

SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING



**Computational analysis of non-coding RNA regulatory functions on a
transcriptome-wide scale**

PhD DISSERTATION

Dimitra Karagkouni

Supervisor: Artemis Hatzigeorgiou, Professor

Volos, 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ



**Υπολογιστική Ανάλυση των Λειτουργιών των μη Κωδικών Μεταγράφων στη
Γονιδιωματική Ρύθμιση**

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Δήμητρα Καραγκούνη

Επιβλέπουσα: Άρτεμις Χατζηγεωργίου, Καθηγήτρια

Βόλος, 2019

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

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

Μπάγκος Παντελεήμων, Καθηγητής, Τμήμα Πληροφορικής με εφαρμογές στην Βιοϊατρική, 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

Κατσαρός Δημήτριος, Επίκουρος Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Χατζιωάννου Αριστοτέλης, Ερευνητής Βαθμίδας Β', Εθνικό Ίδρυμα Ερευνών

Παυλόπουλος Γεώργιος, Ερευνητής Βαθμίδας Β', Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ»

ΕΥΧΑΡΙΣΤΙΕΣ

Η σελίδα με τις ευχαριστίες συνήθως είναι από τις πρώτες σε κάθε διδακτορική διατριβή. Είμαι σίγουρη όμως ότι ο κάθε υποψήφιος διδάκτορας την αφήνει για το τέλος, καθώς θα πρέπει να αναπολήσει όλες εκείνες τις στιγμές που έζησε κατά τη διάρκεια του διδακτορικού του και σε λίγες γραμμές να συνοψίσει τις ευχαριστίες του για όλους εκείνους του ανθρώπους που υπήρξαν οι πυλώνες στην εκπόνησή του. Στο σημείο αυτό λοιπόν, θα ήθελα να ευχαριστήσω τους ανθρώπους που με τη συνεχή τους παρουσία, υποστήριξη αλλά και ανεκτικότητα υλοποιήθηκε η παρούσα διδακτορική διατριβή.

Ο πρώτος άνθρωπος που ευχαριστώ είναι η καθηγήτρια και επιβλέπουσα μου Άρτεμις Χατζηγεωργίου. Την ευχαριστώ θερμά καθώς μου έδωσε την ευκαιρία να ασχοληθώ με την έρευνα, σε πεδία αιχμής του τομέα της βιολογίας και της βιοπληροφορικής από τη πρώτη στιγμή που μπήκα στο εργαστήριο DIANA-Lab το 2014. Μαζί της εκπόνησα τόσο τη μεταπτυχιακή μου μελέτη, όσο και τη διδακτορική μου έρευνα. Είμαι ευγνώμων τόσο για τη διαρκή καθοδήγηση και την τεχνογνωσία που μου προσέφερε όλα αυτά τα χρόνια, όσο και για τη δυνατότητα που μου έδωσε να συνεργαστώ με εξαιρετους συναδέλφους και επιστήμονες. Θέλω να την ευχαριστήσω για την υποστήριξή της, την κατανόησή της αλλά και για τις στιγμές που πίστεψε σε εμένα πιο πολύ από ότι εγώ η ίδια.

Ευχαριστώ θερμά τον Καθηγητή Παντελεήμονα Μπάγκο και τον Αναπληρωτή Καθηγητή Γεράσιμο Ποταμιάνο, που ως μέλη της Τριμελούς Επιτροπής, μου έδωσαν τη δυνατότητα να εργαστώ στο διεπιστημονικό πεδίο της Βιοπληροφορικής και συνέβαλαν στην εκπόνηση της διδακτορικής διατριβής.

Θα ήθελα επίσης να ευχαριστήσω όλα τα υπόλοιπα μέλη της εξεταστικής επιτροπής, την Αναπληρώτρια Καθηγήτρια Τσομπανοπούλου Παναγιώτα, τον Επίκουρο Καθηγητή Κατσαρό Δημήτριο, τον Ερευνητή Βαθμίδας Β' Χατζηγιάννου Αριστοτέλη και τον Ερευνητή Βαθμίδας Β' Παυλόπουλο Γεώργιο για την τιμή που μου έκαναν να συμμετέχουν στην κρίση του διδακτορικού μου.

Η παρούσα διδακτορική διατριβή δε θα μπορούσε να ολοκληρωθεί χωρίς τους συνεργάτες και φίλους του DIANA-Lab.

Ιδιαίτερα, θέλω να ευχαριστήσω τη Δρ. Μαρία Παρασκευοπούλου. Με τη Μαρία αρχίσαμε να συνεργαζόμαστε από τα πρώτα μου βήματα στον τομέα της έρευνας, την περίοδο της μεταπτυχιακής μου μελέτης. Η μοναδική και αξιοθαύμαστη συνεργασία μας, παράλληλα με τη αμοιβαία εκτίμηση και συμπαράσταση, επέφεραν πολλές δημοσιεύσεις. Η ολοκλήρωση του διδακτορικού μου οφείλεται σε μεγάλο βαθμό σε εκείνη. Την ευχαριστώ θερμά για την από κοινού εργασία μας στην ανάπτυξη των αλγορίθμων microCLIP και microT καθώς και στη δημιουργία της βάσης δεδομένων DIANA-TarBase.

Θέλω να ευχαριστήσω θερμά το Δρ. Ιωάννη Βλάχο. Ο Γιάννης μοιράστηκε τις γνώσεις του μαζί μου σε κάθε στιγμή της ερευνητικής μου πορείας. Τον ευχαριστώ για την αगाστή

συνεργασία μας όλα αυτά τα χρόνια. Η αμείωτη συμβολή του υπήρξε καθοριστικός παράγοντας για την ανάπτυξη του αλγορίθμου microCLIP και τη δημιουργία της βάσης DIANA-TarBase. Τον ευχαριστώ για τις πολύτιμες συμβουλές του και για τη στήριξή του σε καθ' όλη τη διάρκεια της συνεργασίας μας.

Ευχαριστώ το συνάδελφο και φίλο Σπύρο Ταστούγλου για τη συνολική συνεισφορά του στη διατριβή, καθώς πραγματοποίησε την ανάλυση πολλών πειραμάτων που ενσωματώθηκαν στο DIANA-TarBase και χρησιμοποιήθηκαν στην ανάπτυξη του microCLIP και microT. Ευχαριστώ επίσης το Γιώργο Σκούφο, αρχικά για τη συμβολή του στη δημιουργία της βάσης δεδομένων DIANA-TarBase, αλλά και για την παρουσία του στο εργαστήριο. Ο Σπύρος με το Γιώργο με στήριξαν ως συνεργάτες αλλά και ως φίλοι, πάντα πρόθυμοι να συζητήσουμε οποιοδήποτε θέμα με απασχολούσε.

Ακόμη θέλω να ευχαριστήσω το Δρ. Γεώργιο Γεωργακίλα για την άριστη συνεργασία μας και τη πολύτιμη βοήθεια του στην ολοκλήρωση αρκετών μελετών.

Για την πραγμάτωση του σχεδιασμού και της διεπαφής της βάσης DIANA-TarBase, ευχαριστώ τους συνεργάτες μας στο Ινστιτούτο «Αθηνά», καθώς η αλληλεπίδραση μαζί τους υπήρξε καθοριστική. Ειδικότερα ευχαριστώ τους Δρ. Θεόδωρο Δαλαμάγκα και Δρ. Θανάση Βεργούλη για τη σημαντική συμβολή τους τόσο στην παραπάνω μελέτη όσο και στις περισσότερες από τις δημοσιεύσεις στις οποίες μετέχω.

Ιδιαίτερη μνεία θα ήθελα να κάνω στο Ελληνικό Ίδρυμα Έρευνας και Καινοτομίας (ΕΛΙΔΕΚ), για την οικονομική στήριξη που μου προσέφερε τον τελευταίο χρόνο της διδακτορικής μου διατριβής.

Κλείνοντας θέλω να ευχαριστήσω από καρδιάς τους δικούς μου ανθρώπους. Τη μητέρα μου Ελένη και τον πατέρα μου Ευστάθιο, τον αδερφό μου Αθανάσιο, τους αγαπημένους μου φίλους και κυρίως τον Οδυσσέα, για την αγάπη, την υποστήριξη και την κατανόηση που πάντα μου δείχνουν. Χωρίς τη δική τους συμβολή τίποτε από όσα έχω καταφέρει δε θα ήταν εφικτό.

ABSTRACT

The emerging technological developments during the past decade enable large scale analyses in the “regulatory RNA” field and have turned non-coding RNA (ncRNA), initially considered as junk, into a research goldmine. ncRNAs play a crucial role in a remarkable variety of physiological and pathological biological processes. The vast production of data has also been the most important factor underlying the accelerated growth of bioinformatics, a field dedicated to the analysis of data and the development of computational tools indispensable for handling, manipulating and interpreting the results. This thesis focuses on the thorough aggregation of high-throughput data and state-of-the-art Machine Learning techniques in order to develop algorithms for the functional characterization of non-coding transcripts.

The current dissertation is specialized on a specific category of RNA transcripts, the microRNAs. microRNAs (miRNAs) are small single stranded non-coding RNA molecules, ~22 nucleotides long, that are loaded into Argonaute (AGO) to induce target cleavage, degradation or translational suppression. Accurate characterization of their targets is considered fundamental to elucidate their regulatory roles. Over the last 15 years, a multitude of *in silico* and experimental procedures have been developed aiming to determine the miRNA interactome. Currently, high-throughput techniques have enabled the identification of novel experimentally-supported miRNA-gene interactions in a transcriptome-wide scale. This wealth of information is dispersed in a great number of publications and raw datasets. During this thesis DIANA-TarBase v8.0, a reference database devoted to the indexing of experimentally-supported miRNA targets, was designed. Its 8th version is the first database to index more than 1 million entries, corresponding to ~700,000 unique miRNA target pairs, supported by more than 33 experimental methodologies, applied to 592 cell types/tissues under ~430 experimental conditions.

AGO-CLIP-Seq experiments are the most widely used high-throughput methodologies. PAR-CLIP variant against AGO proteins methodology has been performed to map miRNA-gene interactions on a transcriptome-wide scale for healthy or disease cell types. Computational methods devoted to AGO-PAR-CLIP present reduced ability to distinguish a large portion of genuine miRNA-targets. To this end, one of the aims of this thesis was to revisit, identify and address current obstacles in AGO-CLIP-Seq analysis. An *in silico* framework for CLIP-guided identification of miRNA interactions, microCLIP model, was developed. microCLIP is the first relevant implementation to employ the innovative super learner ensemble framework and the only available A-to-Z computational approach for the analysis of AGO-PAR-CLIP datasets. It operates on every AGO-enriched cluster, providing previously neglected functional miRNA binding events with strong RNA accessibility.

microCLIP deployment emboldened the development of a next generation de novo miRNA target prediction algorithm. Even the extensive production of relevant approaches observed

during the past few years, leading implementations still achieve a far from perfect predictive accuracy followed by an increased number of false positives predictions. Therefore, microT Super Learning framework is presented that maintains and upgrades the pipeline adopted in microCLIP, by enhancing the training with even more high-throughput experiments under a tissue-specific scheme. The new model characterizes interactions with stronger functional efficacy and correctly detects 1.5-fold more experimentally validated target sites when juxtaposed against leading computational approaches. The increased performance of microCLIP and microT frameworks in the detection of miRNA interactions, uncovers previously elusive regulatory events and miRNA-controlled pathways.

During this thesis, the candidate participated in 9 scientific studies, involving computational approaches for determining the activity of non-coding transcripts and in two of them is first author. The candidate's main research activity and contribution in the publications incorporates the implementation of algorithms and automated pipelines for the analysis of Next Generation Sequencing data, data integration for the elucidation of non-coding RNA function and their involvement in mechanisms of post-transcriptional gene regulation. The studies are published in international journals of high impact factor and a total of 942 citations have been received so far, according to Google Scholar.

SUBJECT AREA: Computational Biology

KEYWORDS: microRNA, high-throughput experiments, AGO-HITS-CLIP, AGO-PAR-CLIP, target prediction, experimentally supported targets, *in silico* predicted targets, Machine Learning

ΠΕΡΙΛΗΨΗ

Οι ραγδαίες τεχνολογικές εξελίξεις την τελευταία δεκαετία επέτρεψαν αναλύσεις μεγάλης κλίμακας στο πεδίο του «ρυθμιστικού RNA», μετατρέποντας τα μη-κωδικά μετάγραφα, που αρχικά θεωρούνταν «σκουπίδια», σε ερευνητικό «χρυσωρυχείο». Τα μη-κωδικά μετάγραφα διαδραματίζουν καθοριστικό ρόλο σε ένα αξιοσημείωτο αριθμό από φυσιολογικές και παθολογικές βιολογικές διεργασίες. Η τεράστια παραγωγή δεδομένων ήταν επίσης ένας από τους σημαντικότερους παράγοντες της επιταχυνόμενης εξέλιξης του τομέα της βιοπληροφορικής, ενός τομέα εξειδικευμένου στην ανάλυση βιολογικών δεδομένων και την ανάπτυξη υπολογιστικών εργαλείων, απαραίτητων για την επεξεργασία και την ερμηνεία των αποτελεσμάτων τους. Αυτή η εργασία επικεντρώνεται στο λεπτομερή και ακριβή συνδυασμό υψηλής διεκπεραιωτικής ικανότητας δεδομένων και σύγχρονων τεχνικών μηχανικής μάθησης για την ανάπτυξη αλγορίθμων με στόχο το λειτουργικό χαρακτηρισμό των μη-κωδικών μεταγραφών.

Η παρούσα διατριβή επικεντρώνεται σε μια συγκεκριμένη κατηγορία μεταγραφών, τα microRNAs. Τα microRNAs (miRNAs) είναι μικρά, μονόκλωνα, μη-κωδικά μόρια RNA, μήκους ~ 22 νουκλεοτιδίων, που προσδένονται στην πρωτεΐνη Αργοναύτη (AGO) για να προκαλέσουν τη διάσπαση του μεταγράφου-στόχου, την αποικοδόμηση ή την καταστολή της μετάφρασής του. Ο ακριβής χαρακτηρισμός των στόχων τους θεωρείται θεμελιώδης για την αποσαφήνιση του ρυθμιστικού τους ρόλου. Τα τελευταία 15 χρόνια, έχει αναπτυχθεί μία πληθώρα υπολογιστικών και πειραματικών προσεγγίσεων με στόχο τον προσδιορισμό των αλληλεπιδράσεων των μικρών RNAs. Επί του παρόντος, οι τεχνικές υψηλής απόδοσης επέτρεψαν την εύρεση νέων πειραματικά υποστηριζόμενων αλληλεπιδράσεων των miRNAs σε όλο το μεταγράφομα. Αυτός ο πλούτος των πληροφοριών είναι διασκορπισμένος σε μεγάλο αριθμό δημοσιεύσεων και ακατέργαστων δεδομένων. Κατά τη διάρκεια αυτής της διατριβής, σχεδιάστηκε το DIANA-TarBase v8.0, μια βάση δεδομένων αναφοράς, αφιερωμένη στην ευρετηρίαση πειραματικά υποστηριζόμενων στόχων των miRNAs. Η 8^η έκδοση είναι η πρώτη βάση δεδομένων που αναφέρει περισσότερες από 1 εκατομμύριο καταχωρήσεις, που αντιστοιχούν σε ~700.000 μοναδικές miRNA-gene αλληλεπιδράσεις, υποστηριζόμενες από περισσότερες από 33 πειραματικές μεθοδολογίες, που έχουν εφαρμοστεί σε 592 κυτταρικούς τύπους/ιστούς, υπό ~ 430 πειραματικές συνθήκες.

Τα πειράματα με ανοσοκατακρήμνηση της πρωτεΐνης AGO (AGO-CLIP-Seq) αποτελούν τις πιο διαδεδομένες μεθοδολογίες υψηλής απόδοσης. Η AGO-PAR-CLIP τεχνική έχει πραγματοποιηθεί ευρέως για τη χαρτογράφηση miRNA-gene αλληλεπιδράσεων σε μεγάλη κλίμακα σε υγιείς ή ασθενείς τύπους κυττάρων. Οι υπολογιστικές μέθοδοι που έχουν αναπτυχθεί με στόχο την ανάλυση αυτών των δεδομένων παρουσιάζουν μειωμένη ικανότητα να διακρίνουν ένα μεγάλο μέρος των πραγματικών miRNA-στόχων. Για το σκοπό αυτό, ένας από τους σκοπούς της παρούσας διατριβής είναι να επανεξετάσει, να εντοπίσει και να αντιμετωπίσει τα τρέχοντα εμπόδια στην ανάλυση AGO-CLIP-Seq δεδομένων.

Παρουσιάζεται, λοιπόν, το μοντέλο microCLIP, μία υπολογιστική προσέγγιση για την κατευθυνόμενη από CLIP-Seq δεδομένα αναγνώριση των αλληλεπιδράσεων των miRNAs. Το microCLIP είναι ένα καινοτόμο ensemble μοντέλο βαθιάς εκμάθησης (super learner) και η μόνη διαθέσιμη υπολογιστική προσέγγιση που αναλύει AGO-PAR-CLIP δεδομένα από το Α έως το Ω. Επεξεργάζεται όλες τις εμπλουτισμένες σε AGO περιοχές, παρέχοντας λειτουργικές περιοχές πρόσδεσης των miRNAs με ισχυρή προσβασιμότητα, που μέχρι πρότινος αγνοούνταν.

Η ανάπτυξη του microCLIP ενέπνευσε τη δημιουργία ενός αλγόριθμου επόμενης γενιάς, για την εύρεση των στόχων των miRNAs απουσία πειράματος. Παρά την εκτενή ανάπτυξη σχετικών προσεγγίσεων που παρατηρείται τα τελευταία χρόνια, ακόμη και οι αλγόριθμοι αιχμής εξακολουθούν να επιτυγχάνουν χαμηλή ακρίβεια και αυξημένο αριθμό ψευδώς θετικών προβλέψεων. Για αυτόν το λόγο, αναπτύχθηκε το μοντέλο microT Super Learning που διατηρεί και αναβαθμίζει τη μεθοδολογία του microCLIP αλγορίθμου, ενισχύοντας την εκπαίδευσή του με ακόμη περισσότερα πειράματα υψηλής απόδοσης υπό έναν ιστο-ειδικό σχεδιασμό. Το νέο μοντέλο χαρακτηρίζει αλληλεπιδράσεις με ισχυρότερη λειτουργικότητα και ανιχνεύει σωστά 1.5 φορές περισσότερες πειραματικά επιβεβαιωμένες περιοχές πρόσδεσης των μικρών RNAs, όταν αντιπαρατίθεται με κορυφαίες υπολογιστικές προσεγγίσεις. Η αυξημένη απόδοση των αλγορίθμων microCLIP και microT στην ανίχνευση των αλληλεπιδράσεων των miRNAs, αναδεικνύει ρυθμιστικά συμβάντα που μέχρι πρότινος αγνοούνταν και νέα μοριακά μονοπάτια που ελέγχονται από τα miRNAs.

Κατά τη διάρκεια της παρούσας εργασίας, η υποψήφια διδάκτωρ συμμετείχε σε 9 επιστημονικές δημοσιεύσεις που αφορούσαν υπολογιστικές προσεγγίσεις για τον προσδιορισμό της λειτουργίας των μη κωδικών μεταγραφών και σε δύο από αυτές είναι η πρώτη συγγραφέας. Η κύρια ερευνητική δραστηριότητα και η συμβολή της υποψήφιας στις δημοσιεύσεις αυτές αφορά την εφαρμογή αλγορίθμων, αυτοματοποιημένων ροών ανάλυσης για την επεξεργασία πειραματικών δεδομένων επόμενης γενιάς και τον κατάλληλο συνδυασμό τους με στόχο την αποσαφήνιση της λειτουργίας των μη-κωδικών RNAs και της συμμετοχής τους σε μηχανισμούς μετα-μεταγραφικής γονιδιακής ρύθμισης. Οι μελέτες έχουν δημοσιευθεί σε διεθνή περιοδικά υψηλής απήχησης και οι συνολικές ετεροαναφορές μέχρι σήμερα, σύμφωνα με το Google Scholar, είναι 942.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: microRNA, πειράματα υψηλής απόδοσης, AGO-HITS-CLIP, AGO-PAR-CLIP, πρόβλεψη στόχων, πειραματικά επιβεβαιωμένοι στόχοι, υπολογιστικά προβλεπόμενοι στόχοι, Μηχανική Μάθηση

*Η εργασία αυτή αφιερώνεται στην οικογένειά μου
και στους ανθρώπους που είναι συνέχεια δίπλα μου,
ακόμη κι αν βρίσκονται χιλιόμετρα μακριά*

CONTENTS

| | |
|---|----|
| LIST OF FIGURES | 18 |
| LIST OF TABLES | 30 |
| CHAPTER 1 | 33 |
| 1.1 ncRNAs – from “junk” DNA into a research goldmine..... | 33 |
| 1.2 The discovery of microRNAs | 34 |
| 1.2.1 Biogenesis of microRNAs | 35 |
| 1.2.2 microRNA function | 36 |
| 1.3 Identification of miRNA targets | 37 |
| 1.3.1 <i>In silico</i> approaches for the identification of miRNA:mRNA interactions | 38 |
| 1.3.1.1 Overview of de novo miRNA Target Prediction Algorithms | 38 |
| 1.3.2 Experimental Methods for the identification of miRNA:gene interactions | 42 |
| 1.3.2.1 AGO-CLIP-Seq experimental methodologies | 43 |
| 1.4 AGO-PAR-CLIP guided implementations | 46 |
| 1.5 Databases indexing miRNA-gene interactions | 48 |
| 1.6 Pattern recognition and Machine Learning in Bioinformatics | 48 |
| 1.6.1 Probabilistic classifiers | 49 |
| 1.6.2 Feature Extraction and Selection | 49 |
| 1.6.3 Machine Learning Algorithms | 51 |
| 1.6.3.1 Generalized Linear Models | 51 |
| 1.6.3.2 Naïve Bayes classifier | 51 |
| 1.6.3.3 Support Vector Machines | 52 |
| 1.6.3.4 Decision Trees | 53 |
| 1.6.3.5 Random Forest | 54 |
| 1.6.3.6 Deep Learning methods | 55 |
| 1.6.3.7 Ensemble learning algorithms | 57 |
| 1.6.3.8 Gradient Boosting | 57 |

| | |
|---|-----|
| CHAPTER 2 | 59 |
| 2.1 Methods for the development of the DIANA-TarBase v8.0 repository | 59 |
| 2.1.1 Collected Data | 59 |
| 2.1.2 Analysis of high-throughput datasets incorporated in DIANA-TarBase v8.0 | 60 |
| 2.1.3 Analysis of AGO-CLIP-Seq datasets incorporated in DIANA-TarBase v8.0..... | 60 |
| 2.1.4 Database interface development | 61 |
| 2.2 Implementation of microCLIP, a novel Super Learning Algorithm for the analysis of AGO-CLIP-Seq data | 62 |
| 2.2.1 Dataset collection | 63 |
| 2.2.2 Analysis of high-throughput experiments | 70 |
| 2.2.2.1 miRNA perturbation experiments | 70 |
| 2.2.2.2 AGO-PAR-CLIP and (s)RNA-Seq expression datasets | 70 |
| 2.2.2.3 PARS experimental data..... | 72 |
| 2.2.3 microCLIP <i>in silico</i> framework | 74 |
| 2.2.4 miRNA interactions from <i>in silico</i> implementations | 84 |
| 2.3 Implementation of microT, a de novo miRNA target prediction algorithm..... | 85 |
| 2.3.1 Dataset collection..... | 86 |
| 2.3.2 microT <i>in silico</i> framework | 91 |
| CHAPTER 3 | 97 |
| 3.1 DIANA-TarBase repository | 97 |
| 3.1.1 DIANA-TarBase update: Database statistics..... | 97 |
| 3.1.2 Interface | 101 |
| 3.1.2.1 Querying the database..... | 101 |
| 3.1.2.2 Database interconnections | 103 |
| 3.2 microCLIP Super Learning framework uncovers functional transcriptome-wide miRNA interactions..... | 105 |
| 3.2.1 T-to-C and non-T-to-C PAR-CLIP clusters share common traits | 106 |
| 3.2.2 Structural sequencing data unveil accessible AGO-bound loci..... | 107 |
| 3.2.3 A super learning approach for AGO-PAR-CLIP analysis | 108 |
| 3.2.4 microCLIP detects novel miRNA interactions from AGO-PAR-CLIP clusters | 114 |
| 3.2.5 Functional enrichment shows importance of non-T-to-C targets..... | 117 |
| 3.2.6 Evaluation of microCLIP against AGO-CLIP-guided models..... | 119 |

| | |
|--|-----|
| 3.3. microT, a Next Generation de novo miRNA-target prediction algorithm | 126 |
| 3.3.1 Feature selection..... | 126 |
| 3.3.2 microT Super Learning framework | 130 |
| 3.3.3 Evaluation of microT against other <i>in silico</i> models | 134 |
| CHAPTER 4 | 137 |
| CHAPTER 5 | 139 |
| ABBREVIATIONS - ACRONYMS..... | 143 |
| REFERENCES..... | 147 |

List of Figures

Figure 1: Biogenesis of microRNAs. miRNAs are transcribed into the nucleus either autonomously or as polycistronic molecules. The miRNA precursor (pri-miRNA) is treated by the microprocessor complex which is composed of the DROSHA and DCGR8 proteins. The resulting pre-miRNA which is extracted from the nucleus by means of Exportin-5 and protein Dicer cuts the loop at the end of miRNA precursor. From the generated double-stranded miRNA, one clone is usually selected, which is incorporated into the RISC complex. The most well-defined functions of miRNAs are observed in the cytoplasm such as: translation suppression and/or degradation of the mRNA target. Recent studies suggest that some mature miRNAs are able to re-enter into the nucleus and interact with other transcripts, something that displays their possible involvement in additional mechanisms that have not been yet characterized. This figure has been designed for the purpose of this dissertation. ... 36

Figure 2: Illustration of miRNA targeting. miRNAs are loaded on AGO and guide the RISC complex to target MRE(s). RISC binding to its target genes can either cease their translation or induce their cleavage and/or degradation (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 37

Figure 3: Overview of AGO-HITS-CLIP (left) and AGO-PAR-CLIP (right) protocols. This figure has been designed for the purpose of this dissertation. 45

Figure 4: Overview of CLASH experiment. This figure has been designed for the purpose of this dissertation. 46

Figure 5: Representation of possible hyperplanes (left) and the optimal hyperplane (right) in a SVM classification scheme. This figure has been designed for the purpose of this dissertation. 53

Figure 6: Random Forest representation with two trees. This figure has been designed for the purpose of this dissertation. 55

Figure 7: Representation of a node in a Deep Learning scheme. This figure has been designed for the purpose of this dissertation..... 56

| | |
|--|----|
| Figure 8: Deep Learning architecture. This figure has been designed for the purpose of this dissertation. | 56 |
| Figure 9: Snapshot from the IGV Genome Browser depicting the adopted pipeline for the analysis of the AGO-CLIP-Seq libraries. Raw CLIP-Seq reads are initially aligned into the reference genome. Regions enriched in AGO are formed by overlapping reads. AGO-CLIP clusters are annotated in a comprehensive set of transcripts. MRE identification is subsequently applied to the annotated peaks. The illustrated peaks are derived from 1 AGO-PAR-CLIP library on HEK293 cells. The brown-and-green vertical lines represent T-to-C transition sites while MREs are detected by microT-CDS algorithm. This figure has been designed for the purpose of this dissertation. | 61 |
| Figure 10: TarBase database schema. This figure has been designed for the purpose of this dissertation. | 62 |
| Figure 11: Peaks derived from 5 AGO-PAR-CLIP libraries on HEK293 cells and from 3 non-RBP background libraries are presented for T-to-C and non-T-to-C AGO-bound regions. The red-and-blue vertical lines represent T-to-C transition sites. Both types of AGO-enriched clusters are clearly distinguished from background signal. Chimeric miRNA-target fragments overlap with (non-)T-to-C peaks providing direct validation for specific miRNA-target pairs (<i>hsa-miR-19a-3p-Ran</i> and <i>hsa-miR-103a-3p-Rps14</i>). microCLIP identifies the aforementioned interactions as a 7-mer (<i>chr12:131,361,200–131,361,400, Ran</i> gene 3' UTR) and an 8-mer with a 3' compensatory site (<i>chr5:149,826,350–149,826,550, Rps14</i> gene CDS) respectively. The 3D depictions of AGO2 were based in the PDB structure 5JS1 (Paraskevopoulou MD and Karagkouni D <i>et al</i> , 2018)[17]. | 63 |
| Figure 12: Dataset collection and methodology for positive and negative MRE identification. More than 6,000 interactions were retrieved from direct techniques and miRNA-target chimeric fragments. Numerous high-throughput experimental data following specific miRNA perturbations enabled the identification of AGO bound or differentially transcribed/translated genes harboring functional binding sites. In order to resolve the exact miRNA binding sites, positive and negative instances were coupled with signal from 24 | |

AGO-PAR-CLIP libraries. The negative set was enhanced by incorporating background CLIP-Seq clusters. sRNA-Seq datasets were included to determine expressed miRNAs and accurately extract positive/negative MREs. This dataset collection was processed to form the training/test sets of microCLIP deployment (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 64

Figure 13: Overview of PARS experiment. This figure has been designed for the purpose of this dissertation. 73

Figure 14: Snapshot of the different miRNA binding types formed according to miRNA specific sub-domains. This figure has been designed for the purpose of this dissertation..... 75

Figure 15: microCLIP *in silico* framework. Separate subsets of the positive/negative miRNA interactions were used to train the distinct levels of the algorithm's modeling. 9 base classifiers in the first layer comprise characteristic feature subsets that assemble into the GBM meta-learner of the second layer. A super learning scheme is utilized in 8 of the 9 base nodes, weighing outputs from seven individual models. 'Region features' node corresponds to an RF classification scheme and consists of CLIP-sequencing-derived features. Five base models (2-6) were designed for MRE specific features: 'Binding Vectors' describe the (un)paired positions along the miRNA/MRE hybrid; 'Matches per miRNA/MRE domain' contain attributes of miRNA-target structure and sub-domains; 'Duplex Features' include free energy, secondary structure and AU base pairing features for miRNA and/or target; 'Base pairing' encompasses composition descriptors of (un)paired nucleotides; 'MRE general' incorporates general MRE-related descriptors. Three supplementary classifiers ('Feature Combination Set 1-3') comprise unique combinations of features found in base nodes 1-6 (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]. 77

Figure 16: Overview of miRNA-target positive/negative instances as identified by different indirect/direct, low and high-throughput experiments. miRNA-targeted regions derived from miRNA perturbation datasets presented an overlap with AGO-bound enriched regions from at least one CLIP-Seq library. Datasets have been combined under a tissue-specific

scheme. No overlap was allowed between positive and negative miRNA-gene interactions and their related MRE-instances..... 90

Figure 17: TarBase entries divided per methodology. Values are plotted in log2 scale. Each grid line corresponds to quadrupling of indexed miRNA interactions. a) Total miRNA-gene entries incorporated in TarBase v8.0. b) Comparison of TarBase v8.0 and TarBase v7.0 entries (Karagkouni D and Paraksevopoulou MD *et al*, 2017)[64]..... 100

Figure 18: Snapshot depicting the DIANA-TarBase v8.0 interface. Users can apply a query with miRNA and/or gene names [1] or navigate in the database content through combinations of the filtering criteria [2]. Positive/negative interactions can be refined with a series of filtering options including species, tissues/cell types, methodologies, type of validation (direct/indirect), database source, publication year as well as in silico predicted score [2]. Brief result statistics are promptly calculated [3]. Interactions can be sorted in ascending or descending order based on gene and/or miRNA names, on the number of experiments, publications and cell types/tissues supporting them [4]. Gene and miRNA details, complemented with active links to Ensembl, miRBase and the DIANA disease tag cloud, are provided [5]. Details regarding the experimental procedures such as the methodology, cell type/tissue, experimental conditions and link to the actual publication are presented [6]. Methods are color-coded, with green and red portraying validation for positive and negative regulation, respectively. Interactions are also accompanied by miRNA-binding site details [7]. Links to DIANA-miRPath functional analysis resource [8] and to an informative Help section [9] are also available. Users can navigate to the separate database statistics page [10] (Karagkouni D and Paraksevopoulou MD *et al*, 2017)[64]. 101

Figure 19: Screen-shot depicting DIANA-TarBase statistics page. The number of interactions, cell types/tissues, publications and low-/high-throughput methodologies are summarized at the top of the page [1]. A pie-chart portraying the database content per species is provided [2]. The user can select any species combination [3] to obtain relevant statistics [4]. The bar-plot [5] and tables [6] at the end of the page show the number of interactions (log2-scaled) per methodology and the cell-type/tissue frequencies respectively. They are also dynamically

| | |
|--|-----|
| populated depending on the user's choice of species (Karagkouni D and Paraksevopoulou MD <i>et al</i> , 2017)[64]. | 103 |
| Figure 20: TarBase integration in ENSEMBL. | 104 |
| Figure 21: TarBase integration in RNAcentral. | 105 |
| Figure 22: Distributions of MRE-related features corresponding to positive miRNA interactions in T-to-C and non-T-to-C AGO-bound regions against the relevant densities of negative binding sites. Assessed characteristics of positive miRNA interactions on (non-)T-to-C clusters significantly diverge from respective feature distributions of negative MREs (two-tailed Wilcoxon rank-sum test) (Paraskevopoulou MD and Karagkouni D <i>et al</i> , 2018)[17]. | 107 |
| Figure 23: Average PARS scores of AGO-bound regions deduced from the analysis of 4 EBV transformed lymphoblastoid PAR-CLIP libraries. RSS base signals were aligned to the start of the miRNA-target binding site. Base 0 corresponds to the 3'-end of the mRNA, at -1 or -2 nt downstream of the initiation of the direct miRNA seed pairing. Negative PARS scores correspond to single stranded RNA structures, while positive scores to double stranded sites. In the examined AGO-PAR-CLIP EF3D-AGO2(a), LCL-BAC-D1(b), LCL-BAC-D3(c) and LCL-BAC(d) datasets, strong structural accessibility occurs in miRNA sites identified on T-to-C (red) and non-T-to-C (green) clusters in the 2-4nt positions (yellow window) of the miRNA seed pairing. These results significantly differ from respective base scores along negative MREs (light blue) located on AGO-enriched peaks (Paraskevopoulou MD and Karagkouni D <i>et al</i> , 2018)[17]. | 108 |
| Figure 24: Evaluation of the accuracy of the 9 base model classifiers. Five-fold cross-validation has been implemented on a separate set of approximately 4,000 instances to test the performance of each node. a) ROC curve of each base model displays the classification of positive/negative miRNA binding sites. b) Distribution of base model scores estimated on positive/negative instances of the test set (Paraskevopoulou MD and Karagkouni D <i>et al</i> , 2018)[17]. | 110 |
| Figure 25: Evaluation of constitutive/internal classifiers of 5 microCLIP base models that adopt a super learning approach. Five-fold cross-validation was applied on a separate set | |

(same as in Figure 23), to test the performance of the seven individual Random Forest (RF), Generalized Linear Model (GLM), Gradient Boosting Model (GBM), Deep Learning (DL) classifiers (2 RF, 2 GBM, 2 DL, 1 GLM models) in each base node. Different colors are consistently utilized to display ROC curves of each sub-classifier incorporated in 'Binding Vectors', 'Matches per miRNA/MRE domain', 'Duplex Features', 'Base pairing' and 'MRE general' base nodes respectively. Information concerning sensitivity, specificity and AUC of each model is shown in the figure legends. The performance of ensemble deep learning models that aggregate the seven independent sub-classifiers in each base node are additionally shown (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]. 111

Figure 26: Evaluation of the accuracy of sub-classifiers included in 'Feature Combination Set 1-3' base nodes. The performance of sub-classifiers (2 RF, 2 GBM, 2 DL, 1 GLM models), along with the performance of the ensemble deep learning models that aggregate their output are displayed in distinct colors (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]. 112

Figure 27: Evaluation of microCLIP performance against 3 alternative classification approaches: a Random Forest classifier comprising all the features; a Random Forest classifier including the top 27 discriminative features ($AUC \geq 65\%$); microCLIP super learner classification scheme including top performing features per base node (70 descriptors in total, $AUC \geq 65\%$). The utilized validation set comprised 1,674 positive miRNA binding sites, derived from experimentally validated direct miRNA interactions. (a) The number of correctly predicted miRNA binding sites for each classification approach is plotted versus the total retrieved predicted sites. (b) A separate comparison captures the models' efficiency to predict correct miRNA-target interactions at different levels of total predictions. The validation set is the same as in (a) collapsed into 1,527 miRNA-gene interactions (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]. 113

Figure 28: Bar plots featuring the average miRNA-target interactions supported by non-T-to-C and/or T-to-C peaks per examined cell type and experimental condition. Mean and standard errors (error bars) of miRNA interactions are shown per library. An average increase of 14% ($\pm 8.8\%$) in the detected interactions was observed across analyzed PAR-CLIP

libraries by the incorporation of non-T-to-C clusters (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 114

Figure 29: Functional efficacy of microCLIP-detected MREs residing on T-to-C and non-T-to-C AGO-bound enriched regions. miRNA binding sites were obtained from the analysis of PAR-CLIP libraries in 3 different cell types. The functional efficiency of predicted targets was examined in 17 public gene expression profiling datasets following miRNA transfection or knockdown. Response of targeted mRNAs to miRNA perturbation experiments was evaluated independently per tested cell type, experimental technique and conditions (a-g). Cumulative distributions of mRNA fold changes for targets comprising at least one predicted MRE on T-to-C clusters or supported only by non-T-to-C peaks were compared to those that lack any site of the considered miRNAs. The number of transcripts included in each category is presented in parentheses. Identified targets supported by T-to-C and non-T-to-C clusters exert a significant difference in expression changes compared to transcripts lacking any predicted binding site (two-tailed Wilcoxon rank-sum test). At same numbers of T-to-C and non-T-to-C sites, the former group relates to more responsive targets at miRNA perturbation experiments in (b-f) (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 117

Figure 30: Functional significance of (non-)T-to-C sites in MCF7 AGO-PAR-CLIP dataset. Top 30 KEGG pathways enriched by T-to-C or (non-)T-to-C (combined T-to-C and non-T-to-C) peak containing genes. X-axis depicts number of genes enriching each term. Pathways are ranked according to the enrichment *P* value shown at the end of each bar. The T-to-C site enrichment rank is provided after pathway description to facilitate comparison with gene set of (non-)T-to-C sites (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]. 118

Figure 31: Correlation analysis of expression of pathway-related miRNA-target interactions across 271 TCGA ductal breast cancer samples (patients). Cumulative distributions of miRNA-target expression relationships, evaluated for interactions supported by T-to-C or non-T-to-C AGO-bound regions were compared to a randomly selected set from all the remaining miRNA-gene interacting pairs lacking any target site of the highly expressed miRNAs. The number of genes considered in each category is presented in parentheses.

Pathway-related miRNA-target interactions supported by T-to-C and non-T-to-C clusters reveal a significant shift towards more negative correlation coefficient values compared to the no-site distribution (two tailed Wilcoxon rank-sum test) (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 119

Figure 32: Assessment of microCLIP prediction efficacy against microCLIP T-to-C, MIRZA, microMUMMIE, PARma and Targetscan v7. miRNA-target pairs for each AGO-CLIP *in silico* approach were obtained from the analysis of 7 PAR-CLIP HEK293 libraries and functional investigation was performed by measuring mRNA responses to miRNA perturbations. Unified sets of (a) 4 microarray and (b) 2 RNA-Seq datasets, in which miRNAs were individually transfected into HEK293 cells, were included in the evaluation process. Median fold change-values (\log_2) of the top predicted targets per tested algorithm were plotted and accordingly compared by applying stepwise cutoffs on total predictions. Performed comparisons additionally incorporate a group comprising mean fold changes of 1000 randomly selected genes (without replacement) by using 100 re-samplings. microCLIP significantly outperforms all the juxtaposed implementations, detecting targets with the strongest median downregulation, from stringent to loose prediction thresholds. microCLIP T-to-C also exhibits greater efficacy than the rest *in silico* approaches (range of P values microarrays: 0 - 2.2×10^{-7} , P values RNA-Seq: 5.5×10^{-265} - 3.6×10^{-29} , two-tailed Wilcoxon signed-rank test, $535 < n_{\text{microarrays}} < 3,223$, $174 < n_{\text{RNA-Seq}} < 1,613$), (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 121

Figure 33: microCLIP performance compared to MIRZA, microMUMMIE, PARma and Targetscan v7 was examined in 7 public gene expression profiling datasets following miRNA transfection or knockdown in HEK293 and HeLa cell lines. miRNA-target interactions for AGO-CLIP *in silico* approaches were obtained from the analysis of PAR-CLIP HEK293 and HeLa libraries. Response of targeted mRNAs to miRNA perturbation experiments was evaluated independently per tested cell type, experimental technique and condition (a-g). Cumulative distributions of mRNA fold changes for targets comprising at least one predicted MRE in the CDS or 3' UTR regions were compared to those that lacked any site of the considered miRNAs (one-sided Kolmogorov-Smirnov test). Functional efficacy was assessed

for equal numbers of top predictions per implementation. Implementations that did not support targets with a fold-change in the examined miRNA perturbation experiments were not included in the relevant cumulative plots. (a-f) Identified targets by microCLIP revealed greater site effectiveness than the rest AGO-CLIP-guided implementations. (g) microCLIP performed similarly as PARma and better than the rest of implementations. Targetscan v7 identifies responsive targets, operating on par with *in silico* approaches based on CLIP data such as PARma, while in (c-d) and (g) it displays analogous efficacy as microCLIP. The number of transcripts included in each comparison is denoted in the parentheses (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]..... 123

Figure 34: Evaluation of microCLIP performance against microCLIP T-to-C, MIRZA, microMUMMIE, PARma, Targetscan v7 (all predictions) and Targetscan v7 conserved predicted sites. The utilized validation set comprised 1,674 positive miRNA binding sites of 125 miRNAs, derived from chimeric miRNA-target fragments and direct miRNA bindings supported by Reporter Gene Assays. The number of correctly predicted miRNA binding sites for each implementation is plotted versus (a) the total retrieved predictions, (b) the top scored miRNA binding site per AGO-bound enriched region. In (a) and (b) comparisons, we restrict each program's predictions on PAR-CLIP clusters overlapping the validation test set. A separate comparison (c) captures algorithms' efficiency to predict correct miRNA-target interactions at different levels of total predictions. The validation set is the same as in (a-b) evaluations, collapsed into 1,527 miRNA-gene interactions. For the latter comparison, seed-baseline methods were operating in the absence of AGO-CLIP data, while CLIP-guided implementations on PAR-CLIP clusters overlapping full transcript regions (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17]. 125

Figure 35: ROC curves of sequence accessibility parameters for the classification of positive/negative miRNA binding sites, i.e. accessibility of the 20nt miRNA binding region and the 30nt region upstream/downstream of the MRE..... 126

Figure 36: ROC curves for the classification of positive/negative miRNA binding sites indicating the a) aggregated MRE seed binding conservation, b) aggregated conservation in

the upstream region of the MRE, c) minimum duplex structure energy and d) MRE-related thermodynamic properties. 127

Figure 37: ROC curves for the classification of positive/negative miRNA binding sites indicating AU base pairs (MRE, seed), seed matches and mismatches per miRNA-target duplex domain, nucleotide and dinucleotide MRE content and binding type. The latter feature comprises an extended set of (non-)canonical miRNA base pairings where smaller values indicate stronger seed matches (9mer to 6mer) and greater values correspond to non-canonical and 3' supplementary sites. 128

Figure 38: Distributions of MRE-related features corresponding to positive miRNA-target pairs against the relevant densities of negative binding sites. The descriptors present higher performance in microT-training set compared to microCLIP-training set. Evaluated descriptors include length of target bulges, start of the binding in the MRE region relative to miRNA binding anchors upon duplex formation, AU base pairs in 3' supplementary region, GC base pairs in tail MRE region, total mismatches per miRNA-target duplex and dinucleotide MRE content. Assessed characteristics of positive miRNA interactions significantly diverge from respective feature distributions of negative MREs (two-tailed Wilcoxon rank-sum test). 129

Figure 39: Evaluation of the accuracy of the 9 base model classifiers. Five-fold cross-validation has been implemented on a separate set of approximately 6,192 instances to test the performance of each node. a) ROC curve of each base model displays the classification of positive/negative miRNA binding sites. b) Distribution of base model scores estimated on positive/negative instances of the test set. 131

Figure 40: Evaluation of microT performance against 4 alternative Super Learning (SL) classification approaches: a model incorporating the same classifiers with microT and without 3 base nodes; a model consisting only Deep Learning classifiers (DL) in the 1st layer; a model combining Deep Learning and Random Forest (RF) classifiers in the 1st layer; a model combining Deep Learning and Random Forest classifiers in the 1st layer and without Base Pairing node. The utilized set comprised 2,092 experimentally validated direct miRNA

binding events (1,805 chimeric fragments and 287 reporter-assay verified), corresponding to 2,032 unique miRNA-gene interactions. (a) The number of correctly predicted miRNA-target interactions for each classification approach is plotted versus the mean prediction per miRNA. (b) A separate comparison captures the models' efficiency to predict correct miRNA binding events at different levels of total predicted sites. 132

Figure 41: Evaluation of microT performance against 7 alternative Deep Learning models. The utilized set comprised 2,092 experimentally validated direct miRNA binding events (1,805 chimeric fragments and 287 reporter-assay verified), corresponding to 2,032 unique miRNA-gene interactions. (a) The number of correctly predicted miRNA-target interactions for each classification approach is plotted versus the mean prediction per miRNA. (b) A separate comparison captures the models' efficiency to predict correct miRNA binding events at different levels of total predicted sites. 133

Figure 42: microT Super Learning performance compared to microT-CDS and Targetscan v7 was examined in 5 public gene expression profiling datasets following miRNA transfection or knockdown in different cell types. Cumulative distributions of mRNA fold changes for targets comprising at least one predicted MRE in the CDS or 3' UTR regions were compared to those that lacked any site of the considered miRNAs (one-sided Kolmogorov-Smirnov test). Functional efficacy was assessed for equal numbers of top predictions per implementation. (a-d) Identified targets by microT revealed greater site effectiveness than the rest de novo approaches. (e) microT performed similarly as microT-CDS and better than Targetscan v7. The number of transcripts included in each comparison is denoted in the parentheses..... 135

Figure 43: Evaluation of microT Super Learning model performance against microT-CDS and Targetscan v7. The utilized set comprised 2,092 experimentally validated direct miRNA binding events (1,805 chimeric fragments and 287 reporter-assay verified), corresponding to 2,032 unique miRNA-gene interactions. (a) The number of correctly predicted miRNA-target interactions for each classification approach is plotted versus the mean prediction per

| | |
|---|-----|
| miRNA. (b) A separate comparison captures the models' efficiency to predict correct miRNA binding events at different levels of total predicted sites | 136 |
|---|-----|

List of Tables

| | |
|---|----|
| Table 1: Non-coding RNA subfamilies – Their Function and Size. | 34 |
| Table 2: De novo miRNA Target Prediction algorithms, published the last decade. | 39 |
| Table 3: Experimental methodologies for miRNA:gene interactions characterization. | 43 |
| Table 4: Summary of the collected experiments in human species upon specific miRNA deregulation. The datasets were utilized to extract independent training and test sets of positive and negative MRE regions for microCLIP deployment. | 65 |
| Table 5: Summary of microarray and RNA sequencing experiments in human species upon specific miRNA deregulation, utilized in benchmarking evaluations of microCLIP model. | 68 |
| Table 6: Summary of the collected AGO-PAR-CLIP experiments in human species, obtained from 9 studies. These datasets provided the source of PAR-CLIP signal (raw reads and transitions) which was integrated with experimentally validated positive/negative instances of miRNA-targeted regions. | 69 |
| Table 7: Description of small RNA-Seq datasets of similar cell types to PAR-CLIP libraries, analyzed to infer expressed miRNAs. The table displays the source of small RNA-Seq libraries along with its ID, cell type, condition and description. | 71 |
| Table 8: Description of RNA-Seq datasets of similar cell types to PAR-CLIP libraries, analyzed to infer expressed transcripts. The table displays the source of RNA-Seq datasets along with its ID, cell type, condition and description. | 72 |
| Table 9: Description of the binding types supported by microCLIP. | 76 |
| Table 10: Description of features incorporated in microCLIP. | 78 |
| Table 11: Summary of training/test sets utilized for microCLIP deployment. | 84 |
| Table 12: Summary of the collected experiments in human species upon specific miRNA deregulation. The datasets were utilized to extract a training set of positive and negative MRE regions for microT deployment. | 86 |

| | |
|---|----|
| Table 13: Summary of the associations regarded between cell types and tissues for the extraction of miRNA-targeted regions incorporated in training/test sets. | 89 |
| Table 14: Overview of miRNA-target positive/negative instances utilized in training set as identified by different indirect/direct, low and high-throughput experiments. | 91 |
| Table 15: Summary of microarray experiments in human species upon specific miRNA deregulation, utilized in benchmarking evaluations of microT Super Learning model. | 91 |
| Table 16: Description of features incorporated in microT. | 92 |
| Table 17: Summary of training set utilized for microT deployment. | 96 |
| Table 18: Summary of miRNA-target instances, located on 3' UTR and CDS regions, utilized in the training/test of microT model. | 96 |
| Table 19: TarBase v8.0 Entries. Statistics regarding the total entries, miRNA-gene interacting pairs derived from low-/high-throughput methodologies, distinct cell types/tissues and curated publications are provided. The number of analyzed datasets and unique studied conditions are presented for high-throughput experiments. The incorporated low-/high-throughput experimental techniques, as well as interface improvements are reported. Newly incorporated experimental methods and interface advancements are marked as bold. | 98 |

CHAPTER 1

Introduction

1.1 ncRNAs – from “junk” DNA into a research goldmine

The term non-coding RNAs (ncRNAs) is commonly employed for RNAs that do not encode proteins. However, this does not confirm that these RNAs do not have a function or do not play a fundamental role in cellular processes. The traditional view of molecular biology is that almost exclusively RNAs transfer genetic information in order to be subsequently translated into protein. However, the discovery of families of ncRNAs, such as ribosomal RNA (rRNA) and transfer RNA (tRNA), comprising a high portion of total RNA and serving necessary organisms functions, broadened the long-established RNA role. The majority of mammalian genomes and other complex organisms are transcribed into ncRNAs and seem to play a key regulatory role in various physiological and pathological processes[1].

ncRNAs are sub-divided according to their size and their biological function (Table 1). There are ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), >200 nucleotide-long non-coding RNAs, also known as long non-coding RNAs (lncRNAs), small non-coding RNAs, such as microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs) etc. These categories also display sub-groups according to the genomic regulatory regions the ncRNAs originate from. ncRNAs may derive from intergenic, intragenic, intronic regions of protein coding genes or even from pseudogenes[2].

