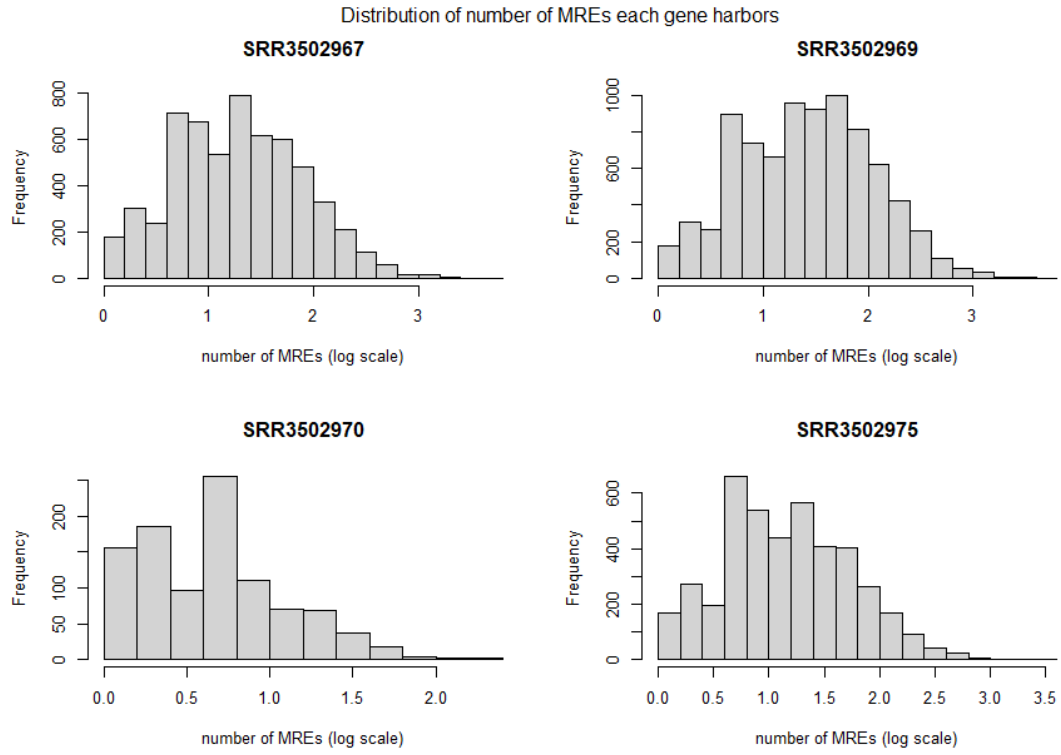

Figure 12. Distribution of the number of MREs identified in each gene.

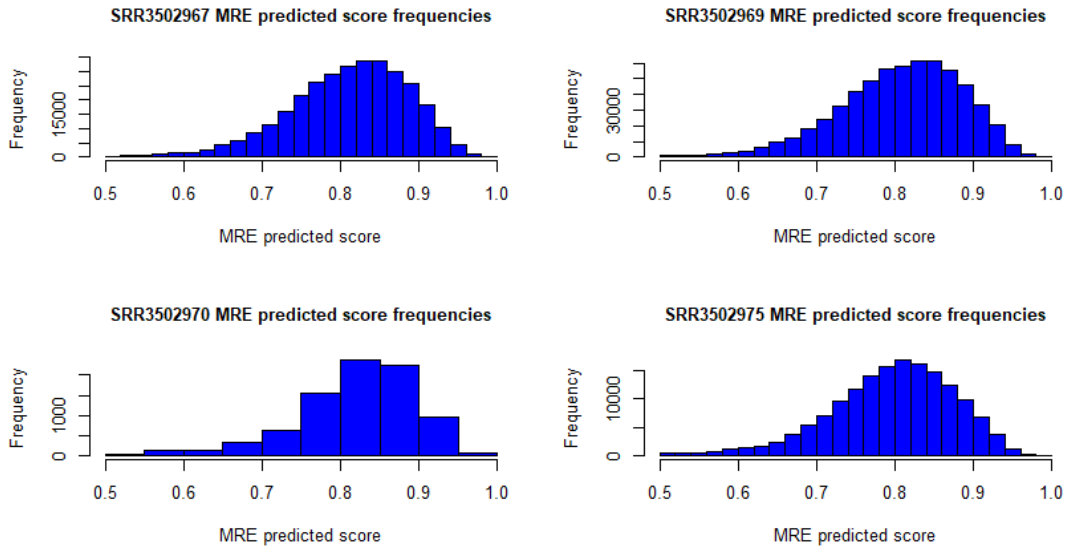

Figure 13. Distribution of the scores predicted by microCLIP for the MREs.

It is evident that dataset SRR3502970 (22Rv1 cell line) is somewhat distinct from the other datasets, probably because of the smaller number of reads it contains, and could be best considered as an outlier. Each of the top-expressed tRFs seems to interact with hundreds of genes, the number ranging from 500 to 1,500 for the majority of tRFs (this number is a little smaller for dataset SRR3502975). Similarly, each gene harbors many MREs, with the number ranging from a few to a hundred MREs for the majority of the genes. As expected, the majority of genes interacting with tRFs are protein-coding genes, and the gene types as well as their proportions are found to be similar between the different datasets. Regarding the distribution of the microCLIP scores, it is left skewed and we can assume that a cut-off of ≥0.8 is reasonable in order to filter for higher quality MREs.

As mentioned in the Materials and Methods section, microCLIP supports the identification of a wide range of canonical and non-canonical MRE binding types, also including in the analysis AGO-enriched clusters without T-to-C conversions. These different types of MREs may have different characteristics and/or functionalities. An exploration of the possible differences between the different MRE types for the different datasets is shown in **Figures 14-17**.

