ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ - ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΔΙΔΑΚΤΟΡΙΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αλγόριθμοι για την Υπολογιστική Ανάλυση της Λειτουργίας των Μη Κωδικών Μεταγράφων

Μαρία Δ Παρασκευοπούλου

Επιβλέπουσα: **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια

ΒΟΛΟΣ

ΙΟΥΝΙΟΣ 2016

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αλγόριθμοι για την Υπολογιστική Ανάλυση της Λειτουργίας των Μη Κωδικών Μεταγράφων

Μαρία Δ Παρασκευοπούλου

Α.Μ.: 399

ΕΠΙΒΛΕΠΟΥΣΑ: **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Άρτεμις Χατζηγεωργίου, Καθηγήτρια
Αντιγόνη Λάζου, Καθηγήτρια
Ιωάννης Τσαμαρδίνος, Επίκουρος Καθηγητής

Ιούνιος 2016

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Χατζηγεωργίου Άρτεμις, Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας (ΕΠΙΒΛΕΠΟΥΣΑ)

Λάζου Αντιγόνη, Καθηγήτρια, Τμήμα Βιολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Τσαμαρδίνος Ιωάννης, Αναπληρωτής Καθηγητής, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Βάβαλης Εμμανουήλ, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Σταμούλης Γεώργιος, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ποταμιάνος Γεράσιμος, Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Τσομπανοπούλου Παναγιώτα, Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

ABSTRACT

The RNA revolution has turned non-coding RNA (ncRNA) from dark-matter into a biological research hotspot. Accumulating evidence from multiple Next Generation Sequencing (NGS) experiments has recently introduced the regulatory roles of ncRNAs in a wide range of biological processes. This thesis focuses on the development of computational algorithms for the functional characterization of non-coding transcripts, while investigating in-depth their in-between interactions. The methodologies developed during this thesis combine advanced next-generation sequencing (NGS) data analyses and state-of-the-art Machine Learning algorithms in order to perform automated analyses and to monitor the corresponding results.

This doctoral thesis studies specific categories of RNA transcripts: microRNAs (miRNAs) and long non-coding RNAs (lncRNAs). miRNAs are single stranded RNA molecules approximately 22 nucleotides long. They have been deemed central post-transcriptional gene regulators and play a key role in numerous biological processes. Therefore, miRNAs are intensively studied for their potential as biomarkers and/or therapeutic targets. Apart from their involvement in physiological processes, microRNAs appear to be associated with a plethora of pathological conditions.

Although microRNAs are mainly considered mRNA repressors, there are studies supporting miRNA-lncRNA interactions. lncRNAs are long non-coding transcripts that can also regulate gene expression. To this end, DIANA-LncBase database was designed in order to characterize the entire spectrum of miRNA interactions with lncRNAs. LncBase supports a compendium of experimentally supported miRNA-lncRNA interactions. It contains more than 70,000 interactions derived from the analysis of numerous NGS experiments and specific low-throughput techniques, across 66 different types spanning 36 tissues in human and mouse species. DIANA-TarBase update was also part of the thesis. TarBase v7 is considered the largest available repository of miRNA-mRNA interactions as compared to any of the relevant databases. It hosts more than half a million interactions from published experiments on 356 different cell types (59 tissues), belonging to 24 species. The detailed cataloguing of RNA interactions unveiled a set of approximately 400 unique viral-miRNA:lncRNA interacting pairs in human virus-infected cells. This type of regulation adds an extra layer of complexity in the miRNA interactome, and perplexes the network with the inclusion of virus-encoded and human transcript interactions.

By analyzing more than 150 raw CLIP-Seq datasets, DIANA-TarBase v7.0 and DIANA-LncBase are the first relevant databases to provide an unprecedented amount of experimentally supported interactions in many different cell types and tissues. Furthermore, RNA sequencing data were analyzed to accurately assess miRNA and transcript expression in the investigated cell types. Optimized pipelines were developed for the analysis of sequencing data, while a machine learning approach has been applied for the identification of miRNA binding sites.

The adopted methodology for AGO-CLIP-Seq data analysis was compared against other available state-of-the-art implementations and has been proven robust and advantageous.

During the course of the Doctoral thesis, the continuous archiving of experimental data from low and high-throughput methodologies, along with extensive evaluation of the available AGO-CLIP-Seq analysis programs, revealed that there was room for algorithms' further improvement and optimization. State-of-the-art CLIP-guided target identification implementations currently manage to identify approximately half of the experimentally validated binding sites. To this end, a novel algorithm was developed for CLIP-Seq data analysis. The algorithm was trained and extensively tested on a comprehensive collection of accurate positive and negative miRNA-target interactions from numerous experimental data sources. It was additionally evaluated against all leading implementations, including CLIP-Seq analysis adopted by TarBase/LncBase. The results depict that the new approach not only significantly outperforms other implementations in terms of accuracy but also manages to increase sensitivity, predicting sites that were not detected by any other algorithm.

The functional significance of miRNA interactions with coding and non-coding transcripts was further assessed with the evolutionary conservation of the miRNA binding sites. The thesis additionally associates the catalogued interactions to diseases and molecular pathways, providing new insights in ncRNA function.

DIANA-microT web server was upgraded and enhanced with automated analyses pipelines (workflows) that can be applied to NGS-derived data. The ready-to-use modules seamlessly integrate DIANA supported algorithms for the identification of miRNA-gene interactions and miRNA-targeted pathway analyses.

During the course of the Doctoral thesis, the candidate took part in 8 scientific studies involving computational approaches for determining the activity of the non-coding transcripts and in four of them the candidate is first author. The studies are published in international peer-reviewed scientific journals, while the total citations received to date are 310.

SUBJECT AREA: Computational Biology

KEYWORDS: microRNA, lncRNA, HITS-CLIP, PAR-CLIP, target prediction, experimentally verified targets

ΠΕΡΙΛΗΨΗ

Η επανάσταση του RNA μετέφερε τα μη κωδικά μετάγραφα (non-coding RNAs) στο επίκεντρο της βιολογικής έρευνας. Τα τελευταία χρόνια, στοιχεία από πολυάριθμα πειράματα αποκαλύπτουν πολλαπλούς ρυθμιστικούς ρόλους των μη κωδικών μεταγράφων στο γονιδίωμα σε ένα ευρύ φάσμα βιολογικών διεργασιών. Η εργασία αυτή εστιάζει στην ανάπτυξη αλγορίθμων για την κατανόηση της λειτουργίας μη κωδικών μορίων και διερευνά εκτενώς τις αλληλεπιδράσεις μεταξύ ομάδων κωδικών και μη κωδικών μεταγράφων. Οι μεθοδολογίες που αναπτύχθηκαν κατά τη διάρκεια της διδακτορικής διατριβής συνδύασαν προηγμένες αναλύσεις δεδομένων αλληλούχησης επόμενης γενεάς και συμπεριέλαβαν αλγορίθμους αιχμής Μηχανικής Μάθησης, για την πραγματοποίηση αυτόματων αναλύσεων καθώς και εποπτείας των αντίστοιχων αποτελεσμάτων.

Η εργασία εστιάζει στη μελέτη ειδικών κατηγοριών μορίων: τα microRNAs (miRNAs) και τα long non-coding RNAs (lncRNAs). Τα miRNAs είναι μονόκλωνα μόρια RNA μήκους περίπου 22 νουκλεοτιδίων. Θεωρούνται βασικοί μετα-μεταγραφικοί ρυθμιστές της έκφρασης των γονιδίων και διαδραματίζουν καθοριστικό ρόλο σε πληθώρα βιολογικών διαδικασιών. Αποτελούν αντικείμενο έντονης μελέτης τα τελευταία χρόνια για τη δυναμική τους ως πιθανοί θεραπευτικοί στόχοι καθώς πέρα από το ρόλο τους σε φυσιολογικές διεργασίες, εμφανίζονται να εμπλέκονται σε ένα ευρύ φάσμα παθολογικών καταστάσεων. Βάσει τελευταίων ερευνών, τα miRNAs στοχεύουν και άλλα μη κωδικά RNAs, τα lncRNAs. Τα lncRNAs είναι μακρά μη κωδικά μετάγραφα και μέρος αυτών σχετίζεται με την ρύθμιση της γονιδιακής έκφρασης.

Προκειμένου να χαρακτηριστεί ολόκληρο το φάσμα των αλληλεπιδράσεων των miRNAs με lncRNAs, σχεδιάστηκε η βάση δεδομένων DIANA-LncBase που υποστηρίζει τον μεγαλύτερο κατάλογο πειραματικά επιβεβαιωμένων miRNA-lncRNA αλληλεπιδράσεων. Περιέχει πάνω από 70.000 αλληλεπιδράσεις από πληθώρα πειραμάτων αλληλούχησης επόμενης γενεάς και ειδικές τεχνικές μικρής διεκπεραιωτικής ικανότητας σε 66 διαφορετικούς τύπους κυττάρων, που εκτείνονται σε 36 ιστούς του ανθρώπου και του μυός. Στη παρούσα διατριβή ανανεώθηκε και η βάση δεδομένων DIANA-TarBase, η βάση με τον εκτενέστερο κατάλογο πειραματικά επιβεβαιωμένων αλληλεπιδράσεων μεταξύ μικρών RNA και κωδικών γονιδίων στόχων παγκοσμίως. Περιέχει περισσότερες από 500.000 αλληλεπιδράσεις από 28 διάφορες πειραματικές μεθοδολογίες, καλύπτοντας 356 κυτταρικούς τύπους και 59 διαφορετικούς ιστούς. Κατά τη λεπτομερή καταγραφή του χάρτη των αλληλεπιδράσεων των μορίων στο επίπεδο του RNA σημειώθηκαν για πρώτη φορά και αλληλεπιδράσεις των μικρών RNAs που παράγονται από ιούς με τα μακρά μη κωδικά μετάγραφα του ανθρώπου. Η αναγνώριση τέτοιων αλληλεπιδράσεων έγινε σε ανθρώπινες κυτταρικές σειρές που έχουν προσβληθεί από κάποιο στέλεχος ιού. Αυτά τα δεδομένα βάζουν ένα ακόμη επίπεδο πολυπλοκότητας στις αλληλεπιδράσεις των μη κωδικών μορίων, καθώς χρειάζεται να μελετηθούν και αυτές μεταξύ των μεταγράφων του ιού και του ανθρώπου.

Τα δεδομένα NGS που αναλύθηκαν, για την ανεύρεση στόχων των microRNAs με τα (μη)κωδικά μετάγραφα για το σχηματισμό των βάσεων LncBase και TarBase, περιλαμβάνουν πάνω από 150 βιβλιοθήκες CLIP-Seq. Παράλληλα, συλλέχθηκαν και αναλύθηκαν δεδομένα αλληλούχησης για την έκφραση των microRNA και των μεταγράφων στα κύτταρα όπου πραγματοποιήθηκαν τα CLIP-Seq πειράματα. Αναπτύχθηκαν αλγόριθμοι για την ανάλυση των δεδομένων αλληλούχησης, ενώ ο εντοπισμός των αναγνωριστικών θέσεων πρόσδεσης των microRNAs στα μετάγραφα έγινε με μηχανική μάθηση. Η μεθοδολογία που υιοθετήθηκε συγκρίθηκε με αντίστοιχους αλγορίθμους αιχμής, ενώ εμφάνισε πληθώρα πλεονεκτημάτων σε κάθε σύγκριση.

Κατά τη διάρκεια της διδακτορικής διατριβής, η συνεχής αρχειοθέτηση και ανάλυση πειραματικών δεδομένων από χαμηλής και υψηλής διεκπεραιωτικής ικανότητας μεθοδολογίες, μαζί με την εκτενή αξιολόγηση των διαθέσιμων CLIP-Seq προγραμμάτων, αποκάλυψε ότι υπήρχε περιθώριο για περαιτέρω βελτίωση. Οι διαθέσιμοι αλγόριθμοι αιχμής που εντοπίζουν στόχους των miRNAs μέσα από την ανάλυση CLIP-Seq δεδομένων επιτυγχάνουν την ορθή αναγνώριση σε περίπου μισές πειραματικά επικυρωμένες αλληλεπιδράσεις. Για το σκοπό αυτό, αναπτύχθηκε ένας καινοτόμος αλγόριθμος για την ανάλυση AGO-CLIP-Seq δεδομένων. Ο αλγόριθμος εκπαιδεύτηκε και δοκιμάστηκε εκτενώς σε μια υψηλής ποιότητας, ολοκληρωμένη συλλογή θετικών και αρνητικών αλληλεπιδράσεων των miRNAs με γονίδια βάσει πολυάριθμων πειραματικών δεδομένων. Επιπλέον αξιολογήθηκε έναντι παρόμοιων εφαρμογών αιχμής, συμπεριλαμβανομένου και του αλγορίθμου ανάλυσης CLIP-Seq δεδομένων των TarBase / LncBase. Τα αποτελέσματα παρουσίασαν ότι η νέα αλγοριθμική προσέγγιση ξεπερνά σημαντικά τις άλλες εφαρμογές όχι μόνο όσον αφορά την ακρίβεια, αλλά παράλληλα καταφέρνει να αυξήσει την ευαισθησία μέσω της πρόβλεψης περιοχών πρόσδεσης των μικρών RNA που δεν είχαν εντοπιστεί από οποιοδήποτε άλλο αλγόριθμο.

Παράλληλα, η λειτουργική σημασία των αλληλεπιδράσεων των miRNAs με τις διάφορες κατηγορίες μεταγράφων μελετήθηκε μέσω της διερεύνησης της εξελικτικής συντήρησης των περιοχών πρόσδεσης σε κωδικές και μη κωδικές ακολουθίες. Η διδακτορική διατριβή περιλαμβάνει και τη μελέτη του χάρτη των αλληλεπιδράσεων των μορίων στο επίπεδο του RNA σε σχέση με ασθένειες και μοριακά μονοπάτια, γεγονός που θα βοηθήσει να προσδιοριστούν άγνωστες μέχρι τώρα πτυχές της δράσης των μικρών RNAs. Παράλληλα, αναβαθμίστηκαν και σχηματίστηκαν νέες λειτουργικότητες για τον εξυπηρετητή του DIANA-microT και πραγματοποιήθηκε η δημιουργία αυτόματων ροών ανάλυσης (workflows), δεδομένων που προκύπτουν από πειράματα NGS. Οι έτοιμες αναλύσεις διασυνδέουν εργαλεία του DIANA που αφορούν αλληλεπιδράσεις των μικρών RNAs με γονίδια και την εμπλοκή τους σε μοριακά μονοπάτια.

Κατά τη διάρκεια της διδακτορικής διατριβής, η υποψήφια έλαβε μέρος σε 8 επιστημονικές μελέτες που περιλαμβάνουν υπολογιστικές προσεγγίσεις για τον

προσδιορισμό της δράσης των μη κωδικών μεταγράφων, και σε τέσσερις από αυτές είναι πρώτη συγγραφέας. Οι μελέτες δημοσιεύτηκαν σε διεθνή έγκριτα περιοδικά και οι συνολικές αναφορές που έχουν λάβει έως τώρα είναι 310.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: microRNA, lncRNA, HITS-CLIP, PAR-CLIP, πρόβλεψη στόχων, πειραματικά επιβεβαιωμένοι στόχοι

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό θα ήθελα να εκφράσω τις ευχαριστίες μου προς τα μέλη της Τριμελούς Συμβουλευτικής Επιτροπής για τη δυνατότητα που μου έδωσαν να ασχοληθώ με ένα διεπιστημονικό θέμα μεγάλης σημασίας και εμβέλειας στο τομέα της Βιολογίας και της Πληροφορικής.

Θα ήθελα να ευχαριστήσω την επιβλέπουσά μου Καθηγήτρια κ. Άρτεμις Χατζηγεωργίου, για την άψογη καθοδήγηση και υποστήριξη που μου παρείχε κατά τη διάρκεια εκπόνησης της εργασίας. Την ευχαριστώ θερμά για την τεχνογνωσία που μου προσέφερε ως μαθητευόμενή της στις μεταπτυχιακές και διδακτορικές μου σπουδές, καθώς και για το ότι με συμπεριέλαβε στην ερευνητική ομάδα του εργαστηρίου DIANA του οποίου και είναι υπεύθυνη. Θα ήθελα να εκφράσω την ευγνωμοσύνη μου για τη δυνατότητα που μου έδωσε να ασχοληθώ με μελέτες αιχμής και να συνεργαστώ με εξαιρετικούς συναδέλφους και επιστήμονες. Ακόμη, την ευχαριστώ γιατί υπήρξε βασικός πυλώνας για την εξέλιξή μου ως επιστήμονα και ως άνθρωπο.

Επίσης, οφείλω να ευχαριστήσω την Καθηγήτρια κ. Λάζου Αντιγόνη και τον Αναπληρωτή Καθηγητή Τσαμαρδίνο Ιωάννη για την τιμή που έκαναν να είναι μέλη της Τριμελούς Συμβουλευτικής Επιτροπής, καθώς και για τη συμβολή τους στην εκπόνηση της διδακτορικής διατριβής.

Η παρούσα εργασία δε θα μπορούσε να έχει υλοποιηθεί χωρίς την αγαστή συνεργασία με τα μέλη του εργαστηρίου DIANA.

Θα ήθελα να ευχαριστήσω θερμά τον συνάδελφό μου Δρ. Ιωάννη Βλάχο για την άψογη από κοινού εργασία για τη δημιουργία της βάσης του TarBase, καθώς και για την καθοριστική συμβολή του στη δημιουργία των βάσεων microT-CDS και LncBase. Τον ευχαριστώ για τις πολύτιμες συμβουλές του για την πραγματοποίηση της διατριβής και τη συνολική στήριξη καθ' όλη τη διάρκεια της συνεργασίας μας.

Ευχαριστώ θερμά την Καραγκούνη Δήμητρα για την εξαιρετική συνεργασία που είχαμε και την αμέριστη συμπαράσταση που μου παρείχε. Η συμβολή της ήταν καθοριστική στη συλλογή και ανάλυση δεδομένων για τις βάσεις TarBase και LncBase. Ακόμη αποτελεί ένα βασικό συνεργάτη, μαζί με τον Δρ. Ιωάννη Βλάχο στον σχεδιασμό ενός καινοτόμου αλγορίθμου για την αναγνώριση στόχων των μικρών RNAs μέσα από την ανάλυση δεδομένων υψηλής διεκπεραιωτικής ικανότητας.

Ευχαριστώ τον συνάδελφο μου Δρ. Γεώργιο Γεωργακίλα για την άριστη συνεργασία μας και τη πολύτιμη βοήθειά του στην ολοκλήρωση αρκετών μελετών, τη συμβολή του στο κομμάτι της πρόβλεψης των στόχων των miRNAs για τις βάσεις microT-CDS και LncBase, και τη συλλογή δεδομένων για τη βάση του TarBase.

Ευχαριστώ ιδιαίτερα τον Σπύρο Ταστσόγλου για τη συνεισφορά του στη διατριβή, καθώς πραγματοποίησε την ανάλυση πολλών πειραματικών δεδομένων. Ευχαριστώ επίσης τους συναδέλφους Κωνσταντίνο Λιάκο και Νεφέλη Ζώττου.

Ακόμη ευχαριστώ τον Δρ. Martin Reczko για την συμβολή του στο κομμάτι της πρόβλεψης των στόχων των miRNAs, απαραίτητη για τη δημιουργία της βάσης microT-CDS.

Για την πραγμάτωση της παρούσας διατριβής και το σχεδιασμό της διεπαφής των βάσεων TarBase, LncBase και microT-CDS καθοριστικό ρόλο έπαιξε η συνεργασία μας με το Ινστιτούτο «Αθηνά». Ευχαριστώ όλους τους συνεργάτες μας για την τεχνογνωσία και τη συμβολή τους στην ανάπτυξη και στο σχεδιασμό των βάσεων. Ειδικότερα ευχαριστώ τους Δρ. Θεόδωρο Δαλαμάγκα και Δρ. Θανάση Βεργούλη για τη σημαντική συμβολή τους σε όλες τις παραπάνω μελέτες.

Η συνεχής χρήση εκτενών υπολογιστικών πόρων και υπερυπολογιστικών συστημάτων ήταν απαραίτητη για την εκπόνηση της παρούσας διατριβής. Θα ήθελα να ευχαριστήσω τον Δρ. Φεύγα και το Υπολογιστικό Κέντρο του Τμήματος Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας για την παροχή των εξυπηρετητών που χρησιμοποιήθηκαν για την ανάλυση σημαντικού μέρους των δεδομένων της μελέτης. Θα ήθελα επίσης να ευχαριστήσω τους υπευθύνους και το προσωπικό του Εθνικού Δικτύου Έρευνας και Τεχνολογίας (ΕΔΕΤ) για την παροχή πρόσβασης στο υπερυπολογιστικό σύστημα "ARIS", η χρήση του οποίου ήταν καθοριστική για την επιτυχία της παρούσης ερευνητικής προσπάθειας.

Κλείνοντας, θα ήθελα να ευχαριστήσω από καρδιάς την οικογένειά μου και τους φίλους, που με στηρίζουν και μου συμπαραστέκονται σε κάθε μου προσπάθεια. Θέλω να τονίσω πως τίποτε από όσα έχω καταφέρει δε θα ήταν εφικτό χωρίς τη δική τους συμβολή.

Η εργασία αυτή αφιερώνεται στην οικογένειά μου,
και σε δύο αγαπημένους φίλους και συνοδοιπόρους

CONTENTS

List of Figures

List of Tables

1. INTRODUCTION

1.1 ncRNAs

The traditional view of molecular biology argued that the primary and almost exclusive role of RNA is to carry genetic information in order to be subsequently translated into protein. However, the discovery of functional non-coding transcripts other than those participating in the translational machinery (ribosomal RNAs and tRNAs) broadened the long-established RNA role and revised the "central dogma". Non-coding RNAs (ncRNAs), although initially considered as "junk", have been deemed as important key regulators in various biological processes. A large percentage of the mammalian genomes and other complex organisms are transcribed into ncRNAs comprising a hidden layer of regulation in a plethora of physiological and pathological processes.

ncRNAs originate from different regulatory regions within the genomes and are characterized by high versatility. It has been observed that ncRNAs may derive from intragenic, intergenic, intronic regions of protein coding genes or even from transposons and pseudogenes (1). ncRNAs can be divided into many subcategories, such as, ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), long noncoding RNAs (lncRNAs), etc.

Currently, we can observe an unprecedented expansion of the so-called "regulatory RNA" field thanks to emerging new technological developments. Extensive sequencing experiments during the past decade and deep sequencing data produced by large consortia, including the Encyclopedia of DNA Elements Consortium (ENCODE) (2,3) revealed that the majority of the transcribed eukaryotic genomes corresponds to functional non-coding RNA elements, while only 3% of these regions produce protein coding transcripts. Numerous high-throughput experiments suggest that ncRNAs define the complexity of an organism and regulate numerous biological processes including splicing, editing, transcription, translation, various levels of gene expression, development and epigenetic mechanisms (4).

1.2 microRNAs

miRNAs are small noncoding RNAs (~22 nts) and are considered central post-transcriptional gene regulators, acting through transcript degradation, cleavage and/or translation suppression in the case of mRNAs (5). Since their first identification in 1993 (6), the number of annotated miRNAs and miRNA-related publications increased in a super linear rate, clearly depicting their central position in the RNA revolution (7). More than 21,000 miRNAs have been identified in various organisms, while their number in the human genome surpasses 2,500 (8).

The first microRNAs were discovered in 1993 in *Caenorhabditis elegans* (9) by Ambros, Lee and Feinbaum. The researchers observed that the lin-4 gene produced a non-coding RNA segment of about 22 bases long, that bound to the 3' untranslated end (3'-

UnTranslated Region, 3'-UTR) of lin-14 mRNA. The interaction between the lin-4 non-coding RNA and lin-14 led to translational repression of the latter. The above phenomenon was reinforced by another research study in *C. elegans*, where the miRNA let-7 was identified to target the 3'UTR region and induce suppression of lin-41 gene expression (10). Let-7 appeared to be conserved in other organisms, supporting the putative existence and regulatory role of other small non-coding RNA molecules (11). These early discoveries inaugurated the detection of large numbers of novel miRNA sequences in various organisms and also established their function as regulators of gene expression (12). Current studies indicate that more than half of human genes are regulated by miRNAs.

1.2.1 miRNA Biogenesis

Most miRNAs in mammals are transcribed by RNA polymerase II (RNA polymerase II, Pol II) (13), while few appear to be transcribed by RNA polymerase III (RNA polymerase III, Pol III) (14). At the same time a large number of transcription factors (TFs) associated with Pol II activity are taking part in the transcription process of miRNA genes (15). More than half of the miRNAs are derived from intragenic loci, embedded in protein coding introns, while ~45% originate from intergenic transcripts. The initially generated long primary miRNA transcripts are of thousand kilobases long (pri-miRNAs) and are 5′ capped, spliced and polyadenylated at the 3′ end.

The first stages of pri-miRNA transcript preprocessing are carried out in the cell nucleus. The pri-miRNAs form local stem-loop structures and usually contain at least one hairpin structure, termed as miRNA precursor sequence (precursor miRNA, pre-miRNA). In the primary maturation step, miRNA transcripts are cleaved by RNase III enzyme Drosha which processes pri-miRNAs into the ~60-100nt hairpin structure of the miRNA precursor (pre-miRNA) (16). The precursor sequences comprise several bulges and regions of imperfect complementarity. The rapid cleavage of pri-miRNAs by Drosha in the nucleus hinders their identification with conventional sequencing techniques.

During the pri-miRNA cleavage process, Drosha cooperates with DiGeorge syndrome Critical Region 8 (DGCR8) in humans and Pasha in *Drosophila melanogaster* and *C. elegans* (17-19). Protein DGCR8 and Drosha form the Microprocessor complex. Precursor sequences are subsequently exported from the nucleus to the cytoplasm, where their nuclear transport is accomplished by exportin-5 (20) and Ran-GTP. miRNA precursors are cleaved by Dicer enzyme in the cytoplasm, a highly conserved protein found in most eukaryotes. The produced double stranded mature transcripts are approximately 19-22nt long (21).

miRNA interacts with the RNA-induced silencing complex (RISC) to form the miRNA-induced silencing complex (miRISC). Both strands of the miRNA duplex-intermediate can be potentially functional. However, usually one strand (guide strand) accumulates as the mature miRNA and is loaded into the RISC complex along with a highly

conserved protein of the Argonaute (AGO) family. The other strand, termed as the "passenger" strand, is released and degraded. Perfect base pairing between the guide strand and the mRNA-target can lead to degradation (22), whereas in cases of imperfect complementarity miRNAs can direct gene silencing by translational repression, which is accompanied by degradation of mRNA in P-bodies (Processing bodies). Figure 1 summarizes the steps of microRNA biogenesis starting from the miRNA gene transcription to the mature miRNA function in the cytoplasm.

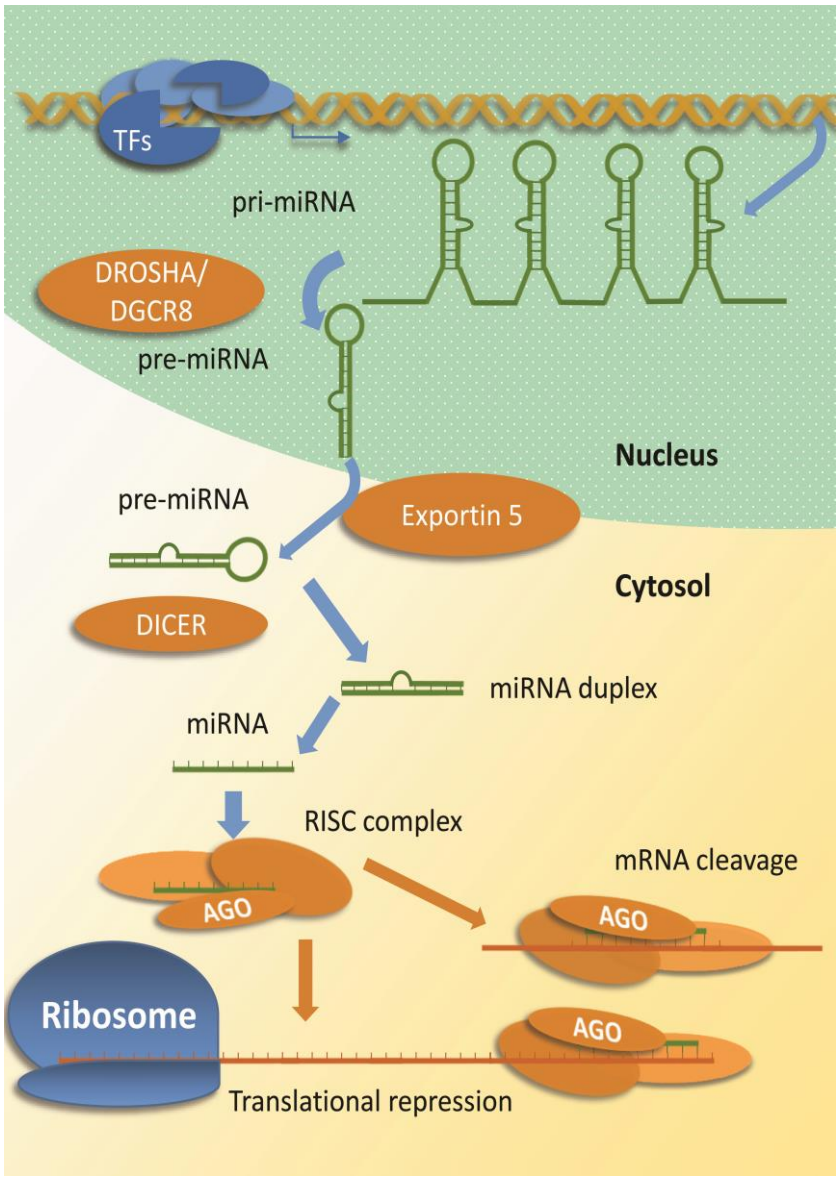

Figure 1: Summary of miRNA biogenesis. (1) miRNA gene transcription and formation of the pri-miRNA, (2) creation of pre-miRNA structures from cooperative Drosha-DGCR8 activity, (3) pre-miRNA nuclear export assisted by exportin 5 and Ran-GTP, (4) pre-miRNA is cleaved by Dicer enzyme to form the mature transcripts. (5) miRNAs loaded in the RISC complex post-transcriptionally regulate protein coding genes through mRNA cleavage, direct translational repression and/or mRNA destabilization in the cytoplasm.

1.2.2 miRNA Function

miRNAs are considered central post-transcriptional regulators of gene expression. As described in the previous section, mature miRNA sequences are incorporated in the RISC complex and induce gene silencing. They target genes usually by partial or complete base pairing with specific miRNA recognition elements (MREs) on the mRNA sequences (23). More precisely, miRNA and target gene interactions usually require 6-8 base-paired nucleotides, the so-called seed (24) at the 5' miRNA end. It should be noted that the degree of base pairing of the miRNA seed region with the mRNA plays a very important role in the efficiency of the interaction. miRNAs were primarily detected to effectively target specific mRNA 3′ untranslated regions (3′-UTRs), where highly conserved MREs exist (25). Recent findings showed *bona fide* miRNA interactions with MREs located also in 5′-UTR regions as well as within the coding sequence (CDS) (26).

miRNAs play a key role in numerous biological processes such as stem cell proliferation, division and differentiation, immunity, cell signaling, apoptosis and metabolism. Apart from their normal role, a large number of studies describe their implication in a vast array of diseases, such as cancer, viral infections, cardiovascular diseases, metabolic disorders, autoimmune pathologies, as well as neuropsychiatric pathological conditions. miRNAs can affect gene expression in various tissues. Therefore, possible changes in the concentration of miRNAs caused by epigenetic silencing or deregulated transcription factors, genetic disorders/abnormalities, deletion and amplification events can lead to the deregulation of their respective target genes (27-31). miRNAs are therefore intensively studied for their potential as therapeutic targets.

1.3 Identification of miRNA targets

One of the most important processes in miRNA research is the detection of their targets. Identification of miRNA-gene interactions can be performed with either computational approaches or experimental methodologies.

Accurate cataloguing of miRNA targets is crucial to the understanding of their function. To this end, numerous wet lab methodologies have been developed, enabling the validation of predicted miRNA interactions or the high-throughput screening and identification of novel miRNA targets (32). Currently available methodologies can elucidate different parts of the equation and are often used complementarily in investigative studies. On the other hand there are multiple programs based on simple to more sophisticated algorithms that perform target prediction.

Despite the contribution of both experimental methodologies and computational approaches, a large part of the miRNA targets, even for the well-studied organisms such as mouse and human, remains unexplored.

1.3.1 *In silico* approaches for the identification of de novo miRNA:mRNA interactions

In silico miRNA target identification is a crucial step in most miRNA-based experiments, since the miRNA interactome has not yet been adequately mapped, even for the most well-studied model organisms. Although the available experimental techniques are utilized to verify genuine targets, the first step in the analysis is the computational determination of miRNA-gene interactions. Early miRNA-related research efforts have highlighted the necessity of computational analyses in order to assist the experimental identification of miRNA targets. This has resulted to the development of numerous miRNA target prediction algorithms (33), which are now considered indispensable for the design of relevant experiments. These algorithms identify *in silico* miRNA targets as candidates for further experimentation or for computational processing, such as target enrichment analyses. Predictions of the available computational algorithms can be acquired from relevant miRNA:gene interaction databases or web servers (33,34).

The first target prediction program was developed in 2003, following the observation that miRNAs present high abundance in the cell, and since then more sophisticated implementations have been developed.

Significant nucleotides for the identification of binding sites are located at the 5' end of the mature miRNA sequence. Statistical analysis conducted by the group of Lewis and collaborators revealed certain highly conserved motifs across species in the 3'UTR region of mRNAs that match 2-7 positions of the miRNA 5’ end (35). These 6 nucleotides constitute the so-called seed region of the miRNA, which until now remains one of the most important features in target prediction (36). Other important features are considered the evolutionary conservation, dinucleotide base content and structural accessibility of the miRNA binding site as well as the base pairing stability (37).

Available algorithms can utilize diverse techniques and features, including machine learning, physics models, target site context and accessibility, pairing stability and conservation. These implementations often produce diverse outcomes as a result of their distinct analysis pipelines. Each algorithm is also trained on different experimental data and utilizes unique sets of features. The best of these algorithms in terms of performance achieve sensitivity and specificity of approximately 60% and 30% respectively. Moreover, most of them are trained to provide *in silico* predictions in the 3’UTR regions of the mRNAs, while very few have been tested for identifying targets in their 5'UTR and coding regions (38,39).

1.3.1.1 Overview of de novo Target Prediction Algorithms

In the following section, a brief overview of the most widely utilized target prediction algorithms is provided.

TargetScan (40) is considered one of the first available programs with high sensitivity and precision. Its algorithm is mainly focused on the identification of miRNA binding sites with perfect complementarity in the seed region (7 or 8 consecutive nucleotides of perfect complementarity). Based on experimental evidence, these sites exhibit the highest repressing activity. Even though TargetScan algorithm follows a seed-dependent scoring system, it additionally identifies centered and offset 6mer miRNA sites. TargetScan algorithm provides a context score for each binding site which derives from a quantitative model that incorporates 14 distinct features, such as the first binding position on the 5′ end of the miRNA, binding type of the target site, local AU content and 3′ supplementary pairing.

It supports an extra mode specifically designed to rank sites from higher to less conserved target sites based on an aggregate conservation score. These two basic modes of the TargetScan algorithm can be jointly used for the assessment of the efficacy of each target site. TargetScan predictions are miRNA-family based, where miRNAs are clustered according to their seed similarity. The latest version can provide predictions for miRNAs of miRBase Release 21.

The algorithm has been trained on microarray data with clear siRNA/miRNA induced repression using a multiple/stepwise linear regression and has been tested in its efficiency to detect targets in the 3′-UTR region of protein coding transcripts.

miRanda (41) implementation scores the candidate target sites with a support vector regression algorithm, mirSVR. miRanda/mirSVR is specifically trained to identify potent repressing miRNA interactions. The model takes into account binding site complementarity, conservation, binding energy, site position in 3′UTRs and A/U flanking content. It was trained on a set of nine miRNA transfection experiments performed on HeLa cells. miRanda utilizes a pre miRBase 18 miRNA nomenclature.

MIRZA-G (42) is a recently developed target prediction algorithm able to detect both canonical and non-canonical miRNA binding sites and siRNA off-targets. Features such as, the flanking nucleotide composition, site structural accessibility, location within the 3′UTR and evolutionary conservation are deemed important for this algorithm. miRZA-G algorithm is based on a generalized linear model that additionally incorporates duplex base binding energy measurements derived from the MIRZA biophysical model (43). The training and the testing of miRZA-G model performance was evaluated against a set of 26 miRNA transfection microarray datasets. MIRZA-G utilizes miRNA sequences downloaded from miRBase version 20.

mirMark (44) is a computational framework that incorporates multiple characteristics of miRNA binding sites in order to assess putative miRNA-mRNA 3′UTR interactions.

Subsequently, the model utilizes distinct levels of classification for the evaluation of the binding sites and miRNA-mRNA interactions, respectively. The authors selected a random forest learning scheme for having the best performance for these two separate classification levels. The initial detection of miRNA-targeted regions is accomplished with the miRanda algorithm. Decisive features for mirMark performance are considered the base pairing, the nucleotide composition, the site structural accessibility and evolutionary conservation. The model has been trained on negative instances of mock miRNA-gene interactions as well as on positive experimentally supported miRNA-mRNA interactions retrieved from miRecords (45) and miRTarBase (46). Finally, the algorithm's performance has been tested on PAR-CLIP data. miRNA sequences utilized by miRmark are obtained from miRBase release 19.

mBSTAR (47) is a multiple instance learning framework developed for the identification of miRNA-gene interacting pairs. The mBSTAR model incorporates a random forest classifier utilizing 40 distinct features, such as nucleotide frequencies, duplex structure internal loops, bulges, and minimum free energy. The training and the testing of the algorithm was assessed on experimentally derived miRNA-mRNA interactions from miRecords (45), Tarbase v6.0 (7) and starBase (48). mBSTAR supports miRNAs obtained from miRBase release 20.

MirTarget (49) is a computational model that identifies canonical miRNA seed binding events in mRNA targets. The algorithm has been applied in 5 different organisms (human, mouse, rat, dog or chicken). MirTarget predictions are being deposited in miRDB web server (http://mirdb.org). It adopts an SVM-recursive feature elimination approach (RFE) in order to detect the most prominent independent features. The model incorporates several features, such as the target site conservation, accessibility (calculated with RNAfold), nucleotide usage per duplex position, location on the 3'UTR as well as other 3'UTR related characteristics. Although it has been developed based on 3'UTR characteristics, it also provides predictions for CDS and 5'UTR regions. MirTarget has been trained on canonical chimeric miRNA-target pairs derived from CLIP-Seq experiments (50,51). The testing of this implementation was performed on miRNA inhibition microarray datasets.

Initial research efforts have unveiled that miRNAs regulated gene expression through their binding on the 3'UTR of protein coding genes (6). However, accumulated experimental evidence has revealed that miRNA binding sites within coding sequences are also functional in controlling gene expression (52).

PACCMIT/PACCMIT-CDS (53) (Prediction of ACcessible and/or Conserved MIcroRNA Targets) is a recently developed algorithm that comprises two modules for the prediction of miRNA binding sites (with seed complementarity) on the 3'UTR and CDS regions of the mRNAs. Candidate miRNA binding regions are pre-filtered based on their structural accessibility and/or evolutionary conservation. Subsequently, the

predictions are scored following a Markov model that includes the information of the overrepresented targets versus a random background. PACCMIT includes the information of weakly and highly conserved miRNAs as introduced from TargetScan algorithm. The model has been trained and evaluated on proteomics and PAR-CLIP data.

DIANA-microT-CDS (54) is specifically trained on both 3′-UTR and CDS regions. The experimental positive and negative sets of MREs are derived from PAR-CLIP data performed in HEK293 cells (26). Candidate miRNA binding sites are subsequently combined in a general linear model trained on a set of 13 distinct microarray datasets that measure mRNA expression changes following transfection or knockout of a specific miRNA. The algorithm identifies (non)canonical 6mer to 9mer binding sites in 3′UTR and CDS regions. Target sites with buldges, G:U wobble, seed mismatches that correspond to non-canonical bindings are supported by additional 3′ pairing. Features of great importance for the microT-CDS algorithm are the target conservation, site accessibility that is estimated with Sfold program, binding free energy as calculated with RNA-Hybrid, AU flanking dinucleotides and the binding type. The algorithm adopts distinct conservation score models for the CDS and 3′UTR regions in 30 and 16 species, respectively. microT-CDS provides a final score for each miRNA-gene interaction combining the synergistic efficiency of MREs detected in CDS and 3′UTR results with a general linear model.

Further details on the microT-CDS algorithm and the utilized training sets, can be found in the relevant publication by Reczko *et al.* (38). DIANA-microT-CDS provides increased accuracy and the highest sensitivity at any level of specificity over other available state-of-the-art implementations, when tested against pulsed stable isotope labeling by amino acids in cell culture (pSILAC) proteomics datasets (55) and HITS-CLIP data . microT-CDS adopts a miRBase v18 nomenclature.

1.3.2 Experimental Methods for the identification of miRNA:mRNA interactions

Experimental techniques are usually divided into low yield and high-throughput methods, depending on their application scope and the number of obtained results per experiment. The most commonly used low yield techniques are reporter genes, qPCR and western blotting. Reporter genes are used for binding site validation, while qPCR and western blot or ELISA assays are usually combined to identify interactions that induce mRNA decay and/or translation suppression.

The first high-throughput techniques that became available could be considered as an increased throughput/lower accuracy version of specific techniques (56). Microarrays can be utilized to identify possible miRNA-gene interactions, as a high-throughput version of qPCR and northern blotting, while quantitative proteomic techniques can be seen as a high yield generalization of ELISA assays and western blots. Novel NGS

experiments have offered a high-throughput/increased accuracy combination that has revolutionized the way we identify miRNA-gene interactions. These techniques are based on NGS sequencing of mRNA sites bound by the Argonaute (AGO) protein and are often accompanied by sequencing of small-RNAs, as well as complementary experiments such as RNA-Seq and ribosome profiling (26,57).

HITS-CLIP (High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) was the first technique that offered for the first time a transcriptome-wide map of miRNA binding sites (57). The identified regions are usually wide and perplex the identification of the exact miRNA binding location, which is performed algorithmically. PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) is a modified CLIP-Seq methodology, incorporating 4-thiouridine in the nascent RNAs, which are subsequently detected as T-to-C transition sites in the AGO-miRNA-RNA cross-linked regions (26). Compared to the results obtained by HITS-CLIP, the boundaries of the identified binding locations are sharper and significantly narrower, while T-to-C mutations close to the region occupied by the RISC complex contribute to the identification of the exact MRE (58). Despite the accurate detection of the crosslinked region, these methods cannot directly reveal the specific miRNA participating in the interaction, which has to be bioinformatically identified. A more recent variant of the PAR-CLIP methodology (51) as well as the CLASH (crosslinking, ligation, and sequencing of hybrids) and CLEAR (covalent ligation of endogenous Argonaute-bound RNAs)-CLIP protocols (50,59) incorporate an extra ligation step, concatenating the miRNA to the mRNA binding region. The derived chimeric miRNA-mRNA fragments are subsequently sequenced and bioinformatically separated for the concurrent identification of the targeted mRNAs, binding sites and interacting miRNAs. Another important distinction between CLIP-Seq approaches is the reliance on either endogenous or exogenous AGO expression for the identification of AGO–miRNA–mRNA complexes. Nevertheless, the class of CLIP-Seq/CLASH experiments can reveal thousands of miRNA-gene interactions in each analysis and has significantly altered the scope and scale of relevant research projects.

Each technique has its own merits and disadvantages. An overview of available experimental techniques is presented in **Table 1**, along with short comments on their intended use, obtained results and expected throughput.

Method	Throughput	Intended Use
Reporter Genes(32)	Low	Validation of miRNA:UTR (or binding region) interaction
Northern Blotting(32)	Low	Relative effect of miRNA on mRNA levels
qPCR(32)	Low	Quantification of miRNA effect on mRNA levels
Western Blot(32)	Low	Relative assessment of miRNA effect on protein concentration
ELISA(32)	Low	Quantification of miRNA effect on protein concentration
5′ RLM-RACE(32)	Low	Identification of cleaved mRNA targets
Microarrays(32)	High	High throughput assessment of miRNA effect on mRNA expression
RNA-Seq(32)	High	High throughput assessment of miRNA effect on mRNA expression
Quantitative Proteomics (e.g. pSILAC(55))	High	High throughput assessment of miRNA effects on protein concentration
RPF-Seq	High	High throughput assessment of ribosome protected fragments
PARE / Degradome-Seq(60)	High	High Throughput identification of cleaved mRNA targets
Biotin miRNA tagging(32)	High/Low	Pull-down of biotin-tagged miRNAs and estimation of bound transcript content using qPCR (Low yield), microarrays (High throughput) and RNA-Seq (High Throughput)
IMPACT-Seq(61)	High	Pull-down of biotin-tagged miRNAs, identification of interacting pairs and binding regions.
PARE / Degradome-Seq(60)	High	High Throughput identification of cleaved mRNA targets
3Life(62)	High	High Throughput Reporter Gene Assay
miTRAP(63)	High	miRNA trapping by RNA baiting
AGO-IP	High	Identification of enriched transcripts (miRNAs and mRNAs) in AGO immunoprecipitates
HITS-CLIP(57)	High	Sequencing of AGO binding regions on targeted transcripts
PAR-CLIP(26)	High	Sequencing of AGO binding regions on targeted transcripts
CLASH(50) / PAR-CLIP + Ligation(51), CLEAR CLIP (59)	High	Sequencing of AGO binding regions on targeted transcripts. Production of chimeric miRNA-mRNA reads for the identification of interacting pairs.

Table 1: Index of experimental techniques utilized for the identification of miRNA-gene interactions. (Vlachos IS and Paraskevopoulou MD *et al*, 2014) (64)

1.3.2.1 Description of a CLIP-Seq protocol

CLIP-Seq techniques, which constitute a combination of AGO immunoprecipitation and NGS, have revolutionized miRNA-gene interactions research and enabled the detection of transcriptome-wide miRNA target sites. Typical PAR/HITS-CLIP protocols include the following steps (Figure 2) (65):

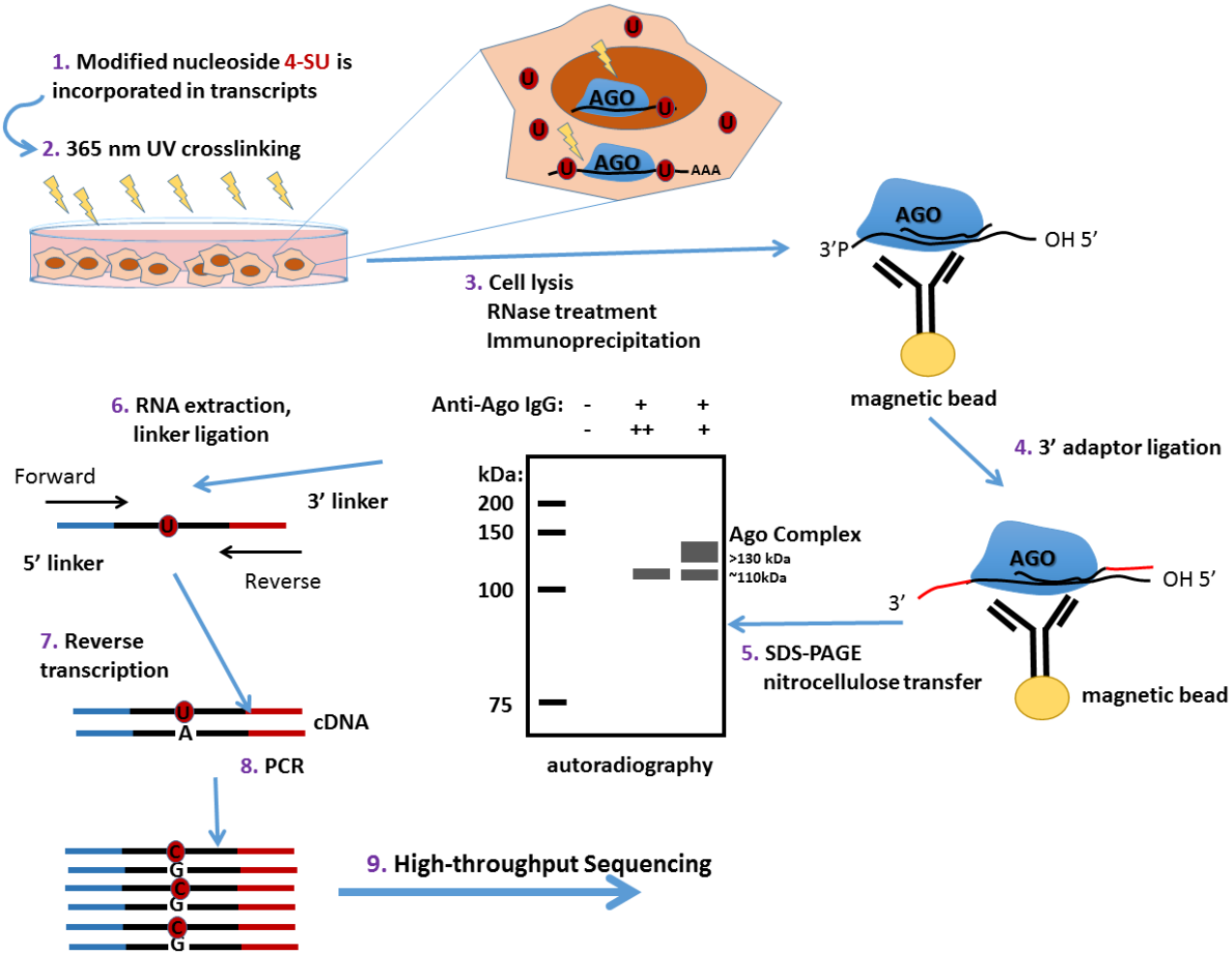

Figure 2: Steps followed in a typical PAR-CLIP-Seq experiment. *(Copyright Paraskevopoulou MD)*

- PAR-CLIP requires the incorporation of photoactivatable thioribonucleosides (4SU) into nascent transcripts.
- Crosslinking by using long-wavelength 365 and 254 nm UV in PAR-CLIP and HITS-CLIP respectively.
- Cell lysis.
- Isolation of crosslinked RNA–AGO complexes is achieved by immunoprecipitation.

- Sample processing by a ribonuclease (RNase) in order to partially digest the covalently bound RNAs.
- Radiolabeling of RNA segments crosslinked to immunoprecipitated AGO proteins.
- 3′ adapter ligation.
- Crosslinked AGO proteins are further purified by SDS–PAGE. After recovery of the RNA from the purified radioactive band the RNA is carried through a small RNA cDNA library preparation protocol for sequencing.
- The co-purified RNA molecules are reverse-transcribed and amplified with the aid of 5′ and 3′ adaptors.
- In PAR-CLIP experiments, the reverse transcription of the crosslinked-modified RNAs followed by PCR amplification, leads to a characteristic mutation that is used to identify the AGO binding sites (T-to-C when using 4SU and G-to-A when using 6SG).

1.4 State-of-the-art implementations for AGO-CLIP-Seq analysis.

PARalyzer (66) is considered the first model dedicated to the analysis of PAR-CLIP data. It was not developed though to identify miRNA binding sites from AGO PAR-CLIP. It is a generic model that identifies enriched regions for RNA-binding proteins from the analysis of PAR-CLIP deep sequencing data. PARalyzer adopts a kernel density estimator to quantify thymine-to-cytosine transitions. The kernel density approach is applied to crosslinked regions with normalized read counts and values of T-to-C conversions along with relative background signal estimations. Notably, for the detection of binding events this implementation has to be complemented with other algorithms such as cERMIT (67), mEAT (26) and MEME (68) depending on the intended use.

There are other algorithms similar to **PARalyzer** such as **CLIPZ** (69), **miCLIP** (70) and **Wavcluster** (71) that can be utilized to identify candidate RBP binding regions from the analysis of PAR/HITS-CLIP sequencing data.