Emerging technological developments during the past decade have revolutionized biomedical research. Extensive sequencing experiments produced by large consortia, including the Encyclopedia of DNA Elements Consortium (ENCODE)[3, 4] enabled large scale analyses in the “regulatory RNA” field and turned non-coding RNA, initially considered as junk, into a research goldmine. Numerous high-throughput experiments suggest that ncRNAs partake crucially in a remarkable variety of biological processes, such as gene expression, editing, splicing, heterochromatin formation, histone modification, DNA methylation etc[1].

Table 1: Non-coding RNA subfamilies – Their Function and Size.

| Category | Definition | Function | Size |
|---------------|-----------------------|---|---|
| miRNA | microRNA | Small ncRNA that interacts with (non-)coding RNAs through the RISC complex to induce target cleavage/degradation or translational suppression | ~22nt |
| piRNA | piwi-interacting RNA | ncRNA mainly characterized in the male germline - directs chromatin modification to repress transcription | 27nt |
| siRNA | small interfering RNA | Product of Dicer cleavage of dsRNA that targets RNAs to induce their cleavage | ~22nt |
| snRNA | small nuclear RNA | ncRNA localized in the eukaryotic cell nucleus | 100-300nt |
| snoRNA | small nucleolar RNA | Guide RNA of chemical modifications of other RNAs | 70nt |
| sRNA | small RNA regulator | Bacterial ncRNA that interacts with mRNAs and regulate gene expression | <300nt |
| rRNA | ribosomal RNA | RNA component of the ribosomal subunit | 120,160,1868, 5025nt, human; 120,1541, 2904nt, <i>E. coli</i> |
| tRNA | transfer RNA | Facilitates protein synthesis by carrying amino-acids to ribosomal units | 70-90nt |
| lncRNA | long non-coding RNA | Transcribed ncRNA, often capped and polyadenylated. Epigenetic gene expression regulator, sponge, transporter | >200nt |

1.2 The discovery of microRNAs

miRNAs are small non-coding RNA molecules, approximately 22 nucleotides long. They are central post-transcriptional regulators of gene expression and play a pivotal role in numerous biological processes. For more than a decade, miRNAs are intensively researched for their involvement in a variety of physiological and pathological conditions[5].

The first microRNAs were discovered in 1993 by *C. elegans* (*Caenorhabditis elegans*)[6] by Ambros, Lee and Feinbaum. The researchers observed that the *lin-4* gene produced a non-coding RNA segment of approximately 22 bases long that binds to the 3'-untranslated region (3' UTR) of *lin-14* mRNA. The interaction between the *lin-4* non-coding and *lin-14* gene led to the translational repression of the latter. The above phenomenon is amplified by another research result in the *C. elegans* organism, where the *let-7* microRNA was identified to target the 3' UTR region and induce suppression of *lin-41* gene expression[7]. *Let-7* microRNA appeared to be conserved in other organisms supporting the existence and regulatory role of other small non-coding RNA molecules[8]. These first discoveries were the beginning of a large number of findings for novel microRNAs in various organisms that have established their function as regulators of gene expression.

1.2.1 Biogenesis of microRNAs

More than 45% of miRNAs are derived from non-coding transcripts, while the rest are transcribed from protein coding regions. The majority of miRNA genes are transcribed from RNA polymerase II (Pol II), generating large primary transcripts (pri-miRNAs). The protein Drosha processes the pri-miRNAs generating ~60-100 bases long hairpin structures, also known as pre-miRNA precursors. Rapid cleavage of pri-miRNAs by Drosha in the nucleus prevents their characterization by conventional sequencing techniques, raising limitations to the clarification of the regulatory mechanisms that control their transcription. The precursor sequences are extracted from the nucleus and transferred to the cytoplasm by means of the exportin-5 and Ran-GTP proteins, which *inter alia* participate in the transport of molecules inside and outside the nuclear membrane. After the pre-miRNA comes out of the nucleus, they are cut with the help of the Dicer enzyme, a highly conserved protein found in most eukaryotic organisms. The effect of the Dicer enzyme by cutting the loop at the end of the microRNAs precursors leads to the release of double-stranded ~22 nucleotide microRNAs[9] (Figure 1).

Both strands of the miRNA duplex-intermediate can be potentially functional. However, usually one strand (guide strand) accumulates as the mature miRNA. The mature single-stranded molecule is loaded into protein Argonaute (AGO) while the other strand, termed as the “passenger” strand, is released and degraded. The main action of microRNAs is observed in the cytoplasm, while recent studies indicate that some mature miRNAs are able to re-enter into the nucleus and interact with other transcripts, something that displays their possible involvement in additional mechanisms that have not been yet characterized[10].

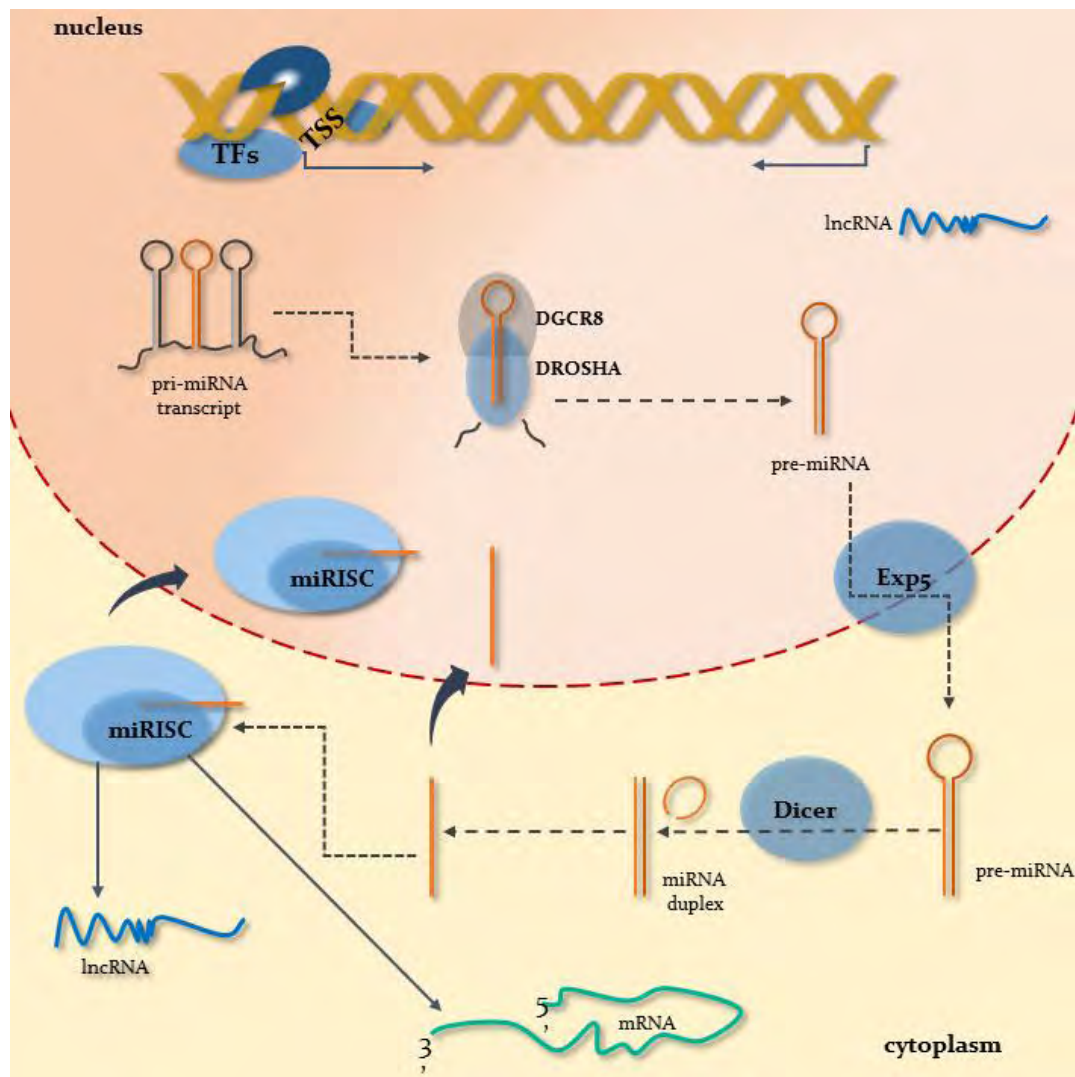


Figure 1: Biogenesis of microRNAs. miRNAs are transcribed into the nucleus either autonomously or as polycistronic molecules. The miRNA precursor (pri-miRNA) is treated by the microprocessor complex which is composed of the DROSHA and DCGR8 proteins. The resulting pre-miRNA which is extracted from the nucleus by means of Exportin-5 and protein Dicer cuts the loop at the end of miRNA precursor. From the generated double-stranded miRNA, one clone is usually selected, which is incorporated into the RISC complex. The most well-defined functions of miRNAs are observed in the cytoplasm such as: translation suppression and/or degradation of the mRNA target. Recent studies suggest that some mature miRNAs are able to re-enter into the nucleus and interact with other transcripts, something that displays their possible involvement in additional mechanisms that have not been yet characterized. This figure has been designed for the purpose of this dissertation.

1.2.2 microRNA function

miRNAs are loaded into protein Argonaute and interact with the RISC complex to form the miRNA-induced silencing complex (miRISC). Since miRNAs are incorporated into the RISC complex, they induce gene silencing with partial or full complementary binding with mRNAs (Figure 2). In particular, interactions of miRNAs with target mRNAs require complementarity

of 6-8 nucleotides, the so-called seed region at the 5' end of the miRNA. It should be noted that the base pairing of the seed with the mRNA plays a very important role in the effectiveness of the interaction.

Initially, miRNAs were demonstrated to systematically and effectively target the 3' untranslated region (3' UTRs) of mRNA, where highly conserved miRNA Recognition Elements (MREs) are identified. However, recent studies have shown new functional miRNA target sites within the 5'-Untranslated Region (5' UTR) and the coding region (CDS) of the mRNA[11].

At the same time, miRNAs play a key regulatory role in a variety of biological processes such as stem cell differentiation, involvement in immune mechanisms and cell signaling. Beyond their physiological role, a large number of studies address the positive or negative role of miRNAs in various diseases. miRNAs affect the expression levels of genes in different tissues. Consequently, possible changes in miRNA concentration by mutation, deletion, amplification, and epigenetic silencing or transcription factors, affect targeted genes, including oncogenes and tumor suppressors, involved in a wide range of pathological conditions in the human body, such as carcinogenesis, cardiovascular diseases, metabolic disorders, autoimmune diseases, etc.[12-16]. miRNAs are therefore intensively studied for their potential as therapeutic targets.

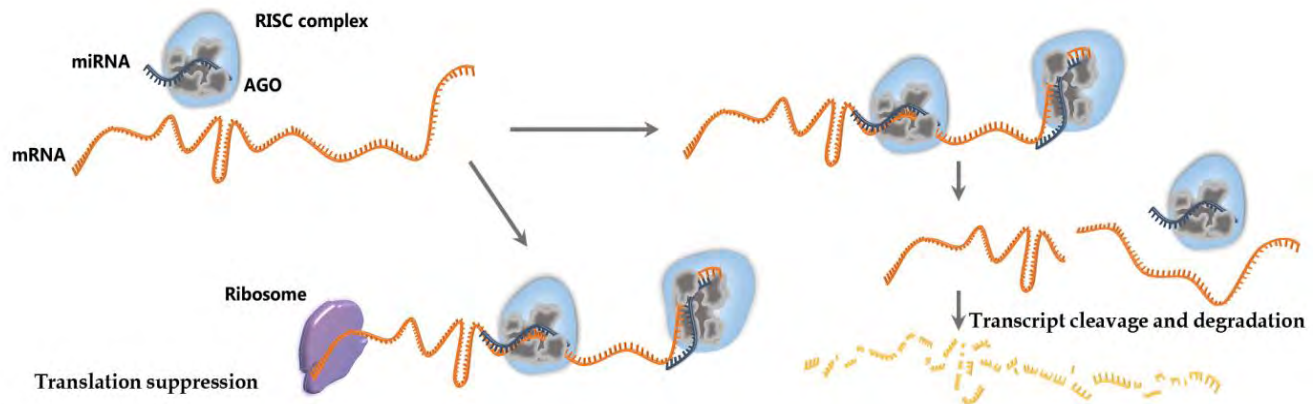


Figure 2: Illustration of miRNA targeting. miRNAs are loaded on AGO and guide the RISC complex to target MRE(s). RISC binding to its target genes can either cease their translation or induce their cleavage and/or degradation (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

1.3 Identification of miRNA targets

Accurate characterization of miRNA targets is considered fundamental to elucidate their regulatory roles. Over the last 15 years, a multitude of *in silico* and experimental procedures have been developed aiming to determine the miRNA interactome[5]. Currently, high-throughput techniques have enabled the identification of novel experimentally-supported

miRNA-gene interactions in a transcriptome-wide scale[18]. The broad use of these experimental methodologies has advanced miRNA target recognition towards the gradual substitution of related computational approaches. Despite the contribution of experimental methods and computational techniques, much of the microRNA targets, even for extensively studied organisms, such as mouse and human, remain unexplored.

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

Target prediction tools constituted the first *in silico* approaches in miRNA research. miRNAs may occupy hundreds of thousands of potential target sites, while their validation with experimental procedures is time consuming and costly. Computational approaches constitute the backbone of miRNA related studies by facilitating the process and proposing potential target sites for downstream analyses.

The first miRNA target prediction algorithm was published in 2003 by Lewis *et al.*[19], who first introduced the concept of the “seed region”. The miRNA “seed region” is a 7 base-long segment, between the 2nd and 8th nucleotide, counting from the 5′ of a miRNA sequence. This region showed perfect Watson-Crick complementarity with the 3′ UTR of the target mRNA and was highly conserved among miRNAs and species. Since then, several miRNA target prediction algorithms have been developed and heavily rely on the complementarity of this region with the respective binding site, as a key biological element for miRNA-target prediction.

Most of the developed algorithms focused from the very beginning on the prediction of miRNA binding sites solely on the 3′ UTR of mRNAs. However, recent advances in high-throughput sequencing revealed a significant portion of target sites in CDS[11]. Currently, there are numerous widely used and promising applications for de novo identification of miRNA-gene interactions. Most of them rely on decisive features for miRNA target recognition, such as nucleotide composition of the binding site, thermodynamic stability, secondary structure and evolutionary conservation. They often produce radically different outcomes due to the incorporation of diverse experimental data and different mathematical models, utilized for the deployment of each algorithm. Therefore, selecting the most appropriate implementation is a common and multifaceted problem.

1.3.1.1 Overview of de novo miRNA Target Prediction Algorithms

Available de novo miRNA Target Prediction algorithms, published in the last decade, are displayed in Table 2 and a concise description of the most widely used and recently developed methods, is reviewed below in more detail.

Table 2: De novo miRNA Target Prediction algorithms, published the last decade.

| Algorithm | URL | Year of the latest update |
|----------------------------|--|---------------------------|
| miRAW | bitbucket.org/account/user/bipous/projects/MIRAW | 2018 |
| DeepMirTar | github.com/Bjoux2/DeepMirTar_SdA | 2018 |
| chimiRic | bitbucket.org/leslielab/chimiric | 2016 |
| MIRZA-G | www.clipz.unibas.ch/index.php?r=tools/sub/mirza_g | 2015 |
| PACCMIT/PACCMIT-CDS | paccmit.epfl.ch | 2015 |
| Targetscan | www.targetscan.org | 2015 |
| MBSTAR | www.isical.ac.in/~bioinfo_miu/MBStar30.htm | 2014 |
| mirMark | github.com/lanagarmire/MirMark | 2014 |
| miRmap | cegg.unige.ch/mirmap | 2013 |
| DIANA-microT-CDS | www.microrna.gr/microT-CDS | 2012 |
| MiRanda/mirSVR | www.microRNA.org | 2010 |

TargetScan[20]: TargetScan is a model with high performance in terms of sensitivity and precision. The first version of the algorithm was introduced in 2003 by Lewis et al. [19] and since then is constantly updated. Targetscan v7 provides a quantitative model that incorporates 14 distinct features, including the target site type, 3' supplementary pairing, local AU content, 3' UTR binding site abundance, predicted seed-pairing stability and conservation. It mainly detects canonical (high level of seed complementarity) sites within 3' UTR regions, according to a seed-dependent scoring system. The latest version of the model also predicts effective non-canonical site types, such as 3' compensatory sites and centered sites. The updated context++ model is applicable to all canonical sites, independently to the evolutionary conservation feature, evaluating not only non-conserved sites to conserved miRNAs but also sites for non-conserved miRNAs, including viral miRNAs. Each target site can be evaluated with a cumulative context and/or an aggregated conservation score. The training and testing of the model was performed on 74 microarray datasets, analyzed from scratch to minimize technical biases, with clear sRNA-induced repression using stepwise regression[21].

DIANA-microT-CDS[22, 23]: DIANA-microT-CDS is a state-of-the-art implementation which identifies seed-based miRNA binding sites with perfect or partial complementarity, both in CDS and 3' UTR regions. It achieves increased performance in terms of sensitivity and precision due to the independent analysis and the distinct feature extraction performed for CDS and 3' UTR regions. Important microT-CDS features are the target site complementarity, upflank AU content, accessibility, pairing stability and conservation of miRNA targeted regions in 30 and 16 species respectively. A dynamic programming algorithm identifies the optimal alignment between the miRNA extended seed sequence (nucleotides 1–9 from the 5' end of the miRNA) and every 9 nt window on the 3' UTR or the CDS region. Positive and negative instances are derived from PAR-CLIP data[11]. The separate prediction models are

combined in a Generalized Linear Model, which is trained on microarray datasets that measure mRNA expression changes after transfection or knockout of a specific miRNA. The potency of each miRNA-gene interaction is described by a combined score that represents the synergistic action of multiple binding sites in the targeted mRNA regions. The overall performance of the algorithm is estimated on quantitative proteomics and HITS-CLIP data[24].

miRAW[25]: miRAW is one of the more recently developed de novo miRNA target prediction algorithms. Its core algorithm identifies (non-)canonical sites within the 3' UTR region. Decisive features of the model include miRNA:target hybrid stability, site accessibility and per-nucleotide base pairing composition. The model adopts a Deep Learning classification scheme of eight dense hidden layers, while the output layer is composed of two softmax nodes. It was trained and tested on experimentally validated miRNA:gene interactions indexed on TarBase v7[26] and mirTarBase[27] repositories. The miRNA:target pairs were further combined with AGO-PAR-CLIP[28] and CLASH[29] experiments to retrieve the exact miRNA binding locations. The model also integrates evolutionary conservation of targets by combining broadly conserved sites from Targetscan[20]. The performance of the algorithm was further evaluated on microarray datasets after miRNA transfection into HeLa cells[30].

DeepMirTar[31]: DeepMirTar model is another Deep Learning approach that was recently developed. The algorithm predicts (non-)canonical miRNA target sites within the 3' UTR region. It incorporates 7 categories of features including sequence composition, duplex free energy, site location, accessibility and evolutionary conservation. It also integrates hot-encoding features representing the per-base nucleotide composition of miRNA binding regions. The model has been trained on experimentally validated miRNA-gene interactions derived from miRecords[32] database and a CLASH[29] experiment, while mock miRNA-gene pairs were included as negative regions. DeepMirTar performance was evaluated on a separate AGO-PAR-CLIP dataset[11].

chimiRic[33]: chimiRic detects seed-based miRNA-target pairs within the 3' UTR region, with perfect or partial complementarity, by adopting a tissue-specific scheme. To address the possibility of cell type specific miRNA binding, the model applied a multi-task learning approach by treating the different cell types separately with related learning tasks. It integrates decisive features, such as base pairing composition, duplex structure and 3' UTR related characteristics. chimiRic utilizes a Support Vector Machine approach while positive and negative miRNA binding sites were extracted from CLASH[29] and AGO-CLIP-Seq[11].[34] experiments. The model was subsequently evaluated on AGO-PAR-CLIP, CLEAR-CLIP[35] datasets and Reporter Gene Assay experiments.

MIRZA-G[36]: MIRZA-G is another tool able to predict seed-based canonical and non-canonical miRNA binding sites, residing on 3' UTR region and siRNA off-targets. Decisive

features for MIRZA-G are the nucleotide composition around putative targeted regions, the site structural accessibility, the evolutionary conservation and the location of the site within the 3' UTR region. miRNA binding affinity in mRNA regions is assessed by the MIRZA biophysical model deduced from AGO-CLIP-Seq data[37]. The latter implementation assigns base binding energies on the candidate miRNA-mRNA duplexes. The training and testing of the algorithm was performed using a generalized linear model against 26 miRNA/siRNA transfection microarray and proteomics datasets.

miRanda/mirSVR[38, 39]: miRanda is the target prediction model provided by microRNA.org. Its core algorithm identifies putative miRNA:gene interactions which are scored by mirSVR model. It provides both canonical and non-canonical miRNA binding sites within the 3' UTR region, by permitting one G:U wobble pair or mismatch in the 6mer seed region, followed by a perfect binding in the 3' compensatory region. mirSVR utilizes a Support Vector Regression approach and is trained on miRNA transfection microarray experiments performed on HeLa cells. The scoring scheme is based on local and global features. Local features incorporate the AU sequence composition and the accessibility of the target site, while global features refer to UTR-relevant features and the conservation level of the targeted region. The performance of miRanda-mirSVR joint usage was assessed on microarray, proteomics and AGO-IP datasets after miRNA perturbation and AGO-PAR-CLIP experiments.

mirMark[40]: mirMark provides both canonical and non-canonical miRNA binding sites in the 3' UTR region, allowing up to 2 G:U wobble pairs. The main characteristic of the algorithm is the extensive list of site and UTR relevant features that incorporates. Site level features refer to miRNA/mRNA duplex energy, complementarity, structural accessibility, composition and evolutionary conservation. The initial identification of candidate miRNA binding sites is performed with miRanda algorithm. mirMark adopts separate levels of classification, trained with a random forest model; a first one for the assessment of the target site and a second one for the evaluation of the miRNA-gene interaction. The training was performed using experimentally verified miRNA-gene targets derived from miRecords[32] and miRTarBase[27], while mock miRNA-gene pairs were included as negative targeted regions. AGO-PAR-CLIP data were used for the evaluation of mirMark's performance.

mBSTAR[41]: mBSTAR constitutes a learning framework designed for predicting seed-based binding sites of miRNAs within the 3' UTR region, allowing a single G:U wobble pair. It incorporates 40 sequence, structural and energy features, including nucleotide frequencies, internal loops, bulges and minimum free energy of the entire flanking region. mBSTAR utilizes a Random Forest classifier, while the training and testing was performed on experimentally supported miRNA-gene targets derived from miRecords[32], Tarbase v6[42] and StarBase[43].

PACCMIT/PACCMIT-CDS[44]: PACCMIT algorithm (Prediction of the Accessible MicroRNA Targets) is based on an overrepresentation ranking system. The original model ranks the candidate seed-based miRNA binding sites which reside on 3' UTR regions, according to their over-representation with respect to a random background. The ranking system is based on a Markov model. The sites are subsequently filtered by considering accessibility and evolutionary conservation. PACCMIT-CDS follows the aforementioned scheme by searching potential miRNA bindings also in the CDS region. The model was tested on AGO-PAR-CLIP data[11] and proteomics experiments, followed by miRNA transfection[45][46].

1.3.2 Experimental Methods for the identification of miRNA:gene interactions

The experimental techniques, utilized to identify novel miRNA targets and validate predicted interactions, can significantly differ in their accuracy and robustness. They are mainly divided into low- and high-throughput experiments according to the amount of information they produce. In low-throughput techniques, Reporter Gene Assays focus on the recognition of the exact miRNA binding location, while indirect methodologies like quantitative Polymerase Chain Reaction (qPCR), Western blot and Enzyme-Linked Immunosorbent Assay (ELISA) infer interactions by taking into consideration the reduction of mRNA or protein concentration[47]. High-throughput techniques, such as microarrays and proteomics are the extension of low-yield methodologies, enabling the indirect detection of numerous miRNA targets. Current advancements in Next Generation Sequencing (NGS) technologies have radically changed the characterization of the miRNA interactome[18]. RNA immunoprecipitation combined with sequencing (RIP-Seq) constitutes one of the first experiments to enable the identification of RNAs bound by a protein of interest[48]. Recently, Ribosome profiling sequencing (RPF-Seq) experiments have been proposed as a sensitive and quantitative protocol, able to measure the efficiency and speed of translation, as well as the ribosome occupancy per transcript. This methodology allows the evaluation of miRNA-mediated translational repression by the analysis of captured ribosome-bound transcripts[49]. These procedures are coupled with overexpression or knockdown of a specific miRNA in order to detect genes quantitatively affected by miRNA expression perturbations. Crosslinking and immunoprecipitation sequencing (CLIP-Seq) methodologies focus on the transcriptome-wide recognition of RNA-protein binding regions and are usually complemented with RNA expression experiments[50]. AGO-CLIP-Seq methodologies inaugurated a new era in miRNA research, providing unprecedented accuracy and multitude of miRNA targets in a transcriptome-wide scale. Recent modified versions of the later techniques, such as CLEAR-CLIP(Covalent Ligation of Endogenous Argonaute-bound RNAs)[35] and CLASH(Crosslinking, Ligation, and Sequencing of Hybrids)[51] protocols,

include an extra ligation step which links miRNA molecules with their respective target binding site, resulting in hundreds of chimeric miRNA-mRNA fragments.

Table 3 summarizes the most widely used experimental methodologies for miRNA target characterization.

Table 3: Experimental methodologies for miRNA:gene interactions characterization.

| Method | Direct technique | Throughput | Experiment context |
|--|------------------|------------|--|
| Reporter Gene Assay | ✓ | Low | Identification of interacting miRNA-gene regions |
| qPCR, Northern Blot | - | Low | miRNA effect on mRNA levels |
| Western Blot, ELISA | - | Low | miRNA effect on protein abundance |
| Microarrays, RNA-Seq | - | High | miRNA effect on mRNA expression |
| CLIP-Seq/CLASH/ CLEAR-CLIP | ✓ | High | MRE binding site sequencing |
| 3LIFE | ✓ | High | High-throughput Reporter Gene Assay |
| RPF-Seq | - | High | Sequencing of actively translated transcripts |
| Biotin miRNA tagging (Biotin-Seq, Biotin-Microarrays, Biotin-qPCR) | - | High/Low | Biotin-tagged miRNA pull down followed by RNA-Seq/Microarrays/qPCR |
| Quantitative Proteomics | - | High | miRNA effect on protein abundance |
| AGO-IP/RIP-Seq | - | High | Enriched transcripts in AGO immunoprecipitates |
| miTRAP | - | High | miRNA trapping by RNA baiting |
| IMPACT-Seq | - | High | Biotin-tagged miRNA pull down |
| PARE/Degradome-Seq | - | High | Cleaved mRNA targets |
| LAMP | - | High | Labeled miRNA pull-down with digoxigenin |

1.3.2.1 AGO-CLIP-Seq experimental methodologies

CLIP-Seq methodologies have revolutionized the study of protein-RNA interactions by enabling the accurate characterization of RNA binding protein (RBP) target sites on a transcriptome-wide scale in different species under psychological or pathological conditions. The inception of the first and original CLIP-Seq protocol was conceived by Ule L et al[52] in 2003 and since then several CLIP-Seq variants have been developed. Photoactivatable

Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) and High-throughput Sequencing of RNA Isolated by Crosslinking Immunoprecipitation (HITS-CLIP) variants against protein AGO are widely used methodologies for miRNA targetome characterization. The last decade, these experiments have been performed to map miRNA-gene interactions on a transcriptome-wide scale for healthy or diseased cell types/tissues and have provided valuable insights into miRNA regulation of pathogen infections and cancer[53, 54]. They are considered among the most powerful high-throughput methods for the characterization of miRNA targets.

The experimental protocol adopted in PAR-CLIP and HITS-CLIP methodologies is summarized in Figure 3. In brief, the implemented steps of the procedures are mentioned below:

1. Protein–RNA complexes are covalently crosslinked in live cells or tissues.
2. Cells/tissues are lysed and treated with RNase leaving small fragments of RNA molecules bound with the protein of interest.
3. Protein–RNA complexes are immunoprecipitated, and non-specific RNAs and proteins are removed by stringent washes.
4. Ligation of the radioactively labeled 5' adapter is performed, while protein-RNA products are attached to beads, allowing the removal of unligated 5' adapter.
5. The purified protein–RNA complexes are radioactively labeled and separated by SDS-PAGE.
6. Bound RNA is isolated either directly from SDS-PAGE gels or from nitrocellulose membranes following transfer by Proteinase K treatment.
7. Eluted RNA is ligated to adapters, reverse transcribed while the resulting cDNA is PCR amplified and subjected to sequencing.
8. Sequencing reads are processed and mapped to reference genomes. Computational steps are following for CLIP-Seq analysis.

HITS-CLIP relies on UV crosslinking of protein–RNA complexes at UV 254 nm. The resulting library usually contains substitutions or deletions at the crosslinking site, induced by reverse transcriptase, facilitating the downstream analysis.

A major difference of PAR-CLIP against HITS-CLIP protocol is the use of 4-thiouridine (4SU) and 6-thioguanosine (6SG) analogs that significantly enhance the efficiency of protein–RNA crosslinking. In PAR-CLIP experiments, cells are typically grown in the presence of ribonucleoside analogs for up to 16 h and UV crosslinking is achieved at UV 365 nm. This procedure limits the application of PAR-CLIP experiment only to cell cultures.

The analogs incorporation provokes T-to-C (4SU) and G-to-A (6SG) substitutions at the crosslinking site during cDNA synthesis, an incident that allows the accurate mapping of protein RNA targets.

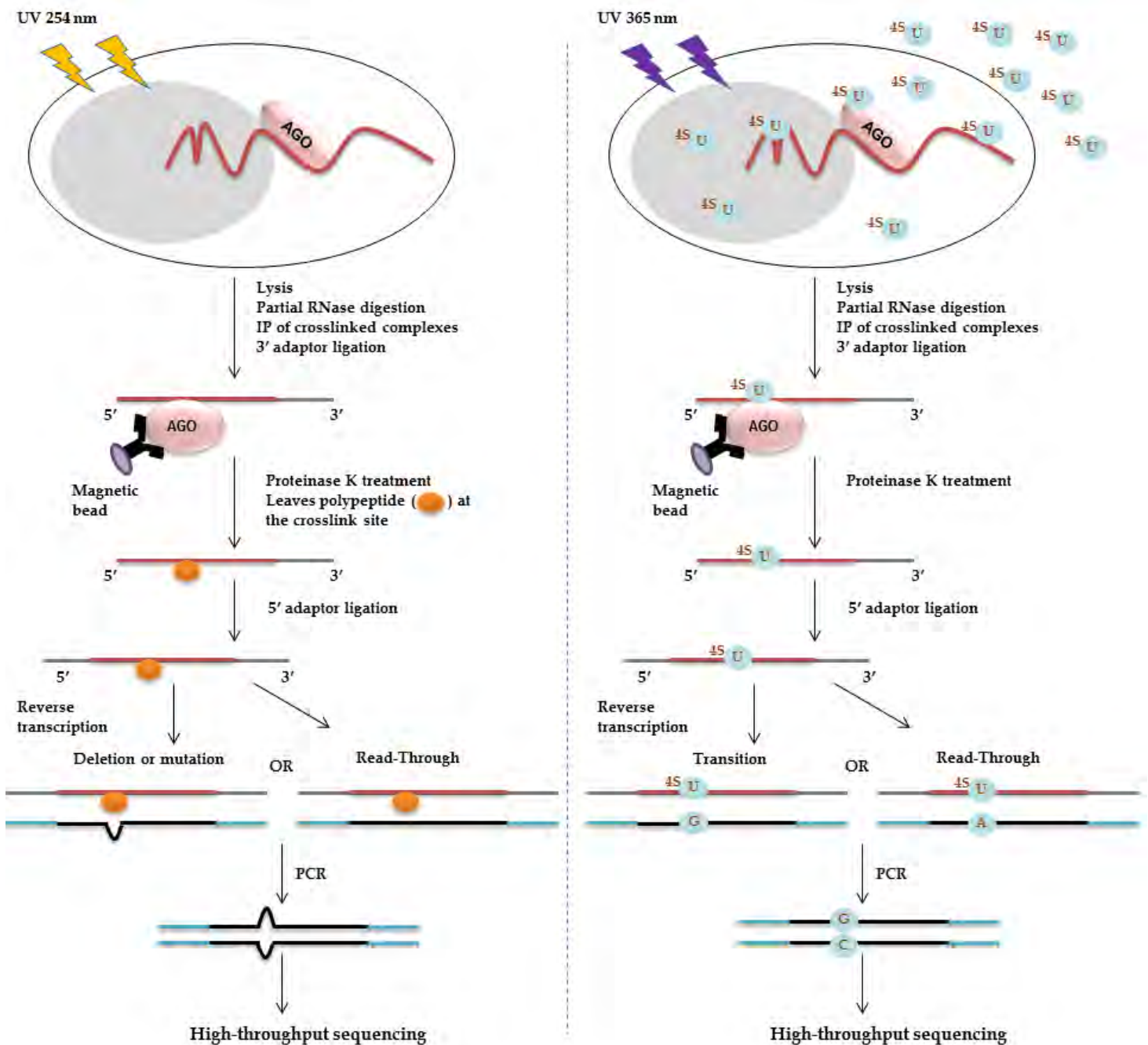


Figure 3: Overview of AGO-HITS-CLIP (left) and AGO-PAR-CLIP (right) protocols. This figure has been designed for the purpose of this dissertation.

Recently, modified protocols of AGO-CLIP-Seq methodologies have been introduced, such as CLEAR-CLIP[35] and CLASH[51], that incorporate extra ligation steps which link miRNA

molecules with their respective target binding site. This step facilitates the computational analysis and characterizes more accurate miRNA binding regions. The ligation step is following after the crosslinking and the AGO-IP process, and is induced by treatment with T4 RNA Ligase I (Figure 4). In the case of CLEAR-CLIP experiment the RNA treatment with RNA Ligase yields miRNA–target chimeric RNAs in two orientations (5' and 3' ends).

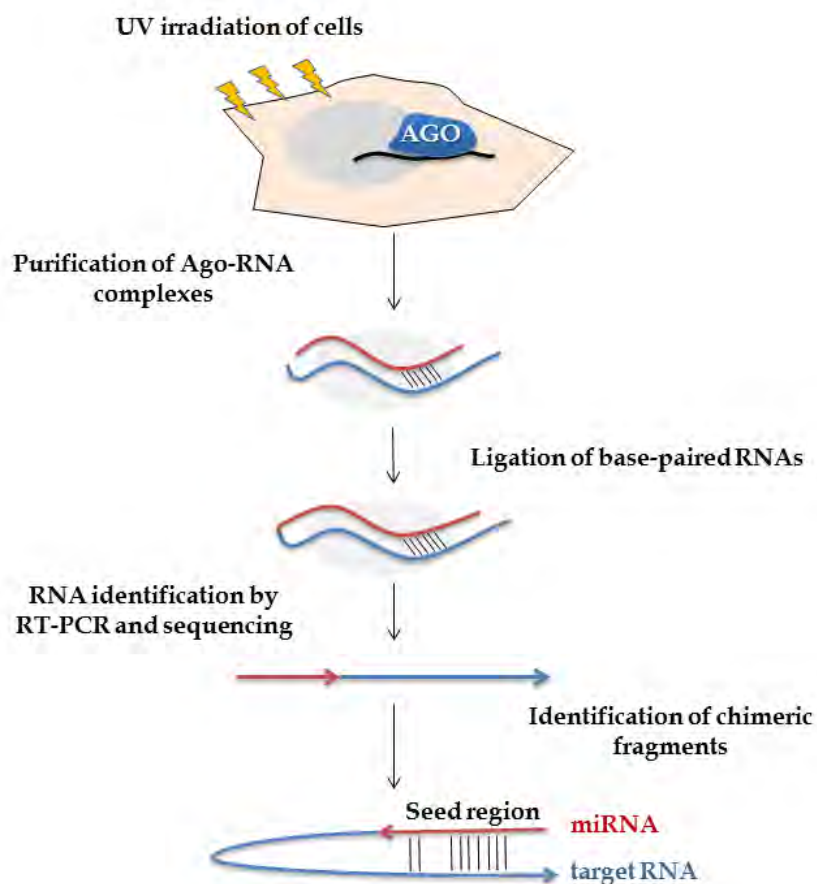


Figure 4: Overview of CLASH experiment. This figure has been designed for the purpose of this dissertation.

1.4 AGO-PAR-CLIP guided implementations

During the past few years, computational methods devoted to AGO-PAR-CLIP data analysis have been elaborated making the complex analysis of these datasets accessible to a broader community. They employ different mathematical models and feature sets and they depend strongly upon the induced T-to-C conversions to pinpoint miRNA binding sites, following the analysis performed in the seminal paper of Hafner *et al*[11]. Current models cannot be readily used on sequencing data, since they require extra pre-processing steps and the creation of non-standard file types.

A concise description of the most widely used methodologies is reviewed below:

MIRZA[55]: MIRZA is one of the first computational approaches devoted to the analysis of AGO-CLIP-Seq datasets. The implementation introduces a biophysical model for the identification of miRNA targets, leaving behind the conventional miRNA seed-based approaches. The model incorporates 27 energy parameters, inferred from AGO-CLIP-Seq data, combined with characteristics associated with base pairs, loops and specific miRNA positions. More precisely, miRNA binding positions 2-7 (seed region), 13-16 and 18-19 show the highest energy contribution, contrary to position 9 which is usually disfavored as is opening a loop. The model characterizes seed-based miRNA binding sites with perfect or partial complementarity. The algorithm utilizes a simulated annealing approach for parameters optimization. It is trained on 2,988 cross-linked regions, derived from 4 AGO-CLIP-Seq datasets[34] and evaluated against 36 microarray experiments after miRNA transfection. The model necessitates extra pre-processing steps by the user to run. It requires 30-51nt long AGO bound fragments and discards miRNA sequences shorter than 21nt.

microMUMMIE[56]: microMUMMIE is another state-of-the-art approach, that pioneered in the analysis of AGO-PAR-CLIP datasets. The algorithm is based on a six-state Hidden Markov Model for characterizing the background, the AGO-bound clusters and their flanking regions. The shape of PAR-CLIP has been modeled in a six-state topology, in which state 5 expands into a 41-state submodel for the detection of different types of miRNA seed pairing. Its core algorithm solely processes T-to-C enriched clusters determined by PARalyzer[57] and recognizes miRNA binding sites with (im)perfect seed complementarity. Evolutionary conservation, sequence composition and location of the miRNA binding site within the AGO-bound region are deemed decisive for this model. Evaluation of the prediction accuracy of the model was performed via the signal-to-noise ratio (SNR), computed by comparing shuffled and non-shuffled sites among a set of predictions. The algorithm was trained and evaluated on AGO-PAR-CLIP data, performed on EBV infected lymphoblastoid and HEK293 cell lines.

PARma[58]: PARma is a leading AGO-PAR-CLIP guided approach which provides canonical miRNA seed family interactions by processing significantly overrepresented kmers. The model adopts an iterative procedure. It identifies statistically overrepresented kmers in AGO-bound regions and all the incorporated parameters, such as seed activity probabilities, are iteratively refined until convergence. Decisive features for miRNA-targets detection are the observed positions of the T-to-C conversions and the RNase T1 cleavage sites upstream and downstream of the seed region. PARma characterizes the most probable miRNA seed in an AGO enriched cluster (MAcore), accompanied with a cluster score (Cscore). The Cscore describes the probability that a cluster is indeed a miRNA-AGO bound region, while the MAcore reflects the efficacy of the miRNA regulator. The model is fitted with an EM algorithm. The algorithm has been trained on AGO-PAR-CLIP experiments performed on B-cells. It is evaluated on DG75 cells, as well as on virus infected cell types, such as BCBL1, a

Kaposi's sarcoma-associated herpesvirus (KSHV) infected cell line and on EBV infected cells.

1.5 Databases indexing miRNA-gene interactions

The emergence of databases devoted to the cataloguing of miRNA-gene interactions has played a pivotal role in the miRNA research field.

miRTarBase[59] constitutes an extensive repository, integrating 422,517 miRNA targets, supported from low-/high-throughput experiments for several species, collected from ~8,500 publications. It provides information regarding the miRNA, the targeted gene, the binding site location, as well as miRNA/mRNA profiles retrieved from the Cancer Genome Atlas (TCGA)[60].

miRecords[32] and **miR2Disease**[61] are smaller and not consistently updated repositories. They contain approximately 3,000 validated interactions from low-yield techniques, while the latter hosts manually curated miRNA targets combined with information for miRNA deregulation in human diseases.

Other repositories, such as **StarBase**[62] and **CLIPZ**[63], substantially differ in their scope, as they provide RNA binding protein (RBP) regions from different CLIP-Seq datasets.

DIANA-TarBase v8.0[64] is an extensive repository with approximately one million miRNA-gene entries corresponding to ~670,000 unique experimentally supported miRNA-gene interactions. This collection of targets, supported by more than 33 experimental methodologies, applied to ~600 cell types/tissues under ~451 experimental conditions. TarBase was initially released in 2006, constituting the first database to catalog experimentally validated miRNA interactions and since then it is constantly updated. The current version has been enhanced with a large compilation of high quality miRNA-binding events derived from chimeric fragments, reporter gene assay and CLIP-Seq experiments. More than 200 high-throughput experiments followed by perturbation of a specific miRNA have been analyzed and integrated in the database. This extension provides an increase of approximately 200,000 interactions and ~300,000 entries since the previous version[50].

1.6 Pattern recognition and Machine Learning in Bioinformatics

Over the past two decades, the dramatic evolution of experimental methodologies has dropped the cost and increased the throughput of the results exponentially. The vast production of data has likely been the most important factor underlying the accelerated growth of bioinformatics, a field dedicated to the analysis of data and the development of computational tools indispensable for handling, manipulating and interpreting the results.

Data-driven approaches are gaining ground over the traditional methods, mainly utilized to test pre-defined hypotheses in a biological phenomenon. In most cases, in spite of the availability of data, a theoretical model, able to study the phenomenon is missing. Thus, the

bioinformatic challenge is to build and generalize predictive models, suited to solve biological problems.

Pattern recognition, in a more engineering-based approach, handles data modeling and algorithms development to effectively solve problems, by using a set of instances, represented by a number of characteristics. These problems are separated into supervised and unsupervised issues and incorporate clustering, classification and dimensionality reduction tasks. Pattern recognition is closely related to machine learning however, the latter constitutes only a part of the first. Supervised machine learning approaches are mainly trained based on characteristics derived from positive and negative instances (training data), under the purpose to effectively characterize a novel set of unknown instances. Unsupervised learning on the other hand, is applied in cases where no positive/negative data are available and unknown patterns have to be discovered. According to Bishop et al[65], *“the field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories”*.

The following section focuses on machine learning supervised classification approaches applied to ncRNA-related studies and discusses in detail the function of the algorithms.

1.6.1 Probabilistic classifiers

Probabilistic classifiers[66] are among the most popular classifiers used in the machine learning community and appear in a wide range of applications. These classifiers are derived from generative probability models that cover the original space or more involved spaces and are assigned to the study of complex statistical classification domains such as natural language and visual processing. A probabilistic classifier is able to predict and classify unknown observations by considering a set of characteristics. Notably, unlike other algorithms, it does not simply detect the “best” classification option but also assigns a probability under which the instance is being described by the label. Probabilistic classifiers provide classification that can also be utilized in ensemble learners that are discussed later in this section.

1.6.2 Feature Extraction and Selection

Descriptors are extracted from a positive and negative set of observations (training set) and are responsible for the training and the performance of the building models. The optimal selection of features, covering the complexity of the problem, is considered fundamental in a classification procedure. Transformation techniques are usually applied to descriptors with respect to the type of the latter (continuous, nominal, dichotomous, ordinal), prior to a machine learning algorithm in order to achieve the optimal model performance. The most common techniques are the (a) transformation of the categorical features in numerical, (b)

scaling or normalizing features within a specific range, e.g. 0-1 and (c) dimensionality reduction. The latter can be achieved with the Principal Components Analysis (PCA)[67] which attempts to reduce a large dimensionality feature vector into a smaller dimensionality vector that encodes less redundancy and can be more efficiently interpreted. Another transformation method usually used for categorical features is the One Hot Encoding technique. The transformation actually takes one column with x categories ($x > 2$) and converts it into x columns, where each one represents one category in the original column. Notably, several of the classification models mentioned below, such as Neural Networks and Support Vector Machines, also transform features internally.

Feature selection methods differentiate from the aforementioned techniques as they are applied to privilege the most optimal subset of the original feature set[68]. The selection of the optimal subset of descriptors not only accelerates the training process but also improves the accuracy of the model and reduces overfitting. A brief description of various feature selection techniques is presented below.

Filtering methods: Statistical tests such as Pearson's Correlation, Linear discriminant analysis, Wilcoxon's exact test etc., investigating the in-between correlations of features, as well as their equivalence with the outcome variable, are mainly used in this phase of the modeling procedure. Additionally, tests estimating the predictive accuracy of descriptors (ROC, AUC) are utilized for feature evaluation and ranking. However, these methods evaluate the behavior of features in one dimension, ignoring their in-between associations in the multidimensional space.

Wrapper methods: Wrapper methods use subsets of features and train the model to retain only descriptors that provide the best performance. The sequential training processes make these methods computationally very expensive. Some of the main wrapper methods are the forward feature selection, the backward, and the recursive feature elimination. In forward feature selection, descriptors are added iteratively until the addition of a new parameter does not improve the model's performance. The backward feature elimination is the exact opposite procedure; starting from the initial set of features, a descriptor is detached until no improvement is observed by removal of another feature. The recursive feature elimination is a greedy procedure aiming to rank features based on their performance. The best and worst performing descriptors are retained in each iteration process until the exhaustion of all features. A great disadvantage of these methods is that by using the selected subset of features, the model becomes more prone to overfitting.

Embedded methods: Embedded methods combine the advantages of the aforementioned categories and are usually implemented by algorithms that internally incorporate their own feature selection methods, performed simultaneously with classification. The most popular

examples are LASSO and RIDGE regression which include penalization functions to reduce overfitting. LASSO regression performs L1 regularization which adds a penalty equivalent to the absolute value of the dimension of coefficients, while RIDGE regression performs L2 regularization which adds a penalty equivalent to the square of the dimension of coefficients.

1.6.3 Machine Learning Algorithms

1.6.3.1 Generalized Linear Models

The generalized linear models (GLMs) are primarily introduced by Nelder and Wedderburn in 1972¹³⁰ and are considered as an extension of the linear regression model to variables that are not normally distributed. The idea was conceived in order to unify other statistical models including linear regression, logistic regression and Poisson regression. In a generalized linear model the response variable is modelled by a linear predictor of explanatory variables (1) followed by a link (**Error! Reference source not found.**) and a variance (**Error! Reference source not found.**) function. The link function describes the dependency of the mean against the linear predictor and the variance associates variance with the mean. In contrast to the simple linear model, “general” refers to the dependence on potentially more than one explanatory variable and to an included error term which is independent and identically distributed.

$$h_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} \quad (1)$$

$$E(Y) = \mu = g^{-1}(h) \quad (2)$$

$$Var(Y) = Var(\mu) = V(g^{-1}(h)) \quad (3)$$

The response variable is assumed to be generated from a particular distribution in the exponential family including the normal, binomial, Poisson and gamma distributions. The explanatory variables β are typically estimated with maximum likelihood or Bayesian techniques.