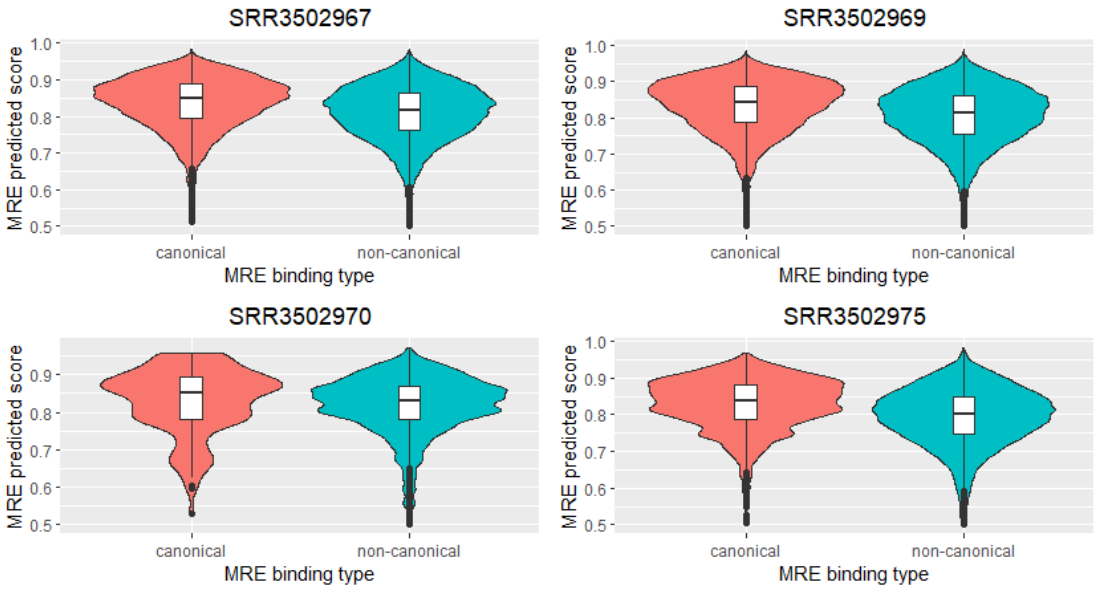

Figure 14. Violin plots of the scores predicted by microCLIP for canonical and non-canonical MREs.

Figure 15. Violin plots of the scores predicted by microCLIP for for MREs identified in TC and non-TC clusters.

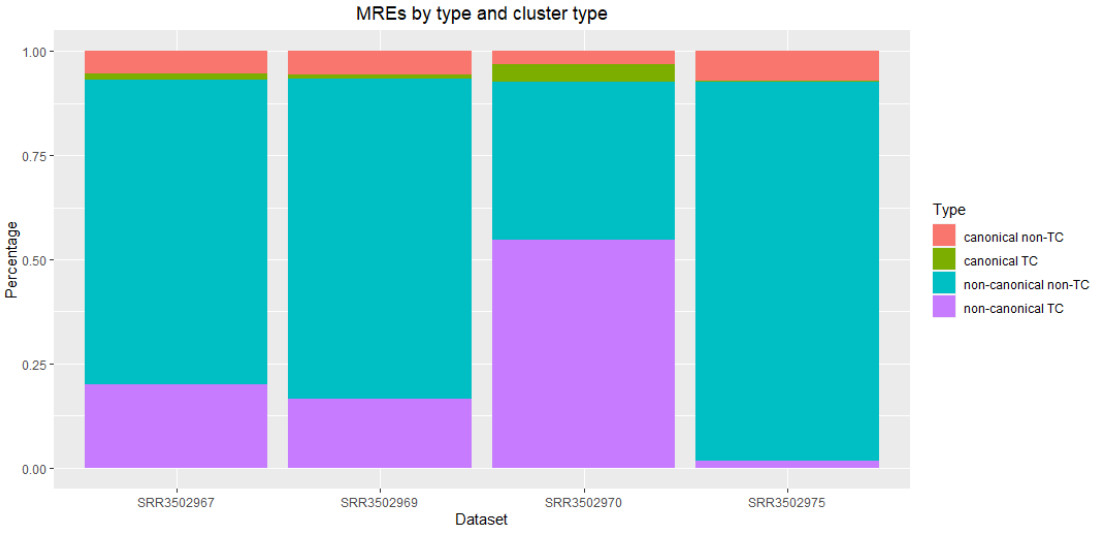

Figure 16. Percentages of the different MRE types in the datasets.

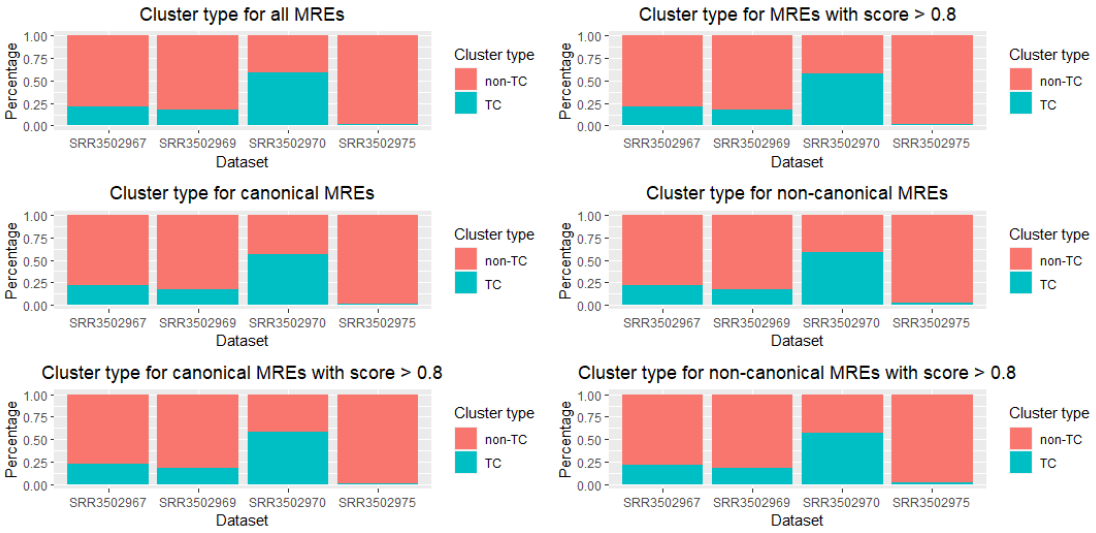

Figure 17. Percentage of TC and non-TC clusters for different subsets of the MREs.