MIRZA (72) is a biophysical model that has been designed to identify miRNA binding sites in Ago2-bound enriched regions. Model parameters have been deduced from PAR-CLIP AGO CLIP-Seq datasets. MIRZA implementation adopts 27 distinct energy parameters for the assessment of putative miRNA-transcript duplexes and assigns position-dependent binding energies. It predicts the frequencies with which RISC complexes are associated to different mRNA fragments and calculates a ‘binding site quality’ quantifying miRNA total affinity for each targeted fragment. Decisive features for MIRZA-adopted scoring scheme include, base pairing, base energies in the miRNA seed region (position 2-7) as well as energies at the 3′ compensatory base-pairing (positions 13-16) and 18-19 positions. It takes into consideration several other

parameters such as, the (a)symmetric loops and bulges formed in miRNA-target hybrids, mRNA fragment abundance as estimated from the CLIP-Seq experiment and miRNA expression values. The MIRZA model additionally identifies non-canonical miRNA sites.

The algorithm utilizes a simulated annealing approach for the optimization of 100 parameters, starting from random initializations. It has been trained on 2,988 crosslinked regions derived from four Ago2-PAR-CLIP datasets (73) and evaluated against 38 transfection microarray datasets comprising 26 distinct miRNAs. It has been compared against several other implementations that perform *de novo* miRNA target prediction and do not depend on CLIP-Seq experiments. The model presents some limitations, such as it discards miRNA sequences shorter than 21 nucleotides and requires AGO bound fragments to have 30-51 nts length and to be centered at the most abundant T-to-C crosslinked position/nucleotide. Moreover, it does not immediately process CLIP-Seq data but requires the input files in the specified format to be prepared by the user.

PARma (74) is a more recent implementation specifically designed to analyze PAR-CLIP datasets for the identification of AGO-miRNA binding events. The algorithm focuses on enriched regions that include T-to-C conversion sites.

PARma initially recognizes clusters of overlapping sequencing reads and subsequently uses these regions to infer statistically significant overrepresented kmers. Retrieved k-mers constitute the initial predictions for miRNA-family binding sites, following a seed-based miRNA clustering, and are subsequently forwarded for further evaluation to the core algorithm. PARma learning framework adopts 3 distinct/independent probabilistic models that consider positions of T-to-C conversions, RNase T1 cleavage sites (guanosines) upstream and downstream of the seed region. A likelihood is assigned to every putative miRNA binding region within the peak, taking into account the observed positions of transitions and guanosine cleavage sites. The adopted learning framework is fitted iteratively with an EM approach to infer required parameters.

The most probable miRNA seed family is associated with each cluster. Each prediction is accompanied by Cscore and MAscore scores for the cluster and miRNA activity, respectively. The first score describes the probability that a cluster is a correct miRNA-AGO bound region, while the latter reflects the efficacy of the miRNA regulator. The algorithm may produce at some cases more than one prediction for wider peaks that may have been produced from distinct clusters in close vicinity with overlapping spurious reads. PARma implementation has been trained on different PAR-CLIP experiments on B-cells. It was evaluated against PAR-CLIP datasets of EBV infected cells, as well as by comparing DG75 and BCBL1 B-cell lines expressing different sets of miRNAs. BCBL1 is a Kaposi's sarcoma-associated herpesvirus (KSHV) infected cell line presenting 25 distinct virus encoded miRNAs. Notably, PARma can perform parallel analysis of multiple PAR-CLIP datasets.

microMUMMIE (75). The group that initially developed PARalyzer, subsequently designed microMUMMIE algorithm to specifically address miRNA activity through the analysis of PAR-CLIP data. The model depends on PARalyzer predicted clusters which are provided as input to the microMUMMIE algorithm. The latter framework is preferentially applied to transcript 3′ untranslated regions.

microMUMMIE utilizes a six-state hidden Markov model (HMM). More precisely, state 1 corresponds to the background modeling, states 2, 4, 5, 6 model the cluster flanking regions and state 3 models the AGO enriched region. The fifth state comprises a 41-metastate submodel that identifies different types of miRNA seed pairing (6mer3-8, 6mer2-7, 7mer-m8, 7mer-m1, 7mer-A1, 8mer-A1 and 8mer-m1). The model is accordingly parameterized in order to prioritize predicted seed bindings near the 3′ cluster ends. miRNA seed complementarity, T-to-C conversions relative to binding sites, evolutionary conservation and sequence characteristics are deemed decisive features for this implementation. Shuffled miRNA sequences are included in the model in order to simulate decoy bindings and evaluate miRNA predicted sites via signal-to-noise ratios (SNR). Estimates of conservation can be optionally derived from TargetScan branch-length conservation scores (BLS).

It was trained and evaluated for its predictive accuracy against other algorithms on EBV infected lymphoblastoid cell lines. In the performed comparisons the authors included the 100 top expressed miRNAs.

1.5 Databases of miRNA-mRNA interactions

Low yield and especially high-throughput experimental techniques have already identified hundreds of thousands of miRNA-gene interactions in different taxa, species, tissues, cell lines and experimental conditions. This wealth of information is fragmented and hidden in thousands of manuscripts, supplemental materials, figures and raw NGS datasets.

DIANA-TarBase (76) was initially released in 2006 and was the first database aiming to catalogue published experimentally validated miRNA-gene interactions. Since then, a handful of similar projects (45,77) index and map experimentally identified miRNA interactions utilizing manual article curation, in order to maintain a high quality level of database entries. The sixth version of DIANA-TarBase (rel. Dec 2011) (56) inaugurated a new generation of such projects, incorporating for the first time novel methodologies, including CLIP-Seq experiments. The release of DIANA-TarBase v6.0 increased the available target space by 16.5 – 175-fold, to 65,000 manually curated experimentally validated interactions. This radical increase in collected interactions was a prelude of the upcoming paradigm shift introduced by the new high-throughput methods.

The current release, DIANA-TarBase v7.0 has pushed the envelope further and provides more than half a million entries derived from 28 different low/high experimental methodologies, 356 cell types and 59 tissues. More than 250 miRNA-related NGS datasets have been analyzed and approximately 7,500 validated specific miRNA-gene interactions have been indexed in 24 species.

miRTarBase (46) is the only similar database that has allocated resources to the curation of targets from high-throughput experiments. The last available version (release 6, Sep 2015) hosts 366,181 entries derived from low yield as well as high-throughput experiments. miRTarbase interactions include ~22,500 genes and >3,500 mature miRNAs from 18 different species. Other databases are updated less frequently or catalogue significantly smaller sets of interactions. **miRecords** (45) was first deployed in 2009 and focuses mostly on curating interactions from low yield experiments. The last update of the database (Apr 2013) comprises 2,705 interactions with 2,028 derived from low yield methodologies. There are also databases hosting CLIP-Seq sequencing results and/or that enable the online analysis of such datasets, such as starBase(48) and CLIPZ(78). These databases differ significantly from the aforementioned repositories, since their aim is to catalogue CLIP datasets and binding regions from any RNA binding protein (RBP).

1.6 LncRNAs

Recent transcriptome-wide NGS studies unveiled a large number of lncRNA transcripts and introduced their regulatory roles in the cell (79). LncRNAs are typically longer than 200nts, and are characterized by compartmental, tissue, disease and developmental stage-specific expression. Even though they generally exhibit poor sequence conservation, and were initially considered "transcriptional noise", recent studies have described lncRNA conserved function (80-85). There are several examples of well-characterized lncRNAs such as Xist and Air that present intrinsic functions, despite their low primary conservation (85).

LncRNAs are spatially classified into four main categories (sense, antisense, intergenic, bidirectional) according to their loci of origin and transcription orientation as compared to protein coding genes (86,87):

- Sense or antisense: lncRNAs overlapping non-intronic parts of protein-coding genes, located in the same or the opposite strand.
- Sense or antisense intronic: lncRNAs overlapping intronic parts of protein-coding genes, located in the same or the opposite strand.
- Bidirectional: transcribed in a "head to head" orientation and located in close proximity to a protein-coding gene.
- Intergenic: lncRNAs located exclusively within intergenic regions.

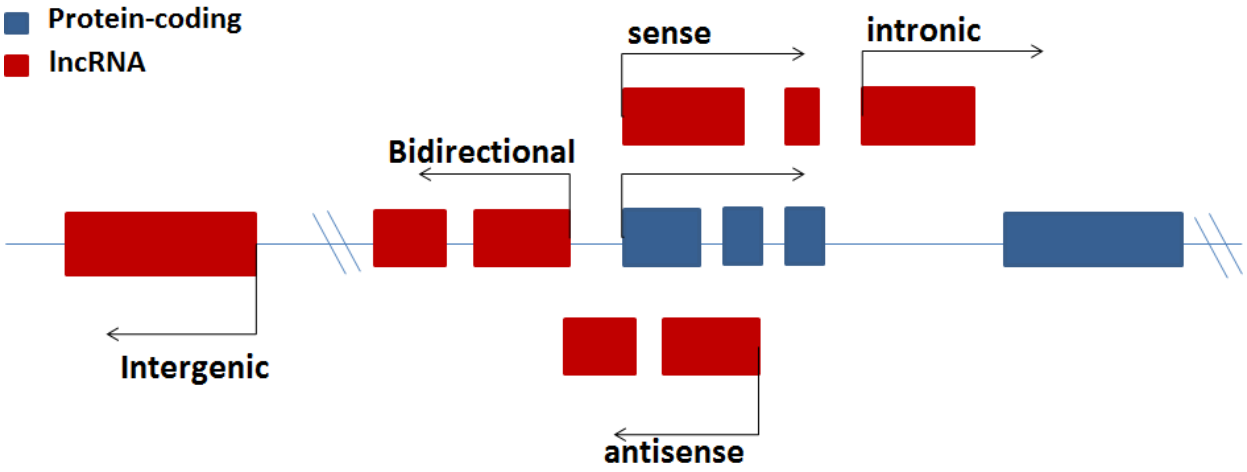

Figure 3: Spatial classification of lncRNAs into four main categories (sense, antisense, intergenic, bidirectional) according to their loci of origin and transcription orientation as compared to protein coding genes. *(Copyright Paraskevopoulou MD)*

LncRNAs have common characteristics with the protein coding transcripts. Many lncRNAs are polyadenylated, 5′ capped and spliced. Most lncRNAs transcribed by RNA polymerase II, and few of the RNA polymerase III (82). Their low abundance is probably connected with the underestimation of lncRNA transcript length and number of exons (88). Even though they generally do not have a clearly defined open reading frame (ORF) (87) and any coding ability, recent studies used ribosome profiling and revealed that some of them may encode short peptides (89). Table 2 summarizes the main similarities / differences of mRNA with lncRNA.

Characteristics	mRNA	lncRNA
Function	protein coding	regulatory, structural roles
ORF	Yes	few or no ORF
Cap structure	Yes	yes /no
Polyadenylation	Yes	Yes
Translation	Yes	No
Splicing	Yes	Yes /No
Subcellular localization	cytoplasm	predominantly nucleus, cytoplasm
Conservation	highly conserved	less conserved than protein coding genes

Table 2: Comparison of lncRNA-mRNA characteristics.

1.6.1 lncRNA Functions

LncRNAs exhibit numerous functions, many of which are under debate or remain to be uncovered (90). They perform different roles in all cell compartments, controlling gene expression in *cis* and/or *trans* by participating in almost every known level of regulation. lncRNAs promote chromatin modifications, mediating gene silencing; can act as guide molecules and scaffolds for proteins, contributing to the formation of cellular substructures; they are also shown to control protein synthesis, RNA maturation and transport (91,92); some of them encode small non coding RNAs.

a. Identification of lncRNAs that regulate miRNA transcription

LncRNAs (such as Meg3, Dleu2, H19, Ftx, etc) can function as pri-miRNA host genes (93). Genomic regions where miRNA transcripts and lncRNAs overlap can have dual/multiple functionality. Different biological processes can either trigger lncRNA function or promote the activation of the miRNA biogenesis pathway. Several well-known polycistronic miRNA gene clusters, including members of let-7 family, derive form intergenic regions that also encode lncRNAs.

The characterization of pri-miRNA transcripts remains widely unknown and is hindered by practical obstacles (94). The rapid cleavage of primary miRNA transcripts by Drosha enzyme in the nucleus does not allow complete transcript annotation with conventional approaches. microTSS (94) is a versatile computational framework that enables tissue specific identification of miRNA transcription start sites. Its current version requires RNA-Seq datasets in order to detect expressed regions upstream of miRNA precursors. The area around the 5′ of the RNA-Seq signal is scanned for H3K4me3, Pol2 and DNase enrichment, corresponding to putative regions for pri-miRNA transcription initiation. The candidate miRNA promoters are scored based on 3 distinct SVM models, trained on deep sequencing data.

The annotation of intergenic pri-miRNA transcripts with microTSS can enable the identification of overlapping lncRNAs, the revision of lncRNA and pri-miRNA annotation that in many cases is considered incomplete, the detection of common lncRNA - miRNA promoter regions as well as further support lncRNA-centered functional analyses. This machine learning approach outperforms any other similar existing methodologies and can be easily applied on any cell line/tissue of human or mouse species utilizing RNA-Seq, Chip-Seq and DNase-Seq data. microTSS is available for free download at www.microrna.gr/microTSS.

b. Experimentally verified miRNA-lncRNA interactions

LncRNAs have also been shown to function as “sponges” coordinating miRNA function. Most of these interactions take place in the cytoplasm, while there are also examples of miRNAs targeting lncRNAs in the nucleus. PTEN pseudogene competes its coding counterpart for miRNA binding; CDR1as/ciRS-7 circular antisense transcript acts as a sponge by harboring multiple miRNA binding sites, while it is also cleaved in the nucleus through a miRNA-AGO mediated mechanism; linc-MD1, a muscle-specific

lncRNA, functions in the nucleus as a pri-miRNA host gene, while it also exports in the cytoplasm acting in a "target mimetic" fashion for two miRNAs. An extended collection of functional direct miRNA-lncRNA interactions is described in Table 3.

Several other lncRNA-miRNA indirect interactions have been identified by low throughput expression experiments that quantify miRNA effect on mRNA levels and *vice versa* (95,96). There are also lncRNAs that originate from highly conserved genomic regions (ultra-conserved regions - UCRs) and are considered candidate miRNA targets (97). In a recent study, authors utilized lentiviral small hairpin RNAs to suppress 147 lncRNAs. The results of their approach demonstrated that lncRNAs, although mainly detected in the cell nucleus, appear to be sensitive in AGO-sRNA-mediated regulatory mechanisms (98).

A significant portion of miRNA-lncRNA interactions remains obscure and unexplored. To this end, new technological advances and NGS experiments can assist the process of miRNA (non)coding target characterization.

lncRNA	miRNA	Cell Type/Tissue	Species	Compartment	Low-throughput experiment
CDR1as/ ciRS-7(99-101)	miR-671, miR-7	embryonic kidney, brain	*H. sapiens, M. musculus, C. elegans, Zebrafish*	nucleus, cytoplasm	Reporter, qPCR, Northern blot
HULC(102)	miR-372, miR-433-3p, miR-557, miR-622, miR-134-5p, miR-613, miR-1236-3p	Liver	*H. sapiens*	cytoplasm, nucleus	Reporter, qPCR
BACE1-AS(103)	miR-485-5p	Brain	*H. sapiens, M. musculus*	cytoplasm	Reporter, qPCR
PTENP1(104)	sequesters miRNAs that target PTEN	Prostate	*H. sapiens*	cytoplasm	Reporter
linc-MD1(105)	miR-133a, miR-135b, miR-206	Muscle	*H. sapiens, M. musculus*	cytoplasm	Reporter, qPCR
H19(106-108)	miR-106a, miR-17-5p, miR-20b, let-7, miR-141, miR-200, miR-429, miR-675	myoblast, muscle, liver, brain	*H.sapiens, M. musculus*	cytoplasm	Reporter, qPCR
MALAT1(109 ,110)	miR-101, miR-217, miR-9, miR-125b	ESCC, brain, bladder	*H. sapiens*	nucleus	Reporter, qPCR
GAS5(111)	miR-21	Breast	*H. sapiens, M. musculus*	cytoplasm	Reporter, qPCR
PCAT-1(112)	miR-3667-3p	Prostate	*H. sapiens*	cytoplasm, nucleus	Reporter, qPCR
MDRL(113)	miR-361	Cardiomyocytes	*M. musculus*	nucleus, cytoplasm	Reporter, qPCR, Northern blot
HOTAIR(114)	miR-34a, miR-130a	prostate, gallbladder	*H. sapiens, M. musculus*	-	Reporter, qPCR, Northern blot
UFC1(115)	miR-34a	liver	*H. sapiens, M. musculus*	cytoplasm	Reporter, qPCR
HOST2(116)	miR-1266,let-7b	Ovary	*H. sapiens*	-	Reporter, qPCR

Table 3: miRNA-lncRNA experimentally verified interactions from different low yield experimental techniques. lncRNA target mimetic function has been recorded in the cytoplasm as well as the cell nucleus. Certain interactions are conserved in more than one species. (Paraskevopoulou MD *et al*, 2015)(117)

1.7 ceRNA Activity

It is hypothesized that "competing endogenous RNA" (ceRNA) interactions exist in the transcriptome (Figure 4). In this level, mRNAs, pseudogenes, and ncRNAs communicate through a competing language, forming a large-scale regulatory network. miRNAs have been considered as the controllers of the ceRNA activity. In this large network of in-between transcript interactions; miRNAs target other RNAs (mRNAs, pseudogenes, lncRNAs), while the latter may act as sponges for miRNAs, mediating their regulatory role. To this end, the formation of a complete map of their endogenous interactions is considered essential. This activity has been reported both in the nucleus and in the cytoplasm.

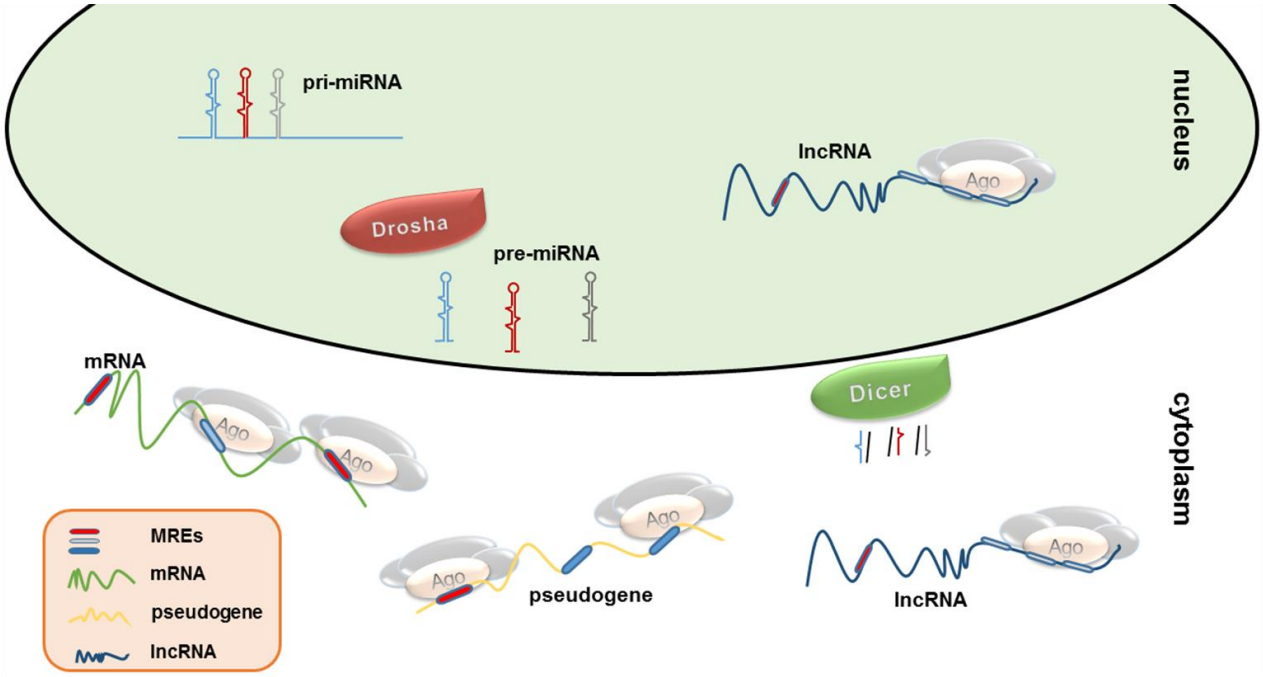

Figure 4: Overview of the ceRNA activity in nucleus and cytoplasm. miRNAs loaded in the RISC complex post-transcriptionally regulate protein coding genes through mRNA cleavage, direct translational repression and/or mRNA destabilization in the cytoplasm. lncRNAs compete with mRNAs for miRNA binding by acting as 'sponge' molecules in both cell compartments. (Paraskevopoulou MD *et al.*, 2016) (118).

1.8 Databases of miRNA-lncRNA interactions

DIANA-LncBase v1 (119) is considered as the first extensive database dedicated to the cataloguing of miRNA-lncRNA interactions and provided the largest collection of experimentally supported entries. LncBase v2 (117) currently hosts more than 10 million *in silico* predicted and ~70,000 experimentally supported interactions for an integrative meticulously curated collection of lncRNA transcripts. The new database enables the identification of miRNA-lncRNA regulatory interactions in numerous tissues, cell types and conditions, validated with low yield or high-throughput experimental methodologies.

miRcode (39) hosts predicted miRNA binding sites on human lncRNA transcripts retrieved from GENCODE v11 (120), while **LNCipedia** (121) accompanies lncRNA entries with miRNA canonical predictions by using the MirTarget2 algorithm (122). **starBase** (48) provides a collection of binding events for different RNA binding proteins. For AGO-binding sites, it can intersect miRanda/mirSVR-predicted (41) miRNA binding sites with the identified CLIP-Seq enriched regions spanning lncRNA transcripts. **NPInter** (123) integrates information from other repositories and literature regarding non coding regulation and interactions, including ncRNA-protein and ncRNA:miRNA binding events. It supports lncRNA annotation from different resources, while lncRNA-miRNA interactions are obtained from external databases such as Starbase. **LncReg** and **lncRNome** (124,125) aim to catalogue lncRNA-associated regulatory events. These databases also host a restricted number of miRNA binding sites on lncRNAs. These sites are either derived by text mining or are *in silico* inferred from high-throughput datasets.

1.9 Pattern Recognition

Pattern Recognition is considered an extremely broad scientific area that aims to detect and classify entities/objects in noisy and complicated environments. It is an intelligent machines system utilized for making data-driven predictions or decisions expressed as outputs. Depending on the type of application these objects can be found in any kind of format such as image, sound and simple measurements. In this field, different Machine learning and Statistical Decision Theory methods are utilized (126). The machine learning field deals with the development of techniques and methodologies, commonly referred as algorithms that allow computers to adopt learning behaviors. It aims to change and adapt the software behavior, based on the experience provided by the analysis of previous cases. Some of the most promising methodologies include, Artificial Neural Networks (126), Support Vector Machines (126,127) and Random Forests (128).

Machine learning has a broad spectrum of applications including text classification, economics, medical diagnosis and bioinformatics.

1.9.1 Machine Learning

Machine learning frameworks are initially developed based on the comprehension of a features dataset, commonly referred to as the "training set". The evaluation of a model on its ability to make the correct decision in an unknown set different from training (test set), is considered crucial. This ability is known as generalization and is a central goal in machine learning models (129). The selection of the right training and test sets is pivotal to a model's predictive performance. There are different ways to assess and optimize a model learning process (bootstrapping, cross validation, Jackknife resampling).
Machine learning applications, where the training data are provided along with their outcomes are referred as supervised. Such cases are further divided in classification and

regression problems if the training samples are categorized on discrete output classes or assigned on one or more continuous values, respectively. On the other hand, on unsupervised learning frameworks (e.g. clustering techniques), the output of the training instances is not known *a priori*. In such models, the goal is to perform exploratory data analysis for the identification of data distributions, rules and patterns that may enable their clustering into groups. Finally, there are machine learning approaches termed as "Reinforcement learning", where the models are interactively developed from the environment and make decisions based on new data observations to maximize the reward/gain.

Notably, the process of designing a learning framework can be assisted by prior knowledge of a theoretic model based on previous observations and experiments. However, many of the machine learning problems are not coupled with such information and therefore require exploratory, data-driven analyses. The lack of this prior model knowledge can be bypassed with the use of advanced non-parametric methodologies (e.g. Support Vector Machines, Neural Networks etc.).

1.9.2 Machine Learning models

This section aims to indicatively present a series of machine learning models adopted to support ncRNA related studies and discuss the intrinsic details of how these algorithms function. In addition, state-of-the-art learning frameworks that were applied during this thesis are described in the following sections.

1.9.2.1 Generalized Linear Models

Generalized Linear Models (GLM) were introduced by Nelder and Wedderburn (1972) (130), and are a broad class of models, which is considered as an extension of the general linear models. This category comprises linear regression, logistic regression and Poisson regression. A simple GLM model utilizes a linear combination of observed variables (linear predictor), in order to infer/predict the expected outcome of unseen inputs (response variable). The response variables can follow an exponential distribution such as Gaussian, binomial, gamma, Poisson or non-exponential distributions. The parameters of the models for maximum likelihood derivation are being calculated iteratively with least squares techniques or Bayesian approaches. GLMs additionally adopt an invertible linearizing link function to capture the association between the linear predictors and the response variable.

GLMs can be highly adapted and expanded to more complex and sophisticated learning models, exhibiting a high plasticity in their analytical properties. However, they face important restrictions when processing high dimensional datasets.

1.9.2.2 Naive Bayes Classifier (131)

Naïve Bayes (NB) Classifier belongs to a family of ML models that have evolved from the strength and elegance of the Bayes Theorem. This methodology is one of the most popular among researchers and has frequently demonstrated its usefulness in solving

difficult bioinformatics problems, many times more accurately than other more sophisticated techniques. NB adopts a simplistic approach, based exclusively on the Bayesian theorem and assuming that each parameter is independent and unrelated with the others. NB seeks to assign in each instance a class that maximizes a product of posterior probabilities (probability of a class occurrence for a set of features). A general technique that can be utilized towards this direction is the expectation-maximization (EM) algorithm. In an NB approach, the prior class probabilities can be assumed equal or be estimated from the training data. The distribution of features can be either considered continuous (e.g. Gaussian) or discrete (e.g. multinomial and Bernoulli)

The "naive" independence hypothesis implies that the (non)existence of a variable does not correlate with the behavior of the others. Even though its foundation is over-simplistic, in practice it demonstrates robustness and excellent generalization capabilities. A series of works have been elaborated in order to locate the intrinsic details of how this simple classifier performs so well in real world situations, including medical and bioinformatics applications. It has been applied to different miRNA-related research projects, including miRNA-target prediction (132).

1.9.2.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is considered a regression or classification tool for two or more groups. It is used in ML to identify a linear association between the features of a model. Using a fairly large number of predictors, the LDA creates optimal dividing lines between instances that define the identity of each group. This method can be additionally applied for dimensionality reduction.

1.9.2.4 Artificial Neural Networks (ANN) (133)

Artificial neural networks are considered as reference supervised/unsupervised machine learning algorithms for regression and classification. ANNs are mathematical models capable of nonlinear statistical data processing. Inspired by the structure and function of mammalian biological neural networks, they comprise interconnecting artificial neurons of adaptive importance for information exchange. In a supervised learning context, the weights of these connections are being tailored on the training data. Weight selection is performed most of the times by the minimization of an error function relevant to the network architecture. This function is usually a metric, describing the deviation of real values that serve as targets and the predicted outcome. ANNs are characterized by their network architecture, topology, number of hidden layers and included neurons. The final decision is characterized by the appropriate weight selection. As soon as the neural network is trained, it exhibits good generalization ability and robust predictive accuracy. They have been successfully applied to many different problems in Bioinformatics (134,135).

Notably, ANNs can be efficiently combined to form Ensemble Classifiers and avoid entrapment in local minima during the training process.

1.9.2.5 Support Vector Machines (SVMs) (127)

Support Vector Machines are a powerful family of supervised learning methods, used for classification and regression purposes. They were initially proposed by Vladimir Vapnik and belong today to the frontline of methodologies in the field of machine learning. They have been successfully applied in numerous problems and are considered as some of the most robust methodologies with excellent generalization capabilities. They belong to the group of kernel based methods that can provide sparse solutions. In brief, a support vector machine constructs a set of hyperplanes to address regression or classification problems in a very high or even infinite dimensional space.

Given a classification problem, an SVM aims to define a hyperplane $wx + b$ that best separates the classes. In order to find the maximum-margin hyperplane (equivalent to maximizing $2/\|w\|$ following specific constraints) that divides the points belonging to the different classes **w, b** should appropriately be chosen. It achieves high accuracy by optimizing the decision hyperplane to be the one that provides maximum margin between the classes (in case of classification). This classification framework is adopted for linearly separable hyperplanes. It is possible to use a nonlinear hyperplane by first mapping the sample points in a higher dimensional space via nonlinear mapping. This procedure, called a 'kernel trick', introduces additional dimensions to enable linear classification in the transformed space.

Many extensions of the original methodology have been proposed that allow mislabeled examples (soft margin classifiers), nonlinear support vector classifiers, multiclass support vector classifiers etc. Common kernels adopted for non-linear SVM are sigmoid, polynomial and Gaussian radial basis functions (RBF).

They have been applied successfully in a very large variety of problems and are rigorously researched, since this category of models provides very high performance, in terms of sensitivity and specificity, and robust generalization.

1.9.2.6 Relevance Vector Machines (RVMs) (136)

Relevance Vector Machines (RVMs) are a classifying method introduced by Tipping, which is in terms of sparsity equivalent to Support Vector Machines. The main difference between SVMs and RVMs, is that the second machine learning algorithm is probabilistic in nature, which is regarded as one of the most important issues in terms of decision making. It is obvious that when probabilistic classifiers capture the uncertainty in the prediction they are preferred from the hard point classifiers like SVM. RVM algorithm proposed by Tipping can achieve significant accuracy, generalize well and are considered as computationally efficient. The main concept in RVM algorithms is that they identify the patterns in the training set that seem to be more representative, which are considered as the Relevance Vectors. These patterns correspond to nonzero weights and are used in the predicting phase. It has been observed that RVMs exhibit a

good performance in the classification procedure with similar generalization abilities as SVMs.

1.9.2.7 Decision trees (137)

Decision Trees are a machine learning framework that assigns observations of variables to target values (predictions). They can be utilized to address classification or regression problems. In decision trees, the leaves correspond to the target values that can be discrete or continuous. In classification trees, the leaves are represented by the predicted classes or probabilities for the classes. The interior nodes are linked to input features, while the interconnecting branches describe the possible outcomes of a specific feature. The learning process in decision trees is accomplished by partitioning and evaluating the training. This step can be performed either recursively or following other splitting criteria of the initial dataset into subsets, such as normalized information gain or entropy. The C5.0 model belongs to the family of decision trees and is an updated version of previous algorithms (C4.5, ID3). It is faster, more efficient in memory usage and additional. Moreover, it provides boosted learning to improve the performance of weak decision trees as well as weighting of variables and misclassified cases. Even though they do not exhibit exceptionally high accuracy (maybe because of the high variance of the data), they can often provide robust predictions for different feature distributions and for datasets comprising missing values and/or correlated features.

1.9.2.8 Random Forests (128)

Random Forests (RFs) are ensemble classifiers that were developed by Leo Breiman and Adele Cutler. They incorporate multiple models to achieve better predictive performance. RFs preserve most of the appealing features of the decision trees with the ability to deal with both classification and regression problems. RFs are considered as a streamlined version of bagging. The basic concept of the algorithm is that it combines Breiman's "bagging" idea and the random selection of features, in order to construct a set of decision trees with controlled variation. Some of the basic RF characteristics render the algorithm preferable against other machine learning methods. RF models can be efficiently applied, since they can process thousands of input variables without prior feature selection and data preprocessing; define the most appropriate/prominent set of descriptors (utilized as an alternative feature selection approach); handle datasets with missing values without downgrading the achieved accuracy. A predicted class in random forest approach is the one that occurs most frequently as an output by individual trees. The construction of each tree during the learning procedure is achieved through a number of specific steps. For instance, if a model has N training instances and M number of variables in the classifier: m input variables (with m<=M) are utilized to determine the decision at a node tree. A set of n training instances is chosen from the pool of N training rows, whereas the rest of the samples are used as the

test set for error rate estimation. For each node, a set of m random variables is chosen in order to make the prediction in the specific node. This procedure is carried out for the definition of the best combination of m variables in the training set. Moreover, each tree is fully grown and not pruned. While the forest building progresses, the algorithm estimates the generalization error. Random forests utilize proximities between pairs of cases in order to detect outliers and provide useful views of the data. In the prediction phase, each test sample traverses the tree till it reaches a leaf node. The result comes from the average vote of all trees, since the procedure is iterated over all trees in the ensemble classifier.

Compared to other machine learning techniques such as SVMs and neural networks, this classifier has relatively fewer applications in bioinformatics, but is rapidly gaining popularity.

1.9.2.9 Gradient Boosting Machines (GBMs) (138)

Gradient boosting is a category of highly adaptive ensemble machine learning models that can be utilized in different regression and classification applications. They are composed by large /small trees that are sequentially fitted to reweighted versions of the training data. This breakthrough invention of Freud and Schapire has a different learning strategy than classic ensemble algorithms such as random forests. GBMs gradually increase the number of included models, adding a new weak learner on each iteration, and finally decide based on weighted average voting. More precisely, they perform sequentially training where initial simple learners fit models to the data, while subsequent ones analyze the data for error cases of prior learners (error residuals), and finally try to provide the correct predictions in the following steps. This procedure is commonly referred as stage-wise additive modeling where the main goal is to minimize a loss function. Boosting models can vary depending on the different optimization approaches and loss function distributions (Bernoulli, Poisson, Adaboost, Gaussian, and Laplace).

GBMs can convert combinations of weighted weak learners into complex predictors, where the results of new trees represent partial solutions to the entire problem. They are sensitive to noise and extreme values. There are different ways to leverage trees for achieving better performance and to avoid overfitting, such as monitoring the number the included trees. Boosting learners are robust algorithms that often achieve better accuracy than random forests and bagging algorithms.

1.9.3 Feature Preprocessing

The preprocessing of a dataset's features is often necessary for many predictive models and is commonly used in cases requiring dimensionality reduction and elimination of outliers. In many machine learning frameworks, the initial variables set is often transformed in a new feature space to achieve their easier interpretation as well as

increased model performance. Features may also be preprocessed, not only for dimensionality reduction but in order to facilitate faster computations.

Spatial sign transformation is a relatively new method, which was proposed in 2006 by Serneels and partners (139). This process projects the predictors into a multidimensional sphere. The transformed features present a more robust behavior to outlying observations. This technique locates all sampled variables in equal distance from the center of a sphere. An interesting characteristic of this process (in contrast to conventional methods, such as centering / scaling) is that the predictive parameters are independently transformed simultaneously and not sequentially. This technique is able to increase the performance of a learning method without the removal of predictors.

1.9.3.1 Methodologies for parameter Selection

In machine learning applications a common issue is that when increasing the number of measured parameters, it forces the necessity to further increase the number of studied instances, in order to provide accurate predictions. This is often described as "the curse of dimensionality". To circumvent this problem, since it is often technically unfeasible in terms of resources and time to increase the instances accordingly, a variety of methods has been devised for selecting the most prominent descriptors. These methods are considered indispensable components in demanding machine learning problems.

Exhaustive search of predictors is considered as a computationally challenging approach for parameter selection. In exhaustive search, all possible subsets of features are evaluated for their performance. Other methods adopt search algorithms and/or utilize score functions to assess the predictive accuracy of subset of features. In many applications, stepwise regression is used to identify promising variables. Moreover, one popular machine learning approach is to combine a Recursive Feature Elimination algorithm to identify the most informative features and iteratively evaluate the performance of Support Vector Machines following lowly weighted predictor removal. For the development of machine learning models it is also highly recommended to reduce features presenting high correlation and to remove non-informative predictors that exhibit near to zero variance. Other techniques that are utilized towards this direction are presented below.

Filtering / selection methodologies (e.g. distance Kullback-Leibler, Wilcoxon's exact test, ROC AUC, etc) evaluate and rank every parameter individually based their predictive accuracy. Disadvantage of these methodologies is that they reflect the behaviors of parameters in one dimension, ignoring the other measured data and their in-between associations.

Information gain methods for feature selection. For a particular set of descriptors, feature selection can be accomplished using the information gain measure of Quinlan (137). This measure considers that higher information is associated with higher separation ability (e.g. active/inactive compounds). Higher information gain is related to lower information entropy of the subsets defined by the presence and absence of

particular features. Another commonly used approach for feature selection is the minimum-redundancy-maximum-relevance (mRMR) (140). It enables the identification of sets of non-redundant features through association, distance and mutual information measures.

Methodologies for extracting novel parameters. These methodologies usually combine the measured characteristics/features in order to extract new variables presenting higher predictive accuracy. Often, the results of these methodologies can be used for selecting the most informative variables and transform the data into a smaller subspace comprising uncorrelated or independent descriptors. Typical algorithms that can be utilized for this purpose are principal component analysis (PCA), discriminant analysis and independent component analysis (ICA).

Optimization Algorithms. They are considered methodologies that can identify the most prominent parameters by optimizing the performance of an algorithm. Genetic algorithms (GAs), swarm optimization algorithms, search algorithms (e.g Best-first search) etc. belong to this class of techniques.

2. Methods

This section provides a complete overview of the implemented computational approaches for the identification of miRNA-mRNA-lncRNA endogenous interactions and their functional interpretation. The algorithmic steps described in the following sub-sections can be summarized accordingly:

1. Identification of *in silico* predicted miRNA targets. Development of a microT web server for the indexing of miRNA-mRNA interactions.
2. Formation of automated analysis pipelines for functional analysis of miRNA targets and the seamless interconnection of workflows with the DIANA microT web server.
3. Development of a DIANA-Taverna Plug-in and deployment of DIANA-related services.
4. *In silico* analysis of raw (small)RNA-Seq datasets and AGO-CLIP-Seq libraries for the identification of miRNA-gene interactions.
5. Applied methods for the development of DIANA-TarBase v7, a database dedicated to the cataloguing of experimentally derived miRNA-mRNA pairs.
6. Applied methods for the release of DIANA-LncBase v2, a repository devoted to the indexing of experimentally supported miRNA-lncRNA interactions.
7. Evaluation of the LncBase/Tarbase AGO-CLIP-Seq algorithm for miRNA target identification.
8. Implementation of a Novel Algorithm for AGO-CLIP-Seq data analysis.
    a. Collection of numerous low/high throughput experiments to reveal the impact of miRNA targeting on gene expression and to deduce putative positive/negative miRNA-target interactions.
    b. Compilation of a training set comprising positive and negative CLIP-Seq-guided miRNA binding sites.
    c. Feature extraction and assessment.
    d. Proposed learning framework for the identification of miRNA targeted regions through the analysis of AGO-CLIP-Seq data.
    e. Evaluation of the proposed algorithm.

2.1 Computational identification of miRNA-target interactions

Collected Transcripts.

Ensembl v75 has been utilized for protein coding transcript annotation, while miRNA identifiers and sequences were obtained from miRBase v18 nomenclature (141).

Annotation for lncRNA transcripts was derived from GENCODE v21 (120). GENCODE provides the largest available collection of high quality lncRNA transcripts, spatially classified into four main categories (sense intronic, sense overlapping, antisense and intergenic) according to their transcription orientation and locus of origin relative to protein coding genes. Transcripts annotated as 'processed transcripts' also clustered in the larger lncRNA family were included in the performed analyses. The finalized lncRNA collection includes all GENCODE indexed transcripts as its main annotation, and also integrates lncRNAs contained in RefSeq (142) and the publication of Cabili *et al*. (88) presenting less than 90% sequence similarity with GENCODE entries. This integration was essential due to the highly dissimilar spliced transcripts that exist between different lncRNA resources. The final set of lncRNA transcripts comprised 1,830 sense, 10,201 antisense, 18,029 long non-coding intergenic RNAs (lincRNAs) and 2,163 processed transcripts for *Homo sapiens*. The respective set for *Mus musculus* consisted of 399 sense, 2,642 antisense, 4,542 lincRNA and 1,689 processed transcripts.

2.1.1 *In silico* predicted interactions.

miRNA-mRNA in silico predicted interactions. *In silico* target prediction for human and mouse spliced mRNA sequences was performed using DIANA-microT-CDS algorithm (54).

miRNA-lncRNA in silico predicted interactions. *In silico* target prediction for human and mouse spliced lncRNA sequences was performed with an appropriately adjusted DIANA-microT algorithm (54). MREs were scored separately and each miRNA:lncRNA interacting pair was characterized by a cumulative score which signifies the interaction strength.

2.2 Methods for the development of DIANA-microT web server

One of the major aims of this thesis goals was to specify a comprehensive catalogue of miRNA-mRNA *in silico* interactions. To this end, DIANA-microT web server v5 was implemented to provide a reference archive of computationally predicted miRNA-mRNA interactions.

DIANA-microT web server (http://www.microrna.gr/webServer) is dedicated to miRNA target prediction/functional analysis and it is being widely used from the scientific community, since its initial launch in 2009. During the thesis course, **DIANA-**

microT v5.0 (54), the new version of the microT server, has been significantly enhanced with an improved target prediction algorithm, DIANA-microT-CDS (38). The new algorithm microT-CDS can identify miRNA targets in 3'UTR, as well as in CDS regions. microT-CDS is the only algorithm available online, specifically designed to identify miRNA targets both in 3' untranslated region (3'UTR) and in coding sequences (CDS).

2.2.1 Release of DIANA-microT web server v5

The web server was completely redesigned, in order to host a series of sophisticated workflows, which can be used directly from the on-line web interface, enabling users without the necessary bioinformatics infrastructure to perform advanced multi-step functional miRNA analyses. DIANA-microT web server v5.0 also supports a complete integration with the Taverna Workflow Management System (WMS) (143), using an in-house developed DIANA-Taverna Plug-in. This plugin provides ready-made modules for miRNA target prediction and functional analysis, which can be used to form advanced high throughput analysis pipelines.

2.2.2 Formation of Automated Analysis pipelines

As high-throughput data have become the new backbone of biological research, there is an increasing need to support advanced high throughput analysis pipelines. DIANA-microT web server v5.0 was completely redesigned in order to provide the necessary building blocks to easily incorporate miRNA functional analyses in complex pipelines. The new DIANA-microT web server facilitates users not having access to extensive computational infrastructures and support, in order to perform ready-to-deploy sophisticated analyses.

A series of workflows have been prepared, which can be used as standalone modules, as a foundation for custom pipelines or to be incorporated into pre-existing algorithms. These pipelines can be utilized to analyze user data derived from small scale and high throughput experiments directly from the DIANA-microT web server interface, without the necessity to install or implement any kind of software. For the identification of miRNAs having functional impact in differentially expressed genes, the user can specify the species and two lists of differentially expressed mRNAs (microarray/RNA-Seq) and miRNAs (microarray/sRNA-Seq), respectively. The gene list has to contain ENSEMBL gene identifiers, while the miRNA list should be composed of miRNA names/identifiers according to miRBase nomenclature. miRNA and gene identifiers can optionally be followed by fold change values. In this case, the workflows automatically match suppressed genes with overexpressed miRNAs (and *vice versa*).

Detailed descriptions of the automated analysis pipelines are provided in the relevant results section.

2.2.3 DIANA-microT web server integration with Taverna WMS

DIANA-microT web server enables advanced users to create novel or to enhance existing pipelines with miRNA target identification and functional analysis tools. To this end, DIANA-microT web server v5.0 provides a complete integration with the Taverna Workflow Management System, using our in-house developed DIANA-Taverna Plug-in.

2.2.3.1 Description of the DIANA-Taverna Plugin Services

DIANA-microT-ANN (v4) service. The user can directly access the web server and identify miRNAs predicted to target selected genes AND/OR to find gene targets of selected miRNAs. The input/output ports of the DIANA-microT-ANN (v4) service are described below.

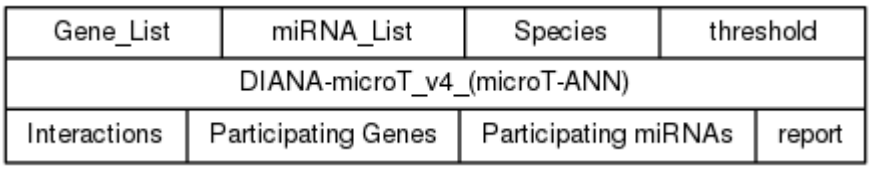

Figure 5: DIANA-microT-ANN (v4) service

The user has to specify the input ports of the DIANA-microT_v4 (microT-ANN) service in the Taverna plugin:

- *Gene_List*: DIANA-microT_v4 can be queried using a gene name/identifier or with a list of gene names/identifiers (gene names OR Ensembl v69 gene ids separated by a carriage return / newline character). Example value: FBgn0086758.
- *miRNA_List*: DIANA-microT_v4 can be queried with a miRNA name/identifier, or with a list of miRNA names/identifiers (miRNA names OR MIMAT ids are separated by a carriage return / newline character). Example value: dme-let-7-5p.
- threshold: A prediction score cut off value for presented predictions, ranging from 0.3 to 1. If no threshold is defined by the user, prediction results are provided for a default value of 0.7.

The output ports (provided results) of the DIANA-microT_v4 (microT-ANN) service in the Taverna plugin are presented below:

- *Interactions:* Predicted microRNA-gene interactions.
- *Participating Genes:* Ensembl v69 gene ids of the targets present in the predicted interactions.
- *Participating miRNAs:* mature miRNA names (miRBase v18) of the miRNAs taking part in the predicted interactions.
- *report:* General information about the provided results.

DIANA-microT-CDS (v5) Service. DIANA-microT-CDS service follows the exact same syntax as DIANA-microT v4, presenting the same input/output ports as shown below.

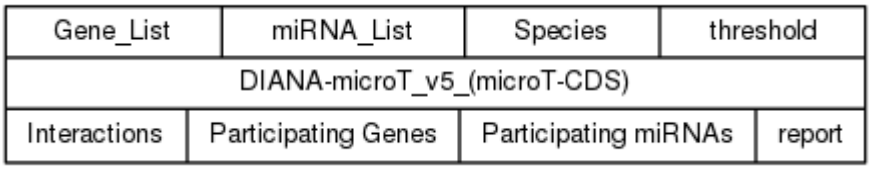

Figure 6: DIANA-microT-CDS (v5) Service

DIANA-TarBase v6.0 Service. This is a service to query directly DIANA-TarBase v6.0, the database indexing manually curated experimentally validated miRNA-gene interactions. The user can perform a query by using a gene name or Ensembl gene identifier (preferred) AND/OR miRNA name (miRBase 18+ nomenclature) / MIMAT ID. The input ports of the DIANA-TarBase v6.0 service are described below.

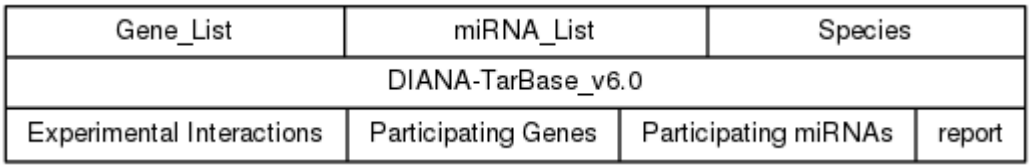

Figure 7: DIANA-TarBase v6.0 Service

The user has to specify at least one of the input ports of the DIANA-TarBase v6.0 service in the Taverna plugin:

- *Gene_List*: DIANA-TarBase v6.0 can be queried with a gene name/identifier or with a list of gene names/identifiers (gene names OR Ensembl 69 gene ids separated by a newline character). Example value: TUSC2.
- *miRNA_List*: DIANA-TarBase v6.0 can be queried with a miRNA name/identifier or with a list of miRNA names/identifiers (miRNA names OR MIMAT ids are separated by a newline character). Example value: hsa-let-7a-5p.

DIANA-miRPath v2.1 service. This service queries DIANA-miRPath server and identifies significantly targeted pathways by the selected miRNA(s). The miRNA-gene interactions can be derived directly from TarBase or can be computationally predicted using DIANA-microT-CDS. In case where more than one miRNAs are queried, DIANA miRPath identifies significantly targeted pathways by assessing the combinatorial effect of the selected miRNAs. The input/output ports of the DIANA-miRPath service are described below.

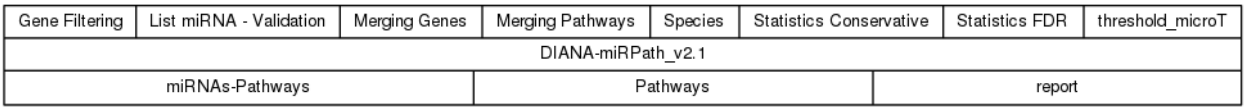

Figure 8: DIANA-miRPath v2.1 service

The user has to specify at least one of the input ports of the DIANA-miRPath service in the Taverna plugin:

- *Gene filtering:* A *url* to a file containing a list of gene name/identifiers separated by a newline character. The user can upload to a public url a predefined list of genes that are expressed in investigated tissues. MiRNA targets will be automatically filtered based on this list and will use only the expressed subset of genes for the pathway enrichment analysis.
- *List miRNA - Validation*: DIANA-miRPath can be queried with a miRNA name/identifier followed by the source of interactions (Tarbase or microT-CDS) or with a list of miRNA names/identifiers, each accompanied by the relevant interaction source (miRNA names OR MIMAT ids – interaction source pairs are separated by a newline character). miRNA name – interaction source terms can be separated by commas, spaces or tab characters. If no interaction source is provided for a miRNA then the service enables the microT-CDS as a default. Example value: hsa-mir-125b-5p Tarbase.
- *Merging Genes*: union/intersection
- *Merging Pathways*: union/intersection
- *Species*: E.g. human, mouse
- *Statistics Conservative*: true/false
- Statistics FDR: true/false
- *threshold_microT*: A cut off value for presented predictions (when microT-CDS is utilized as an interaction source), ranging from 0.3 to 1.

The output ports (provided results) of the DIANA-miRPath service in the Taverna plugin are the following:

- *miRNAs-Pathways:* DIANA-miRPath results, containing information such as Pathway KEGGid, Pathway description, Number of Associated genes, Gene Names, pValue Participating miRNAs.
- *Pathways:* A list with the pathways KEGGids significantly targeted by the selected miRNA(s).

report: General information about the provided results.

2.3 AGO-CLIP-Seq guided analysis for miRNA-target identification

The complex network of miRNA-lncRNA-mRNA regulatory machinery is difficult to be determined by exploring individual pairs of interactions. To this end, an in house algorithm was implemented in order to analyze CLIP-Seq data on different cell types and tissues for mouse and human species.

The analysis of CLIP-seq data is summarized in the following steps (Figure 9):

Preprocessing of deep sequencing data. Raw CLIP-Seq data were initially quality checked with FastQC (144) and further processed for contaminant removal with a combined use of Minion (145), Trimgalore (146) and Trimmomatic (147).

Alignment of reads. Alignment of CLIP-Seq reads against the reference genome was performed with GMAP/GSNAP (148), accordingly parameterized in order to identify reads in splice junctions.

Identification of CLIP-Seq enriched regions. Regions enriched in CLIP-Seq reads were formed by overlapping reads. In PAR-CLIP data, peaks were filtered to retain only regions with adequate T-to-C (sense strand) or A-to-G (antisense strand) incorporation in the same position (>5% of the reads).

Annotation of peaks. A comprehensive reference set of transcripts including mRNAs, lncRNAs and pseudogenes was utilized for the annotation of enriched CLIP-Seq regions.