The GLM models presume that the incorporated descriptors should be uncorrelated. Extensions of the methodology, such as Generalized estimating equations (GEEs) and Generalized Linear Mixed Models (GLMMs) permit in-between parameter associations.

1.6.3.2 Naïve Bayes classifier

Naïve Bayes classifier is a probabilistic machine learning model, ideal for classification tasks[69]. It is based on the Bayesian theorem in which the final predictions are displayed by combining prior and likelihood probabilities to form and maximize the probabilities of a class occurrence for a set of features, known as posterior probabilities. The main assumptions, adopted by the classifier, are that the input descriptors are independent and they have an

equal effect on the outcome. Naïve Bayes classifier was introduced in 1960s and since then several extensions of the algorithm have been developed. Some of the dominant types are the Multinomial Naïve Bayes, which is mostly utilized in document classification problems, the Bernoulli Naïve Bayes, which is quite similar with the former type, with the difference that the predictors are boolean variables and the Gaussian Naïve Bayes, which is ideal in cases of continuous descriptors that follow a Gaussian distribution.

Naïve Bayes classifier, despite its simplicity, is still a popular baseline method mostly used in sentiment analysis, spam filtering, recommendation systems, bioinformatics, medical diagnosis etc. It is fast and easy to implement as it necessitates only a small number of training data. However, the worst drawback is the requirement of predictors to be independent, something practically impossible in most real cases.

1.6.3.3 Support Vector Machines

Support Vector Machines (SVMs) are supervised learning models widely used both in classification and regression analysis. They were initially introduced by Vapnik in 1963[70] and belong to the frontline in the machine learning field due to their high accuracy within a low computational cost. The purpose of this algorithm is to define a hyperplane in N-dimensional space (N is the number of descriptors) that distinctly classifies the data points. The optimal hyperplane acquires the maximum margin i.e. the maximum distance between the data points of both classes (Figure 5), in order to confidently classify future unknown data points and reduce generalization error. Hyperplanes act as decision boundaries for the classification of new data, i.e. each side represents the different classes. Also, the dimension of the hyperplane is associated with the number of features. Support vectors are the data points that define hyperplane's limits. SVMs perform a non-linear classification and define their inputs into high-dimensional feature space in terms of a kernel function $k(x,y)$.

Many extensions of the original SVMs have been proposed providing different options such as Support-Vector Clustering (SVC) ideal for unsupervised learning, Transductive Support-Vector Machines adopted in semi-supervised learning, multiclass SVM, Bayesian SVM etc.

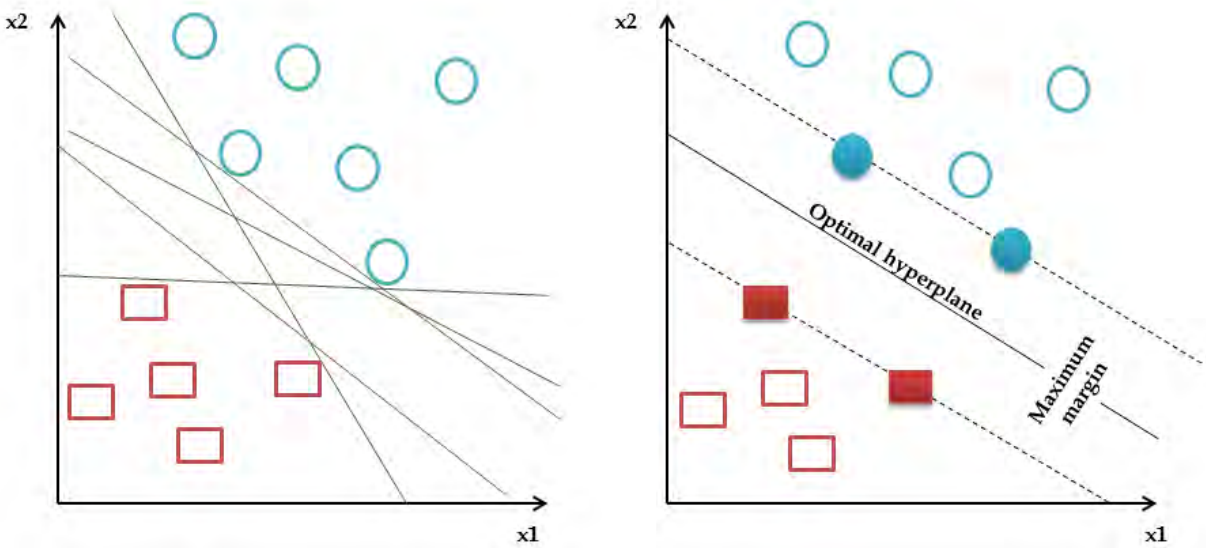


Figure 5: Representation of possible hyperplanes (left) and the optimal hyperplane (right) in a SVM classification scheme. This figure has been designed for the purpose of this dissertation.

1.6.3.4 Decision Trees

Decision Trees is a widely used machine learning algorithm, ideal for classification and regression problems. The classification scheme of the algorithm imitates human thinking and logic with a tree-like approach. A decision tree incorporates nodes assigned to each feature, links representing a decision rule and leafs attributed to a categorical or continuous outcome. The modeling of the tree presumes the optimal selection of features and conditions in each step, followed by its trimming to avoid overfitting. Recursive binary splitting strategy is usually adopted where all the features are considered and different split strategies are tried and evaluated to minimize the cost. A large number of features can lead to complex trees and unavoidable to overfitting. Therefore, parameters indicating the minimum number of utilized training inputs on each leaf or the maximum depth of the model, i.e. the longest path from a root to a leaf, should be considered. Pruning is a subsequent method that can be utilized to remove features (nodes/branches) with low importance with the purpose to reduce tree complexity.

The most commonly used decision trees are CART and ID3[71]. Decision trees are simple and can be easily interpreted. They handle both numerical and categorical data, are not affected by non-linear associations between descriptors and perform internally feature selection. Therefore, the data preparation process is eliminated. However, the use of this algorithm also demonstrates several drawbacks. The incorporation of high-dimensional data may lead to over-complex trees and to overfitting. Decision trees are very unstable – small variations in

the data may generate a totally different outcome. Also, if some classes dominate, the model becomes biased.

1.6.3.5 Random Forest

Random Forest (RF) is introduced by Ho in 1995[72] and evolved by Breiman[73] and Culter in 2001, as an ensemble learning method that constructs and combines a multitude of decision trees (bagging) at training time. “Bagging” is coupled with the random selection of features to control variance (Figure 6). It is one of the most widely used machine learning approaches which can be applied on several tasks including classification and regression. The Random Forest model displays high performance even with its default parameters and as an upgrade to the conventional decision trees it avoids overfitting.

The Random Forest algorithm displays the general techniques of bagging/bootstrap aggregating. On each round, a random sample of the training set is selected with replacement and trees are subsequently constructed. After the training process, the unseen samples are predicted and scored by considering the majority vote of the individual trees (classification) or by averaging their predictions (regression). This procedure controls the variance without increasing the bias. The bagging process is complemented with the selection of a random subset of features at each candidate split, in order to avoid the correlation among the resulting trees. The adopted hyperparameters are nearly the same as in a decision tree or a bagging classifier.

Random forest can also measure the relative importance of features internally. The estimate is conducted by evaluating the range of noise redundancy across all trees, achieved by each node-descriptor. The main limitation of the algorithm is the run-time performance in case of high-dimensional data processing.

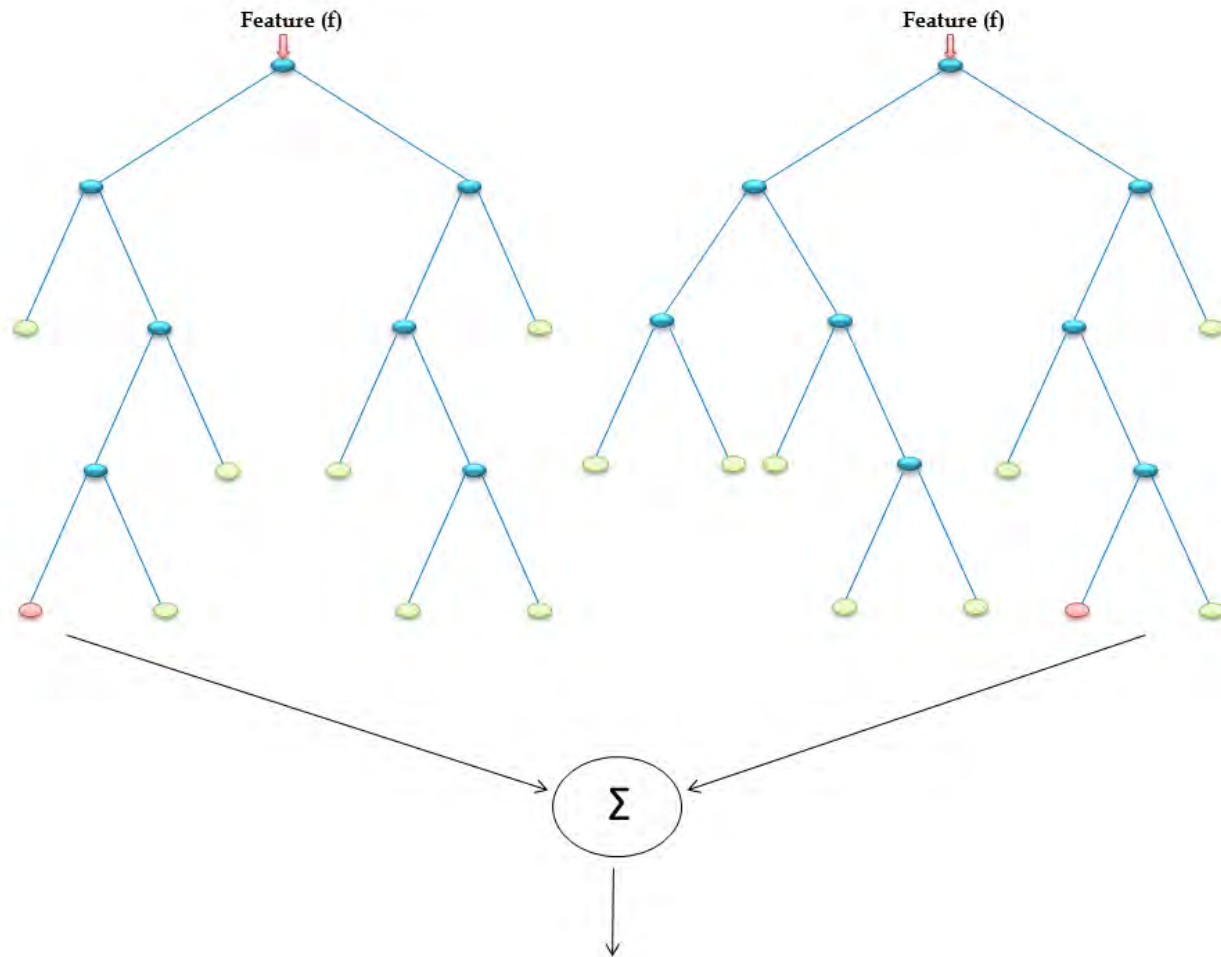


Figure 6: Random Forest representation with two trees. This figure has been designed for the purpose of this dissertation.

1.6.3.6 Deep Learning methods

Deep Learning methods[74], also known as deep neural networks, are state-of-the-art algorithms that are widely used in supervised, semi-supervised and unsupervised tasks. They are inspired by the human brain to interpret sensory numerical data under a machine perception system. Deep Learning as a “universal approximator” can easily define associations between inputs and outputs in classification, clustering and regression analysis. Walter Pitts and Warren McCulloch were the first who introduced a computational model based on neural networks of the human brain in 1943. Since then, Deep Learning is constantly evolving.

Deep learning is composed of several layers, while each one displays several nodes. A node combines input data with a set of coefficients/weights that either amplify or scale down these input parameters. The weighted input data are summed and non-linearly processed by an activation function to determine nodes impact throughout the whole network and decide their activation or not. The activation functions are usually s-shaped functions, such as sigmoid, tanh, hard tanh etc. Deep Learning also utilizes a gradient descent optimization

function in order to adjust the weights according to the error they provoke. A diagram of a node representation is displayed in Figure 7.

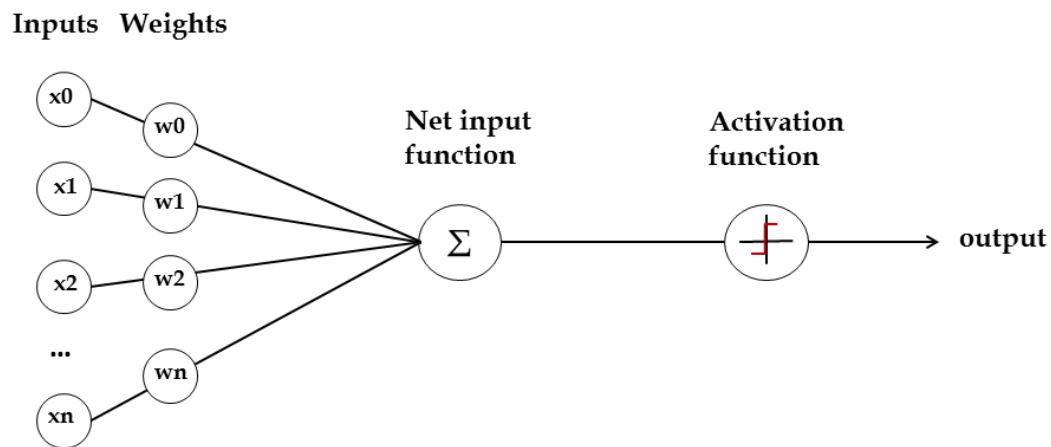


Figure 7: Representation of a node in a Deep Learning scheme. This figure has been designed for the purpose of this dissertation.

Deep neural networks are characterized by their depth, which indicates the number of node layers (hidden layers) through which data are processed. The first neural networks consisted one hidden layer apart from the input and output. More than one hidden layers mark the “deep” learning condition, where each layer’s output is the input of the subsequent hidden layer. A representation of deep learning architecture is presented below.

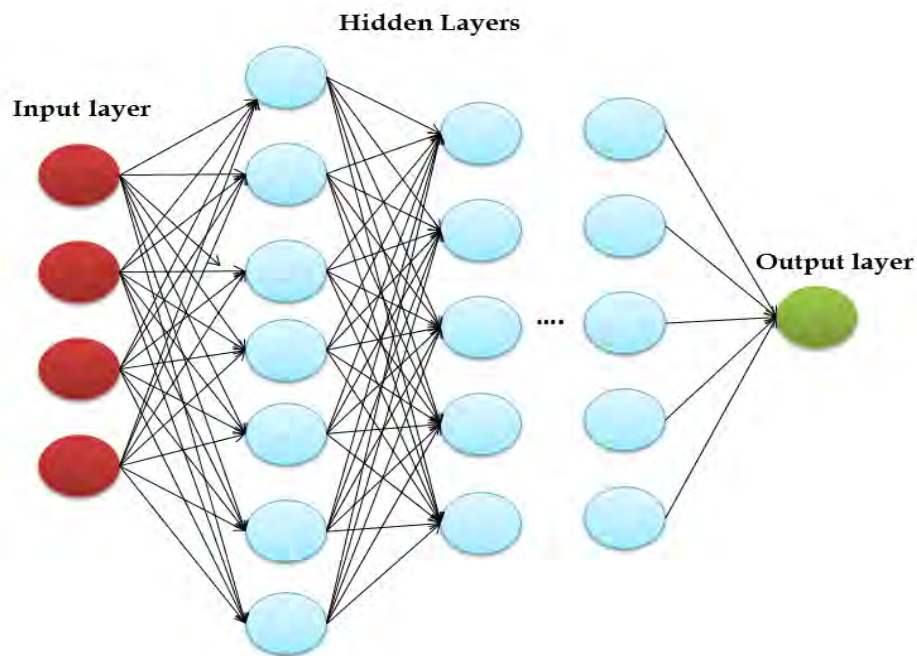


Figure 8: Deep Learning architecture. This figure has been designed for the purpose of this dissertation.

The ability to aggregate and recombine features successfully constitutes Deep Learning methods ideal for high-dimensional data manipulation. Notably, these algorithms perform automatic feature extraction without the user's intervention, unlike most of the traditional machine learning techniques. Overfitting can be avoided by applying regularization methods, such as weight decay (L1 regularization) or sparsity (L2 regularization), as well as dropout regularization that randomly omits units from the hidden layers during training.

Deep Learning methods have been applied in a multitude of different fields such as computer vision, natural language processing, audio recognition, social network filtering, bioinformatics, drug design, medical image analysis and demonstrate high accuracy by producing results comparable to human experts.

1.6.3.7 Ensemble learning algorithms

In machine learning, Ensemble learning algorithms combine multiple models to achieve better predictive performance than any individual classifier. A machine learning ensemble classifier is composed of a concrete finite set of alternative models that can be combined in a flexible structure. Not only slow algorithms may benefit from ensemble techniques but also fast algorithms such as decision trees are commonly utilized in ensemble methods e.g. Random Forests. Ensemble learning is separated into 4 main categories: boosting, bootstrap aggregating/bagging, ensemble averaging, mixture of experts. Boosting is an ensemble meta-algorithm, mainly utilized for bias and variance redundancy, able to convert weak learners to strong ones. Bootstrap aggregating or bagging is specifically designed to improve stability and accuracy, as well as to evade overfitting. As discussed in the aforementioned sections, bagging is specifically utilized in decision tree methods. Ensemble averaging is particularly adopted in neural networks where different models are generated and combined. The final model displays the best performance because the various errors of the models are averaged out. In the final category "mixture of experts", multiple experts (learners) divide the problem space into homogeneous regions. Therefore, the model decides which experts/learners are utilized in the different input regions.

1.6.3.8 Gradient Boosting

Gradient Boosting algorithm (GBM) is an ensemble machine learning boosting approach that combines weak models in a stage-wise fashion, typically decision trees, and generalizes them by optimizing a loss function. It is used both in regression and classification problems. It was initially introduced by Breiman[75] while the more recent form of the model was developed by Bartlett and Frean[76], who presented an iterative gradient descent algorithm.

In contrast to the bagging ensemble methods, gradient boosting is generating trees gradually, additively and sequentially, i.e. each decision tree is a fit on a modified version of the original dataset, emerged after the evaluation of the former tree. Final predictions are the weighted sum of the predictions that were displayed by the previous trees. The algorithm defines the shortcomings of each learner/decision tree by utilizing gradients in the loss function.

One of the commonly used regularization techniques to control overfitting is the number of gradient boosting iterations attributed to the number of trees. An optimal value of iterations is often privileged by monitoring prediction error on a separate validation dataset.

Several variants of Gradient Boosting have been developed that are widely used in a multitude of scientific fields.

CHAPTER 2

Methods

This section provides an overview of the implemented computational approaches for the accurate characterization of miRNA-mRNA interactions and their indexing in a comprehensive repository. The applied methods are summarized below:

1. Methods for the development of DIANA-TarBase v8.0[64], a database dedicated to the cataloguing of experimentally supported miRNA-mRNA pairs.
2. Implementation of microCLIP[17], a novel Super Learning Algorithm for the analysis of AGO-CLIP-Seq data.
3. Implementation of microT, a Next Generation de novo framework for the detection of miRNA-target pairs.

2.1 Methods for the development of the DIANA-TarBase v8.0 repository

DIANA-TarBase[64] is a database devoted to the indexing of experimentally supported miRNA targets. One of the major aims of this thesis was to extensively study and characterize miRNA targets. To this end, the 8th version of TarBase has been developed providing more than a million of entries. It integrates information on cell-type specific miRNA-gene regulation and hundreds of thousands of miRNA binding locations are reported. The repository enables users to extract miRNA interactions derived from 33 experimental methodologies, applied to 603 distinct cell types/tissues under 88 experimental conditions. A completely redesigned intuitive interface is also introduced, constituting a user-friendly application with flexible options to different queries.

2.1.1 Collected Data

In DIANA-TarBase v8.0 approximately 419 publications have been manually curated and added, while more than 245 high-throughput datasets harboring (in-)direct interactions have been collected and analyzed. Emphasis was placed on extracting extensive meta-data to accompany indexed entries. Each miRNA-target interaction is coupled with information regarding the relevant publications and methodologies, tissues, cell types as well as the positive or negative type of regulation. In the case of direct techniques, the exact miRNA binding locations have been archived and complementary information of the cloning primers and the targeted regulatory regions on the transcripts (e.g. 3' UTR, CDS) are included. Interactions supported from high-throughput experiments, have been extracted either from relevant publications or from the analysis of raw libraries retrieved from GEO[77] and DDBJ[78] repositories. Descriptions regarding the experimental procedures/conditions are also available to the users.

2.1.2 Analysis of high-throughput datasets incorporated in DIANA-TarBase v8.0

High-throughput experiments were analyzed to retrieve gene expression alterations upon specific miRNA treatment. Raw microarray datasets have been processed with a standardized *in silico* pipeline developed in R[79]. In Affymetrix arrays, Robust Multi-Array Average (RMA) from Bioconductor packages *affy*[80] or *oligo*[81] was utilized to perform probe set summarization. Agilent and Illumina microarray data sets were background corrected using *normexp* method and quantile normalization[82]. Probe sets were mapped to Ensembl gene IDs[83] utilizing chip-specific Bioconductor R packages[84]. Differential expression was assessed with *limma*[82], using moderated t-statistics and adjusting the associated p-values with Benjamini-Hochberg method to control the false discovery rate. The log₂ fold change values of probe sets mapped on the same gene were averaged to calculate its expression alteration. Positive and negative interactions from each set were inferred using a ± 0.5 log₂ fold change threshold, according to the perturbation type.

Processed RPF-Seq, RNA-Seq and RIP-Seq libraries, submitted to specific miRNA treatment were collected from the respective publications. Positive/negative miRNA interactions were formed from genes presenting >10 RPKM and >50% expression change.

2.1.3 Analysis of AGO-CLIP-Seq datasets incorporated in DIANA-TarBase v8.0

The CLIP-Seq analysis has been performed using an in-house developed pipeline. Regions formed by at least five overlapping reads were included to the analysis. For PAR-CLIP data, peaks containing adequate T-to-C (sense strand) or A-to-G (antisense strand) incorporation were selected. At least two transitions in the same position for peaks with less than 50 reads were required, while for the remaining regions a threshold of >5% was applied, as indicated by Hafner *et al.*[11]. For all CLIP-Seq data sets having replicates, a peak had to be present in at least two replicates in order to be considered as valid. Where available, top expressed miRNAs were retrieved from the original publication. In all other instances, publicly available small-RNA-Seq libraries derived from the relevant cell lines were analyzed. The adopted pipeline for the pre-processing of the AGO-PAR-CLIP libraries and the analysis of the sRNA-Seq datasets is described in detail in section Methods 2.2.2.

miRNA:gene interactions were inferred using a modified a version of microT-CDS algorithm[85] which considers decisive features for the accurate MRE characterization such as the miRNA:mRNA binding type, binding free energy, MRE conservation and AU flanking content. In cases where replicates were available, an interaction had to be present in at least two replicates, in order to be included to the database. Figure 9 depicts the adopted pipeline for the analysis of the AGO-CLIP-Seq libraries. The snapshot has been retrieved from the IGV[86] Genome Browser.

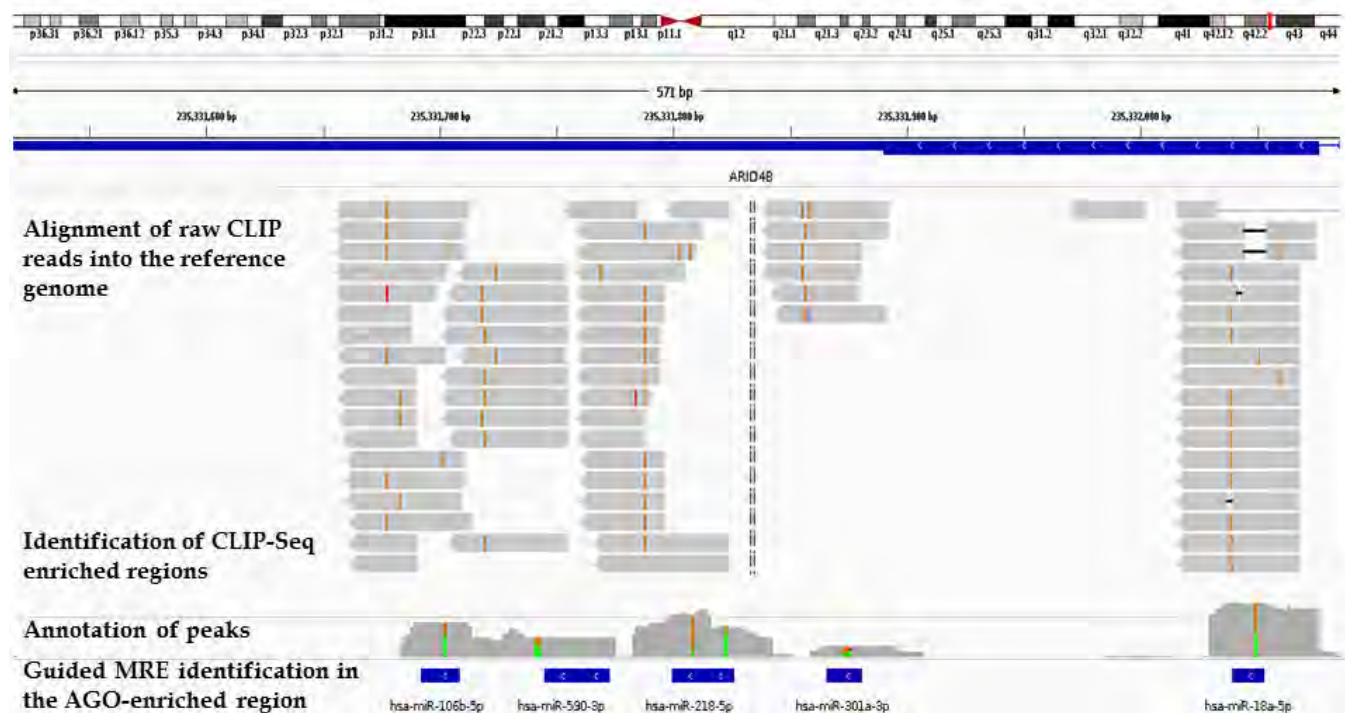


Figure 9: Snapshot from the IGV Genome Browser depicting the adopted pipeline for the analysis of the AGO-CLIP-Seq libraries. Raw CLIP-Seq reads are initially aligned into the reference genome. Regions enriched in AGO are formed by overlapping reads. AGO-CLIP clusters are annotated in a comprehensive set of transcripts. MRE identification is subsequently applied to the annotated peaks. The illustrated peaks are derived from 1 AGO-PAR-CLIP library on HEK293 cells. The brown-and-green vertical lines represent T-to-C transition sites while MREs are detected by microT-CDS algorithm. This figure has been designed for the purpose of this dissertation.

2.1.4 Database interface development

A new relational schema was designed to host TarBase v8.0 data (Figure 10). Indices were created to guarantee the efficient execution of the system and foreign keys were added to avoid integrity violations in the data. PostgreSQL was utilized to implement the hosting database. The web interface of TarBase was designed around the new database schema and effort was put into making it adaptable to a wide variety of screen formats and devices (PCs, tablets, smartphones, etc.). The interface was developed using the Yii 2.0 PHP framework. The interactive charts were implemented using the D3.js JavaScript library.

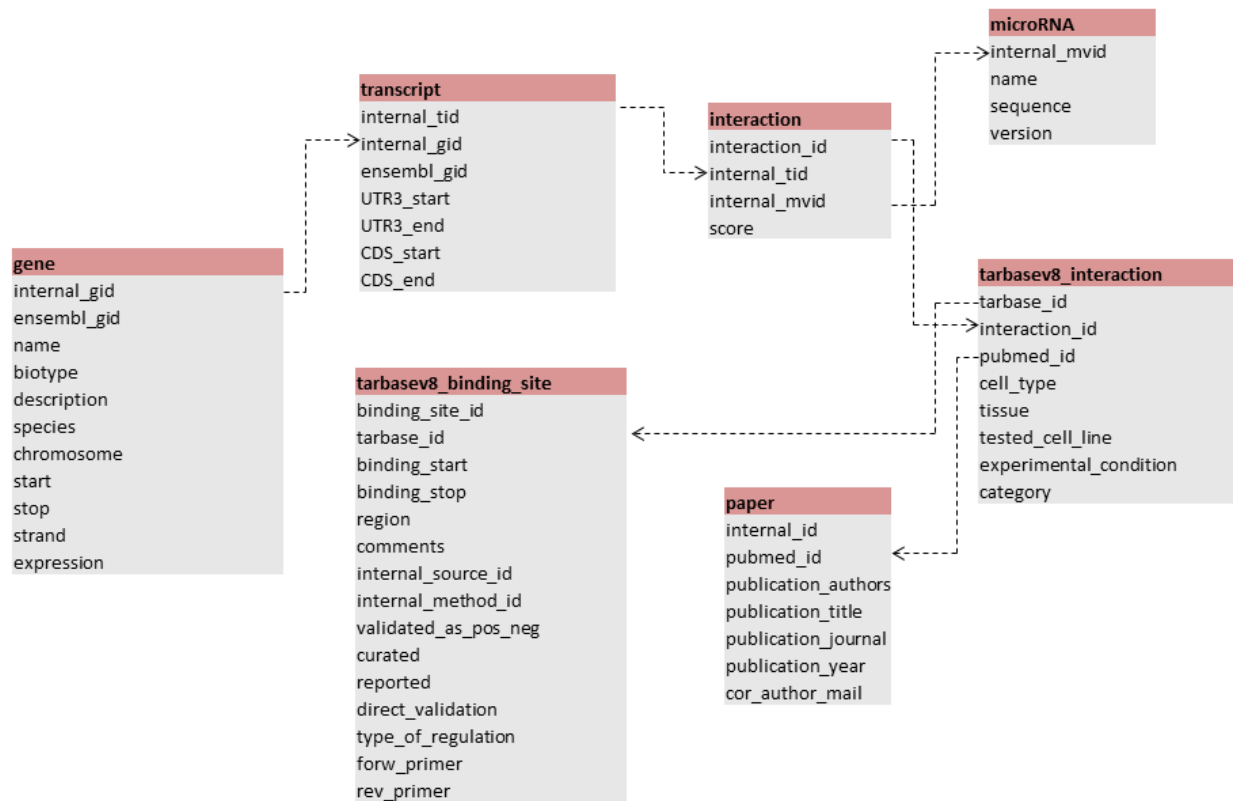


Figure 10: TarBase database schema. This figure has been designed for the purpose of this dissertation.

2.2 Implementation of microCLIP, a novel Super Learning Algorithm for the analysis of AGO-CLIP-Seq data

One of the aims of the current thesis was to revisit, identify and address current obstacles in AGO-CLIP-Seq analysis, in order to enable the accurate determination of experimentally supported functional miRNA targets. To this end, microCLIP[17] was developed, an *in silico* framework for CLIP-guided identification of miRNA interactions. microCLIP incorporates novel aspects in PAR-CLIP analysis and increases the experiment's scope and robustness. Computational approaches for AGO-CLIP-Seq data analysis incorporate machine learning techniques and thus rely heavily on training/validation dataset selection. Therefore, an extensive experimental collection of miRNA interactions was created in order to boost the proper optimization of microCLIP algorithm and its exposure to the actual search space complexity. The most remarkable finding was that clusters depleted on T-to-C conversions, which are always filtered out in such analyses, can aid in the identification of functional miRNA binding events (Figure 11).

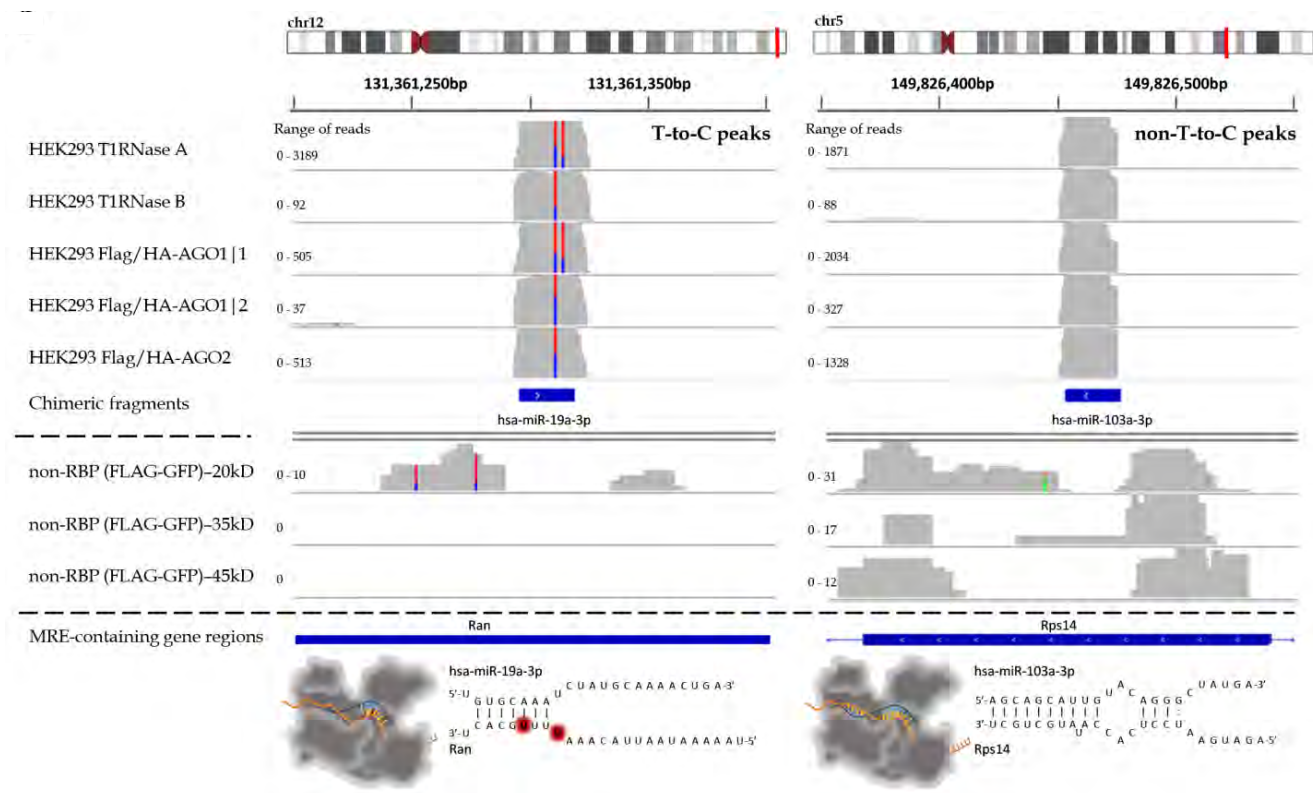


Figure 11: Peaks derived from 5 AGO-PAR-CLIP libraries on HEK293 cells and from 3 non-RBP background libraries are presented for T-to-C and non-T-to-C AGO-bound regions. The red-and-blue vertical lines represent T-to-C transition sites. Both types of AGO-enriched clusters are clearly distinguished from background signal. Chimeric miRNA-target fragments overlap with (non-)T-to-C peaks providing direct validation for specific miRNA-target pairs (hsa-miR-19a-3p-Ran and hsa-miR-103a-3p-Rps14). microCLIP identifies the aforementioned interactions as a 7-mer (*chr12:131,361,200–131,361,400*, *Ran* gene 3' UTR) and an 8-mer with a 3' compensatory site (*chr5:149,826,350–149,826,550*, *Rps14* gene CDS) respectively. The 3D depictions of AGO2 were based in the PDB structure 5JS1 (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

microCLIP provides a robust pipeline for the analysis of all AGO-enriched regions. It encompasses an approach based on a super learning scheme and employs combinations of deep learning, random forest and gradient boosting classifiers. The super learner approach was introduced by van der Laan *et al.* in 2007 and has been shown to be an asymptotically optimal system for machine learning[87]. By using multiple combinations of classifiers, super learning outperforms a single prediction model.

2.2.1 Dataset collection

microCLIP was trained and evaluated against an extensive set of interactions from hundreds of miRNA specific low/high-throughput experiments across ~50 different cell types. A high quality set, composed of direct miRNA binding events retrieved from Reporter Gene Assays and chimeric miRNA-target fragments[28, 29, 50, 88, 89], was incorporated in the algorithm's development and evaluation process (Figure 12).

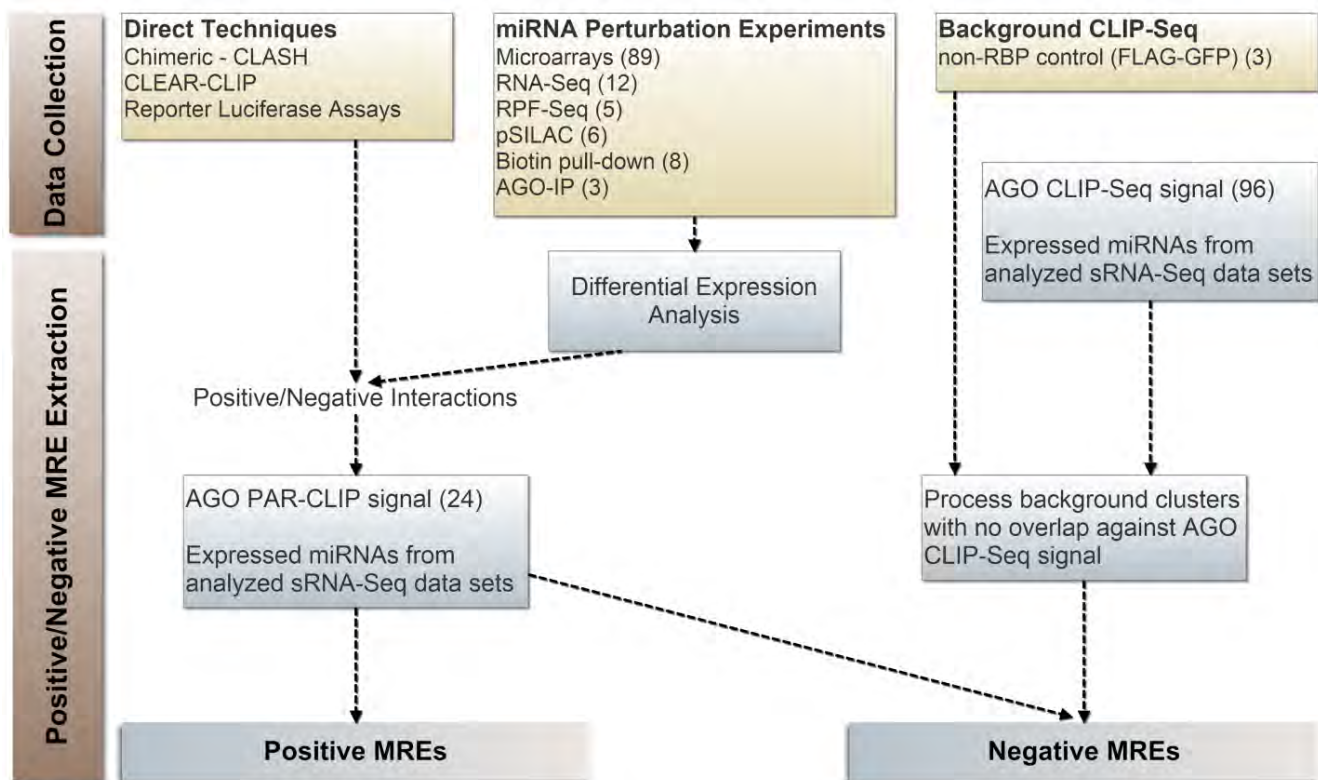


Figure 12: Dataset collection and methodology for positive and negative MRE identification. More than 6,000 interactions were retrieved from direct techniques and miRNA-target chimeric fragments. Numerous high-throughput experimental data following specific miRNA perturbations enabled the identification of AGO bound or differentially transcribed/translated genes harboring functional binding sites. In order to resolve the exact miRNA binding sites, positive and negative instances were coupled with signal from 24 AGO-PAR-CLIP libraries. The negative set was enhanced by incorporating background CLIP-Seq clusters. sRNA-Seq datasets were included to determine expressed miRNAs and accurately extract positive/negative MREs. This dataset collection was processed to form the training/test sets of microCLIP deployment (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

6,724 high confidence MREs were retrieved from direct experiments, including reporter gene assay techniques indexed in DIANA-TarBase repository[50, 89], miRNA-chimeras from CLASH[29] and CLEAR-CLIP[88] experiments, as well as additional miRNA-target chimeric fragments derived from a previous meta-analysis of published AGO-CLIP datasets[28]. In order to quantify miRNA-induced mRNA expression changes and to identify functional binding sites, 101 miRNA perturbation experiments were analyzed (89 microarray and 12 RNA-Seq experiments, Table 4, Table 5). This process enabled the formation of approximately 3,900 and 4,000 positive and negative miRNA-target pairs respectively. A set of 5 ribosome profiling sequencing (RPF-Seq) libraries after miRNA overexpression, capturing differentially ribosome-bound transcripts, and 6 pSILAC (quantitative proteomics) experiments were an additional source for detecting more than 5,900 miRNA effects at protein expression level

(Table 4). The inclusion of AGO-IP and biotin pull-down high-throughput experiments upon specific miRNA perturbation yielded approximately 2,600 miRNA binding events (Table 4). The aforementioned miRNA perturbation experiments enabled the detection of deregulated targets without specifying the exact binding sites[50]. miRNA-targeted regions were extracted from AGO-bound enriched regions present in at least 1 of 24 AGO-PAR-CLIP sequencing libraries (Table 6). Published background PAR-CLIP libraries[90], stably expressing a commonly utilized non-RBP control (FLAG-GFP), were incorporated in our pipeline to identify non-specific AGO-bound transcripts and deduce more than 24,000 negative miRNA binding sites. A compendium of 96 AGO-CLIP-Seq experiments was derived from DIANA-TarBase and used to further select background-derived MREs displaying no overlap with AGO-enriched regions.