From **Figures 14-15** it is apparent that canonical MREs possess a slightly higher score than non-canonical MREs, while there is no difference in the score between MREs identified from TC and non-TC clusters. Non-canonical MREs are predominant over canonical MREs, with their number being much larger, and the same is the case with MREs identified from TC and non-TC clusters, with non-TC clusters being the majority (**Figure 16**). The TC to non-TC proportion is unaltered from that of the total MREs when filtering for canonical MREs, non-canonical MREs, higher quality MREs ($score_{microCLIP} \geq 0.8$), or a combination of these (**Figure 17**). This observation implies that MREs identified from non-TC clusters may have similar efficacy and functionality as the MREs identified from TC clusters. It is again evident that dataset SRR3502970 stands out from the rest of the datasets, and is also noticeable that dataset SRR3502975 has a lower proportion of TC clusters compared to the other datasets, something that could be attributed to technical and/or biological factors.

3.4 Functional analysis of microCLIP-derived tRF-gene interactions

A next step was to perform enrichment analysis, in an effort to determine if the genes harboring MREs for the top-100 expressed tRFs are enriched for genes associated with specific GO (Gene Ontology) Biological Processes for each dataset. The enrichment analysis resulted in a number of enriched Biological Processes (n = 1054, 1176, 109, 986 respectively) shown in **Table 3**. Hierarchical clustering (with pheatmap R package) of the log_{10}-transformed adjusted p-values resulted to the heatmap depicted in **Figure 18**. It is evident that datasets SRR3502967 and SRR3502967 (22Rv1 cell line) cluster together, something to be expected, while the 22Rv1 SRR3502970 dataset seems to be clustering together with the DU145 dataset SRR3502975, confirming its outlier state.

Table 3. Number of enriched GO biological processes with adjusted p-value < 0.01 per dataset.

Dataset	Number of enriched biological processes with adjusted p-value < 0.01
SRR3502967	1054
SRR3502969	1176
SRR3502970	109
SRR3502975	986

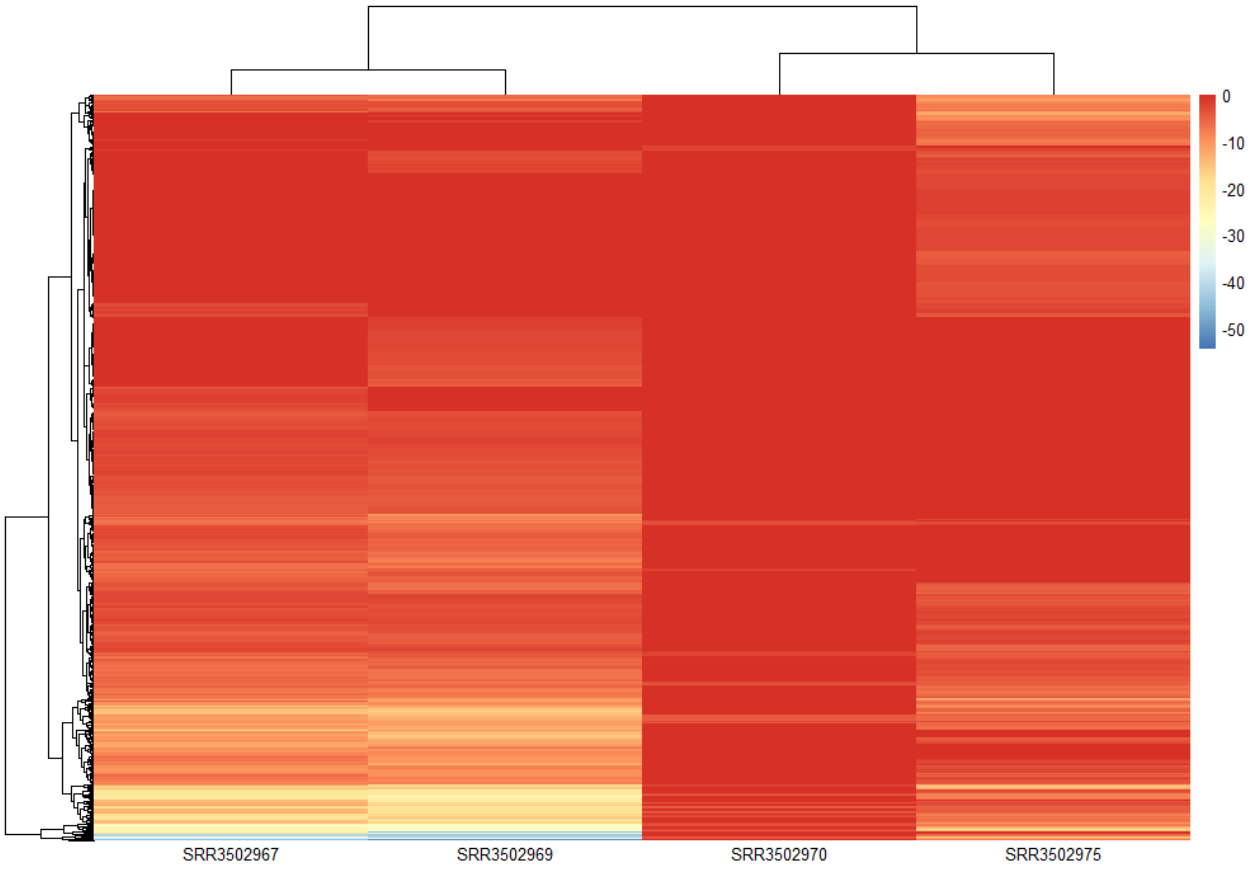

Figure 18. Heatmap of the log_{10}-transformed adjusted p-values of enriched biological processes.

Interestingly, focusing on the down left corner of the above heatmap reveals a set of 13 Biological Processes that are extremely enriched (i.e., adjusted p-value < 10^{-30}) in both SRR3502967 and SRR3502969 22Rv1 datasets (**Figure 19**).