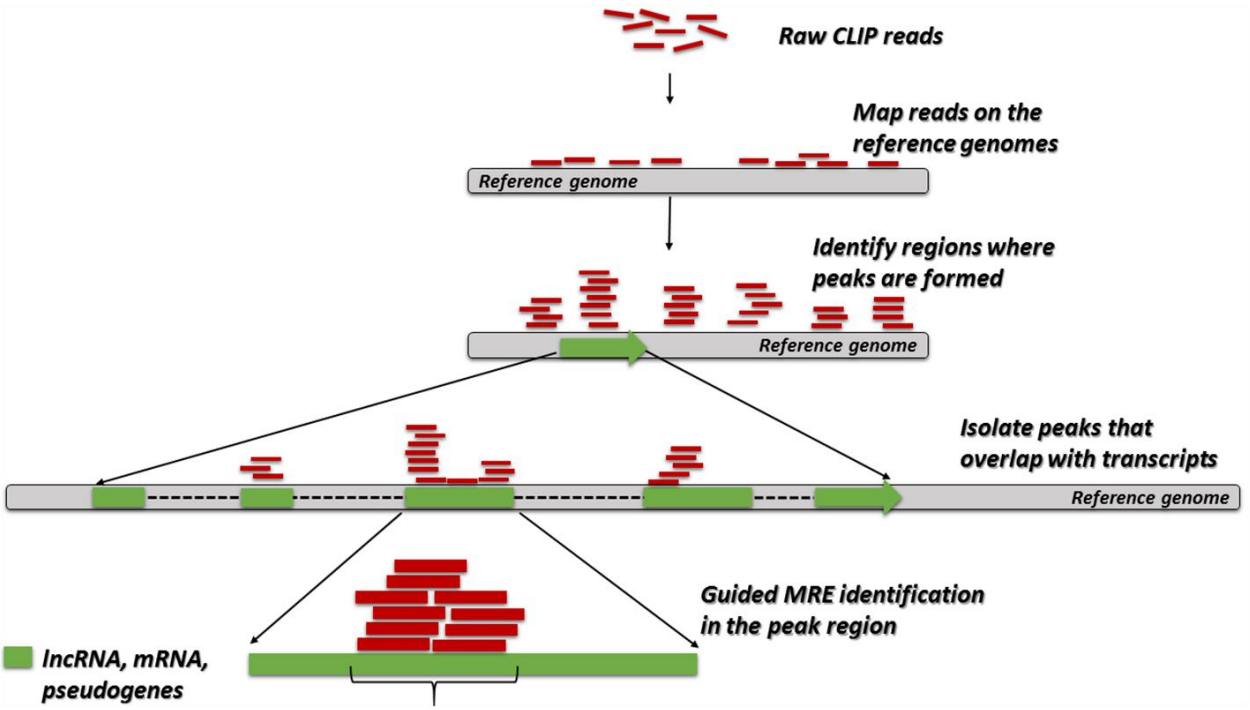

Figure 9: Raw CLIP-seq data were initially processed for contaminant removal and reads were aligned against the reference genome. Enriched regions in CLIP-Seq signal are formed from overlapping reads. Peaks were annotated in transcript loci. A CLIP-peak-guided MRE search algorithm was utilized to compute interactions of expressed miRNAs. (Paraskevopoulou MD *et al.*, 2016). (118)

Guided MRE identification. An in-house developed CLIP-peak-guided MRE search algorithm was subsequently utilized to identify interactions of expressed miRNAs. The algorithm utilizes the search space that is defined by the AGO binding peaks for MRE identification. It takes into account CLIP-Seq-induced mutations and the number of reads in peaks. The AGO enriched regions are subjected to an MRE detection algorithm. The implemented dynamic algorithm slides a 9 nucleotide-long window along each transcript and identifies the best possible alignment with the miRNA "extended" seed (nucleotides 1-9 on the miRNA 5'end). This procedure can detect different binding types, ranging from 6mer to 9mer (Table 4). The adopted pipeline includes features of miRNA binding type, miRNA-lncRNA/miRNA-mRNA duplex free energy, site accessibility, AU flanking content, and conservation. The CLIP-Seq-based characteristics are used to pinpoint the MRE location, while the miRNA-target binding features are combined and scored by a general linear model (GLM) classifier, as initially described by Rezcko *et al.* in microT-CDS algorithm (38), in order to identify the microRNA responsible for the binding (Figure 9).

Binding types
binding type 1 : 9mer Canonical (perfect seed match)
binding type 2 : 9mer
binding type 3 : 8mer Canonical (perfect seed match)
binding type 4 : 8mer
binding type 5 : 7mer Canonical (perfect seed match)
binding type 6 : 7mer
binding type 7 : 9mer with G:U wobble (8 matches + wobble + 3′ binding)
binding type 8 : 8mer with G:U wobble (7 matches + wobble + 3′ binding)
binding type 9 : 8mer with target bulge (8 matches + bulge + 3′ binding)
binding type 10 : 8mer with miRNA bulge (8 matches + bulge + 3′ binding)
binding type 11 : 8mer with mismatch
binding type 12 : 7mer with G:U wobble (6 matches + wobble + 3′ binding)
binding type 13 : 6mer Canonical
binding type 14 : 6mer (6 matches + 3′ binding)

Table 4: Different binding types from 6mer to 9mer identified by the adopted algorithm.

Use of biological replicates in CLIP-Seq data. In CLIP-Seq experiments that comprised biological replicates, an interaction had to be present in at least 2 samples (Figure 10). This filtering step increases the robustness of the identified results, minimizing technical and biological variability. The variance among different libraries was estimated as well as the percentage of target sites that were rejected after applying this filtering step. For variance estimation, the identified MREs were compared between CLIP-Seq replicates and the most deeply sequenced replicate (i.e. having the highest number of detected interactions). The use of the aforementioned filtering step removed a median of 63% (IQR = 3) of the total identified MREs in human CLIP-Seq libraries, and 40% (IQR = 14) in mouse. Certainly, a percentage of these MREs can be *bona fide* binding sites and not technical noise but it has been decided to prioritize robust and high quality experimentally supported results.

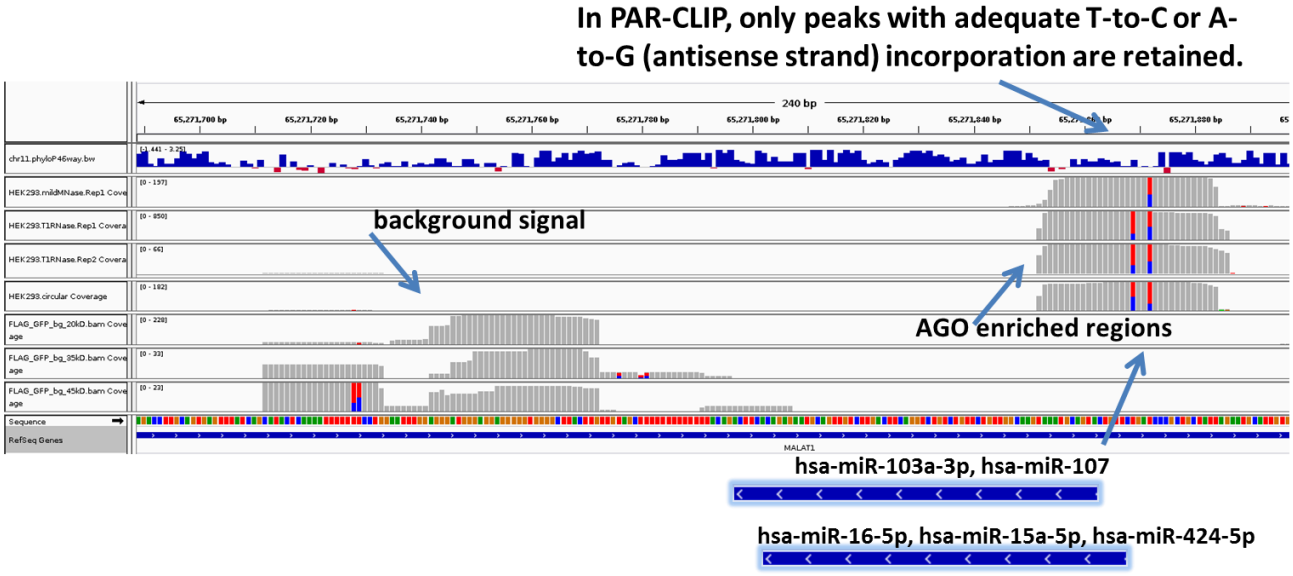

Figure 10: Example of identified MREs in PAR-CLIP AGO enriched regions. The peaks have adequate T-to-C incorporation and do not overlap with CLIP background signal.

2.4 Methods for the development of the DIANA-TarBase repository

Despite the evident advancements in the process of cataloguing miRNA targets, the majority of studies examining miRNA regulatory networks and their effect on molecular pathways usually rely on *in silico* predictions, since they require increased numbers of interactions. Aim of TarBase v7.0 (149) was to push the envelope further and to provide for the first time hundreds of thousands of high quality manually curated experimentally validated miRNA-gene interactions, enhanced with the most detailed meta-data available to date.

2.4.1 Text-mining pipeline selection of miRNA related articles

The number of publications that describe miRNA-mRNA regulation is increasing. The collection of the related literature is already considered as a demanding and time consuming practice. The manual curation can be assisted by text-mining pipelines successfully applied for the inquiry of miRNA-gene interactions.

The selection of the most information-rich articles for manual curation is a complex process, since thousands of manuscripts published per year have "microRNA" or "miRNA" keywords in their abstract or title. DIANA-TarBase 6.0 introduced a text-mining assisted pipeline for identification of articles which would be subsequently subjected to manual curation. This pipeline has been significantly extended and enhanced, in order to be able to capture all the advancements in the experimental methodologies. The text mining algorithm has been iteratively fine-tuned based on the feedback of curators following the analysis of hundreds of manuscripts.

In brief, the subset of MedLine articles having the terms "microRNA" or "miRNA" (and variations) in their title, abstract, keywords or MeSH terms are selected for analysis by the text mining algorithm. Abstracts and publication meta-data are downloaded in XML format from MedLine and subjected to Named Entity Recognition. Gene mentions were initially identified using AIIAGMT (150). The pipeline was subsequently updated to utilize GNAT libraries and online services (151) for gene name tagging and normalization. An extensive in-house-developed dictionary comprising all established, as well as novel experimental methodologies is utilized to recognize miRNAs, methods, important verbs and interaction terms. Sentences with a high probability for interaction (e.g. hosting gene, miRNA, and interaction terms) are scored based on the methods found within the text. Highest scored articles will be forwarded for manual curation, as well as articles containing high throughput methods relevant to miRNA function (e.g. AGO PAR-CLIP). The developed methodology has now been enhanced in order to be able to analyze freely-available complete articles and meta-data from PubMed Central. This pipeline has diminished the probability of curators analyzing low or no interaction articles, which pose a significant overhead in manual curation processes.

2.4.2 Collected Data

The types of collected data and meta-data have significantly increased, in order to facilitate the extensive testing and validation of prediction algorithms, as well as to empower regulatory investigations with experimentally derived miRNA-gene interactions. Each interaction is now accompanied with detailed information regarding the performed experimental procedure, including tissue, cell type and condition. Furthermore, a more relaxed database schema permits the description of more complex experiments and interactions involving multiple cell types or even species (e.g. miRNA-gene interactions between the host and a viral miRNA or *vice versa*, experiments where a 3'UTR from one species is being tested using a cell type of a different species).

Until now, databases usually distinguished experimental protocols into basic categories (e.g. specific and high-throughput) or into a handful of major methodology classes (e.g. Sequencing, Proteomics, Blotting, etc.). The new database schema enables the characterization of each methodology with two identifiers: a) a methodology class (12 classes) and b) a specific subtype (20 method subtypes). By utilizing twin-identifiers, it is now possible to distinguish two closely related methods that have different information content (e.g. biotin pull-down of miRNA targets + microarray transcript quantification vs biotin pull-down + qPCR transcript quantification).

A new field has been introduced to the database schema for marking interactions derived from chimeric reads from CLASH or modified CLIP-Seq experiments. These interactions have higher information content, since miRNA and mRNA sequences reside on the same read, enabling the accurate identification of both actors, as well as the exact site of the interaction. Even though these high quality interactions are currently limited, the new database schema enables their detailed cataloguing.

Specific attention was paid on archiving the exact binding site of each interaction, since such information is crucial for testing target prediction algorithms or for identifying regulatory regions on a transcript (e.g. deciphering the effect of a variant on a 3'UTR region). The curation pipeline was extended with tools and techniques that enabled the curators to identify targeted regions using any relevant information available within the manuscript or supplemental material (genomic/transcript coordinates, cloning primers, mutation sites, etc). Any experimental information used by the curators for the identification of the targeted regions is kept within the database. Binding sites were also identified by analyzing an extensive array of CLIP-Seq methods. By including binding-site level data into the database, TarBase v7.0 can present positive/negative results from experimental validations of distinct binding sites on the same transcript.

Details concerning the database of experimentally supported miRNA-mRNA interactions and the updated interface of TarBase v7.0 are provided in the relevant result sections.

2.5 Methods for the development of the DIANA-LncBase repository

DIANA-LncBase v1 (119) is considered as the first extensive compendium dedicated to cataloguing miRNA-lncRNA interactions and providing the largest collection of experimentally supported entries. One of the major aims of this thesis was to extensively study miRNA-(non)coding targets and to provide further insights for this still obscure mechanism. To this end, the largest collection of (in)direct low and high-throughput methodologies and relevant publications was compiled. The analyzed experiments span numerous cell types across different experimental conditions for human and mouse species. Since lncRNA function is characterized by tissue specificity, a large number of RNA sequencing data was processed to complement miRNA-lncRNA putative interactions with transcript expression. This wealth of information and results inferred from the analysis were included in the updated version of LncBase.

2.5.1 Collected Data

An extensive collection of manuscripts has been manually curated, while more than 150 raw NGS datasets harboring miRNA interactions with (non)coding transcripts were analyzed, in order to unveil and explore the lncRNA target-mimetic function.

Experimental methodologies. miRNA-lncRNA experimentally supported interactions from low yield and high-throughput methodologies were extracted from manually curated publications and raw sequencing data. LncBase v2 supports miRNA-lncRNA interactions derived from more than 150 CLIP-Seq (24 PAR-CLIP, 129 HITS-CLIP) libraries across a wide range of cell types, corresponding to the largest collection of AGO-CLIP data compared to any other relevant resource.

2.5.2 Tissue/cell type expression

Collected expression data. Raw RNA-Seq datasets were retrieved from ENCODE (2,3), UCSC (152) and Gene Expression Omnibus (GEO) (153) repositories in order to assess lncRNA transcript expression in a wide range of cell types for both human and mouse species. RNA-Seq data corresponding to similar cell types with those in CLIP-Seq samples were preferentially selected. All RNA-Seq libraries were depleted of ribosomal RNAs. Whole transcriptome and poly-A selected libraries were analysed. The analysis of deeply sequenced RNA samples enabled the extensive identification of expression patterns for targeted lncRNAs. Details concerning the accession codes of the processed RNA-Seq samples and library specifications are provided in **Table 5**. Raw datasets were quality checked and pre-processed to minimize contaminant sequences. Expression at transcript level was estimated using RSEM (154). Raw reads were aligned against human transcriptomes compiled from Ensembl 75 (GRCh37), RefSeq Release 106 (GRCh38) (142) and Cabili *et al.* (88) as well as mouse transcriptomes derived from Ensembl 81 (GRCm81) (155) and RefSeq Release 104 (GRCm38.p2). Transcript expression information, extracted from analysed RNA-Seq data across 24 tissues and cell types in Cabili *et al.*, was also incorporated in LncBase v2.

Accession	Repository	Cell Type/Tissue	Total Reads	Species	Sequencing
ENCFF001REK ENCFF001REJ	encodeproject.org	GM12878	195M	*Homo sapiens*	PE, 101bp
ENCFF000FOM ENCFF000FOV	encodeproject.org	HeLa-S3	242M	*Homo sapiens*	PE,76bp
ENCFF000GET ENCFF000GEQ	encodeproject.org	HMepC	293M	*Homo sapiens*	PE, 101bp
ENCFF002DKX ENCFF002DKY	encodeproject.org	MCF-7	121M	*Homo sapiens*	PE,100bp
ENCFF109IUU ENCFF322VHJ	encodeproject.org	HREpiC	212M	*Homo sapiens*	PE,101bp
ENCFF000GHA ENCFF000GGZ	encodeproject.org	hMSC-BM	379M	*Homo sapiens*	PE, 101bp
wgEncodeCshlLongRnaSeqA549CellPapFastq - Rep1	hgdownload.cse.ucsc.edu	A549	190M	*Homo sapiens*	PE,76bp
wgEncodeCshlLongRnaSeqHepg2CellPapFastq – Rep1	hgdownload.cse.ucsc.edu	HepG2	248M	*Homo sapiens*	PE,76bp
wgEncodeCshlLongRnaSeqHuvecCellPapFastq – Rep1	hgdownload.cse.ucsc.edu	HUVEC	174M	*Homo sapiens*	PE,76bp
wgEncodeCshlLongRnaSeqK562CellPapFastq – Rep1	hgdownload.cse.ucsc.edu	K562	227M	*Homo sapiens*	PE,76bp
wgEncodeCshlLongRnaSeqSknshraCellPapFastq – Rep2	hgdownload.cse.ucsc.edu	SK-N-SH	234M	*Homo sapiens*	PE,76bp
wgEncodeCshlLongRnaSeqH1hescCellPapFastq – Rep1	hgdownload.cse.ucsc.edu	H1 hESC	250M	*Homo sapiens*	PE,76bp
wgEncodeCshlLongRnaSeqImr90CellPapFastq – Rep1	hgdownload.cse.ucsc.edu	IMR90	217M	*Homo sapiens*	PE,101bp
wgEncodeCshlLongRnaSeqWbrainE14halfFastq – Rep1	hgdownload.cse.ucsc.edu	Brain	341M	*Mus musculus*	PE,101bp
wgEncodeCshlLongRnaSeqHeartAdult8wksFastq –Rep1	hgdownload.cse.ucsc.edu	Heart	149M	*Mus musculus*	PE,76bp
wgEncodeCshlLongRnaSeqKidneyAdult8wksFastq – Rep1	hgdownload.cse.ucsc.edu	Kidney	186M	*Mus musculus*	PE,76bp
wgEncodeCshlLongRnaSeqLiverAdult8wksFastq – Rep1	hgdownload.cse.ucsc.edu	Liver	160M	*Mus musculus*	PE,76bp
wgEncodeCshlLongRnaSeqLungAdult8wksFastq – Rep1	hgdownload.cse.ucsc.edu	Lung	141M	*Mus musculus*	PE,76bp
wgEncodeCshlLongRnaSeqThymusAdult8wks – Rep1	hgdownload.cse.ucsc.edu	Thymus	160M	*Mus musculus*	PE,76bp
ENCFF001IDD ENCFF001ICW	encodeproject.org	C2C12 (60h)	280M	*Mus musculus*	PE,75bp
ENCFF001IUF ENCFF001IUD	encodeproject.org	Frontal Cortex	371M	*Mus musculus*	PE, 101bp
ENCFF001NEG NCFF001NEC	encodeproject.org	MEL	281M	*Mus musculus*	PE, 101bp
GSM973235	ncbi.nlm.nih.gov/geo	ES-E14	341M	*Mus musculus*	PE, 101bp
GSM1370364	ncbi.nlm.nih.gov/geo	HEK-293	395M	*Homo sapiens*	PE, 50bp
GSM1133247	ncbi.nlm.nih.gov/geo	LCLBAC	68M	*Homo sapiens*	PE, 50bp
GSM1133250	ncbi.nlm.nih.gov/geo	LCLBACD2	45M	*Homo sapiens*	PE, 50bp
GSM1133251	ncbi.nlm.nih.gov/geo	LCLBACD3	74M	*Homo sapiens*	PE, 50bp
GSM1133248	ncbi.nlm.nih.gov/geo	LCLBACD1	77M	*Homo sapiens*	PE, 50bp
GSM1133249	ncbi.nlm.nih.gov/geo	LCLBACD1	74M	*Homo sapiens*	PE, 50bp

Table 5: Details concerning the analysed RNA-Seq samples. The table presents accession codes and sequencing specifications for each library. RNA-Seq datasets were retrieved from ENCODE(2,3), UCSC(152) and Gene Expression Omnibus (GEO)(153) repositories in order to assess lncRNA transcript expression in various cell types and tissues. (Paraskevopoulou MD *et al*, 2015)(117)

Further information concerning the database of experimentally supported miRNA-lncRNA interactions and the updated LncBase v2 interface are provided in the relevant result sections.

2.6 Comparison of TarBase/LncBase AGO-CLIP-Seq data analysis algorithm with other CLIP-Seq Target Identification applications

For the evaluation of the in-house developed CLIP-Seq data analysis algorithm, different implementations identifying miRNA targets with CLIP-Seq have been utilized. The list of tested algorithms included microMUMMIE (75), MIRZA (72), and PARMA (74) The evaluation of programs' performance was based on their accuracy in predicting both miRNA-mRNA interactions, as well as their ability to correctly identify experimentally verified miRNA binding sites.

The computational algorithms were assessed for their performance in distinct high quality validation sets comprising ~300 Luciferase Reporter Gene Assays and ~1,700 chimeric interactions in HEK293T cells, respectively. The chimeric interactions were retrieved from 1 CLASH library (156). An additional evaluation of TarBase/LncBase AGO-CLIP-Seq algorithm incorporated an extended set of ~850 interactions validated with Luciferase Reporter Gene Assays.

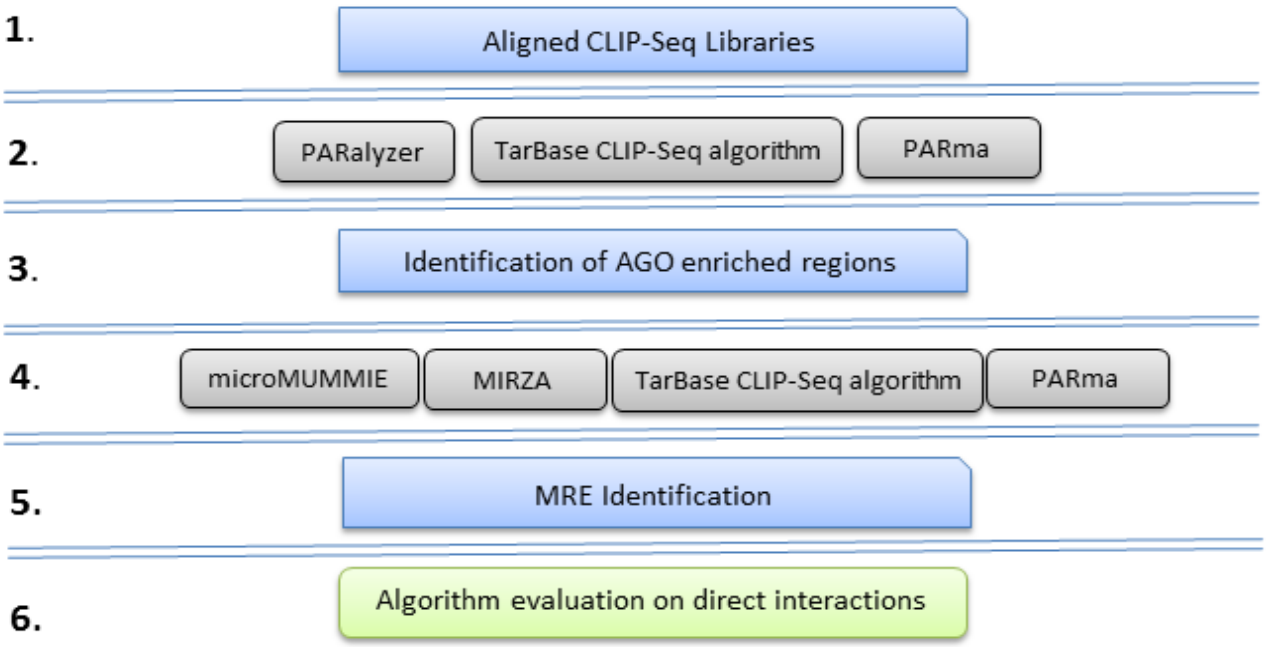

Figure 11: Summary of the performance evaluation pipeline for CLIP-Seq analysis algorithms. SAM files produced by different aligners were utilized for CLIP target identification. Total predicted MREs (miRNA Recognition elements) in CLIP-Seq enriched regions were filtered in order to retain only miRNAs and transcripts contained in the validation set composed of 2,000 Reporter gene and chimeric miRNA interactions. *(Copyright Paraskevopoulou Maria)*

2.7 Implementation of a novel Algorithm for the AGO CLIP-Seq data analysis

Most of the current *in silico* implementations devoted to the analysis of AGO-CLIP sequencing data still lack a robust A-to-Z pipeline in order to sufficiently catalogue miRNA-target interactions.

MIRZA algorithm does not support the direct processing of the raw or aligned CLIP-Seq data, and requires to be supplied with a specific format for the clusters and the miRNAs. In addition, it has several limitations such as the length of the provided clusters (30-51 nts) and miRNAs (21 nts), the required supplementary information of miRNA expression values and the formation of clusters centered on the position with maximum T-to-C conversion sites.

microMUMMIE has to be complemented with other implementations, which are considered essential for its core algorithm. However, these extra steps of calculations are not seamlessly incorporated in the microMUMMIE pipeline but have to be generated independently by the user. Moreover, microMUMMIE mainly focuses on the analysis of miRNA binding sites in the 3'UTR regions, even though CLIP-Seq experiments can be efficiently applied to discover transcriptome-wide miRNA interactions.

The main restriction of PARma is that it adopts a family miRNA-seed clustering approach and relies on the identification of miRNA-seed binding sites in AGO-peaks comprising statistically significant overrepresented kmers. It also requires a specific input format of AGO enriched regions with relevant conversion sites that has to be prepared by the user. Notably, the latter two implementations do not cover the whole spectrum of miRNA binding types.

All the aforementioned algorithms are preferably applied for the analysis of PAR-CLIP data. They are not appropriate for the processing of other CLIP-Seq experiments including, HITS-CLIP, CLASH or iCLIP. This is due to the fact that they strongly depend on the induced T-to-C conversions in the AGO crosslinked regions to pinpoint miRNA binding sites. Moreover, they do not process AGO enriched regions that do not have T-to-C substitutions, omitting a large amount of highly covered PAR-CLIP clusters. Importantly, the evaluation of the described implementations against the adopted AGO-CLIP-Seq analysis pipeline of TarBase and LncBase repositories revealed that there is room for further improvement for all algorithms and optimization in order to attain increased accuracy.

It should also be noted that there are no available implementations incorporating the wealth of high/low throughput released experiments specific for miRNA-gene interactions. To this end, a novel Algorithm was developed primarily for PAR-CLIP data analysis, with the potential to be generalized for other CLIP-Seq variants. The collected low-yield and high-throughput experimental data sources for the derivation of positive and negative miRNA-target interactions as well as the algorithm's deployment and testing are described in the following sections.

2.7.1 Collection of experimental datasets

A comprehensive collection of low/high-throughput experimental datasets was created in order to extract putative miRNA-target interactions. More precisely, Reporter gene Assays, CLASH, CLEAR-CLIP, PAR-CLIP, RNA-Seq, microarrays, quantitative proteomics (pSILAC), Ribosome profiling sequencing (Ribo-Seq) were utilized to generate positive and negative instances. Direct interactions retrieved from Reporter Gene Assay techniques and high quality miRNA-target chimeras derived from CLASH and CLEAR-CLIP constitute a source of specific MRE regions and were included as positive cases. On the other hand, indirect high-throughput methodologies such as RNA-Seq and microarrays are experiments that measure mRNA expression changes after transfection, silencing or knockout of a specific miRNA and therefore were processed for the derivation of both positive and negative instances. Ribosome profiling sequencing after miRNA overexpression can reveal differences in ribosome-bound transcripts and for that reason it is a valuable component for detecting functional (positive) miRNA effects or negative instances. pSILAC experiments were also included in the training set since they can reveal the strong or weak impact of a miRNA deregulation on protein concentration.

Friedersdorf M and Keene J (157) generated background PAR-CLIP libraries aiming to study non-specific RBP binding events and reveal patterns of true protein-RNA interactions. These datasets were additionally incorporated to deduce negative miRNA binding sites in the respective CLIP-Seq clusters.

Finally, random CLIP data, at the level of raw reads, were generated in order to provide an extra source for the creation of decoy clusters and MRE regions.

2.7.1.1 Direct miRNA-target interactions derived from high/low throughput techniques

The positive collection incorporates interactions retrieved from 377 publications and comprises more than 30,000 direct miRNA-target interactions, spanning approximately 200 cell types. Positive cases validated with Luciferase Reporter assays are obtained from DIANA-TarBase v7.0 (64). Luciferase expression vectors are usually tested with whole 3'UTR mRNA sequences that may harbor more than one candidate miRNA binding sites. However, TarBase repository also indexes a considerable amount of luciferase data, where specific candidate binding regions are cloned in the relevant vectors. To this end, only short RNA fragments (<200 nts) tested with reporter assays were included in the positive set. These instances correspond to miRNA-target interactions spanning more than 40 tissues, while the majority of them are tested on Human Embryonic Kidney (HEK-293), Mammary Gland (MCF7 or MDAMB231) and Cervix tissue (HeLa).

Chimeric miRNA-target fragments are derived from two CLASH (50) and CLEAR-CLIP (59) experiments. CLEAR-CLIP has been performed on a neoplastic cell line (Huh7.5) in

liver tissue, while CLASH-supported interactions correspond to T-REx 293 cells. These two datasets comprise 28,000 miRNA direct interactions.

Moreover, Grosswendt S *et al.* (51) observed the existence of miRNA-target ligated pairs in already published PAR-CLIP experiments. The authors introduced an *in silico* pipeline for the identification of chimeric miRNA-gene fragments, and they applied their method on already released experiments to form a collection of such events. A selected set of these precompiled chimeric miRNA interactions were included in the algorithm. These interactions cover 5 different cell lines: Human Embryonic Kidney cells (HEK293, Kishore et al. (73)), BC-1 and BC-3 primary effusion lymphoma-derived cell lines infected with Epstein-Barr virus (EBV) and Kaposi's sarcoma-associated herpesvirus (KSHV) (Gottwein et al. (158)), EBV-infected lymphoblastoid cell lines (Skalsky et al. (159)) and Human Embryonic stem Cells (hESC, Lipchina et al. (160)). A concise description of the positive miRNA interactions from the different experiments is provided in Table 6.

Experiment	Species	Cell Line	Number of miRNAs	Interactions	Studies
Luciferase Reporter	human	197	165	714	371
CLASH	human	1	176	1,573	1
CLEAR-CLIP	human	1	482	27,335	1
Chimeric miRNA-targets (Grosswendt S et al.)	human	4	262	12,511	4

Table 6: Summary of the positive miRNA interactions and associated cell types, derived from the different direct experiments.

2.7.1.2 RNA-Seq datasets

A set of 9 different experimental conditions (shown in Table 7), corresponding to RNA sequencing datasets, were analyzed in order to infer positive and negative gene changes after miRNA overexpression. In total, the transcriptome-wide differential expression in three human cell lines (HEK-293T, HeLa and U2OS) was calculated for two miRNAs (miR-1 and miR-155). These datasets were released from a recent publication by Eichhorn *et al.* (161).

RNA Sequencing Datasets				
#	**Cell Line**	**miRNA**	**miRNA treatment**	**Post-Transfection Cell Harvest Time/Experimental Condition**
1	HEK-293T	hsa-miR-1-3p	Overexpression	24h
2	HELA	hsa-miR-1-3p	Overexpression	24h
3	HELA	hsa-miR-155-5p	Overexpression	24h
4	U2OS (total)	hsa-miR-1-3p	Overexpression	32h/poly(A)-selected total RNA
5	U2OS (total)	hsa-miR-155-5p	Overexpression	32h/poly(A)-selected total RNA
6	U2OS (cyto)	hsa-miR-1-3p	Overexpression	32h/poly(A)-selected cytoplasmic RNA
7	U2OS (cyto)	hsa-miR-155-5p	Overexpression	32h/poly(A)-selected cytoplasmic RNA
8	U2OS (ribo)	hsa-miR-1-3p	Overexpression	tRNA and rRNA depleted RNA
9	U2OS (ribo)	hsa-miR-155-5p	Overexpression	tRNA and rRNA depleted RNA

Table 7: Description of RNA Sequencing datasets after miRNA overexpression utilized to extract positive and negative instances for the training of a novel AGO-CLIP-Seq-guided Algorithm for miRNA-target identification.

2.7.1.3 Microarray datasets

Different experimental conditions (shown in Table 8) were analyzed from 52 microarray studies. In total, the transcriptome-wide differential expression in 53 human cell lines was calculated for 65 miRNAs that were either overexpressed or knocked-down/out. Human cell lines from Affymetrix chips were analyzed. Affymetrix microarray raw files (.CEL) from experiments listed in Supplementary Table 8 were analyzed in-house. miRNA-treated and control samples were appropriately combined in order to perform background correction, quantile normalization and log2 expression calculation. These processing steps were implemented using Robust Multi-Array Average (RMA) with *affy* (162) or *oligo* (163) R-packages. Annotation enrichment of each probe set was accomplished using the chip-specific annotation R-packages *hgu133a2.db, hgu133plus2.db* or *hugene10sttranscriptcluster.db*. Each experiment was examined independently of other cell lines or miRNA treatments. log2(FC) and p-values were calculated with *limma* package (164), following the guidelines for Single-Channel Designs.

Importantly, since microarray analyses were performed at a probe set level, there was a considerable portion of gene instances comprising one-to-many associations (gene referring to multiple probe sets). In these cases, a majority rule was applied to same-gene probe sets in order to determine up/down-regulation of transcript expression. Subsequently, a median log2(FC) was calculated including only the gene-associated probe sets that exceeded either a positive or negative threshold (>0.5 or <-0.5, respectively), depending on the type of the regulation decided by the majority rule in the previous step. This probe-to-gene level transition allowed the incorporation of deregulated genes derived from microarray analyses, into the positive and negative training sets.

Summary of Microarray Datasets				
#	**Cell Line**	**miRNA**	**miRNA treatment**	**Post-Transfection Cell Harvest Time**
1-2	113/6-4L, 131/4-5B1	hsa-miR-30d-5p	Overexpression	60h
3	AGS	hsa-miR-210-3p	Overexpression	36h
4	CALU3	hsa-miR-138-5p	Overexpression	48h
5-7	CCL86, CRL1432, CRL1596	hsa-miR-26a-5p	Overexpression	72h
8	DLD1	hsa-miR-143-3p	Overexpression	24h
9	DLD1	hsa-miR-145-5p	Overexpression	24h
10	DU145	hsa-miR-224-5p	Overexpression	48h
11	DU145	hsa-miR-452-5p	Overexpression	48h
12	H4	hsa-miR-103a-3p	Overexpression	48h
13	H4	hsa-miR-107	Overexpression	48h
14	H4	hsa-miR-15b-3p	Overexpression	48h
15	H4	hsa-miR-16-5p	Overexpression	48h
16	H4	hsa-miR-195-5p	Overexpression	48h
17	H4	hsa-miR-320b	Overexpression	48h
18	HEK-293	hsa-miR-212-3p	Overexpression	-
19	HEK-293	hsa-miR-124-3p	Overexpression	15h
20	HEK-293	hsa-miR-7-5p	Overexpression	15h
21-22	HELA	hsa-let-7b-5p	Overexpression	8h, 32h
23-24	HELA	hsa-miR-1-3p	Overexpression	8h, 32h
25-26	HELA	hsa-miR-155-5p	Overexpression	8h, 32h
27-28	HELA	hsa-miR-16-5p	Overexpression	8h, 32h
29-30	HELA	hsa-miR-30a-5p	Overexpression	8h, 32h
31	HEPG2	hsa-miR-191-5p	Anti-miR	-
32-38	HEPG2	hsa-miR-124-3p	Overexpression	4h, 8h, 16h, 24h, 32h, 72h , 120h
39	HEY	hsa-miR-429	Overexpression	48h
40	HEY	hsa-miR-128-3p	Overexpression	48h
41	HEY	hsa-miR-7-5p	Overexpression	48h
42	HUH7	hsa-miR-517a-3p	Overexpression	-
43	HUH7.5	hsa-miR-27a-3p	Anti-miR	-
44	HUH7.5	hsa-miR-27a-3p	Overexpression	-
45-46	HUVEC	hsa-miR-210-3p	Anti-miR, Overexpression	24h
47	HUVEC	hsa-miR-126-3p	Anti-miR	72h
48	IMR90	hsa-miR-29a-3p	Knock-down	48h
49	K562	hsa-miR-34a-5p	Overexpression	24h
50	LNCAP	hsa-miR-106b-5p	Overexpression	24h
51	LNCAP	hsa-miR-130a-3p	Overexpression	24h
52	LNCAP	hsa-miR-203a-3p	Overexpression	24h
53	LNCAP	hsa-miR-205-5p	Overexpression	24h
54	LNCAP	hsa-miR-1-3p	Overexpression	24h
55	LNCAP	hsa-miR-206	Overexpression	24h
56	LNCAP	hsa-miR-27b-3p	Overexpression	24h
57-60	MCF10A	hsa-miR-20a-5p	Silencing	0h, 0.5h, 1h, 2h post EGF stimulation
61-64	MCF10A	hsa-miR-671-5p	Silencing	0h, 0.5h, 1h, 2h post EGF stimulation

65	MCF7	hsa-miR-95a-3p	Overexpression	24h
66	MCF7	hsa-miR-101-3p	Overexpression	24h
67	MCF7FR	hsa-miR-221-3p	Silencing	72h
68	MCF7FR	hsa-miR-222-3p	Silencing	72h
69	MHH-ES-1	hsa-miR-483-5p	Overexpression	48h
70	MHH-ES-1	hsa-miR-483-3p	Overexpression	48h
71	MKN45	hsa-miR-210-3p	Overexpression	36h
72	MSK543	hsa-miR-124-3p	Overexpression	24h
73	MSK543	hsa-miR-380-3p	Overexpression	24h
74	MSK543	hsa-miR-433-3p	Overexpression	24h
75	MSK543	hsa-miR-448	Overexpression	24h
76	MSK543	hsa-miR-132-3p	Overexpression	24h
77	PAG C81-61	hsa-miR-20a-5p	Overexpression	3d
78	PAG C81-61	hsa-miR-17-5p	Overexpression	3d
79	PC3	hsa-miR-224-5p	Overexpression	48h
80	PC3	hsa-miR-452-5p	Overexpression	48h
81	SKHEP1	hsa-miR-21-5p	Anti-miR	16h
82	SW1783	hsa-miR-376a-5p	Overexpression	24h
83-84	U87	hsa-miR-376a-5p	Overexpression	24h, 72h
85-86	U87, HS683	hsa-miR-20a-5p	Overexpression	-
87	HTERT-RPE1	hsa-miR-129-2-3p	Overexpression	72h
88	FLS	hsa-miR-23b-3p	Overexpression	-
89	HAEC	hsa-miR-34a-5p	Overexpression	48h
90	HAEC	hsa-miR-34b-5p	Overexpression	48h
91	HAEC	hsa-miR-34c-5p	Overexpression	48h
92	HAEC	hsa-miR-449b-5p	Overexpression	48h
93	HAEC	hsa-miR-449a	Overexpression	48h
94	HDF	hsa-miR-29a-3p	Inhibition	48h
95-97	GBM4, GBM6, GBM8	hsa-miR-10b-5p	Inhibition	24h
98-100	HEK-293, HEK-293T, HSF2	hsa-miR-941	Overexpression	24h
101	HT29	hsa-miR-146a-5p	Overexpression	2w after lentiviral infection
102	H929	hsa-miR-214-3p	Overexpression	-
103	MDAMB231	hsa-miR-200c-3p	Overexpression	-
104	MDAMB231	hsa-miR-205-5p	Overexpression	-
105	MDAMB231	hsa-mir-375	Overexpression	-
106	U1810	hsa-miR-214-3p	Antagomir	24h
107	HCT116	hsa-miR-34a-5p	Overexpression	2w after retroviral infection
108	HCT116	hsa-miR-147a	Overexpression	3d
109	SUM159	hsa-miR-203a-3p	Overexpression	-
110	U87-2M1	hsa-miR-10b-5p	Inhibition	-
111	A549	hsa-miR-7-5p	Overexpression	24h
112-113	Jurkat	hsa-miR-146a-5p	Overexpression, Knock-down	48h
114	Melanoma-metastatic Liver Cells	hsa-miR-182-5p	Anti-miR	administered twice per week over 4 weeks
115-116	P3HR1	hsa-miR-28-5p	Overexpression	12h, 24h

Table 8: Description of miRNA inhibition/overexpression/KO microarray datasets utilized to extract positive and negative instances for the training of a novel AGO-CLIP-Seq-guided algorithm for miRNA-target identification.

2.7.1.4 Ribosome Profiling Datasets

Ribosome Profiling Datasets. Ribo-Seq datasets that correspond to 5 experimental conditions were retrieved from a recent publication by Eichhorn *et al.* (161). As described in Table 9, the data refer to three human cell lines (HEK-293T, HeLa and U2OS) and two human miRNAs (miR-1 and miR-155) that were overexpressed in the relevant experiments. Fold change values as calculated from differential expression analyses of control vs post-transfection states enabled the formation of positive and negative miRNA-mRNA interactions.

	Ribosome Profiling Datasets			
#	**Cell Line**	**miRNA**	**miRNA treatment**	**Post-Transfection Cell Harvest Time**
1-2	HEK-293T, HELA	hsa-miR-1-3p	Overexpression	24h
3	HELA	hsa-miR-155-5p	Overexpression	24h
4	U2OS	hsa-miR-1-3p	Overexpression	32h
5	U2OS	hsa-miR-155-5p	Overexpression	32h

Table 9: Description of ribosome profiling datasets after overexpression of a specific miRNA. These sets were utilized to extract positive and negative instances for the training of a novel Algorithm for the analysis of AGO CLIP-Seq data.

2.7.1.5 Quantitative Proteomics Datasets

A collection of 6 distinct pSILAC (provided in Table 10) experimental datasets were derived from the Selbach *et al.* publication (55). In this study, quantitative proteome-wide profiles were assessed in HeLa cells following the individual overexpression of 5 human miRNAs (let-7b, miR-1, miR-16, miR-30a and miR-155) or knock-down of let-7b. The precompiled median log2(Fold-change) values from relevant publication were accordingly processed to deduce miRNA-gene associations reflecting the positive/negative impact of miRNA overexpression to protein concentration.

	pSILAC Datasets			
#	**Cell Line**	**miRNA**	**miRNA treatment**	**Post-Transfection Cell Harvest Time**
1	HELA	hsa-let-7b-5p	Overexpression	8h post-transfection and 24h pSILAC labelling
2	HELA	hsa-miR-1-3p	Overexpression	8h post-transfection and 24h pSILAC labelling
3	HELA	hsa-miR-16-5p	Overexpression	8h post-transfection and 24h pSILAC labelling
4	HELA	hsa-miR-30a-5p	Overexpression	8h post-transfection and 24h pSILAC labelling
5	HELA	hsa-miR-155-5p	Overexpression	8h post-transfection and 24h pSILAC labelling
6	HELA	hsa-let-7b-5p	Knock-down	8h post-transfection and 24h pSILAC labelling

Table 10: Description of miRNA overexpression/KO pSILAC datasets utilized to extract positive and negative instances for training a novel Algorithm for the analysis of AGO CLIP-Seq data.

Differential expression analyses of miRNA-(un)treated cell lines were performed for the aforementioned high-throughput experimental datasets. More precisely, the entries corresponding to positive and negative miRNA-mRNA interactions for microarrays and quantitative proteomics (pSILAC) experiments were defined by applying a strict -1 or +1 log2(Fold Change) threshold.

For Ribosome profiling and RNA sequencing experiments, gene expression values were initially filtered with a threshold of >10 RPKM. Subsequently, the remaining genes in these experiments (Ribo/RNA-Seq) were selected with a -0.5 or 0.5 log2(Fold Change) threshold for positive and negative interactions respectively.

The selection of fold change thresholds was performed after observation of their distribution in each of the processed datasets. Notably, since multiple datasets were integrated for the algorithm development, it has been observed that specific miRNA-gene interactions appeared to be both positive and negative in different experimental settings. Such conflicting outcomes were removed.

2.7.1.6 CLIP deep sequencing datasets

A collection of 24 PAR-CLIP datasets derived from 8 studies were incorporated to the pipeline (Table 11). Each independent experiment provided AGO cluster information comprising the signal of raw aligned reads and transition sites.

These AGO-bound clusters were combined with the positive and negative miRNA-target interactions, as identified by different low and high-throughput experiments, in order to infer multiple descriptors for each targeted region. MRE regions located within PAR-CLIP peaks were subsequently utilized for feature extraction.

It should be noted that indirect experiments cannot provide the exact MRE region. In order to address this issue, an extra step was included to scan transcripts participating is indirect interactions for miRNA-specific binding sites. This analysis, in many cases, revealed more than one candidate MRE per miRNA-target pair. In such instances, identifying overlapping MREs with AGO clusters introduced one or more positive or negative instances in the training set.

The following sections describe the derivation of extra negative miRNA-target instances from background PAR-CLIP and randomly simulated PAR-CLIP experiments respectively.

	Experiment	Species	Cell line	Cell condition	Samples
1-2	PAR-CLIP	human	HEK293	enzymatic digestion: complete T1 digestion, protein: Ago2	2
3-4	PAR-CLIP	human	HEK293	enzymatic digestion: mild MNase digestion, protein: Ago2	2
5	PAR-CLIP	human	MCF7	monoclonal anti-AGO2 (C1.9E8.2)	1
6	PAR-CLIP	human	hESC	monoclonal anti-AGO2 (C1.9E8.2)	1
7	PAR-CLIP	human	C8166	hiv-1 strain: NL4-3, length of infection (days): 3	1
8	PAR-CLIP	human	TZM-bl	hiv-1 strain: WT/BaL, length of infection (days): 3	1
9	PAR-CLIP	human	TZM-bl	hiv-1 strain: WT/BaL, length of infection (days): 3, engineered cells to stably express HIV-1-specific amiRNAs	1
10	PAR-CLIP	human	BC-1	primary effusion lymphoma (PEL) cell line, latently infected with both KSHV and EBV	1
11	PAR-CLIP	human	BC-3	primary effusion lymphoma (PEL) cell line, latently infected only with KSHV	1
12	PAR-CLIP	human	EF3DAGO2	EBV B95-8-infected lymphoblastoid cells, antibody: Anti-Ago2 (clone 9E8)	1
13	PAR-CLIP	human	LCL35	EBV B95-8-infected lymphoblastoid cells, antibody: Anti-Ago2 (clone 9E8)	1
14	PAR-CLIP	human	LCLBAC	LCL-BAC lymphoblastoid cells infected by EBV B95-8 BACmid, antibody: Anti-Ago2 (clone 9E8)	1
15	PAR-CLIP	human	LCLBACD1	LCL-BACD1 lymphoblastoid cells infected by EBV B95-8 BACmid, mutationally inactivated for miR-BHRF1-1 expression, antibody: Anti-Ago2 (clone 9E8)	1
16	PAR-CLIP	human	LCLBACD3	LCL-BACD3 lymphoblastoid cells infected by EBV B95-8 BACmid, mutationally inactivated for miR-BHRF1-3 expression, antibody: Anti-Ago2 (clone 9E8)	1
17-20	PAR-CLIP	human	HEK-293	3 samples stable expressing Flag/HA–AGO1- antibody: FLAG, 1 sample with antibody: AGO2 11A9	4
21-24	PAR-CLIP	human	HEK-293	immunoprecipitated protein: AGO1,AGO2, AGO3, AGO4 respectively	4

Table 11: Summary of the collected PAR-CLIP experiments in human species, obtained from 8 studies. These datasets provided the source of PAR-CLIP signal (raw reads and transitions) which was combined with experimentally validated positive/negative instances of miRNA-targeted regions.

2.7.1.7 Background CLIP deep sequencing datasets

PAR-CLIP sequencing experiments of HEK-293 cells, stable expressing a non-RBP control (FLAG-GFP) and treated with FLAG-tagged antibody, enabled the detection of non-specific protein-bindings. These recently published experiments can be exploited to decipher background CLIP signal (157).

Consequently, they were accordingly processed to produce negative PAR-CLIP regions for miRNA binding. The identification of negative MREs within the background clusters was performed for miRNAs expressed in the HEK-293 cell line.

2.7.1.8 Random CLIP-Seq

An *in silico* pipeline was implemented to simulate PAR-CLIP libraries. The randomly produced CLIP-Seq data, at the level of raw reads, provided an extra source for negative clusters and MRE regions. MRE control instances derived from the simulated PAR-CLIP clusters were generated for all the miRNAs encountered on positive or negative interactions and were supported by (in)direct, low/high-throughput experiments.

2.7.2 Compilation of positive and negative training sets

The unified set of positive MRE-instances was compiled from chimeric miRNA-target fragments, direct miRNA bindings supported by Reporter Gene Assays as well as miRNA-target interactions derived from RNA sequencing experiments, quantitative proteomics and ribosome profiling. Positive MREs exceeded 11 thousand and are mainly placed in coding and 3′ untranslated regions of the mRNAs. miRNA-targeted regions presented an overlap with clusters from at least one AGO-PAR-CLIP sequencing library. The respective negative set, defined by different indirect high-throughput experiments, background PAR-CLIP libraries as well as by randomly generated CLIP datasets, was appropriately filtered to avoid any conflict with positive instances (both at interaction and at miRNA binding site level). Notably, specific attention was paid to create positive and negative sets with similar ratios in terms of MRE biotype annotation (Table 12).

This comprehensive collection of miRNA interactions enabled the development of a novel AGO-CLIP-Seq-guided Algorithm intended for miRNA-target identification.

Positive Instances	miRNAs in interactions	targeted regions	miRNA-target pairs
Chimeric	238	5,201	5,885
Reporter	94	168	179
RNA-Seq	2	309	309
Microarrays	52	4,553	4,663
pSILAC	5	123	123
RPF	2	507	507
Negative Instances	**miRNAs in interactions**	**targeted regions**	**miRNA-target pairs**
RNA-Seq	2	274	274
Microarrays	51	3,106	3,118
pSILAC	5	36	36
RPF	2	497	497
Background CLIP-Seq	360	9,946	10,031
Simulated CLIP-Seq	200	8407	8523

Table 12: Overview of miRNA-target positive/negative instances as identified by different indirect/direct low and high-throughput experiments as well as by randomly simulated CLIP datasets. miRNA-targeted regions presented an overlap with clusters from at least one PAR-CLIP sequencing library. No overlap was allowed between positive and negative miRNA-gene interactions and their related MRE-instances.

2.7.3 Feature set description

A set of approximately 300 descriptors was created for the comprehensive compendium of positive and negative instances. The extracted features comprised coverage measurements derived from the CLIP-Seq signal; substitution ratios and distance of substitutions from the MRE start; base and dinucleotide contents for the miRNA site as well as its respective flanking regions; location of the MRE within the cluster; complexity features for the MRE and proximal upstream/downstream sequences; energy-related variables for the duplex structure; paired positions and nucleotides of the miRNA-target hybrid; (mis)matches, bulges, loops and wobble pairs for miRNA and MRE sub-domains that participate in the duplex formation (seed, after-seed, 3′ compensatory and tail region); binding type; conservation scores for the MRE and upflank/downflank-MRE regions. There are also features describing binding length ratios of miRNA and/or target regions, as well as metrics for sequence content skewness/asymmetry and biases of codon usage. Major categories are described in more detail in the following paragraphs.

Expression Features. The first category of features corresponds to coverage measurements derived from the analyzed PAR-CLIP experimental datasets. The descriptors were designed for AGO enriched regions (clusters) as well as the relevant miRNA targeted regions (MREs). Cluster and MRE RPKM measurements correspond to the normalized read coverage of the peak and the miRNA binding region, respectively. Aligned reads residing within the cluster or the MRE are normalized for CLIP sequencing depth and relevant region length. In addition to the RPKM values, the raw number of overlapping reads is included as an extra feature. MRE coverage relative to cluster coverage is another informative feature especially useful for binding sites located on broader peaks or near the cluster's 3′ or 5′ end.