Table 4: Summary of the collected experiments in human species upon specific miRNA deregulation. The datasets were utilized to extract independent training and test sets of positive and negative MRE regions for microCLIP deployment.

| Accession | Repository | Authors | Experiment | Cell Type | miRNA | miRNA treatment | Post-Transfection Cell Harvest Time/Experimental Condition |
|-----------|----------------------|-----------------------------|-------------|---------------------|-----------------|-----------------|--|
| GSE27718 | ncbi.nlm.nih.gov/geo | Gaziel-Sovran <i>et al.</i> | microarrays | 113/6-4L, 131/4-5B1 | hsa-miR-30d-5p | Overexpression | 60h |
| GSE58004 | ncbi.nlm.nih.gov/geo | Kiga <i>et al.</i> | microarrays | AGS | hsa-miR-210-3p | Overexpression | 36h |
| GSE38956 | ncbi.nlm.nih.gov/geo | Ramachandran <i>et al.</i> | microarrays | CALU3 | hsa-miR-138-5p | Overexpression | 48h |
| GSE12400 | ncbi.nlm.nih.gov/geo | Sander <i>et al.</i> | microarrays | CCL86, CRL1432 | hsa-miR-26a-5p | Overexpression | 72h |
| GSE51053 | ncbi.nlm.nih.gov/geo | Kristensen <i>et al.</i> | microarrays | DU145 | hsa-miR-452-5p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson <i>et al.</i> | microarrays | H4 | hsa-miR-103a-3p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson <i>et al.</i> | microarrays | H4 | hsa-miR-107 | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson <i>et al.</i> | microarrays | H4 | hsa-miR-15b-3p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson <i>et al.</i> | microarrays | H4 | hsa-miR-16-5p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson <i>et al.</i> | microarrays | H4 | hsa-miR-195-5p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson <i>et al.</i> | microarrays | H4 | hsa-miR-320b | Overexpression | 48h |
| GSE22790 | ncbi.nlm.nih.gov/geo | Elyakim <i>et al.</i> | microarrays | HEPG2 | hsa-miR-191-5p | Anti-miR | NA |
| GSE6207 | ncbi.nlm.nih.gov/geo | Wang <i>et al.</i> | microarrays | HEPG2 | hsa-miR-124-3p | Overexpression | 4h, 8h, 16h, 24h, 32h, 72h, 120h |
| GSE56973 | ncbi.nlm.nih.gov/geo | Hill <i>et al.</i> | microarrays | HEY | hsa-miR-429 | Overexpression | 48h |
| GSE23392 | ncbi.nlm.nih.gov/geo | Shahab <i>et al.</i> | microarrays | HEY | hsa-miR-128-3p | Overexpression | 48h |

| | | | | | | | |
|----------|--------------------------|------------------------------------|-------------|----------------|----------------------|-----------------------------|---------------------|
| GSE23392 | ncbi.nlm.nih.gov/ geo | Shahab <i>et al.</i> | microarrays | HEY | hsa-miR-7-5p | Overexpression | 48h |
| GSE41737 | ncbi.nlm.nih.gov/ geo | Shirasaki <i>et al.</i> | microarrays | HUH7.5 | hsa-miR-27a-3p | Anti-miR, Overexpression | NA |
| GSE16962 | ncbi.nlm.nih.gov/ geo | Fasanaro <i>et al.</i> | microarrays | HUVEC | hsa-miR-210-3p | Anti-miR, Overexpression | 24h |
| GSE18651 | ncbi.nlm.nih.gov/ geo | Cushing <i>et al.</i> | microarrays | IMR90 | hsa-miR-29a-3p | Knockdown | 48h |
| GSE16674 | ncbi.nlm.nih.gov/ geo | Navarro <i>et al.</i> | microarrays | K562 | hsa-miR-34a-5p | Overexpression | 24h |
| GSE17362 | ncbi.nlm.nih.gov/ geo | Boll <i>et al.</i> | microarrays | LNCAP | hsa-miR-130a-3p | Overexpression | 24h |
| GSE17362 | ncbi.nlm.nih.gov/ geo | Boll <i>et al.</i> | microarrays | LNCAP | hsa-miR-203a-3p | Overexpression | 24h |
| GSE17362 | ncbi.nlm.nih.gov/ geo | Boll <i>et al.</i> | microarrays | LNCAP | hsa-miR-205-5p | Overexpression | 24h |
| GSE31620 | ncbi.nlm.nih.gov/ geo | Hudson <i>et al.</i> | microarrays | LNCAP | hsa-miR-1 | Overexpression | 24h |
| GSE31620 | ncbi.nlm.nih.gov/ geo | Hudson <i>et al.</i> | microarrays | LNCAP | hsa-miR-27b-3p | Overexpression | 24h |
| GSE33538 | ncbi.nlm.nih.gov/ geo | Bossel Ben- Moshe <i>et al.</i> | microarrays | MCF10A | hsa-miR-20a-5p | Silencing | 0h, 0.5h, 1h, 2h |
| GSE33538 | ncbi.nlm.nih.gov/ geo | Bossel Ben- Moshe <i>et al.</i> | microarrays | MCF10A | hsa-miR-671-5p | Silencing | 0h, 1h, 2h |
| GSE58142 | ncbi.nlm.nih.gov/ geo | Frankel <i>et al.</i> | microarrays | MCF7 | hsa-miR-95a-3p | Overexpression | 24h |
| GSE31397 | ncbi.nlm.nih.gov/ geo | Frankel <i>et al.</i> | microarrays | MCF7 | hsa-miR-101-3p | Overexpression | 24h |
| GSE19777 | ncbi.nlm.nih.gov/ geo | Rao <i>et al.</i> | microarrays | MCF7FR | hsa-miR-221-3p | Silencing | 72h |
| GSE58004 | ncbi.nlm.nih.gov/ geo | Kiga <i>et al.</i> | microarrays | MKN45 | hsa-miR-210-3p | Overexpression | 36h |
| GSE32876 | ncbi.nlm.nih.gov/ geo | Setty <i>et al.</i> | microarrays | MSK543 | hsa-miR-124-3p | Overexpression | 24h |
| GSE32876 | ncbi.nlm.nih.gov/ geo | Setty <i>et al.</i> | microarrays | MSK543 | hsa-miR-132-3p | Overexpression | 24h |
| GSE57158 | ncbi.nlm.nih.gov/ geo | Greenberg <i>et al.</i> | microarrays | PAG C81-61 | hsa-miR-20a-5p | Overexpression | 3d |
| GSE51053 | ncbi.nlm.nih.gov/ geo | Kristensen <i>et al.</i> | microarrays | PC3 | hsa-miR-224-5p | Overexpression | 48h |
| GSE51053 | ncbi.nlm.nih.gov/ geo | Kristensen <i>et al.</i> | microarrays | PC3 | hsa-miR-452-5p | Overexpression | 48h |
| GSE65892 | ncbi.nlm.nih.gov/ geo | Wagenaar <i>et al.</i> | microarrays | SKHEP1 | hsa-miR-21-5p | Anti-miR | 16h |
| GSE19693 | ncbi.nlm.nih.gov/ geo | Chen <i>et al.</i> | microarrays | U87, HS683 | hsa-miR-20a-5p | Overexpression | NA |
| GSE34846 | ncbi.nlm.nih.gov/ geo | Cao <i>et al.</i> | microarrays | HTERT- RPE1 | hsa-miR-129-2- 3p | Overexpression | 72h |
| GSE37427 | ncbi.nlm.nih.gov/ geo | Zhu <i>et al.</i> | microarrays | FLS | hsa-miR-23b-3p | Overexpression | NA |
| GSE22143 | ncbi.nlm.nih.gov/ geo | Marcet <i>et al.</i> | microarrays | HAEC | hsa-miR-34a-5p | Overexpression | 48h |
| GSE22143 | ncbi.nlm.nih.gov/ geo | Marcet <i>et al.</i> | microarrays | HAEC | hsa-miR-34c-5p | Overexpression | 48h |
| GSE22143 | ncbi.nlm.nih.gov/ geo | Marcet <i>et al.</i> | microarrays | HAEC | hsa-miR-449b- 5p | Overexpression | 48h |
| GSE22143 | ncbi.nlm.nih.gov/ geo | Marcet <i>et al.</i> | microarrays | HAEC | hsa-miR-449a | Overexpression | 48h |

| | | | | | | | |
|-----------------|--------------------------|--------------------------|-------------|--|-----------------|------------------------------|---|
| GSE68424 | ncbi.nlm.nih.gov/ geo | Teplyuk <i>et al.</i> | microarrays | GBM4, GBM6 | hsa-miR-10b-5p | Inhibition | 24h |
| GSE35621 | ncbi.nlm.nih.gov/ geo | Hu <i>et al.</i> | microarrays | HEK293T, HSF2 | hsa-miR-941 | Overexpression | 24h |
| GSE37596 | ncbi.nlm.nih.gov/ geo | Hwang <i>et al.</i> | microarrays | HT29 | hsa-miR-146a-5p | Overexpression | 2w after lentiviral infection |
| GSE40058 | ncbi.nlm.nih.gov/ geo | Luo <i>et al.</i> | microarrays | MDA-MB- 231 | hsa-miR-200c-3p | Overexpression | NA |
| GSE40058 | ncbi.nlm.nih.gov/ geo | Luo <i>et al.</i> | microarrays | MDA-MB- 231 | hsa-miR-205-5p | Overexpression | NA |
| GSE7754 | ncbi.nlm.nih.gov/ geo | Chang <i>et al.</i> | microarrays | HCT116 | hsa-miR-34a-5p | Overexpression | 2w after retroviral infection |
| GSE51875 | ncbi.nlm.nih.gov/ geo | Lee <i>et al.</i> | microarrays | HCT116 | hsa-miR-147a | Overexpression | 3d |
| GSE50697 | ncbi.nlm.nih.gov/ geo | Taube <i>et al.</i> | microarrays | SUM159 | hsa-miR-203a-3p | Overexpression | NA |
| GSE35208 | ncbi.nlm.nih.gov/ geo | Lin <i>et al.</i> | microarrays | U87-2M1 | hsa-miR-10b-5p | Inhibition | NA |
| GSE14507 | ncbi.nlm.nih.gov/ geo | Webster <i>et al.</i> | microarrays | A549 | hsa-miR-7-5p | Overexpression | 24h |
| GSE21132 | ncbi.nlm.nih.gov/ geo | Li <i>et al.</i> | microarrays | Jurkat | hsa-miR-146a-5p | Overexpression, Knockdown | 48h |
| GSE24824 | ncbi.nlm.nih.gov/ geo | Huynh <i>et al.</i> | microarrays | Melanoma- metastatic Liver Cells | hsa-miR-182-5p | Anti-miR | administered twice per week over 4 weeks |
| GSE56268 | ncbi.nlm.nih.gov/ geo | Schneider <i>et al.</i> | microarrays | P3HR1 | hsa-miR-28-5p | Overexpression | 12h, 24h |
| GSE52531 | ncbi.nlm.nih.gov/ geo | Nam <i>et al.</i> | | HEK293 | hsa-miR-155-5p | Overexpression | 24h |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RNA-Seq | HeLa | hsa-miR-155-5p | Overexpression | 32h |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RNA-Seq | U2OS (total) | hsa-miR-155-5p | Overexpression | 32h/poly(A) -selected total RNA |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RNA-Seq | U2OS (cyto) | hsa-miR-155-5p | Overexpression | 32h/poly(A) -selected cytoplasmic RNA |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RNA-Seq | U2OS (ribo) | hsa-miR-155-5p | Overexpression | tRNA and rRNA depleted RNA |
| GSE37918 | ncbi.nlm.nih.gov/ geo | Pellegrino <i>et al.</i> | RNA-Seq | MDA-MB- 231 | hsa-miR-23b-3p | Overexpression | NA |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RPF-Seq | HEK293T | hsa-miR-1-3p | Overexpression | 24h |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RPF-Seq | HeLa | hsa-miR-155-5p | Overexpression | 32h |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RPF-Seq | U2OS | hsa-miR-1-3p | Overexpression | 32h |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RPF-Seq | U2OS | hsa-miR-155-5p | Overexpression | 32h |
| GSE60426 | ncbi.nlm.nih.gov/ geo | Eichhorn <i>et al.</i> | RPF-Seq | HeLa | hsa-miR-1-3p | Overexpression | 32h |
| NA | psilac.mdc- | Selbach <i>et al.</i> | pSILAC | HeLa | hsa-let-7b-5p | Overexpression, | 8h post- |

| | | | | | | | |
|----------|---------------------------------------|-----------------------|------------------|---------|----------------|----------------|---|
| | berlin.de | | | | | Knockdown | transfection and 24h pSILAC labelling 8h post-transfection and 24h pSILAC labelling 8h post-transfection and 24h pSILAC labelling 8h post-transfection and 24h pSILAC labelling |
| | psilac.mdc-berlin.de | Selbach <i>et al.</i> | pSILAC | HeLa | hsa-miR-1-3p | Overexpression | |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | pSILAC | HeLa | hsa-miR-16-5p | Overexpression | |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | pSILAC | HeLa | hsa-miR-30a-5p | Overexpression | |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | pSILAC | HeLa | hsa-miR-155-5p | Overexpression | |
| GSE40408 | ncbi.nlm.nih.gov/geo | Martin et al. | Biotin pull-down | HEK293T | hsa-miR-23b-3p | Overexpression | 24h |
| GSE40408 | ncbi.nlm.nih.gov/geo | Martin et al. | Biotin pull-down | HEK293T | hsa-miR-27a-3p | Overexpression | 24h |
| GSE40408 | ncbi.nlm.nih.gov/geo | Martin et al. | Biotin pull-down | HEK293T | hsa-miR-17-5p | Overexpression | 24h |
| GSE29101 | ncbi.nlm.nih.gov/geo | Cloonan et al. | Biotin pull-down | HEK293T | hsa-miR-10a-5p | Overexpression | 24h |
| GSE40411 | ncbi.nlm.nih.gov/geo | Krishnan et al. | Biotin pull-down | MCF7 | hsa-miR-139-5p | Overexpression | 24h |
| GSE38593 | ncbi.nlm.nih.gov/geo | Krishnan et al. | Biotin pull-down | HEK293T | hsa-miR-182-5p | Overexpression | 24h |
| GSE11082 | ncbi.nlm.nih.gov/geo | Hendrickson et al. | AGO-IP | HEK293T | hsa-miR-1 | Overexpression | 48h |
| GSE11082 | ncbi.nlm.nih.gov/geo | Hendrickson et al. | AGO-IP | HEK293T | hsa-miR-124-3p | Overexpression | 48h |
| GSE39227 | ncbi.nlm.nih.gov/geo | Hu et al. | AGO-IP | HEK293T | hsa-miR-941 | Overexpression | NA |
| NA | doi:10.1371/journal.pgen.1002363.s006 | Lal et al. | Biotin pull-down | K562 | hsa-miR-34a-5p | Overexpression | 24h |
| NA | doi:10.1371/journal.pgen.1002363.s006 | Lal et al. | Biotin pull-down | HTC116 | hsa-miR-34a-5p | Overexpression | 24h |

Table 5: Summary of microarray and RNA sequencing experiments in human species upon specific miRNA deregulation, utilized in benchmarking evaluations of microCLIP model.

| Accession | Repository | Authors | Cell Type | miRNA | miRNA treatment | Post-Transfection Cell Harvest Time/Experimental Condition |
|-----------|----------------------|----------------------|-----------|----------------|-----------------|--|
| GSE46039 | ncbi.nlm.nih.gov/geo | Helwak <i>et al.</i> | HEK293 | hsa-miR-92a-3p | Knockdown | 48h |

| | | | | | | |
|----------|----------------------|--------------------------|---------|----------------|----------------|---|
| GSE21577 | ncbi.nlm.nih.gov/geo | Hafner <i>et al.</i> | HEK293 | hsa-miR-20a-5p | Knockdown | simultaneous miRNA knockdown using inhibitor cocktail |
| GSE21901 | ncbi.nlm.nih.gov/geo | Hollander <i>et al.</i> | HEK293 | hsa-miR-212-3p | Overexpression | NA |
| GSE14537 | ncbi.nlm.nih.gov/geo | Hausser <i>et al.</i> | HEK293 | hsa-miR-124-3p | Overexpression | 15h |
| GSE14537 | ncbi.nlm.nih.gov/geo | Hausser <i>et al.</i> | HEK293 | hsa-miR-7-5p | Overexpression | 15h |
| GSE35621 | ncbi.nlm.nih.gov/geo | Hu <i>et al.</i> | HEK293 | hsa-miR-941 | Overexpression | 24h |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | HeLa | hsa-let-7b-5p | Overexpression | 32h |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | HeLa | hsa-miR-1 | Overexpression | 32h |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | HeLa | hsa-miR-155-5p | Overexpression | 32h |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | HeLa | hsa-miR-16-5p | Overexpression | 32h |
| NA | psilac.mdc-berlin.de | Selbach <i>et al.</i> | HeLa | hsa-miR-30a-5p | Overexpression | 32h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson <i>et al.</i> | HeLa | hsa-miR-7-5p | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam <i>et al.</i> | HEK293 | hsa-miR-124-3p | Overexpression | 24h |
| GSE68987 | ncbi.nlm.nih.gov/geo | Zhang <i>et al.</i> | HeLa | hsa-miR-603 | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam <i>et al.</i> | HeLa | hsa-miR-155-5p | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam <i>et al.</i> | HeLa | hsa-miR-124-3p | Overexpression | 24h |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn <i>et al.</i> | HEK293T | hsa-miR-1 | Overexpression | 24h |
| GSE37918 | ncbi.nlm.nih.gov/geo | Pellegrino <i>et al.</i> | MCF7 | hsa-miR-23b-3p | Overexpression | NA |

Table 6: Summary of the collected AGO-PAR-CLIP experiments in human species, obtained from 9 studies. These datasets provided the source of PAR-CLIP signal (raw reads and transitions) which was integrated with experimentally validated positive/negative instances of miRNA-targeted regions.

| Accession | Repository | Authors | Experiment | Species | Cell line | Samples |
|------------|----------------------|------------------------|------------|---------|------------|--|
| GSE28859 | ncbi.nlm.nih.gov/geo | Kishore <i>et al.</i> | PAR-CLIP | human | HEK293 | GSM714644, GSM714645, GSM714646, GSM714647 |
| SRR1045082 | ncbi.nlm.nih.gov/sra | Farazi <i>et al.</i> | PAR-CLIP | human | MCF7 | SRA110557 |
| SRR359787 | ncbi.nlm.nih.gov/sra | Lipchina <i>et al.</i> | PAR-CLIP | human | hESC | SRA047324 |
| GSE59944 | ncbi.nlm.nih.gov/geo | Whisnant <i>et al.</i> | PAR-CLIP | human | C8166 | GSM1462572 |
| GSE59944 | ncbi.nlm.nih.gov/geo | Whisnant <i>et al.</i> | PAR-CLIP | human | TZMBL | GSM1462573, GSM1462574 |
| GSE32109 | ncbi.nlm.nih.gov/geo | Gottwein <i>et al.</i> | PAR-CLIP | human | BC-1 | GSM796037, GSM796038 |
| GSE32109 | ncbi.nlm.nih.gov/geo | Gottwein <i>et al.</i> | PAR-CLIP | human | BC-3 | GSM796039, GSM796040 |
| GSE41437 | ncbi.nlm.nih.gov/geo | Skalsky <i>et al.</i> | PAR-CLIP | human | EF3D-AGO2 | GSM1020021 |
| GSE41437 | ncbi.nlm.nih.gov/geo | Skalsky <i>et al.</i> | PAR-CLIP | human | LCL35 | GSM1020022 |
| GSE41437 | ncbi.nlm.nih.gov/geo | Skalsky <i>et al.</i> | PAR-CLIP | human | LCL-BAC | GSM1020023 |
| GSE41437 | ncbi.nlm.nih.gov/geo | Skalsky <i>et al.</i> | PAR-CLIP | human | LCL-BAC-D1 | GSM1020024 |
| GSE41437 | ncbi.nlm.nih.gov/geo | Skalsky <i>et al.</i> | PAR-CLIP | human | LCL-BAC-D3 | GSM1020025 |
| GSE21578 | ncbi.nlm.nih.gov/geo | Hafner <i>et al.</i> | PAR-CLIP | human | HEK293 | GSM545212, GSM545213, GSM545214, GSM545215 |
| GSE43573 | ncbi.nlm.nih.gov/geo | Memczak <i>et al.</i> | PAR-CLIP | human | HEK293 | GSM1065667, GSM1065668, GSM1065669, GSM1065670 |
| GSE43909 | ncbi.nlm.nih.gov/geo | Erhard <i>et al.</i> | PAR-CLIP | human | BCBL-1 | GSM1074233, GSM1074234 |

2.2.2 Analysis of high-throughput experiments

2.2.2.1 miRNA perturbation experiments

High-throughput experiments were collected to measure gene expression alterations after specific miRNA transfection, silencing or knockout. Log₂ fold change values as calculated from differential expression analyses of control versus post-treatment state enabled the formation of miRNA-mRNA positive and negative interactions. 44 microarray studies of distinct experimental conditions (Table 4, Table 5) covering 43 human cell lines and 49 miRNAs were examined to deduce positive and negative miRNA-target interactions. In-house analysis was initiated from microarray raw data (Affymetrix .CEL files). Probe set summarization was implemented using Robust Multi-Array Average (RMA) with R packages *affy*[80] or *oligo*[81]. Annotation of probe sets to Ensembl Gene IDs was accomplished using the chip-specific annotation R-packages *hgu133a2.db*, *hgu133plus2.db* or *hugene10sttranscriptcluster.db*. miRNA-treated and control samples in each experiment were analyzed independently of other cell lines or miRNA treatments. Log₂ fold change ratios and p-values were calculated with *limma* package[82], following package instructions on Single-Channel Designs. Probe sets mapping to the same gene were averaged to calculate its fold change. A log₂ fold change cutoff of ± 1 (>1 or <-1 , respectively), depending on the type of regulation, was applied to determine negative and positive interaction subsets. For GSE8501 experiment conducted in Rosetta-Merck microarrays, error-weighted log₁₀ intensity ratios were retrieved and transformed to log₂-scale.

Ribosome profiling sequencing (RPF-Seq) and RNA-Seq libraries treated with specific miRNA overexpression, 12 experimental conditions in total were retrieved from Eichhorn *et al.*[91], Nam *et al.*[92], Pellegrino *et al.*[93], Zhang *et al.*[94]. To identify positive/negative miRNA interactions, a ± 0.5 log₂ fold change threshold was applied to genes presenting >10 RPKM expression.

Quantitative proteomics datasets (pSILAC) in HeLa cells following the individual overexpression of 5 human miRNAs (let-7b, miR-1, miR-16, miR-30a and miR-155) or knockdown of let-7b (Table 4) were derived from Selbach *et al.*[46]. Positive/negative miRNA interactions were deduced using a ± 1 log₂ fold change threshold respectively.

2.2.2.2 AGO-PAR-CLIP and (s)RNA-Seq expression datasets

AGO-PAR-CLIP datasets from 9 studies, corresponding to 13 cell lines in human species, were derived from GEO[57, 95] and DDBJ[96] repositories. 15 small RNA-Seq and 9 RNA-Seq experiments of similar cell types with PAR-CLIP libraries were analyzed to infer expressed miRNAs and transcripts. (s)RNA-Seq datasets were derived from the ENCODE repository and from a series of studies (Table 7, Table 8). Whole transcriptome depleted from ribosomal RNAs and poly-A selected RNA-Seq libraries were analyzed.

The libraries were initially quality checked using FastQC

(www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapter sequences were retrieved from the original publication or GEO/SRA entries, when available. Contaminants were detected utilizing in-house-developed algorithms and the Kraken suite[97]. Pre-processing was performed utilizing Cutadapt[98]. PAR-CLIP reads were aligned against human reference genome (GRCh37/hg19) with GMAP/GSNAP[99] spliced aligner, appropriately parameterized to identify known and novel splice junctions. microRNA expression was quantified using miRDeep2[100]. Ensembl v75[101] and miRBase v18[102] were used as annotation for genes and microRNAs, respectively. Top expressed miRNAs and AGO-PAR-CLIP data in each cell type, were jointly utilized as input to microCLIP *in silico* framework for miRNA-target identification.

Table 7: Description of small RNA-Seq datasets of similar cell types to PAR-CLIP libraries, analyzed to infer expressed miRNAs. The table displays the source of small RNA-Seq libraries along with its ID, cell type, condition and description.

| Accession | Repository | Cell Type/Tissue | Description |
|----------------|----------------------|------------------|----------------------|
| GSM897079_Rep1 | ncbi.nlm.nih.gov/geo | HeLaS3 | Cervical Carcinoma |
| GSM897079_Rep2 | ncbi.nlm.nih.gov/geo | HeLaS3 | Cervical Carcinoma |
| GSM897073_Rep1 | ncbi.nlm.nih.gov/geo | H1hESC | Embryonic Stem Cells |
| GSM897073_Rep2 | ncbi.nlm.nih.gov/geo | H1hESC | Embryonic Stem Cells |
| GSM973690_Rep3 | ncbi.nlm.nih.gov/geo | MCF7 | Adenocarcinoma |
| GSM973690_Rep4 | ncbi.nlm.nih.gov/geo | MCF7 | Adenocarcinoma |
| GSM897081_Rep1 | ncbi.nlm.nih.gov/geo | MCF7 | Adenocarcinoma |
| GSM897081_Rep2 | ncbi.nlm.nih.gov/geo | MCF7 | Adenocarcinoma |
| SRR2084358 | ncbi.nlm.nih.gov/sra | MCF7 | Adenocarcinoma |
| GSM1020026 | ncbi.nlm.nih.gov/geo | EF3D-AGO2 | Adenocarcinoma |
| GSM1020028 | ncbi.nlm.nih.gov/geo | LCL-BAC | Adenocarcinoma |
| GSM1020029 | ncbi.nlm.nih.gov/geo | LCL-BAC-D1 | Adenocarcinoma |
| GSM1020030 | ncbi.nlm.nih.gov/geo | LCL-BAC-D3 | Adenocarcinoma |

Table 8: Description of RNA-Seq datasets of similar cell types to PAR-CLIP libraries, analyzed to infer expressed transcripts. The table displays the source of RNA-Seq datasets along with its ID, cell type, condition and description.

| Accession | Repository | Cell Type/Tissue | Condition | Description |
|--|-------------------------|------------------|--------------------------|---------------------------------------|
| SRR837795 | ncbi.nlm.nih.gov/sra | LCL-BAC-D1 | miR-BHRF1-1 mutant virus | LCL infected with an EBV B95-8 BACmid |
| SRR837796 | ncbi.nlm.nih.gov/sra | LCL-BAC-D1 | miR-BHRF1-1 mutant virus | LCL infected with an EBV B95-8 BACmid |
| SRR837797 | ncbi.nlm.nih.gov/sra | LCL-BAC-D2 | miR-BHRF1-2 mutant virus | LCL infected with an EBV B95-8 BACmid |
| SRR837798 | ncbi.nlm.nih.gov/sra | LCL-BAC-D3 | miR-BHRF1-3 mutant virus | LCL infected with an EBV B95-8 BACmid |
| SRR837794 | ncbi.nlm.nih.gov/sra | LCL-BAC | NA | LCL infected with an EBV B95-8 BACmid |
| ENCFF002DKY & ENCFF002DKX | encodeproject.org | MCF7 | NA | Adenocarcinoma |
| ENCFF000FOV & ENCFF000FOM | encodeproject.org | HeLaS3 | NA | Cervical Carcinoma |
| wgEncodeCshlLongRnaSeqH1hescCellPapFastqRep1 | hgdownload.cse.ucsc.edu | H1hESC | NA | Embryonic Stem Cells |
| GSM1370364 | ncbi.nlm.nih.gov/geo | HEK293 | NA | Embryonic Kidney Cells |

2.2.2.3 PARS experimental data

In order to demarcate RNA Secondary Structures (RSS) of AGO-bound regions compared to a set of negative miRNA sites on mRNA transcripts, respective PARS scores as introduced by Wan *et al.*[103] were estimated (GEO accessions GSM1226157, GSM1226158). In this approach, AGO-binding efficiency is revealed by RSS signatures observed on mRNA transcripts, since increased structural accessibility is expected in functional conformations (Figure 13).

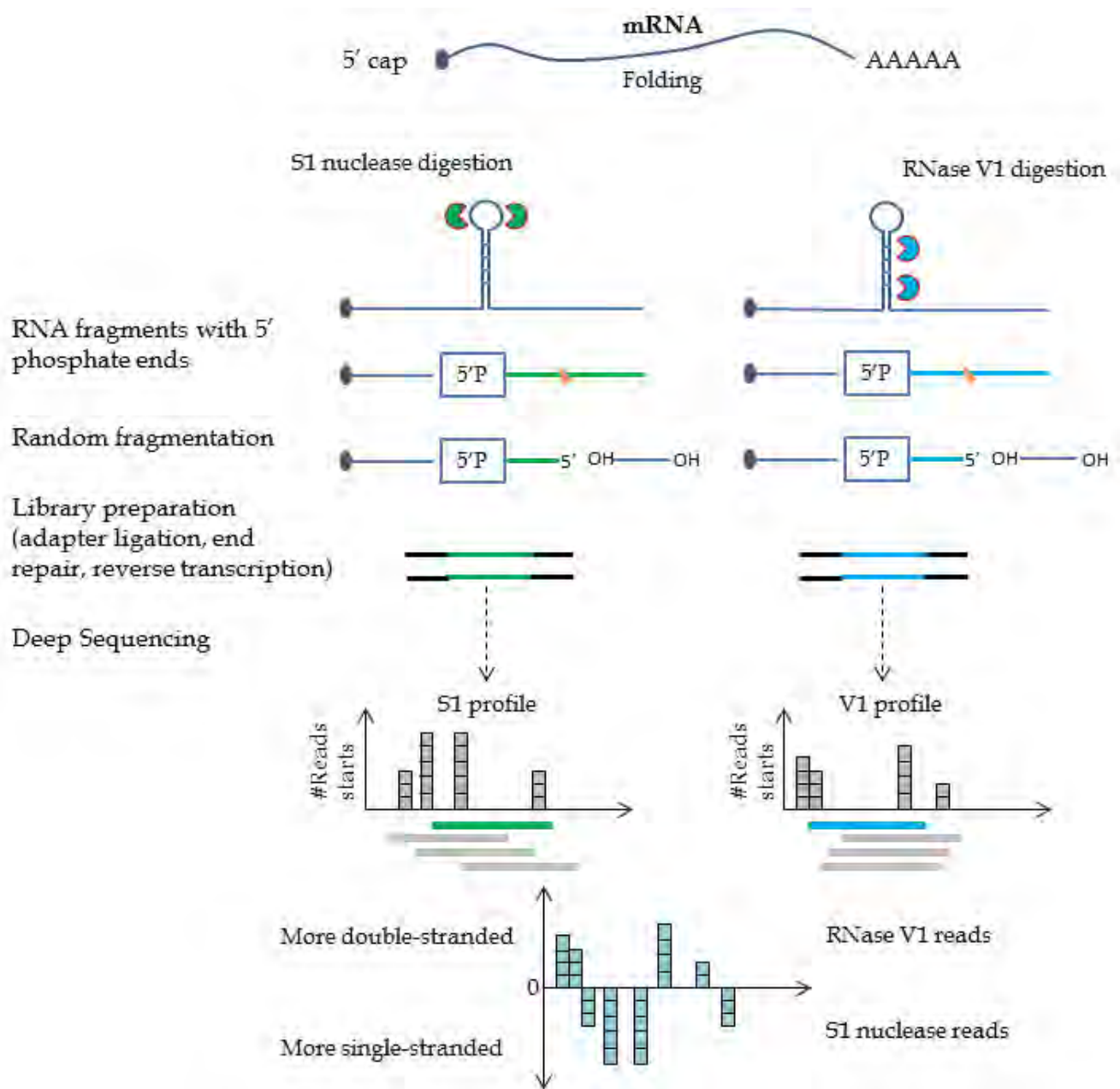


Figure 13: Overview of PARS experiment. This figure has been designed for the purpose of this dissertation.

The identification of single- or double-stranded regions across the human transcriptome was derived from deeply sequenced RNA fragments generated from RNase S1 or V1 nuclease treatment of GM12878 cells respectively.

Raw reads of 51nt length, accordingly pre-processed for quality control and contaminant removal, were aligned against human reference genome (GRCh37/hg19) with GSNAP spliced aligner. This analysis resulted in ~130M uniquely mapped PE-sequenced fragments per sample. In order to derive structural signals in RNase S1 or V1 nuclease experiments at single base resolution, single (S1) and double (V1) stranded raw reads initiating on each

nucleotide were calculated. The number of PARS tags per sample starting at each base were normalized by sequencing library depth. These base intensities were subsequently combined into the formula described by Wan *et al.* to compute PARS scores.

RNA secondary structures (RSS) were defined by estimated PARS scores in the vicinity of PAR-CLIP-derived miRNA binding sites in 4 lymphoblastoid cell lines from the study of Skalsky *et al.*[53]. miRNA-mRNA interactions were identified in both T-to-C and non-T-to-C PAR-CLIP clusters, corresponding to transcripts with >1 TPM expression in GM12878 cells. For expressed miRNAs (≥ 50 aligned reads per miRNA) in respective EFD3-AGO2, LCL-BAC, LCL-BAC-D1 and LCL-BAC-D3 EBV infected lymphoblastoid cells, collapsed miRNA binding sites residing within the PAR-CLIP clusters were included. For the performed comparisons, negative MREs extracted from different high-throughput miRNA perturbation experiments were incorporated. MREs utilized for the assessment of RSS signatures on AGO-bound clusters and the derivation of (non-)functional conformations of miRNA-target base pairings, were localized on coding and 3'UTR regions. The examined sites had to present S1 and V1 signals in at least half of their occupied bases.

sRNA-Seq and RNA-Seq datasets were retrieved from ENCODE consortium (GEO accession numbers GSM605625, GSM1020026, GSM1020027, GSM1020028, GSM1020029, GSM1020030).

2.2.3 microCLIP *in silico* framework

Feature set description. A set of 131 descriptors with non-zero variance was included in microCLIP. The extracted features were retrieved from positive/negative miRNA interactions, identified on AGO-bound locations in different PAR-CLIP datasets. They comprised PAR-CLIP-specific descriptors, such as substitution ratios and distance of conversions from the MRE start, as well as coverage metrics. Aggregate substitution ratios, positions and distances independent of the transition type were also included. In order to estimate MRE and AGO-peak respective sequencing coverage, normalized RPKM values for miRNA-target sites and clusters were calculated.

Moreover, single base and dinucleotide contents for miRNA binding and respective flanking regions, complexity features for the MRE and proximal upstream/downstream sequences were introduced to microCLIP model. BLAST's DUST score[104] and Shannon-Wiener Index[105] constituted measurements for masking sequence complexity. Other descriptors were formed to represent energy-related variables for the duplex structure, while metrics capturing sequence content skewness/asymmetry (GC-skew, AT-skew, purine-skew, Ks-skew) and biases of codon usage were added. Entropy, enthalpy, free energy and melting temperature (T_m) thermodynamic properties were calculated for MRE sequences in R.

miRNA-target hybrids were associated with different descriptors, such as the binding type, duplex structure energy calculated with the Vienna package[106], positions and nucleotide composition of (un)paired nucleotides. Distinct features have been established to model

(mis)matches, bulges, loops and wobble pairs for miRNA-MRE hybrid structure and sub-domains encountered in the duplex. The distinct domains for miRNA sequences, as defined by microCLIP, are: (i)seed region (2-8 positions), (ii)central region (9-12 positions), (iii)3' supplementary region (13-16 positions), (iv)tail region (17-3'miRNA end) (Figure 14). Similar regions were designated on the MREs based on the miRNA binding anchors upon duplex formation.

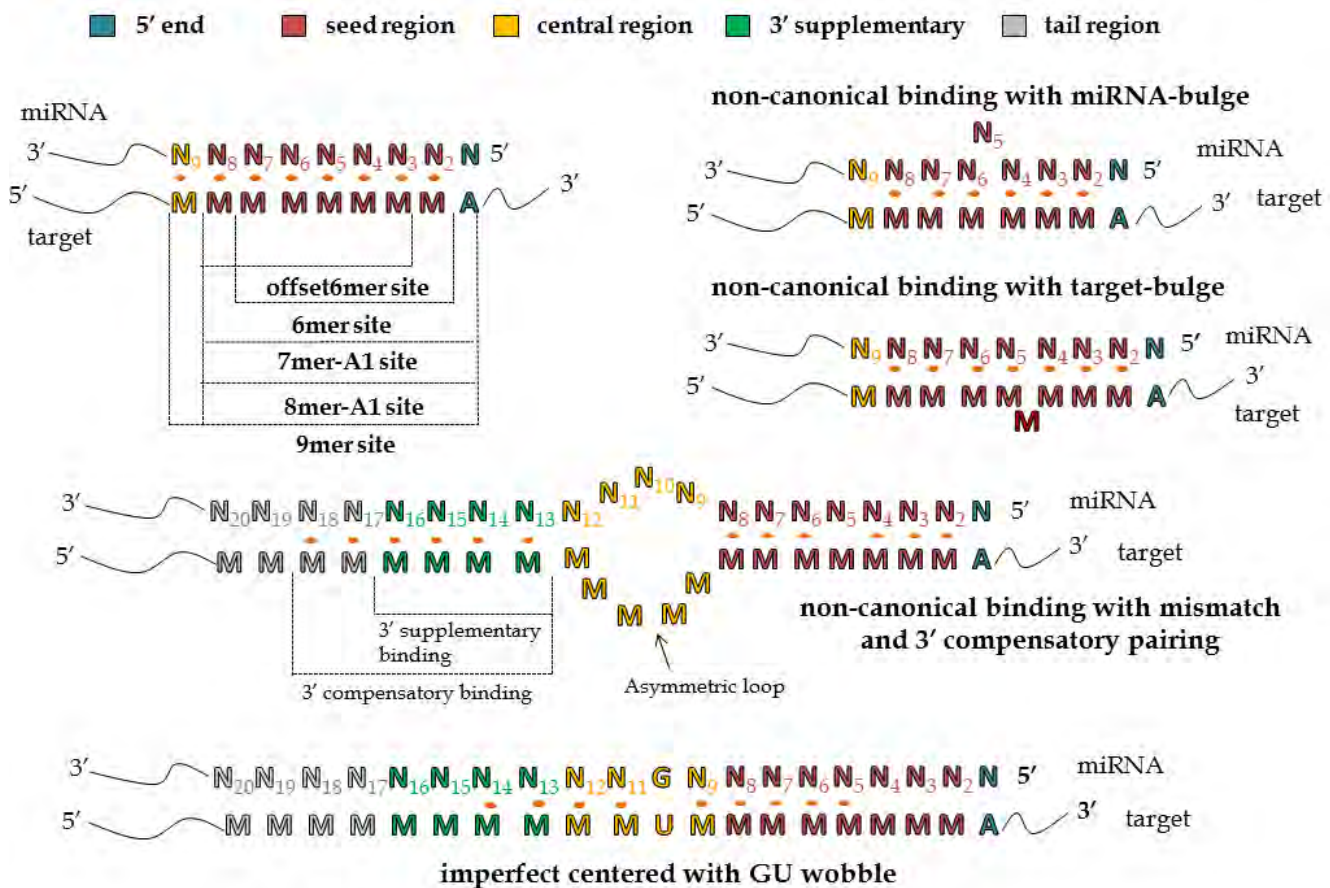


Figure 14: Snapshot of the different miRNA binding types formed according to miRNA specific sub-domains. This figure has been designed for the purpose of this dissertation.

Current approach incorporates conservation of the MRE and upflank/downflank-MRE regions. phastCons pre-computed scores from genome-wide multiple alignments were downloaded from the UCSC repository[107] in bigwig format and were utilized to deduce respective evolutionary rates. Conservation signals were computed as mean intensities of the phastCons base-wise scores on miRNA targeted regions as well as their flanking regions. The conservation of the 5' MRE binding nucleotides was independently modeled. microCLIP integrates additional features corresponding to the location of the MRE within the AGO-enriched cluster and binding length ratios of miRNA and/or target regions.

Description of the algorithm. microCLIP operates on AGO-PAR-CLIP sequencing reads, requiring a SAM/BAM alignment file and a list of miRNAs as minimum input. It initially seeks for AGO-enriched regions and resolves coverage and observed transitions. A sensitive pipeline is adopted to scan read clusters for putative targeted sites including a wide range of binding types. The algorithm supports an extended set of (non-)canonical matches including 6mer to 9mer, offset 6mer, 3'supplementary and compensatory sites as well as (im)perfect centered bindings (Table 9).

Table 9: Description of the binding types supported by microCLIP.

| Binding Type | Description |
|---------------------------------|---|
| 9mer 3prime | 9mer canonical with 3' supplementary binding |
| 9mer | 9mer canonical |
| 9mer GU | base pairing in 1-9 positions with a GU wobble pair |
| 9mer nonCanonical | base pairing in 1-9 positions with a target bulge and/or a GU wobble pair |
| 8mer 3prime | 8mer/8mer1A canonical with 3' supplementary binding |
| 8mer | 8mer canonical |
| 8mer1A | 7mer canonical with additional A in position 1 |
| 8mer GU | base pairing in 1-8 or 2-9 positions with a GU wobble pair |
| 8mer nonCanonical | 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/7mer1A canonical with 3' supplementary binding |
| 7mer | 7mer canonical |
| 7mer1A | 6mer canonical with an additional A in position 1 |
| 7mer GU | base pairing in 2-8 positions with a GU wobble pair |
| 7mer nonCanonical | base pairing in 1-8 positions with a mismatch or miRNA bulge and/or a target bulge |
| 7mer nonCanonical GU | 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 with 3' supplementary binding |
| 6mer | 6mer canonical |
| offset6mer | 6mer base pairing in 3-8 positions |
| 6mer nonCanonical 3prime | base pairing in 2-8 positions with a mismatch or miRNA bulge and/or a target bulge, with 3' supplementary binding |
| centered | base pairing in 4-15 positions with at least 8 consecutive matches |
| imperfect centered | base pairing in 4-15 positions with at least 8 matches and/or less than 2 GU wobble pairs |

microCLIP extracts features for each candidate MRE and subsequently scores sites through a super learning scheme. The adopted framework incorporates two distinct levels of classification. The first layer comprises a group of 9 different nodes (base classifiers), which are aggregated in the meta-classifier of the second layer. The learning procedure is

decentralized through the distinct nodes and relevant base classifiers that specialize in different subsets of features (Figure 15).

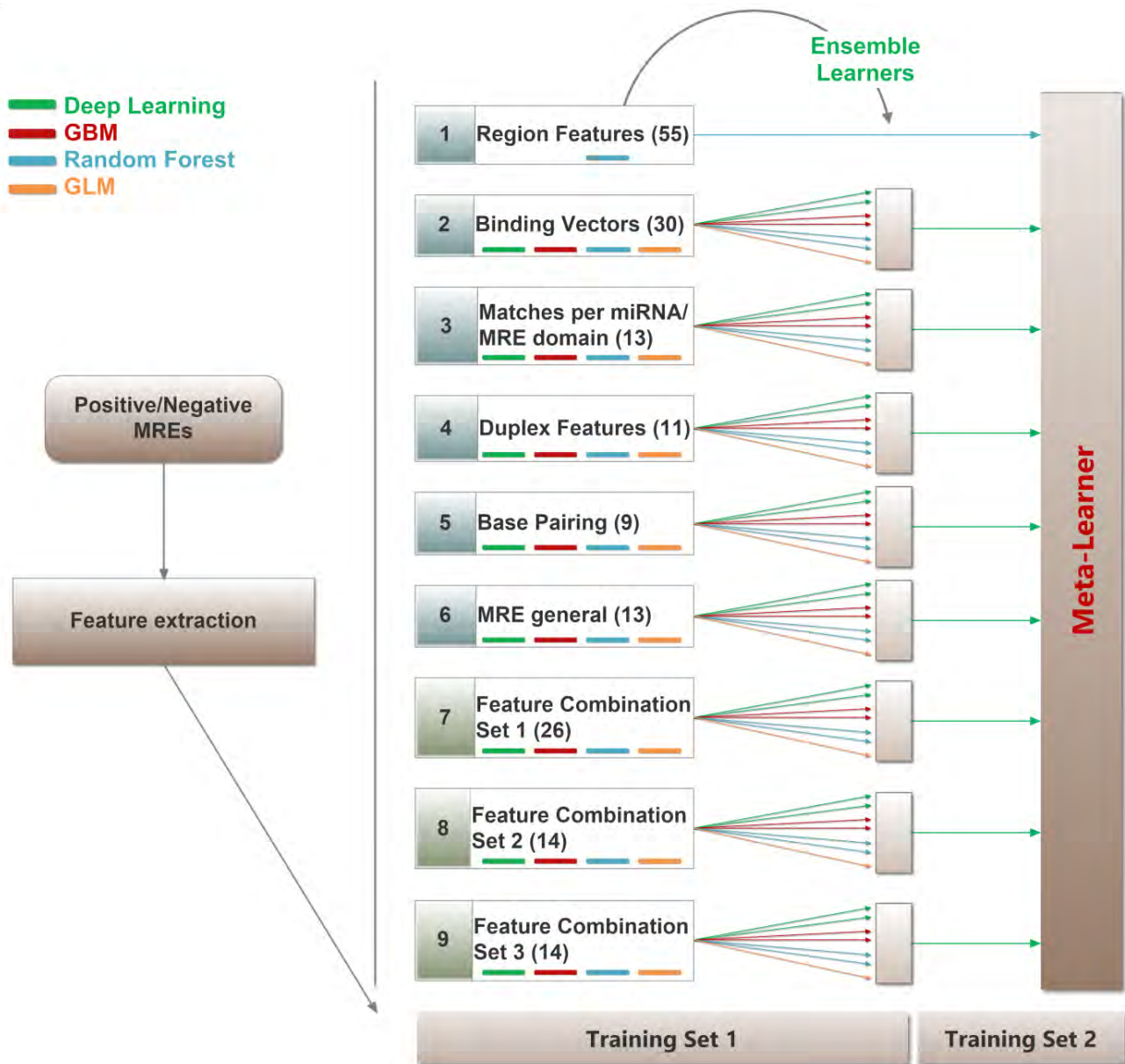


Figure 15: microCLIP *in silico* framework. Separate subsets of the positive/negative miRNA interactions were used to train the distinct levels of the algorithm's modeling. 9 base classifiers in the first layer comprise characteristic feature subsets that assemble into the GBM meta-learner of the second layer. A super learning scheme is utilized in 8 of the 9 base nodes, weighing outputs from seven individual models. 'Region features' node corresponds to an RF classification scheme and consists of CLIP-sequencing-derived features. Five base models (2-6) were designed for MRE specific features: 'Binding Vectors' describe the (un)paired positions along the miRNA/MRE hybrid; 'Matches per miRNA/MRE domain' contain attributes of miRNA-target structure and sub-domains; 'Duplex Features' include free energy, secondary structure and AU base pairing features for miRNA and/or target; 'Base pairing' encompasses composition descriptors of (un)paired nucleotides; 'MRE general' incorporates general MRE-related descriptors. Three supplementary classifiers

(‘Feature Combination Set 1-3’) comprise unique combinations of features found in base nodes 1-6 (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

‘Region Features’ node comprises CLIP-Seq-derived features, such as RPKM coverage, substitution frequencies and region-related descriptors, including nucleotide composition, conservation, sequence energy, complexity, content asymmetry, and biases of codon usage. A set of five additional base classifiers were designed for MRE specific features. Binary binding vectors of miRNA/MRE hybrid were separately incorporated in a base classifier (‘Binding Vectors’). Each vector element corresponds to one (un)paired position in the duplex. Matches per miRNA/MRE sub-domain were added to a distinct base classifier introducing a group of 13 features regarding total and consecutive matches in the miRNA-target structure as well as in MRE and miRNA relevant sub-domains. Another base model consists of miRNA-target duplex descriptors (‘Duplex Features’) including miRNA-target duplex structure energy, bulges, internal loops, GU wobbles and AU base pairing features for the specified miRNA and/or target and relevant sub-domains. The ‘Base pairing’ node encompasses composition descriptors (A, T, G, C) of the (un)paired nucleotides. An extra base learner incorporates MRE general descriptors such as the degree of overlap with the respective cluster, conservation of MRE bound nucleotides, MRE location within the cluster, MRE binding type as well as metrics for duplex paired nucleotides content asymmetry/skewness. The latter five base models are dedicated to the determination of genuine miRNA binding sites. Non-overlapping feature sets from the aforementioned base nodes are combined into three supplementary classifiers also incorporated into microCLIP framework. A table summarizing the incorporated features, associated with the conceptual framework they belong, is presented below:

Table 10: Description of features incorporated in microCLIP.

| feature# | base classifier node | feature description |
|----------|---|---|
| 1 | Region Features (55 features) | MRE region GC-skew |
| 2 | | MRE region Purine-skew |
| 3 | | MRE region Ks-skew |
| 4 | | Upflank MRE region Purine-skew |
| 5 | | Upflank MRE region Ks-skew |
| 6 | | MRE DUST score |
| 7 | | MRE region AT-skew |
| 8 | | MRE dS |
| 9 | | MRE Tm |
| 10 | | Codon Adaptation Index per codon usage bias |
| 11 | | Dinucleotide AA MRE content |
| 12 | | Dinucleotide AC MRE content |
| 13 | | Dinucleotide AG MRE content |

| | |
|----|--|
| 14 | Dinucleotide AT MRE content |
| 15 | Dinucleotide CA MRE content |
| 16 | Dinucleotide CC MRE content |
| 17 | Dinucleotide CT MRE content |
| 18 | Dinucleotide GA MRE content |
| 19 | Dinucleotide GC MRE content |
| 20 | Dinucleotide GT MRE content |
| 21 | Dinucleotide TA MRE content |
| 22 | Dinucleotide AC content upflank of MRE |
| 23 | Dinucleotide AT content upflank of MRE |
| 24 | Dinucleotide GA content upflank of MRE |
| 25 | Dinucleotide GC content upflank of MRE |
| 26 | Dinucleotide GT content upflank of MRE |
| 27 | Dinucleotide TA content upflank of MRE |
| 28 | Dinucleotide TC content upflank of MRE |
| 29 | Dinucleotide TG content upflank of MRE |
| 30 | A content upflank of MRE |
| 31 | C content upflank of MRE |
| 32 | G content upflank of MRE |
| 33 | T content upflank of MRE |
| 34 | A or G content upflank of MRE |
| 35 | A or T content upflank of MRE |
| 36 | A content in MRE |
| 37 | A or G content in MRE |
| 38 | A or T content in MRE |
| 39 | C content in MRE |
| 40 | G content in MRE |
| 41 | G or T content in MRE |
| 42 | T content in MRE |
| 43 | Average conservation in MRE seed region |
| 44 | Average conservation downstream of MRE region |
| 45 | Average conservation upstream of MRE region |
| 46 | Minimum conservation in MRE seed region |
| 47 | Minimum conservation downstream of MRE region |
| 48 | Minimum conservation upstream of MRE region |
| 49 | MRE coverage (RPKM) |
| 50 | Cluster overlapping reads |
| 51 | MRE coverage (RPKM) per cluster coverage (RPKM) |
| 52 | Cluster length |
| 53 | Minimum distance - sum of all substitutions +/- 20nt of MRE start |
| 54 | Sum of all substitutions +/- 20nt of MRE start - minimum distance |

| | | |
|----|--|--|
| 55 | | T-to-C substitutions in MRE region |
| 56 | | Binding event in MRE position 2 |
| 57 | | Binding event in MRE position 3 |
| 58 | | Binding event in MRE position 4 |
| 59 | | Binding event in MRE position 5 |
| 60 | | Binding event in MRE position 6 |
| 61 | | Binding event in MRE position 7 |
| 62 | | Binding event in MRE position 8 |
| 63 | | Binding event in MRE position 10 |
| 64 | | Binding event in MRE position 11 |
| 65 | | Binding event in MRE position 12 |
| 66 | | Binding event in MRE position 17 |
| 67 | | Binding event in MRE position 18 |
| 68 | | miRNA unpaired position 5 |
| 69 | | miRNA unpaired position 6 |
| 70 | Binding Vectors (30 features) | miRNA unpaired position 7 |
| 71 | | miRNA unpaired position 8 |
| 72 | | Base at MRE position 25 |
| 73 | | Base at MRE position 26 |
| 74 | | Base at MRE position 27 |
| 75 | | Base at MRE position 28 |
| 76 | | Base at MRE position 29 |
| 77 | | Base at miRNA position 13 |
| 78 | | Base at miRNA position 14 |
| 79 | | Base at miRNA position 15 |
| 80 | | Base at miRNA position 19 |
| 81 | | Base at miRNA position 4 |
| 82 | | Base at miRNA position 5 |
| 83 | | Base at miRNA position 6 |
| 84 | | Base at miRNA position 7 |
| 85 | | Base at miRNA position 12 |
| 86 | | Total matches |
| 87 | | Max consecutive matches |
| 88 | | Match in position 10 of miRNA |
| 89 | | Match in position 2 of miRNA |
| 90 | Matches per miRNA/MRE domain (13 features) | Consecutive unpaired bases of non-seed region |
| 91 | | Consecutive matches per total matches |
| 92 | | Matches in seed region per total matches |
| 93 | | Consecutive matches in seed region |
| 94 | | Consecutive matches in non-seed region per total matches |
| 95 | | Matches in seed region |

| | | |
|-----|---|--|
| 96 | | Matches in central MRE region |
| 97 | | Matches in 3' MRE |
| 98 | | Consecutive matches in seed region per max consecutive matches |
| 99 | | AU frequency in MRE region |
| 100 | | GC frequency in MRE region |
| 101 | | GU wobble frequency in MRE region |
| 102 | | Internal loop max length in MRE region |
| 103 | Duplex Features (11 features) | AU frequency in seed region |
| 104 | | AU frequency in MRE region excluding seed |
| 105 | | GC frequency in central MRE region |
| 106 | | GC frequency in MRE region excluding seed |
| 107 | | GU wobble frequency in tail region |
| 108 | | Length of MRE binding region |
| 109 | | Bulge positions in MRE region (%) |
| 110 | | Matches in MRE Ks-skew |
| 111 | | miRNA unpaired A |
| 112 | | miRNA unpaired C |
| 113 | Base Pairing (9 features) | miRNA unpaired G |
| 114 | | miRNA unpaired T |
| 115 | | miRNA matches A |
| 116 | | miRNA matches T |
| 117 | | miRNA matches C |
| 118 | | miRNA matches G |
| 119 | | miRNA unpaired position 2 |
| 120 | | miRNA unpaired position 3 |
| 121 | | miRNA unpaired position 4 |
| 122 | | Matches in MRE Purine-skew |
| 123 | | Duplex structure energy |
| 124 | | Length of miRNA binding region |
| 125 | MRE General (13 features) | Distance of MRE start from cluster start |
| 126 | | Nucleotides of MRE that overlap with cluster region |
| 127 | | Length of MRE binding region per cluster length |
| 128 | | Nucleotides of MRE that overlap with cluster region (%) |
| 129 | | Average conservation of whole MRE |
| 130 | | Average conservation of 5' MRE |
| 131 | | Binding Type |
| 1 | | Binding event in MRE position 4 |
| 2 | Feature Combination Set 1 (26 features) | Binding event in MRE position 11 |
| 3 | | Binding event in MRE position 18 |
| 4 | | miRNA unpaired position 5 |
| 5 | | Base at MRE position 28 |

| | | |
|-------|--|---|
| 6 | | Base at MRE position 29 |
| 7 | | Base at miRNA position 13 |
| 8 | | Base at miRNA position 6 |
| 9 | | Base at miRNA position 7 |
| 10 | | Base at miRNA position 12 |
| 11 | | Max consecutive matches |
| 12 | | Binding Type |
| 13 | | Consecutive matches per total matches |
| 14 | | Consecutive matches in seed region per max consecutive matches |
| 15 | | Matches in 3' MRE |
| 16 | | AU frequency in seed region |
| 17 | | GU wobble frequency in MRE region |
| 18 | | AU frequency in MRE region excluding seed |
| 19 | | Length of MRE binding region |
| 20 | | miRNA unpaired C |
| 21 | | miRNA matches A |
| 22 | | miRNA matches C |
| 23 | | miRNA matches G |
| 24 | | Duplex structure energy |
| 25 | | Length of MRE binding region per cluster length |
| 26 | | Average conservation of whole MRE |
| <hr/> | | |
| 1 | Feature Combination Set 2 (14 features) | MRE dS |
| 2 | | G content upflank of MRE |
| 3 | | A or T content upflank of MRE |
| 4 | | Average conservation upstream of MRE region |
| 5 | | MRE coverage (RPKM) |
| 6 | | Sum of all substitutions +/- 20nt of MRE start - minimum distance |
| 7 | | T-to-C substitutions in MRE region |
| 8 | | Binding event in MRE position 3 |
| 9 | | AU frequency in MRE region |
| 10 | | GC frequency in MRE region |
| 11 | | miRNA matches T |
| 12 | | Matches in seed region per total matches |
| 13 | | Consecutive matches in seed region |
| 14 | | Nucleotides of MRE that overlap with cluster region |
| <hr/> | | |
| 1 | Feature Combination Set 3 (14 features) | Binding event in MRE position 6 |
| 2 | | Match in position 10 of miRNA |
| 3 | | Match in position 2 of miRNA |
| 4 | | Matches in seed region |
| 5 | | Matches in central MRE region |
| 6 | | GC frequency in central MRE region |

| | |
|----|---|
| 7 | GC frequency in MRE region excluding seed |
| 8 | Matches in MRE Ks-skew |
| 9 | miRNA unpaired A |
| 10 | miRNA unpaired T |
| 11 | Matches in MRE Purine-skew |
| 12 | Length of miRNA binding region |
| 13 | Distance of MRE start from cluster start |
| 14 | Consecutive unpaired bases of non-seed region |

8 of the 9 base nodes adopt a super learning scheme that assembles the output of seven individual Random Forest (RF), Generalized Linear Model (GLM), Gradient Boosting Model (GBM), Deep Learning (DL) classifiers (2 RF, 2 GBM, 2 DL, 1 GLM models). The 'Region features' is analyzed by an RF classification scheme. The retrieved scores from each node are aggregated in a final GBM meta-classifier.