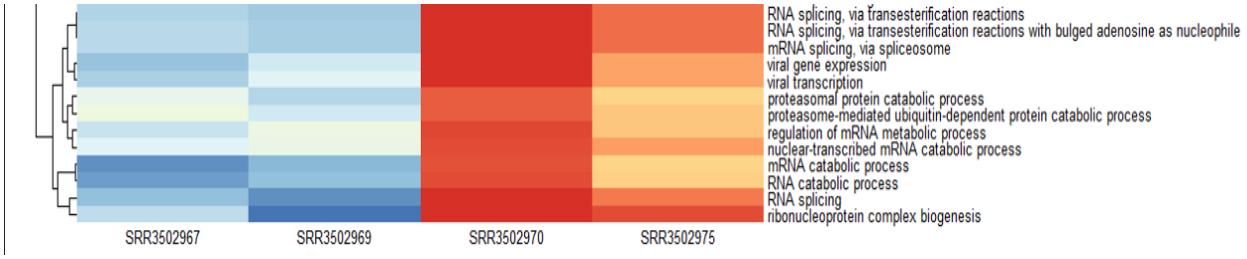

Figure 19. Subset of the most significantly enriched biological processes in two 22Rv1 cell datasets.

In a further investigation of the characteristics of the MREs that are canonical, non-canonical, residing in TC and residing in non-TC clusters, the enrichment analysis was repeated, as described above, for those 4 distinct sets of MREs, while also retaining only MREs with a microCLIP score ≥ 0.8. The resulting heatmap is depicted in **Figure 20**.

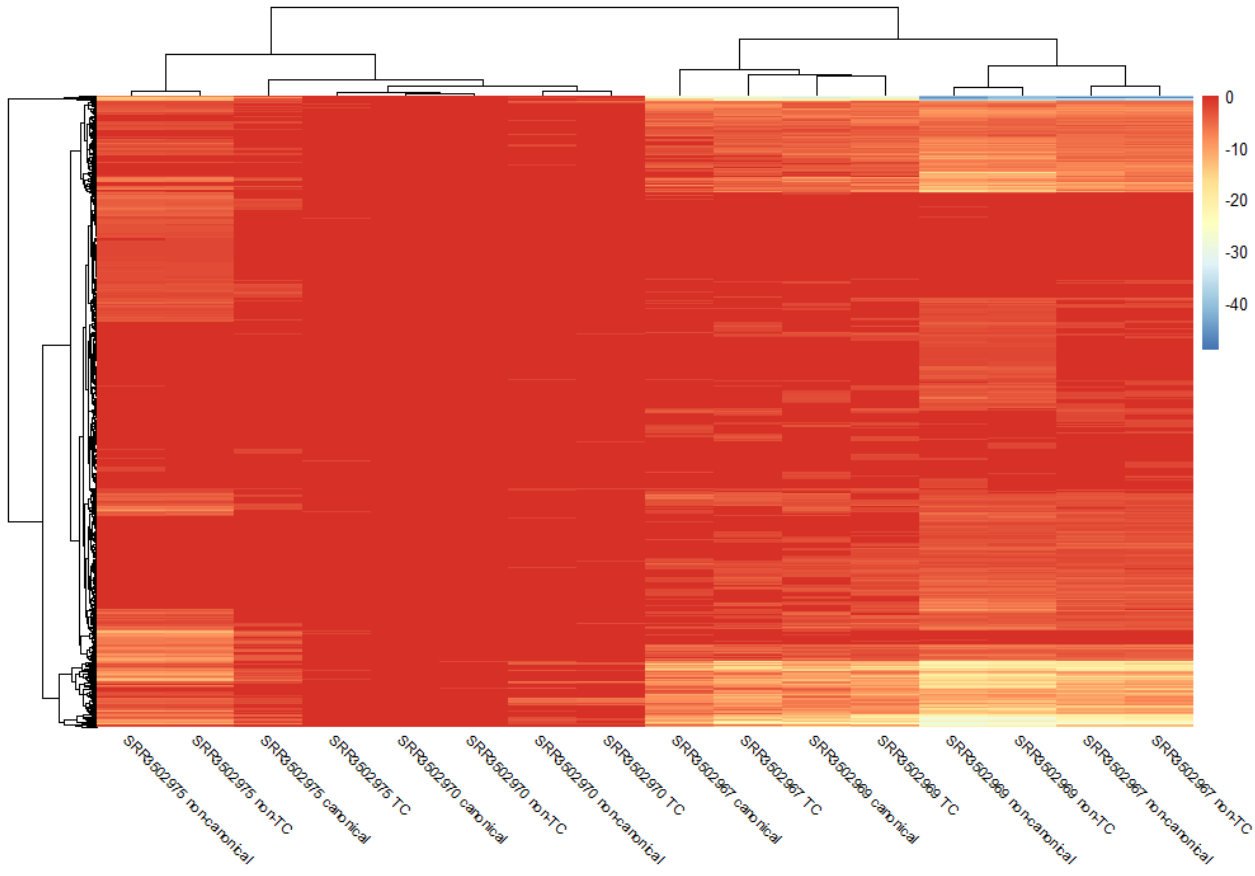

Figure 20. Heatmap of the $\log_{10}$-transformed adjusted p-values of enriched biological processes for the canonical, non-canonical, TC, non-TC MREs with a score ≥ 0.8.

The above heatmap clusters non-canonical and non-TC MREs of 22Rv1 datasets (SRR3502967 and SRR3502969) on one side, and the canonical and TC MREs on the other side, irrespective of which dataset they originate from. Also, the extreme enrichment of the same set of biological processes described previously is retained, although the non-canonical and non-TC MREs exhibit less pronounced adjusted p-values than the canonical and TC MREs (**Figure 21**).