Substitution Features. Another category of features was created to describe substitution ratios based on CLIP-Seq aligned reads. In PAR-CLIP experiments it is expected to observe T-to-C conversion sites in the AGO-miRNA crosslinked regions. Other transitions may also be detected in the vicinity of a binding site (MRE). These non T-to-C events may constitute false positive sources of conversions due to sequencing artifacts or cell type-specific variations (165). However, it is possible that they correspond to other crosslinking-induced mutation sites, generated during the reverse transcription. Therefore, information of every putative substitution ratio upstream/downstream the MRE start and different mutation positions was included in the developed model. Additional features describing substitution distances from relative MRE start were also added. Substitution ratios and distances for each or all transition types were combined to extract other meta-descriptors.

Sequence Complexity and Energy Features. A set of thermodynamic properties including entropy (dS), enthalpy (dH), free energy (dG) and melting temperature (Tm) were estimated for the MRE sequences. Additional sequence measurements were incorporated in the model such as BLAST's DUST score for masking low complexity sequences (166), MRE complexity calculated with the Shannon-Wiener Index (167), as well as quantitative metrics of nucleotide/base composition asymmetry (GC-skew, AT-skew, purine-skew, Ks-skew).

Conservation Features. Conservation is a feature that is deemed important in miRNA-target interactions and therefore it has been adopted from many *in-silico* prediction algorithms. In the specific model, phastCons pre-computed scores from genome-wide multiple alignments were utilized to deduce evolutionary rates of miRNA targeted regions as well as their flanking regions (Figure 12). Regions conservation signal were estimated as mean intensities of the overlapping phastCons base-wise scores. Moreover, separate descriptors were utilized to describe conservations of the most 5' MRE binding nucleotides and all binding nucleotides of the MRE in each miRNA-target duplex. PhastCons precompiled values were downloaded from the UCSC repository (152) in bigwig format.

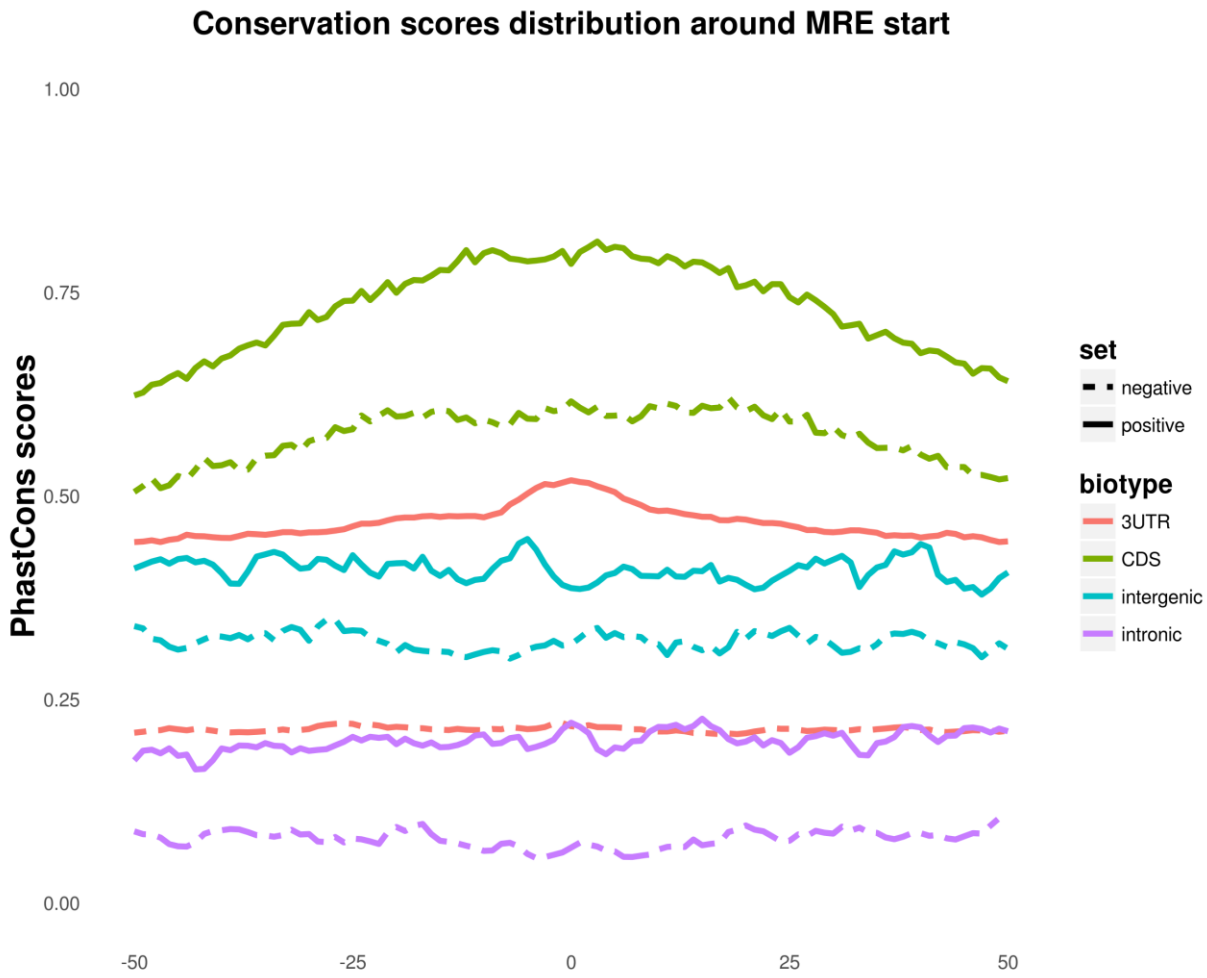

Figure 12: Pre-calculated phastCons base-wise conservation scores (mean values) overlapping positive/negative MRE start sites along with upstream/downstream flanking regions (± 50 nts). Positive/negative MRE conservation scores are spatially classified to 3'UTR, CDS, intergenic and intronic transcript regions. Distribution of conservation base scores are centered in the MRE start sites (position 0). Notably, positive MREs residing on CDS and 3'UTR regions present a significant increase of conservation scores around the MRE-start. *(Copyright Paraskevopoulou Maria)*

Content Features of MRE and flanking regions. Single/di-nucleotide composition descriptors were generated for the miRNA binding site and the upstream or downstream MRE regions.

miRNA-target duplex Features. The duplex structure energy of putative miRNA-target pairs was estimated using the RNAduplex algorithm of the Vienna package (168). Different features have been established to describe loops, miRNA or MRE bulges and mismatches, GU wobbles and AU base pairing features. Several publications discuss the varying impact of mismatches, internal loop formations, miRNA or target bulges in conjunction with their position within the duplex structure (169-171). Moreover, miRNA sequence can be divided into distinct domains with different levels of importance after the 5' anchor (nt 1): (i)seed region (2-8 positions), (ii) central region (9-

12 positions), (iii) 3' supplementary/compensatory region (13-16 positions), (iv) tail region (17-3'miRNA end) (Figure 13). Following a miRNA-target duplex construction, relevant domains can be defined in the MRE region based on the binding anchors of miRNA sub-regions. To this end, the aforementioned descriptors were designed for each miRNA and/or target sub-domains as well as for the entire duplex structure. Finally, miRNA binding and MRE binding length were incorporated in the feature set.

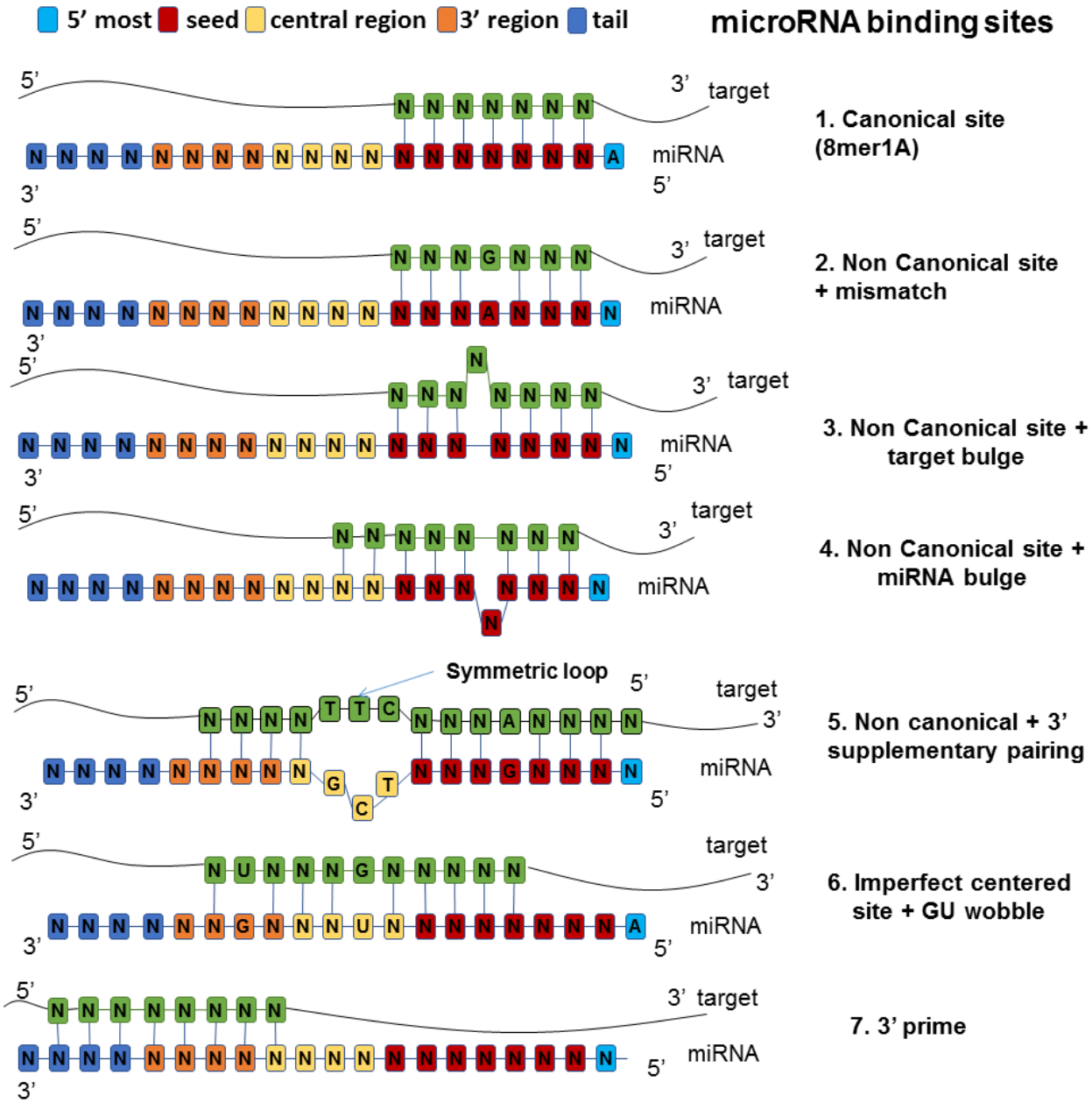

Figure 13: Snapshot of the different binding types identified by the novel Algorithm for CLIP-guided miRNA-target identification. *(Copyright Paraskevopoulou Maria)*

Matches per miRNA/MRE region. Binary binding vectors of miRNA/MRE position base pairing were added to the model, where each element in the vectors constitutes a distinct descriptor. Moreover, extra features were created to describe the total and consecutive matches in the miRNA-target structure as well as in MRE and miRNA relevant domains (seed, central, 3′ supplementary region, tail). Base composition descriptors (A, T, G, C) of the (un)paired nucleotides were also included.

Binding Type descriptor. Accumulating evidence from low-yield and sequencing experiments revealed the high abundance of non-canonical miRNA binding sites. For instance, the analysis of CLASH-Seq experimental data has shown that a significant portion of the identified miRNA-target chimeras correspond to non-canonical base pairings. Moreover, another high-throughput experiment enabled the detection of centered miRNA binding events (5-15 position) that may be potent sites for target repression. Other non-canonical sites with nucleation target bulges in the seed region are considered also effective to mediate mRNA repression (40,171-173). To this end, the adopted binding categories in the TarBase/LncBase CLIP-Seq algorithm were revisited in order to cover the whole spectrum of the putative miRNA-target base pairings. The extended binding codes incorporated in the novel CLIP-Seq learning framework are described in Table 13.

New Binding Codes	Description
9mer.3prime	9mer canonical site (matches in 1-9 positions of the miRNA) with additional compensatory 3′ binding
9mer	9mer canonical site (matches in 1-9 positions of the miRNA)
9mer.GU	miRNA base pairing in 1-9 positions with a GU wobble pair
9mer.nonCanonical	miRNA non canonical base pairing in 1-9 positions, with a target bulge and/or a GU wobble pair
8mer.3prime	8mer canonical site (matches in 1-8 or 2-9 positions of the miRNA) or 8mer1A with additional compensatory 3′ binding
8mer	8mer canonical site (matches in 1-8 or 2-9 positions of the miRNA)
8mer1A	7mer canonical site (matches in 2-8 positions of the miRNA) with additional A in position 1 (match or mismatch)
8mer.GU	miRNA base pairing in 1-8 or 2-9 positions with a GU wobble pair
8mer.nonCanonical	miRNA non canonical base pairing in 1-9 positions with mismatch or mirna bulge and/or a target bulge and/or a GU wobble pair
7mer.3prime	7mer canonical site (matches in 2-8 positions of the miRNA) or 7mer1A, with additional compensatory 3′ binding

7mer	7mer canonical site (matches in 2-8 positions of the miRNA)
7mer1A	6mer canonical site (matches in 2-7 positions of the miRNA) with an additional A in position 1 (match or mismatch)
7mer.GU	miRNA base pairing in 2-8 positions with a GU wobble pair
7mer.nonCanonical	miRNA non canonical base pairing in 1-8 positions with a mismatch or miRNA bulge and/or a target bulge
7mer.nonCanonical.GU	miRNA non canonical base pairing in 1-8 positions with a mismatch or miRNA bulge and/or a target bulge and/or a GU wobble pair
6mer.3prime	6mer canonical site (matches in 2-7 positions of the miRNA) with additional compensatory 3′ binding
6mer	6mer canonical site (matches in 2-7 positions of the miRNA)
offset6mer	6mer canonical site (matches in 3-8 positions of the miRNA)
6mer.nonCanonical.3prime	miRNA non canonical base pairing in 2-8 positions with a mismatch or miRNA bulge and/or a target bulge, with additional compensatory 3′ binding
6mer.nonCanonical	miRNA non canonical base pairing in 2-8 positions with a mismatch or miRNA bulge and/or a target bulge
5mer	5mer canonical site (matches in 2-6 or 3-7 positions of the miRNA) with additional compensatory 3′ binding
5mer.nonCanonical	miRNA non canonical base pairing in 2-8 positions with a mismatch and/or a target bulge and/or miRNA bulge, with additional compensatory 3′ binding
seedless	miRNA non canonical base pairing after position 4 with at least 7 matches after the seed region
seedless.3prime	miRNA non canonical base pairing after position 4 with at least 7 matches after the seed region and additional compensatory 3′ binding
centered	miRNA base pairing with at least 8 consecutive matches in 4-15 positions
imperfect.centered	miRNA base pairing with at least 8 matches in 4-15 positions and/or less than 2 GU wobble pairs.
3prime	miRNA base pairing after the position 13 with at least 7 matches

Table 13: Detailed description of the updated miRNA binding type categories that can be recognized by the novel Algorithm developed for the analysis of AGO CLIP-Seq data.

2.7.4 Feature Preprocessing and Assessment

The identification of informative descriptors from a primary large feature set collection constitutes a hard and demanding task. Since there is no silver bullet for the selection of the most prominent feature subsets, a hybrid approach was adopted comprising different techniques for dimensionality reduction. More precisely, automated methodologies (such as information gain measures and minimum-redundancy-maximum-relevance technique), hierarchical parameter selection and heuristics, including distance Kullback-Leibler, Wilcoxon's exact test, ROC AUC were implemented. Notably, the different sub-groups of negative and positive datasets were compared and evaluated to identify (dis)similar patterns in their respective feature distributions. For non-parametric multiple group comparisons, Kruskal-Wallis test along with Mann-Whintey's U test (as a non-parametric post-hoc test) and the Benjamini-Hochberg's False Discovery Rate (FDR) correction (in order to control family-wise type I error rate and to identify significant differences between groups) were utilized.

Moreover, specific attention was paid to eliminate highly correlated descriptors as well as features presenting close to zero variance. Correlations were assessed using the non-parametric Spearman's rho coefficient. All tests were two-sided. Differences were considered as statistically significant if the null hypothesis could be rejected with >95% confidence ($p<0.05$).

This combinatorial process of feature evaluation enabled the ranking of every parameter individually based on its predictive accuracy and additionally facilitated the identification of possible associations between the input variables.

2.7.5 Novel algorithm Learning Framework for CLIP-Seq analysis

The adopted pipeline revealed different candidate feature vectors that were assessed for their predictive performance on independent test sets with several machine learning models including SVM, Naïve Bayes, Random forest, Adaboost and Gradient Boosting.

Several feature subsets attained higher predictive accuracy in distinguishing the true AGO bound regions (cluster). On the other hand, others were proven of greater efficacy for predicting the correct miRNA binding sites.

Moreover, many of the cluster/region related descriptors had a strong impact and were favored by several learning frameworks compared to binding and MRE-derived features, in case of co-occurrence. These models usually resulted in a high number of predicted MREs per peak, presenting a weak ability to recognize the true miRNA binding sites. Thus region-related features were included in a separate base classifier and were subsequently combined with binding features in a meta-classifier. After the evaluation of different feature sub-vectors with different classifiers we concluded in the following learning model.

The proposed implementation comprises 6 distinct base Random Forest classifiers (Figure 14):

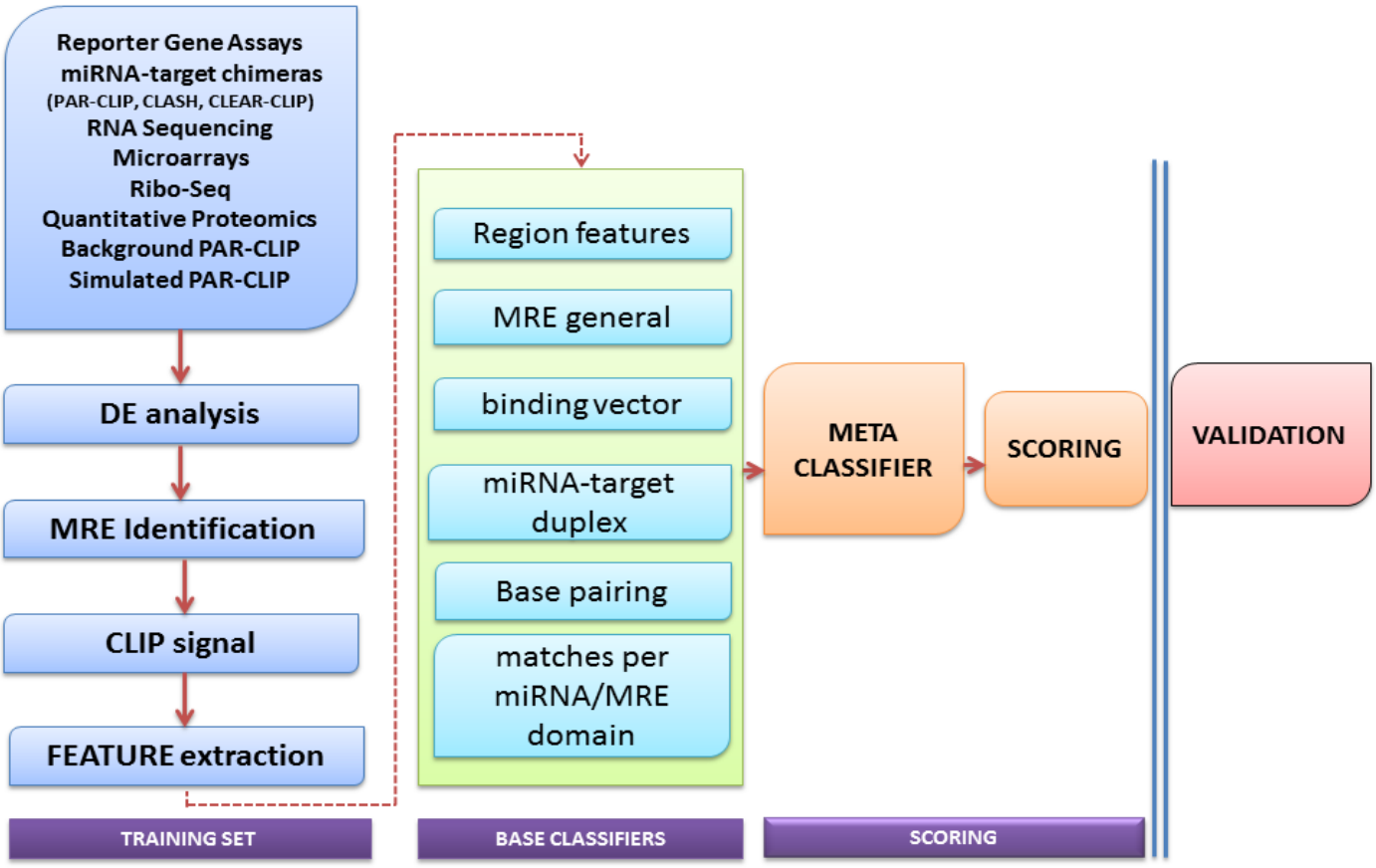

Figure 14: Overview of the adopted pipeline for the development of a novel learning framework for CLIP-guided miRNA-target identification. *(Copyright Paraskevopoulou MD)*

1) **Region features:** CLIP-sequencing-derived features, such as RPKM coverage, substitution frequencies and distances from the MRE start as well as overlapping/upstream/downstream MRE region content, conservation, sequence energy, complexity, content asymmetry, and biases of codon usage.

2) **MRE general:** MRE-related descriptors including the degree of overlap with the respective cluster, conservation of the most 5' MRE binding nucleotides and all MRE binding nucleotides, MRE location within the cluster, MRE binding type well as metrics for duplex matched nucleotide content skewness.

3) **Binding Vector:** Binary binding vectors of miRNA/MRE position base pairing were added to the model, where each element in the vectors constitutes a distinct descriptor.

4) **miRNA-target duplex:** miRNA-target duplex structure energy, miRNA or MRE bulges and mismatches, GU wobbles and AU base pairing features for the specified miRNA and/or target and relevant sub-domains.

5) **Base pairing:** base composition descriptors (A, T, G, C) of the (un)paired nucleotides were also included.

6) **Matches per miRNA/MRE domain:** total and consecutive matches in the miRNA-target structure as well as in MRE and miRNA relevant sub-domains.

A boosting meta-classifier was implemented to assemble the generated output of the base classifiers. The predictive accuracy of the final model as well as its evaluation against other state-of-the-art implementations is presented in the relevant result section.

3. Results

3.1 DIANA-microT web server v5

The updated microT web server incorporates miRBase version 18 (174) and Ensembl version 69 (175) nomenclature. The *in silico*-predicted miRNA-gene interactions in *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans* exceed 11 million in total.

3.1.1 Web Server Update and Extension

The selection of DIANA-microT-CDS as its core algorithm renders the new web server the only available online resource capable of incorporating miRNA targets in 3'UTR as well as in CDS regions. The new web server enables users to attain high quality predicted miRNA-gene interactions in all relevant *in silico* pipelines.

The server is compatible with the new miRNA nomenclature (3p/5p) introduced in miRBase v18, as well as with previous miRNA naming conventions. It currently supports $7.3 \cdot 10^6$ *H. sapiens*, $3.5 \cdot 10^6$ *M. musculus*, $4.4 \cdot 10^5$ *D. melanogaster* and $2.5 \cdot 10^5$ *C. elegans* interactions between 3,876 miRNAs and 64,750 protein coding genes. Gene (175) and miRNA (176) expression data have been incorporated into the web server, enabling the user to perform advanced result filtering based on tissue expression. Furthermore, users can also restrict predictions between uploaded lists of expressed genes and/or miRNAs. For example, this feature can be used to identify interactions between a list of repressed (or overexpressed) genes and overexpressed (or repressed) miRNAs, in the case of a differential expression analysis pipeline.

Moreover, the web server hosts an updated version of the KEGG database providing a relevant search module based on KEGG pathway descriptions (177). A redesigned optional user space has also been implemented, which provides personalized features and facilitates the interconnection between the web server and the available DIANA software and databases (Figure 15).

3.1.2 DIANA-microT web server v5 Interface

The DIANA-microT web server provides *in silico* predictions of miRNA:mRNA interactions in a user-friendly interface. Specific attention has been paid to the web server interface, which follows the DIANA design framework, in order to be instantly familiar to users of previous versions or other DIANA tools. On the other hand, online help, informative tooltips and easy-to-use menus, minimize the learning curve of new users. A snapshot of the DIANA-microT web server interface is provided in Figure 15.

The interface hosts extensive information for predicted miRNA:target gene interactions such as, a global score for each interaction, as well as detailed information for all predicted target sites. Each target site can be individually visualized and the user can

examine its local prediction score, target site conservation and the miRNA-mRNA binding structure. The server provides also connectivity to online biological databases and offers links to nomenclature, sequence and protein databases.

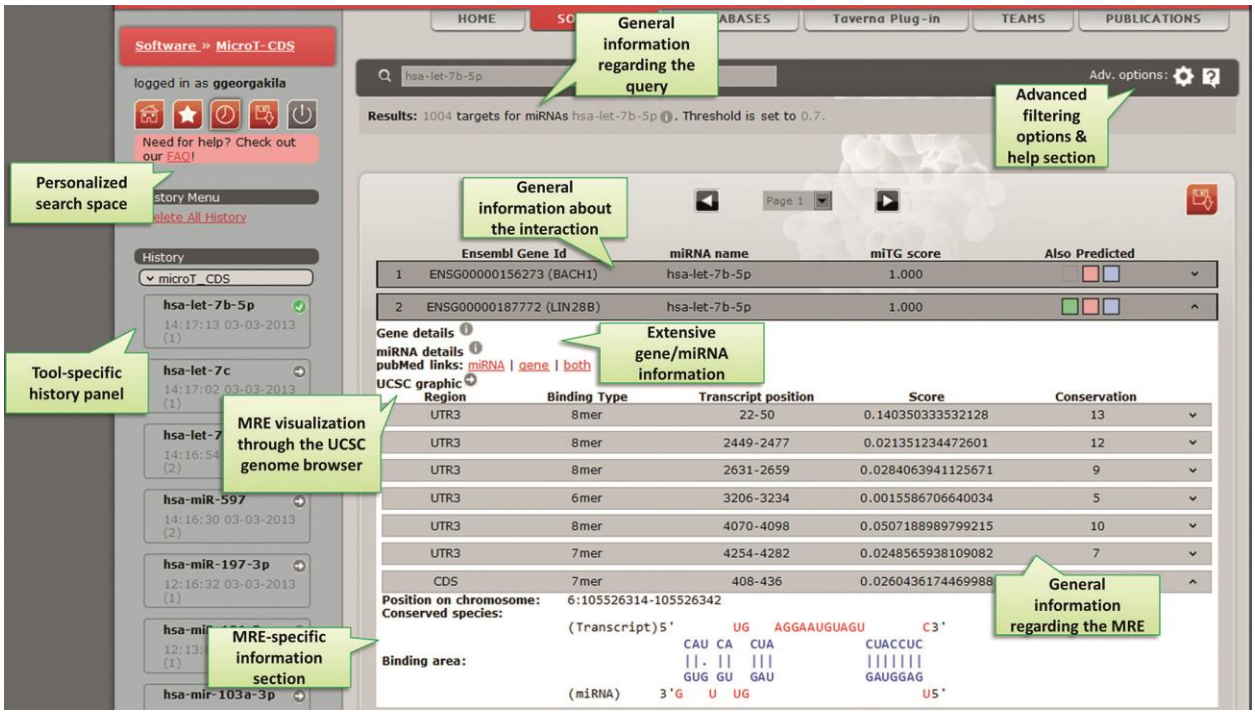

Figure 15: Example of a submitted query in the DIANA-microT web server v5.0. The interface presents information regarding each predicted miRNA:mRNA interactions. miRNA and gene-related information, as well as advanced search options have been expanded. Links to external databases, graphical representation of the binding sites as well as miRNA recognition element (MRE) conservation and prediction scores are displayed in the relevant sections. The left side of the page is devoted to the personal user space, reporting latest searches and bookmarks (Paraskevopoulou MD *et al, 2013*)(54).

3.1.3 Advanced pipelines supported by the microT-web server v5

DIANA-microT web server v5.0 hosts integrated analyses in the form of ready-made advanced pipelines, covering a wide range of inquiries regarding predicted or validated miRNA-gene interactions and their impact on metabolic and signaling pathways. These pipelines can be utilized to analyze user data derived from small scale or high throughput experiments directly from the DIANA-microT web server interface, without the necessity to install or implement any kind of software.

The supported advanced workflows can perform extensive miRNA-related analyses on results derived from high throughput techniques, such as microarrays or NGS. More precisely, workflows can analyze mRNA and miRNA expression data (expression and fold change) with suppressed genes automatically matched with overexpressed miRNAs and *vice versa*.

Supported workflows can perform enrichment analyses of experimentally validated targets derived from DIANA-TarBase v6.0 (7) and/or predicted interactions from microT-CDS. This enrichment analysis methodology is considered crucial in order to identify miRNAs that regulate the differentially expressed genes.

The prediction score threshold can significantly affect the analysis steps that follow. One available pipeline performs miRNA target prediction for the differentially expressed genes using different microT score thresholds and meta-analysis statistics, followed by pathway enrichment analysis. This pipeline is optimized by automatic repetitions of different prediction thresholds (from sensitive to more stringent), in order to minimize the effect of the selected settings to the derived results. By utilizing meta-analysis statistics, the server combines the p-values from each repetition into a total p-value for each miRNA, signifying its effect on the selected genes for all utilized thresholds (178,179). In the last step of the pipeline, the identified miRNAs are subjected to a functional analysis, where pathways controlled by the combined action of these miRNAs are detected using DIANA-miRPath v2.1 (179).

Other supported pipelines can handle miRNA and gene lists, in order to perform the enrichment analysis or even select the type of utilized interactions (predicted or experimentally validated). In the latter, the algorithm "personalizes" the target identification module for each miRNA. It initially identifies the number of available interactions in DIANA-TarBase and DIANA-microT-CDS (validated *vs* predicted) and automatically selects to use validated targets only for well-annotated miRNAs. Computationally identified interactions are used otherwise.

The new DIANA-microT web server enables users to perform such analyses directly from the on-line user interface, and/or create even more extensive pipelines programmatically or by using visual tools (Taverna WMS).

Furthermore, the web server also supports direct programmatic access to all aforementioned utilities in the form of services, in order to facilitate users having already implemented pipelines with scripting or programming languages.

3.1.3.1 Example workflows

The implemented workflows make use of DIANA-Lab services through the DIANA-web-server plug in. In order for this workflow to work properly the plug-in has to be installed in the compatible Taverna versions. The workflow can run automatically, as soon as the necessary input values are provided. Examples of implemented pipelines are presented below.

1. Enrichment analysis of *in silico* predicted miRNA-gene interactions followed by a targeted pathway analysis.

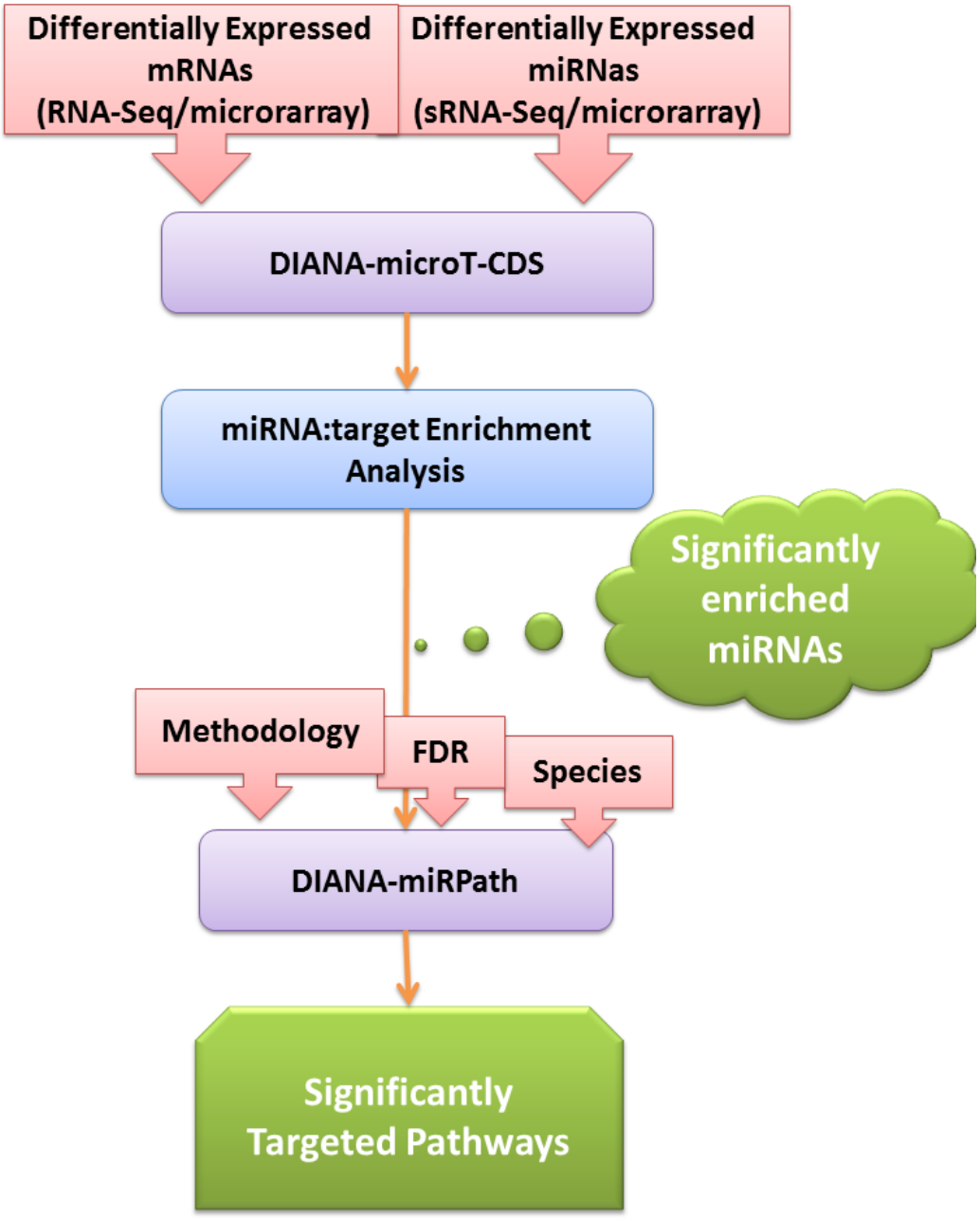

Figure 16: The implemented workflow initially performs enrichment analysis of *in-silico* predicted targets derived from DIANA-microT-CDS and identifies miRNAs significantly controlling the set(s) of differentially expressed genes. Subsequently, a miRNA-targeted pathway analysis is implemented with DIANA-miRPath v2.

2. Optimized enrichment analysis of predicted miRNA-gene interactions followed by targeted Pathway analysis.

The pipeline is automatically repeated for different prediction thresholds (from sensitive to more stringent), in order to minimize the effect of the selected settings to the derived result.

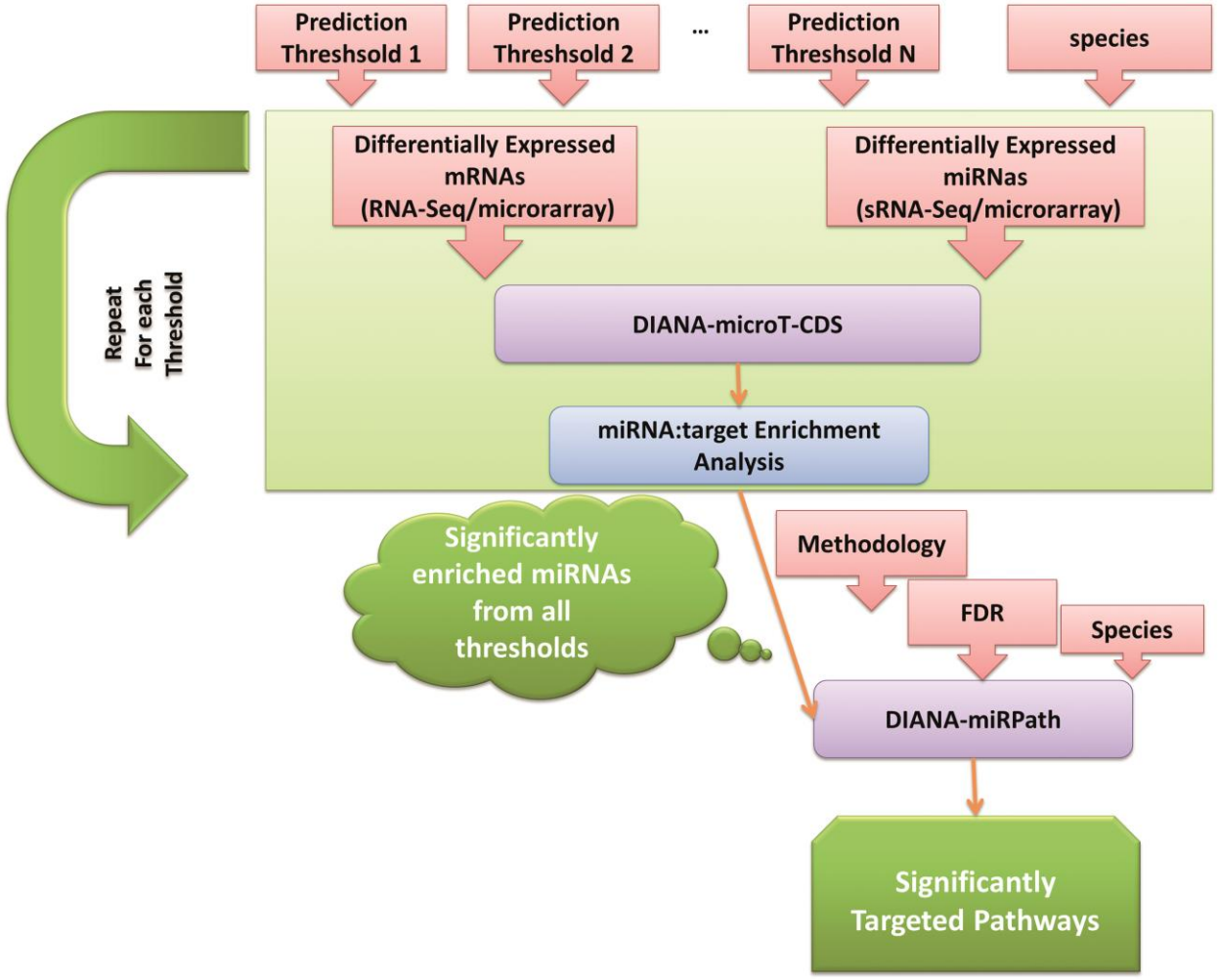

Figure 17: Flowchart depicting an analysis pipeline directly available from the web server interface. Interactions between user-defined miRNA and gene sets are *in silico* identified in 3'UTR and CDS regions using DIANA-microT-CDS. A subsequent miRNA target enrichment analysis identifies miRNAs controlling significantly the sets of differentially expressed genes. The pipeline is automatically repeated for different prediction thresholds (from more sensitive, to more stringent). By utilizing meta-analysis statistics, the server combines the p-values from each repetition into a total p-value for each miRNA, signifying its effect on the selected genes for all utilized thresholds. In the last step of the pipeline, miRNA-targeted pathway analysis is implemented with DIANA-miRPath v2. Paraskevopoulou MD *et al, 2013*) (54)

3. "Personalizing" the selection of miRNA-specific validated/predicted interactions, followed by miRNA Pathway analysis.

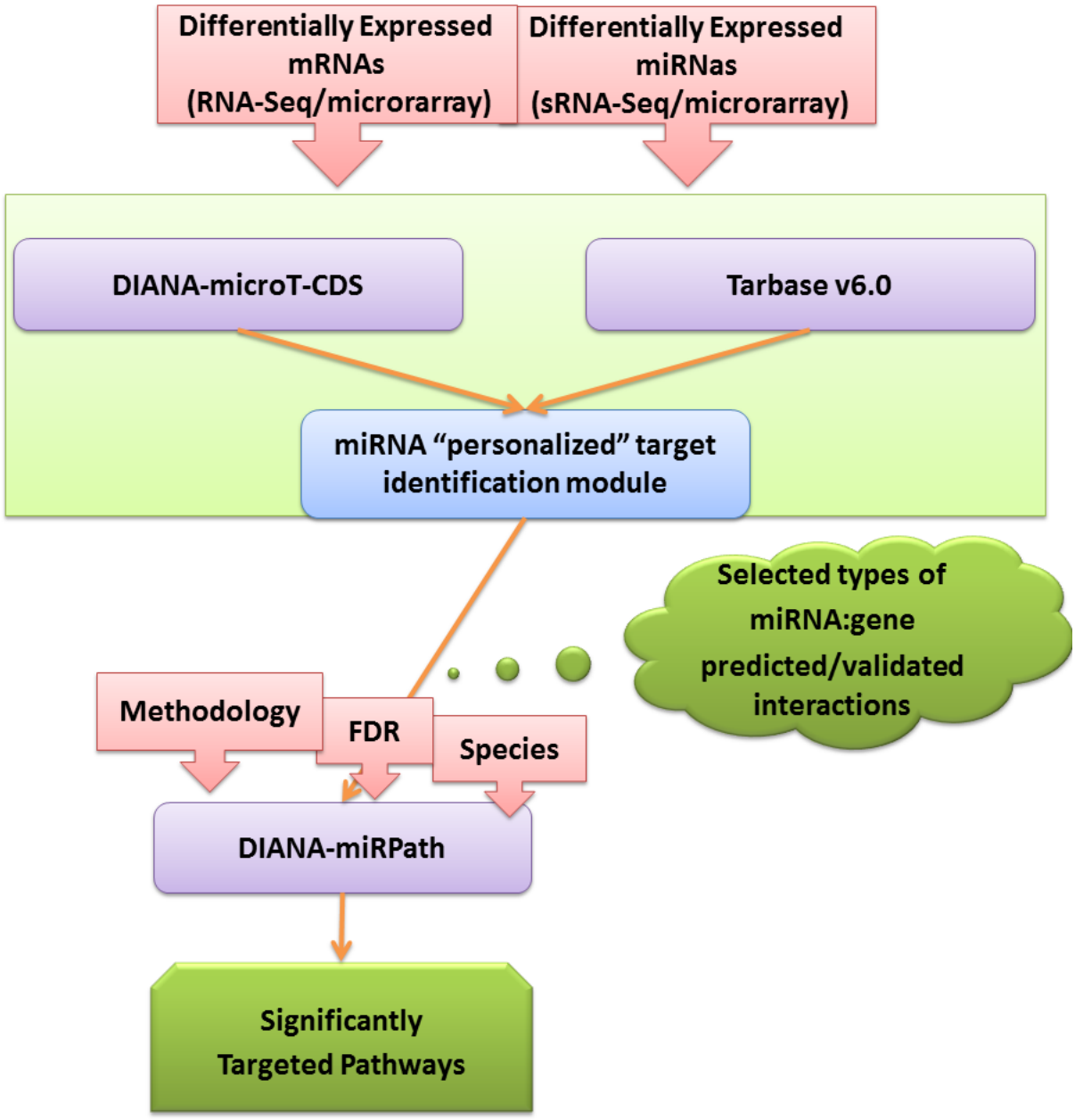

Figure 18: In this workflow, the algorithm "personalizes" the target identification module for each miRNA. It initially identifies the number of available interactions in DIANA-TarBase and DIANA-microT-CDS (validated vs predicted) and automatically selects to use validated targets only in the cases of well-annotated miRNAs. Otherwise, computationally identified interactions are used for the analysis. In the final step of the pipeline the selected miRNAs are subjected to a functional analysis with DIANA-miRPath v2, where pathways controlled by the combined action of these miRNAs are detected. The pipeline selects to use targets predicted with DIANA-microT-CDS or experimentally verified targets from TarBase v6 based on the analysis performed in the previous step.

3.2 DIANA-Taverna plugin

DIANA-Taverna-Plugin enables the user to directly access our target prediction server (microT-CDS) from the graphic interface of Taverna and incorporate advanced miRNA analysis functionalities into custom pipelines. Furthermore, the plug-in enables the extension of such pipelines through the use of other DIANA tools and databases, providing access to an extensive collection of validated miRNA targets and to DIANA-miRPath v2.1, a tool designed for the identification of miRNA targeted pathways.

The DIANA-Taverna Plugin provides optimized use of the DIANA-web server and databases. It can be installed in compatible Taverna versions (v2.3 and v2.5) through the "add plugin site" functionality of the Taverna Workbench.

Following the plugin installation, DIANA services are added under the local Taverna "Available Services" panel section (Figure 19) along with the other provided tools. The DIANA services can be incorporated to develop multistep analysis workflows by 'drag and drop' of each service to the workflow design window.

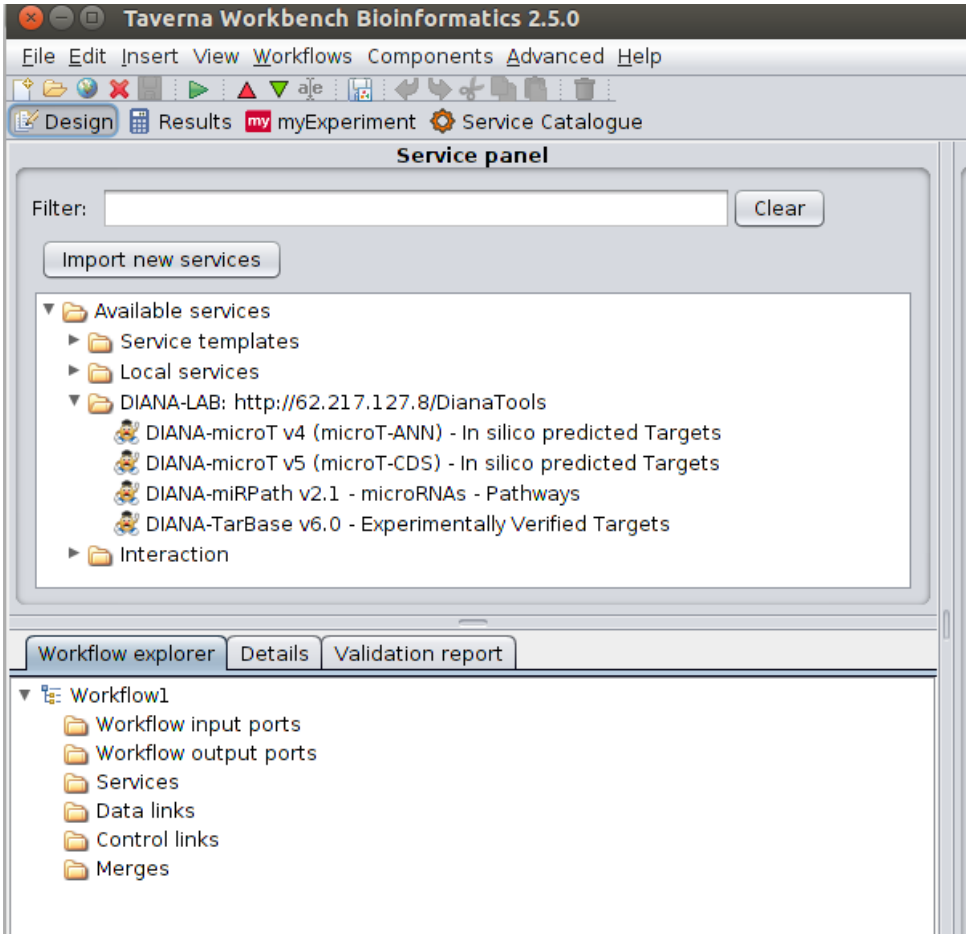

Figure 19: DIANA-Taverna plugin is installed in Taverna WMS. The DIANA services are added under the local Taverna "Available Services" panel section along with the other provided tools.

3.3 DIANA-TarBase repository

DIANA-TarBase currently indexes more than half a million entries, 9-250 times more than any other relevant database. All entries are accompanied by rich detailed meta-data that can also be used as search and filtering terms from the new application-like user interface. For instance, DIANA-TarBase v7.0 collects data regarding the experimental conditions, such as the exposition of cells to stressors, drugs or other agents, since these can alter miRNA regulatory networks. Another novel aspect of the database is its ability to include detailed information regarding the experimental methodologies utilized for the identification of each interaction, since experimental techniques cannot be considered as having equal information content.

The number of targets derived from major method classes is depicted in Figure 20.

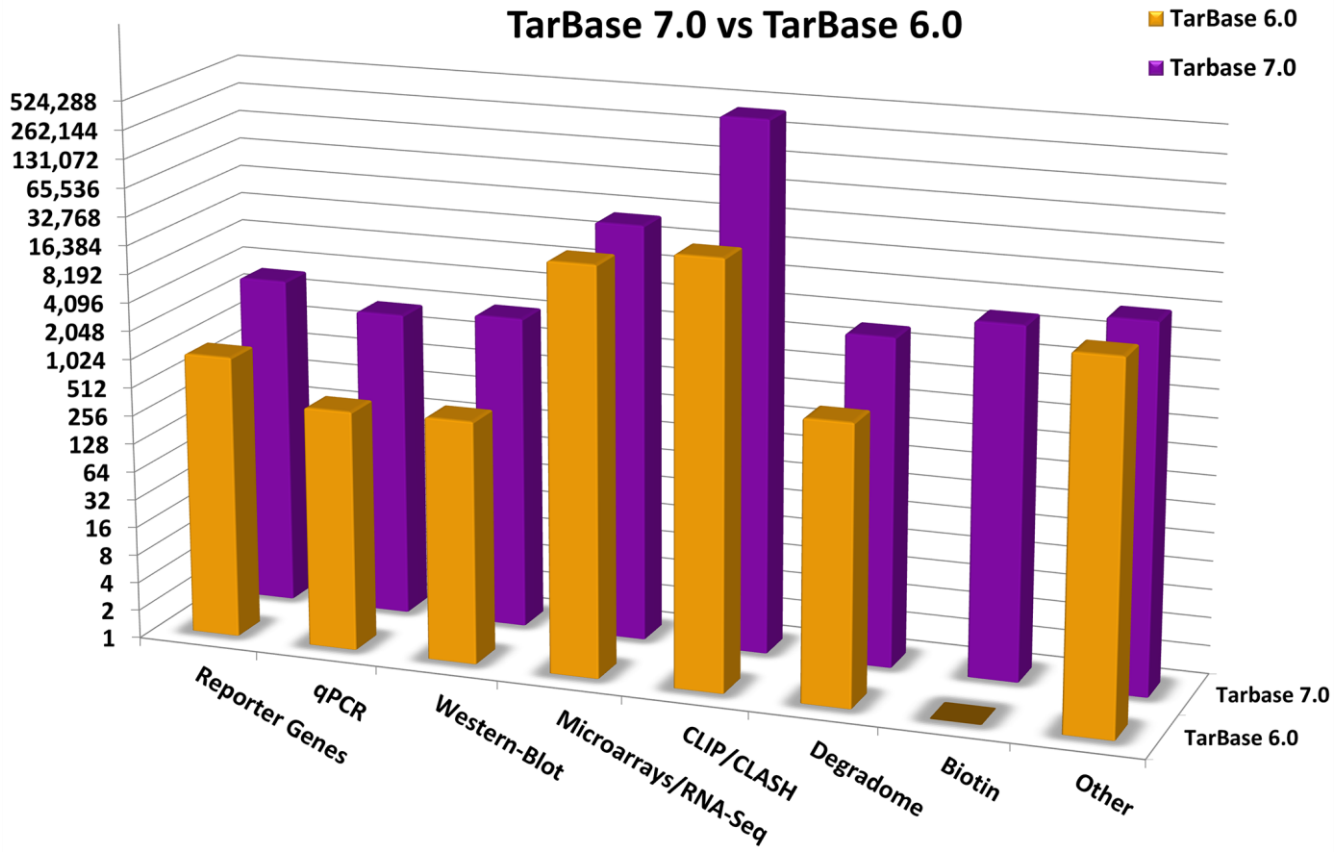

Figure 20: Entries per methodology for TarBase v7.0 and TarBase v6.0. The y-axis (number of entries) is in log2 scale and each mark signifies doubling of available entries. (Vlachos IS and Paraskevopoulou MD *et al*, 2014) (64)

3.3.1 Database Statistics

The database comprises more than half a million interactions spanning 24 species; a 9 to 250-fold increase compared to TarBase 6.0 and other manually curated databases, including miRTarBase and miRecords. The database encompasses interactions derived from the widest variety of experiments to date, which are performed utilizing 28 different experimental techniques, 356 cell types and 59 different tissues.