Model Training. The DL models developed for the microCLIP framework adopt a feed-forward multi-layer architecture. The input layers match the respective feature space and values are subsequently propagated within three hidden layers. A rectifier activation function was utilized to retrieve weighted combinations of the inputs transmitted to interconnected neuron units. Dropout regularization was added to achieve model optimization and avoid overfitting. A cross entropy cost-function was selected to adapt weights during the learning process by minimizing the loss. Bernoulli distribution function was used along with cross entropy (log-loss) to model the response variables. The output layer at the end of the network applies a Softmax activation function so that each neuron (predicted class) results in a probabilistic interpretation. The DL network depth, width and topology as well as activation functions and learning parameters were modeled with a tuning-in grid search algorithm using H2O[108] R package. The RF, GBM, GLM learning models were developed, parameterized and tuned with the caret[109] and H2O[108] R packages.

Base classifiers were trained against a collection of 8,693 positive and 21,789 negative miRNA interactions. The final GBM meta-learner that aggregates the base classifier outcomes was trained against an independent dataset comprising 3,276 and 6,702 positive and negative instances respectively. Ten-fold cross-validation was performed on the training data to estimate each model's accuracy and finalize the algorithm's learning architecture. The performance of the model was assessed against independent test sets comprising ~5,495 instances in total. The composition of respective training/test sets is provided in Table 11. Training and testing of microCLIP have been performed on independent sets of targeted MRE regions.

Table 11: Summary of training/test sets utilized for microCLIP deployment.

| | miRNAs in interactions | | miRNA-target pairs | |
|---------------------------------------|------------------------|------|--------------------|-------|
| | Training | Test | Training | Test |
| Positive Instances | | | | |
| <i>Direct Techniques</i> | 244 | 158 | 4,707 | 2,017 |
| <i>miRNA perturbation experiments</i> | 47 | 5 | 7,262 | 679 |
| Negative Instances | | | | |
| <i>Background CLIP-Seq</i> | 393 | 122 | 22,575 | 1,591 |
| <i>miRNA perturbation experiments</i> | 44 | 23 | 5,916 | 1,208 |

2.2.4 miRNA interactions from *in silico* implementations

microCLIP performance was evaluated against MIRZA[55], microMUMMIE[56] and PARma[58] computational approaches. A set of 7 PAR-CLIP HEK293 libraries obtained from Kishore *et al.*[34] and Memczak *et al.*[110] studies (GEO accessions GSM714644, GSM714645, GSM714646, GSM714647, GSM1065667, GSM1065668, GSM1065669 and GSM1065670) was utilized. The proposed settings for each implementation were retrieved from the relevant publications.

The MIRZA biophysical model was executed in the “nouupdate” mode. The algorithm provides an optional parameterization to introduce miRNA expression profiles. Two different runs of MIRZA were realized, with and without cell type-specific miRNA expression values that were extracted from the CLIPZ web server (<http://www.clipz.unibas.ch>). MIRZA input data were 51-nt AGO-bound sequences centered on T-to-C sites and mature miRNA sequences of 21nt length as reported in the model’s restrictions. The “target frequency” score was utilized to evaluate the quality of MIRZA-detected sites.

microMUMMIE algorithm was tested in both Viterbi and posterior decoding modes. Following microMUMMIE’s constraints, PARalyzer v1.5[57] was utilized to define the set of T-to-C AGO-enriched peaks. An extra prerequisite annotation step to complement PARalyzer detected clusters was implemented with the PARpipe tool (<https://github.com/ohlerlab/PARpipe>). Derived files, comprising annotated AGO clusters with positions of T-to-C transitions, constituted the input of the microMUMMIE core algorithm. Predictions with signal-to-noise ratio (SNR, generally correlated with sensitivity) equal to 9.95 were retained, while posterior probabilities were utilized for the evaluation of microMUMMIE’s performance.

PARma was applied on AGO-PAR-CLIP aligned data that were prepared following the

algorithm's described format. The required input files contained genomic locations of aligned CLIP-reads along with positions of observed conversion sites. PARma predictions are coupled with Cscore and MAScore scores for the cluster and miRNA-seed family activity, respectively. The latter score was utilized for PARma-detected miRNA-target sites evaluation. Precompiled (non)conserved miRNA-site context++ scores for representative transcripts were downloaded from the Targetscan v7.2 site (http://www.targetscan.org/cgi-bin/targetscan/data_download.vert72.cgi). Targetscan v7 algorithm was additionally executed following the proposed settings in order to cover a greater transcript collection, as well as the whole spectrum of Targetscan-detected interactions including 6mer sites. Gene annotation files were retrieved from the Targetscan v7.2 official download page, and the miRNA seed sequence file that is a prerequisite for the execution of the model was provided by Targetscan developers. The local Targetscan run complements the precompiled data with miRNA-target interactions on transcripts presenting the longest 3'UTR, in cases they are not deposited on the online repository.

2.3 Implementation of microT, a de novo miRNA target prediction algorithm

Computational methodologies devoted to miRNA-target characterization unambiguously provide the backbone for many miRNA-related studies. An accurate de novo miRNA target prediction algorithm contributes as an extra boost to study miRNA function, by eliminating time and experimental cost. The last 15 years, a multitude of computational approaches have emerged, aiming to accurately characterize miRNA targets. However, even the most sophisticated implementations still achieve a far from perfect predictive accuracy[5] followed by an increased number of false positive predictions.

Most of the current approaches heavily rely on decisive features towards miRNA target detection such as miRNA seed complementarity, secondary structure and evolutionary conservation. Their predictions are often radically diverse, due to the incorporation of different experimental datasets and mathematical models in the training process. Targetscan[20] is a leading *in silico* target prediction method, however it detects miRNA-target pairs with perfect seed complementarity and ignores non-canonical sites. Therefore, a large portion of functional miRNA binding-events is disregarded. Also, it does not include in its training process recently developed experimental procedures, such as AGO-CLIP-Seq and CLASH methods, that provide a wealth of characteristics regarding the AGO-bound enriched/preferred regions. Recently developed models try to fill the existing gap by incorporating only a small part of the publicly available CLIP-Seq datasets. Most of them prefer to detect seed-based binding sites to scale down the false positive rate, however their predictive accuracy still remains low[20].

To this end, a novel miRNA target prediction algorithm is presented in this thesis, that circumvents pitfalls and limitations of current approaches. microT Super Learning

framework maintains and upgrades the pipeline adopted in microCLIP by enhancing the training with even more high-throughput experiments under a tissue-specific scheme. The new model characterizes interactions with stronger functional efficacy and correctly detects 1.5-fold more experimentally validated target sites when juxtaposed against Targetscan v7 and DIANA-microT-CDS^{22,23}.

2.3.1 Dataset collection

miRNA-targeted regions, utilized in the training and evaluation of microT Super Learning framework, were extracted following the methodology described in microCLIP (Methods 2.2.1). More precisely, AGO-enriched regions derived from AGO-CLIP-Seq libraries were coupled with differentially expressed mRNAs extracted from miRNA specific high-throughput experiments across 34 different cell types and 11 tissues. 113 CLIP-Seq libraries (80 HITS-CLIP, 33 PAR-CLIP), derived from DIANA-TarBase and microCLIP deployment, as well as from the analysis of a subsequent set of 9 publicly available datasets corresponding to 4 different cell types[111], were incorporated.

In order to quantify miRNA-induced mRNA expression changes and to identify functional binding sites, 110 miRNA perturbation experiments were incorporated (91 microarrays, 15 RNA-Seq, 4 RIP-Seq) and 3 ribosome profiling sequencing (RPF-Seq) libraries (Table 12). Approximately 40 of the aforementioned datasets were re-analyzed according to the methodology described in section Methods 2.2.2, while the rest were derived from the analysis displayed in microCLIP deployment.

Table 12: Summary of the collected experiments in human species upon specific miRNA deregulation. The datasets were utilized to extract a training set of positive and negative MRE regions for microT deployment.

| Accession | Repository | Authors | Experiment | Cell Type | miRNA | miRNA treatment | Post-Transfection Cell Harvest Time/Experimental Condition |
|-----------|----------------------|----------------------------|-------------|-----------|-----------------|-----------------|--|
| GSE12400 | ncbi.nlm.nih.gov/geo | Sander et al | microarrays | CCL86 | hsa-miR-26a-5p | Overexpression | 72h |
| GSE12400 | ncbi.nlm.nih.gov/geo | Sander et al | microarrays | CRL1432 | hsa-miR-26a-5p | Overexpression | 72h |
| GSE12400 | ncbi.nlm.nih.gov/geo | Sander et al | microarrays | CRL1596 | hsa-miR-26a-5p | Overexpression | 72h |
| GSE35948 | ncbi.nlm.nih.gov/geo | Misiewicz-Krzeminska et al | microarrays | H929 | hsa-miR-214-3p | Overexpression | NA |
| GSE16674 | ncbi.nlm.nih.gov/geo | Navarro et al | microarrays | K562 | hsa-miR-34a-5p | Overexpression | 24h |
| GSE56268 | ncbi.nlm.nih.gov/geo | Schneider et al | microarrays | P3HR1 | hsa-miR-28-5p | Overexpression | 12h, 24h |
| GSE27718 | ncbi.nlm.nih.gov/geo | Gaziel-Sovran et al | microarrays | 131/4-5B1 | hsa-miR-30d-5p | Overexpression | 60h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson et al | microarrays | H4 | hsa-miR-103a-3p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson et al | microarrays | H4 | hsa-miR-107 | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson et al | microarrays | H4 | hsa-miR-15b-3p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson et al | microarrays | H4 | hsa-miR-16-5p | Overexpression | 48h |
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson et al | microarrays | H4 | hsa-miR-195-5p | Overexpression | 48h |

| | | | | | | | |
|-----------------|----------------------|------------------------|-------------|------------|-----------------|----------------|---------------------------------------|
| GSE42823 | ncbi.nlm.nih.gov/geo | Nelson et al | microarrays | H4 | hsa-miR-320b | Overexpression | 48h |
| GSE34482 | ncbi.nlm.nih.gov/geo | Choudhury et al | microarrays | SW1783 | hsa-miR-376a-5p | Overexpression | 24h |
| GSE34482 | ncbi.nlm.nih.gov/geo | Choudhury et al | microarrays | U87 | hsa-miR-376a-5p | Overexpression | 24h, 72h |
| GSE19693 | ncbi.nlm.nih.gov/geo | Chen et al | microarrays | U87 | hsa-miR-20a-5p | Overexpression | NA |
| GSE19693 | ncbi.nlm.nih.gov/geo | Chen et al | microarrays | HS683 | hsa-miR-20a-5p | Overexpression | NA |
| GSE35208 | ncbi.nlm.nih.gov/geo | Lin et al | microarrays | U87-2M1 | hsa-miR-10b-5p | Inhibition | NA |
| - | psilac.mdc-berlin.de | Selbach et al | microarrays | HeLa | hsa-let-7b-5p | Overexpression | 8h, 32h |
| - | psilac.mdc-berlin.de | Selbach et al | microarrays | HeLa | hsa-miR-1-3p | Overexpression | 8h, 32h |
| - | psilac.mdc-berlin.de | Selbach et al | microarrays | HeLa | hsa-miR-155-5p | Overexpression | 8h, 32h |
| - | psilac.mdc-berlin.de | Selbach et al | microarrays | HeLa | hsa-miR-16-5p | Overexpression | 8h, 32h |
| - | psilac.mdc-berlin.de | Selbach et al | microarrays | HeLa | hsa-miR-30a-5p | Overexpression | 8h, 32h |
| GSE14537 | ncbi.nlm.nih.gov/geo | Hausser et al | microarrays | HEK293 | hsa-miR-124-3p | Overexpression | 15h |
| GSE14537 | ncbi.nlm.nih.gov/geo | Hausser et al | microarrays | HEK293 | hsa-miR-7-5p | Overexpression | 15h |
| GSE46039 | ncbi.nlm.nih.gov/geo | Helwak et al | microarrays | HEK293 | hsa-miR-92a-3p | Knock-down | 48h |
| GSE35621 | ncbi.nlm.nih.gov/geo | Hu et al | microarrays | HEK293 | hsa-miR-941 | Overexpression | 24h |
| GSE35621 | ncbi.nlm.nih.gov/geo | Hu et al | microarrays | HEK293T | hsa-miR-941 | Overexpression | 24h |
| GSE57158 | ncbi.nlm.nih.gov/geo | Greenberg et al | microarrays | PAG C81-61 | hsa-miR-20a-5p | Overexpression | 3d |
| GSE57158 | ncbi.nlm.nih.gov/geo | Greenberg et al | microarrays | PAG C81-61 | hsa-miR-17-5p | Overexpression | 3d |
| GSE33538 | ncbi.nlm.nih.gov/geo | Bossel Ben-Moshe et al | microarrays | MCF10A | hsa-miR-20a-5p | Silencing | 0h, 0.5h, 1h, 2h post EGF stimulation |
| GSE33538 | ncbi.nlm.nih.gov/geo | Bossel Ben-Moshe et al | microarrays | MCF10A | hsa-miR-671-5p | Silencing | 0h, 0.5h, 1h, 2h post EGF stimulation |
| GSE58142 | ncbi.nlm.nih.gov/geo | Frankel et al | microarrays | MCF7 | hsa-miR-95a-3p | Overexpression | 24h |
| GSE31397 | ncbi.nlm.nih.gov/geo | Frankel et al | microarrays | MCF7 | hsa-miR-101-3p | Overexpression | 24h |
| GSE19777 | ncbi.nlm.nih.gov/geo | Rao et al | microarrays | MCF7FR | hsa-miR-221-3p | Silencing | 72h |
| GSE19777 | ncbi.nlm.nih.gov/geo | Rao et al | microarrays | MCF7FR | hsa-miR-222-3p | Silencing | 72h |
| GSE40058 | ncbi.nlm.nih.gov/geo | Luo et al | microarrays | MDA-MB-231 | hsa-miR-200c-3p | Overexpression | NA |
| GSE40058 | ncbi.nlm.nih.gov/geo | Luo et al | microarrays | MDA-MB-231 | hsa-miR-205-5p | Overexpression | NA |
| GSE40058 | ncbi.nlm.nih.gov/geo | Luo et al | microarrays | MDA-MB-231 | hsa-mir-375 | Overexpression | NA |
| GSE50697 | ncbi.nlm.nih.gov/geo | Taube et al | microarrays | SUM159 | hsa-miR-203a-3p | Overexpression | NA |
| GSE51053 | ncbi.nlm.nih.gov/geo | Kristensen et al | microarrays | DU145 | hsa-miR-224-5p | Overexpression | 48h |
| GSE51053 | ncbi.nlm.nih.gov/geo | Kristensen et al | microarrays | DU145 | hsa-miR-452-5p | Overexpression | 48h |
| GSE34893 | ncbi.nlm.nih.gov/geo | Hudson et al | microarrays | LNCAP | hsa-miR-106b-5p | Overexpression | 24h |
| GSE17362 | ncbi.nlm.nih.gov/geo | Boll et al | microarrays | LNCAP | hsa-miR-130a-3p | Overexpression | 24h |
| GSE17362 | ncbi.nlm.nih.gov/geo | Boll et al | microarrays | LNCAP | hsa-miR-203a-3p | Overexpression | 24h |
| GSE17362 | ncbi.nlm.nih.gov/geo | Boll et al | microarrays | LNCAP | hsa-miR-205-5p | Overexpression | 24h |
| GSE31620 | ncbi.nlm.nih.gov/geo | Hudson et al | microarrays | LNCAP | hsa-miR-1-3p | Overexpression | 24h |
| GSE31620 | ncbi.nlm.nih.gov/geo | Hudson et al | microarrays | LNCAP | hsa-miR-206 | Overexpression | 24h |
| GSE31620 | ncbi.nlm.nih.gov/geo | Hudson et al | microarrays | LNCAP | hsa-miR-27b-3p | Overexpression | 24h |
| GSE51053 | ncbi.nlm.nih.gov/geo | Kristensen et al | microarrays | PC3 | hsa-miR-224-5p | Overexpression | 48h |
| GSE51053 | ncbi.nlm.nih.gov/geo | Kristensen et al | microarrays | PC3 | hsa-miR-452-5p | Overexpression | 48h |
| GSE12039 | ncbi.nlm.nih.gov/geo | Fish et al | microarrays | HUVEC | hsa-miR-126-3p | Anti-miR | 72h |
| GSE18438 | ncbi.nlm.nih.gov/geo | Coutler E et al | microarrays | JSC1 | hsa-miR-221-3p | Overexpression | NA |
| GSE25215 | ncbi.nlm.nih.gov/geo | Ikedo Y et al | microarrays | PaCa-2 | hsa-miR-193b-3p | Overexpression | 48h |

| | | | | | | | |
|-----------------|----------------------|-----------------------------|-------------|----------|-----------------|----------------|-----|
| GSE40189 | ncbi.nlm.nih.gov/geo | Ouyang H et al | microarrays | PANC-1 | hsa-miR-10b-5p | Knockdown | NA |
| GSE40189 | ncbi.nlm.nih.gov/geo | Ouyang H et al | microarrays | PANC-1 | hsa-miR-10b-5p | Overexpression | NA |
| GSE13460 | ncbi.nlm.nih.gov/geo | Tzur G et al | microarrays | hESC | hsa-miR-122-5p | Overexpression | NA |
| GSE86432 | ncbi.nlm.nih.gov/geo | Dzikiewicz-Krawczyk A et al | microarrays | DG75 | hsa-miR-150-5p | Overexpression | NA |
| GSE86432 | ncbi.nlm.nih.gov/geo | Dzikiewicz-Krawczyk A et al | microarrays | ST486 | hsa-miR-150-5p | Overexpression | NA |
| GSE19232 | ncbi.nlm.nih.gov/geo | Tome M et al | microarrays | hMSC | hsa-miR-335-5p | Overexpression | NA |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-7-5p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-9-5p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-122-5p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-128-3p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-132-3p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-133a-3p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-142-3p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-148b-3p | Overexpression | 24h |
| GSE8501 | ncbi.nlm.nih.gov/geo | Grimson A et al | microarrays | HELA | hsa-miR-181a-5p | Overexpression | 24h |
| GSE39359 | ncbi.nlm.nih.gov/geo | Cai J et al | microarrays | MCF7 | hsa-374a-5p | Overexpression | 36h |
| GSE40411 | ncbi.nlm.nih.gov/geo | Krishnan K et al | microarrays | MCF7 | hsa-miR-139-5p | Overexpression | NA |
| GSE32999 | ncbi.nlm.nih.gov/geo | Mazda M et al | microarrays | PC3 | hsa-miR-302a-3p | Overexpression | NA |
| GSE32999 | ncbi.nlm.nih.gov/geo | Mazda M et al | microarrays | PC3 | hsa-miR-372-3p | Overexpression | NA |
| GSE32999 | ncbi.nlm.nih.gov/geo | Mazda M et al | microarrays | PC3 | hsa-miR-373-3p | Overexpression | NA |
| GSE32999 | ncbi.nlm.nih.gov/geo | Mazda M et al | microarrays | PC3 | hsa-miR-520c-3p | Overexpression | NA |
| GSE32999 | ncbi.nlm.nih.gov/geo | Mazda M et al | microarrays | PC3 | hsa-miR-520f-3p | Overexpression | NA |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn et al | RNA-Seq | HEK293T | hsa-miR-1-3p | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam et al | RNA-Seq | HEK293 | hsa-miR-124-3p | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam et al | RNA-Seq | HEK293 | hsa-miR-155-5p | Overexpression | 24h |
| GSE68987 | ncbi.nlm.nih.gov/geo | Zhang et al | RNA-Seq | HELA | hsa-miR-603 | Overexpression | 24h |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn et al | RNA-Seq | HELA | hsa-miR-1-3p | Overexpression | 24h |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn et al | RNA-Seq | HELA | hsa-miR-155-5p | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam et al | RNA-Seq | HELA | hsa-miR-155-5p | Overexpression | 24h |
| GSE52531 | ncbi.nlm.nih.gov/geo | Nam et al | RNA-Seq | HELA | hsa-miR-124-3p | Overexpression | 24h |
| GSE68987 | ncbi.nlm.nih.gov/geo | Zhang et al | RNA-Seq | HELA | hsa-miR-603 | Overexpression | 24h |
| GSE37918 | ncbi.nlm.nih.gov/geo | Pellegrino et al | RNA-Seq | MCF7 | hsa-miR-23b-3p | Overexpression | NA |
| GSE37918 | ncbi.nlm.nih.gov/geo | Pellegrino et al | RNA-Seq | MDAMB231 | hsa-miR-23b-3p | Overexpression | NA |
| GSE64615 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RNA-Seq | HELA | has-miR-103a-3p | Overexpression | NA |
| GSE64615 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RNA-Seq | HELA | has-miR-494 | Overexpression | NA |
| GSE64615 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RNA-Seq | HELA | has-miR-503 | Overexpression | NA |
| GSE63555 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RNA-Seq | HELA | hsa-miR-191-5p | Overexpression | NA |

| | | | | | | | |
|-----------------|----------------------|-------------------|---------|---------|-----------------|----------------|-----|
| GSE63555 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RIP-Seq | HELA | hsa-miR-191-5p | Overexpression | NA |
| GSE64615 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RIP-Seq | HELA | has-miR-103a-3p | Overexpression | NA |
| GSE64615 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RIP-Seq | HELA | has-miR-494 | Overexpression | NA |
| GSE64615 | ncbi.nlm.nih.gov/geo | Polioudakis et al | RIP-Seq | HELA | has-miR-503 | Overexpression | NA |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn et al. | RPF-Seq | HEK293T | hsa-miR-1-3p | Overexpression | 24h |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn et al. | RPF-Seq | HELA | hsa-miR-155-5p | Overexpression | 32h |
| GSE60426 | ncbi.nlm.nih.gov/geo | Eichhorn et al. | RPF-Seq | HELA | hsa-miR-1-3p | Overexpression | 32h |

To retrieve a concise training set and reduce noise, datasets were combined under a tissue-specific scheme. Table 13 summarizes the associations regarded between cell types and tissues.

Table 13: Summary of the associations regarded between cell types and tissues for the extraction of miRNA-targeted regions incorporated in training/test sets.

| Tissue | Cell Type | CLIP-Seq libraries | miRNA perturbation experiments |
|-----------------------|---|--------------------|--------------------------------|
| Kidney | HEK293,HEK293T,PAG C81_61 | 43 | 11 |
| B lymphocyte | CCL86,CRL1432,CRL1596,H929,P3HR1,JSC1,D G75,ST486 | 18 | 9 |
| Bone Marrow | HMSC,K562 | 16 | 2 |
| Pancreas | PANC1,PACA2 | 1 | 4 |
| Brain | H4, SW1783,U87,HS683,U872M1,131_4_5B1 | 11 | 13 |
| Mammary Gland | MCF10A,MCF7,MCF7FR,MDAMB231,MDAM B468, SUM159 | 10 | 20 |
| Cervix | HELA | 6 | 34 |
| Embryo | HESC | 1 | 1 |
| Umbilical Vein | HUVEC | 2 | 1 |
| Prostate | LNCAP, PC3, DU145 | 5 | 16 |

Direct miRNA-target pairs derived from Reporter Gene Assay techniques and miRNA chimeric fragments were also incorporated in microT deployment (Methods 2.2.1). Published background PAR-CLIP libraries (Methods 2.2.1), stably expressing a commonly utilized non-RBP control (FLAG-GFP) were utilized to characterize negative miRNA-targeted pairs. The retrieved miRNA-binding events were annotated against a reference set of coding and 3' UTR exons. In cases of multiple transcript-gene associations, the transcript with the longest 3' UTR was selected. The adopted methodology is depicted in Figure 16 while Table 14 summarizes the miRNA-target positive/negative instances utilized in the training set of the microT model, as identified by different indirect/direct, low and high-throughput experiments. Table 15 outlines the independent test datasets included in the benchmarking evaluations.

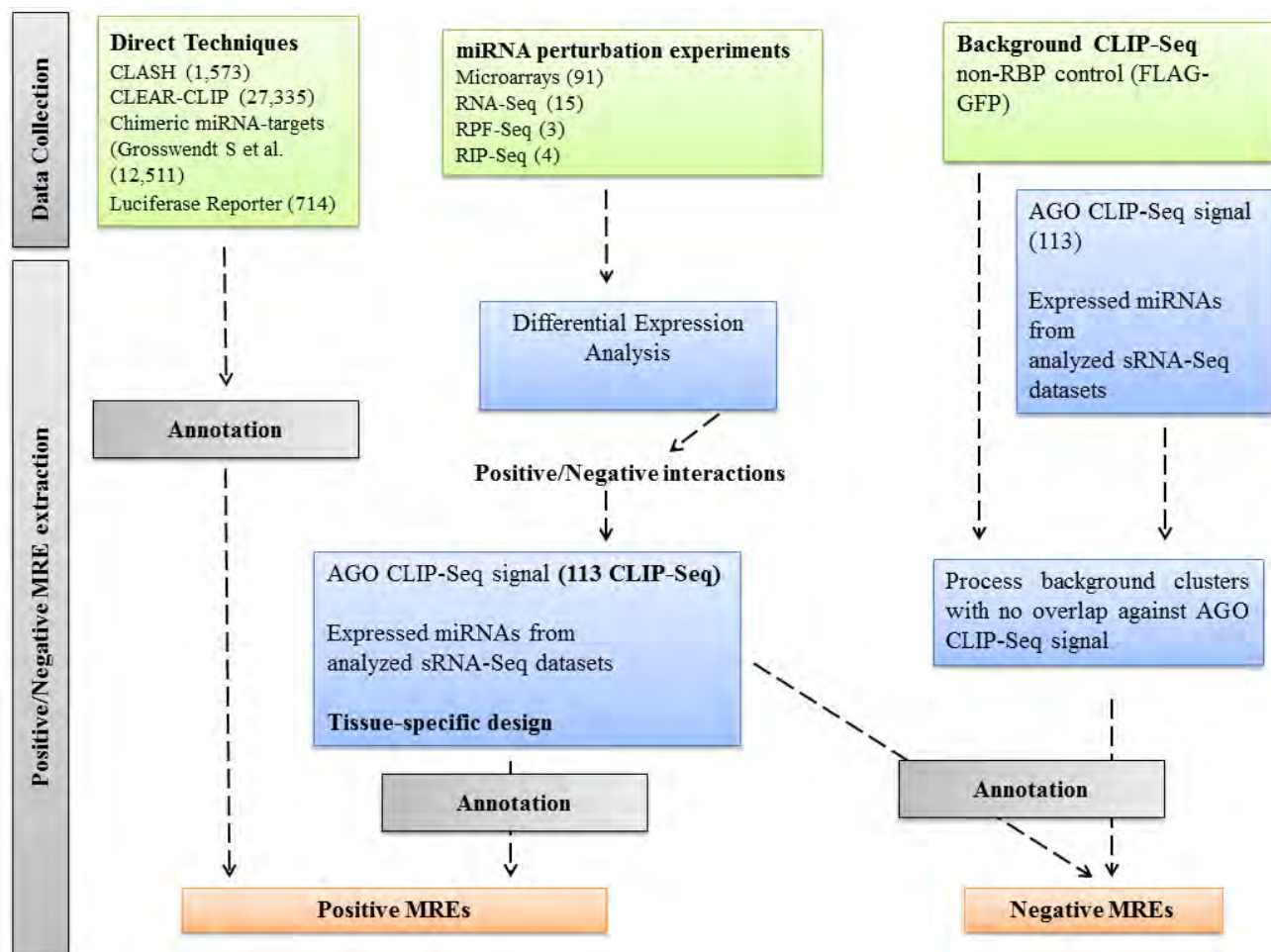


Figure 16: Overview of miRNA-target positive/negative instances as identified by different indirect/direct, low and high-throughput experiments. miRNA-targeted regions derived from miRNA perturbation datasets presented an overlap with AGO-bound enriched regions from at least one CLIP-Seq library. Datasets have been combined under a tissue-specific scheme. No overlap was allowed between positive and negative miRNA-gene interactions and their related MRE-instances.

Table 14: Overview of miRNA-target positive/negative instances utilized in training set as identified by different indirect/direct, low and high-throughput experiments.

| Positive Instances | miRNAs in interactions | Genes in interactions | miRNA-target instances |
|--------------------|------------------------|-----------------------|------------------------|
| Chimeric | 313 | 3,987 | 9,119 |
| RNA-Seq | 9 | 1,942 | 4,244 |
| Microarrays | 55 | 2,414 | 5,194 |
| RPF | 2 | 1,383 | 2,872 |
| RIP | 2 | 322 | 667 |
| Negative Instances | miRNAs in interactions | Genes in interactions | miRNA-target instances |
| RNA-Seq | 8 | 948 | 3,005 |
| Microarrays | 28 | 189 | 835 |
| RPF | 2 | 804 | 2,651 |
| FLAG_GFP_bg_20kD | 393 | 715 | 12,371 |
| FLAG_GFP_bg_35kD | 393 | 2,026 | 24,832 |
| FLAG_GFP_bg_45kD | 393 | 2,219 | 24868 |

Table 15: Summary of microarray experiments in human species upon specific miRNA deregulation, utilized in benchmarking evaluations of microT Super Learning model.

| Accession | Repository | Authors | Cell Type | miRNA | miRNA treatment | Post-Transfection Cell Harvest Time/Experimental Condition |
|-----------|----------------------|-----------------|-----------|-----------------|-----------------|--|
| GSE16962 | ncbi.nlm.nih.gov/geo | Fasanaro et al | HUVEC | hsa-miR-210-3p | Overexpression | 24h |
| GSE21901 | ncbi.nlm.nih.gov/geo | Hollander et al | HEK293 | hsa-miR-212-3p | Overexpression | - |
| GSE22790 | ncbi.nlm.nih.gov/geo | Elyakim et al | HEPG2 | hsa-miR-191-5p | Anti-miR | - |
| GSE21132 | ncbi.nlm.nih.gov/geo | Li et al | Jurkat | hsa-miR-146a-5p | Overexpression | 48h |
| GSE42749 | ncbi.nlm.nih.gov/geo | Salim et al | U1810 | hsa-miR-214-3p | Antagomir | 24h |

2.3.2 microT *in silico* framework

Feature set description. A set of 117 descriptors with non-zero variance was included in microT. The extracted features were retrieved from positive/negative miRNA interactions, identified on AGO-bound locations in different CLIP-Seq datasets and on chimeric fragments. microCLIP descriptors have been re-evaluated in the new enhanced training set by implementing feature selection methods. Statistical tests and metrics estimating the predictive accuracy of descriptors, e.g. AUC plots, were used to evaluate the behavior of features in one-dimension. 30 characteristics representing per nucleotide base pairing composition of miRNA binding region with weak performance, as well as PAR-CLIP associated descriptors were totally discarded. The model integrates 16 new features that outline dinucleotide content and sequence accessibility of miRNA binding region and of respective upstream/downstream

regions. The location of the MRE within the 3' UTR/CDS, the length of the respective exon and the distance of the adjacent MREs were also estimated and incorporated. Accessibility scores were computed using RNAplfold[112] with the following parameters: $w = 150$, $L = 100$ and $u = 31$.

Description of the algorithm. microT operates on the whole transcript. It initially identifies putative (non-)canonical MREs located within the 3' UTR and CDS regions by adopting a sensitive pipeline and subsequently scores them following the microCLIP classification scheme. Base node components have been re-arranged. Characteristics have been removed or inserted based on their conceptual framework. Table 16 describes microT features separated into the base nodes. The new entries are denoted with a bold font.

Table 16: Description of features incorporated in microT.

| feature# | base classifier node | feature description |
|----------|---|---|
| 1 | Region Features (53 Features) | MRE region GC-skew |
| 2 | | MRE region Purine-skew |
| 3 | | MRE region Ks-skew |
| 4 | | Upflank MRE region Purine-skew |
| 5 | | Upflank MRE region Ks-skew |
| 6 | | MRE DUST score |
| 7 | | MRE region AT-skew |
| 8 | | MRE dS |
| 9 | | MRE Tm |
| 10 | | Codon Adaptation Index per codon usage bias |
| 11 | | Dinucleotide AA MRE content |
| 12 | | Dinucleotide AC MRE content |
| 13 | | Dinucleotide AG MRE content |
| 14 | | Dinucleotide AT MRE content |
| 15 | | Dinucleotide CA MRE content |
| 16 | | Dinucleotide CC MRE content |
| 17 | | Dinucleotide CT MRE content |
| 18 | | Dinucleotide GA MRE content |
| 19 | | Dinucleotide GC MRE content |
| 20 | | Dinucleotide CG MRE content |
| 21 | | Dinucleotide GG MRE content |
| 22 | | Dinucleotide TT MRE content |
| 23 | | Dinucleotide TA MRE content |
| 24 | | Dinucleotide AC content upflank of MRE |
| 25 | | Dinucleotide AT content upflank of MRE |
| 26 | | Dinucleotide GC content upflank of MRE |
| 27 | | Dinucleotide GT content upflank of MRE |
| 28 | | Dinucleotide TA content upflank of MRE |
| 29 | | Dinucleotide AA content upflank of MRE |
| 30 | | Dinucleotide TG content upflank of MRE |
| 31 | | A content upflank of MRE |

| | | |
|----|---|---|
| 32 | | C content upflank of MRE |
| 33 | | G content upflank of MRE |
| 34 | | T content upflank of MRE |
| 35 | | A or T content upflank of MRE |
| 36 | | A content in MRE |
| 37 | | A or G content in MRE |
| 38 | | A or T content in MRE |
| 39 | | C content in MRE |
| 40 | | G content in MRE |
| 41 | | G or T content in MRE |
| 42 | | T content in MRE |
| 43 | | Average conservation in MRE seed region |
| 44 | | Average conservation downstream of MRE region |
| 45 | | Average conservation upstream of MRE region |
| 46 | | Minimum conservation in MRE seed region |
| 47 | | Minimum conservation downstream of MRE region |
| 48 | | Minimum conservation upstream of MRE region |
| 49 | | Number of MREs per 3' UTR/CDS length |
| 50 | | 3' UTR/CDS length |
| 51 | | Accessibility of the 30nt region upstream of MRE |
| 52 | | Accessibility of the 30nt region downstream of MRE |
| 53 | | Accessibility of the 20nt MRE region |
| 54 | | Binding event in MRE position 2 |
| 55 | | Binding event in MRE position 3 |
| 56 | | Binding event in MRE position 4 |
| 57 | | Binding event in MRE position 5 |
| 58 | | Binding event in MRE position 6 |
| 59 | | Binding event in MRE position 7 |
| 60 | | Binding event in MRE position 8 |
| 61 | | Binding event in MRE position 10 |
| 62 | | Binding event in MRE position 11 |
| 63 | | Binding event in MRE position 12 |
| 64 | | Binding event in MRE position 17 |
| 65 | Binding Vectors (23 Features) | Binding event in MRE position 18 |
| 66 | | Base at MRE position 25 |
| 67 | | Base at MRE position 26 |
| 68 | | Base at MRE position 27 |
| 69 | | Base at MRE position 28 |
| 70 | | Base at MRE position 29 |
| 71 | | Base at miRNA position 13 |
| 72 | | Base at miRNA position 15 |
| 73 | | Base at miRNA position 19 |
| 74 | | Base at miRNA position 4 |
| 75 | | Base at miRNA position 5 |
| 76 | | Base at miRNA position 6 |
| 77 | Matches per miRNA/MRE | Total mismatches |
| 78 | domain | Max consecutive matches |
| 79 | (13 Features) | Match in position 10 of miRNA |

| | | |
|-----|---|--|
| 80 | | Match in position 2 of miRNA |
| 81 | | Consecutive unpaired bases of non-seed region |
| 83 | | Consecutive matches per total matches |
| 83 | | Matches in seed region per total matches |
| 84 | | Consecutive matches in seed region |
| 85 | | Consecutive matches in non-seed region per total matches |
| 86 | | Matches in seed region |
| 87 | | Matches in central MRE region |
| 88 | | Matches in 3' MRE |
| 89 | | Consecutive matches in seed region per max consecutive matches |
| 90 | | AU frequency in MRE region |
| 91 | | GC frequency in MRE region |
| 92 | | MRE binding start |
| 93 | | Internal loop max length in MRE region |
| 94 | Duplex Features (10 Features) | AU frequency in seed region |
| 95 | | AU frequency in 3' MRE region |
| 96 | | GC frequency in tail MRE region |
| 97 | | GC frequency in MRE region excluding seed |
| 98 | | Length of bulges in MRE region |
| 99 | | Bulge positions in MRE region (%) |
| 100 | | Matches in MRE Ks-skew |
| 101 | | miRNA unpaired A |
| 102 | | miRNA unpaired C |
| 103 | | miRNA unpaired G |
| 104 | Base Pairing (9 Features) | miRNA unpaired T |
| 105 | | miRNA matches A |
| 106 | | miRNA matches T |
| 107 | | miRNA matches C |
| 108 | | miRNA matches G |
| 109 | | MRE distance from 3' UTR/CDS end |
| 110 | | Distance of adjacent MREs |
| 111 | | Matches in MRE Purine-skew |
| 112 | MRE General (9 Features) | Duplex structure energy |
| 113 | | Length of miRNA binding region |
| 114 | | Length of MRE binding region per cluster length |
| 115 | | Average conservation of whole MRE |
| 116 | | Average conservation of 5' MRE |
| 117 | | Binding Type |
| 1 | Feature Combination Set 1 (26 Features) | Binding event in MRE position 4 |
| 2 | | Binding event in MRE position 11 |
| 3 | | Binding event in MRE position 18 |
| 4 | | Base at MRE position 28 |
| 5 | | Base at MRE position 29 |
| 6 | | Base at miRNA position 13 |
| 7 | | Base at miRNA position 6 |

| | | |
|----|---|--|
| 8 | | Max consecutive matches |
| 9 | | Binding Type |
| 10 | | Consecutive matches per total matches |
| 11 | | Consecutive matches in seed region per max consecutive matches |
| 12 | | Matches in 3' MRE |
| 13 | | AU frequency in seed region |
| 14 | | MRE binding start |
| 15 | | AU frequency in 3' MRE region |
| 16 | | Length of bulges in MRE region |
| 17 | | miRNA unpaired C |
| 18 | | miRNA matches A |
| 19 | | miRNA matches C |
| 20 | | miRNA matches G |
| 21 | | Duplex structure energy |
| 22 | | Length of MRE binding region per cluster length |
| 23 | | Average conservation of whole MRE |
| 24 | | Base at miRNA position 15 |
| 25 | | Binding event in MRE position 5 |
| 26 | | Binding event in MRE position 7 |
| 1 | Feature Combination Set 2 (14 Features) | MRE dS |
| 2 | | G content upflank of MRE |
| 3 | | A or T content upflank of MRE |
| 4 | | miRNA unpaired C |
| 5 | | Average conservation upstream of MRE region |
| 6 | | Accessibility of the 30nt region upstream the MRE |
| 7 | | Accessibility of the 30nt region downstream the MRE |
| 8 | | Accessibility of the 20nt MRE region |
| 9 | | Binding event in MRE position 3 |
| 10 | | AU frequency in MRE region |
| 11 | | GC frequency in MRE region |
| 12 | | Matches in seed region per total matches |
| 13 | | Consecutive matches in seed region |
| 14 | | Distance of adjacent MREs |
| 1 | Feature Combination Set 3 (14 Features) | Binding event in MRE position 6 |
| 2 | | Match in position 10 of miRNA |
| 3 | | Match in position 2 of miRNA |
| 4 | | Matches in seed region |
| 5 | | Matches in central MRE region |
| 6 | | GC frequency in tail MRE region |
| 7 | | GC frequency in MRE region excluding seed |
| 8 | | Matches in MRE Ks-skew |
| 9 | | miRNA unpaired A |
| 10 | | miRNA unpaired T |
| 11 | | Matches in MRE Purine-skew |
| 12 | | Length of miRNA binding region |
| 13 | | MRE distance from 3' UTR/CDS end |
| 14 | | Consecutive unpaired bases of non-seed region |

microT was trained against a collection of 22,096 positive and 68,562 negative miRNA interactions (Table 17, Table 18). The first layer comprised 16,062 positive and 52,592 negative instances, while the second one was trained on an independent set of 6,034 positive and 15,970 negative miRNA-target pairs. Ten-fold cross-validation was performed on the training data to estimate each model's accuracy. An independent test set of 6,192 positive/negative instances was utilized in the benchmarking evaluations.

Table 17: Summary of training set utilized for microT deployment.

| | miRNAs in interactions | | Genes in interactions | | miRNA-target instances | |
|---------------------------------------|------------------------|------|-----------------------|-------|------------------------|-------|
| | Training | Test | Training | Test | Training | Test |
| Positive Instances | | | | | | |
| <i>Direct Techniques</i> | 313 | 292 | 3,987 | 2,010 | 9,119 | 3,092 |
| <i>miRNA perturbation experiments</i> | 60 | - | 4,700 | - | 12,977 | - |
| Negative Instances | | | | | | |
| <i>Background CLIP-Seq</i> | 393 | 391 | 3,883 | 1,758 | 62,071 | 2,801 |
| <i>miRNA perturbation experiments</i> | 34 | 15 | 1,698 | 264 | 6,491 | 299 |

Table 18: Summary of miRNA-target instances, located on 3' UTR and CDS regions, utilized in the training/test of microT model.

| Biotype | Positive set | | Negative set | |
|---------|--------------|-------|--------------|-------|
| | Training | Test | Training | Test |
| UTR3 | 13,441 | 2,215 | 41,276 | 1,820 |
| CDS | 8,655 | 877 | 27,286 | 1,280 |

CHAPTER 3

Results

3.1 DIANA-TarBase repository

DIANA-TarBase v8.0 is a reference database devoted to the indexing of experimentally-supported microRNA (miRNA) targets. Its 8th version is the first database indexing more than 1 million entries, corresponding to ~670,000 unique miRNA target pairs. The interactions are supported by more than 33 experimental methodologies, applied to ~600 cell types/tissues under ~451 experimental conditions. It integrates information on cell-type specific miRNA-gene regulation, while hundreds of thousands of miRNA binding locations are reported. TarBase is coming of age, with more than a decade of continuous support in the non-coding RNA field. A new module has been implemented that enables the browsing of interactions through different filtering combinations. It permits easy retrieval of positive and negative miRNA targets per species, methodology, cell type and tissue. An incorporated ranking system is utilized for the display of interactions based on the robustness of their supporting methodologies. Statistics, pie-charts and interactive bar-plots depicting the database content are available through a dedicated result page. An intuitive interface is introduced, providing a user-friendly application with flexible options to different queries.

3.1.1 DIANA-TarBase update: Database statistics

The current version has been enhanced with a large compilation of high quality miRNA-binding events derived from chimeric fragments, reporter gene assay and CLIP-Seq experiments. More than 200 high-throughput experiments followed by perturbation of a specific miRNA have been analyzed and integrated in the database. This extension provides an increase of approximately 200,000 interactions and ~300,000 entries since the previous version[50]. A concise description of TarBase v8.0 is presented in Table 19.

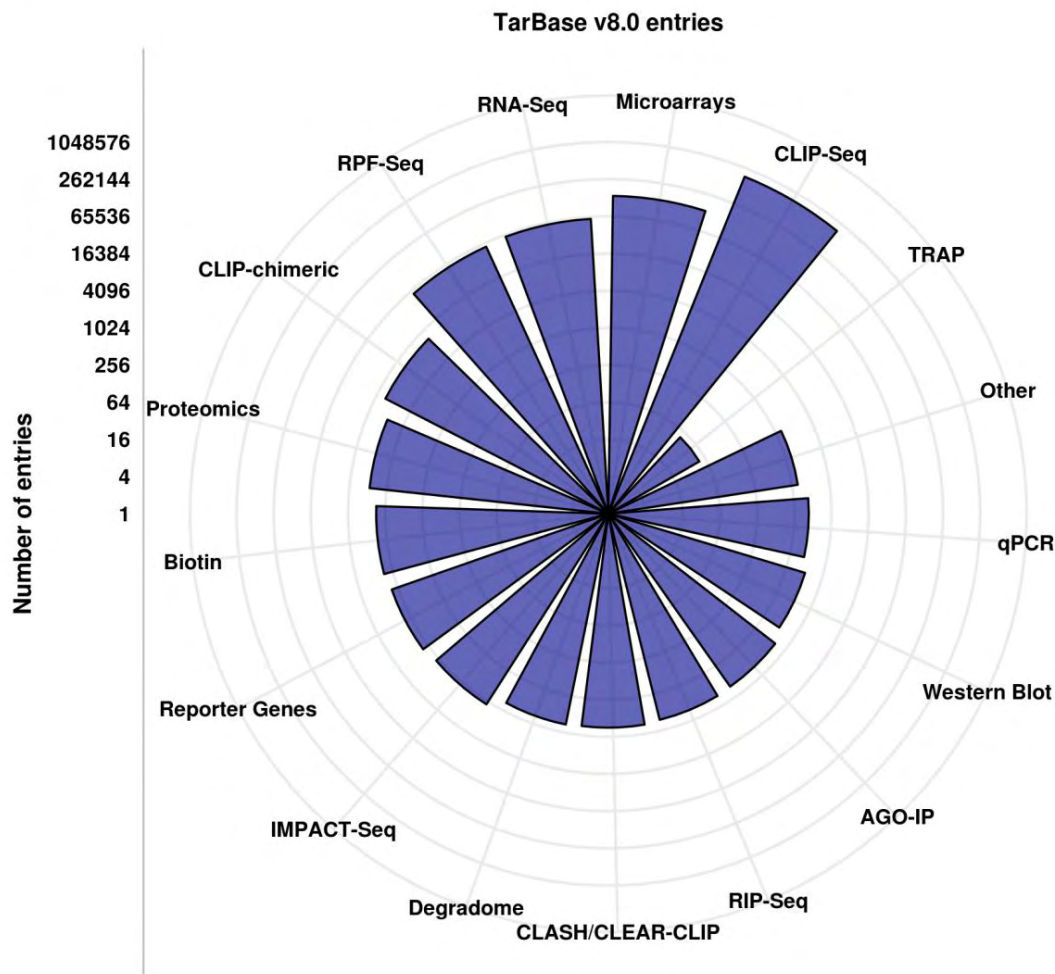
Table 19: TarBase v8.0 Entries. Statistics regarding the total entries, miRNA-gene interacting pairs derived from low-/high-throughput methodologies, distinct cell types/tissues and curated publications are provided. The number of analyzed datasets and unique studied conditions are presented for high-throughput experiments. The incorporated low-/high-throughput experimental techniques, as well as interface improvements are reported. Newly incorporated experimental methods and interface advancements are marked as bold.

| TarBase v8.0 | | |
|-----------------------------------|--------------------------------------|---|
| Database | Total entries | >1,080,000 |
| | Entries from low-yield methods | 10,339 |
| | Entries from high-throughput methods | ~1,069,000 |
| | Cell types | 516 |
| | Tissues | 85 |
| | Publications | 1,208 |
| Support from direct experiments | miRNA-gene entries | ~790,300 |
| | miRNAs | 1,761 |
| | Targeted genes | 27,613 |
| | Publications | 968 |
| Analyzed high-throughput datasets | Datasets | 353 |
| | Conditions | ~230 |
| | Publications | 102 |
| Experimental Methods | Description of major classes | Reporter Genes, Western Blot, qPCR, Proteomics, Biotin miRNA tagging , CLIP-Seq, CLEAR-CLIP , CLASH, CLIP-chimeric , IMPACT-Seq, AGO-IP, RPF-Seq , RIP-Seq , Degradome, RNA-Seq, TRAP, Microarrays, Other |
| Interface | Data visualization | Re-designed interface , support of specific queries, Browsing Mode , Ranking System , customizable sorting of results, advanced interactive statistics, advanced filtering options, cell type/tissue combinations , detailed meta-data, interconnection with DIANA-Tools, ENSEMBL integration |

DIANA-TarBase v8.0 caters more than one million entries, corresponding to the largest compilation of experimentally supported miRNA targets. This collection of miRNA-gene

interactions has been derived from experiments employing more than 33 distinct low-yield and high-throughput techniques, spanning 85 tissues, 516 cell types and ~451 experimental conditions from 18 species (Figure 17a). Approximately 1,200 publications were manually curated and more than 350 high-throughput datasets have been analyzed. The new database version incorporates an assortment of positive and negative direct miRNA interactions. It comprises more than 10,000 interactions derived from specific techniques. Approximately 5,100 of these miRNA targets are verified by reporter gene assays, extracted from ~950 publications, providing a 1.6-fold increase compared to relevant entries in TarBase v7.0. More than 14,000 direct miRNA-mRNA chimeric fragments defined from CLASH and CLEAR-CLIP experiments, as well as from a previous meta-analysis of published AGO-CLIP datasets[113], have been integrated to the repository. Approximately 90,000 new entries were generated from the analysis of additional AGO CLIP-Seq libraries from 3 studies. More than 233,000 interactions have been extracted from miRNA-specific transfection/knockdown microarray, RPF-Seq, RIP-Seq and RNA-Seq experiments which were performed in 28 tissues and 82 cell types under 206 experimental conditions. Updated entries derived from the aforementioned methodologies are summarized in (Figure 17b).