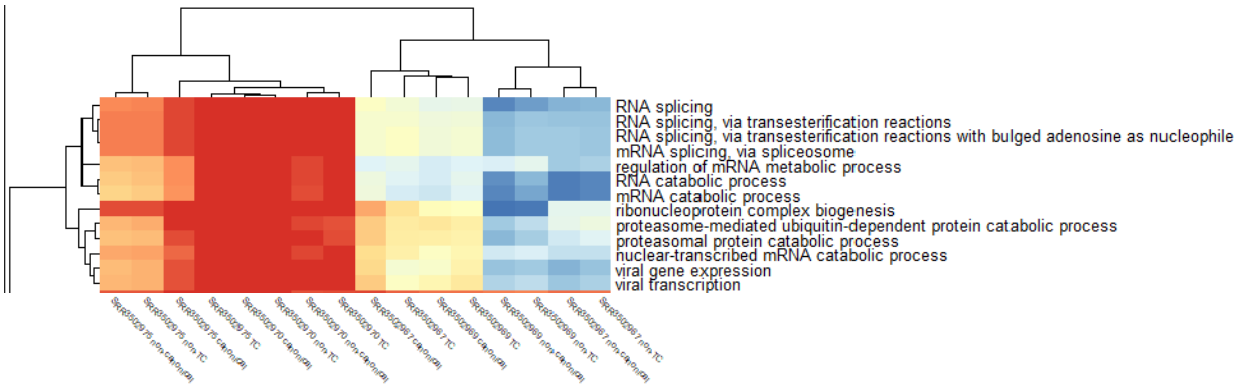

Figure 21. The enrichment of the specific set of biological processes is retained even after grouping MREs into their different types.

The enriched Biological Processes in the SRR3502967 and SRR3502969 22Rv1 datasets are shown in **Table 4.**

Table 4. List of enriched biological processes in the two 22Rv1 datasets.

GO ID	Description	Number of genes of the process in the datasets (SRR3502967/ SRR3502969)	Number of genes included in the process
GO:0006402	mRNA catabolic process	243/277	376
GO:0006401	RNA catabolic process	258/297	415
GO:0008380	RNA splicing	281/345	487
GO:0019080	viral gene expression	149/162	195
GO:0019083	viral transcription	138/149	178
GO:0000375	RNA splicing, via transesterification reactions	236/282	393
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	233/279	390
GO:0000398	mRNA splicing, via spliceosome	233/279	390
GO:0022613	ribonucleoprotein complex biogenesis	270/348	482
GO:1903311	regulation of mRNA metabolic process	211/239	344
GO:0000956	nuclear-transcribed mRNA catabolic process	147/166	210
GO:0010498	proteasomal protein catabolic process	259/326	483
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	232/289	424

Focusing on these most significantly enriched processes, a supplemental analysis was conducted to interrogate whether any tRFs act predominantly as major regulators for them, i.e., possessing a significant proportion of MREs in the associated genes. Isolation of genes that are associated with these processes resulted in 918 (SRR3502967) and 1140 (SRR3502969) genes, with 910 of them being common to both datasets. There is also considerable overlap of the associated genes between the different processes. The distribution of the number of genes of the 13 enriched processes that seem to be regulated by each tRF (harboring MREs for it) is presented in the upper half of **Figure 22**. The percentage of genes each single tRF seems to regulate for each specific process (i.e., amount of genes of the process a tRF regulates over the number of genes of the process in the dataset including all 100 tRFs) was calculated and the mean percentage of genes for all the processes has a distribution shown in the lower half of **Figure 22**. Both distributions are right-skewed, indicating there is a small number of tRFs that regulate a greater number of genes or a greater proportion of genes per process in average. The top 10 tRFs in terms of number of genes or mean proportion for all the processes are shown in **Table 5**.

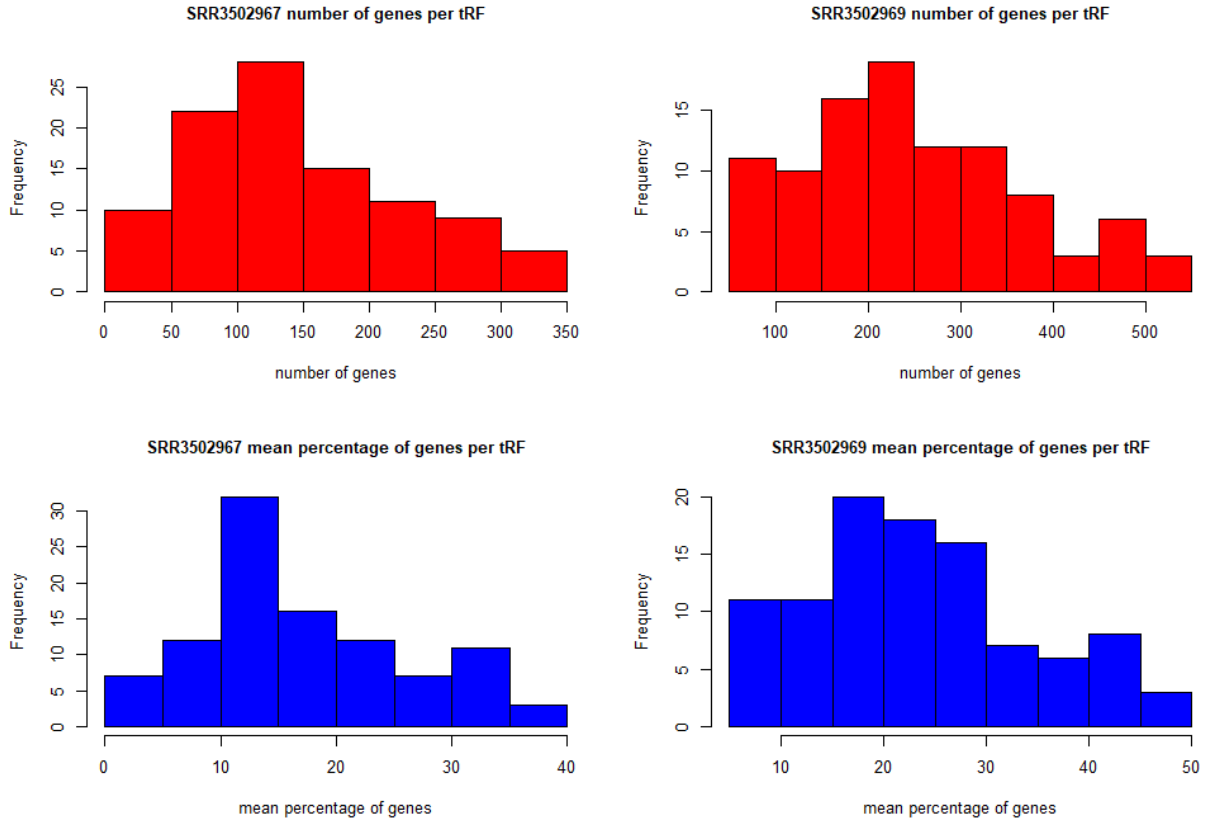

Figure 22. Distributions for the number of genes in total for all the enriched processes (upper half) and the mean of the proportion of genes for each process (lower half) that each tRF seems to regulate.