Importantly, we have paid significant attention to curate articles utilizing highly specific low yield techniques such as reporter genes, as well as state-of-the-art methodologies, such as CLIP-Seq and CLASH experiments. The updated database contains more than 7,500 interactions derived from specific techniques (4-fold increase vs TarBase v6.0) and more than 500,000 interactions derived from high throughput experiments (8-fold increase vs TarBase v6.0). Specifically, DIANA-TarBase v7.0 incorporates data derived from 154 CLIP-Seq/CLASH datasets, as well as more than a hundred other high throughput datasets including Degradome-Seq (60), AGO-IP (32), biotin pull-down (32), miTRAP (63), 3'Life (62) and IMPACT-Seq (61), which is the highest number to be included in a manually curated database. The number of incorporated miRNA-related NGS datasets (e.g. CLIP-Seq, CLASH, Degradome-Seq) is also the highest ever reported.

3.3.2 DIANA-Tarbase Interface

DIANA-TarBase v7.0 is the first of DIANA databases and applications to utilize the new user interface, which is implemented using PHP (under Yii Framework), MySQL and JavaScript (JQuery).

The new DIANA-TarBase interface offers a friendlier, application-like user experience, minimizing the necessity to load/refresh web pages following user selections. The new interface brings the most common as well as advanced functions into the main pane, enabling users to perform simple or complex tasks, without leaving their results page.

Advanced Searching and Filtering

TarBase v7.0 supports advanced real-time search and filtering. All relevant options have been incorporated in the main result screen, in order to enable users to easily filter and query the database. The provided search and filtering options include: miRNA/gene combinations, species, experimental methodology class and subtype, type of regulation and validation, selection of positive or negative experimental results, year of publication and DIANA-microT-CDS threshold for interactions which are also predicted *in silico*. As in the previous version, DIANA-TarBase also integrates interactions from the latest available versions of external databases, including miRTarBase and miRecords. Users can easily filter results and include/exclude external sources or data derived from previous TarBase versions.

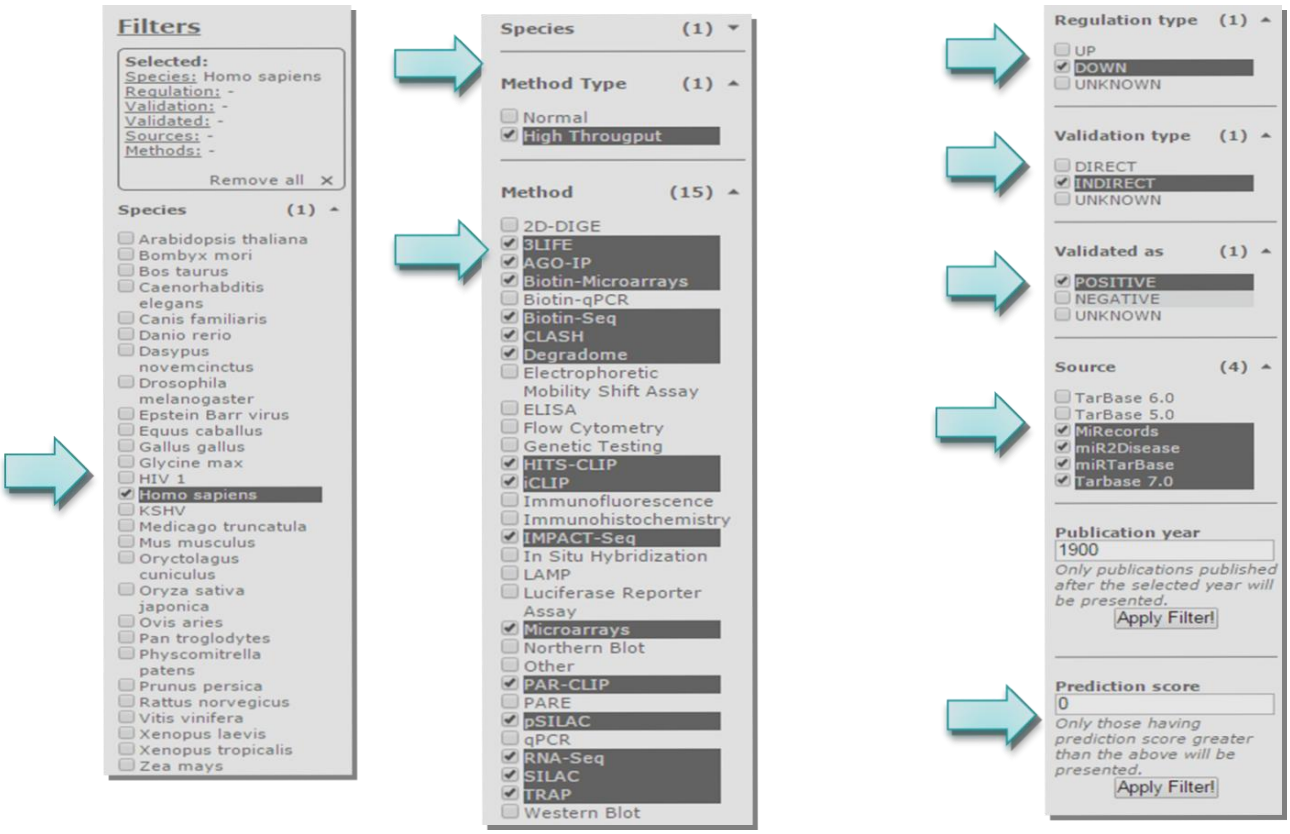

Figure 21: Advanced filtering options in Tarbase v7.

Querying the Database

The database query can be performed by entering any combination of miRNAs and/or gene names or supported identifiers (ENSEMBL (180) gene ids for genes and miRBase (181) MIMAT accessions for miRNAs). If genes and miRNAs are concurrently provided, TarBase will return all indexed interactions of the selected miRNAs with any of the provided genes.

The new interface (Figure 22) is designed around the new database schema, in order to cater to users extended meta-data regarding each interaction. Users can easily identify positive or negative experimental results, the utilized experimental methodology, experimental conditions including cell/tissue type and treatment. The new interface provides also advanced information ranging from the binding site location, as identified experimentally as well as *in silico*, to the primer sequences used for cloning experiments.

This version is also seamlessly incorporated to other DIANA-Tools. The DIANA-TarBase v7.0 user can easily perform a pathway analysis for the miRNA(s) under investigation, identify their predicted targets or examine if they have been identified experimentally or *in silico* to target long non-coding RNAs, using DIANA-miRPath v2.0(182), microT-CDS(183) and LncBase(184), respectively.

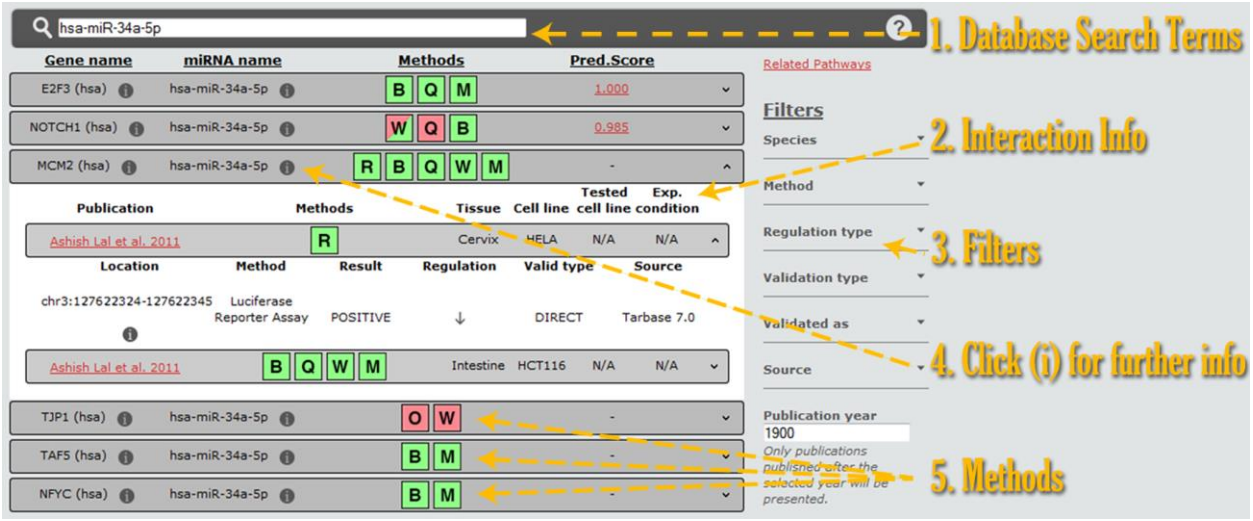

Figure 22: Screen-shot depicting the DIANA-TarBase v7.0 interface. Users can enter the query terms in the simple search box (1). Interaction information is presented below (2), while further details are accessible by expanding the result panel or by selecting the information links (4). All results are color coded, with green and red showing positive and negative experimental outcomes, respectively (5). Mixed results are presented using both colors. Users can filter the query results using any combination of the filtering options (3). (Vlachos IS and Paraskevopoulou MD *et al*, 2014) (64)

Since January 2014, DIANA-TarBase has been integrated in the official ENSEMBL (180) distribution. All TarBase entries having binding site coordinates can be explored directly from the ENSEMBL genome browser. Each DIANA-TarBase entry has a link pointing to the relevant browser view and coordinates, facilitating user interaction with both databases.

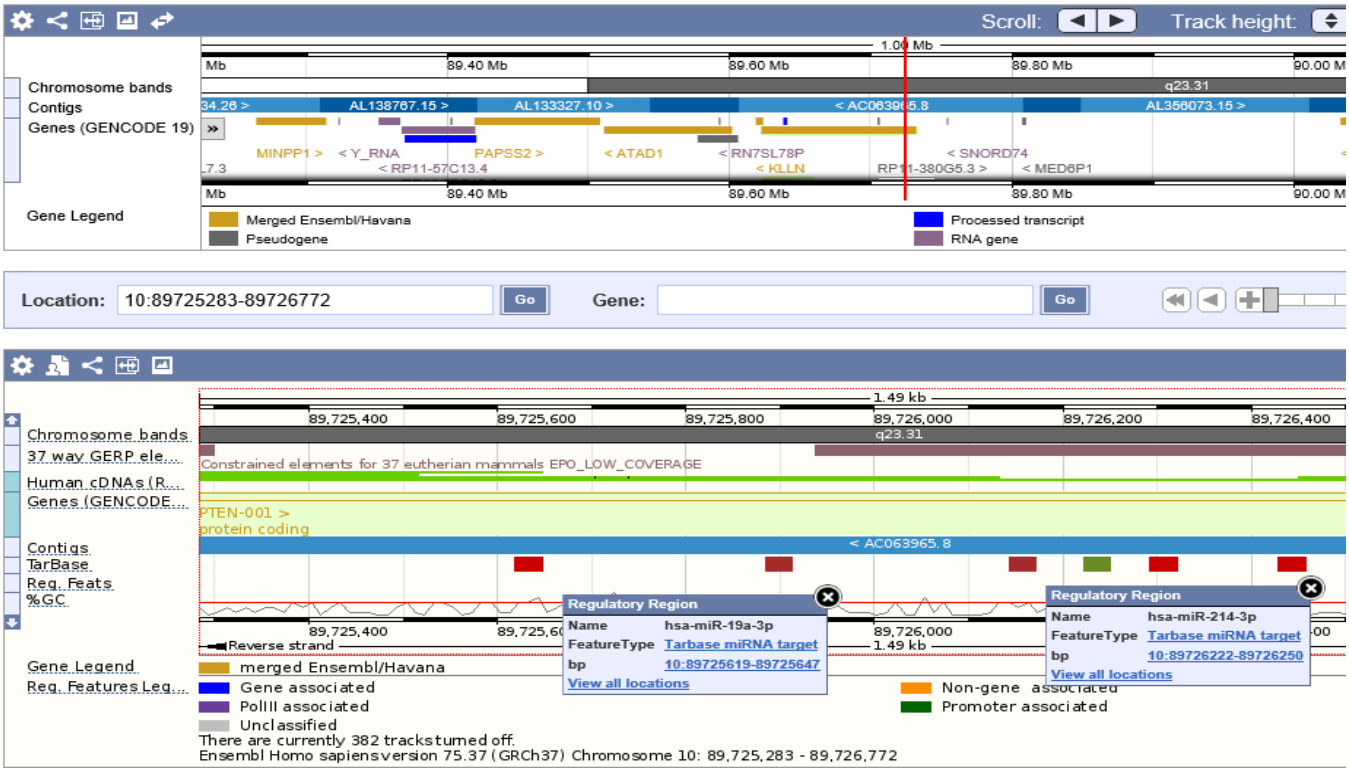

Figure 23: TarBase has been integrated in ENSEMBL since 2014, substituting the *in silico* miRNA predicted targets track.

3.4 DIANA-LncBase repository

LncBase v2 has been significantly extended, compared to the previous release (Table 14). LncBase v2 currently hosts ~70,000 experimentally supported interactions for an integrative meticulously curated collection of lncRNA transcripts. The new database enables the identification of miRNA-lncRNA regulatory interactions in numerous tissues, cell types and conditions, validated with low yield or high-throughput experimental methodologies (more than 150 raw NGS datasets). This compilation of high-throughput datasets corresponds to a 16-fold increase compared to the processed CLIP-Seq libraries available in LncBase v1.

LncBase v2 facilitates the charting of tissue and cell-type-specific miRNA-lncRNA interactions with state-of-the-art experimental techniques. Database entries are enriched with detailed metadata, including information on experimental methodologies, evolutionary conservation of miRNA-targeted regions and lncRNA transcript expression profiles, assessed by analyzing in-house 58 raw RNA-Seq libraries comprising ~6.1 billion reads.

The analysis of CLIP-Seq libraries resulted in a set of approximately 12,900 lncRNA transcripts harboring at least one MRE. More than half of the MREs identified on lncRNAs resided on intronic regions, which may be explained by the underestimation of their spliced length and number of exons. MREs detected on lncRNA introns are appropriately tagged and provided in the current release. LncBase v2 also hosts 14 PAR-CLIP libraries derived from virus infected cells. For these datasets, host lncRNA transcripts were additionally searched for interactions with viral miRNAs. Expressed viral miRNAs were found to participate in more than 400 miRNA-lncRNA unique interacting pairs.

Computationally predicted interactions, on the other hand, exceed 10 million between 41,229 lncRNAs and 4,503 miRNAs, for human and mouse. A subset of these interactions, approximately 5 million, represent a set of highly scored predictions composed of 22,073 lincRNAs, 12,485 antisense, 14,681 sense, 3,664 processed transcripts with at least one MRE.

A concise description of the updated database can be found in the following table.

		LncBase v1	LncBase v2
Database Entries	**miRNA species**	2	4
	miRNAs in interactions	127	~1,400
	unique interacting miRNA:lncRNA pairs	4,982	~51,000
	Cell lines	5	53
	Tissues	5	20
	Total interactions	4,994	>70,000
Analyzed High-Throughput Datasets	**Studies**	2	22
	Conditions	6	67
	Libraries	9	153
Experimental Methodologies	**Number of Methods**	4	12
	Description	CLIP-Seq, qPCR, Reporter Assay, Northern blot	CLIP-Seq, AGO-IP, Biotin miRNA tagging , RNA-Seq, Microarrays, Northern blot, qPCR, Reporter Assay

Table 14: Comparison between LncBase v2 and LncBase v1. The table summarizes the experimental module entries of the two databases, including the number of miRNAs targeting lncRNA transcripts, the unique miRNA:lncRNA interacting pairs, different cell lines and tissues supporting miRNA-related experimental methodologies, analyzed CLIP-Seq libraries and associated studies, experimental conditions, as well as the included low/high throughput experimental methodologies. (Paraskevopoulou MD *et al*, 2015)(117)

The unique features of DIANA-LncBase are highlighted in Table 15, along with a comprehensive summary of other leading repositories indexing experimentally supported miRNA-lncRNA interactions.

	LncBase v2.0	lncRNome	lncReg	NPInter v2.0	Starbase v2.0
miRNA species	*Human, Mouse, Epstein-Barr virus, KSHV*	*Human*	*Human, Mouse, Arabidopsis Thaliana*	*Human, Mouse, Danio rerio*	*Human, Mouse, C.elegans*
lncRNAs in interactions	>3,500	66	14	~1,400	1,149
miRNAs in interactions	~1,400	1,205	24	~25	383
viral miRNA-lncRNA interactions	✓				
Total interactions	>70,000	>3,700	34	>1,500	>10,000
Experimental Methodologies	CLIP-Seq, AGO-IP, Biotin miRNA tagging, RNA-Seq, Microarrays, Northern blot, qPCR, Reporter Assay	CLIP-Seq	(AGO) RNA pull-down, Northern Blot, qPCR, Reporter Assay, FISH	miR-CLIP(185), Microarrays, qPCR, Reporter Assay	CLIP-Seq
Analyzed Raw High-Throughput Libraries	153 AGO CLIP-Seq libraries				108 RNA-binding Protein CLIP-Seq datasets
Cell Types/Tissues	✓	✓		✓	✓
LncRNA expression information	RNA-Seq	Microarrays			RNA-Seq
miRNA Binding Site conservation	✓				
Pathways-Disease association	✓ (miRPath v3.0)	✓	✓		✓
Competing endogenous RNA interactions	✓ (TarBase v7.0)		✓	✓	✓
lncRNA Resources	GENCODE v21, Refseq, Cabili *et al.*	GENCODE v12, HGNC(186), literature	literature	NONCODE(187), LncRNADisease(188)	GENCODE v17
Version	v2.0	Accessed (April 2013)	Accessed (August 2015)	v2.0	v2.0

Table 15: Comparison of included data, as well as basic features and functionalities of online leading repositories indexing experimentally supported miRNA-lncRNA interactions. (Paraskevopoulou MD *et al*, 2015)(117)

3.4.1 DIANA-LncBase Interface

The database interface has been completely redesigned to provide an intuitive and easy to use application as well as high flexibility to different user queries (Figure 24). DIANA-LncBase v2 interface comprises two distinct modules for *in silico* predicted and experimentally supported miRNA-lncRNA interactions.

Module for Experimentally supported interactions.

Indexed interactions were enhanced with extensive metadata regarding the supporting publication, type of regulation, experimental methodologies used for miRNA-lncRNA interaction validation, experimental design (including treatment and conditions), as well as cell types and tissue information. Most of the experimentally supported interactions are now coupled with information regarding their genomic location. An advanced filtering/query panel for experimental methodologies, relevant cell types and species is also provided, in order to enable users to identify cell type and tissue-specific miRNA-lncRNA interactions.

Module for in silico predicted interactions.

Predictions are enriched with information concerning MRE binding sites, structures and conservation. miRNA-lncRNA interactions can be visualized upon selection in an interactive UCSC genome browser (152) graphic (Figure 25), where the user is facilitated with all browser options and additional informative tracks. Prediction interaction score and lncRNA tissue/cell type expression can be utilized for filtering the displayed results.

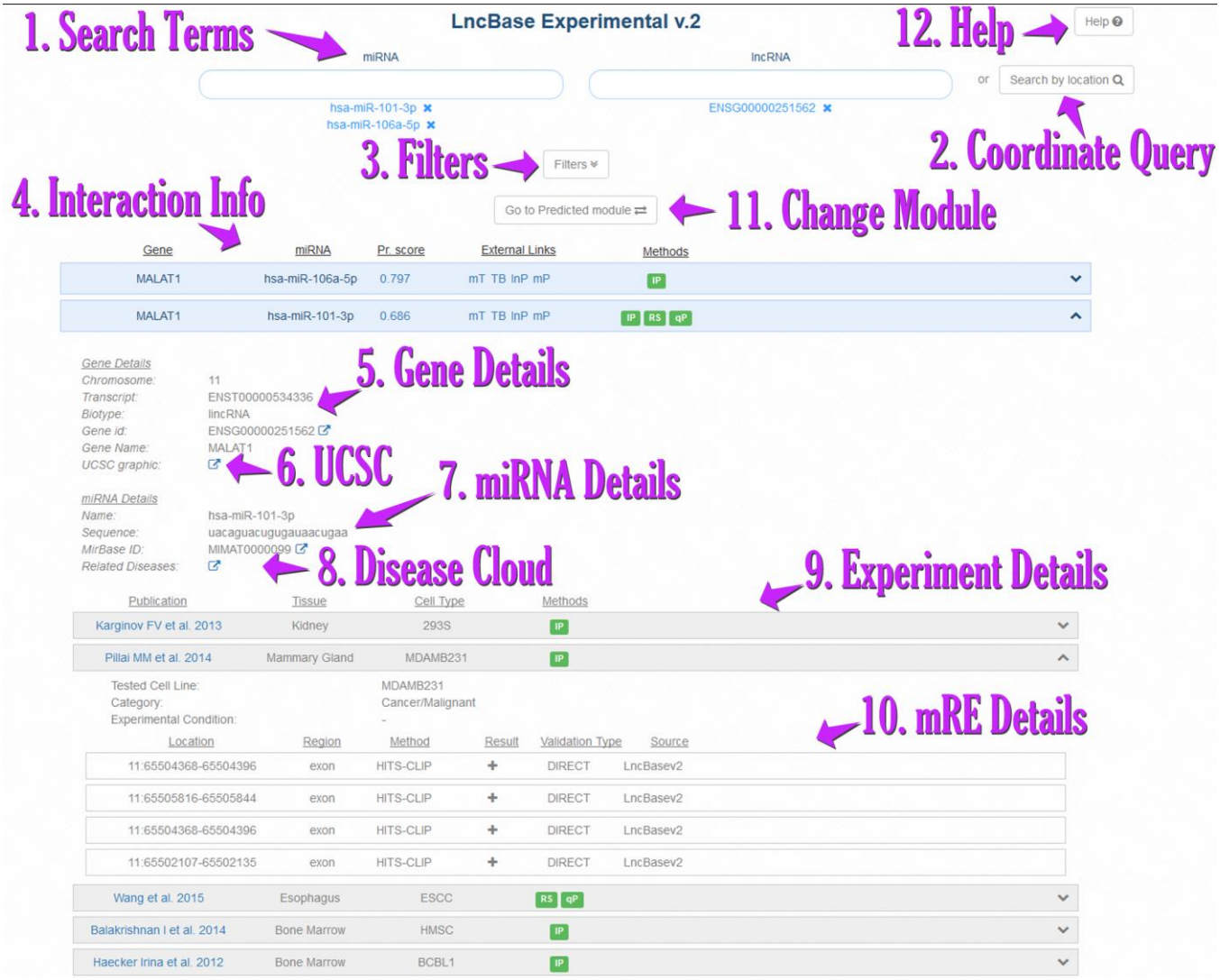

Figure 24: Snapshot depicting the DIANA-LncBase v2 interface. Queries using one or more miRNAs and/or lncRNAs (1) or even the coordinates of a genomic location (2) are supported. Users can add and remove search terms or filter (3) their results based on cell/tissue type and experimental methodology, as well as the experimental outcome (positive/negative) or type of validation (direct/indirect). LncBase offers extensive information for each identified interaction, such as gene/miRNA details (4,5), as well as active links to UCSC graphical representation (6), Ensembl, miRBase and DIANA disease tag cloud (8). LncBase also provides useful information for each performed experiment (9), including the methodology, cell or tissue that was utilized, as well as a link to the original publication. There are direct links to external applications, such as microT, TarBase, miRPath, where the studied miRNAs can be further examined. Interactions are also coupled with miRNA binding site details (10). Users can navigate between the Experimental and Predicted LncBase v2 modules (11). The Help button (12) leads to the LncBase Help section. (Paraskevopoulou MD *et al*, 2015) (117)

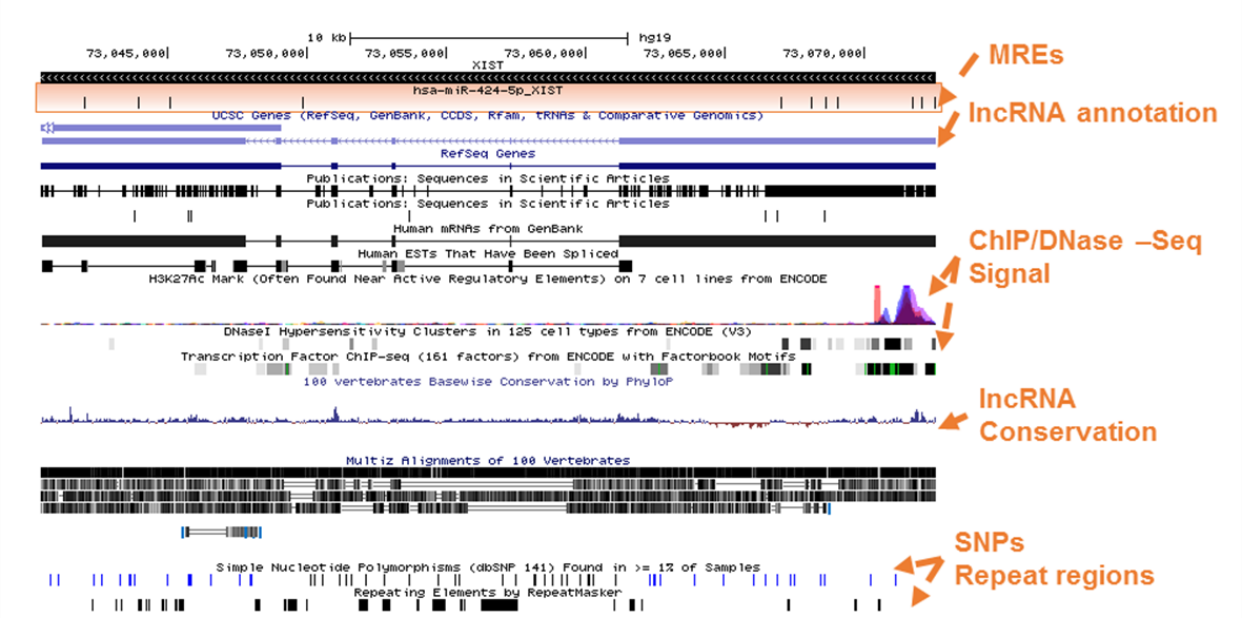

Figure 25: Visualization of a miRNA-lncRNA interaction in UCSC genome browser graphic upon user selection in the LncBase interface. MREs are shown along with the annotated (un)spliced lncRNA transcript. Extra information tracks regarding ChIP/DNase-Seq signal, sequence conservation, SNPs and repeat regions are also provided. The graphical representation is an active link to the UCSC genome browser where the user is facilitated with all the available browser options. (Paraskevopoulou MD *et al.*, 2016) (118)

LncBase v2 indexed interactions are seamlessly interconnected with other available tools in DIANA suite, including TarBase (64) and/or microT-CDS (54) for the identification of competing coding counterparts for miRNA binding and DIANA-miRPath (189) for functional characterization of miRNAs in molecular pathways (Figure 26).

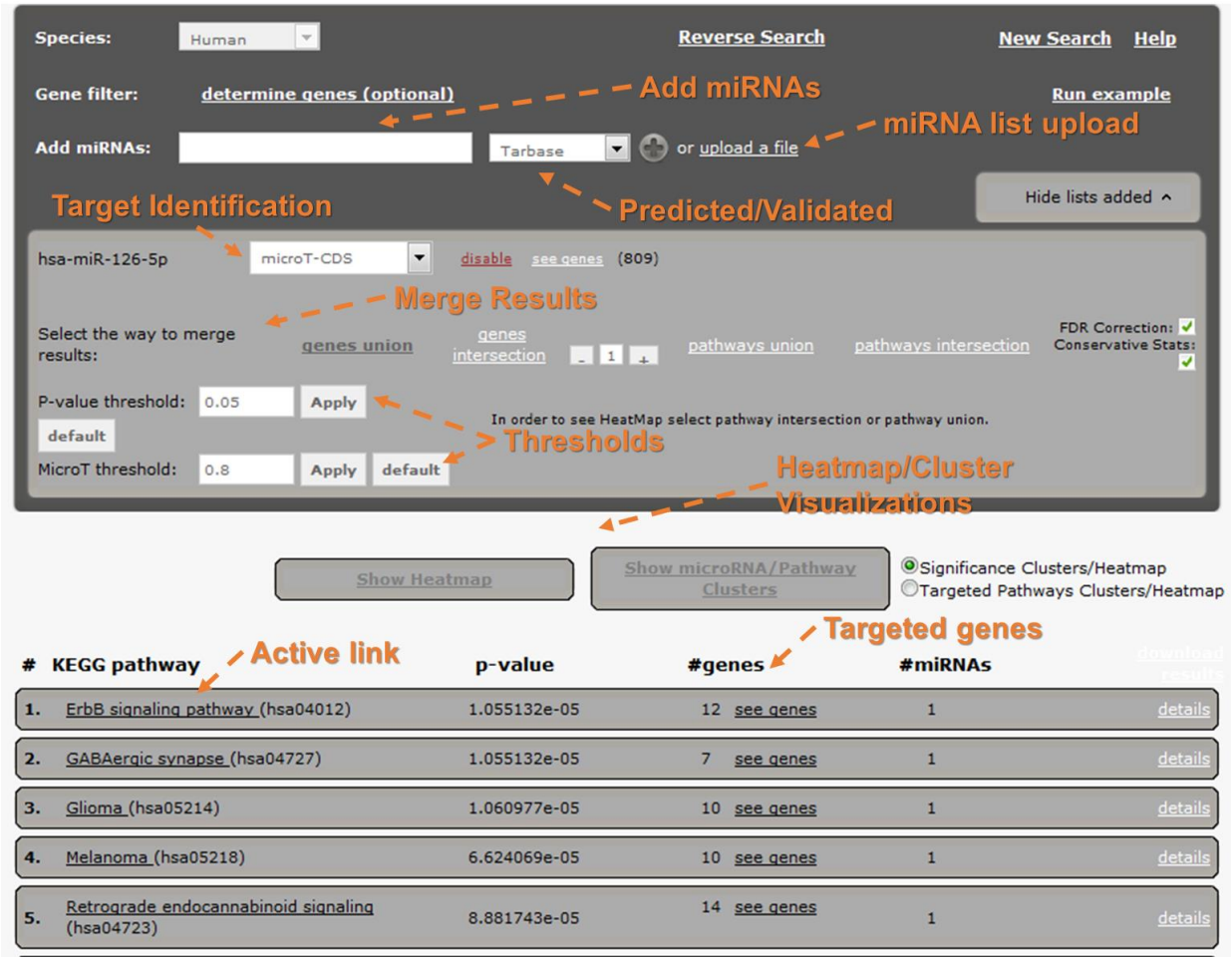

Figure 26: miRNA hsa-miR-126-5p which targets MALAT1 based on LncBase experimentally supported interactions and *in silico* predictions is subjected to a pathway analysis using DIANA-miRPath. Optionally, the user can upload more miRNAs and select to either include their validated or predicted mRNA targets in the functional analysis. Several user-defined options are provided, including, merging method selection, enrichment calculation methodologies as well as parameterization of microT score and p-values of targeted pathways. Sophisticated heatmap/cluster visualizations are available along with pathways merging methods selection. Underlined pathway descriptions are active links to enriched KEGG representations. (Paraskevopoulou MD *et al.*, 2016) (118)

3.5 CLIP-Seq-guided miRNA binding site analysis

The analysis of numerous CLIP-Seq libraries across different cell types and experimental conditions enabled the charting of miRNA-mRNA-lncRNA competing endogenous interactions. This wealth of information has assisted the study of miRNA target repertoire on different gene biotypes as well as of the conservation of MREs in (non)coding regions. The analysis of MREs residing on lncRNA exons additionally unveiled tissue specific miRNA-lncRNA interactions.

3.5.1 Distribution of MREs in (non)coding regions

miRNA binding sites overlapping transcript exons were predominantly encountered in CDS and 3'UTR regions of mRNAs, which was consistent in all cell types and tissues. The analyses of >100 CLIP-Seq libraries in human revealed that 91 ± 5% of the identified MREs were found on CDS and 3'UTR regions, and 5% ± 2% on intergenic, sense, antisense and processed lncRNA transcripts (Figure 27). A similar distribution of miRNA targeted regions was observed in the HITS-CLIP datasets in mouse (Figure 28).

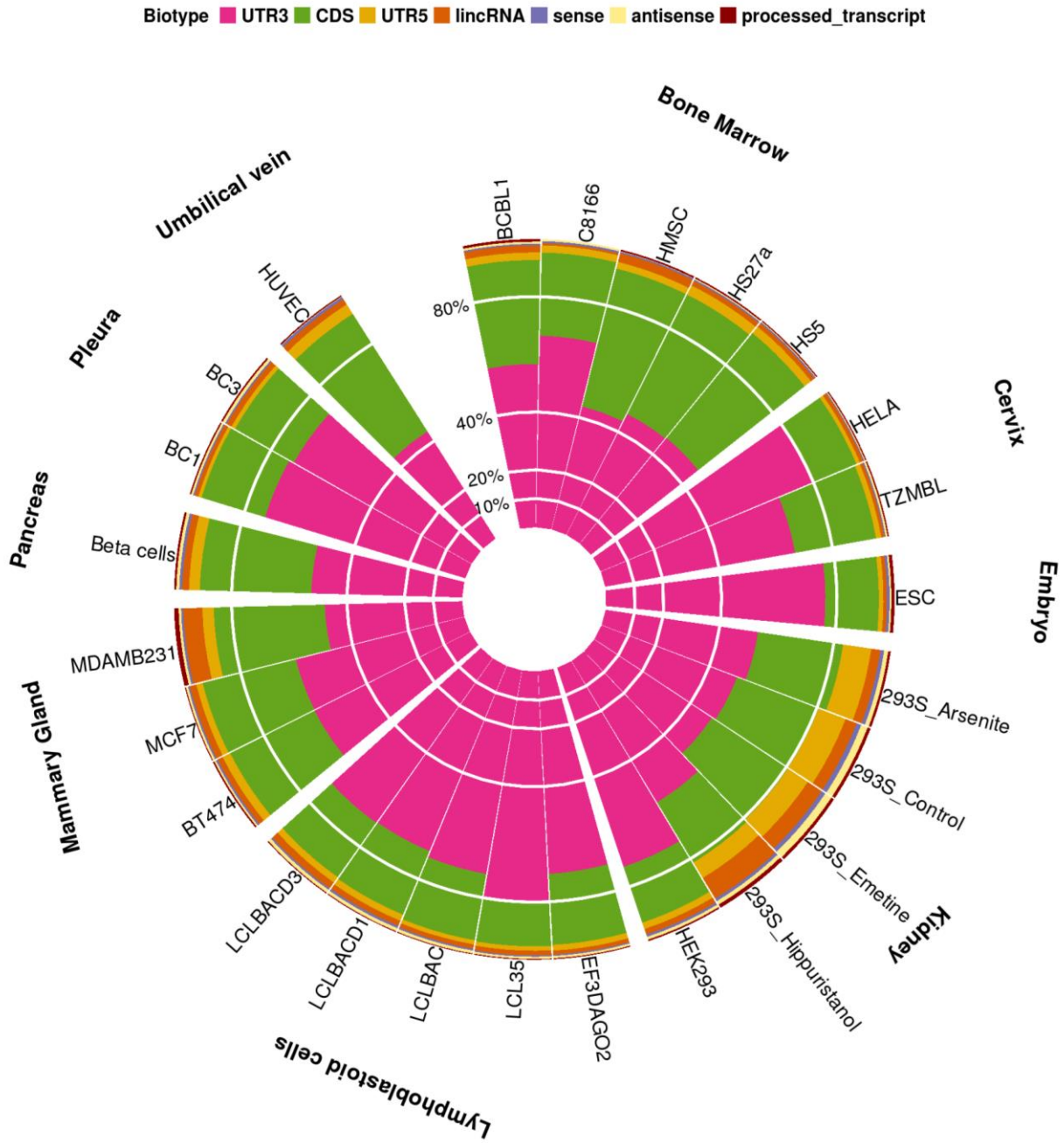

Figure 27: Spatial classification of miRNA-targeted regions as identified in human CLIP-Seq libraries. MREs are being distributed in 3'UTR, 5'UTR, CDS, lincRNA, (anti)sense and processed lncRNA transcript regions across different cell types, with 5 ± 2% of the exonic MREs were annotated on lncRNAs. (Paraskevopoulou MD *et al*, 2015) (117)

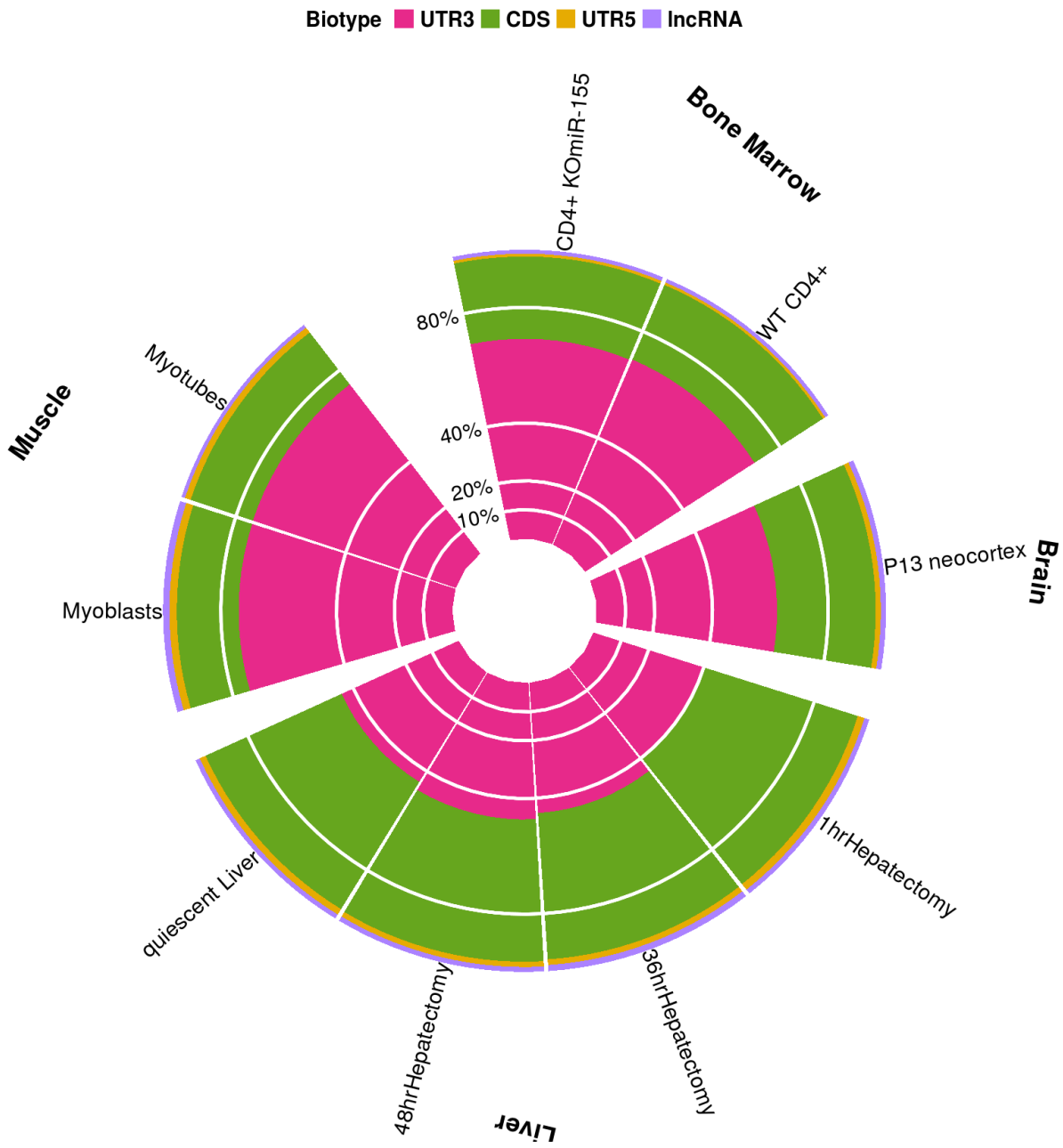

Figure 28: Spatial classification of miRNA-targeted regions as identified in mouse CLIP-Seq libraries. MREs are distributed in 3′UTR, 5′UTR, CDS and lncRNA transcript regions across different cell types. 2 ± 0.3% of the exonic MREs were annotated on lncRNAs. LincRNA, sense, antisense and processed transcripts are grouped together under the umbrella term lncRNA. (Paraskevopoulou MD *et al*, 2015) (117)

3.5.2 Clustering of cell types on targeted lncRNAs

CLIP-Seq libraries from different cell types were hierarchically clustered based on the identified miRNA-lncRNA interactions. Specific cell type groups such as lymphoblastoid, HeLa and bone marrow-derived cell lines in human were found clustered together in the resulting dendrogram; depicting a high similarity in the identified interactions (Figure 29). Similar clusters were also observed in targeted mouse lncRNAs of muscle cognate cell lines and thymocytes which are also densely grouped in the dendrogram (Figure 30).

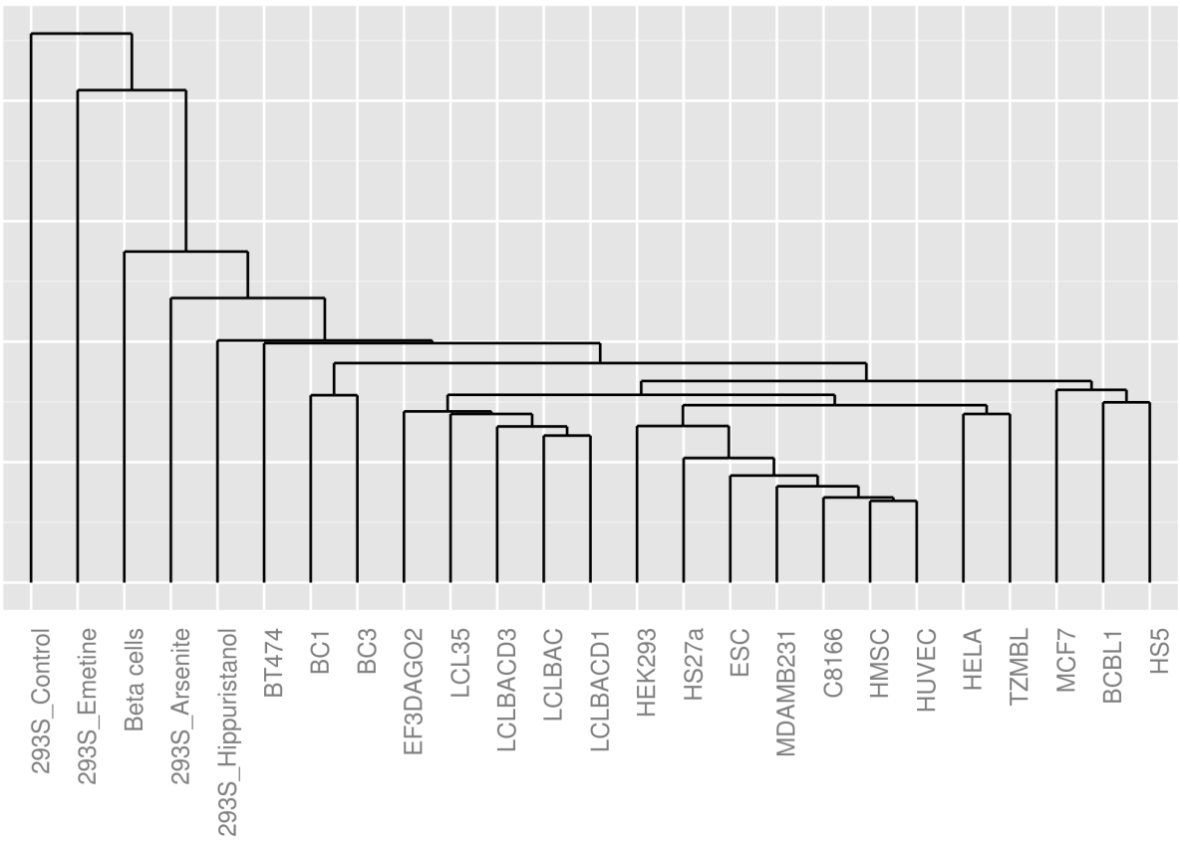

Figure 29: Cell types hierarchically clustered based on targeted human sense, antisense, intergenic and processed lncRNA transcripts. All data included in the dendrogram have been retrieved from analyzed CLIP-Seq libraries spanning different cell types. (Paraskevopoulou MD *et al*, 2015) (117)

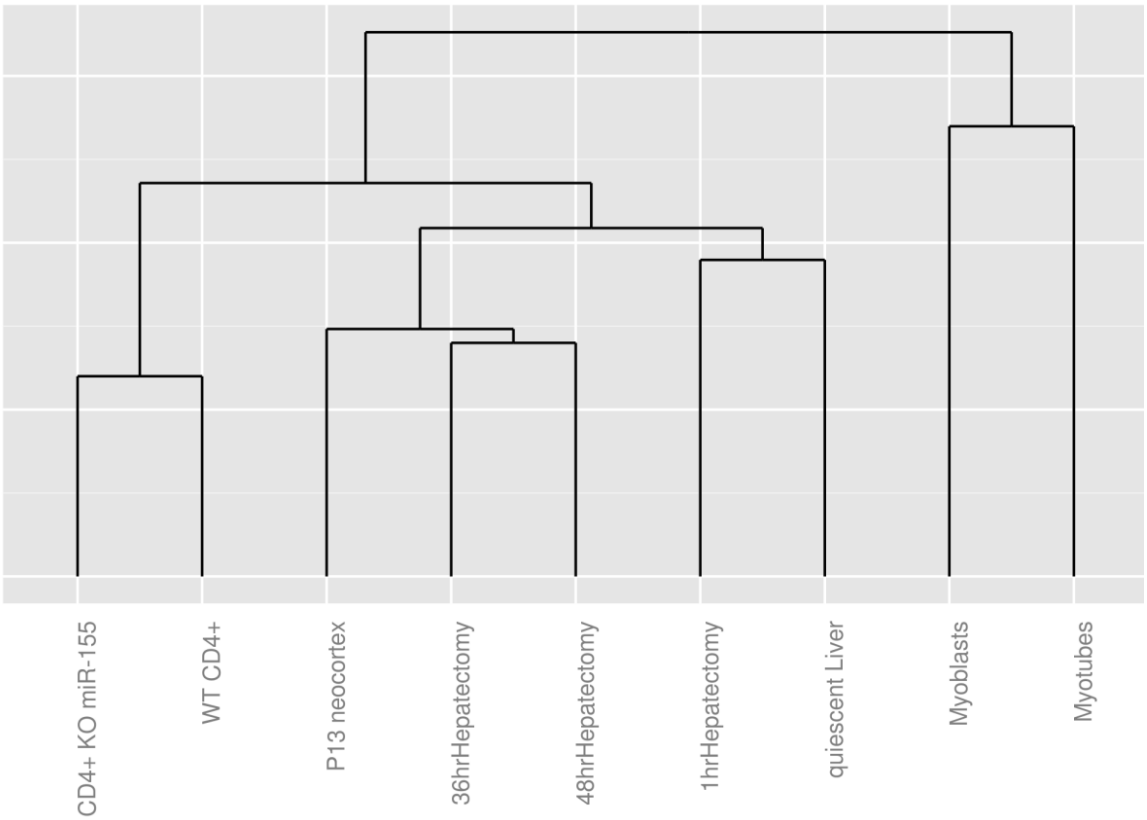

Figure 30: Cell types hierarchically clustered based on targeted mouse lncRNAs. All interactions included in the dendrogram have been derived from analyzed CLIP-Seq libraries across different cell types and tissues. (Paraskevopoulou MD *et al*, 2015)(117)

3.5.3 Conservation of MRE regions

PhyloP (190) pre-computed scores from genome-wide multiple alignments of 46 and 60 vertebrate species for human and mouse, respectively, were utilized to assess evolutionary rates of miRNA targeted regions. PhyloP precompiled values were downloaded from the UCSC repository (152). Conservation signals of MRE regions were estimated as mean intensities of the overlapping PhyloP base-wise scores.

A non-redundant set of collapsed MREs collected from all analyzed CLIP-Seq datasets was defined and annotated accordingly to (non)coding exons. MREs with dual annotation due to overlapping transcript regions were excluded from the analysis. In all pairwise comparisons of conservation, binding sites positioned on lincRNA introns were considered as a separate category. Stronger evolutionary pressure was observed in miRNA binding sites identified on coding and untranslated mRNA regions. MREs on lncRNA exons were significantly more conserved than those residing in introns, while no differences were observed in substitution rates of MREs on intergenic, sense, antisense and processed lncRNA transcripts (Figure 31, Table 16). Statistical analysis of MRE conservation has also been performed for experimentally supported binding sites on mouse lncRNAs (Figure 32, Table 17).