A



B

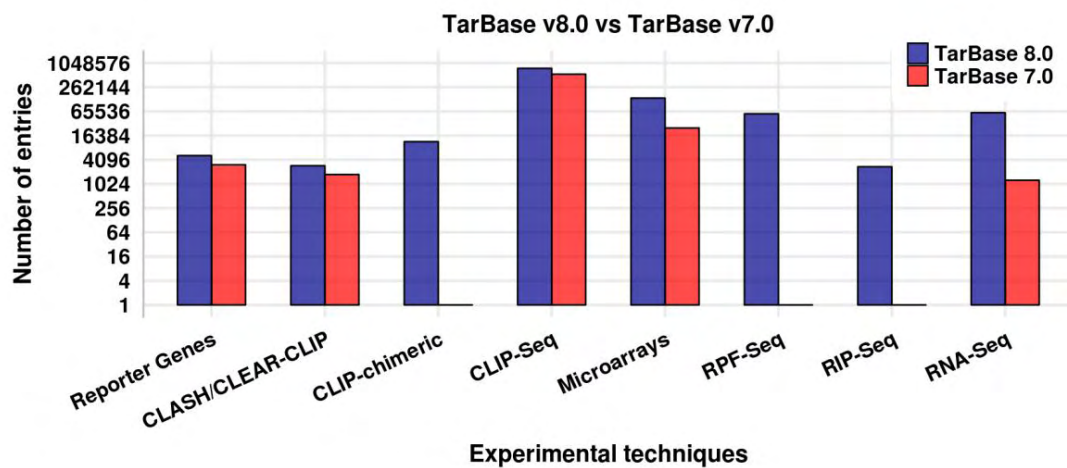


Figure 17: TarBase entries divided per methodology. Values are plotted in log2 scale. Each grid line corresponds to quadrupling of indexed miRNA interactions. a) Total miRNA-gene entries incorporated in TarBase v8.0. b) Comparison of TarBase v8.0 and TarBase v7.0 entries (Karagkouni D and Paraksevpoulou MD *et al*, 2017)[64].

3.1.2 Interface

3.1.2.1 Querying the database

A new relational schema, developed in PostgreSQL, is introduced to host TarBase v8.0 data. The database interface has also been redesigned using the Yii 2.0 PHP framework and enhanced to provide an intuitive user-friendly application as well as flexible options to different queries (Figure 18). Users can retrieve interactions by performing a query with miRNA and/or gene names. Identifiers from ENSEMBL[83] and miRBase[114] are supported. Positive and/or negative miRNA targets can be retrieved through the combination of distinct filters such as experimental methodology, cell type and tissue according to the user's needs. Results can be sorted in ascending or descending order based on gene and/or miRNA names as well as on the number of experiments, publications and cell types/tissues supporting these interactions. Detailed meta-data including the binding location and experimental conditions are displayed in the relevant result sections.

The screenshot displays the DIANA-TarBase v8.0 interface, which is divided into several functional areas:

- (1) Search Fields - Query mode:** Includes input fields for miRNAs (hsa-miR-1-3p, hsa-miR-221-3p) and Genes (ZEB2, SELE, TKT).
- (2) Filters - Browsing mode:** A sidebar with various filters including Species (Homo Sapiens), Method Type, Method (Chimeric fragments, Luciferase Reporter Assay, RPF-Seq), Regulation type, Validation Type, Validated as, Cell Type, Tissue, Source (TarBase 8.0), Publication Year, and Prediction score.
- (3) Result statistics:** Displays summary statistics for the current query: Interactions: 3, Experiments: 7 (low: 2, high: 5), Cell lines: 5, Tissues: 5, Publications: 4.
- (4) Interactive result sorting:** Allows sorting by Experiments throughput, Publications, Cell lines, Tissues, and Pred. Score.
- (5) Gene/miRNA details:** Shows details for TKT and hsa-miR-221-3p, including low-throughput experiments (1 positive, 0 negative) and high-throughput experiments (4 positive, 0 negative).
- (6) Experiment details:** Provides a detailed view of experiments, including publication, methods, tissue, cell line, tested cell line, and experimental conditions.
- (7) Binding site details:** Shows details for ZEB2 and hsa-miR-221-3p, including low-throughput experiments (0 positive, 1 negative) and high-throughput experiments (1 positive, 0 negative).
- (8) miRPath interconnection:** A link to related pathways.
- (9) Help:** A link to the help page.
- (10) Database statistics:** A link to database statistics.

The interface is powered by the Yii Framework and is copyrighted by 2017 Univ. of Thessaly, Pasteur Institute & IMIS - "Athena" RC.

Figure 18: Snapshot depicting the DIANA-TarBase v8.0 interface. Users can apply a query with miRNA and/or gene names [1] or navigate in the database content through combinations of the filtering criteria [2]. Positive/negative interactions can be refined with a series of filtering options including species, tissues/cell

types, methodologies, type of validation (direct/indirect), database source, publication year as well as in silico predicted score [2]. Brief result statistics are promptly calculated [3]. Interactions can be sorted in ascending or descending order based on gene and/or miRNA names, on the number of experiments, publications and cell types/tissues supporting them [4]. Gene and miRNA details, complemented with active links to Ensembl, miRBase and the DIANA disease tag cloud, are provided [5]. Details regarding the experimental procedures such as the methodology, cell type/tissue, experimental conditions and link to the actual publication are presented [6]. Methods are color-coded, with green and red portraying validation for positive and negative regulation, respectively. Interactions are also accompanied by miRNA-binding site details [7]. Links to DIANA-miRPath functional analysis resource [8] and to an informative Help section [9] are also available. Users can navigate to the separate database statistics page [10] (Karagkouni D and Paraksevopoulou MD *et al*, 2017)[64].

Ranking system: A novel ranking system has been incorporated in the interface. miRNA targets are by default sorted according to the robustness of the respective experimental techniques. In brief, miRNA-gene interactions determined from low-throughput experiments are reported first, followed by those derived from high-throughput techniques. More precisely, miRNA-binding events retrieved from reporter gene assays, the gold standard of methodologies in miRNA target recognition, are prioritized, followed by those defined from any other low-yield technique. Direct interactions inferred from chimeric fragments are subsequently presented, followed by those determined from CLIP-Seq methods. miRNA targets supported from any other indirect miRNA-specific transfection/knockdown high-throughput technique are finally displayed. In cases of miRNA-target pairs derived from the same category of methods, ranking is performed based on the number of distinct experiments they have been validated with.

Browsing mode: A novel aspect in the new interface is the browsing mode (Figure 18). Users can easily retrieve the top targets (up to a maximum of 3,000) without applying any specific query. Positive or negative interactions can be obtained based on different combinations of the filtering criteria including species, tissues/cell types and methodologies.

Advanced statistics: DIANA-TarBase v8.0 also provides statistics, advanced interactive pie-charts and bar plots, implemented using the D3.js JavaScript library, to portray the database content and extent for the different species (Figure 19).

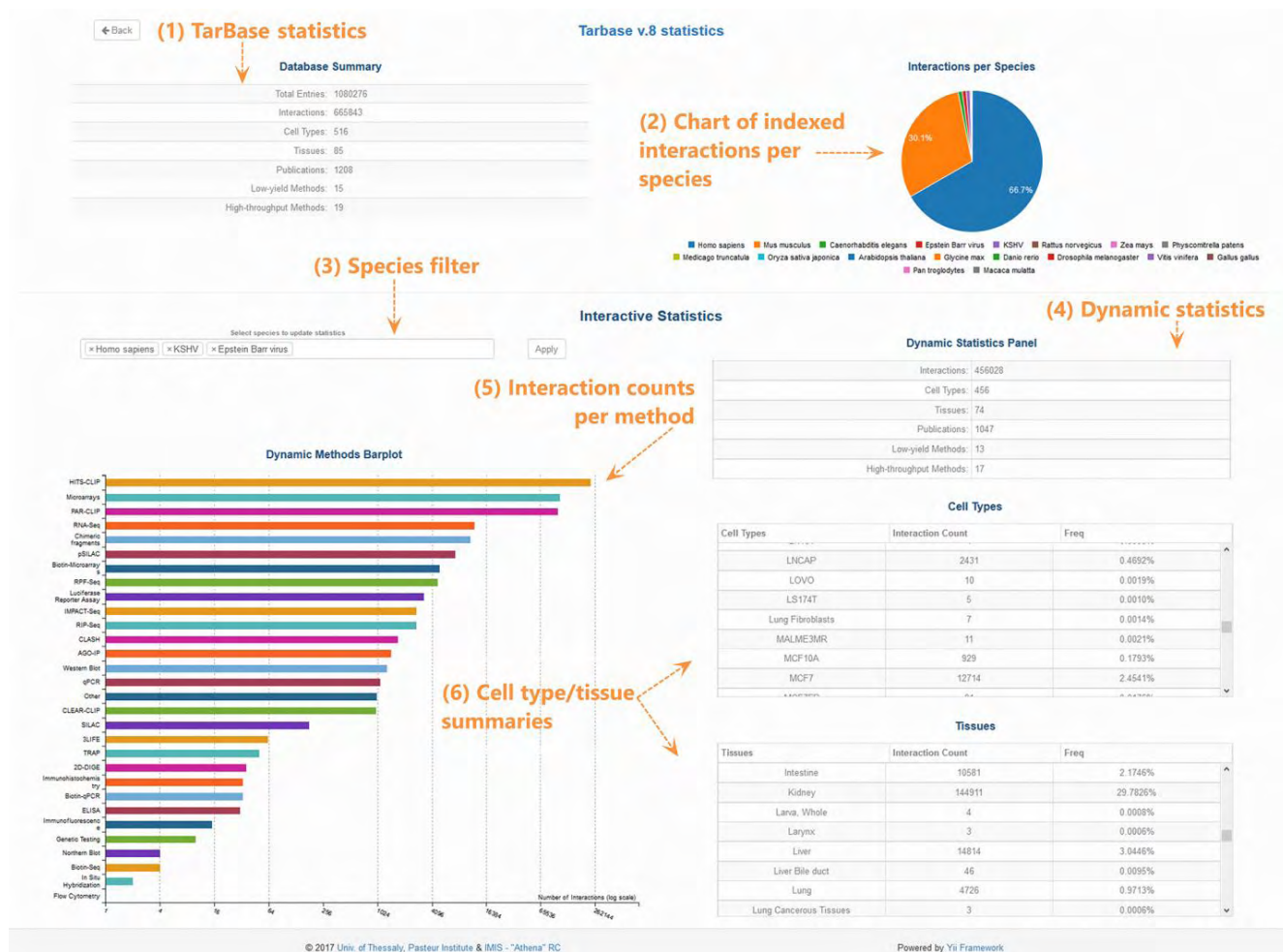


Figure 19: Screen-shot depicting DIANA-TarBase statistics page. The number of interactions, cell types/tissues, publications and low-/high-throughput methodologies are summarized at the top of the page [1]. A pie-chart portraying the database content per species is provided [2]. The user can select any species combination [3] to obtain relevant statistics [4]. The bar-plot [5] and tables [6] at the end of the page show the number of interactions (log2-scaled) per methodology and the cell-type/tissue frequencies respectively. They are also dynamically populated depending on the user's choice of species (Karagkouni D and Paraksevopoulou MD *et al*, 2017)[64].

3.1.2.2 Database interconnections

Since the sixth version, DIANA-TarBase has been integrated in ENSEMBL[83] and RNACentral[115] (Figure 20, Figure 21). Interactions accompanied with the exact binding location can be viewed in the ENSEMBL Genome Browser via the dedicated "TarBase" track. The database is also seamlessly interconnected with other available DIANA-tools, including microT-CDS[85] for *in silico* identification of miRNA targets, LncBase v2.0[116] for the display of miRNA-lncRNA interactions and DIANA-miRPath v3.0[117] for functional characterization of miRNAs.

Additionally to the ~1 million entries indexed in TarBase, miRNA targets retrieved from other relevant databases, including miRTarBase[118] and miRecords[32], are also provided to users. These entries are disregarded from database statistics.

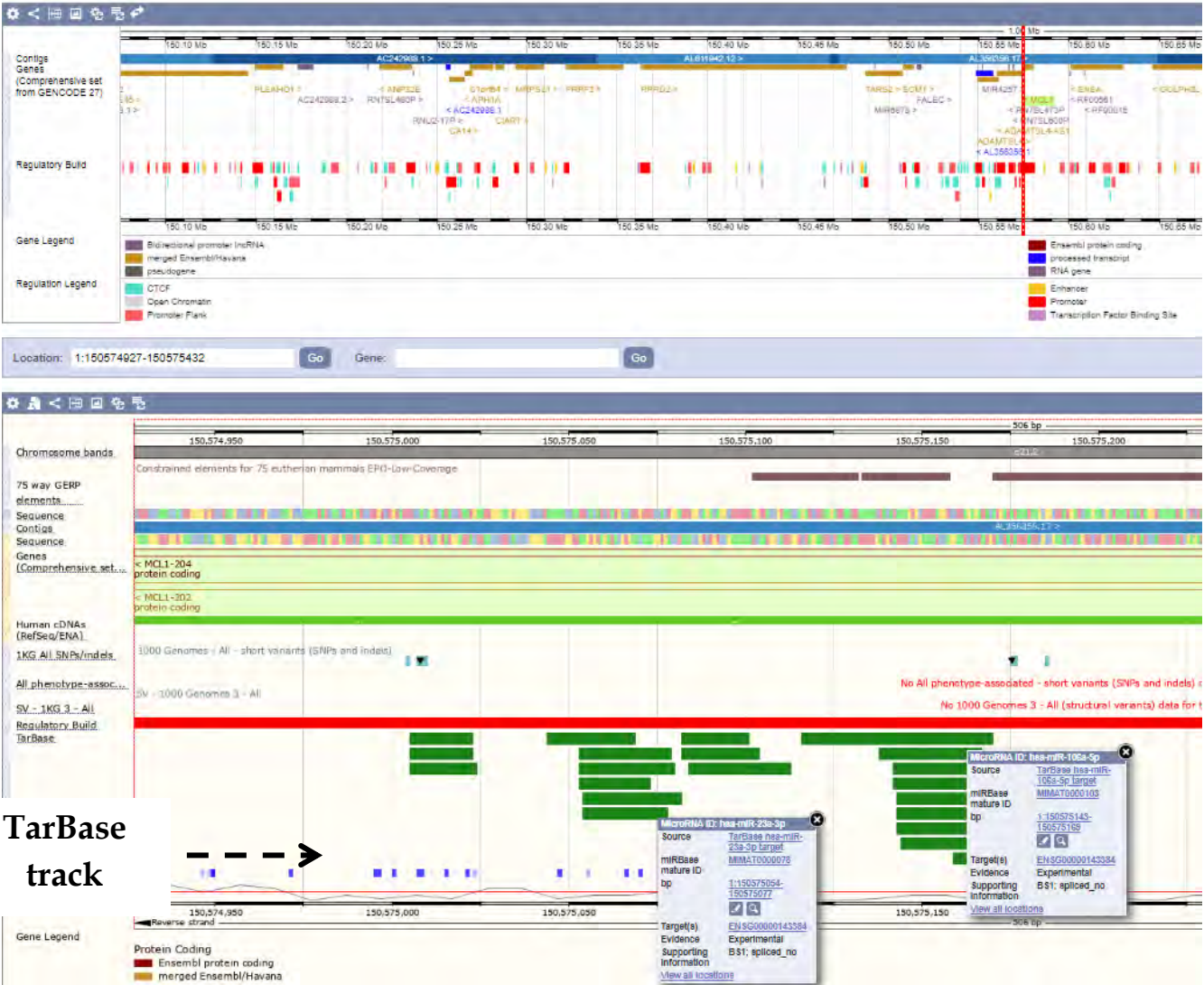


Figure 20: TarBase integration in ENSEMBL.

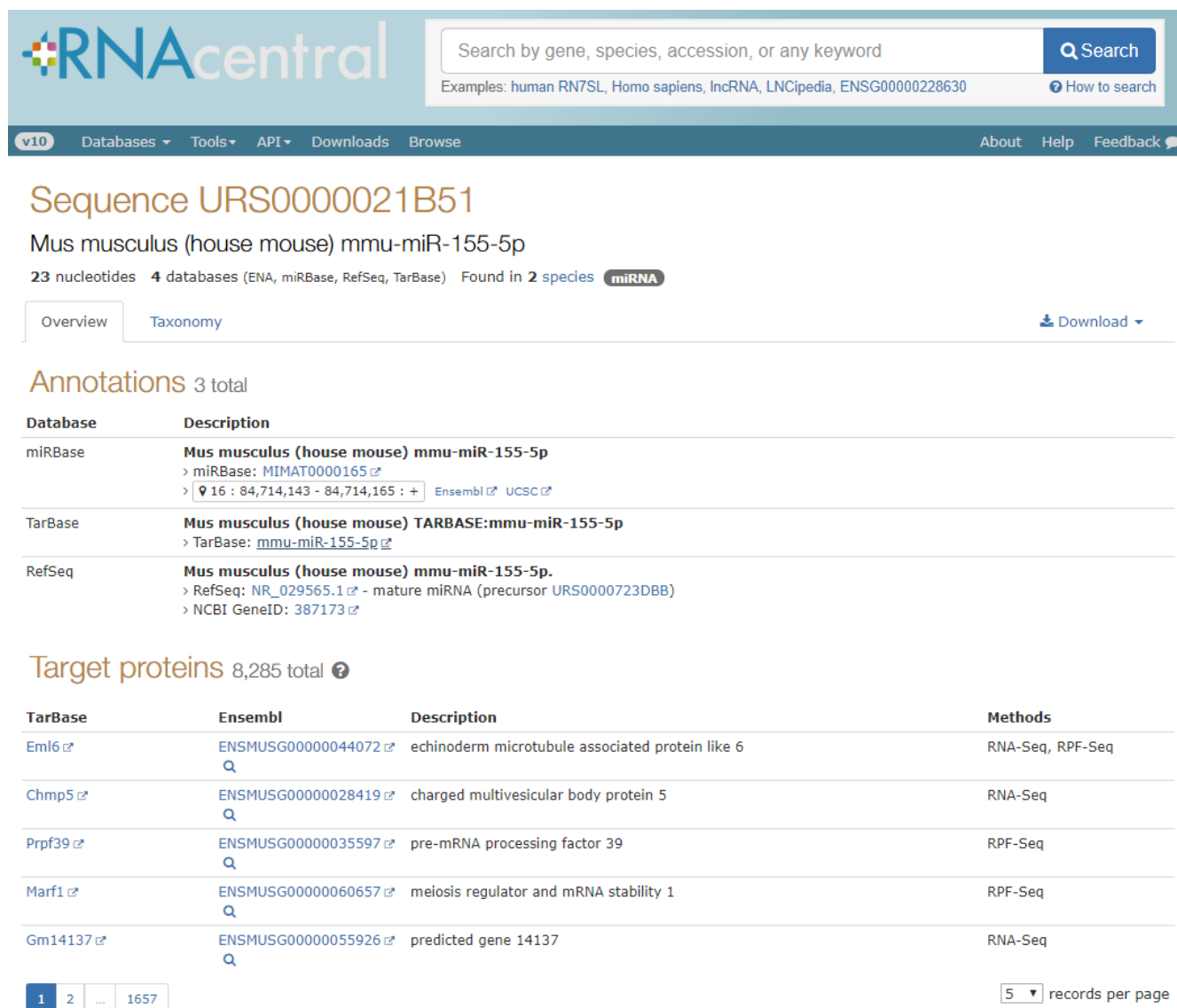


Figure 21: TarBase integration in RNAcentral.

3.2 microCLIP Super Learning framework uncovers functional transcriptome-wide miRNA interactions

microCLIP is a cutting-edge framework, dedicated to the analysis of AGO-CLIP-Seq experiments, that combines deep learning classifiers under a super learning scheme. The analysis of PAR-CLIP methodology focuses on sequence clusters containing T-to-C conversions. In this thesis, it is demonstrated that the non-T-to-C clusters, frequently observed in PAR-CLIP experiments, exhibit functional miRNA binding events and strong RNA accessibility. This discovery is based on the analysis of an extensive compendium of *bona fide* miRNA-binding events, and is further supported by numerous miRNA perturbation experiments and structural sequencing data. The incorporation of these previously neglected clusters yields an average of 14% increase in miRNA-target interactions per PAR-CLIP

library. The increased performance of microCLIP in CLIP-Seq-guided detection of miRNA interactions, uncovers previously elusive regulatory events and miRNA-controlled pathways.

3.2.1 T-to-C and non-T-to-C PAR-CLIP clusters share common traits

Clusters depleted on T-to-C conversions, which are always filtered out in PAR-CLIP analysis, seem to aid in the identification of functional miRNA binding events (Figure 11).

One of the most important steps in PAR-CLIP analysis is the identification of AGO-bound regions for further investigation. This process is mainly based on the presence and percentage of reads harboring T-to-C mutations within a cluster, while all other peaks are omitted from the analysis. Importantly, including only T-to-C enhanced cross-linked regions led to a significant loss (60-80%) of the AGO-PAR-CLIP reads across 24 libraries. non-T-to-C containing regions are examined for the possibility to pinpoint functional miRNA binding events. The applied approach assessed a random set of 4,310 and 1,700 miRNA binding sites, supported by T-to-C and non-T-to-C clusters respectively, located in 3'UTR and CDS regions. More than 65% of miRNA recognition elements (MREs) were derived from direct experimental techniques, while the rest originated from the analyzed miRNA high-throughput perturbation datasets (64 microarray and 12 RNA-Seq experiments).

Importantly, approximately 28% of the positive MREs, including 1,131 chimeric and reporter assay-verified interactions, were observed to be exclusively resolved by non-T-to-C AGO-enriched clusters. Consequently, downstream evaluations were initially centered on the comparison of MRE-specific feature distributions between clusters lacking or containing T-to-C sites. Known important attributes were calculated for miRNA-target recognition such as the AU flanking content, binding type, matches per miRNA-target duplex domain, minimum free energy, GU wobble pairs and MRE conservation. Evaluated descriptors of miRNA positive interactions residing on T-to-C clusters significantly diverge from respective densities observed in negative MREs (Figure 22, range of P values $_{T-to-C}$: 5.9×10^{-198} - 4×10^{-7} , two-tailed Wilcoxon rank-sum test, $n_{T-to-C} = 4,310$, $n_{negative} = 1,423$). It is shown that features related to miRNA targeted sites on non-T-to-C clusters also significantly differentiate from relevant estimates corresponding to negative miRNA-target instances (Figure 22, range of P values $_{non-T-to-C}$: 7.8×10^{-139} - 14×10^{-5} , two-tailed Wilcoxon rank-sum test, $n_{non-T-to-C} = 1,700$, $n_{negative} = 1,423$).

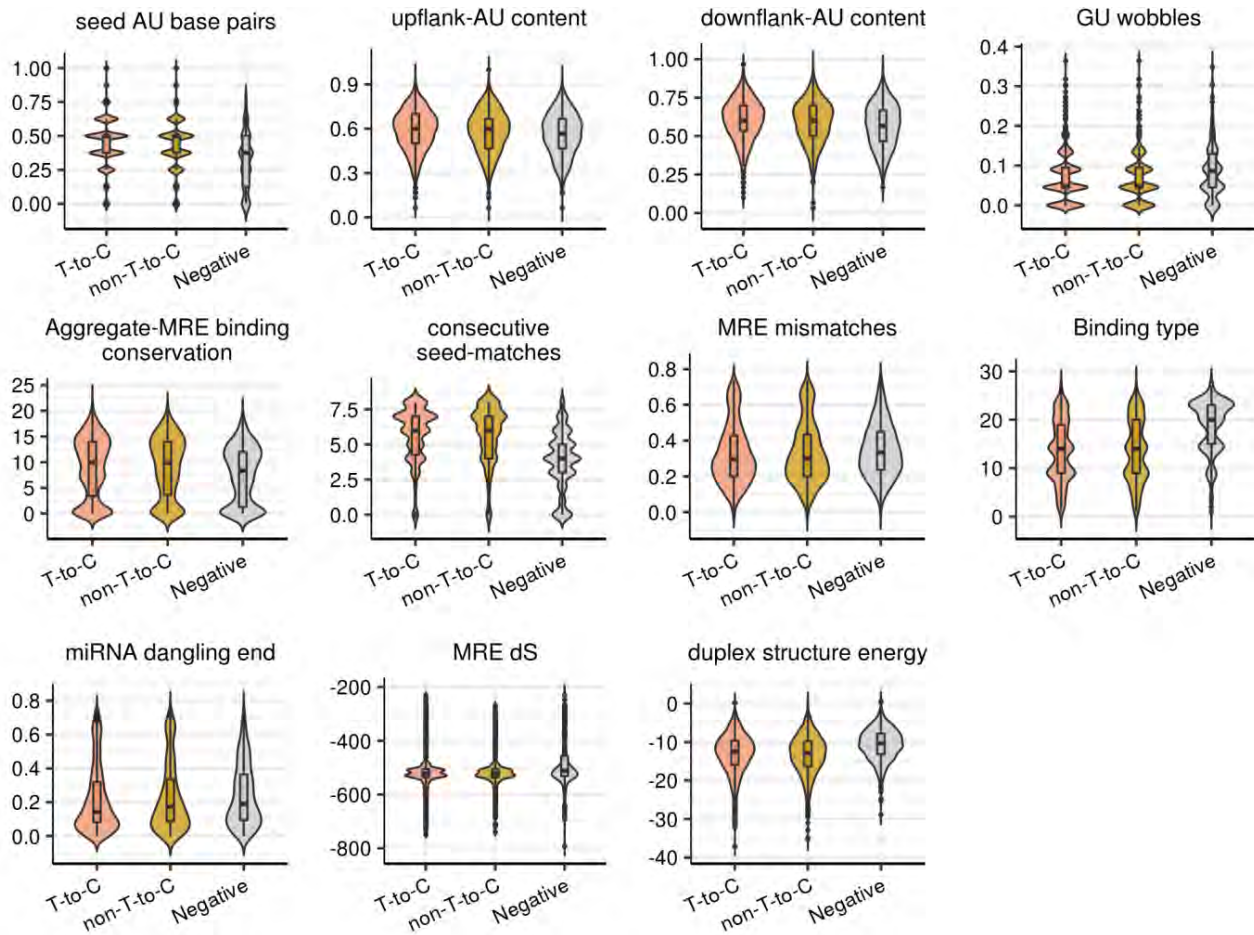


Figure 22: Distributions of MRE-related features corresponding to positive miRNA interactions in T-to-C and non-T-to-C AGO-bound regions against the relevant densities of negative binding sites. Assessed characteristics of positive miRNA interactions on (non-)T-to-C clusters significantly diverge from respective feature distributions of negative MREs (two-tailed Wilcoxon rank-sum test) (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

3.2.2 Structural sequencing data unveil accessible AGO-bound loci

PARS sequencing profiles were calculated around AGO-PAR-CLIP-derived miRNA binding sites in 4 EBV transformed lymphoblastoid cell lines[53]. The analysis of the respective RNase S1 or V1 nuclease signals/intensities at single base resolution enabled the assessment of miRNA site accessibilities in both T-to-C and non-T-to-C clusters. These measurements were juxtaposed against negative MREs comprising miRNAs expressed in the examined lymphoblastoid cell types. The per base averaged PARS scores indicate that strong structural accessibility occurs in the 3' end of miRNA-target sites and specifically on 2-4nt positions of the miRNA seed region. These results were identified on interactions residing on (non-)T-to-C clusters and significantly differ from respective base scores along negative MREs located on AGO-enriched peaks (Figure 23, yellow window; Methods, range of P values $T\text{-to-C}$: 0.03 - 3.7×10^{-5} , P values non-T-to-C : 0.01 - 2.4×10^{-5} , two-tailed Wilcoxon rank-sum test, $3,260 < n_{T\text{-to-C}}$ sites

$< 9,159, 2,119 < n_{\text{non-T-to-C sites}} < 6,473, n_{\text{negative sites}} = 3,059$). The outcome of this analysis is consistent with previous observations[119] and demonstrates that the highest accessibility segregating functional from non-functional binding sites resides towards the initiation of the direct miRNA seed pairing.

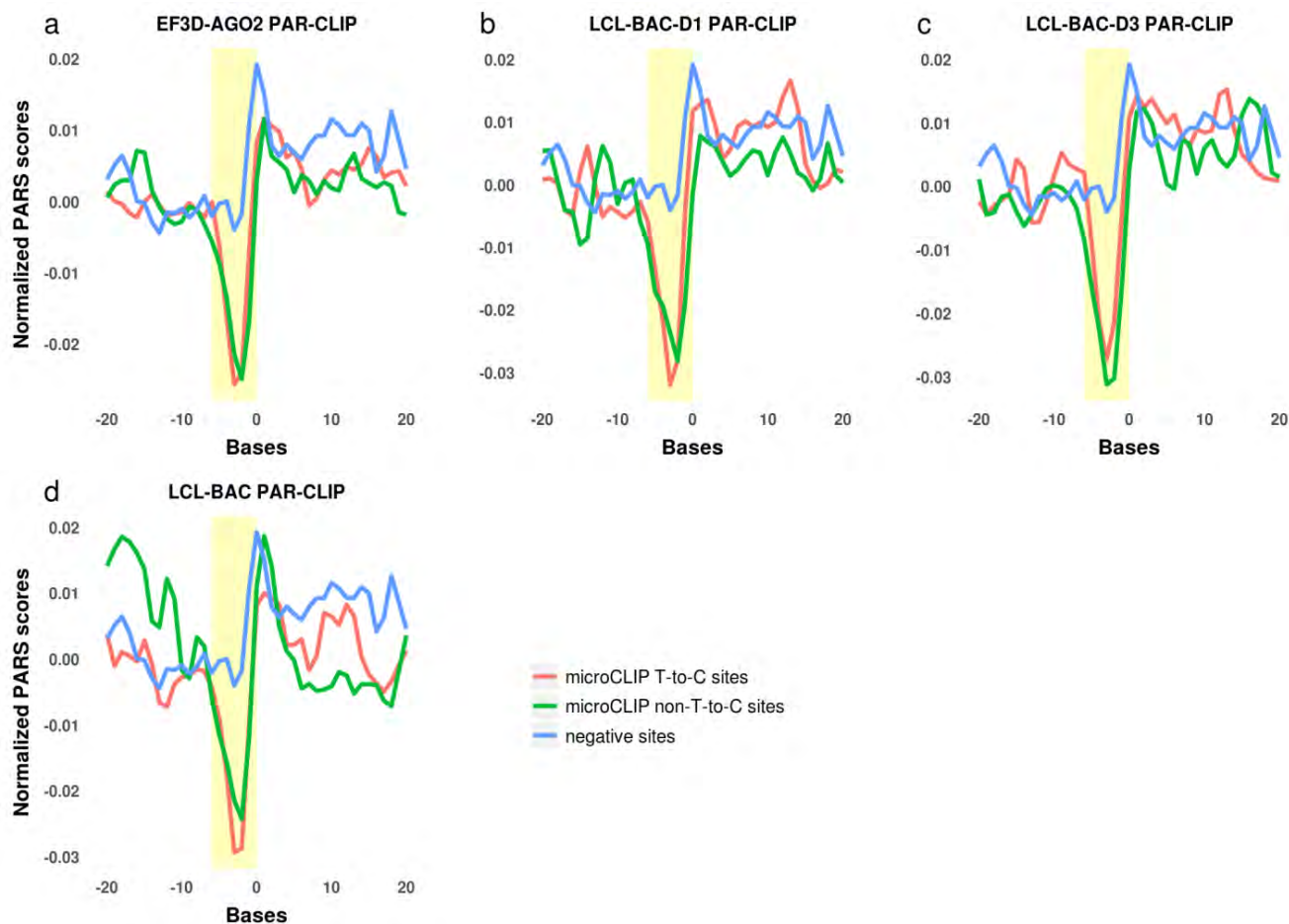


Figure 23: Average PARS scores of AGO-bound regions deduced from the analysis of 4 EBV transformed lymphoblastoid PAR-CLIP libraries. RSS base signals were aligned to the start of the miRNA-target binding site. Base 0 corresponds to the 3'-end of the mRNA, at -1 or -2 nt downstream of the initiation of the direct miRNA seed pairing. Negative PARS scores correspond to single stranded RNA structures, while positive scores to double stranded sites. In the examined AGO-PAR-CLIP EF3D-AGO2(a), LCL-BAC-D1(b), LCL-BAC-D3(c) and LCL-BAC(d) datasets, strong structural accessibility occurs in miRNA sites identified on T-to-C (red) and non-T-to-C (green) clusters in the 2-4nt positions (yellow window) of the miRNA seed pairing. These results significantly differ from respective base scores along negative MREs (light blue) located on AGO-enriched peaks (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

3.2.3 A super learning approach for AGO-PAR-CLIP analysis

All the aforementioned observations have been incorporated in an extensive *in silico* framework. microCLIP is based on ensemble super learning and provides a complete pipeline for experimentally supported miRNA targetome annotation, initiating from aligned

(.sam/.bam) PAR-CLIP sequencing reads. This algorithm, contrary to existing leading implementations, operates on every AGO-enriched cluster, utilizing the previously neglected non-T-to-C clusters.

Distribution of base model scores on positive and negative instances and their respective performance, in terms of sensitivity and specificity in an independent test set of approximately 4,000 instances, are depicted in Figure 24. All the classifiers achieved high performance in a range of sensitivity 73.4% - 92.7% and specificity 67.6% - 86.8% (range of AUC: 75.3% - 95.4%). Their aggregated outcome in the meta-learner of microCLIP framework is provided in a separate curve and exhibits the highest performance in terms of sensitivity and specificity (sensitivity: 96.0, specificity: 87.4, AUC: 95.5%). The individual performance of internal classifiers (DL, RF, GBM, GLM) in microCLIP base models adopting a super learner approach is shown using the same set in Figure 25 and Figure 26.

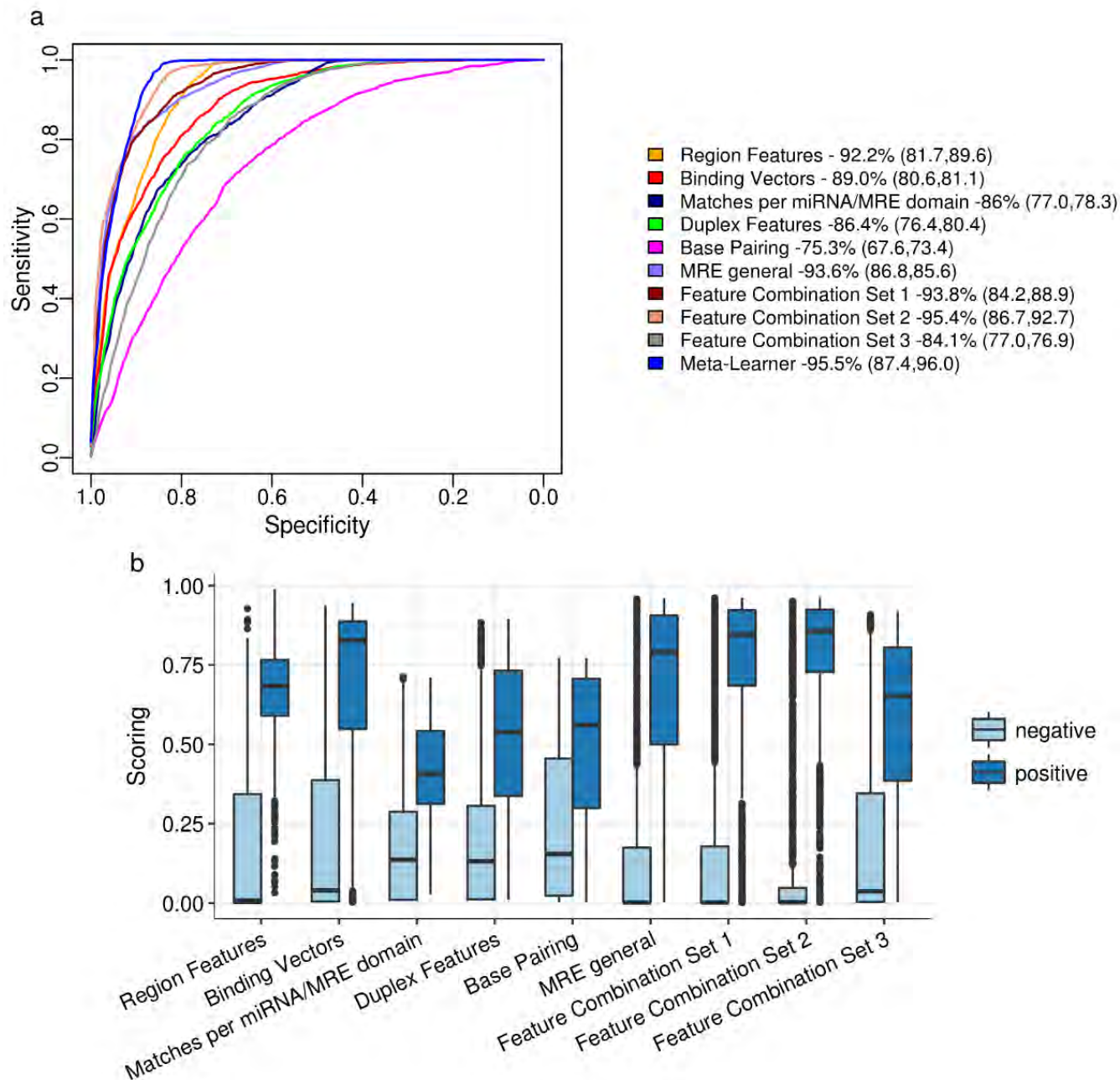


Figure 24: Evaluation of the accuracy of the 9 base model classifiers. Five-fold cross-validation has been implemented on a separate set of approximately 4,000 instances to test the performance of each node. a) ROC curve of each base model displays the classification of positive/negative miRNA binding sites. b) Distribution of base model scores estimated on positive/negative instances of the test set (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

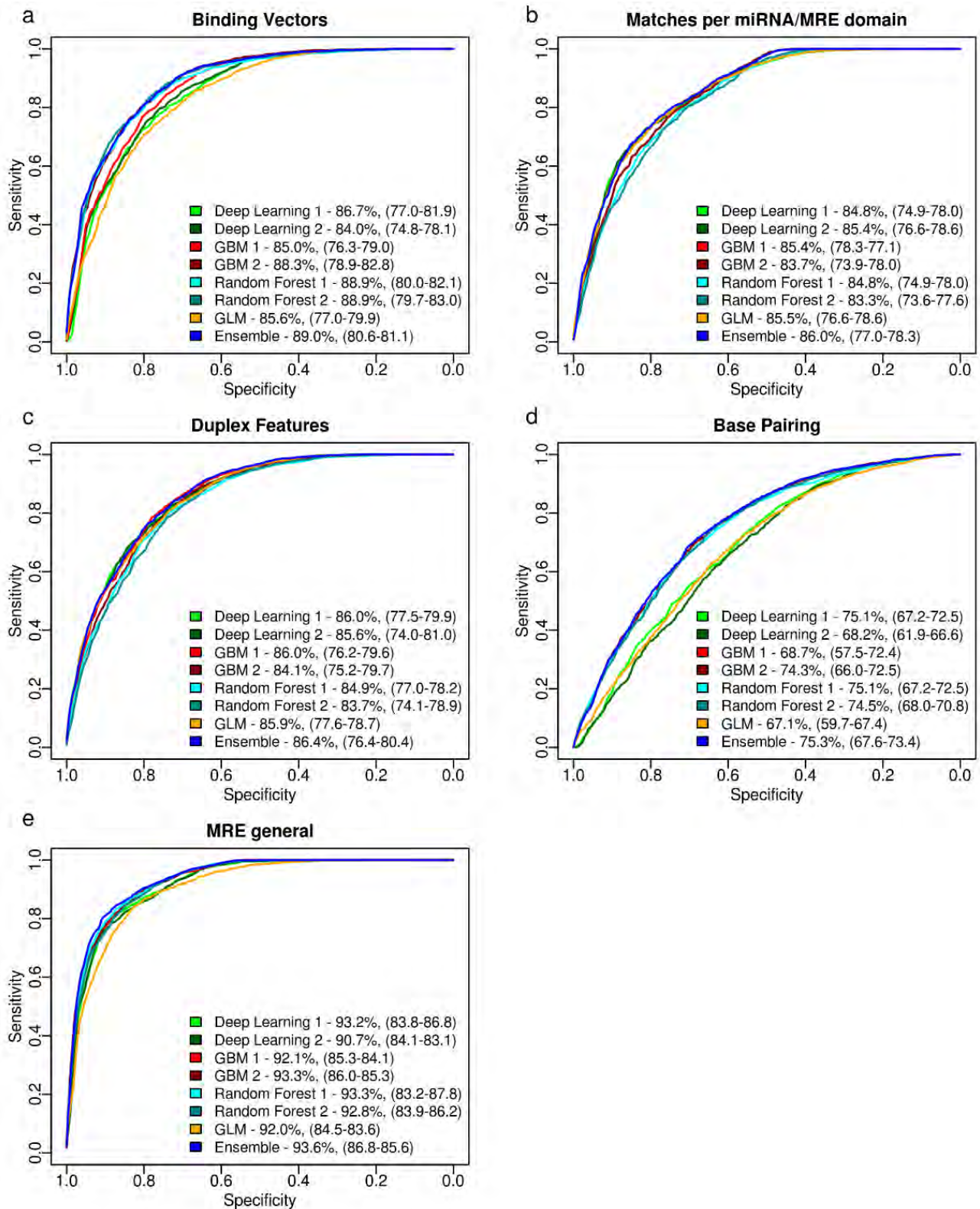


Figure 25: Evaluation of constitutive/internal classifiers of 5 microCLIP base models that adopt a super learning approach. Five-fold cross-validation was applied on a separate set (same as in Figure 23), to test the performance of the seven individual Random Forest (RF), Generalized Linear Model (GLM), Gradient Boosting Model (GBM), Deep Learning (DL) classifiers (2 RF, 2 GBM, 2 DL, 1 GLM models) in each base

node. Different colors are consistently utilized to display ROC curves of each sub-classifier incorporated in 'Binding Vectors', 'Matches per miRNA/MRE domain', 'Duplex Features', 'Base pairing' and 'MRE general' base nodes respectively. Information concerning sensitivity, specificity and AUC of each model is shown in the figure legends. The performance of ensemble deep learning models that aggregate the seven independent sub-classifiers in each base node are additionally shown (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

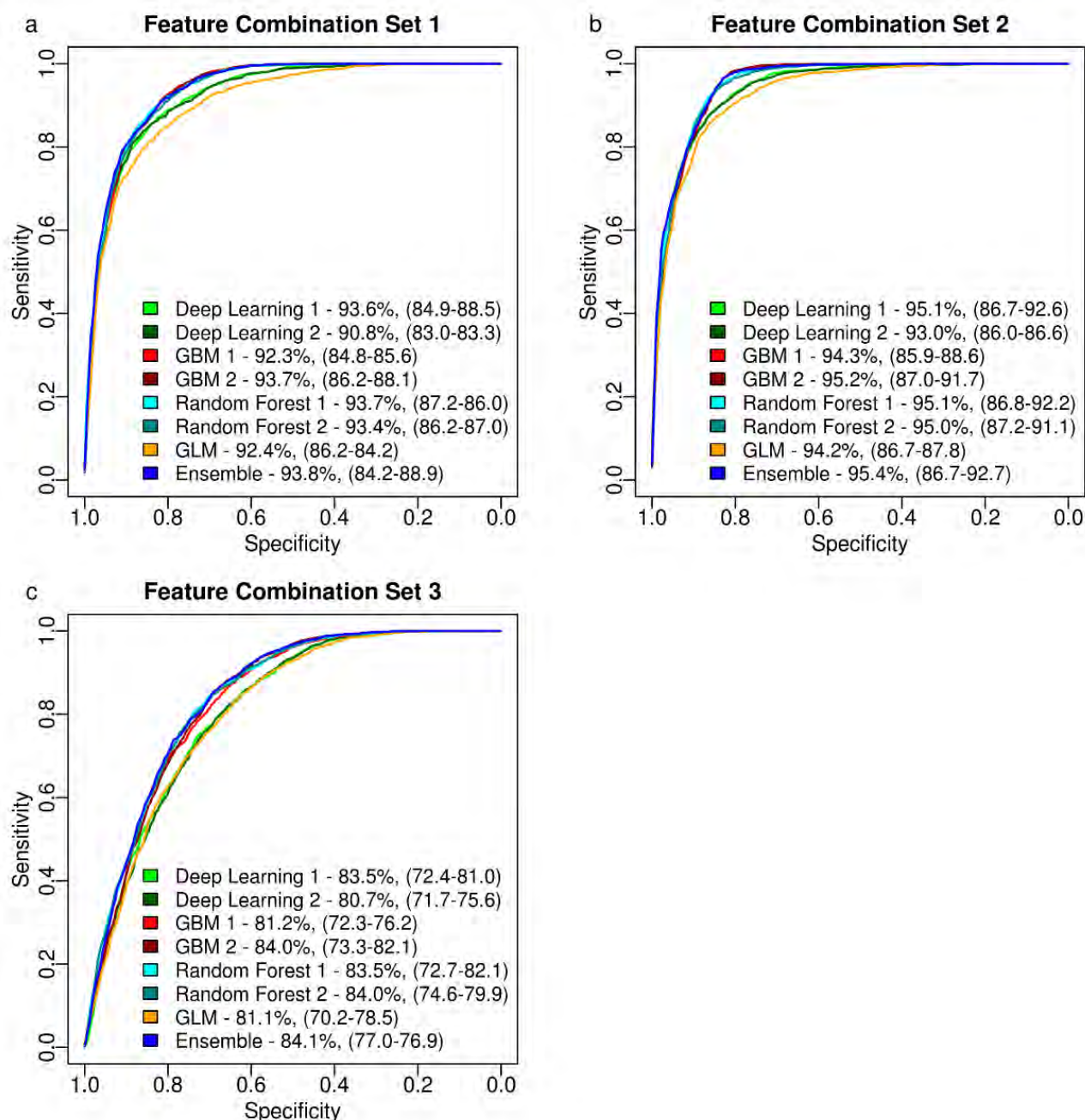


Figure 26: Evaluation of the accuracy of sub-classifiers included in 'Feature Combination Set 1-3' base nodes. The performance of sub-classifiers (2 RF, 2 GBM, 2 DL, 1 GLM models), along with the performance of the ensemble deep learning models that aggregate their output are displayed in distinct colors (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

The multi-layer super learner classification scheme of microCLIP benefits from the incorporation of the complete array of features, maximizing their contribution through their parallel use in different classification models in every node. The impact of weaker features and classifiers in optimal super learner design and behavior is shown in Figure 27, where microCLIP performance was compared to three different classification schemes using an independent validation set of 1,674 positive miRNA binding sites, corresponding to 1,527 miRNA-gene interactions.