Table 5. The top-10 tRFs in terms of number of total genes or mean proportion for all processes.

SRR3502967 dataset		SRR3502969 dataset	
tRF	**Number of genes from the total of 918 genes for all the 13 processes**	**tRF**	**Number of genes from the total of 1140 genes for all the 13 processes**
tRF-18-SP5830D4	337	tRF-16-SP5830D	523
tRF-16-SP5830D	334	tRF-18-SP5830D4	522
tRF-18-P4R8YP04	314	tRF-18-P4R8YP04	520
tRF-25-SP58309MUK	305	tRF-25-SP58309MUK	488
tRF-21-RXF4P2PS0	303	tRF-25-SP5830MMUK	464
tRF-17-18YKISM	298	tRF-19-6998LOJX	463
tRF-25-SP5830MMUK	291	tRF-17-18YKISM	462
tRF-18-S5R83004	278	tRF-18-S5R83004	458
tRF-19-R118LOJX	278	tRF-21-RXF4P2PS0	458
tRF-19-6998LOJX	269	tRF-16-MBQ4NKD	430
SRR3502967 dataset		**SRR3502969 dataset**	
tRF	**Mean proportion of genes for each of the 13 processes**	**tRF**	**Mean proportion of genes for each of the 13 processes**
tRF-18-SP5830D4	38.7%	tRF-18-P4R8YP04	48.7%
tRF-16-SP5830D	37.5%	tRF-16-SP5830D	48.1%
tRF-18-P4R8YP04	36.4%	tRF-18-SP5830D4	47.3%
tRF-21-RXF4P2PS0	34.6%	tRF-18-S5R83004	44.4%
tRF-25-SP58309MUK	34.2%	tRF-19-6998LOJX	44.2%
tRF-17-18YKISM	33.7%	tRF-25-SP58309MUK	43.9%
tRF-18-S5R83004	32.8%	tRF-17-18YKISM	42.6%
tRF-25-SP5830MMUK	32.4%	tRF-25-SP5830MMUK	41.4%
tRF-19-R118LOJX	32.0%	tRF-21-RXF4P2PS0	41.2%
tRF-18-H9R8B7D2	31.3%	tRF-19-R118LOJX	41.0%

Evidently, there is significant overlap of the top 10 tRFs between the two 22Rv1 datasets. No single tRF appears to dominate the regulation of the 13 enriched processes, but rather a number of tRFs with significant contribution in the regulation of all the enriched processes. In order to investigate if there is a tRF that dominates the regulation of a single or a few of the 13 enriched processes the proportion of genes regulated by each tRF for each individual process was calculated. The top 10 tRFs in terms of this proportion are shown in **Table 6**.

Table 6. The top-10 tRFs in terms of gene proportion for each process.

SRR3502967 dataset		
tRF	**Proportion of genes for the process**	**Biological process**
tRF-16-SP5830D	46.3%	viral gene expression
tRF-16-SP5830D	44.9%	viral transcription
tRF-18-P4R8YP04	43.6%	regulation of mRNA metabolic process
tRF-18-SP5830D4	43.0%	viral gene expression
tRF-18-SP5830D4	41.3%	viral transcription
tRF-25-SP58309MUK	40.9%	proteasome-mediated ubiquitin-dependent protein catabolic process
tRF-25-SP58309MUK	40.5%	proteasomal protein catabolic process
tRF-21-RXF4P2PS0	40.3%	regulation of mRNA metabolic process
tRF-19-6998LOJX	39.8%	regulation of mRNA metabolic process
tRF-18-SP5830D4	39.8%	regulation of mRNA metabolic process
SRR3502969 dataset		
tRF	**Proportion of genes for the process**	**Biological process**
tRF-18-P4R8YP04	56.1%	regulation of mRNA metabolic process
tRF-16-SP5830D	53.7%	viral gene expression
tRF-19-6998LOJX	53.1%	regulation of mRNA metabolic process
tRF-18-SP5830D4	53.1%	regulation of mRNA metabolic process
tRF-16-SP5830D	52.3%	viral transcription
tRF-17-18YKISM	51.5%	regulation of mRNA metabolic process
tRF-16-SP5830D	51.5%	regulation of mRNA metabolic process
tRF-18-P4R8YP04	51.1%	RNA splicing, via transesterification reactions
tRF-18-P4R8YP04	51.0%	viral transcription
tRF-18-P4R8YP04	50.9%	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile

Although there is a relative increase in the proportion, and in some cases the proportion of genes regulated by a tRF for a single process is above 50%, again no single tRF exists that dominates the regulation of a single process, as these tRFs exhibit comparable proportions of regulated genes for the rest of the processes. Finally, to further expand this investigation, the proportion of genes regulated by each tRF for each individual biological process that can considered significantly enriched (adjusted p-value < 0.01) was calculated. The top-10 tRFs in terms of this proportion are shown in **Table 7**.

Table 7. The top-10 tRFs in terms of gene proportion for all processes with adjusted p-value < 0.01.