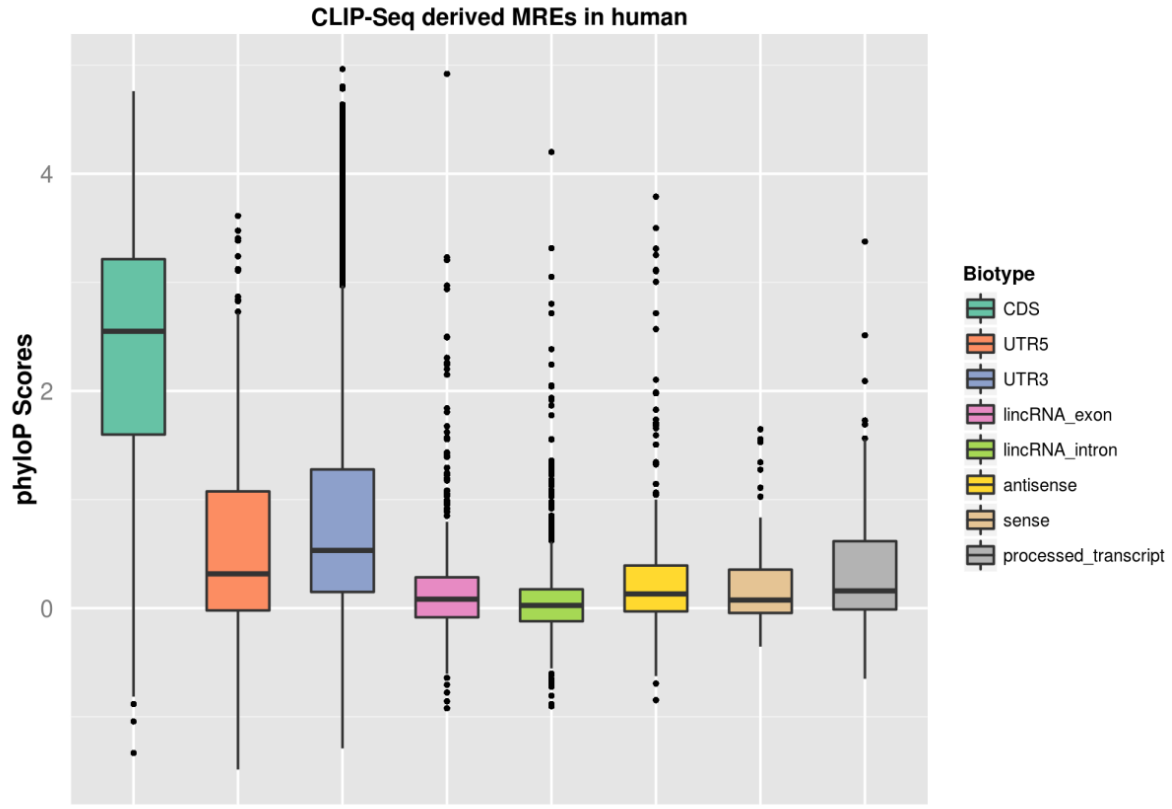

Figure 31: Evaluation of human MRE substitution rates. CLIP-Seq-supported miRNA binding sites on human were spatially classified on CDS, 3'UTR, 5'UTR, lincRNA exons, lincRNA introns, processed transcripts and (anti)sense lncRNA regions. MRE conservation was estimated using PhyloP pre-computed base-wise values from genome-wide multiple alignments of 46 vertebrate species. Binding sites on mRNA regions (CDS, 3'UTR, 5'UTR) were significantly more conserved than the MREs found on lncRNA exons. LincRNA, sense, antisense and processed transcripts presented similar substitution rates. Weaker evolutionary pressure ($p<0.05$) was observed in MREs on lincRNA introns compared to those located on lincRNA exons. (Paraskevopoulou MD *et al*, 2015) (117)

Human	Antisense	CDS	LincRNA_ exon	LincRNA_ intron	Processed_tr anscript	Sense	UTR3
CDS	2.0e-16	-	-	-	-	-	-
lincRNA_ exon	0.24796	2.0e-16	-	-	-	-	-
lincRNA_ intron	1.80e-08	2.0e-16	4.20e-05	-	-	-	-
processed_tr anscript	0.56288	2.0e-16	0.1919	0.00032	-	-	-
Sense	0.56288	2.0e-16	0.56288	0.00805	0.56288	-	-
UTR3	2.0e-16	2.0e-16	2.0e-16	2.0e-16	0.00052	1.60e-11	-
UTR5	0.00107	2.0e-16	5.80e-10	2.0e-16	0.56288	0.00805	3.20e-09

Table 16: FDR-adjusted p-values derived from the statistical analysis of CLIP-Seq-supported human MRE evolutionary rates, spatially classified on CDS, 3′UTR, 5′UTR, lincRNA exons, lincRNA introns, processed transcripts and (anti)sense lncRNA regions. (Paraskevopoulou MD *et al*, 2015) (117)

Mouse	Antisense	CDS	LincRNA_ exon	LincRNA_ intron	Processed_t ranscript	Sense	UTR3
CDS	2.0e-16	-	-	-	-	-	-
lincRNA_ exon	0.20229	2.0e-16	-	-	-	-	-
lincRNA_ intron	0.00411	2.0e-16	2.80e-14	-	-	-	-
processed_t ranscript	1.30e-05	2.0e-16	0.00019	2.0e-16	-	-	-
Sense	0.4401	2.0e-16	0.8879	0.00011	0.00739	-	-
UTR3	2.0e-16	2.0e-16	2.0e-16	2.0e-16	7.30e-12	2.0e-16	-
UTR5	2.0e-16	2.0e-16	2.0e-16	2.0e-16	3.60e-07	6.90e-14	2.10e-05

Table 17: FDR-adjusted p-values derived from the statistical analysis of CLIP-Seq-supported mouse MRE conservation, spatially classified on CDS, 3′UTR, 5′UTR, lincRNA exons, lincRNA introns, processed transcripts and (anti)sense lncRNA regions. (Paraskevopoulou MD *et al*, 2015) (117)

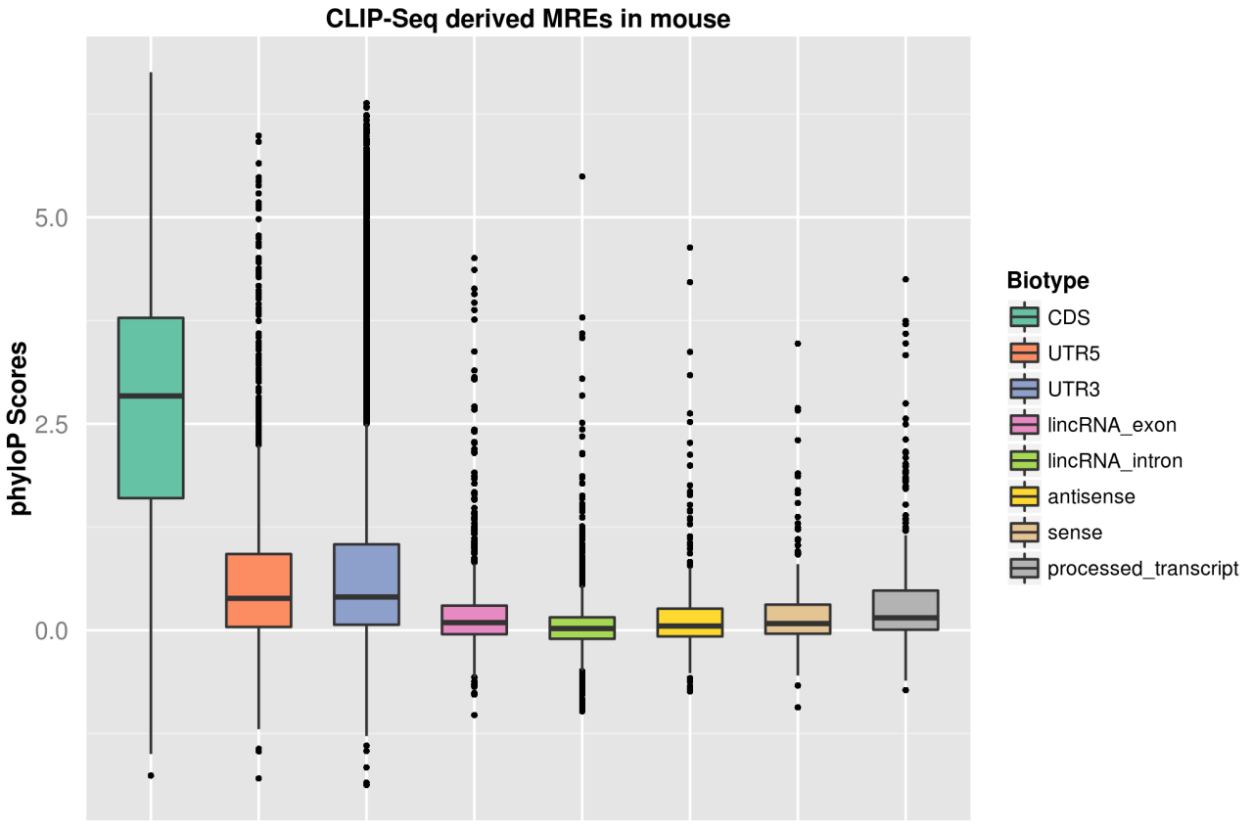

Figure 32: Evaluation of mouse MRE substitution rates. MRE conservation was estimated using PhyloP pre-computed base-wise values from genome-wide multiple alignments of 60 vertebrate species. CLIP-Seq-supported miRNA binding sites were spatially classified on CDS, 3'UTR, 5'UTR, lincRNA exons, lincRNA introns, processed transcripts and (anti)sense lncRNA regions. Binding sites on mRNA regions (CDS, 3'UTR, 5'UTR) were significantly more conserved than the MREs found on lncRNA exons. LincRNA, sense and antisense transcripts presented similar substitution rates. Weaker evolutionary pressure ($p<0.05$) was observed in MREs on lincRNA introns compared to those located on lincRNA exons. (Paraskevopoulou MD *et al*, 2015) (117)

Random background regions retrieved from each spatially classified genomic group were additionally utilized as controls for the assessment of MRE evolutionary pressure. Pairwise comparisons revealed that CLIP-Seq-supported miRNA binding sites in human, even in lncRNA regions, are significantly more conserved than their background sequences (Figure 33), which is a phenomenon previously known to occur in MREs located in mRNA 3'UTRs (25). The evaluation of MRE evolutionary rates among different genomic classes compared to their background in mouse species produced similar results and is presented in Figure 34.

Non-parametric comparisons were performed with Kruskal-Wallis test in order to detect significant differences on substitution rates between multiple groups. Pairwise Mann-Whitney's U tests were adopted as a post-hoc non-parametric test. All p-values were FDR-adjusted to control family-wise error rates due to multiple comparisons (191). All tests were two-sided and p-values < 0.05 were considered as statistically significant.

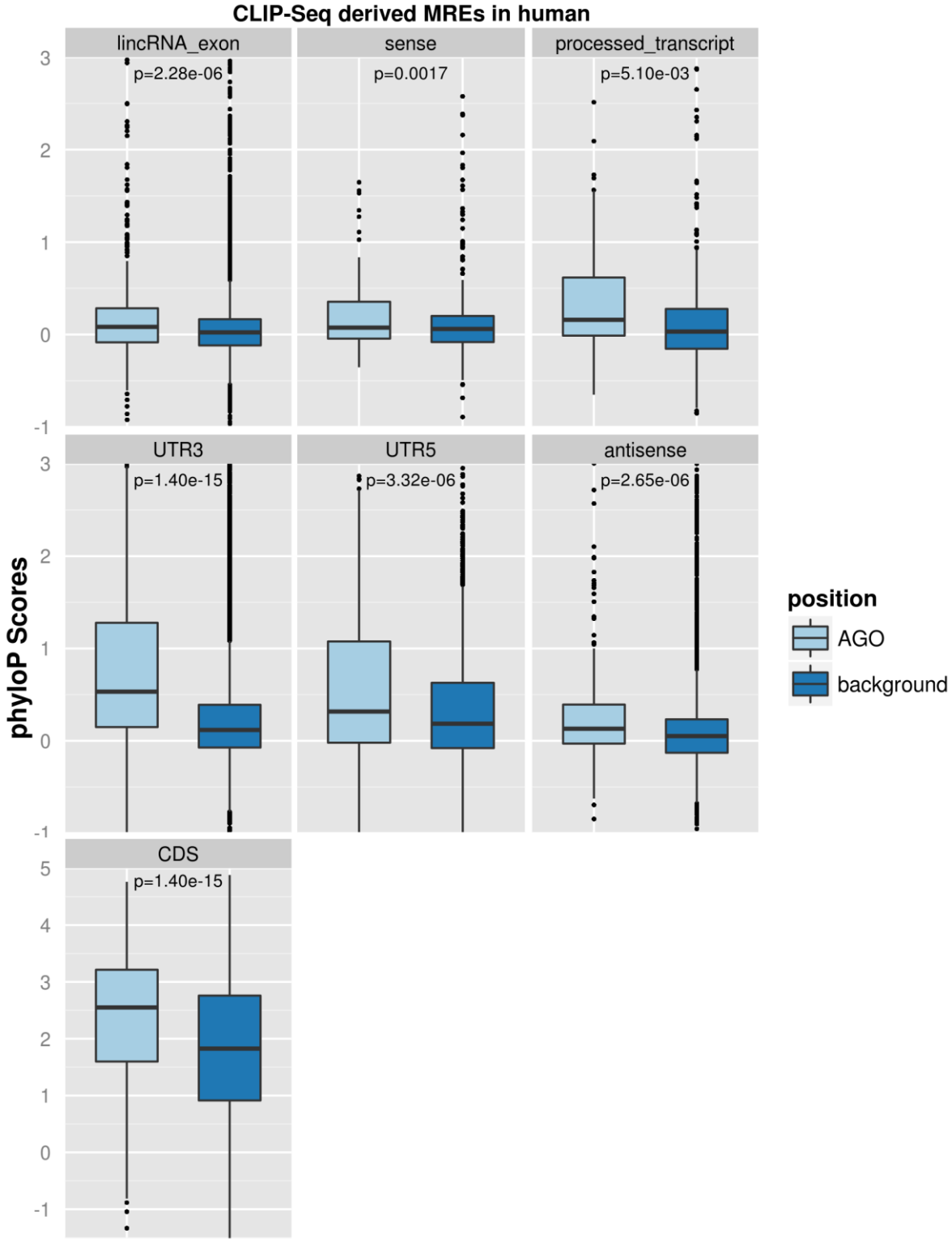

Figure 33: Evaluation of CLIP-Seq-supported human MRE substitution rates. miRNA binding sites were spatially classified on CDS, 3'UTR, 5'UTR, lincRNA exons, processed transcripts and (anti)sense lncRNA regions. Random background regions retrieved from each spatially classified genomic group were additionally utilized as controls for the assessment of MRE evolutionary pressure. MRE and background region conservation were estimated using PhyloP pre-computed base-wise values from genome-wide multiple alignments of 46 vertebrate species. Pairwise comparisons revealed that MREs, even in lncRNA regions, are significantly more conserved than their background sequences, which is

a phenomenon previously known to occur in MREs located in mRNA 3′UTRs. P-values derived from statistical analyses are marked in the relevant panels. (Paraskevopoulou MD *et al*, 2015) (117)

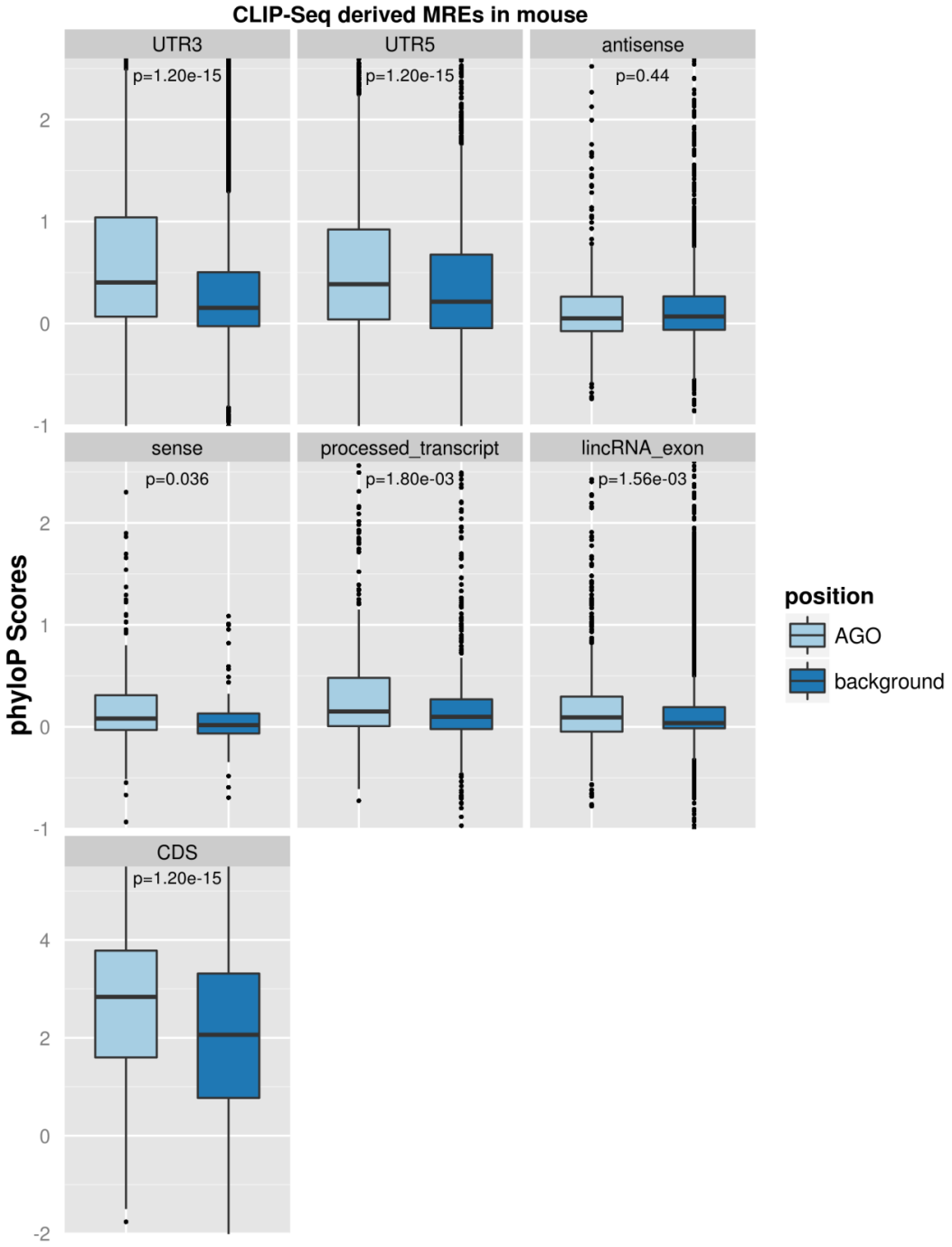

Figure 34: Evaluation of CLIP-Seq-supported mouse MRE substitution rates. Random background regions were utilized as control evolutionary pressure measurements in each group of spatially classified miRNA binding sites on CDS, 3′UTR, 5′UTR, lincRNA exons, processed transcripts and (anti)sense lncRNA regions. MRE and background region conservation was estimated using PhyloP pre-computed base-wise values from genome-wide multiple alignments of 60 vertebrate species. Pairwise comparisons revealed that MREs (even in most lncRNA subgroups) are significantly more conserved than their background sequences. P-values derived from statistical analyses are marked in the relevant panels. (Paraskevopoulou MD *et al*, 2015)(117).

3.5.4 Identification of competing endogenous interactions

By analyzing the experimentally supported interactions available in TarBase and LncBase repositories, we identified thousands of cell type specific miRNA-lncRNA-mRNA trios that can be considered as candidate ceRNAs. The following table summarizes the competing interactions identified per cell type. LncRNAs and mRNAs participating in the interactions are reported only if they have more than 2 miRNA binding sites.

Cell line	Number of lncRNAs in competing interactions	Mean miRNA binding sites per lncRNA	Number of mRNAs in competing interactions	Mean miRNA binding sites per mRNA
293S	38	7.5	826	3.6
BC1	2	3	16	3.2
BC3	1	3	1	3
BCBL1	2	4.3	45	3.7
Beta cells	18	4.8	448	3.6
Brain	69	4	2,683	3.6
BT474	8	9	420	3.6
HEK293	3	3.2	17	3.1
HELA	7	3.6	100	3.5
hMSC	2	3.5	18	3.4
HS27a	2	3.3	17	3.6
HS5	9	5	360	3.5
HUVEC	2	3.5	21	3.6
MCF7	7	5.4	205	3.5
MDAMB231	1	3	1	3
TZMBL	2	3	2	3.2

Table 18: Competing interactions identified per cell type. Interactions are derived from the analysis of more than 150 raw AGO-CLIP-Seq libraries. LncRNAs and mRNAs participating in the interactions are reported only if they have more than 2 miRNA binding sites.

3.6 Evaluation of Tarbase/LncBase AGO-CLIP-Seq data Analysis performance against other CLIP-Seq Target Identification Algorithms

The in-house implemented algorithm for the analysis of AGO-CLIP-Seq data is central to DIANA-tools. Therefore, it has been extensively tested against collections of experimental targets. The evaluation of AGO-CLIP implementations is a complex and laborious procedure. Even if thousands of experimentally verified miRNA-gene interactions have been already indexed, only a small portion corresponds to validated specific negative interactions. Therefore, in the following comparisons (Figure 35, Figure 36) correctly predicted experimentally supported interactions are included.

From the performed tests, the algorithm outperforms state-of-the-art approaches for MRE identification in CLIP-Seq data, such as MIRZA, microMUMMIE and PARMA (Figure 35). CLIP target identification implementations currently manage to identify ~25% of the experimentally validated binding sites and to provide one valid miRNA binding site in approximately every 4 predicted targets. This result shows that state of the art implementations need further optimization and improvement.

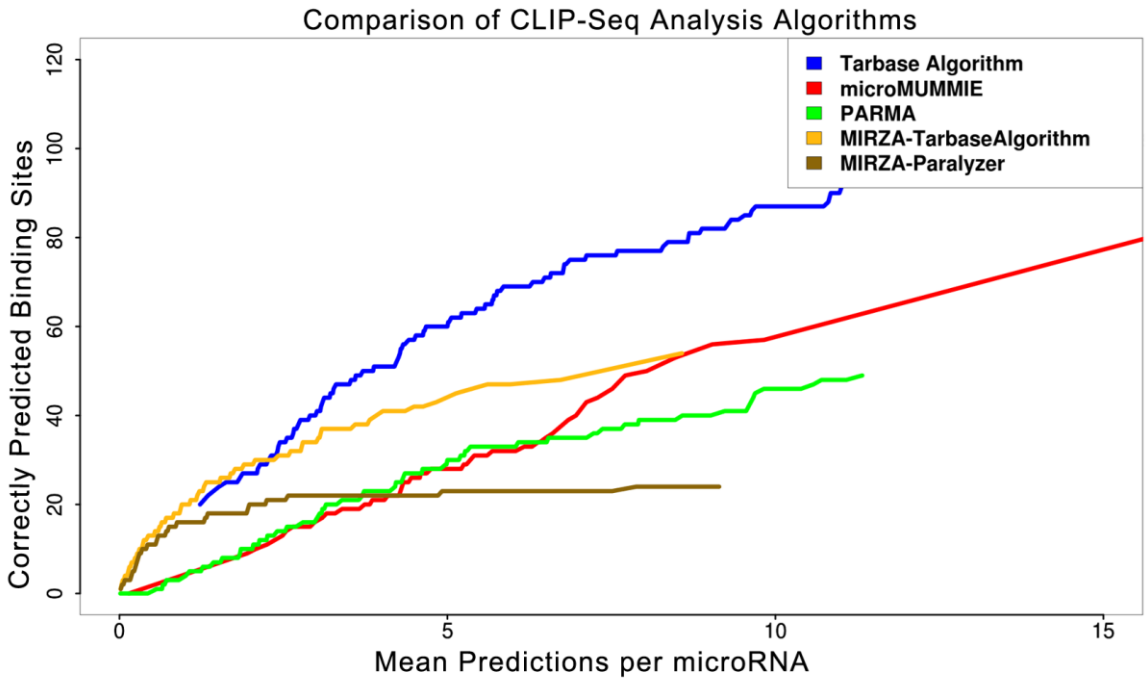

Figure 35: CLIP-Seq algorithm comparison against a unified positive set of Reporter Luciferase Gene Assays and Chimeric interactions. The number of correctly predicted miRNA binding sites vs mean predicted interactions per miRNA is shown for different interaction score thresholds.

In another evaluation, CLIP-Seq adopted pipeline performance has been tested against the biophysical model MIRZA. In this comparison, two distinct high quality sets of experimentally verified interactions with positive regulation, derived from DIANA-TarBase v7, were utilized. The first comprised 1,655 TarBase v7.0 indexed interactions from ~300 Luciferase Reporter Gene Assays and ~1,300 chimeric interactions (CLASH)(173) in HEK293T cells. The second incorporated an extended set of ~850

interactions validated with Luciferase Reporter Gene Assays. For all selected interactions, the exact binding site coordinates had to be known.

The number of correctly predicted miRNA binding sites versus total predictions for different prediction score thresholds is depicted in the following figure (Figure 36a,b). The results demonstrate that CLIP-Seq analysis algorithms are more efficient in stricter prediction scores. It should be noted that the MIRZA implementation provides true positive predictions approximately for 30% of the included miRNAs, while our approach identifies correctly more than half of the miRNAs (50+%).

Since MIRZA requires miRNA expression values, it cannot be used also in the second dataset (1b). The DIANA CLIP algorithm manages to identify more than half of the experimentally supported interactions and to provide approximately one externally validated (with another technique) miRNA binding site in every 4 predicted MREs.

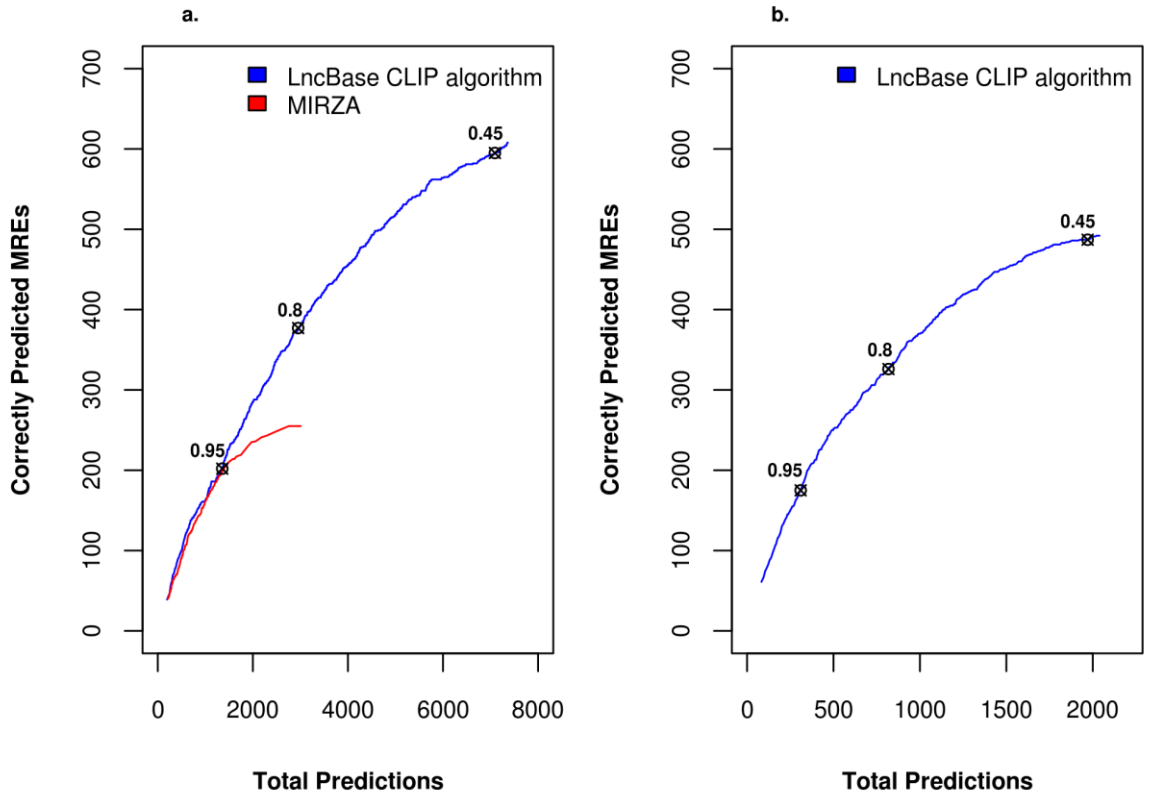

Figure 36: Evaluation of CLIP-Seq algorithm performance. The selected points indicate the performance of the implementations from loose to strict prediction scores. a) The number of correctly predicted miRNA binding sites by our in-house-developed CLIP algorithm and MIRZA versus total predictions for different interaction score thresholds. The utilized validation set comprised 1,655 experimentally validated interactions from ~300 Luciferase Reporter Gene Assays and ~1300 chimeric CLASH interactions. b) LncBase CLIP-Seq algorithm performance evaluation in a set of ~850 Luciferase Reporter Gene Assays spanning different cell types. Approximately 1 externally validated miRNA binding site is provided in every 2 predicted MREs by using score thresholds of moderate stringency.

3.7 Evaluation of a novel algorithm for CLIP-Seq-guided miRNA-target identification.

The subsequent sections describe retrieved outcomes from the descriptor preprocessing and assessment prior to feature selection. Each descriptor is independently evaluated for its predictive accuracy using ROC curves. ROC plots are selectively presented below for a handful of prominent and top performing descriptors (1 dimension). Broad sub-groups of the initial feature set are also explored for in-between associations (data are shown on the following correlation heat maps). The accuracy of base Random Forest classifiers coupled with each model internal feature ranking is additionally presented. The final GBM meta-classifier is evaluated for its performance to accurately predict positive and negative instances derived from an independent test set. Consclusively, the performance of the introduced algorithm is evaluated against other state-of-the art implementations, including the computational approach adapted by TarBase/LncBase for the AGO-CLIP-Seq data analysis.

3.7.1 Feature ROC curves

Several features derived from CLIP-Seq experiments, such as the cluster length, RPKM expression values for MRE regions (Figure 37), descriptors of substitution frequencies (especially T-to-C conversion-related features - Figure 38) as well as substitution distances from relative MRE start sites have presented high predictive performance.

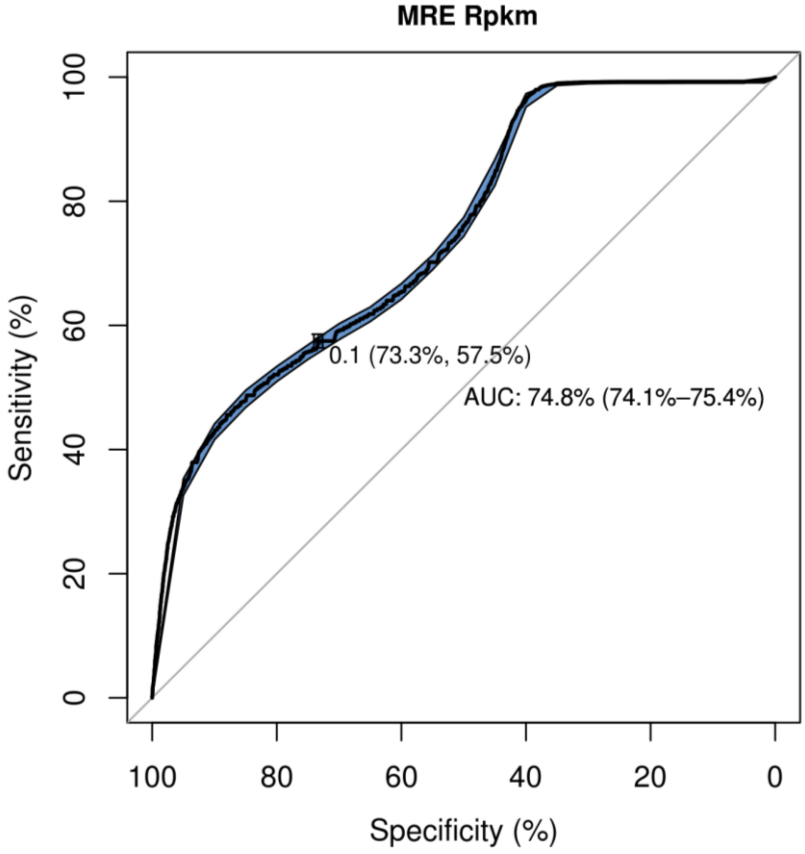

Figure 37: ROC curve of 'MRE RPKM' parameter for the classification of positive/negative miRNA binding sites.

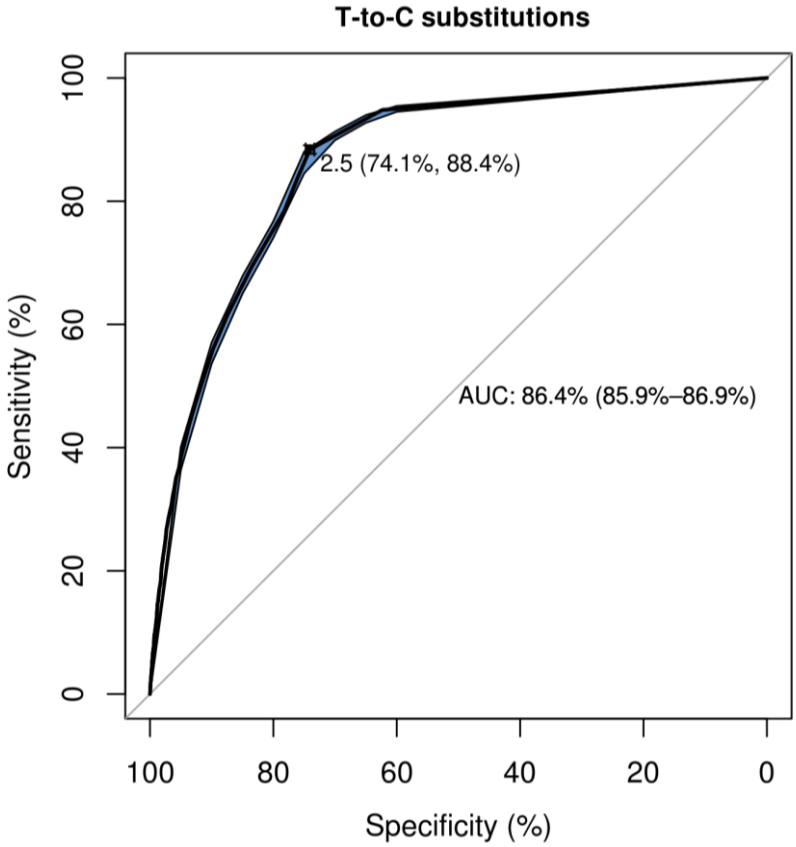

Figure 38: ROC curve of 'T-to-C transitions' parameter for the classification of positive/negative miRNA binding sites.

Certain single or di-nucleotide composition descriptors for overlapping/upstream/downstream MRE regions presented high performance in the one dimensional feature evaluation. Below, ROC curves of upflank-MRE "A or T" (Figure 39) and upflank-MRE "G" (Figure 40) are indicatively provided. Notably, A/U flanking content is deemed important by many miRNA target prediction approaches and it has been associated with accessible miRNA sites. Moreover, "G" enrichment in upflank-MRE region has been associated with RNase cleavage sites.

Thermodynamic MRE properties and MRE content asymmetry including, entropy (dS), enthalpy (dH), free energy (dG), and melting temperature (Tm) and purine skew, exhibited significant difference between CLIP-derived positive/negative miRNA binding sites. Relevant ROC curves of Tm, dS and purine skews are shown in Figures 41-43. These three features are for the first time incorporated in a relevant learning framework.

Finally, ROC AUC curves of prominent features, describing the miRNA binding site are selectively displayed in the Figures 44-48. More precisely, the interaction binding type, consecutive miRNA-target matches in the seed, binding of the first seed nucleotide (MRE position 2), "miRNA C-matches" and AU base pairing in the seed region appeared to significantly differ between CLIP-Seq positive and negative MREs.

Notably, most of the presented descriptors in the ROC curves were also highly ranked in the implemented base classifiers.

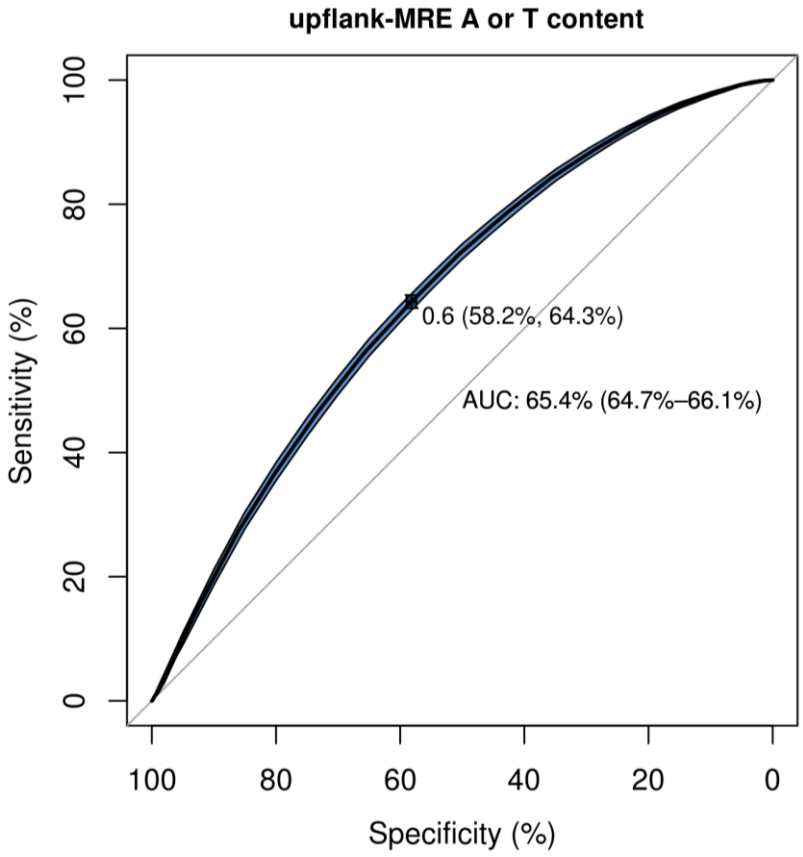

Figure 39: ROC curve of "upflank-MRE A or T content" parameter for the classification of positive/negative miRNA binding sites.

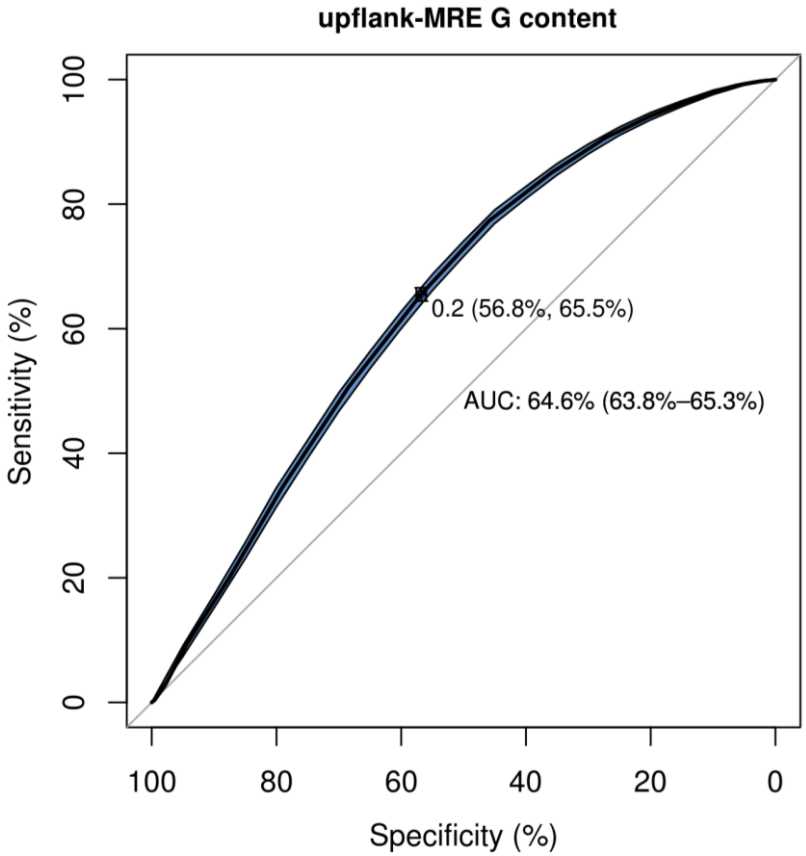

Figure 40: ROC curve of "upflank-MRE G content" parameter for the classification of positive/negative miRNA binding sites.

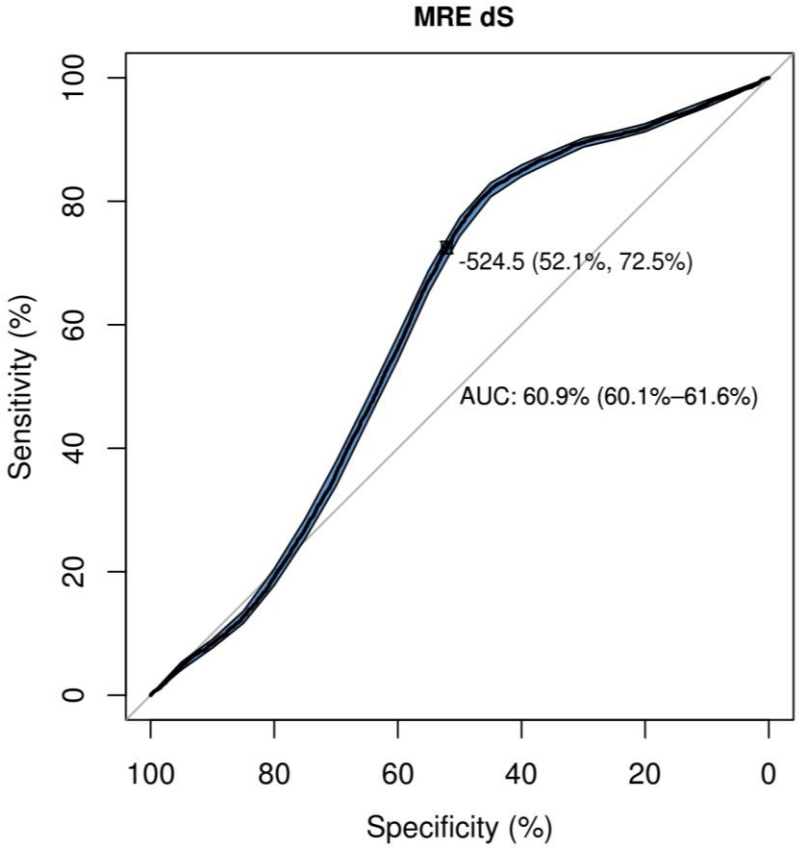

Figure 41: ROC curve of "MRE dS" parameter for the classification of positive/negative miRNA binding sites.

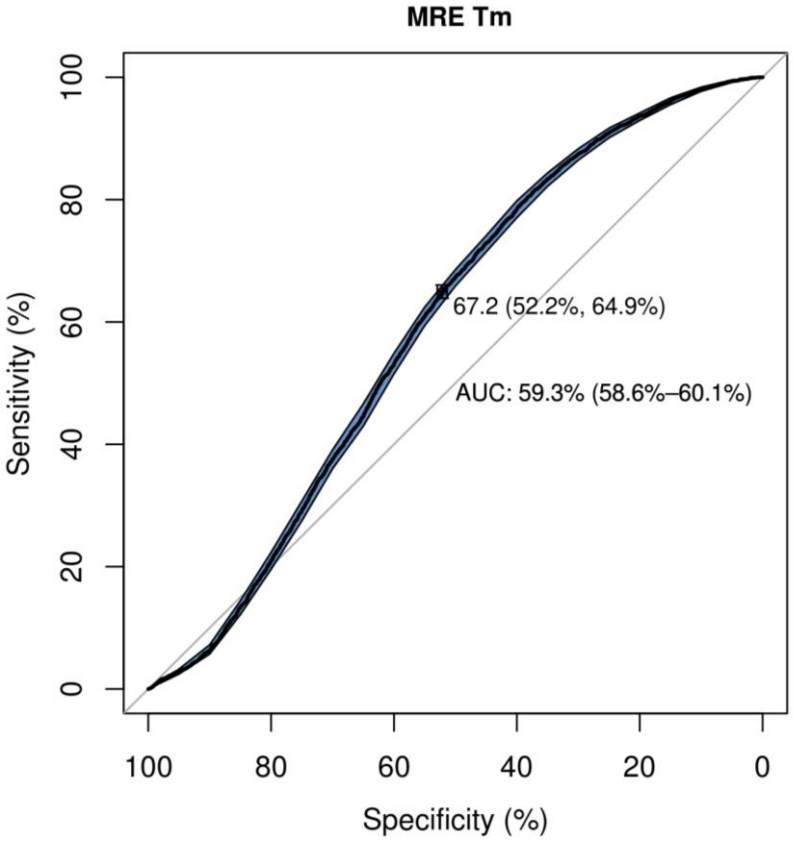

Figure 42: ROC curve of "MRE Tm" parameter for the classification of positive/negative miRNA binding sites.

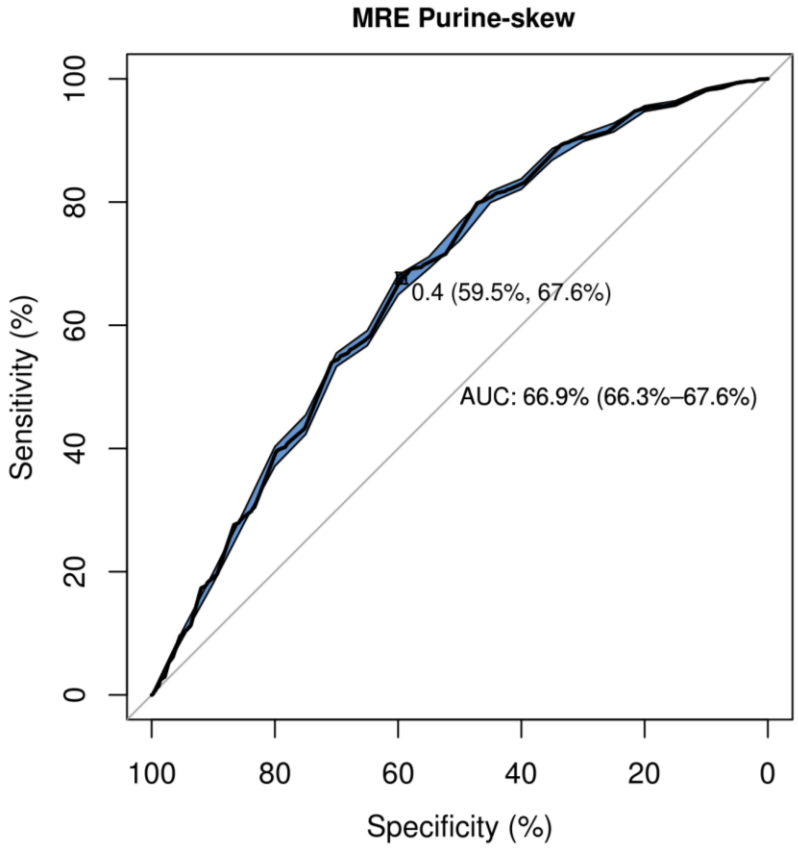

Figure 43: ROC curve of "MRE Purine-skew" parameter for the classification of positive/negative miRNA binding sites.

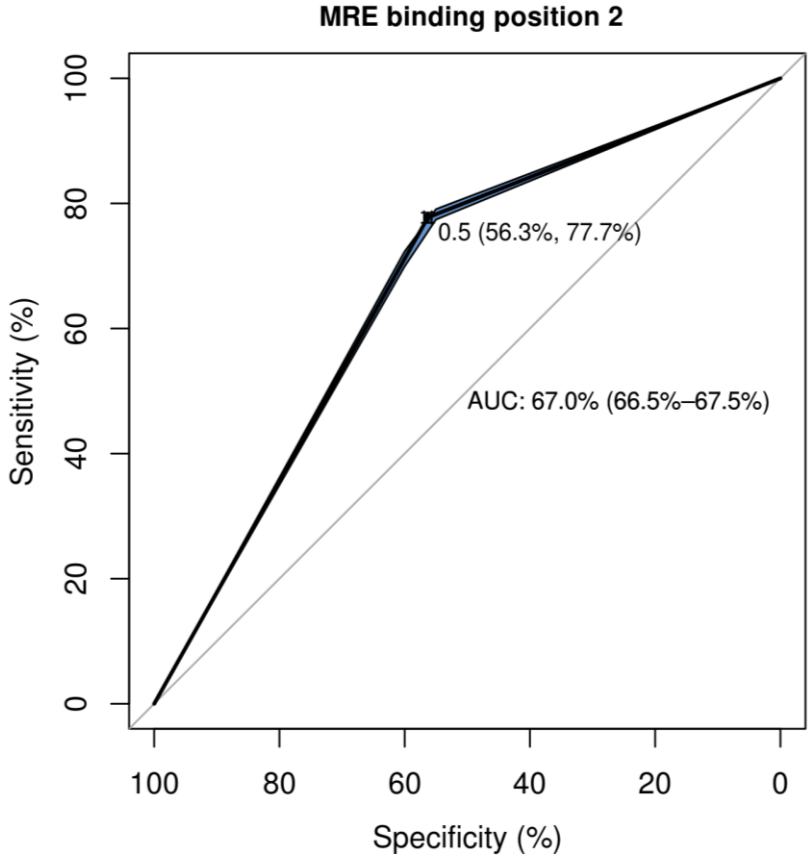

Figure 44: ROC curve of "MRE binding position 2" parameter for the classification of positive/negative miRNA binding sites.

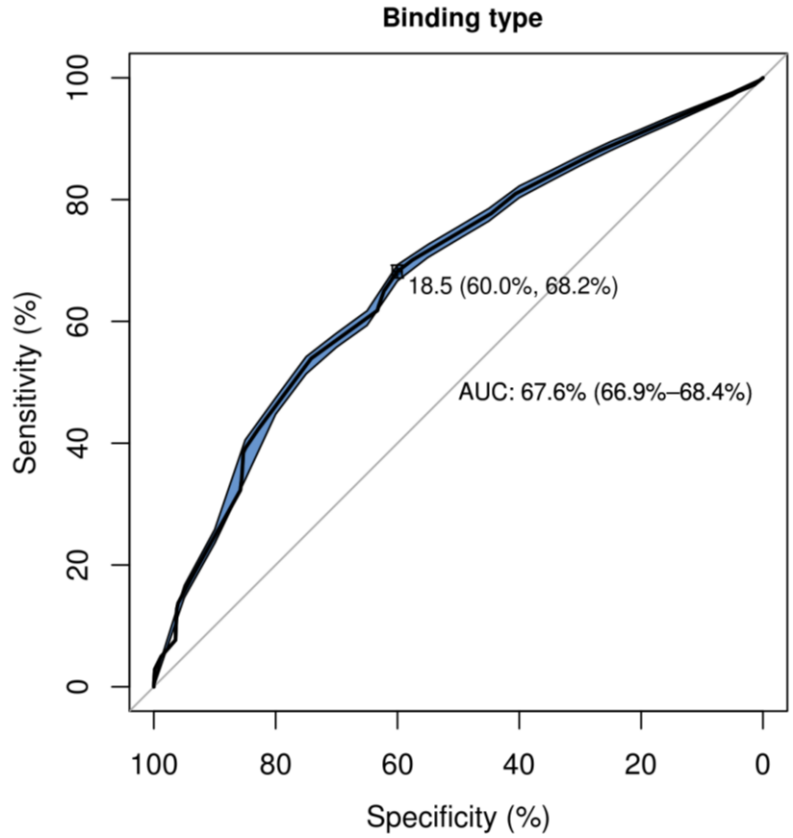

Figure 45: ROC curve of "Binding type" parameter for the classification of positive/negative miRNA binding sites.

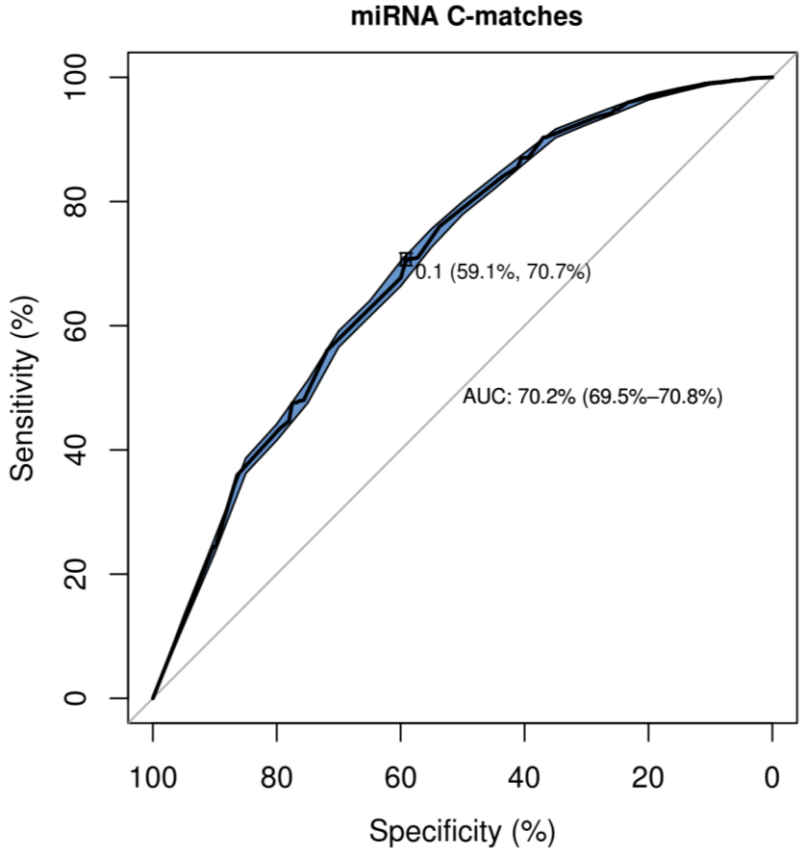

Figure 46: ROC curve of "miRNA C-matches" parameter for the classification of positive/negative miRNA binding sites.