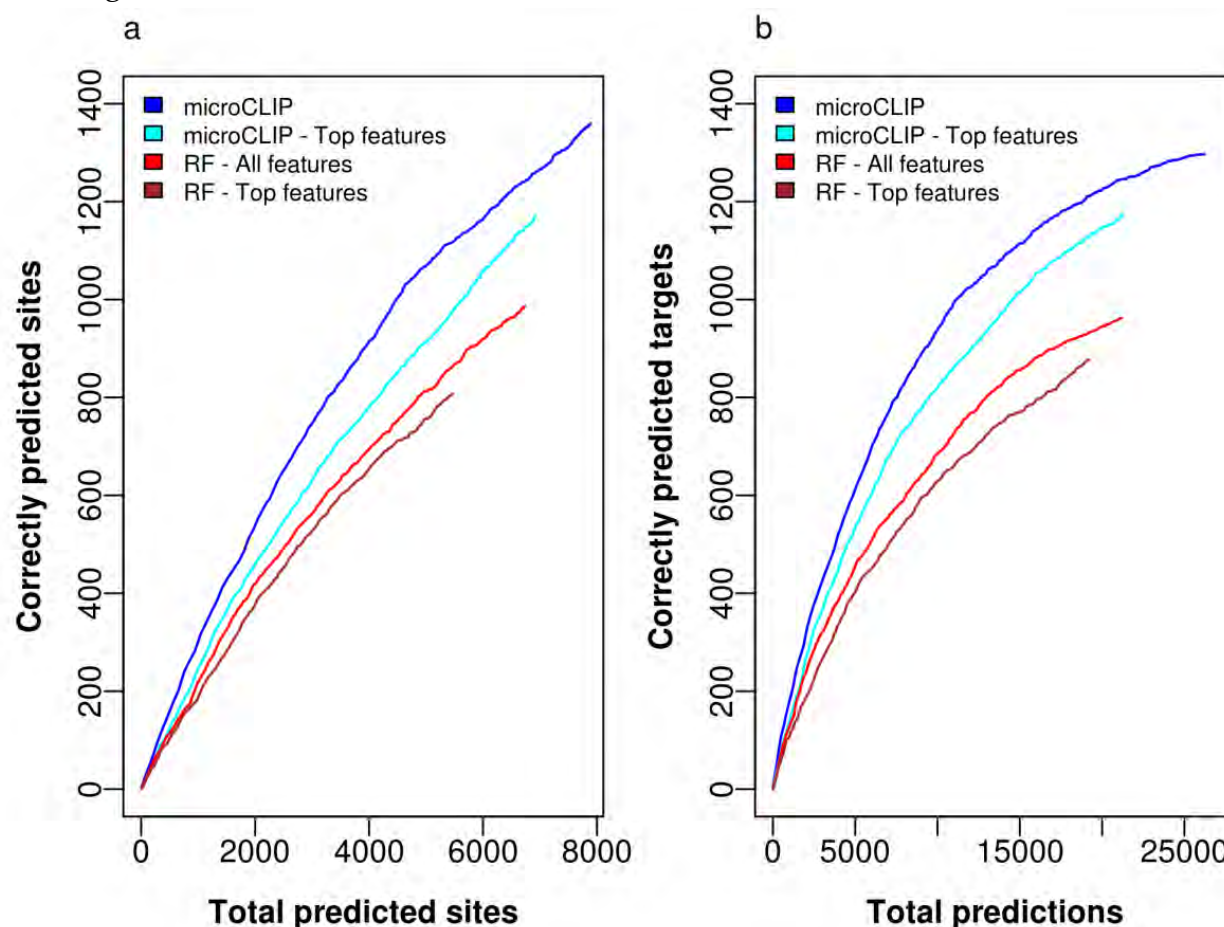


Figure 27: Evaluation of microCLIP performance against 3 alternative classification approaches: a Random Forest classifier comprising all the features; a Random Forest classifier including the top 27 discriminative features ($AUC \geq 65\%$); microCLIP super learner classification scheme including top performing features per base node (70 descriptors in total, $AUC \geq 65\%$). The utilized validation set comprised 1,674 positive miRNA binding sites, derived from experimentally validated direct miRNA interactions. (a) The number of correctly predicted miRNA binding sites for each classification approach is plotted versus the total retrieved predicted sites. (b) A separate comparison captures the models' efficiency to predict correct miRNA-target interactions at different levels of total predictions. The validation set is the same as in (a) collapsed into 1,527 miRNA-gene interactions (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

3.2.4 microCLIP detects novel miRNA interactions from AGO-PAR-CLIP clusters

The analysis of 10 public datasets across different experimental conditions (GEO/SRA accessions GSE28859, GSE59944, GSE41437, SRR1045082, SRR359787) was revisited with microCLIP, in order to explore the extent of miRNA-target pairs that remain uncovered using standard AGO-PAR-CLIP computational approaches. Processed CLIP-Seq libraries were accompanied by RNA-Seq and small RNA-Seq (sRNA-Seq) data to determine the set of expressed transcripts and miRNAs per cell type. By screening every AGO-enriched region, microCLIP reveals a significant portion of targeted genes distinguished only from CLIP clusters presenting no conversion sites. An average $11 \pm 6.4\%$ increase of detected targets was observed across the analyzed experiments. Figure 28 summarizes the miRNA-target interactions per library, supported by T-to-C and/or non-T-to-C peaks, respectively. The retrieved results suggest that the miRNA targetome is not sufficiently covered by inferring targets solely in T-to-C enriched cross-linked regions. The impact of the unrecognized miRNA interactions is also reflected in functional analyses.

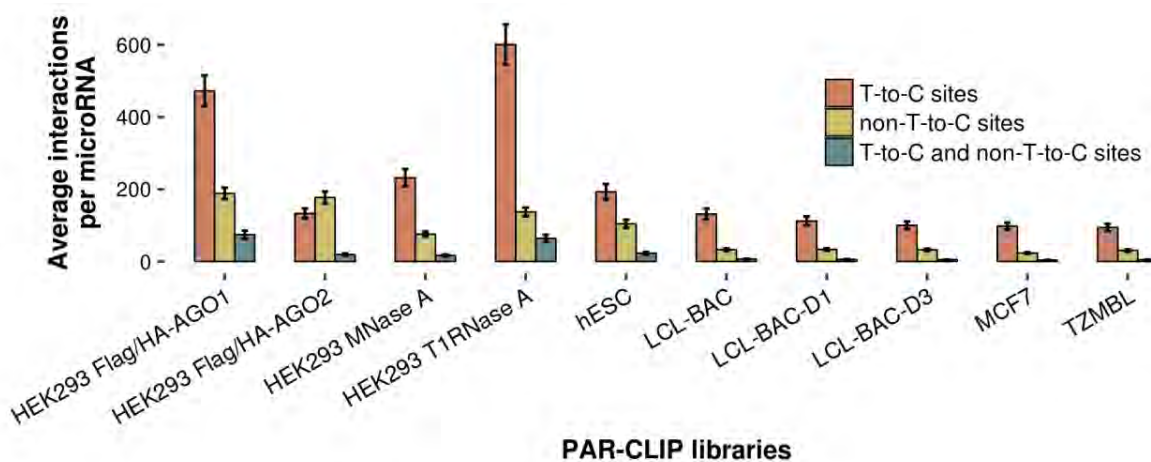


Figure 28: Bar plots featuring the average miRNA-target interactions supported by non-T-to-C and/or T-to-C peaks per examined cell type and experimental condition. Mean and standard errors (error bars) of miRNA interactions are shown per library. An average increase of $14\% (\pm 8.8\%)$ in the detected interactions was observed across analyzed PAR-CLIP libraries by the incorporation of non-T-to-C clusters (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

To investigate the functional importance of miRNA sites residing on AGO-enriched regions presenting insufficient T-to-C substitutions, 17 public high-throughput gene expression profiling datasets following transfection or knockdown of specific miRNAs (GEO accessions GSE60426, GSE52531, GSE68987, GSE37918, GSE21901, GSE14537, GSE35621, GSE46039, GSE21577, microarrays from the study of Selbach *et al*. [46]) were utilized. These experiments were complemented with AGO-PAR-CLIP datasets conducted in relevant cell types. microCLIP was applied to detect miRNA-gene interactions on HEK293, MCF7 and TZMBL PAR-CLIP libraries (Kishore *et al*. [34], Farazi *et al*. [54], Whisnant *et al*. [120]). Response of

targeted mRNAs to miRNA deregulation was evaluated independently per tested cell type. In the conducted comparisons, target fold changes in 3 distinct groups were measured: (i) mRNAs presenting at least one predicted MRE on T-to-C clusters, (ii) mRNAs participating in interactions resolved only by non-T-to-C clusters, (iii) transcripts lacking sites for the examined miRNAs. In all miRNA perturbation experiments, detected targets overlapping (non-)T-to-C clusters were significantly downregulated or upregulated upon transfection or knockdown of different miRNAs compared to transcripts having no miRNA binding site (Figure 29, range of P values $_{T-to-C}$: 5.1×10^{-138} - 11×10^{-3} , P values $_{non-T-to-C}$: 8.5×10^{-29} - 37×10^{-3} , two-tailed Wilcoxon rank-sum test, $51 < n_{T-to-C} < 1,569$, $11 < n_{non-T-to-C} < 344$, $2,677 < n_{no-site} < 12,330$). Regardless of the perturbation type, T-to-C clusters were observed to relate to more responsive targets at equal numbers of predicted sites (Figure 29, range of P values $_{(b-f)}$: 2.7×10^{-11} - 3.9×10^{-2} , two-tailed Wilcoxon rank-sum test, $11 < n_{T-to-C/non-T-to-C} < 344$).

The definition of T-to-C locations varies in relevant publications and describes T-to-C loci as those that are covered with reads having at least 5-25% T-to-C substitutions[11, 121-124]. For the analyses presented in the aforementioned figures, a minimum 20% T-to-C incorporation ratio defines T-to-C clusters. The selected T-to-C percentage threshold is considered of medium stringency to confidently identify clusters following the experiment's specifications.

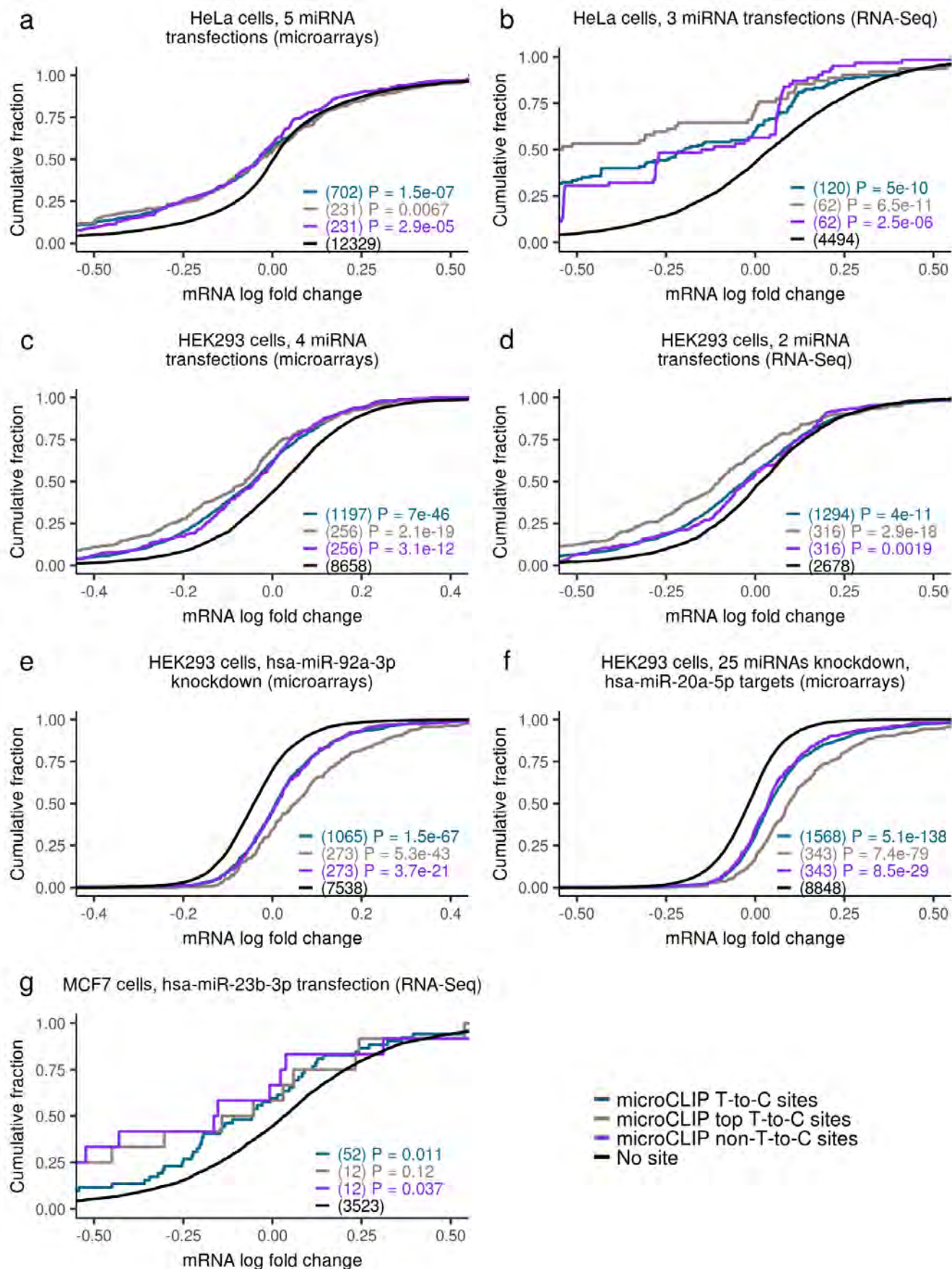


Figure 29: Functional efficacy of microCLIP-detected MREs residing on T-to-C and non-T-to-C AGO-bound enriched regions. miRNA binding sites were obtained from the analysis of PAR-CLIP libraries in 3 different cell types. The functional efficiency of predicted targets was examined in 17 public gene expression profiling datasets following miRNA transfection or knockdown. Response of targeted mRNAs to miRNA perturbation experiments was evaluated independently per tested cell type, experimental technique and conditions (a-g). Cumulative distributions of mRNA fold changes for targets comprising at least one predicted MRE on T-to-C clusters or supported only by non-T-to-C peaks were compared to those that lack any site of the considered miRNAs. The number of transcripts included in each category is presented in parentheses. Identified targets supported by T-to-C and non-T-to-C clusters exert a significant difference in expression changes compared to transcripts lacking any predicted binding site (two-tailed Wilcoxon rank-sum test). At same numbers of T-to-C and non-T-to-C sites, the former group relates to more responsive targets at miRNA perturbation experiments in (b-f) (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

3.2.5 Functional enrichment shows importance of non-T-to-C targets

To demonstrate the ability of detected non-T-to-C interactions to statistically empower downstream analyses, a functional enrichment investigation on KEGG pathways was conducted in highly scored miRNA-target pairs from an independent AGO-PAR-CLIP dataset in MCF7 cells (Farazi *et al.*[54]). The dataset was analyzed with microCLIP, while the 100 most highly expressed miRNAs and their targets in 3' UTR regions were retained.

8,921 and 846 unique interactions retrieved from T-to-C and non-T-to-C peaks, respectively, were utilized to form two gene sets: one containing unique T-to-C targets ($n = 396$), and one combining T-to-C and non-T-to-C targets ($n = 491$). 391 genes were common between the two. Pathway analysis of T-to-C targets resulted in 63 significantly enriched terms ($P < 0.01$, one-sided Fisher's exact test, Benjamini-Hochberg adjustment, $6 < n_{\text{T-to-C}} < 51$), while the combined set yielded 67 enriched terms ($P < 0.01$, one-sided Fisher's exact test, Benjamini-Hochberg adjustment, $6 < n_{(\text{non-})\text{T-to-C}} < 58$). An average of 2.4 more targets per pathway was observed when non-T-to-C interactions were included.

In both analyses, top-ranking terms were pathways modulating endocrine resistance, growth factor receptor signaling and typical tumor-related processes, like cell growth, migration and apoptosis. Numerous cancer pathways occupied top positions based on P value scores (Figure 30). This elementary analysis indicated that non-T-to-C peaks assisted in discovering more targeted pathway members.

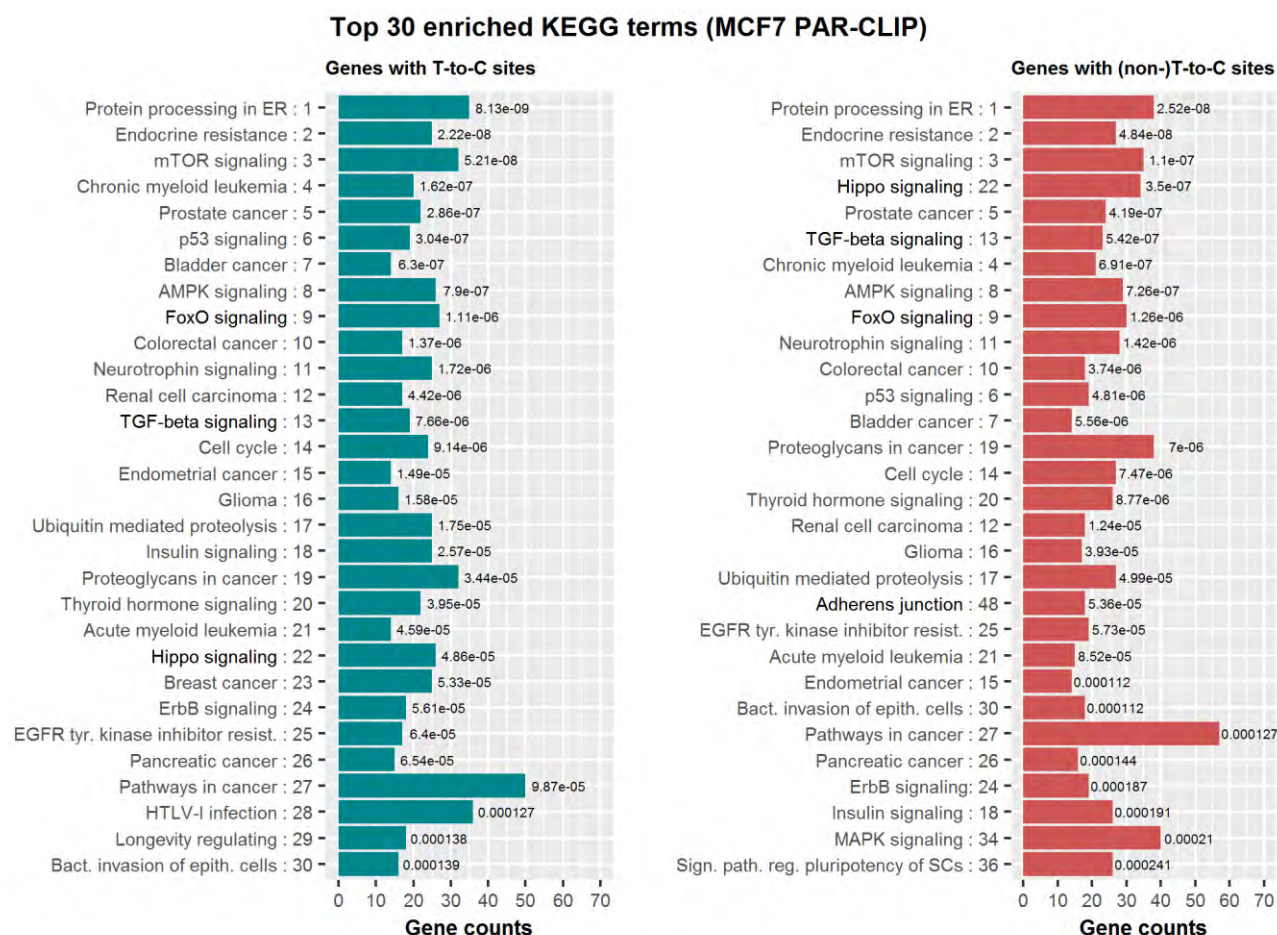


Figure 30: Functional significance of (non-)T-to-C sites in MCF7 AGO-PAR-CLIP dataset. Top 30 KEGG pathways enriched by T-to-C or (non-)T-to-C (combined T-to-C and non-T-to-C) peak containing genes. X-axis depicts number of genes enriching each term. Pathways are ranked according to the enrichment *P* value shown at the end of each bar. The T-to-C site enrichment rank is provided after pathway description to facilitate comparison with gene set of (non-)T-to-C sites (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

To further validate pathway-related interactions from (non-)T-to-C clusters, we investigated miRNA-target expression associations in 271 breast cancer patient samples indexed in TCGA[125]. miRNA and mRNA expression profiles were measured by ductal breast cancer sRNA-Seq and RNA-Seq samples obtained from Firehose (http://gdac.broadinstitute.org/runs/stddata_2016_01_28). In downstream analysis 13,346 mRNAs and 322 expressed miRNAs were incorporated. Pearson correlation analysis of expression across samples was conducted for each miRNA-target pair contained in enriched KEGG terms. miRNA-gene expression associations, evaluated separately for interactions resolved by T-to-C and non-T-to-C clusters, are depicted in cumulative distribution plots (Figure 31). The analysis confirmed a significant shift of pathway-related miRNA-target interactions towards more negative correlation coefficients, when compared against a randomly selected subset from all miRNA-gene interacting pairs lacking target sites for the

highly expressed miRNAs ($P_{\text{T-to-C}} = 6.7 \times 10^{-22}$, $P_{\text{non-T-to-C}} = 8 \times 10^{-4}$, two-tailed Wilcoxon rank-sum test, $n_{\text{T-to-C}} = 2,299$, $n_{\text{non-T-to-C}} = 494$, $n_{\text{no-site}} = 4,000$).

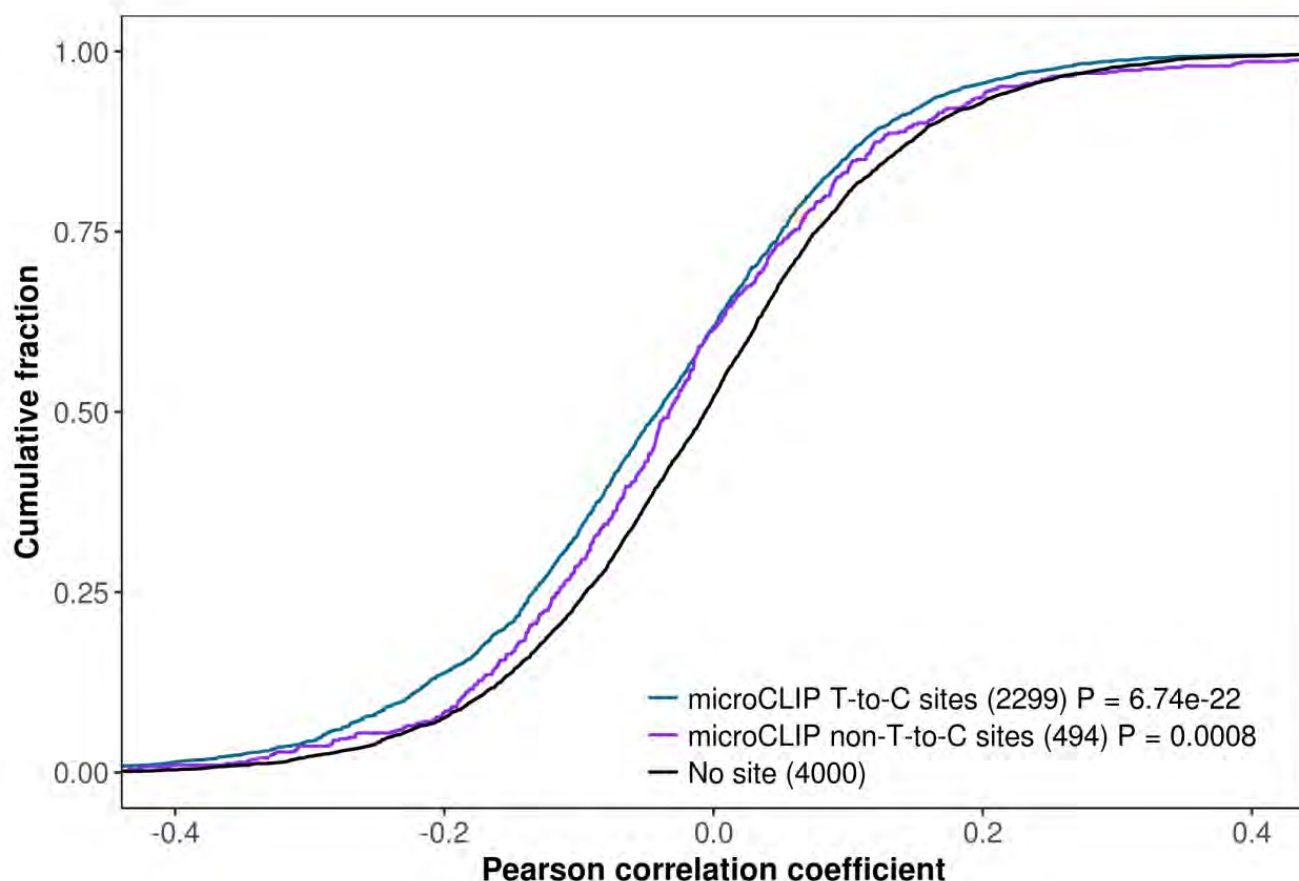


Figure 31: Correlation analysis of expression of pathway-related miRNA-target interactions across 271 TCGA ductal breast cancer samples (patients). Cumulative distributions of miRNA-target expression relationships, evaluated for interactions supported by T-to-C or non-T-to-C AGO-bound regions were compared to a randomly selected set from all the remaining miRNA-gene interacting pairs lacking any target site of the highly expressed miRNAs. The number of genes considered in each category is presented in parentheses. Pathway-related miRNA-target interactions supported by T-to-C and non-T-to-C clusters reveal a significant shift towards more negative correlation coefficient values compared to the no-site distribution (two tailed Wilcoxon rank-sum test) (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

3.2.6 Evaluation of microCLIP against AGO-CLIP-guided models

To assess microCLIP accuracy and to estimate the information gain with the incorporation of non-T-to-C AGO-enriched regions, we compared the model against MIRZA[55], microMUMMIE[56] and PARma[58]. In the evaluation process, AGO-CLIP-guided algorithm performance was also contrasted with Targetscan v7[20] de novo miRNA-target prediction algorithm. A model adopting the same super learning scheme, including information only from T-to-C enriched sites, microCLIP T-to-C, was also deployed. Clusters from the training set incorporating adequate T-to-C transition sites were selected as input to re-train the super

learning classifier. Additional support for the robustness of CLIP-guided super learner classification irrespective of non-T-to-C site inclusion is provided through the inclusion of microCLIP T-to-C algorithm in the evaluation process.

The performance evaluation was initially accomplished against unified sets of 4 microarray and 2 RNA-Seq public datasets in which miRNAs were individually transfected into HEK293 cells (GEO accessions GSE60426, GSE52531, GSE21901, GSE14537, GSE35621). An extensive list of interactions for each CLIP-guided program was derived from the analysis of 7 PAR-CLIP HEK293 libraries (Kishore *et al.*[34], Memczak *et al.*[110]). Each miRNA-target pair was characterized by the highest scored miRNA binding site overlapping coding or 3'UTR exons, since utilized algorithms provided MRE-oriented prediction scores. In cases of multiple transcript-gene associations, the transcript with the longest 3'UTR was selected. The retrieved MREs were juxtaposed with deregulated targets identified in the gene expression profiling experiments. To determine the ability of each method to identify the most strongly downregulated targeted genes, detected interactions were ranked according to their provided scores. The median fold changes (\log_2) of the top predicted targets for the different algorithms were subsequently estimated and accordingly compared by applying stepwise thresholds of total predictions. The performance of implementations was additionally evaluated against averaged log fold changes of 1000 randomly selected genes (without replacement). The mean \log_2 fold change values of the randomly selected genes in different stepwise thresholds were taken and the median curve derived from these values was calculated. Genes with zero fold-change indication were filtered out from the random selection process.

In the examined miRNA perturbation experiments, microCLIP-detected targets revealed the strongest repression, compared to all the assessed approaches (range of P values_{microarrays}: 0 – 8.2×10^{-74} , P values_{RNA-Seq}: 0 – 8.1×10^{-30} , two-tailed Wilcoxon signed-rank test, $535 < n_{\text{microarrays}} < 5,529$, $174 < n_{\text{RNA-Seq}} < 3,129$; Figure 32) and to randomly selected genes ($P_{\text{microarrays}} = 3.3 \times 10^{-165}$, $P_{\text{RNA-Seq}} = 3.3 \times 10^{-165}$, two-tailed Wilcoxon signed-rank test, $n_{\text{microarrays}} = 1,000$, $n_{\text{RNA-Seq}} = 1,000$; Figure 32). microCLIP uncovered interactions with stronger functional impact, when equivalent numbers of top predictions, ordered from highest to lowest scores, were compared. Importantly, the predictions of the tested algorithms were significantly more responsive than expected by chance (range of P values_{microarrays}: 3.3×10^{-165} – 2×10^{-89} , P values_{RNA-Seq}: 3.3×10^{-165} – 1.8×10^{-30} , two-tailed Wilcoxon signed-rank test, $535 < n_{\text{microarrays}} < 1,001$, $174 < n_{\text{RNA-Seq}} < 1,001$; Figure 32).

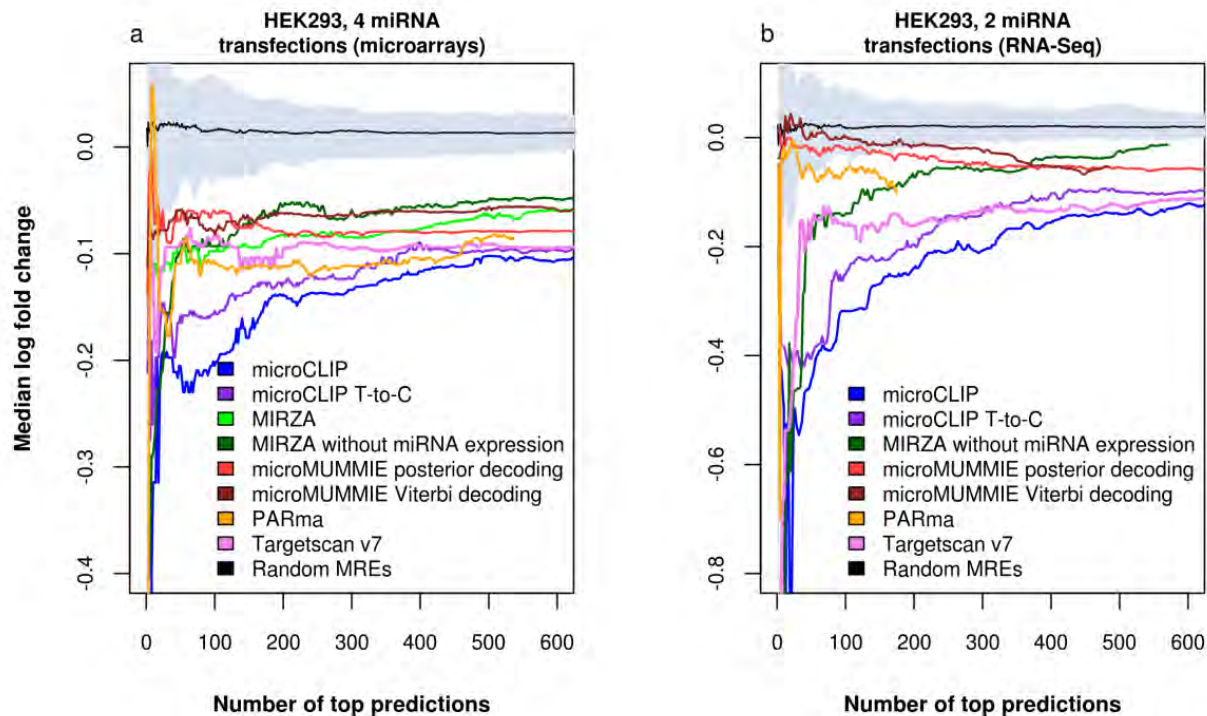


Figure 32: Assessment of microCLIP prediction efficacy against microCLIP T-to-C, MIRZA, microMUMMIE, PARma and Targetscan v7. miRNA-target pairs for each AGO-CLIP *in silico* approach were obtained from the analysis of 7 PAR-CLIP HEK293 libraries and functional investigation was performed by measuring mRNA responses to miRNA perturbations. Unified sets of (a) 4 microarray and (b) 2 RNA-Seq datasets, in which miRNAs were individually transfected into HEK293 cells, were included in the evaluation process. Median fold change-values (\log_2) of the top predicted targets per tested algorithm were plotted and accordingly compared by applying stepwise cutoffs on total predictions. Performed comparisons additionally incorporate a group comprising mean fold changes of 1000 randomly selected genes (without replacement) by using 100 re-samplings. microCLIP significantly outperforms all the juxtaposed implementations, detecting targets with the strongest median downregulation, from stringent to loose prediction thresholds. microCLIP T-to-C also exhibits greater efficacy than the rest *in silico* approaches (range of P values microarrays: 0 - 2.2×10^{-7} , P values RNA-Seq: 5.5×10^{-265} - 3.6×10^{-29} , two-tailed Wilcoxon signed-rank test, $535 < n_{\text{microarrays}} < 3,223$, $174 < n_{\text{RNA-Seq}} < 1,613$), (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

The performance of microCLIP, MIRZA, microMUMMIE, PARma and Targetscan v7 was also tested using 3 HEK293 and 4 HeLa expression profiling datasets following miRNA perturbation. Interactions were obtained by analyzing HEK293 and HeLa AGO-PAR-CLIP libraries (GEO accessions: GSM714644, GSM1462574) reported in studies by Kishore *et al.*[34] and Whisnant *et al.*[120], while each miRNA-target pair was characterized by its associated miRNA binding site with the highest score. To ascertain an impartial evaluation, cumulative distributions of fold changes were compared for equivalent sets of top predicted targets, i.e. genes with one or more predicted MRE, against genes lacking any site(s) for the considered

miRNAs. microCLIP exerted significant differences in expression changes compared to transcripts lacking any predicted binding site (range of P values _(a-g): 3.2×10^{-71} – 1.3×10^{-6} , one-sided Kolmogorov-Smirnov test, $6,764 < n_{\text{no-site}} < 13,122$). Compared to the other CLIP-guided implementations, microCLIP displayed the greatest site effectiveness in most cases (range of P values _(a-f): 3.1×10^{-13} – 0.031 , one-sided Kolmogorov-Smirnov test, $70 < n < 321$; Figure 33a-f). In Figure 33g, it performed similarly as PARma and better than the rest implementations (range of P values _(g): 0.0005 – 0.1 , one-sided Kolmogorov-Smirnov test, $n = 192$). In this evaluation, Targetscan achieved similar site efficacy as microCLIP in Figure 33c,d,g. microCLIP demonstrated overall more robust performance compared to this sequence-based predictor (range of P values _(a-g): 0.002 – 0.5 , one-sided Kolmogorov-Smirnov test, $70 < n < 321$).

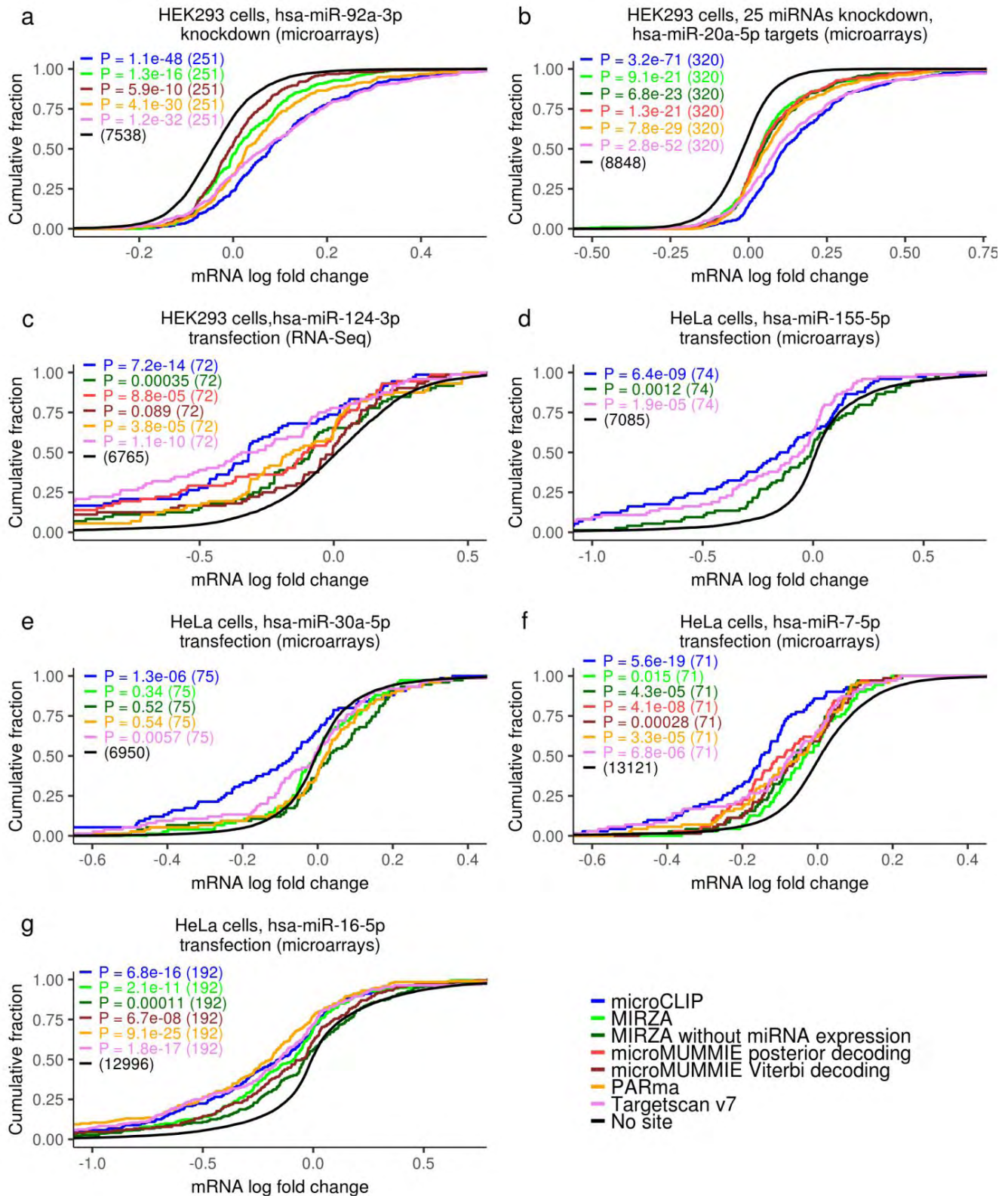


Figure 33: microCLIP performance compared to MIRZA, microMUMMIE, PARma and Targetscan v7 was examined in 7 public gene expression profiling datasets following miRNA transfection or knockdown in HEK293 and HeLa cell lines. miRNA-target interactions for AGO-CLIP *in silico* approaches were obtained

from the analysis of PAR-CLIP HEK293 and HeLa libraries. Response of targeted mRNAs to miRNA perturbation experiments was evaluated independently per tested cell type, experimental technique and condition (a-g). Cumulative distributions of mRNA fold changes for targets comprising at least one predicted MRE in the CDS or 3' UTR regions were compared to those that lacked any site of the considered miRNAs (one-sided Kolmogorov-Smirnov test). Functional efficacy was assessed for equal numbers of top predictions per implementation. Implementations that did not support targets with a fold-change in the examined miRNA perturbation experiments were not included in the relevant cumulative plots. (a-f) Identified targets by microCLIP revealed greater site effectiveness than the rest AGO-CLIP-guided implementations. (g) microCLIP performed similarly as PARma and better than the rest of implementations. Targetscan v7 identifies responsive targets, operating on par with *in silico* approaches based on CLIP data such as PARma, while in (c-d) and (g) it displays analogous efficacy as microCLIP. The number of transcripts included in each comparison is denoted in the parentheses (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

A significant aspect of AGO-CLIP-guided implementations, aside from their ability to detect functionally relevant miRNA interactions, is their efficiency to correctly determine *bona fide* miRNA binding sites at a low number of total predictions. Therefore, an extra evaluation was implemented against a validation set of experimentally verified direct miRNA-target pairs to investigate the accuracy of microCLIP-detected interactions compared to existing methods. microCLIP T-to-C model was also tested. The utilized validation set is composed of 1,674 chimeric and reporter assay-verified interactions from 125 miRNAs. The list of predictions for CLIP-guided implementations was obtained from an AGO-PAR-CLIP dataset in HEK293 cells (GEO accession GSM714644), while Targetscan (all predictions) and Targetscan conserved predicted sites were utilized. PARma adopts a seed-based approach and identifies miRNA-families with a perfect k-mer match within the PAR-CLIP regions. Accordingly, its predictions have been transformed from miRNA-family sites to miRNA-targeted sites, where every binding region is assigned to each one of the miRNA-family members. The number of correctly predicted MREs per tested *in silico* method is plotted against the total predictions for different score thresholds (Figure 34a). MIRZA algorithm provides the most probable prediction per cluster. Therefore, an additional evaluation was performed by including only the top scored miRNA binding site per AGO-peak region, in order to ascertain fairness against all implementations (Figure 34b). Since PARma cannot provide a single top prediction at the miRNA level, all miRNAs bound at a specific site with the same score were considered as top predictions. A separate comparison capturing algorithms' efficiency to predict correct miRNA-target interactions at different levels of total predictions was also conducted (Figure 34c). The validation set was the same as in the aforementioned evaluations, collapsed into 1,527 miRNA-gene interactions. Targetscan operated in the absence of AGO-CLIP data, while predicted interactions of CLIP-guided implementations were defined from PAR-CLIP clusters overlapping full transcript regions. The results demonstrate that although Targetscan methods perform well, *in silico* approaches based on CLIP data, like microCLIP and PARma, have a significantly better performance. Baseline seed methodologies with and without conservation only identify a small proportion of the MREs presented in the positive test set

when they operate on AGO-CLIP enriched regions (Figure 34a,b). microCLIP exhibits a markedly greater ability to discriminate miRNA interactions at equivalent numbers of total predictions, providing a significantly higher sensitivity in the algorithm's complete predictions set (Figure 34a,c).

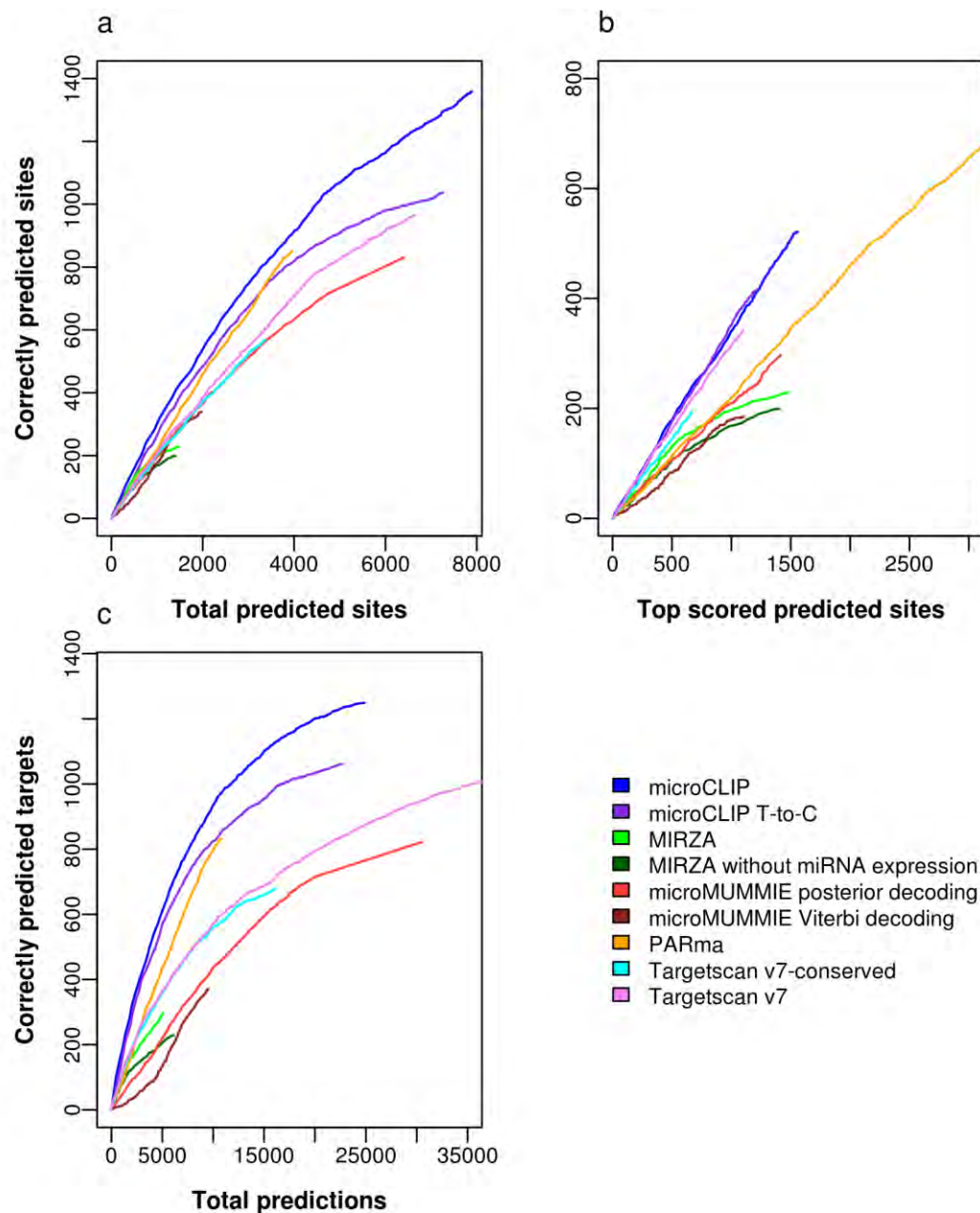


Figure 34: Evaluation of microCLIP performance against microCLIP T-to-C, MIRZA, microMUMMIE, PARma, Targetscan v7 (all predictions) and Targetscan v7 conserved predicted sites. The utilized validation set comprised 1,674 positive miRNA binding sites of 125 miRNAs, derived from chimeric miRNA-target fragments and direct miRNA bindings supported by Reporter Gene Assays. The number of correctly predicted miRNA binding sites for each implementation is plotted versus (a) the total retrieved predictions, (b) the top scored miRNA binding site per AGO-bound enriched region. In (a) and (b) comparisons, we

restrict each program's predictions on PAR-CLIP clusters overlapping the validation test set. A separate comparison (c) captures algorithms' efficiency to predict correct miRNA-target interactions at different levels of total predictions. The validation set is the same as in (a-b) evaluations, collapsed into 1,527 miRNA-gene interactions. For the latter comparison, seed-baseline methods were operating in the absence of AGO-CLIP data, while CLIP-guided implementations on PAR-CLIP clusters overlapping full transcript regions (Paraskevopoulou MD and Karagkouni D *et al*, 2018)[17].

3.3. microT, a Next Generation de novo miRNA-target prediction algorithm

microT is a Next Generation target prediction algorithm that maintains and upgrades the pipeline adopted in microCLIP deployment. This section describes the retrieved outcome from the assessment of descriptors in the pre-processing steps, as well as the performance of the new model in terms of sensitivity and specificity, evaluated on independent test sets. The performance of the model is also assessed against Targetscan v7 and microT-CDS, leading *in silico* approaches in miRNA-target detection field^{22,23}.

3.3.1 Feature selection

In order to demarcate descriptors with high performance, statistic tests and metrics (ROC curves) were implemented in the enhanced training set. Most of the features incorporated into microCLIP, such as AU base pairs, matches and mismatches per miRNA-target duplex domain, binding type, MRE conservation and minimum free energy, presented the same or even higher predictive accuracy. Accessibility features in miRNA binding and in upstream/downstream regions presented also high performance. ROC curves and respective AUC measurements of prominent features, as well as distributions of MRE-related features, corresponding to positive miRNA-target pairs against the relevant densities of negative binding sites, are selectively displayed in the following Figures.