SRR3502967 dataset					
tRF	**Proportion of genes for the process**	**Biological process**	**Adjusted p-value**	**Number of genes of the process in the dataset**	**Number of genes included in the process**
tRF-25-SP5830MMUK	87.5%	positive regulation of oxidative phosphorylation	0.0042	8	10
tRF-16-MBQ4NKD	80.0%	protein localization to cell-cell junction	0.0016	10	13
tRF-17-18YKISM	78.6%	positive regulation by host of viral transcription	0.000028	14	17
tRF-19-R118LOJX	77.8%	chromatin-mediated maintenance of transcription	0.0093	9	13
tRF-16-MBQ4NKD	75.0%	mitotic nuclear envelope reassembly	0.0042	8	10
tRF-18-H9R8B7D2	75.0%	mitotic nuclear envelope reassembly	0.0042	8	10
tRF-16-KPM43RB	75.0%	regulation of cyclic-nucleotide phosphodiesterase activity	0.0042	8	10
tRF-17-18YKISM	75.0%	regulation of cyclic-nucleotide phosphodiesterase activity	0.0042	8	10
tRF-19-6998LOJX	75.0%	positive regulation of oxidative phosphorylation	0.0042	8	10
tRF-16-SP5830D	75.0%	positive regulation of oxidative phosphorylation	0.0042	8	10
SRR3502969 dataset					
tRF	**Proportion of genes for the process**	**Biological process**	**Adjusted p-value**	**Number of genes of the process in the dataset**	**Number of genes included in the process**
tRF-19-R118LOJX	90.9%	regulation of nuclear-transcribed mRNA poly(A) tail shortening	0.0087	11	14
tRF-18-S5R83004	90.9%	RNA decapping	0.0087	11	14
tRF-18-S5R83004	90.9%	methylguanosine-cap decapping	0.0087	11	14
tRF-16-SP5830D	88.9%	mitotic nuclear envelope reassembly	0.0044	9	10
tRF-25-SP5830MMUK	88.9%	mitotic nuclear envelope reassembly	0.0044	9	10
tRF-16-MBQ4NKD	88.9%	mitotic nuclear envelope reassembly	0.0044	9	10
tRF-18-P4R8YP04	86.7%	mitotic chromosome condensation	0.000029	15	16
tRF-18-SP5830D4	83.3%	SREBP signaling pathway	0.0014	12	14
tRF-18-S5R83004	82.4%	P-body assembly	0.00030	17	21
tRF-17-18YKISM	82.4%	stress granule assembly	0.0018	17	23

In this case it is notable that there is a further increase in the proportion of genes regulated by a tRF for a single process, and in some cases the proportion is above 90%, but all these processes include a relative small number of genes and this subsequently results in much larger adjusted p-value (compared to adjusted the p-value of the 13 extremely enriched processes that lies in the range of 10^{-30} to 10^{-50}), and whether the enrichment of these processes is really significant should be further investigated.

4.CONCLUSIONS AND FUTURE WORK

tRNA-derived fragments (tRFs) are short (13-32nt) RNAs that are produced through endonucleolytic cleavage of mature and precursor tRNAs, and are classified into into four main types based on the region of the pre-tRNA or mature tRNA they originate from (5'-, 3'-, i-, 3'-U-tRFs. Despite their deep conservation and universal presence in almost every branch of life, tRfs were initially considered as random degradation products, but increasing evidence indicates that tRFs are functional molecules and more specifically that are loaded into AGO proteins and guide RISC-dependent post-transcriptional repression, in the same manner that microRNAs do.

In this thesis we quantified the expression of tRF species from small RNA-Seq (sRNA-Seq) datasets for two prostate cancer cell lines (i.e., 22Rv1 and DU145) and 10 prostate cancer samples from TCGA. Quantification was performed by incorporating the genomic coordinates of tRFs, extracted from MINTbase v2.0, to the workflow of Manatee. The top-100 expressed tRFs were subsequently utilized to guide the analysis of AGO-CLIP datasets for relevant cell-lines using the microCLIP tool, to identify genes interacting with those tRFs. The identified tRF-interacting genes were further investigated for enrichment for specific biological processes

Our results demonstrate the use, for the first time, of the Manatee workflow for the quantification of tRFs. One possible drawback is that the current Manatee workflow does not account for the post-transcriptional modifications of tRFs, and that may affect the accuracy of the quantification. The analysis of the tRF-gene interactions demonstrated that each of the top-expressed tRFs seems to interact with hundreds of genes and similarly each gene harbors hundreds of MREs. Interestingly the enrichment analysis for the tRF-interacting genes highlighted a set of 13 biological processes that are extremely enriched (adjusted p-value < 10^{-30}) in two of 22Rv1 datasets, an enrichment that is retained even if we filter for subsets of the MREs. Further analysis revealed that no single tRF appears to dominate the regulation of the 13 enriched processes, but rather a combination of tRFs with significant contribution in the regulation of all the enriched processes.

Regarding future plans, one first step would be to expand the analysis to a larger number of datasets, including more cell lines. A second step would be to try and incorporate miRNAs into the sRNA-gene interaction analysis and the subsequent functional analysis.

ABBREVIATIONS - ACRONYMS

tRNA	transfer-RNA
mRNA	messenger-RNA
tRFs	tRNA-derived fragments
miRNAs	micro-RNAs
AGO	Argonaute proteins
RISC	RNA-Induced Silencing Complex
sRNA-Seq	small RNA sequencing
AGO-CLIP	AGO-Cross-Linking ImmunoPrecipitation
MRE	microRNA Response Element
GO	Gene Ontology

REFERENCES

[1] Mattick JS, Makunin IV. Non-coding RNA. *Hum Mol Genet.* 2006 Apr 15;15 Spec No 1:R17-29. doi: 10.1093/hmg/ddl046.

[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 Dec 3;75(5):843-54. doi: 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 Dec 3;75(5):855-62. doi: 10.1016/0092-8674(93)90530-4.

[4] Nestler, E.J., and Hyman, S.E. Regulation of gene expression. *Neuro-psychopharmacology: The Fifth Generation of Progress*, Philadelphia, Penn, 2002 .

[5] Bhattacharjee S, Renganaath K, Mehrotra R, Mehrotra S. Combinatorial control of gene expression. *Biomed Res Int.* 2013;2013:407263. doi: 10.1155/2013/407263.

[6] Carvalho Barbosa C, Calhoun SH, Wieden HJ. Non-coding RNAs: what are we missing? *Biochem Cell Biol.* 2020 Feb;98(1):23-30. doi: 10.1139/bcb-2019-0037.

[7] Kapranov P *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science.* 2007 Jun 8;316(5830):1484-8. doi: 10.1126/science.1138341.

[8] Djebali S *et al.* Landscape of transcription in human cells. *Nature.* 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233.

[9] Hombach S, Kretz M. Non-coding RNAs: Classification, Biology and Functioning. *Adv Exp Med Biol.* 2016;937:3-17. doi: 10.1007/978-3-319-42059-2_1.