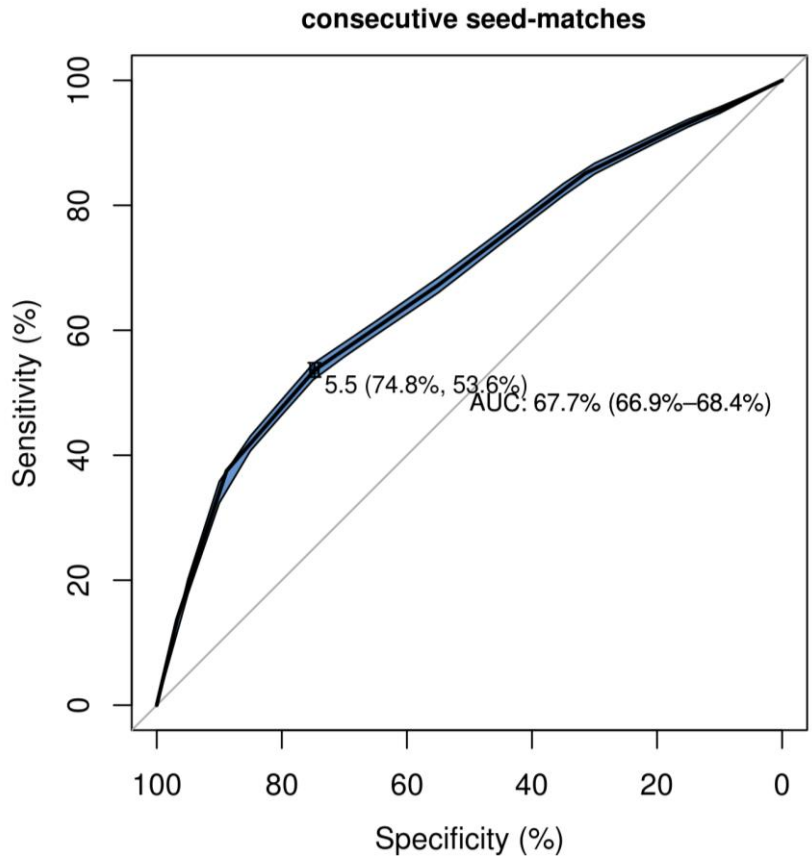

Figure 47: ROC curve of "consecutive seed-matches" parameter for the classification of positive/negative miRNA binding sites.

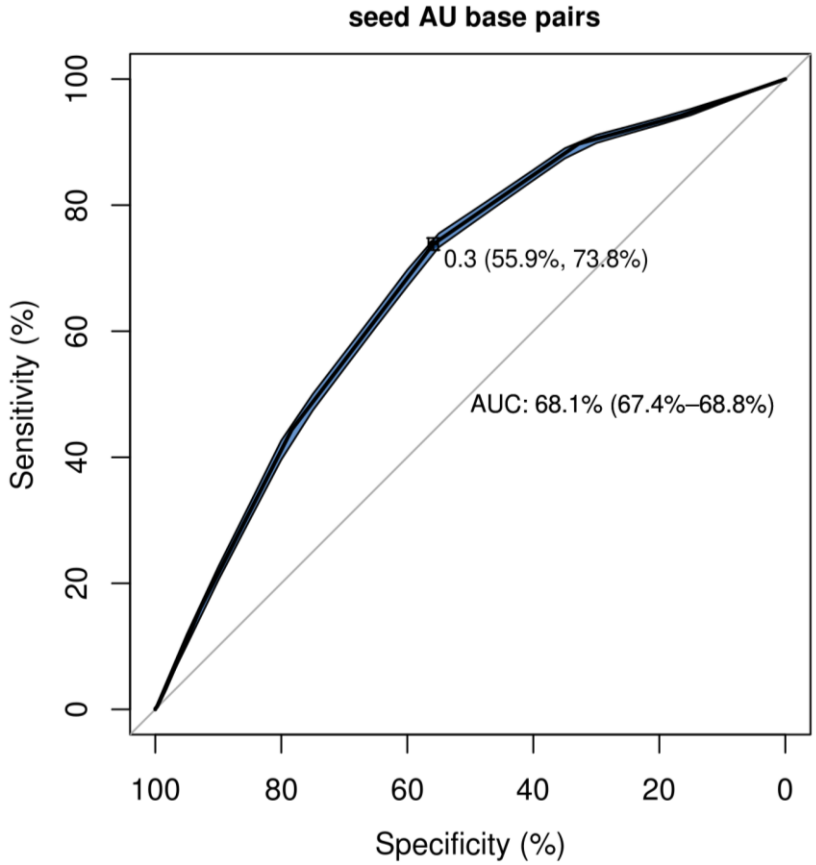

Figure 48: ROC curve of "seed AU base pairs" parameter for the classification of positive/negative miRNA binding sites.

3.7.2 Feature Correlation plots

A common problem of large datasets is the existence of highly correlated parameters. Thus, the primary descriptors collection was appropriately filtered in order to include only unrelated features and to avoid correlation-induced biases in the implemented learning models. Feature correlation estimations revealed several parameters presenting increased (anti)correlation. Figures 49-55 correspond to correlation plots comprising sub-groups of the initial feature set.

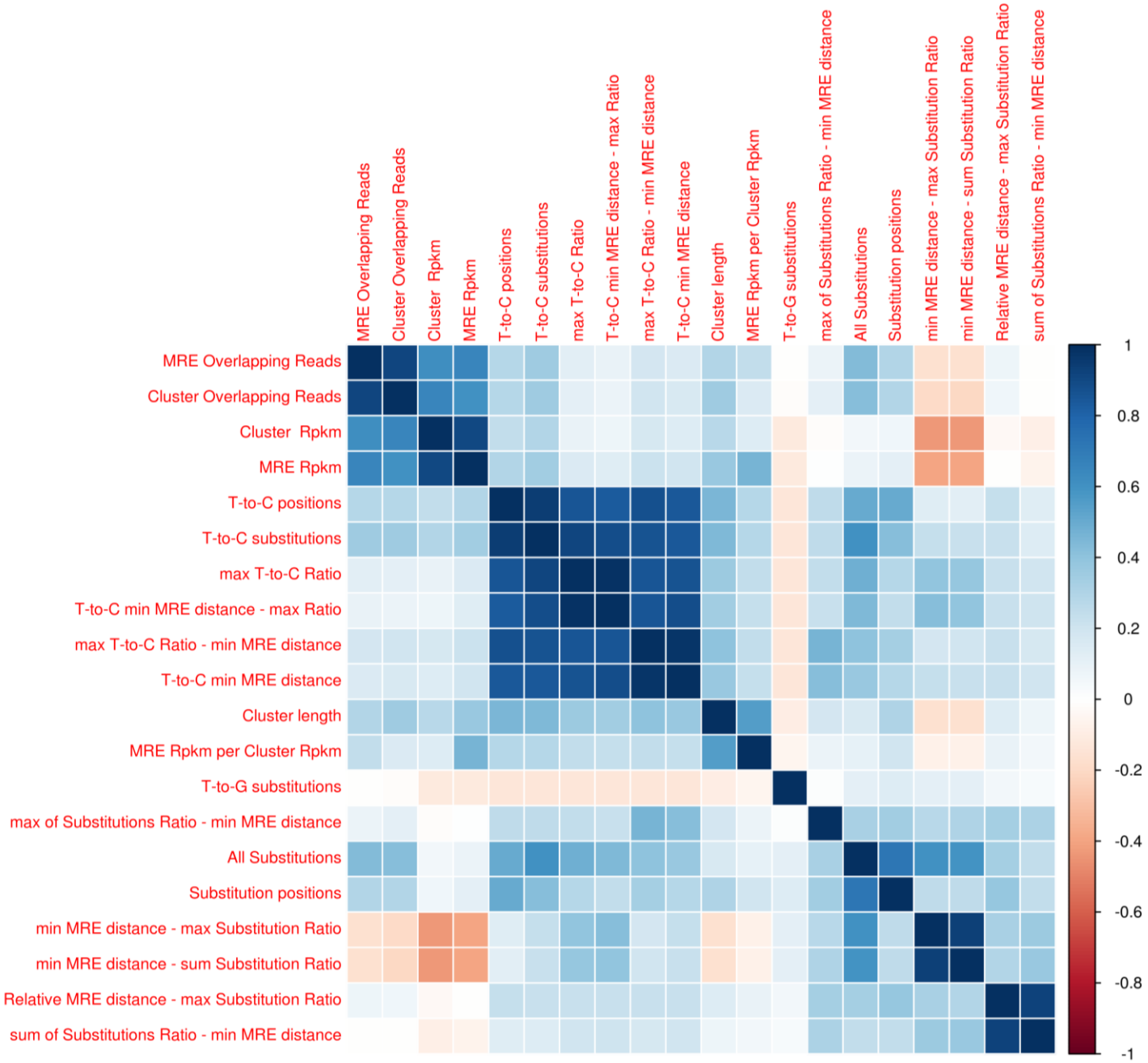

Figure 49: Correlation plot of expression and substitution parameters derived by the processed CLIP-Seq experiments. Cluster overlapping reads and cluster RPKM expression were removed due to high correlation with relative descriptors of the MRE region. Features designed to portray characteristics of

transition events, especially T-to-C related features, were appropriately filtered to retain only unrelated and top performing descriptors. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level p< 0.05.

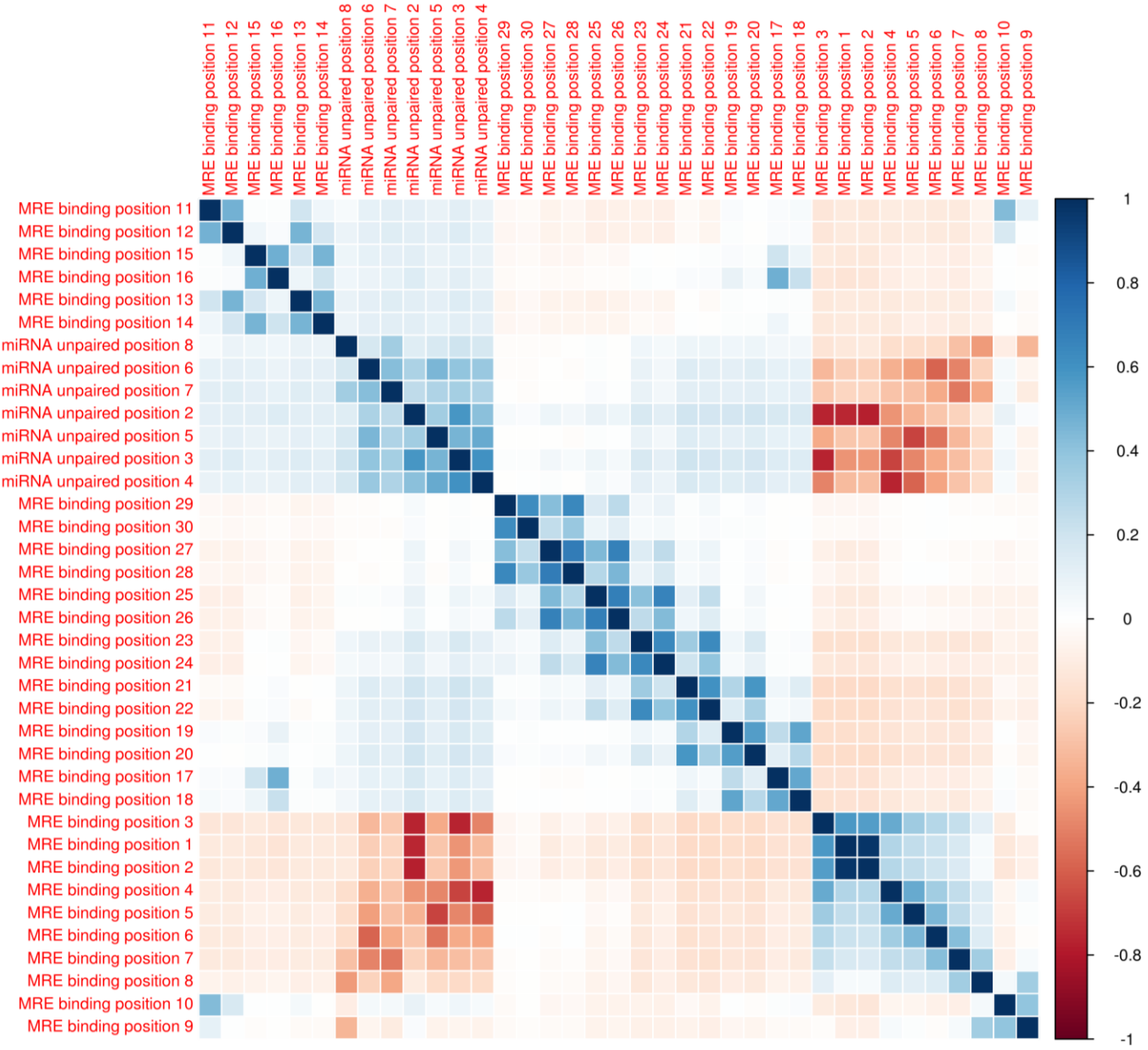

Figure 50: Correlation plot of parameters that reflect the base-wise binding affinity of the MRE and miRNA respectively. miRNA and MRE first binding positions (2-4 seed positions) on the corresponding binary vectors were highly correlated. These features were retained only for the MRE binding vector. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level p< 0.05.

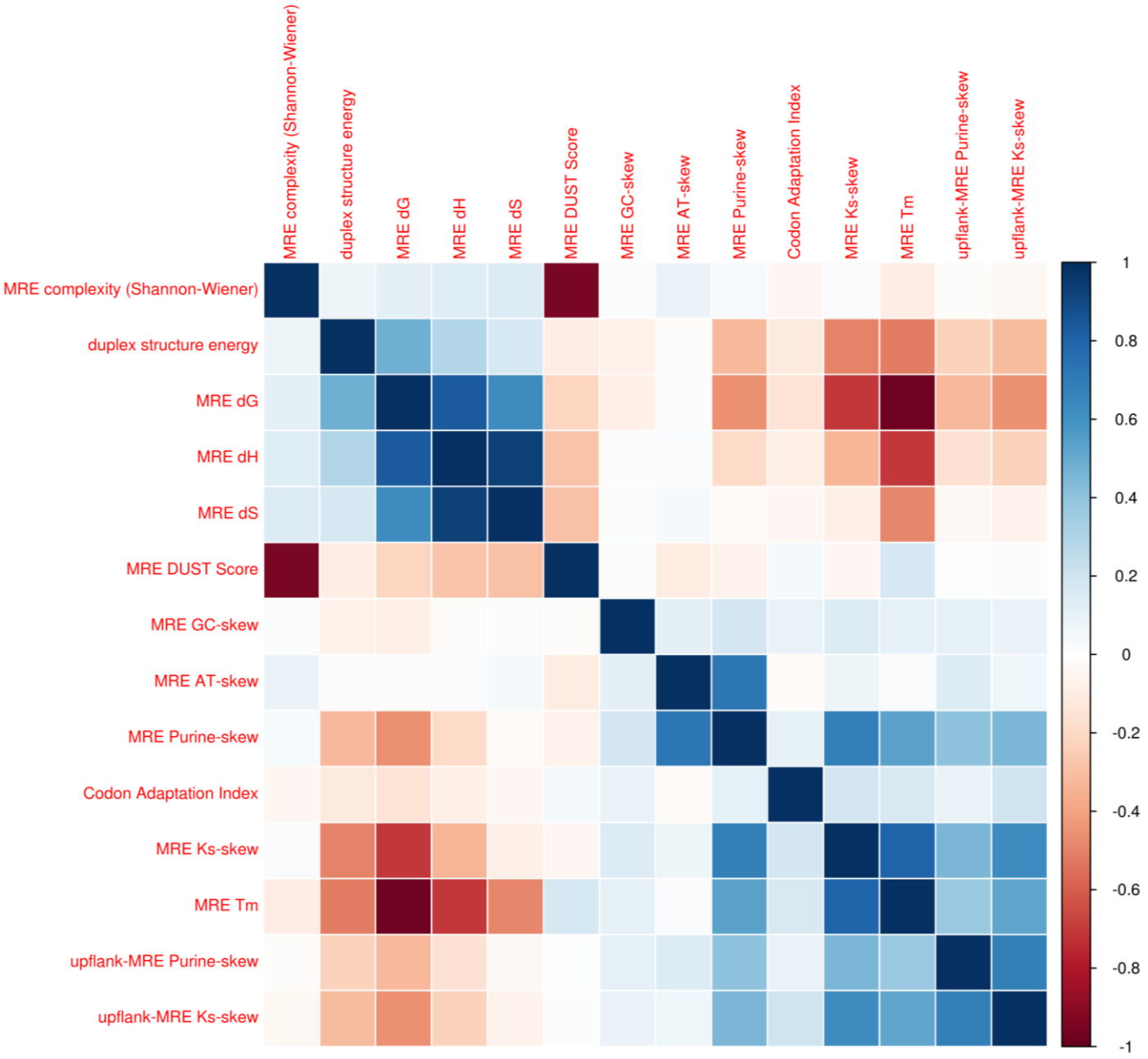

Figure 51: Correlation plot of parameters referring to thermodynamic properties, energy, sequence complexity and content asymmetry of miRNA targeted regions. MRE free energy (dG) and enthalpy (dH) were excluded from the descriptors due to increased (anti-)correlation with MRE melting temperature (Tm) and entropy (dS), respectively. Similarly, only MRE DUST score was retained as a metric of MRE sequence complexity. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level $p<0.05$.

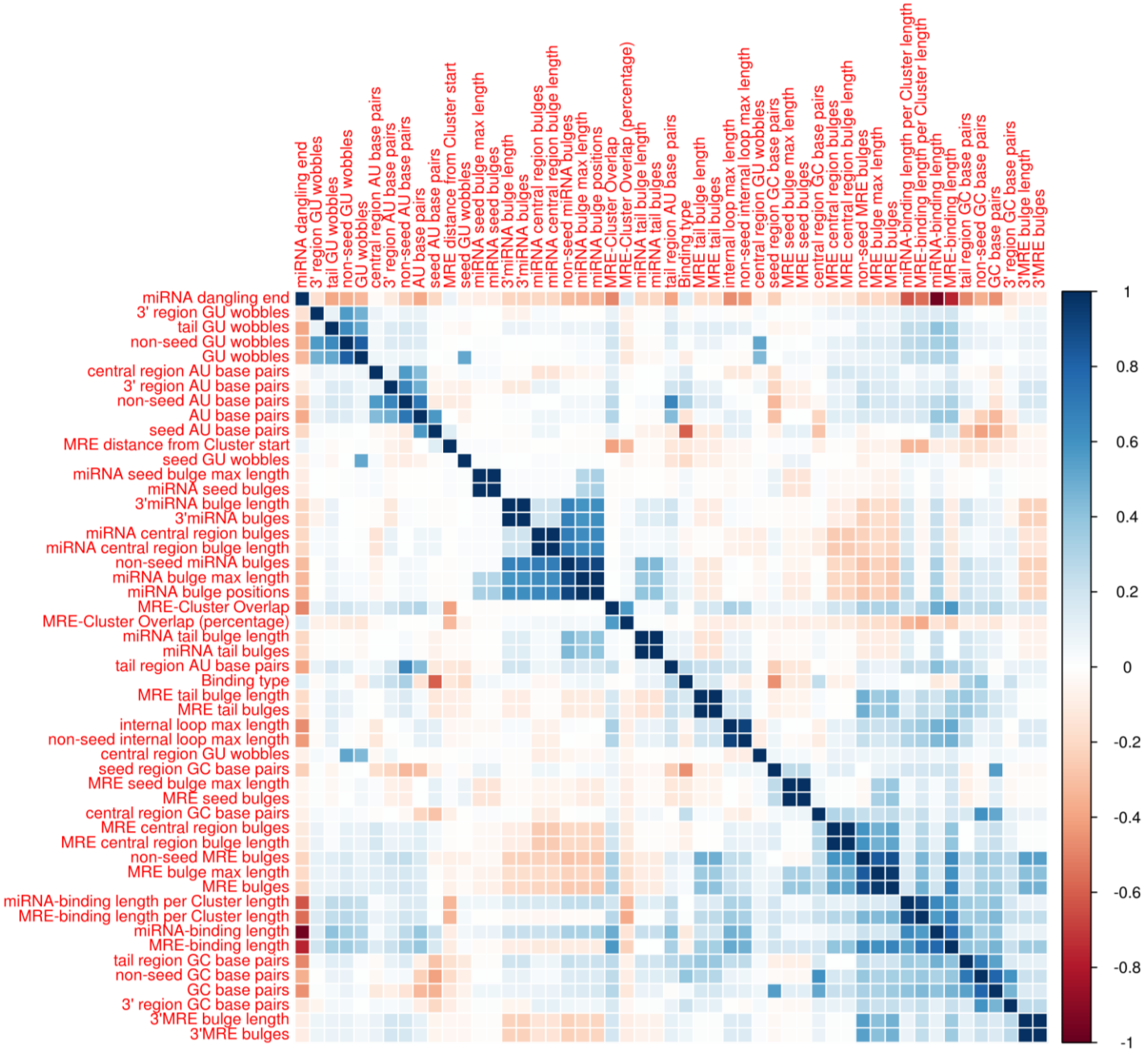

Figure 52: Correlation plot of parameters that characterize the miRNA-target entire duplex structure and relative sub-domains. Highly correlated features describing miRNA or MRE bulges, GU wobbles and AU base pairs were appropriately filtered. miRNA binding length appeared to be highly anti-correlated with 'miRNA dangling end' and therefore only the first parameter was included in the developed learning model. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level $p < 0.05$.

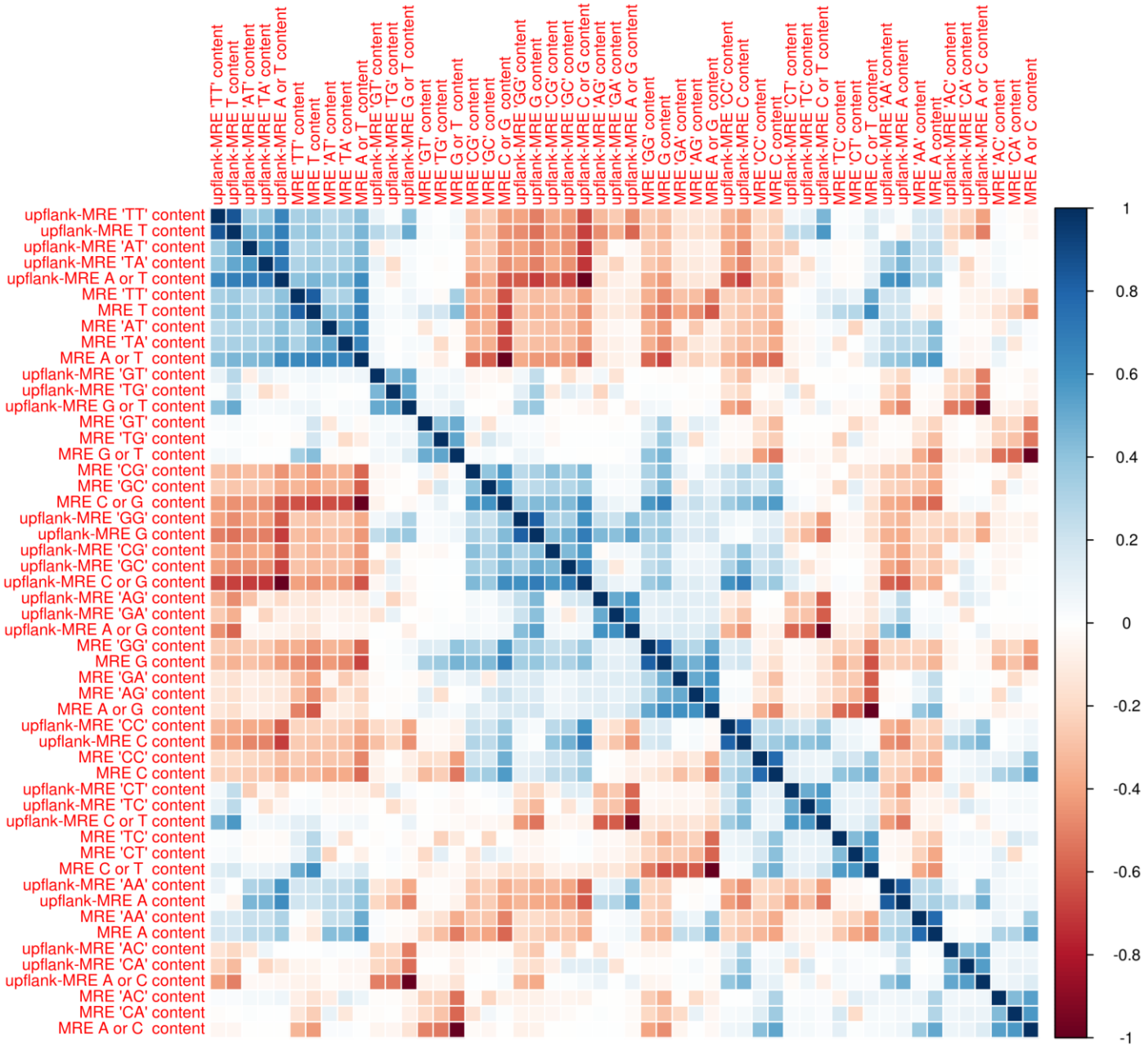

Figure 53: Correlation matrix of content descriptors assigned to the overlapping, upstream and downstream regions of the miRNA binding site. This group of features embodies many highly (anti)correlated single/di-nucleotide composition descriptors, which were appropriately filtered. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level $p < 0.05$.

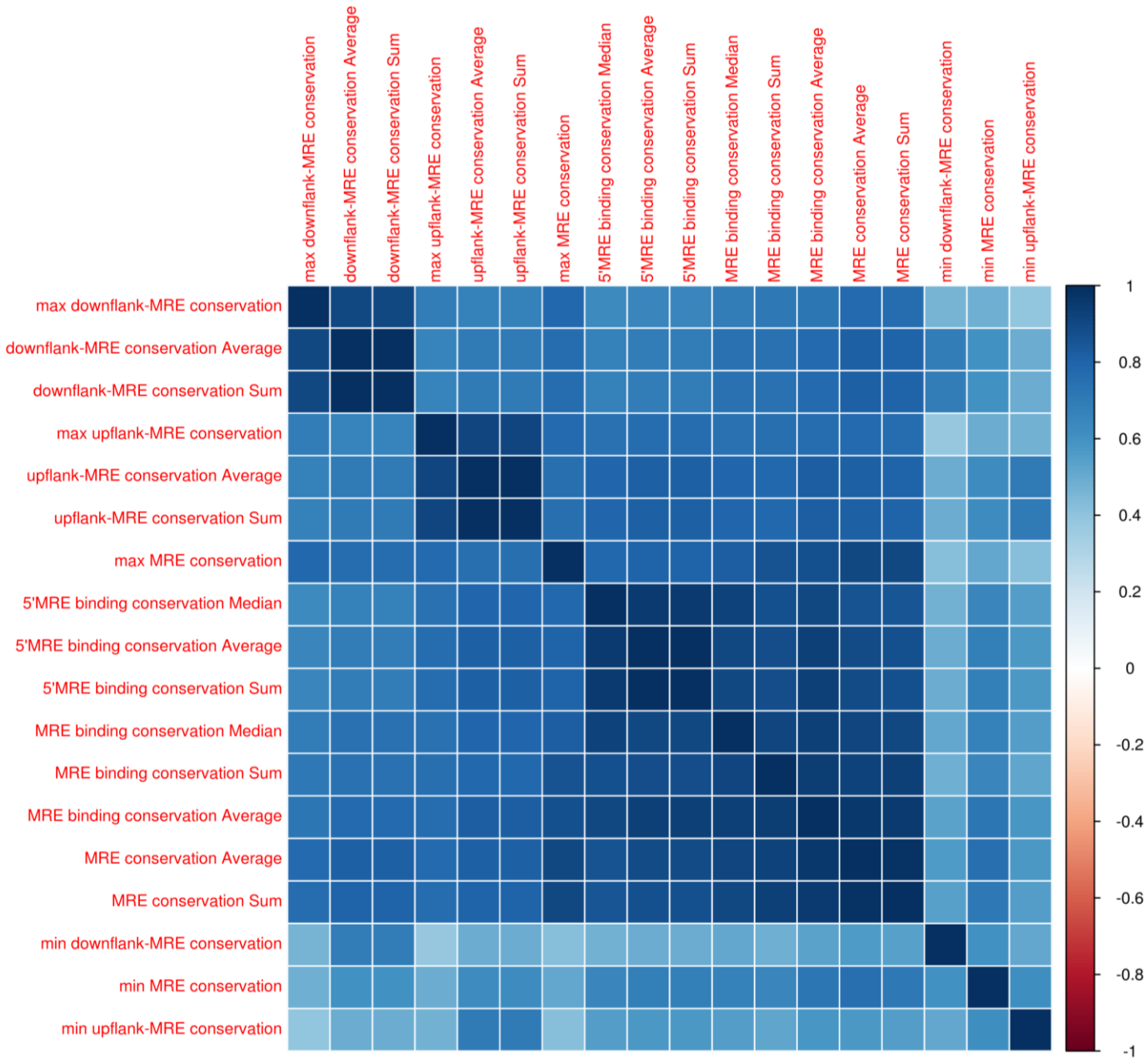

Figure 54: Correlation matrix of conservation features calculated for the respective MRE, upflank-MRE, downflank-MRE regions. Conservation parameters corresponding to max or sum of phastCons pre-computed values presented increased correlation coefficients (>0.9) with relative average scores in MRE regions. The highly correlated features were eliminated. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level $p < 0.05$.

Figure 55: Correlation matrix comprising features for the miRNA-target duplex and miRNA/MRE sub-domains. Base composition descriptors (A, T, G, C) of the (un)paired nucleotides are also included in the plot. Highly correlated parameters including "miRNA mismatches", "miRNA seed" and "miRNA tail" were removed. Possible correlations were estimated by calculating the non-parametric Spearman's rho coefficient using two-sided tests with a significance level p< 0.05.

3.7.3 Base Classifier Models

This section describes the performance of base classifier models in the proposed learning framework for CLIP-Seq-guided miRNA-target identification. The implemented 6 base models ("Region features", "MRE general", "Binding Vector", "miRNA-target duplex", "Base pairing", "Matches per miRNA/MRE domain"),

comprise different sets from uncorrelated parameters. The composition of the feature vector incorporated in each base classifier was optimized against a considerable number of candidate vectors. These models adopt a Random Forest learning approach and are included in the first layer of positive/negative instance classification. Every base classifier assigns a probability score in candidate MREs reflecting its potency of being a true binding site.

3.7.3.1 "Region features" Classifier

The "Region features" base classifier incorporates 55 distinct features, including CLIP-Seq-derived features such as expression, substitution frequencies and distances from the MRE start; content descriptors assigned to the overlapping, upstream and downstream MRE region; conservation, sequence energy, complexity, content asymmetry, and biases of codon usage. These parameters are utilized to characterize the MRE and proximal regions profile. The first top ranked descriptors as specified by the Random forest model are presented in Table 19. The highest importance is assigned to 'MRE RPKM', 'T-to-C substitutions' and 'min MRE distance - sum Substitution Ratio' (i.e. aggregate ratio of substitutions located in minimum distance from the MRE start). This model achieves the best performance among the 6 classifiers in the first layer of positive/negative instances classification. The predictive model exhibited 87.3% sensitivity and 72.2% specificity (AUC 0.862) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio) (Figure 56).

	Importance
MRE Rpkm	2271.286
T-to-C substitutions	1986.001
min MRE distance - sum Substitution Ratio	1127.73
sum of Substitutions Ratio - min MRE distance	613.1598
MRE G content	484.2459
MRE dS	334.596
MRE Overlapping Reads	282.5212
MRE conservation Average	198.6979
Cluster length	198.4758
upflank-MRE conservation Average	179.4623
MRE Rpkm per Cluster Rpkm	161.306
downflank-MRE conservation Average	144.5406
Codon Adaptation Index	140.1478
MRE Tm	122.0494
MRE GC-skew	91.25585
MRE DUST Score	84.03686
MRE A or G content	74.82929
MRE AT-skew	72.38096
MRE G or T content	71.97509
upflank-MRE G content	68.69797

Table 19: The first top 20 ranked descriptors as specified by the "Region features" classifier that adopts an RF learning model. Importance values are provided in decreasing order and signify each parameter's contribution to the classification process.

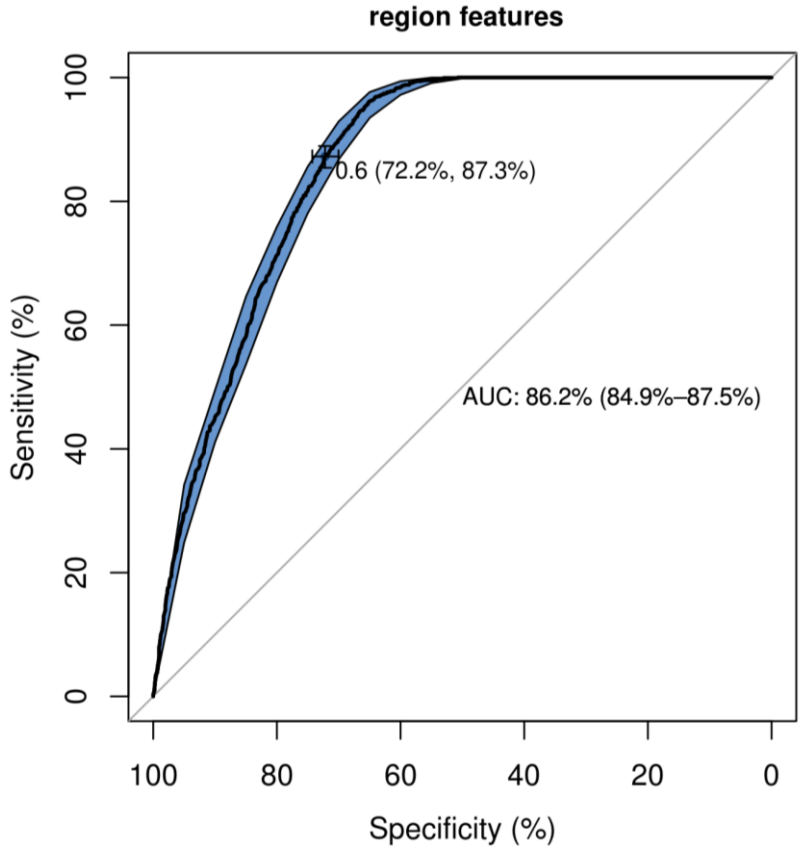

Figure 56: ROC curve of the "Region features" Random Forest model for the classification of positive/negative MREs. The predictive model comprised 55 distinct parameters and exhibited 87.3% sensitivity and 72.2% specificity (AUC 0.862) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.3.2 "Base pairing" Classifier

The "Base pairing" classifier encompasses base composition descriptors (A, T, G, C) of the (un)paired miRNA nucleotides. This predictive model comprised 8 distinct parameters and exhibited 72.1% sensitivity and 56.3% specificity (AUC 0.691) (Figure 57). The importance of the incorporated variables, as estimated by the RF classifier, is shown in Table 20. The highest importance is assigned to the parameters describing matched nucleotides for the miRNA.

	Importance
miRNA C-matches	1527.883
miRNA A-matches	1394.119
miRNA G-matches	1114.7722
miRNA T-matches	1052.1875
miRNA unpaired-T	952.5725
miRNA unpaired-A	944.2197
miRNA unpaired-C	881.6211
miRNA unpaired-G	836.6257

Table 20: "Base pairing" classifier variable importance, as estimated by the RF model. Importance scores are provided in decreasing order and signify each parameter's contribution to the classification process.

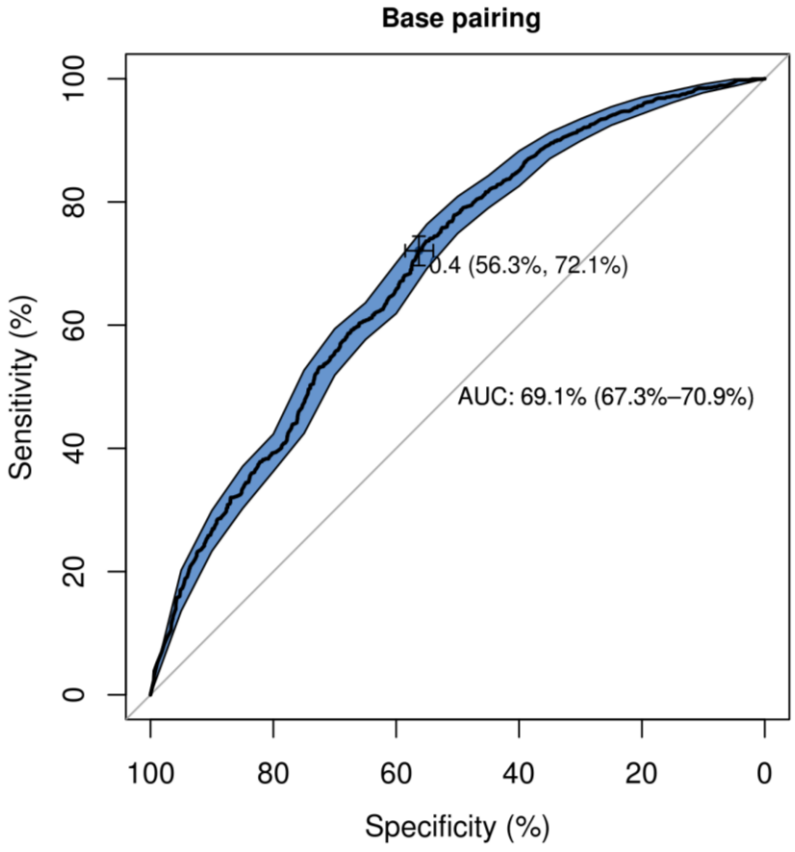

Figure 57: ROC curve of the "Base pairing" Random Forest model for the classification of positive/negative miRNA binding sites. The predictive model comprising 8 distinct parameters exhibited 72.1% sensitivity and 56.3% specificity (AUC 0.691) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.3.3 "MRE general" Classifier

The "MRE general" classifier includes miRNA binding site-related descriptors such as, MRE-cluster overlap, conservation of the most 5' MRE binding nucleotides and all MRE binding nucleotides, MRE location within the cluster, MRE binding type, and variables describing the asymmetry of the duplex matched nucleotides. This predictive model

comprises 8 parameters and presents 74.5% sensitivity and 77.1% specificity (AUC 0.832) (Figure 58).The importance of the included variables, as estimated by the RF classifier is shown in Table 21. The highest importance is assigned to the parameter describing the conservation level of the paired MRE bases.

	Importance
MRE binding conservation Average	2129.3759
MRE matches Ks-skew	1732.8131
Binding type	1515.0408
MRE-binding length per Cluster length	1357.1499
MRE matches Purine-skew	1273.4089
MRE distance from Cluster start	1043.9392
MRE-Cluster Overlap	908.894
MRE-Cluster Overlap (percentage)	748.9639

Table 21: Variable importance scores as estimated by the 'MRE general' classifier that adopts an RF learning model. Importance values are provided in decreasing order and signify each parameter's contribution to the classification process.

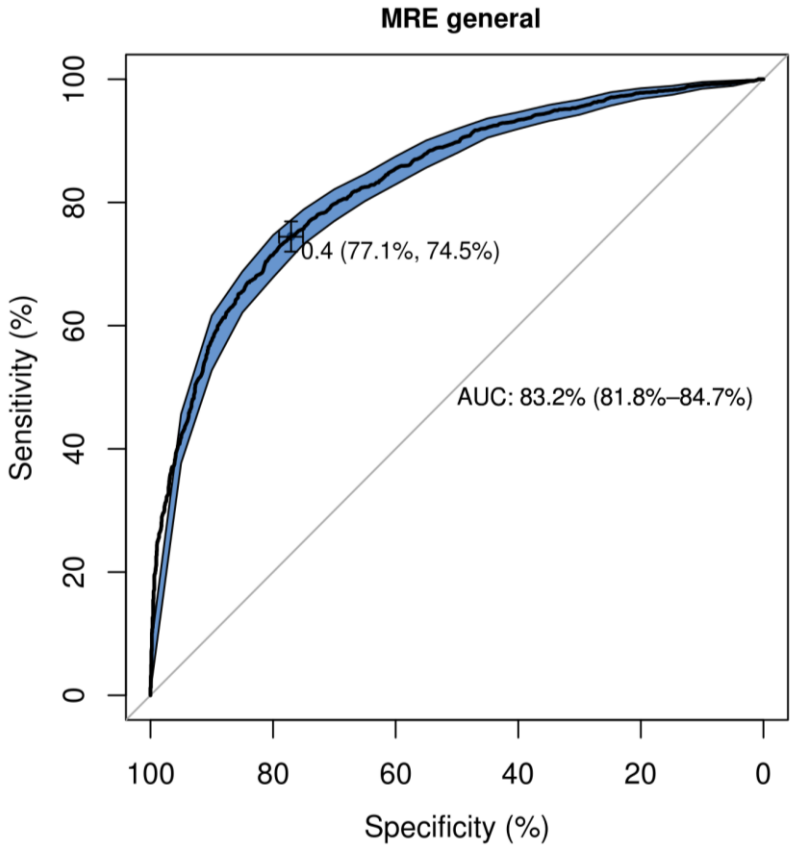

Figure 58: ROC curve of the 'MRE general' Random Forest model for the classification of positive/negative miRNA binding sites. The predictive model comprising 8 distinct parameters presented 74.5% sensitivity and 77.1% specificity (AUC 0.832) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.3.4 "Binding Vector" Classifier

"Binding Vector" classifier comprises 14 distinct descriptors associated with the base pairing per miRNA/MRE position. This predictive model presented 66.4% sensitivity and 80.7% specificity (AUC 0.788) when tested against the independent test set (Figure 59).

	Importance
MRE binding position 2	970.5826
MRE binding position 3	297.8217
MRE binding position 6	219.5371
MRE binding position 7	199.0366
miRNA unpaired position 7	196.3089
miRNA unpaired position 6	187.8222
miRNA unpaired position 5	187.3171
MRE binding position 4	184.3857
MRE binding position 5	168.9264
miRNA unpaired position 8	143.2214
MRE binding position 18	137.2115
MRE binding position 17	136.9804
MRE binding position 11	130.47
MRE binding position 10	127.7633

Table 22: "Binding Vector" classifier variable importance, as estimated by the RF model. Importance scores signify each parameter's contribution to the classification process.

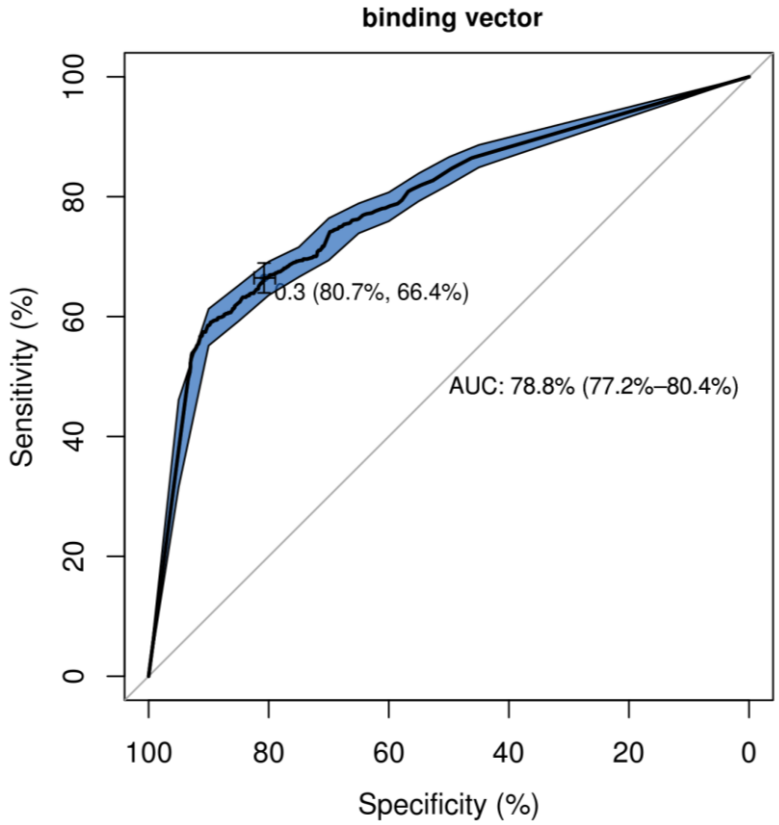

Figure 59: ROC curve of the "Binding Vector" Random Forest model for the classification of positive/negative miRNA binding sites. The predictive model comprising 14 distinct parameters

exhibited 66.4% sensitivity and 80.7% specificity (AUC 0.788) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.3.5 "Matches per miRNA/target domain" Classifier

The "Matches per miRNA/MRE domain" classifier contains 11 parameters that describe the matches in the miRNA-target structure and in MRE/miRNA relevant sub-domains. The importance of the included variables, as estimated by the RF classifier is shown in Table 23. This model exhibited 70.8% sensitivity and 75.5% specificity (AUC 0.793) Figure 60).

	Importance
Binding type	990.8313
consecutive matches per total matches	918.3636
MRE central region matches	767.9645
consecutive seed-matches	763.5292
non-seed consecutive unpaired bases	719.4423
3'MRE matches	708.5513
seed matches per total matches	674.143
max consecutive matches	667.3362
MRE seed matches	657.0682
non-seed consecutive matches per total matches	572.1283
miRNA matches	540.3774

Table 23: "Matches per miRNA/MRE domain" classifier variable importance, as estimated by the RF model. Importance scores are provided in decreasing order.

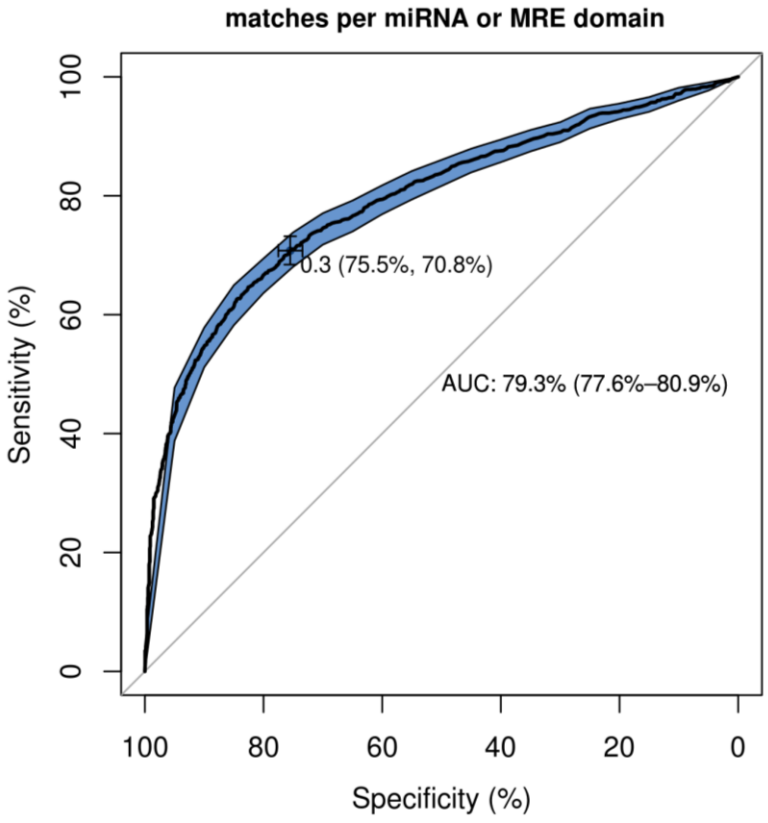

Figure 60: ROC curve of the "Matches per miRNA or MRE domain" Random Forest model for the classification of positive/negative miRNA binding sites. The predictive model comprising 11 distinct

parameters exhibited 70.8% sensitivity and 75.5% specificity (AUC 0.793) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.3.6 "miRNA-target duplex" Classifier

The "miRNA-target duplex" classifier comprises 13 parameters that describe the duplex structure energy, miRNA or MRE bulges, GU wobbles and GC/AU base pairing features for the specified miRNA and/or target and relevant sub-domains. This model presented 70.8% sensitivity and 75.5% specificity (AUC 0.793) (Figure 60). The ranking of included parameters based on the implemented RF classifier is provided in Table 24.

	Importance
duplex structure energy	1491.9206
AU base pairs	1013.4779
GC base pairs	873.9221
non-seed GC base pairs	857.5021
seed AU base pairs	794.0966
MRE-binding length	727.615
GU wobbles	698.3376
non-seed AU base pairs	675.0161
miRNA-binding length	674.5717
MRE bulges	668.571
internal loop max length	545.639
central region GC base pairs	473.0839
tail GU wobbles	329.2324

Table 24: Variable importance scores as estimated by the "miRNA-target duplex" classifier that adopts an RF learning model. Importance values are provided in decreasing order and signify each parameter's contribution to the classification process.

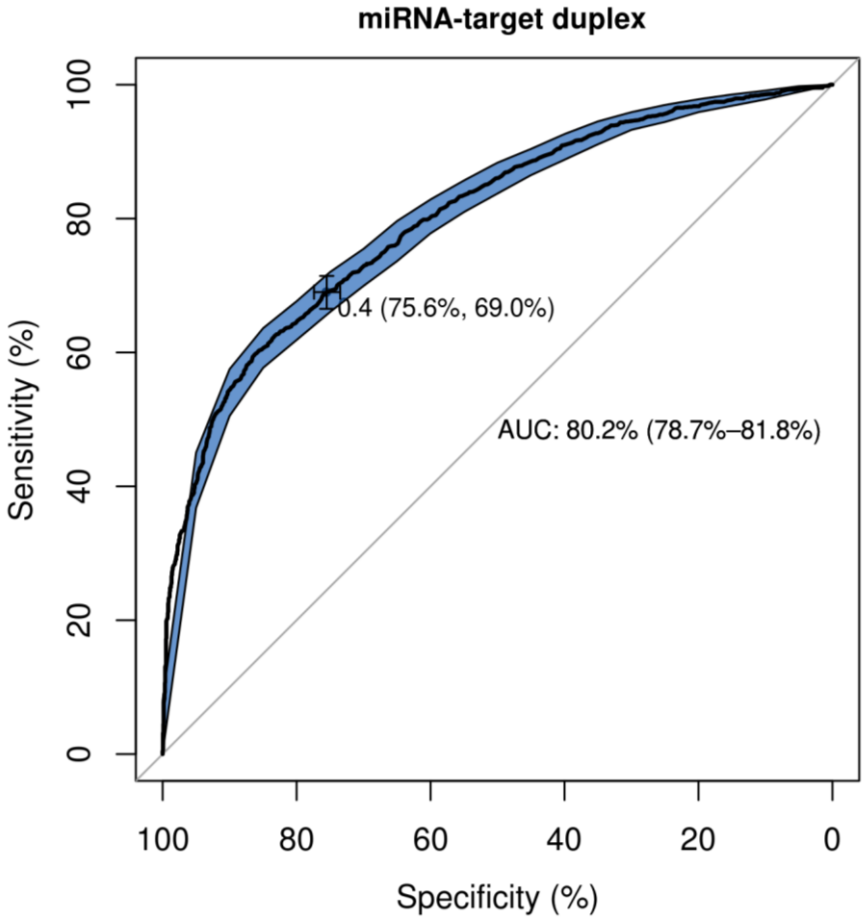

Figure 61: ROC curve of "miRNA-target duplex" Random Forest model for the classification of positive/negative miRNA binding sites. The predictive model comprising 13 distinct parameters exhibited 69% sensitivity and 75.6% specificity (AUC 0.802) in the control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.4 Meta-classifier

Each base classifier presented in the previous section generates a probability score that is subsequently forwarded to the second layer of classification. The 6 distinct probability scores are aggregated in a meta-classifier scoring model that derives the miRNA binding affinity within the cluster regions. The use of a GBM model as the meta-classifier outperforms every other tested algorithm including RFs and SVMs. The GBM classifier achieved 81.6% sensitivity and 80.6% specificity (AUC 0.908) (Figure 62); RF presented 83.8% sensitivity and 76.5% specificity (AUC 0.897) (Figure 63), while the SVM exhibited 86.5% sensitivity and 72.7% specificity (AUC 0.859).