[10] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet.* 2009 Mar;10(3):155-9. doi: 10.1038/nrg2521.

[11] Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet.* 2019 Feb;20(2):89-108. doi: 10.1038/s41576-018-0073-3.

[12] Mattick JS, Makunin IV. Small regulatory RNAs in mammals. *Hum Mol Genet.* 2005 Apr 15;14 Spec No 1:R121-32. doi: 10.1093/hmg/ddi101.

[13] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* 2004 Jan 23;116(2):281-97. doi: 10.1016/s0092-8674(04)00045-5.

[14] Saliminejad K, Khorram Khorshid HR, Soleymani Fard S, Ghaffari SH. An overview of microRNAs: Biology, functions, therapeutics, and analysis methods. *J Cell Physiol.* 2019 May;234(5):5451-5465. doi: 10.1002/jcp.27486.

[15] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D155-D162. doi: 10.1093/nar/gky1141.

[16] Ha M, Kim VN. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol.* 2014 Aug;15(8):509-24. doi: 10.1038/nrm3838.

[17] Fabian MR, Sonenberg N. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat Struct Mol Biol.* 2012 Jun 5;19(6):586-93. doi: 10.1038/nsmb.2296.

[18] Iwakawa HO, Tomari Y. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Mol Cell.* 2022 Jan 6;82(1):30-43. doi: 10.1016/j.molcel.2021.11.026.

[19] Mayya VK *et al*. microRNA-mediated translation repression through GYF-1 and IFE-4 in C. elegans development. *Nucleic Acids Res.* 2021 May 21;49(9):4803-4815. doi: 10.1093/nar/gkab162.

[20] Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet.* 2011 Nov 18;12(12):861-74. doi: 10.1038/nrg3074.

[21] Kumar P, Kuscu C, Dutta A. Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). *Trends Biochem Sci.* 2016 Aug;41(8):679-689. doi: 10.1016/j.tibs.2016.05.004.

[22] Fu H *et al*. Stress induces tRNA cleavage by angiogenin in mammalian cells. *FEBS Lett.* 2009 Jan 22;583(2):437-42. doi: 10.1016/j.febslet.2008.12.043.

[23] Telonis AG *et al.* Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget.* 2015 Sep 22;6(28):24797-822. doi: 10.18632/oncotarget.4695.

[24] Keam SP, Hutvagner G. tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. *Life (Basel).* 2015 Nov 27;5(4):1638-51. doi: 10.3390/life5041638.

[25] Gebetsberger J, Zywicki M, Künzi A, Polacek N. tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in Haloferax volcanii. *Archaea.* 2012;2012:260909. doi: 10.1155/2012/260909.

[26] Sobala A, Hutvagner G. Small RNAs derived from the 5' end of tRNA can inhibit protein translation in human cells. *RNA Biol.* 2013 Apr;10(4):553-63. doi: 10.4161/rna.24285.

[27] Maute RL *et al.* tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. *Proc Natl Acad Sci U S A.* 2013 Jan 22;110(4):1404-9. doi: 10.1073/pnas.1206761110.

[28] Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* 2009 Nov 15;23(22):2639-49. doi: 10.1101/gad.1837609.

[29] Couvillion MT, Bounova G, Purdom E, Speed TP, Collins K. A Tetrahymena Piwi bound to mature tRNA 3' fragments activates the exonuclease Xrn2 for RNA processing in the nucleus. *Mol Cell.* 2012 Nov 30;48(4):509-20. doi: 10.1016/j.molcel.2012.09.010.

[30] Goodarzi H, Liu X, Nguyen HC, Zhang S, Fish L, Tavazoie SF. Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. *Cell.* 2015 May 7;161(4):790-802. doi: 10.1016/j.cell.2015.02.053.

[31] Kuscu C, Kumar P, Kiran M, Su Z, Malik A, Dutta A. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA.* 2018 Aug;24(8):1093-1105. doi: 10.1261/rna.066126.118.

[32] Shigematsu M, Kirino Y. tRNA-Derived Short Non-coding RNA as Interacting Partners of Argonaute Proteins. *Gene Regul Syst Bio.* 2015 Sep 10;9:27-33. doi: 10.4137/GRSB.S29411.

[33] Yeung ML, Bennasser Y, Watashi K, Le SY, Houzet L, Jeang KT. Pyrosequencing of small non-coding RNAs in HIV-1 infected cells: evidence for the processing of a viral-cellular double-stranded RNA hybrid. *Nucleic Acids Res.* 2009 Oct;37(19):6575-86. doi: 10.1093/nar/gkp707.

[34] Wang Q, Lee I, Ren J, Ajay SS, Lee YS, Bao X. Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection. *Mol Ther.* 2013 Feb;21(2):368-79. doi: 10.1038/mt.2012.237.

[35] Haussecker D, Huang Y, Lau A, Parameswaran P, Fire AZ, Kay MA. Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA.* 2010 Apr;16(4):673-95. doi: 10.1261/rna.2000810.

[36] Rounge TB, Furu K, Skotheim RI, Haugen TB, Grotmol T, Enerly E. Profiling of the small RNA populations in human testicular germ cell tumors shows global loss of piRNAs. *Mol Cancer.* 2015 Aug 12;14:153. doi: 10.1186/s12943-015-0411-4.

[37] Pliatsika V *et al.* MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D152-D159. doi: 10.1093/nar/gkx1075.

[38] Loher P, Telonis AG, Rigoutsos I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Sci Rep.* 2017 Feb 21;7:41184. doi: 10.1038/srep41184.

[39] Handzlik JE, Tastsoglou S, Vlachos IS, Hatzigeorgiou AG. Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data. *Sci Rep.* 2020 Jan 20;10(1):705. doi: 10.1038/s41598-020-57495-9.

[40] Hafner M *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell.* 2010 Apr 2;141(1):129-41. doi: 10.1016/j.cell.2010.03.009.

[41] Paraskevopoulou MD, Karagkouni D, Vlachos IS, Tastsoglou S, Hatzigeorgiou AG. microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. *Nat Commun.* 2018 Sep 6;9(1):3601. doi: 10.1038/s41467-018-06046-y.