All candidate meta-classifiers were evaluated in a control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

Base classifiers	GBM variable importance	RF variable importance
region features	3312.5	1987.5
Base pairing	163.03	820.5
miRNA-target duplex	33.63	383.7
binding vector_features	22.2	229
matches per miRNA or MRE domain	21.98	301.8
MRE general	16	453.1

Table 25: Variable importance scores as estimated by the meta-classifier that adopts a GBM or an RF learning model respectively. The included parameters in these classifiers correspond to the output of the base classifiers (first layer of classification). Importance values are provided in decreasing order and signify each parameter's contribution to the classification process. The highest importance is assigned by both models to the "region features" classifier probability scores.

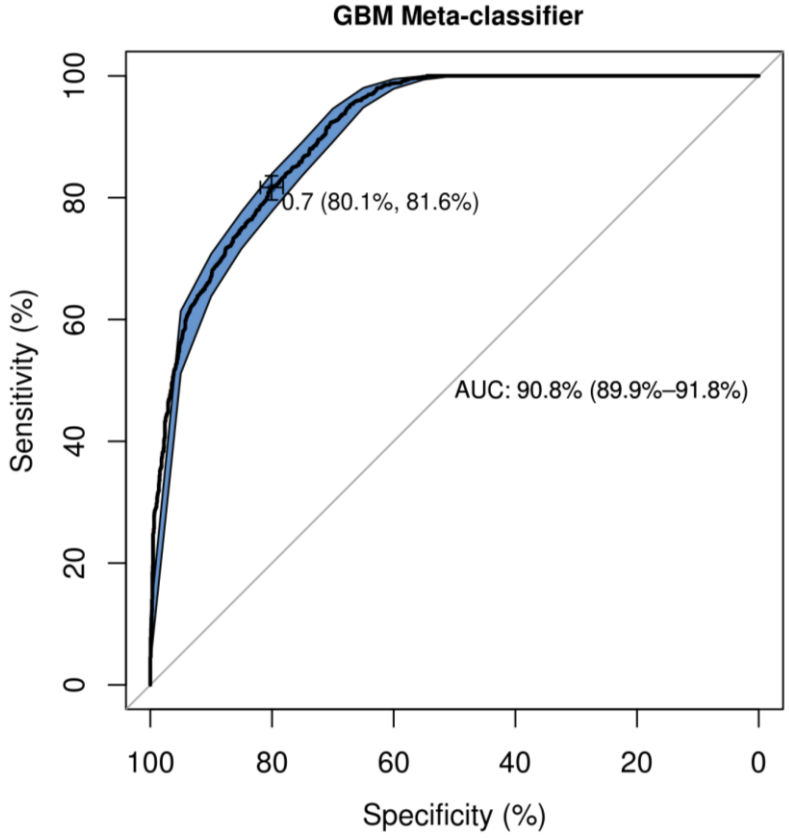

Figure 62: ROC curve of "GBM meta-classifier" model for the classification of positive/negative miRNA binding sites. This learning approach achieved the highest performance, presenting 81.6% sensitivity and 80.6% specificity (AUC 0.908). GBM was evaluated against a control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

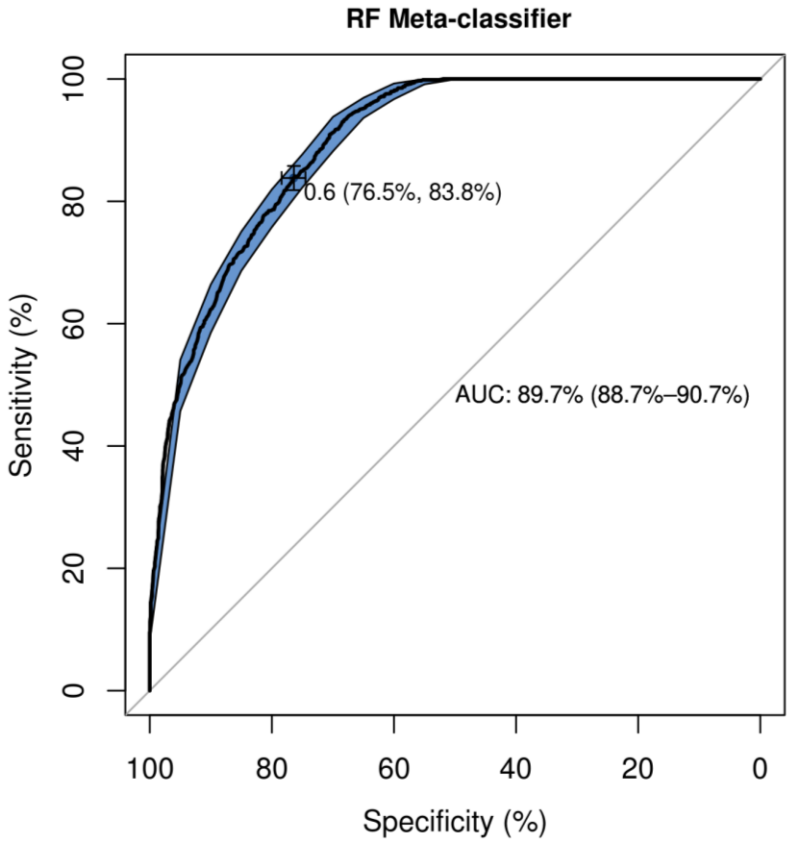

Figure 63: ROC curve of "RF meta-classifier" model for the classification of positive/negative miRNA binding sites. This model exhibited 83.8% sensitivity and 76.5% specificity (AUC 0.897) when tested against a control set (test set) of approximately 3000 instances (1:1 positive-negative ratio).

3.7.5 Evaluation of the Novel Learning framework against other state-of-the-art implementations.

In order to evaluate the novel learning framework for CLIP-Seq guided miRNA-target identification, it was compared against different implementations such as, microMUMMIE (75), MIRZA (72), PARMA (74) and the TarBase/LncBase analysis algorithm. The assessment of their performance was performed against a control set of (in)direct miRNA-target interactions in an embryonic kidney cell line (HEK293) supported by low and high-throughput methodologies. More precisely, the validation set comprises 1,365 positive interactions including 138 highly expressed miRNAs in HEK293 cells. In order to obtain a complete list of interactions for all the tested implementations, each algorithm has been executed on a comprehensive set of PAR-CLIP HEK-293 libraries. The proposed settings for each algorithm were retrieved from the relevant publications, in order to attain high quality results for the conducted comparisons.

A major concern with CLIP-Seq algorithms, excepting their ability to correctly identify experimentally verified miRNA binding sites, is the number of provided predictions per AGO-peak region. Therefore, in the presented evaluation (Figure 64, Figure 65) the number of correctly predicted MRE regions is plotted versus total predictions for different prediction score thresholds.

Moreover, microMUMMIE and especially PARma implementations do not cover the whole spectrum of miRNA binding types. Therefore, an extra evaluation test was realized that included only positive miRNA interactions with canonical seed matches, in order to render the obtained results as comparable as possible, (Figure 64b, Figure 65b).

A primary evaluation was implemented to demonstrate the performance of the novel algorithm compared to the CLIP-Seq guided analysis adopted by TarBase/LncBase. The results depict that the new approach not only significantly outperforms the former implementation in terms of accuracy but also manages an impressive increase in sensitivity, predicting almost twice as many validated sites. Most of these sites were not detected by any other algorithm (Figure 64).

The novel algorithm also achieved the best performance in any metric when juxtaposed against other state-of-the-art implementations (Figure 65). This evaluation was generated separately for canonical and non-canonical miRNA positive interactions.

All the leading algorithms proved to be far from perfect and suffered from a low ability to identify a high percentage of true miRNA-target interactions with a high cost in the total predictions. The novel implementation that has been trained on an unprecedented collection of high quality low/high-throughput experiments, breaks this barrier by providing true positive predictions for more than 80% of the included miRNAs.

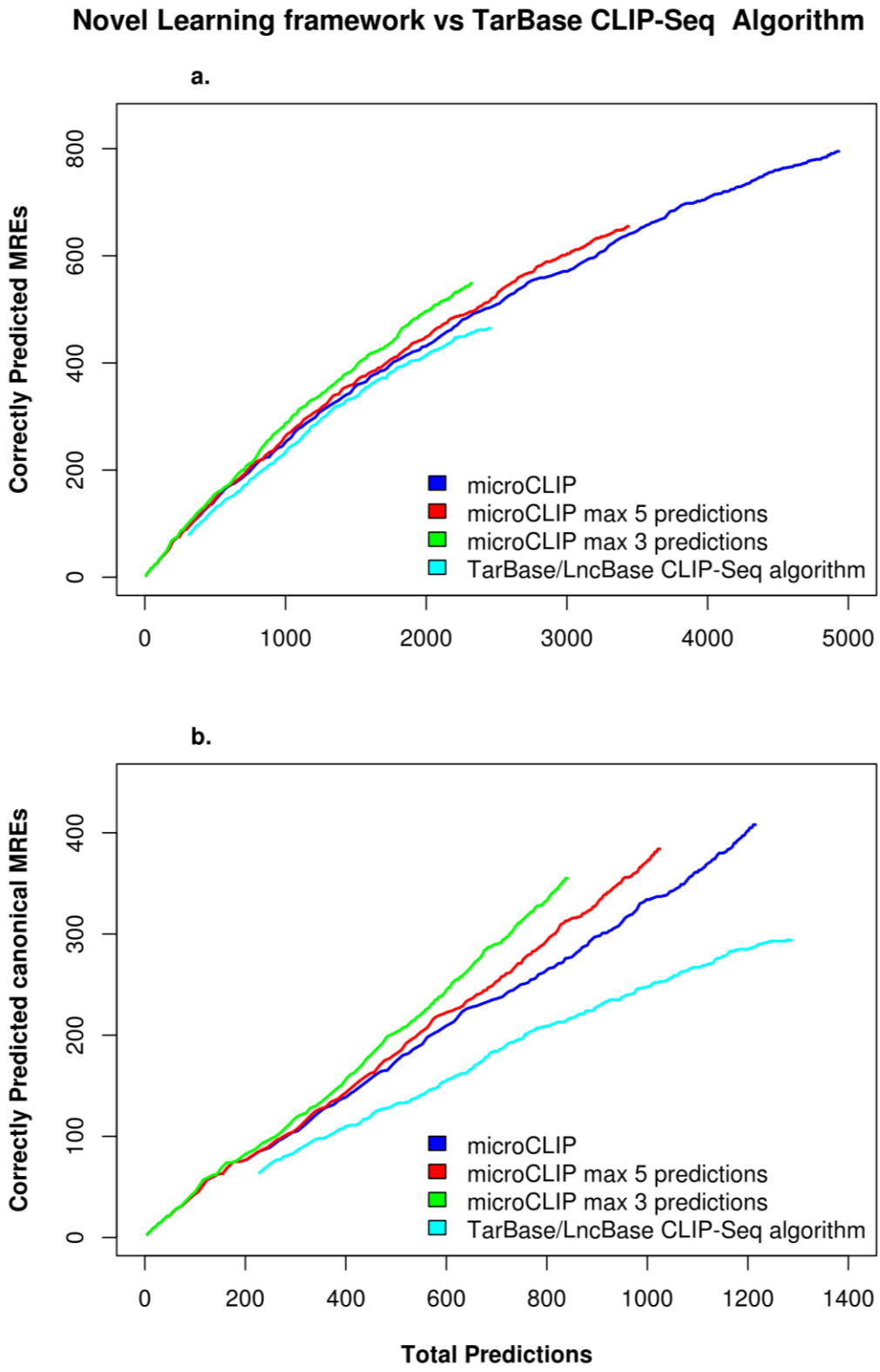

Figure 64: Evaluation of the novel AGO-CLIP learning framework (microCLIP) against the TarBase/LncBase adopted algorithm. The number of correctly predicted miRNA binding sites for each implementation is plotted versus the total retrieved predictions for different interaction score thresholds. The performance of the novel algorithm is additionally provided for the top 5 and top 3 predictions per cluster region. The utilized validation set comprised 1,072 positive miRNA interactions derived from direct and indirect experimental methodologies (a). The new algorithmic approach significantly outperforms the former implementation and manages a 2-fold increase in the

correct identification of experimentally verified miRNA binding sites. An extra evaluation was realized including only positive miRNA interactions (~500) with canonical seed match (b). The novel algorithm managed to identify ~90% of the positive canonical miRNA interactions, a ~30% increase compared to TarBase/LncBase CLIP-Seq implementation and provides one valid miRNA canonical binding site in approximately every 2 predicted targets.

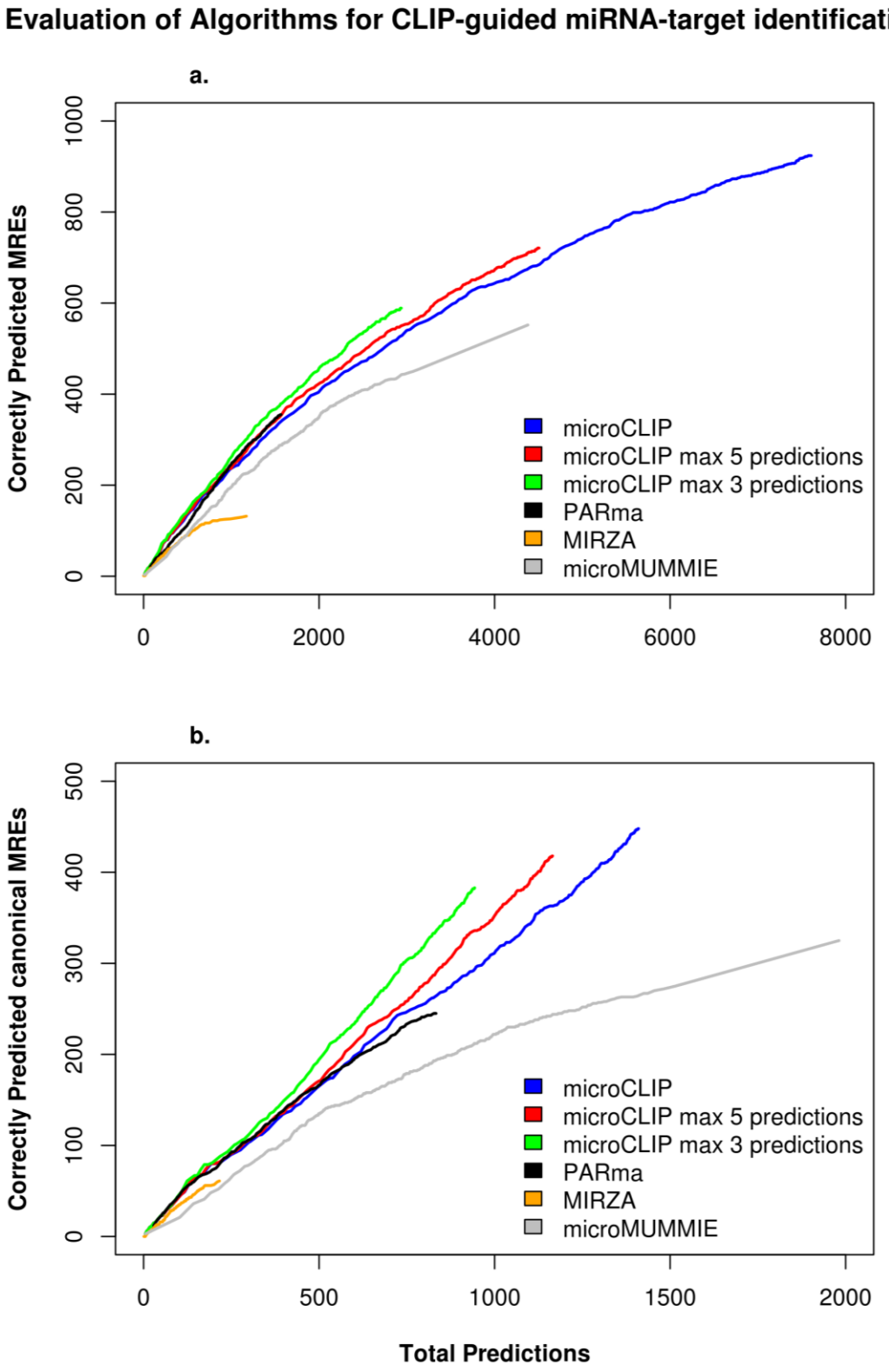

Figure 65: Evaluation of the novel AGO-CLIP learning framework (microCLIP) against the leading implementations of PARma, MIRZA and microMUMMIE. The number of correctly predicted miRNA binding sites for each implementation is plotted versus the total retrieved predictions for different

interaction score thresholds. The performance of the novel algorithm is additionally provided for the top 5 and top 3 predictions per cluster region. The utilized validation set comprised 1,365 positive miRNA interactions derived from direct and indirect experimental methodologies (a). The results demonstrate that the novel AGO-CLIP implementation has a significant greater ability to discriminate correct interactions compared to other approaches. An extra evaluation was realized including only positive miRNA interactions (~500) with canonical seed matches (b). The novel algorithm managed to identify ~90% of the positive canonical miRNA interactions.

4. Conclusion

One of the most important processes in miRNA research is their target detection. Identification of miRNA-gene interactions can be performed with either computational approaches or experimental methodologies.

During the thesis course, DIANA-microT v5.0 (54), the new version of the microT server, has been significantly enhanced with an improved target prediction algorithm, DIANA-microT-CDS (38). microT-CDS is the only algorithm available online, specifically designed to identify miRNA targets both in 3' untranslated region (3'UTR) and in coding sequences (CDS). The web server was also completely redesigned, in order to host a series of sophisticated workflows, enabling users to perform advanced multi-step functional miRNA analyses. DIANA-microT web server v5.0 additionally supports a complete integration with the Taverna Workflow Management System (WMS) (143), using an in-house developed DIANA-Taverna Plug-in. This plugin provides ready-made modules for miRNA target prediction and functional analysis, which can be used to form advanced high throughput analysis pipelines.

Computational methodologies unambiguously provide a valuable resource for miRNA oriented studies. However, even the most advanced implementations include an increased number of false positive interactions and do not allow the derivation of functional downstream analyses. *In silico* implementations can be further improved if coupled with technological breakthroughs of sequencing experiments.

Numerous wet lab methodologies have been developed, enabling the validation of predicted miRNA interactions or the high-throughput screening and identification of novel miRNA targets (32). Moreover, during the past few years, NGS methodologies have revolutionized almost every aspect of biological research. Novel NGS-based high-throughput miRNA target identification techniques have enabled the identification of thousands of interactions present in specific cell types or experimental conditions.

Despite the contribution of both experimental methodologies and computational approaches, a large part of the miRNA targets, even for the well-studied organisms such as mouse and human, remains unexplored. The wealth of information provided by experimental methodologies remains fragmented and hidden in thousands of manuscripts, supplemental materials and raw sequencing datasets.

Accurate cataloguing of miRNA targets is crucial to the understanding of their function. However, the complex network of miRNA-lncRNA-mRNA regulatory machinery is difficult to be determined by exploring individual pairs of interactions and relies on the analysis of extensive NGS datasets. By analyzing more than 250 miRNA-related NGS datasets (e.g. 150 CLIP-Seq, CLASH, microarrays, Degradome-Seq) and extracting interactions from hundreds of meticulously curated articles, DIANA-TarBase v7.0 is the first database to provide an unprecedented amount of experimentally supported miRNA-mRNA interactions in many different cell types and tissues. DIANA-TarBase v7.0 breaks the barrier of 300,000 entries indexed by relevant repositories, providing

more than half a million interactions in 24 species, 9-250 times more than any other manually curated database. These interactions can enforce or even at cases substitute *in silico* predicted interactions.

LncRNA functions still remain widely uncovered, while others are currently under debate. The recently introduced sponge/decoy role of lncRNAs has been characterized for a few transcripts in specific tissue and/or disease conditions. LncBase v2 provides an extensive compendium of miRNA-lncRNA *in silico* inferred and experimentally supported interactions covering a wide range of cell types and tissues for human and mouse. The analysis of extensive sequencing data unveiled thousands of miRNA-lncRNA interactions, including lncRNAs harboring multiple miRNA binding sites and a set of approximately 400 unique viral-miRNA-lncRNA interacting pairs in virus infected cells. Spatial classification of miRNA-targeted regions in CLIP-Seq experiments revealed similar percentages of targeted lncRNA transcripts across different cell types. A considerable amount of MREs residing on lncRNA transcript regions were highly conserved presenting stronger evolutionary pressure than their background regions, while miRNA sites located in lncRNA intronic regions presented accelerated evolutionary rates compared to those in lncRNA exons. AGO-CLIP-Seq cognate cell lines were densely grouped by targeted lncRNAs, possibly indicating a tissue specific miRNA-lncRNA regulation mechanism.

During the thesis course, an in house algorithm was implemented in order to analyze CLIP-Seq data on different cell types and tissues for mouse and human species. It was thoroughly tested against state-of-the-art implementations and was utilized for TarBase and LncBase updates.

The continuous archiving of experimental data from low and high-throughput methodologies, along with the extensive evaluation of the available AGO-CLIP-Seq analysis programs, revealed that there was room for further improvement and optimization of the relevant algorithms in order to attain increased accuracy. State-of-the-art CLIP-Seq target identification implementations currently manage to identify approximately half of the experimentally validated binding sites. To this end, a novel algorithm was developed for CLIP-Seq data analysis. The algorithm was trained and extensively tested on a comprehensive collection of accurate positive and negative miRNA-target interactions from low-yield and high-throughput experimental data sources. The novel algorithm was evaluated against all leading implementations, including CLIP-Seq guided analysis adopted by TarBase/LncBase. Former algorithms proved to be far from perfect and suffered from a low ability to identify a high percentage of positive miRNA-target sites. The results depict that the new approach not only significantly outperforms other implementations in terms of accuracy but also manages to increase sensitivity, predicting sites that were not detected by any other algorithm.

The novel algorithm will enable the accurate identification of miRNA coding and non-coding target repertoire, which is crucial to the detection of competing endogenous

interactions. This information can be utilized for multiple exploratory studies and in-depth analyses for the creation of tissue specific lncRNA-miRNA-mRNA/TF regulatory networks. Moreover, functional interpretation of the interaction networks can boost the understanding of unexplored regulatory mechanisms and the elucidation of key players in different biological processes.

5. Thesis Publications

During the course of the thesis, the candidate participated in 8 scientific studies, involving computational approaches for determining the activity of the non-coding transcripts and in four of them the candidate is first author. The studies are published in international journals of high impact factor and total citations received so far are 310. The publications achieved are presented in chronological order.

1. Vlachos IS, Vergoulis T, **Paraskevopoulou MD**, Lykokanellos F, Georgakilas G, Georgiou P, Chatzopoulos S, Karagkouni D, Christodoulou F, Dalamagas T, Hatzigeorgiou A.G. (2016) DIANA-mirExTra v2.0: Uncovering microRNAs and transcription factors with crucial roles in NGS expression data. Nucleic Acids Res. (**9.112 Impact Factor**)
2. **Paraskevopoulou MD** and Hatzigeorgiou AG. Analyzing MiRNA-LncRNA Interactions. Methods Mol Biol. 2016 (2 citations)
3. **Paraskevopoulou MD**, Vlachos IS, Karagkouni D, Georgakilas G, Kanellos I, Vergoulis T, Zagganas K, Tsanakas P, Floros F, Dalamagas T, Hatzigeorgiou AG. (2015) DIANA-DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic Acids Research (**9.112 Impact Factor**) (2 citations)
4. Georgakilas G and Vlachos IS, Zagganas K, Vergoulis T, **Paraskevopoulou MD**, Kanellos I, Tsanakas P, Dellis D, Feygas A, Dalamagas T, Hatzigeorgiou AG. (2016) DIANA-miRGen v3.0: extensive characterization of microRNA promoters and their regulation. Nucleic Acids Research (**9.112 Impact Factor**) (3 citations)
5. Vlachos IS, Zagganas K, **Paraskevopoulou MD**, Georgakilas G, Karagkouni D, Vergoulis T, Dalamagas T, Hatzigeorgiou AG. (2015) DIANA-miRPath v3.0: Deciphering microRNA function with experimental support. Nucleic Acids Research (**9.112 Impact Factor**) (28 citations)
6. Georgakilas G, Vlachos IS, **Paraskevopoulou MD**, Yang P, Zhang Y, Economides AN, Hatzigeorgiou AG. (2014) microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. Nature Communications (**10.7 Impact Factor**) (16 citations)
7. Vlachos IS and **Paraskevopoulou MD**, Karagkouni D, **Georgakilas G**, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, Dalamagas T, Hatzigeorgiou AG. (2014) DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Research (**9.112 Impact Factor**) (**joint first authorship**) (90 citations)
8. **Paraskevopoulou MD** and Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Research (**9.112 Impact Factor**) (**joint first authorship**) (169 citations)

6. ABBREVIATIONS - ACRONYMS

3′-UTR	3′-UnTranslated Region
3Life	Luminescent Identification of Functional Elements in 3'UTRs
5′-UTR	5′-UnTranslated Region
5′ RLM-RACE	Rapid amplification of cDNA ends
AGO	Argonaute
AGO-IP	AGO Immunoprecipitation
ANN	Artificial Neural Networks
AUC	Area Under Curve
BLAST	Basic Local Alignment Search Tool
BLS	Branch-length conservation scores
C. elegans	Caenorhabditis elegans
CDS	Coding Sequence
ceRNA	Competing endogenous RNA
Chip-Seq	Chromatin Immunoprecipitation Sequencing
CLASH	Crosslinking, ligation, and sequencing of hybrids
CLEAR-CLIP	Covalent ligation of endogenous Argonaute-bound RNAs
CLIP-Seq	Cross-linking immunoprecipitation sequencing
D. melanogaster	Drosophila melanogaster
dG	Free energy
DGCR8	DiGeorge syndrome Critical Region 8
dH	Enthalpy
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
DNase-Seq	DNase I hypersensitive sites sequencing
dS	Entropy
EBV	Epstein-Barr virus
ELISA	Enzyme-linked immunosorbent assay
EM	Expectation Maximization
ENCODE	Encyclopedia of DNA Elements Consortium
FDR	False Discovery Rate
GAs	Genetic algorithms

GBMs	Gradient Boosting Machines
GEO	Gene Expression Omnibus
GFP	Green Fluorescent Protein
GLM	Generalized Linear Models
H. sapiens	Homo sapiens
H3K4me3	Histone 3 lysine 4 trimethylation
HEK-293	Human Embryonic Kidney Cells
HELA	Human Cervical Cancer Cells
hESC	Human Embryonic stem Cells
HITS-CLIP	High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
HMM	hidden Markov model
Huh7.5	Hepatocarcinoma cells
ICA	Independent component analysis
iCLIP	Individual-nucleotide resolution UV crosslinking and immunoprecipitation
ID3	Iterative Dichotomiser 3
IMPACT-Seq	Pull-down sequencing of biotin-tagged miRNAs
KEGG	Kyoto Encyclopedia of Genes and Genomes
KSHV	Kaposi's sarcoma-associated herpesvirus
Ks-skew	Keto skew
LDA	Linear Discriminant Analysis
lncRNAs	long non-coding RNAs
M. musculus	Mus musculus
MCF7	Human Mammary Gland Cancer Cells / Michigan Cancer Foundation-7
MDAMB231	Human Mammary Gland Cancer Cells
MeSH	Medical Subject Headings
miRISC	miRNA-induced silencing complex
miRNA	microRNA
miTRAP	miRNA trapping by RNA in vitro affinity purification
ML models	Machine Learning model
MNase	Micrococcal Nuclease
MREs	miRNA Recognition Elements
mRMR	Minimum-redundancy-maximum-relevance

mRNA	messenger RNA
NB	Naïve Bayes
ncRNAs	non-coding RNAs
NGS	Next Generation Sequencing
nt	nucleotide
ORF	Open Reading Frame
PAR-CLIP	Photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation
PARE/ Degradome-Seq	Parallel analysis of RNA ends/ Degradome sequencing
P-bodies	Processing bodies
PCA	Principal component analysis
Pol II/III	RNA polymerase II/III
poly-A	Polyadenylation
pre-miRNA	precursor miRNA
pri-miRNA	primary miRNA
qPCR	Quantitative real-time polymerase chain reaction
RBF	Radial basis function
RBPs	RNA-binding proteins
RF	Random Forest
RISC	RNA-induced silencing complex
RMA	Robust Multi-Array Average
RNA	Ribonucleic Acid
RNase	Ribonuclease
RNA-Seq	RNA sequencing
ROC	Receiver operating characteristic
RPF-Seq	Ribosome profiling sequencing
RPKM	Reads Per Kilobase of transcript per Million mapped reads
rRNA	Ribosomal RNA
RVM	Relevance Vector Machine
SDS–PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SILAC	Stable isotope labeling by amino acids in cell culture
SNPs	Single Nucleotide Polymorphism

SNR	Signal-to-noise ratios
sRNA	Small RNA
sRNA-Seq	Small RNA sequencing
SVM	Support Vector Machine
T/Thy	Thymine
Tm	Melting temperature
tRNA	transfer RNA
url	Uniform Resource Identifier
WMS	Workflow Management System
XML	Extensible Markup Language

7. References

1. **Eddy, S.R. (1999) Noncoding RNA genes. *Curr Opin Genet Dev*, 9, 695-699.**
2. **Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515, 355-364.**
3. **Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57-74.**
4. **Mattick, J.S. (2004) RNA regulation: a new genetics? *Nat Rev Genet*, 5, 316-323.**
5. **Huntzinger, E. and Izaurralde, E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12, 99-110.**
6. **Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75, 843-854.**
7. **Vergoulis, T., Vlachos, I.S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T. and Hatzigeorgiou, A.G. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*, 40, D222-229.**
8. **Griffiths-Jones, S. (2010) miRBase: microRNA sequences and annotation. *Curr Protoc Bioinformatics*, Chapter 12, Unit 12 19 11-10.**
9. **Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75, 843-854.**
10. **Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. and Ruvkun, G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403, 901-906.**
11. **Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408, 86-89.**
12. **Huntzinger, E. and Izaurralde, E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12, 99-110.**
13. **Cai, X., Hagedorn, C.H. and Cullen, B.R. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna*, 10, 1957-1966.**
14. **Borchert, G.M., Lanier, W. and Davidson, B.L. (2006) RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol*, 13, 1097-1101.**
15. **Lee, Y.S. and Dutta, A. (2009) MicroRNAs in cancer. *Annu Rev Pathol*, 4, 199-227.**
16. **Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425, 415-419.**
17. **Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. and Hannon, G.J. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432, 231-235.**
18. **Gregory, R.I., Chendrimada, T.P., Cooch, N. and Shiekhattar, R. (2005) Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell*, 123, 631-640.**
19. **Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H. and Kim, V.N. (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev*, 18, 3016-3027.**
20. **Kim, V.N. (2004) MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends Cell Biol*, 14, 156-159.**
21. **Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T. and Zamore, P.D. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293, 834-838.**
22. **Khvorova, A., Reynolds, A. and Jayasena, S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115, 209-216.**

23. Kim, V.N., Han, J. and Siomi, M.C. (2009) Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol*, 10, 126-139.

24. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, 136, 215-233.

25. Friedman, R.C., Farh, K.K., Burge, C.B. and Bartel, D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19, 92-105.

26. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141, 129-141.

27. Dai, R. and Ahmed, S.A. (2011) MicroRNA, a new paradigm for understanding immunoregulation, inflammation, and autoimmune diseases. *Transl Res*, 157, 163-179.

28. Wang, H. and Peng, D.Q. (2011) New insights into the mechanism of low high-density lipoprotein cholesterol in obesity. *Lipids Health Dis*, 10, 176.

29. Erson, A.E. and Petty, E.M. (2008) MicroRNAs in development and disease. *Clin Genet*, 74, 296-306.

30. Guay, C., Roggli, E., Nesca, V., Jacovetti, C. and Regazzi, R. (2011) Diabetes mellitus, a microRNA-related disease? *Transl Res*, 157, 253-264.

31. Ono, K., Kuwabara, Y. and Han, J. (2011) MicroRNAs and cardiovascular diseases. *Febs J*, 278, 1619-1633.

32. Thomson, D.W., Bracken, C.P. and Goodall, G.J. (2011) Experimental strategies for microRNA target identification. *Nucleic Acids Res*, 39, 6845-6853.

33. Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M. and Hatzigeorgiou, A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25, 3049-3055.

34. Witkos, T.M., Koscianska, E. and Krzyzosiak, W.J. (2011) Practical Aspects of microRNA Target Prediction. *Curr Mol Med*, 11, 93-109.

35. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, 115, 787-798.

36. Bartel, D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136, 215-233.

37. Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M. and Hatzigeorgiou, A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25, 3049-3055.

38. Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I. and Hatzigeorgiou, A.G. (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics*, 28, 771-776.

39. Jeggari, A., Marks, D.S. and Larsson, E. (2012) miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics*, 28, 2062-2063.

40. Agarwal, V., Bell, G.W., Nam, J.W. and Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4.

41. Betel, D., Koppal, A., Agius, P., Sander, C. and Leslie, C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*, 11, R90.

42. Gumienny, R. and Zavolan, M. (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res*, 43, 9095.

43. Khorshid, M., Hausser, J., Zavolan, M. and van Nimwegen, E. (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 10, 253-255.

44. Menor, M., Ching, T., Zhu, X., Garmire, D. and Garmire, L.X. (2014) mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome Biol*, 15, 500.

45. **Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. and Li, T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, 37, D105-110.**

46. **Chou, C.H., Chang, N.W., Shrestha, S., Hsu, S.D., Lin, Y.L., Lee, W.H., Yang, C.D., Hong, H.C., Wei, T.Y., Tu, S.J. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res*, 44, D239-247.**

47. **Bandyopadhyay, S., Ghosh, D., Mitra, R. and Zhao, Z. (2015) MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets. *Sci Rep*, 5, 8004.**

48. **Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. and Yang, J.-H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42, D92-D97.**

49. **Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, 32, 1316-1322.**

50. **Helwak, A. and Tollervey, D. (2014) Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat Protoc*, 9, 711-728.**

51. **Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E. and Rajewsky, N. (2014) Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol Cell*, 54, 1042-1054.**

52. **Tay, Y., Zhang, J., Thomson, A.M., Lim, B. and Rigoutsos, I. (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455, 1124-1128.**

53. **Sulc, M., Marin, R.M., Robins, H.S. and Vanicek, J. (2015) PACCMIT/PACCMIT-CDS: identifying microRNA targets in 3' UTRs and coding sequences. *Nucleic Acids Res*, 43, W474-479.**

54. **Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T. and Hatzigeorgiou, A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res*, 41, W169-173.**

55. **Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455, 58-63.**

56. **Vergoulis, T., Vlachos, I.S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T. and Hatzigeorgiou, A.G. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*, 40, D222-229.**

57. **Chi, S.W., Zang, J.B., Mele, A. and Darnell, R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460, 479-486.**

58. **Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*, 8, 559-564.**

59. **Moore, M.J., Scheel, T.K., Luna, J.M., Park, C.Y., Fak, J.J., Nishiuchi, E., Rice, C.M. and Darnell, R.B. (2015) miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature communications*, 6, 8864.**

60. **German, M.A., Luo, S., Schroth, G., Meyers, B.C. and Green, P.J. (2009) Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc*, 4, 356-362.**

61. **Tan, S.M., Kirchner, R., Jin, J., Hofmann, O., McReynolds, L., Hide, W. and Lieberman, J. (2014) Sequencing of Captive Target Transcripts Identifies the Network of Regulated Genes and Functions of Primate-Specific miR-522. *Cell Rep*, 8, 1225-1239.**

62. **Wolter, J.M., Kotagama, K., Pierre-Bez, A.C., Firago, M. and Mangone, M. (2015) 3'LIFE: a functional assay to detect miRNA targets in high-throughput. *Nucleic Acids Res*, 42, e132.**

63. Braun, J., Misiak, D., Busch, B., Krohn, K. and Hüttelmaier, S. (2014) Rapid identification of regulatory microRNAs by miTRAP (miRNA trapping by RNA in vitro affinity purification). *Nucleic Acids Research*, 42, e66.

64. Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.L., Maniou, S., Karathanou, K., Kalfakakou, D. *et al.* (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res*, 43, D153-159.

65. Jaskiewicz, L., Bilen, B., Hausser, J. and Zavolan, M. (2012) Argonaute CLIP - A method to identify in vivo targets of miRNAs. *Methods*.

66. Corcoran, D.L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R.L., Keene, J.D. and Ohler, U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol*, 12, R79.

67. Georgiev, S., Boyle, A.P., Jayasurya, K., Ding, X., Mukherjee, S. and Ohler, U. (2010) Evidence-ranked motif identification. *Genome Biol*, 11, R19.

68. Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34, W369-373.

69. Khorshid, M., Rodak, C. and Zavolan, M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res*, 39, D245-252.

70. Wang, T., Chen, B., Kim, M., Xie, Y. and Xiao, G. (2014) A model-based approach to identify binding sites in CLIP-Seq data. *PLoS One*, 9, e93248.

71. Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012) Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res*, 40, e160.

72. Khorshid, M., Hausser, J., Zavolan, M. and van Nimwegen, E. (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat Methods*, 10, 253-255.

73. Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*, 8, 559-564.

74. Erhard, F., Dolken, L., Jaskiewicz, L. and Zimmer, R. (2013) PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biol*, 14, R79.

75. Majoros, W.H., Lekprasert, P., Mukherjee, N., Skalsky, R.L., Corcoran, D.L., Cullen, B.R. and Ohler, U. (2013) MicroRNA target site identification by integrating sequence and binding information. *Nat Methods*, 10, 630-633.

76. Sethupathy, P., Corda, B. and Hatzigeorgiou, A.G. (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12, 192-197.

77. Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res*, 42, D78-85.

78. Khorshid, M., Rodak, C. and Zavolan, M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res*, 39, D245-252.

79. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57-74.

80. Johnsson, P., Lipovich, L., Grander, D. and Morris, K.V. (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta*, 1840, 1063-1071.

81. Ponjavic, J., Ponting, C.P. and Lunter, G. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, 17, 556-565.

82. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458, 223-227.

83. Marques, A.C. and Ponting, C.P. (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol*, 10, R124.

84. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*, 22, 1775-1789.

85. Pang, K.C., Frith, M.C. and Mattick, J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*, 22, 1-5.

86. Gibb, E.A., Brown, C.J. and Lam, W.L. (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer*, 10, 38.

87. Ponting, C.P., Oliver, P.L. and Reik, W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, 136, 629-641.

88. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915-1927.

89. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *Embo J*, 33, 981-993.

90. Baker, M. (2011) Long noncoding RNAs: the search for function. *Nat Meth*, 8, 379-383.

91. Rinn, J.L. and Chang, H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, 81, 145-166.

92. Gutschner, T. and Diederichs, S. (2012) The hallmarks of cancer: A long non-coding RNA point of view. *RNA Biology*, 9, 703-719.

93. Cai, X. and Cullen, B.R. (2007) The imprinted H19 noncoding RNA is a primary microRNA precursor. *Rna*, 13, 313-316.

94. Georgakilas, G., Vlachos, I.S., Paraskevopoulou, M.D., Yang, P., Zhang, Y., Economides, A.N. and Hatzigeorgiou, A.G. (2014) microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat Commun*, 5, 5700.

95. Braconi, C., Kogure, T., Valeri, N., Huang, N., Nuovo, G., Costinean, S., Negrini, M., Miotto, E., Croce, C.M. and Patel, T. (2011) microRNA-29 can regulate expression of the long non-coding RNA gene MEG3 in hepatocellular cancer. *Oncogene*, 30, 4750-4756.

96. Fan, M., Li, X., Jiang, W., Huang, Y., Li, J. and Wang, Z. (2013) A long non-coding RNA, PTCSC3, as a tumor suppressor and a target of miRNAs in thyroid cancer cells. *Exp Ther Med*, 5, 1143-1146.

97. Calin, G.A., Liu, C.G., Ferracin, M., Hyslop, T., Spizzo, R., Sevignani, C., Fabbri, M., Cimmino, A., Lee, E.J., Wojcik, S.E. *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell*, 12, 215-229.

98. Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L. *et al.* (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477, 295-300.

99. Hansen, T.B., Wiklund, E.D., Bramsen, J.B., Villadsen, S.B., Statham, A.L., Clark, S.J. and Kjems, J. (2011) miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *Embo J*, 30, 4414-4422.

100. Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495, 333-338.

101. Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K. and Kjems, J. (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, 495, 384-388.

102. Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F. and Fan, Q. (2010) CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res*, 38, 5366-5383.

103. Faghihi, M.A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M.P., Nalls, M.A., Cookson, M.R., St-Laurent, G., 3rd and Wahlestedt, C. (2010) Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol*, 11, R56.

104. Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J. and Pandolfi, P.P. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465, 1033-1038.

105. Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A. and Bozzoni, I. (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147, 358-369.

106. Imig, J., Brunschweiger, A., Brummer, A., Guennewig, B., Mittal, N., Kishore, S., Tsikrika, P., Gerber, A.P., Zavolan, M. and Hall, J. (2014) miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nat Chem Biol*.

107. Zhang, L., Yang, F., Yuan, J.H., Yuan, S.X., Zhou, W.P., Huo, X.S., Xu, D., Bi, H.S., Wang, F. and Sun, S.H. (2013) Epigenetic activation of the MiR-200 family contributes to H19-mediated metastasis suppression in hepatocellular carcinoma. *Carcinogenesis*, 34, 577-586.

108. Consortium, F., the, R.P., Clst, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, 507, 462-470.

109. Wang, X., Li, M., Wang, Z., Han, S., Tang, X., Ge, Y., Zhou, L., Zhou, C., Yuan, Q. and Yang, M. (2014) Silencing of long noncoding RNA MALAT1 by miR-101 and miR-217inhibits proliferation, migration and invasion of esophageal squamous cell carcinoma cells. *J Biol Chem*.

110. Leucci, E., Patella, F., Waage, J., Holmstrom, K., Lindow, M., Porse, B., Kauppinen, S. and Lund, A.H. (2013) microRNA-9 targets the long non-coding RNA MALAT1 for degradation in the nucleus. *Sci Rep*, 3, 2535.

111. Zhang, Z., Zhu, Z., Watabe, K., Zhang, X., Bai, C., Xu, M., Wu, F. and Mo, Y.Y. (2013) Negative regulation of lncRNA GAS5 by miR-21. *Cell Death Differ*, 20, 1558-1568.

112. Prensner, J.R., Chen, W., Han, S., Iyer, M.K., Cao, Q., Kothari, V., Evans, J.R., Knudsen, K.E., Paulsen, M.T., Ljungman, M. *et al.* (2014) The Long Non-Coding RNA PCAT-1 Promotes Prostate Cancer Cell Proliferation through cMyc. *Neoplasia*, 16, 900-908.

113. Wang, K., Sun, T., Li, N., Wang, Y., Wang, J.X., Zhou, L.Y., Long, B., Liu, C.Y., Liu, F. and Li, P.F. (2014) MDRL lncRNA regulates the processing of miR-484 primary transcript by targeting miR-361. *PLoS Genet*, 10, e1004467.

114. Chiyomaru, T., Yamamura, S., Fukuhara, S., Yoshino, H., Kinoshita, T., Majid, S., Saini, S., Chang, I., Tanaka, Y., Enokida, H. *et al.* (2013) Genistein inhibits prostate cancer cell growth by targeting miR-34a and oncogenic HOTAIR. *PLoS One*, 8, e70372.

115. Cao, C., Sun, J., Zhang, D., Guo, X., Xie, L., Li, X., Wu, D. and Liu, L. (2014) The Long Intergenic Noncoding RNA UFC1, A Target of MicroRNA 34a, Interacts With the mRNA Stabilizing Protein HuR to Increase Levels of beta-Catenin in HCC Cells. *Gastroenterology*.

116. Gao, Y., Meng, H., Liu, S., Hu, J., Zhang, Y., Jiao, T., Liu, Y., Ou, J., Wang, D., Yao, L. *et al.* (2015) LncRNA-HOST2 regulates cell biological behaviors in epithelial ovarian cancer through a mechanism involving microRNA let-7b. *Human molecular genetics*, 24, 841-852.

117. Paraskevopoulou, M.D., Vlachos, I.S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., Zagganas, K., Tsanakas, P., Floros, E., Dalamagas, T. *et al.* (2016) DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res*, 44, D231-238.

118. Paraskevopoulou, M.D. and Hatzigeorgiou, A.G. (2016) Analyzing MiRNA-LncRNA Interactions. *Methods Mol Biol*, 1402, 271-286.

119. Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Reczko, M., Maragkakis, M., Dalamagas, T.M. and Hatzigeorgiou, A.G. (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res*, 41, D239-245.

120. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22, 1760-1774.

121. Volders, P.J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J. and Mestdagh, P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res*, 43, 4363-4364.

122. Wang, X. and El Naqa, I.M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24, 325-332.

123. Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y. and Chen, R. (2014) NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res*, 42, D104-108.

124. Bhartiya, D., Pal, K., Ghosh, S., Kapoor, S., Jalali, S., Panwar, B., Jain, S., Sati, S., Sengupta, S., Sachidanandan, C. *et al.* (2013) lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database (Oxford)*, 2013, bat034.

125. Zhou, Z., Shen, Y., Khan, M.R. and Li, A. (2015) LncReg: a reference resource for lncRNA-associated regulatory networks. *Database (Oxford)*, 2015.

126. Theodoridis, S. and Koutroumbas, K. (2006) *Pattern recognition*. 3rd ed. Academic Press, San Diego, CA.

127. Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine learning*, 20, 273-297.

128. Breiman, L. (2001) Random forests. *Machine learning*, 45, 5-32.

129. Bishop, C.M. (2006) Pattern recognition. *Machine Learning*, 128.

130. Quinlan, J.R. (1986) Induction of decision trees. *Machine learning*, 1, 81-106.

131. Lewis, D. (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. *Machine Learning: ECML-98*, 4-15.

132. Amirkhah, R., Farazmand, A., Gupta, S.K., Ahmadi, H., Wolkenhauer, O. and Schmitz, U. (2015) Naive Bayes classifier predicts functional microRNA target interactions in colorectal cancer. *Mol Biosyst*, 11, 2126-2134.

133. Moller, M.F. (1993) A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6, 525-533.

134. Reczko, M., Maragkakis, M., Alexiou, P., Papadopoulos, G.L. and Hatzigeorgiou, A.G. (2011) Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data. *Front Genet*, 2, 103.

135. Lancashire, L.J., Lemetre, C. and Ball, G.R. (2009) An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in bioinformatics*, 10, 315-329.

136. Tipping, M.E. (2001) Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1, 211-244.

137. Quinlan, J.R. (1987) Simplifying decision trees. *International journal of man-machine studies*, 27, 221-234.

138. Natekin, A. and Knoll, A. (2013) Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7, 21.

139. Serneels, S., De Nolf, E. and Van Espen, P.J. (2006) Spatial sign preprocessing: a simple way to impart moderate robustness to multivariate estimators. *Journal of chemical information and modeling*, 46, 1402-1409.

140. Peng, H., Ding, C. and Long, F. (2005). IEEE COMPUTER SOC 10662 LOS VAQUEROS CIRCLE, PO BOX 3014, LOS ALAMITOS, CA 90720-1314 USA.

141. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39, D152-157.

142. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*, 42, D756-763.

143. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34, W729-732.

144. Andrews, S. (2015) FastQC: A quality control tool for high throughput sequence data.

145. Davis, M.P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. and Enright, A.J. (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63, 41-49.

146. Krueger, F. (2015) Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

147. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114-2120.

148. Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26, 873-881.

149. Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.L., Maniou, S., Karathanou, K., Kalfakakou, D. *et al.* (2014) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res*.

150. Smith, L., Tanabe, L., Ando, R., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C., Ganchev, K. *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biology*, 9, S2.

151. Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., Gonzalez, G., Nenadic, G. and Bergman, C.M. (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27, 2769-2771.

152. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*, 43, D670-681.

153. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*, 41, D991-995.

154. Li, B. and Dewey, C. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.

155. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res*, 43, D662-669.

156. Helwak, A. and Tollervey, D. (2014) Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat Protoc*, 9, 711-728.

157. Friedersdorf, M.B. and Keene, J.D. (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol*, 15, R2.

158. Gottwein, E., Corcoran, D.L., Mukherjee, N., Skalsky, R.L., Hafner, M., Nusbaum, J.D., Shamulailatpam, P., Love, C.L., Dave, S.S., Tuschl, T. *et al.* (2011) Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell Host Microbe*, 10, 515-526.

159. Skalsky, R.L., Corcoran, D.L., Gottwein, E., Frank, C.L., Kang, D., Hafner, M., Nusbaum, J.D., Feederle, R., Delecluse, H.J., Luftig, M.A. *et al.* (2012) The viral and cellular microRNA targetome in lymphoblastoid cell lines. *PLoS Pathog*, 8, e1002484.

160. Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L. and Betel, D. (2011) Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. *Genes & development*, 25, 2173-2186.

161. Eichhorn, S.W., Guo, H., McGeary, S.E., Rodriguez-Mias, R.A., Shin, C., Baek, D., Hsu, S.H., Ghoshal, K., Villen, J. and Bartel, D.P. (2014) mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell*, 56, 104-115.

162. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20, 307-315.

163. Carvalho, B.S. and Irizarry, R.A. (2010) A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26, 2363-2367.

164. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43, e47.

165. Comoglio, F., Sievers, C. and Paro, R. (2015) Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics*, 16, 32.

166. Morgulis, A., Gertz, E.M., Schaffer, A.A. and Agarwala, R. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol*, 13, 1028-1040.

167. Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27, 863-864.

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

169. Wee, L.M., Flores-Jasso, C.F., Salomon, W.E. and Zamore, P.D. (2012) Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, 151, 1055-1067.

170. Schirle, N.T., Sheu-Gruttadauria, J. and MacRae, I.J. (2014) Structural basis for microRNA targeting. *Science*, 346, 608-613.

171. Chi, S.W., Hannon, G.J. and Darnell, R.B. (2012) An alternative mode of microRNA target recognition. *Nat Struct Mol Biol*, 19, 321-327.

172. Martin, H.C., Wani, S., Steptoe, A.L., Krishnan, K., Nones, K., Nourbakhsh, E., Vlassov, A., Grimmond, S.M. and Cloonan, N. (2014) Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. *Genome Biol*, 15, R51.

173. Helwak, A., Kudla, G., Dudnakova, T. and Tollervey, D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153, 654-665.

174. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39, D152-157.

175. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. *et al.* (2012) Ensembl 2012. *Nucleic acids research*, 40, D84-90.

176. Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129, 1401-1414.

177. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40, D109-114.

178. Ramon, C.L. and Folks, J.L. (1971) Asymptotic Optimality of Fisher's Method of Combining Independent Tests. *Journal of the American Statistical Association*, 66, 802-806.

179. Vlachos, I.S., Kostoulas, N., Vergoulis, T., Georgakilas, G., Reczko, M., Maragkakis, M., Paraskevopoulou, M.D., Prionidis, K., Dalamagas, T. and Hatzigeorgiou, A.G. (2012) DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Research*, 40, W498-W504.

180. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Research*, 42, D749-D755.

181. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42, D68-D73.

182. Vlachos, I.S., Kostoulas, N., Vergoulis, T., Georgakilas, G., Reczko, M., Maragkakis, M., Paraskevopoulou, M.D., Prionidis, K., Dalamagas, T. and Hatzigeorgiou, A.G. (2012) DIANA

miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res*, 40, W498-504.

183. Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T. and Hatzigeorgiou, A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res*, 41, W169-173.

184. Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Reczko, M., Maragkakis, M., Dalamagas, T.M. and Hatzigeorgiou, A.G. (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res*, 41, D239-245.

185. Imig, J., Brunschweiger, A., Brummer, A., Guennewig, B., Mittal, N., Kishore, S., Tsikrika, P., Gerber, A.P., Zavolan, M. and Hall, J. (2015) miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19-miR-106a interaction. *Nat Chem Biol*, 11, 107-114.

186. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W. and Bruford, E.A. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic acids research*, 43, D1079-1085.

187. Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R. and Zhao, Y. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic acids research*, 42, D98-103.

188. Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G. and Cui, Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic acids research*, 41, D983-986.

189. Vlachos, I.S., Zagganas, K., Paraskevopoulou, M.D., Georgakilas, G., Karagkouni, D., Vergoulis, T., Dalamagas, T. and Hatzigeorgiou, A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res*, 43, W460-466.

190. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20, 110-121.

191. Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289-300.