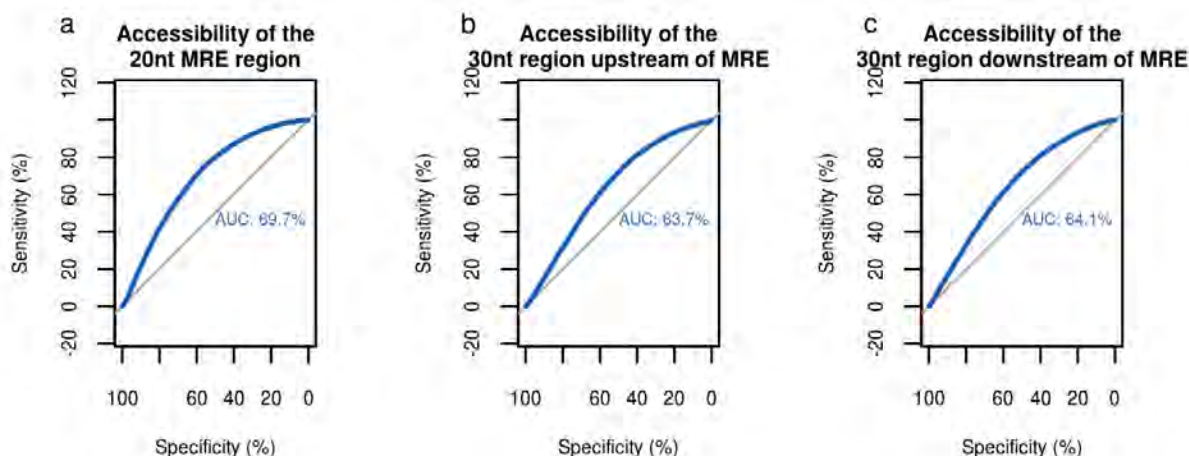


Figure 35: ROC curves of sequence accessibility parameters for the classification of positive/negative miRNA binding sites, i.e. accessibility of the 20nt miRNA binding region and the 30nt region upstream/downstream of the MRE.

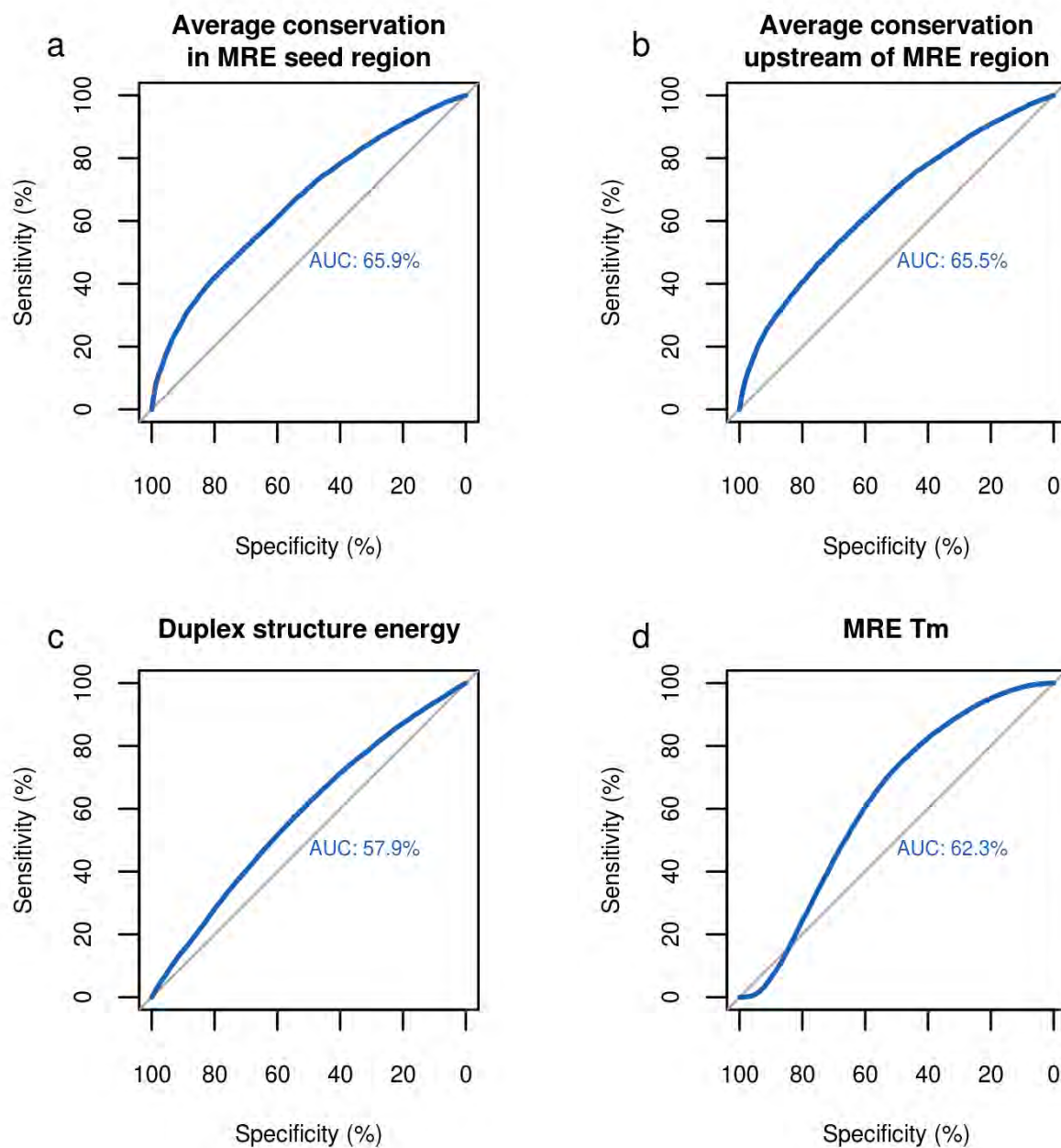


Figure 36: ROC curves for the classification of positive/negative miRNA binding sites indicating the a) aggregated MRE seed binding conservation, b) aggregated conservation in the upstream region of the MRE, c) minimum duplex structure energy and d) MRE-related thermodynamic properties.

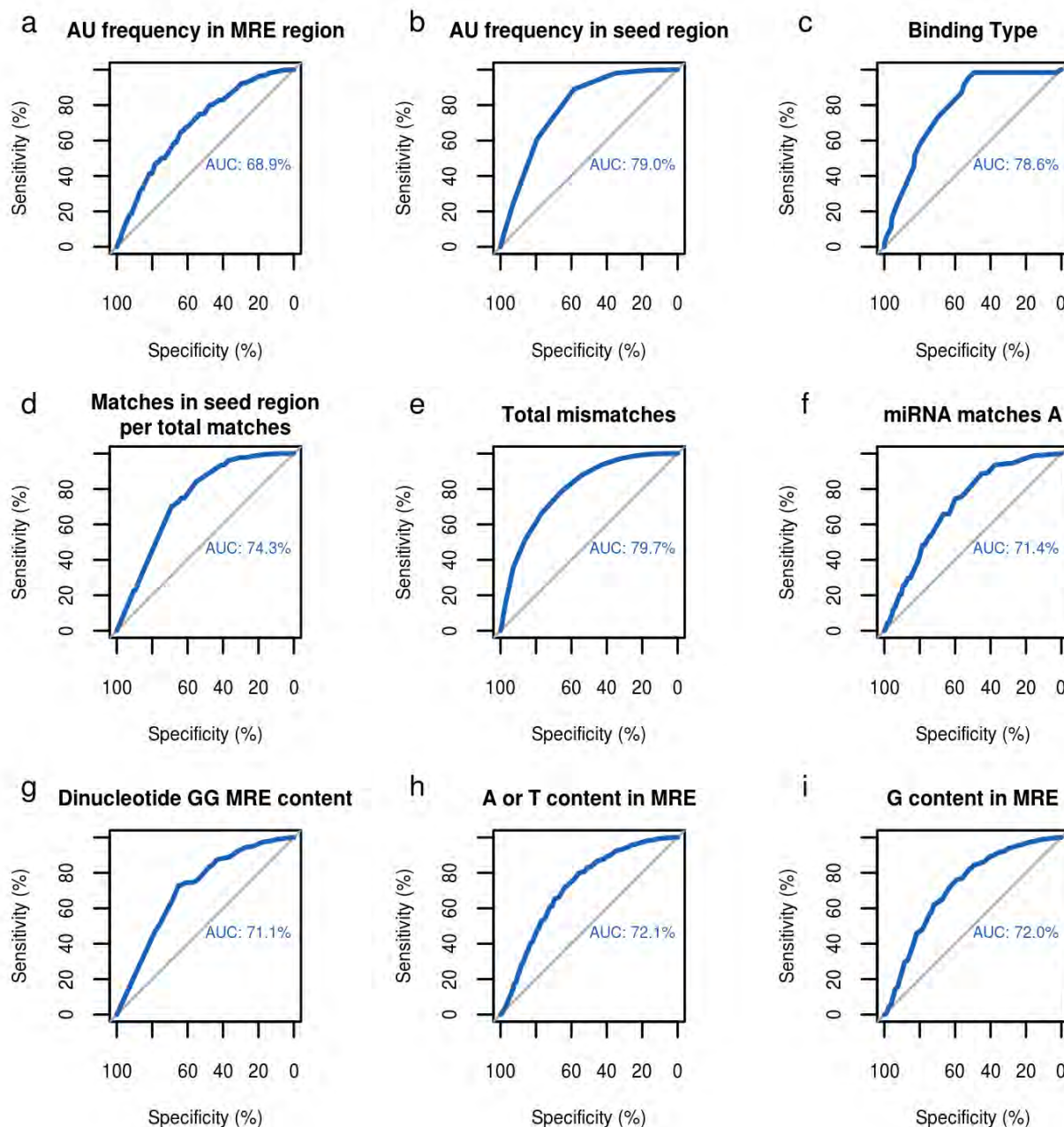


Figure 37: ROC curves for the classification of positive/negative miRNA binding sites indicating AU base pairs (MRE, seed), seed matches and mismatches per miRNA-target duplex domain, nucleotide and dinucleotide MRE content and binding type. The latter feature comprises an extended set of (non-)canonical miRNA base pairings where smaller values indicate stronger seed matches (9mer to 6mer) and greater values correspond to non-canonical and 3' supplementary sites.

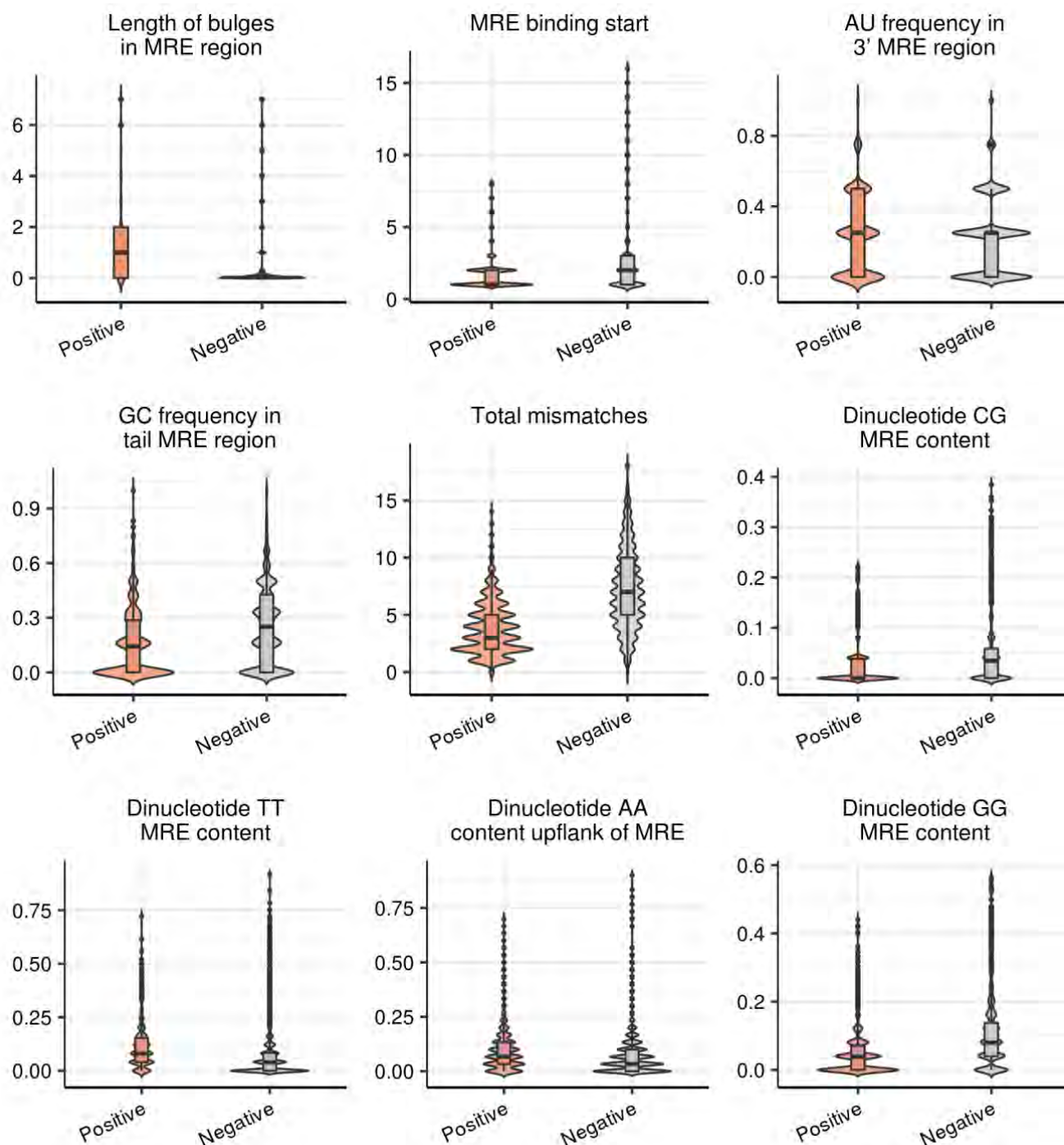


Figure 38: Distributions of MRE-related features corresponding to positive miRNA-target pairs against the relevant densities of negative binding sites. The descriptors present higher performance in microT-training set compared to microCLIP-training set. Evaluated descriptors include length of target bulges, start of the binding in the MRE region relative to miRNA binding anchors upon duplex formation, AU base pairs in 3' supplementary region, GC base pairs in tail MRE region, total mismatches per miRNA-target duplex and dinucleotide MRE content. Assessed characteristics of positive miRNA interactions significantly diverge from respective feature distributions of negative MREs (two-tailed Wilcoxon rank-sum test).

3.3.2 microT Super Learning framework

microT identifies putative MREs within the 3' UTR and CDS regions. The model adopts the microCLIP classification scheme with several updates, by incorporating an enhanced training set and re-arrangements in features. Evaluation of the accuracy, in terms of sensitivity and specificity, of the 9 base nodes and the meta-learner has been performed on a separate independent test set of 6,192 instances. Base nodes seem to achieve a better prediction accuracy compared to microCLIP (Methods 2.2.3). All the classifiers achieved high performance in a range of sensitivity 78.6% - 90.9% and specificity 77.4% - 91.9% (range of AUC: 83.7% - 97.6%). Their aggregated outcome in the meta-learner of microT exhibits the highest performance in terms of sensitivity and specificity (sensitivity: 95.0, specificity: 93.3, AUC: 98.7%). Also the relative distributions of base model scores, estimated on positive/negative instances of the test set, demonstrate greater in-between disrelations compared to microCLIP relevant evaluation (Figure 39).

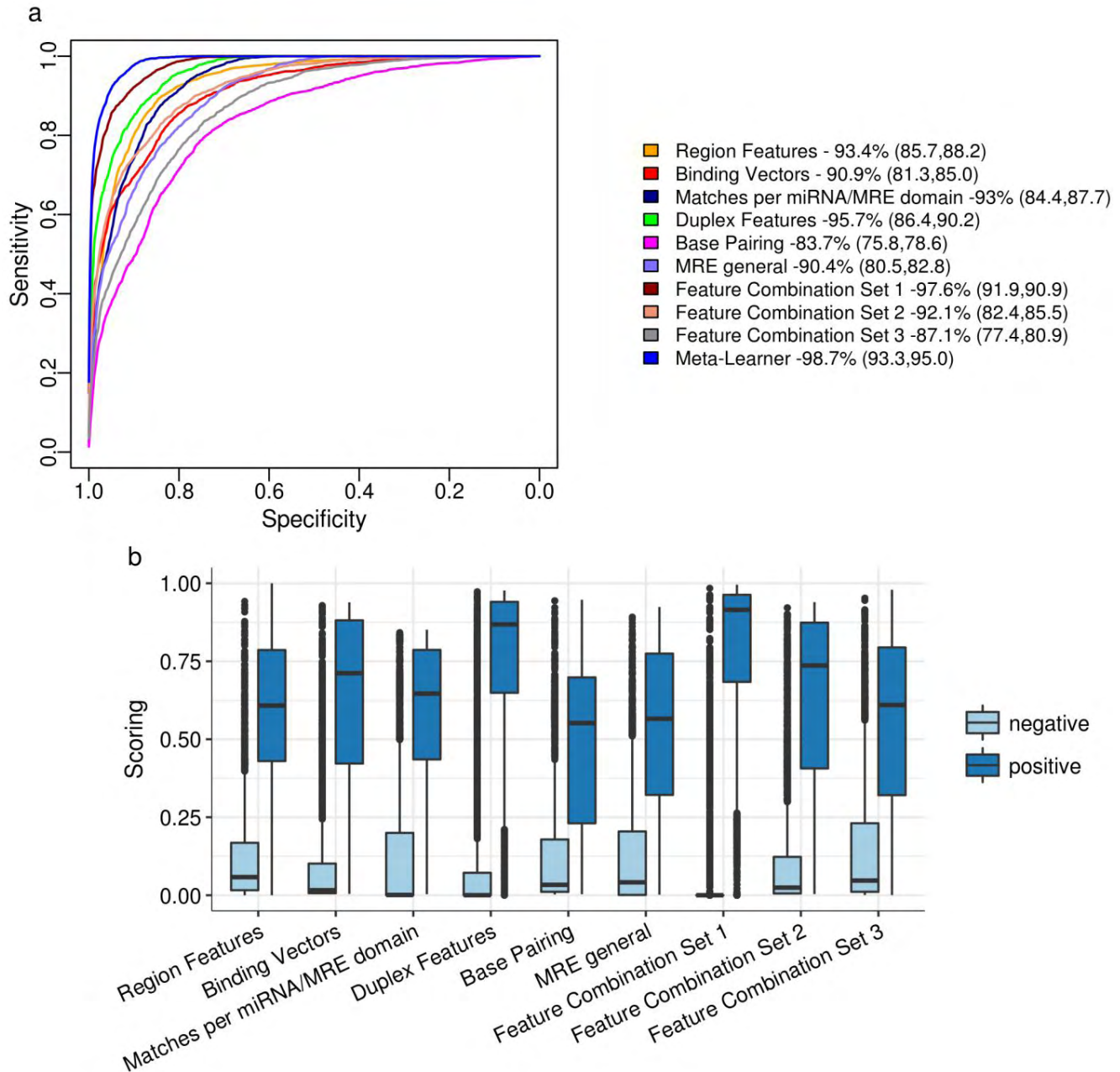


Figure 39: Evaluation of the accuracy of the 9 base model classifiers. Five-fold cross-validation has been implemented on a separate set of approximately 6,192 instances to test the performance of each node. a) ROC curve of each base model displays the classification of positive/negative miRNA binding sites. b) Distribution of base model scores estimated on positive/negative instances of the test set.

To further estimate the predictive accuracy of the multi-layer super learner classification scheme and to validate the proper partition of the features into the base nodes, 4 different super learning models have been deployed. The first one (microT SL - 6 base classifiers) combines the same classifiers with microT but without the three supplementary nodes (Feature Combination 1, 2, 3). The other three models combine Deep Learning (DL) and/or

Random Forest (RF) classifiers in the 1st layer, either with the whole set of nodes, or by eliminating the one with the weakest performance (Base Pairing Classifier, microT SL – 8 Base Classifiers – DL, RF). The performance of the models was evaluated against an independent test set of 2,092 positive chimeric and reporter assay-verified miRNA binding events, corresponding to 2,032 miRNA-gene interactions. The number of correctly predicted miRNA targets for each classification approach is plotted versus the mean predictions per miRNA. A separate comparison captures the models' efficiency to predict correct miRNA binding sites in different levels of total predicted sites. The results indicate that even if all the approaches have similar efficiency to correctly determine *bona fide* miRNA binding sites, microT Super Learning framework demonstrates better sensitivity to correctly predict miRNA-gene interactions, i.e. lower false positive rate (Figure 40).

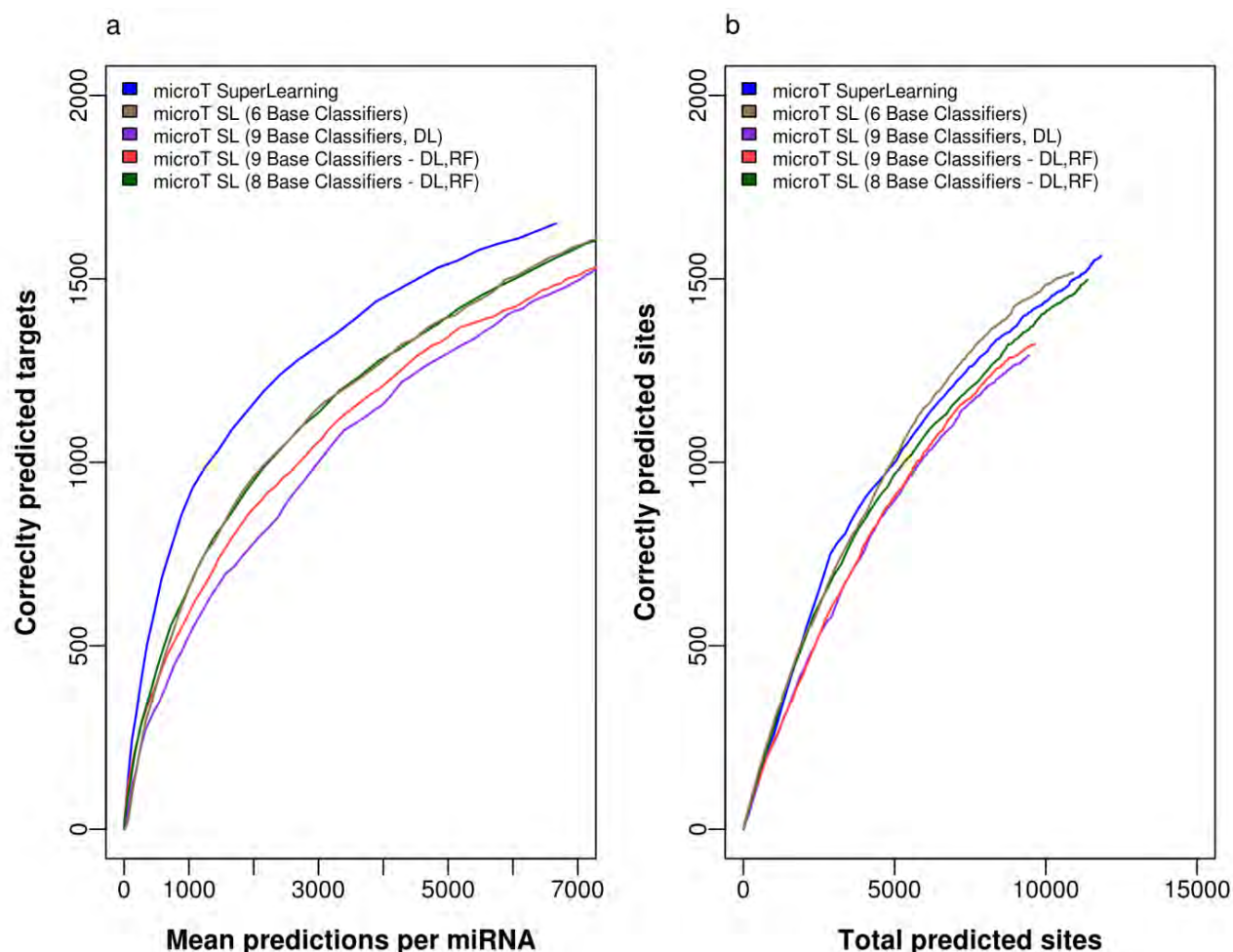


Figure 40: Evaluation of microT performance against 4 alternative Super Learning (SL) classification approaches: a model incorporating the same classifiers with microT and without 3 base nodes; a model consisting only Deep Learning classifiers (DL) in the 1st layer; a model combining Deep Learning and Random Forest (RF) classifiers in the 1st layer; a model combining Deep Learning and Random Forest

classifiers in the 1st layer and without Base Pairing node. The utilized set comprised 2,092 experimentally validated direct miRNA binding events (1,805 chimeric fragments and 287 reporter-assay verified), corresponding to 2,032 unique miRNA-gene interactions. (a) The number of correctly predicted miRNA-target interactions for each classification approach is plotted versus the mean prediction per miRNA. (b) A separate comparison captures the models' efficiency to predict correct miRNA binding events at different levels of total predicted sites.

To display the impact of features and classifiers under an optimal super learner design, different Deep Learning models incorporating all the features were deployed. Deep learning models are composed of different number of hidden layers and units, while input dropout of descriptors (ID) was allowed up to 20% percentage. The training of the models was executed according to the methodology described in Methods 2.2.3. Models with high predictive accuracy (AUC \geq 0.99, 10-fold cross-validation) were retained. The performance of the models was evaluated against the independent validation set described in Figure 40. The results indicate that the super learning classification scheme outperforms all the Deep Learning models, reinforcing the hypothesis that the contribution of features is maximized through their parallel use in different classification models and nodes (Figure 41).

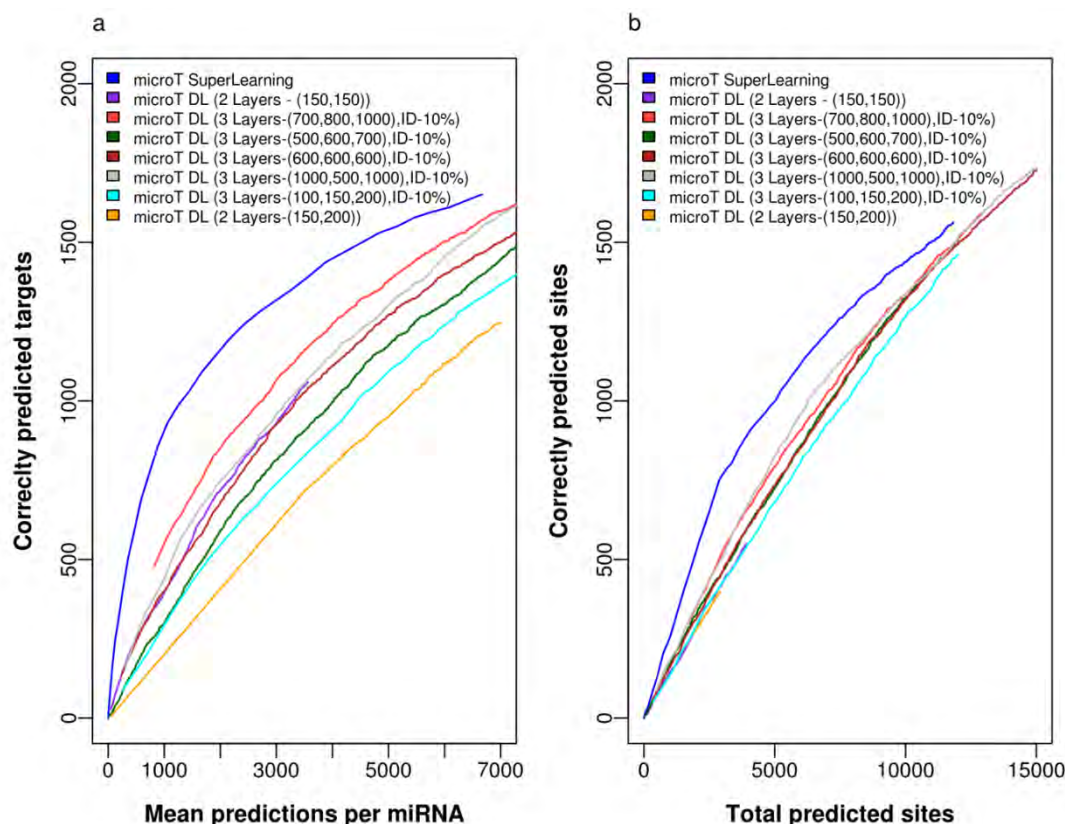


Figure 41: Evaluation of microT performance against 7 alternative Deep Learning models. The utilized set comprised 2,092 experimentally validated direct miRNA binding events (1,805 chimeric fragments and 287 reporter-assay verified), corresponding to 2,032 unique miRNA-gene interactions. (a) The number of correctly predicted miRNA-target interactions for each classification approach is plotted versus the mean

prediction per miRNA. (b) A separate comparison captures the models' efficiency to predict correct miRNA binding events at different levels of total predicted sites.

3.3.3 Evaluation of microT against other *in silico* models

To assess the predictive accuracy of microT, we compared the model against leading implementations in miRNA-target characterization, Targetscan v7 and microT-CDS. We utilized precompiled data from the microT-CDS site (www.microrna.gr/microT-CDS), while for Targetscan v7 we incorporated the unified set of predictions, described in section "Methods 2.2.4". In cases of multiple transcript-gene associations, the predictions of the models were filtered to retain interactions for the transcripts with the longest 3'UTR.

The performance of the models was initially tested using 5 profiling datasets following miRNA perturbation in different cell types. To estimate the generalization ability of microT, 3 of the 5 tested cell types were not included in the training process. We followed the methodology described in section "Results 3.2.8", where each miRNA-target pair was characterized by its associated miRNA binding site with the highest score and equivalent sets of top predicted targets were integrated to compare cumulative distributions of fold changes. microT detected targets yielded significant differences in expression changes compared to transcripts lacking any predicted binding site (Figure 42, range of P values (a-e): 4×10^{-38} - 4×10^{-18} , one-sided Kolmogorov-Smirnov test, $987 < \text{no-site} < 4,254$). Compared to the other two implementations, microT displayed the greatest site effectiveness in most cases (Figure 42, range of P values (a-d): 7.3×10^{-21} -0.04, one-sided Kolmogorov-Smirnov test, $171 < n < 889$). In Figure 42e, it performed similarly as microT-CDS and better than Targetscan (range of P values (e): 0.035 - 0.16, one-sided Kolmogorov-Smirnov test, $n = 270$).

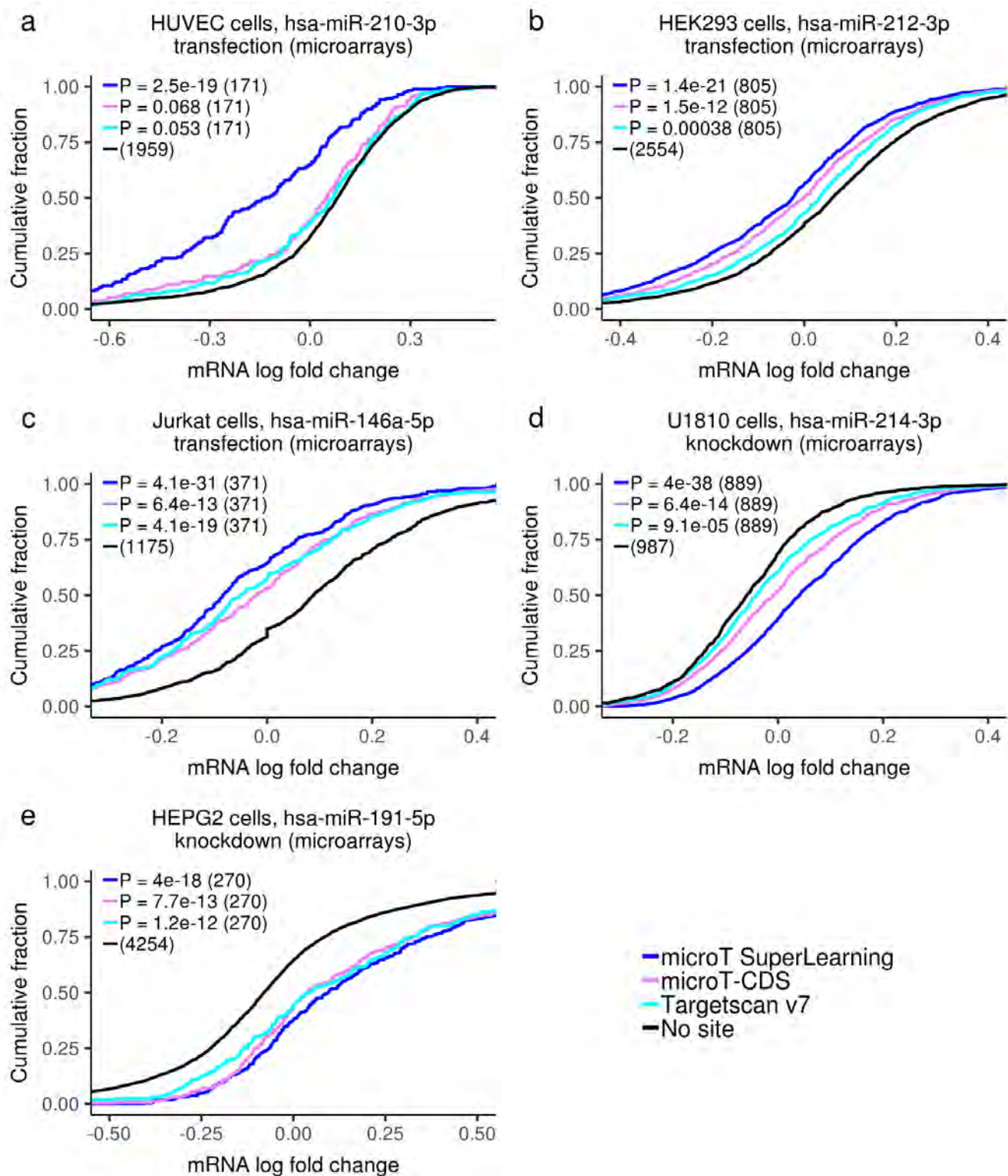


Figure 42: microT Super Learning performance compared to microT-CDS and Targetscan v7 was examined in 5 public gene expression profiling datasets following miRNA transfection or knockdown in different cell types. Cumulative distributions of mRNA fold changes for targets comprising at least one predicted MRE in the CDS or 3' UTR regions were compared to those that lacked any site of the considered miRNAs (one-sided

Kolmogorov-Smirnov test). Functional efficacy was assessed for equal numbers of top predictions per implementation. (a-d) Identified targets by microT revealed greater site effectiveness than the rest de novo approaches. (e) microT performed similarly as microT-CDS and better than Targetscan v7. The number of transcripts included in each comparison is denoted in the parentheses.

In silico de novo miRNA target prediction approaches were further evaluated for their efficiency to correctly determine *bona fide* miRNA binding sites/target pairs at a low number of total predictions. The utilized independent validation set was composed of 2,092 positive chimeric and reporter assay-verified miRNA binding events from 186 miRNAs, corresponding to 2,032 miRNA-gene interactions. The number of correctly predicted miRNA targets for each classification approach is plotted versus the mean predictions per miRNA. A separate comparison captures the models' efficiency to predict correct miRNA binding sites in different levels of total predicted sites. The results demonstrate that although all the methods perform well, microT has a significantly better performance (Figure 43). The new model detects 1.5-fold more experimentally validated miRNA binding events compared to the other approaches, verifying that the generalization of our AGO-CLIP-guided model to the whole transcript achieves equivalent high predictive accuracy.

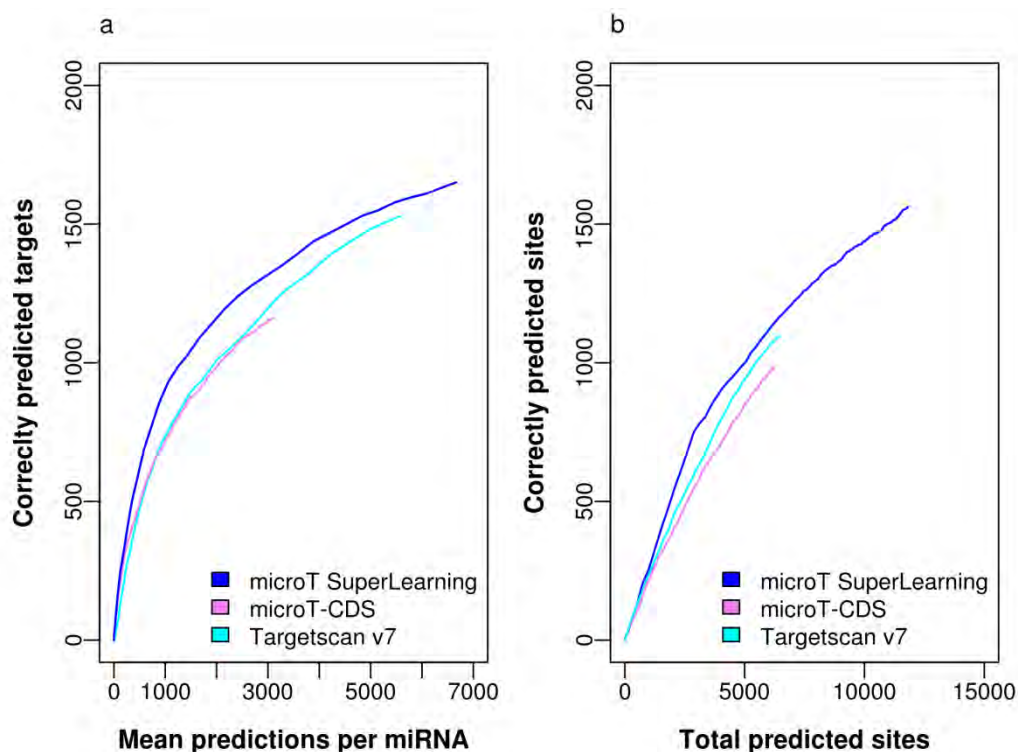


Figure 43: Evaluation of microT Super Learning model performance against microT-CDS and Targetscan v7. The utilized set comprised 2,092 experimentally validated direct miRNA binding events (1,805 chimeric fragments and 287 reporter-assay verified), corresponding to 2,032 unique miRNA-gene interactions. (a) The number of correctly predicted miRNA-target interactions for each classification approach is plotted versus the mean prediction per miRNA. (b) A separate comparison captures the models' efficiency to predict correct miRNA binding events at different levels of total predicted sites

CHAPTER 4

Conclusion

Accurate characterization of miRNA targets is considered fundamental to elucidate their regulatory roles. The identification of miRNA targets can be realized with either computational or experimental approaches.

During the last 15 years a multitude of experimental techniques have been emerged. High-throughput techniques have enabled the identification of novel experimentally-supported miRNA-gene interactions in a transcriptome-wide scale. However, the information of validated miRNA targets is dispersed in a great number of publications and raw datasets from high-throughput experiments.

During the course of this thesis, DIANA-Tarbase v8.0, the first new version since the 10th anniversary of the database inauguration, was developed. The repository indexes approximately one million entries, the largest compilation of miRNA-gene interactions compared to any relevant database. The new re-designed interface facilitates the extraction of miRNA interactions derived from more than 33 experimental methodologies, applied to ~600 distinct cell types/tissues under ~451 experimental conditions. The direct interconnection with DIANA-miRPath v3.0, simplifies the investigation of miRNA exerted regulation in physiological/pathological molecular pathways. DIANA-TarBase v8.0 is an important asset to the research community, empowering experimental investigations as well as *in silico* miRNA-related exploratory studies.

CLIP-Seq methodologies have revolutionized the study of protein-RNA interactions by enabling the accurate characterization of RBP target sites on a transcriptome-wide scale in different species under psychological or pathological conditions. PAR-CLIP variant against AGO proteins is considered among the most powerful high-throughput methods for the characterization of miRNA targets. During the past few years, computational methods devoted to AGO-PAR-CLIP data analysis have been elaborated by employing different mathematical models and feature sets. However, even the leading implementations present reduced ability to distinguish a large portion of genuine miRNA-targets.

In the current thesis, microCLIP framework was deployed, a cutting-edge algorithm for the identification of transcriptome-wide functional AGO-occupied clusters and associated miRNA-target pairs. microCLIP model circumvents pitfalls and limitations of existing implementations dedicated to PAR-CLIP data analysis, with the ability to be generalized to other CLIP-Seq variants. It is the first relevant implementation to employ the innovative super learner ensemble framework and the only available A-to-Z computational approach for the analysis of AGO-PAR-CLIP data initiating from aligned sequence reads (.sam/.bam files). Until now, miRNA-gene interactions derived from AGO-bound regions with inadequate T-to-C substitution rates were excluded from the target identification pipeline. By

implementing an extensive and thorough investigation, non-T-to-C clusters were shown to exhibit functional miRNA binding events and strong RNA accessibility.

microCLIP integrates these findings and provides a model that operates on every AGO-enriched cluster. The model detects interactions with the strongest functional efficacy and provides 1.6-fold more validated target sites when juxtaposed against leading implementations.

microCLIP deployment emboldened the development of a next generation de novo miRNA target prediction algorithm that will provide accurate miRNA targets and will guide miRNA-related studies with limited time and experimental cost. Currently, a multitude of computational approaches have been emerged aiming to accurately characterize miRNA targets. However, even the most sophisticated implementations still achieve a far from perfect predictive accuracy followed by an increased number of false positive predictions.

During this thesis, a novel miRNA target prediction algorithm is presented that overcomes limitations of current approaches. microT Super Learning framework maintains and upgrades the pipeline adopted in microCLIP by enhancing the training with even more high-throughput experiments under a tissue-specific scheme. The new model characterizes interactions with stronger functional efficacy and correctly detects 1.5-fold more experimentally validated target sites when juxtaposed against leading computational approaches.

The increased accuracy of microCLIP and microT frameworks in the multifaceted problem of miRNA-target identification can be attributed to the integration of meticulously curated high/low-throughput experimental datasets in an avant-garde super learner framework. The comprehensive construction of miRNA interactomes can guide downstream investigations towards the elucidation of unexplored regulatory mechanisms and key components in different biological processes.

CHAPTER 5

Thesis Publications

During this thesis, the candidate participated in 9 scientific studies, involving computational approaches for determining the activity of the non-coding transcripts and in two of them the candidate is first author. The candidate's main research activity and contribution in the publications incorporates the implementation of algorithms and automated pipelines for the analysis of Next Generation Sequencing data (small-RNA-Seq, RNA-Seq, CLIP-Seq), data integration for the elucidation of non-coding RNA function and their involvement in mechanisms of post-transcriptional gene regulation.

The studies are published in international journals of high impact factor and a total of 942 citations have been received so far, according to Google Scholar. The publications are separated and presented below according to their related research field.

miRNA target prediction

1. Paraskevopoulou MD* and **Karagkouni D***, Vlachos IS, Tastsoglou S, Hatzigeorgiou AG, **microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions**, Nature Communications, 2018 IF: 12,353 (*joint first authorship)

Databases of experimentally supported microRNA (non-)coding targets

2. **Karagkouni D***, Paraskevopoulou MD*, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T, Dalamagas T, Hatzigeorgiou AG, **DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions**, Nucleic Acids Res. 2017 IF: 11.561 (*joint first authorship)
3. BA Sweeney, AI Petrov,..., **D. Karagkouni, et al., RNAcentral: a hub of information for non-coding RNA sequences**, Nucleic Acids Res. 2018 IF: 11.561

RNAcentral (<https://rnacentral.org/>) is a comprehensive database of non-coding transcripts that incorporates information of all ncRNA types from a broad range of organisms. RNAcentral Consortium collaborates a group of 44 Expert Database, while 31 of them have been totally imported. DIANA-TarBase v8 and DIANA-LncBase v2 have been included in the latest version of RNAcentral repository, providing miRNA-mRNA and miRNA-lncRNA interactions on separate intuitive report pages that can be easily queried by users. The candidate realized this integration by providing the

databases content in special formats and also participated in the design of user interface.

4. Paraskevopoulou MD, Vlachos IS*, **Karagkouni D***, Georgakilas G, Kanellos I, Vergoulis T, Zagganas K, Tsanakas P, Floros E, Dalamagas T, Hatzigeorgiou AG, **DIANA-LncBase v2: Indexing microRNA targets on non-coding transcripts**, Nucleic Acids Res. 2015 IF: 11.561 (*joint second authorship)

DIANA-LncBase (www.microrna.gr/LncBase) is a reference repository, dedicated to the cataloguing of miRNA targets on long non-coding transcripts. The latest version incorporates more than 70,000 experimentally supported interactions in human and mouse species, derived from 13 distinct low/high - throughput techniques, accompanied with extensive meta-data. miRNA:lncRNA experimentally supported interactions were extracted from manually curated publications and the analysis of 153 AGO-CLIP-Seq libraries. LncBase v2 also hosts ~1 million of *in silico* predicted miRNA targets on lncRNAs. The candidate collected/combined lncRNA transcripts from different repositories and was entrusted with the annotation of the miRNA binding events into the reference transcriptome. She also participated in the analysis of the AGO-CLIP-Seq datasets and in the statistical investigation concerning the evolutionary conservation of miRNA-lncRNA binding events. The manual curation process and the import of the experimentally supported interactions into the repository were also part of the candidate's responsibilities.

5. Vlachos IS, Paraskevopoulou MD, **Karagkouni D**, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, Dalamagas T, Hatzigeorgiou AG, **DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions**, Nucleic Acids Res. 2015 IF: 11.561

DIANA-TarBase v7.0 (www.microrna.gr/tarbasev7) is the first relevant database with hundreds of thousands of high-quality experimentally supported miRNA-gene interactions, extracted from the manual curation of hundreds of publications and the analysis of raw AGO-CLIP-Seq libraries. The interactions are enhanced with detailed meta-data and tissue/cell type specific information. The candidate was entrusted with the manual curation of numerous publications, the identification of expressed miRNAs in numerous cell types/tissues, as well as the data preparation.

Elucidating the combinatorial effect of microRNAs on molecular pathways

6. Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, **Karagkouni D**, Vergoulis T, Dalamagas T, Hatzigeorgiou AG, **DIANA-miRPath v3.0: Deciphering microRNA function with experimental support**, Nucleic Acids Res. 2015 IF: 11.561

DIANA-mirPath v3.0 (<http://www.microrna.gr/miRPathv3>) is an on-line software suite, dedicated to the assesement of the combinatorial effect of multiple miRNAs on molecular pathways. The functional annotation of miRNAs is determined by using hypergeometric and unbiased empirical distributions, accompanied with meta-analysis statistics. DIANA-mirPath supports KEGG molecular pathway analysis and Gene Ontology terms in seven species. The suite also incorporates experimentally supported and *in silico* predicted miRNA targets. The candidate participated in the deployment of a modified version of the unbiased empirical distributions algorithm and the incorporation of experimenatlly supported miRNA targets into the database.

TF:miRNA:mRNA:TF networks

7. Vlachos IS, Vergoulis T, Paraskevopoulou MD, Lykokanellos F, Georgakilas G, Georgiou P, Chatzopoulos S, **Karagkouni D**, Christodoulou F, Dalamagas T, Hatzigeorgiou AG, **DIANA-mirExTra v2.0: Uncovering microRNAs and transcription factors with crucial roles in NGS expression data**. Nucleic Acids Res. 2016 IF: 11.561

DIANA-mirExTra v2.0 (<http://www.microrna.gr/mirextrav2>) is an online software suite, dedicated to uncover TF:miRNA:mRNA:TF networks. The suite supports A-to-Z functional analysis, initiating from NGS expression data to identify important regulators with crucial roles in the processed libraries. It enables state-of-the-art investigation of miRNAs controlling mRNAs and TFs controlling (activating, repressing or regulating) mRNA or miRNA expression. The candidate participated in (a) the design of user interface, (b) the data preparation and (c) the integration of experimentally supported miRNA targets into the database.

Annotation of microRNAs

8. Papanicolaou A,..., **Karagkouni D**, et al., **The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species**, Genome biology, 2016 IF: 13.214

The candidate participated in the genetic analysis of a major destructive insect pest, the Mediterranean fruit fly, *Ceratitis capitata*. She was entrusted with a part of the annotation of previously uncharacterized miRNAs in this species. The annotation was based on (a) the analysis of (s)RNA-Seq datasets, (b) the homology of miRNAs in relative species, (c) the appropriate adjustment of publicly available algorithms for the characterization of pre-miRNAs and their hairpin structure.

Book Chapters

9. Vlachos IS, Georgakilas G, Tastsoglou S, Paraskevopoulou MD, **Karagkouni D**, Hatzigeorgiou AG, **Computational challenges and -omics approaches for the identification of miRNAs and targets** *Essentials of microRNAs in neurogenesis* Academic Press (Elsevier), 2017

ABBREVIATIONS - ACRONYMS

| | |
|--------------------|---|
| 3' UTR | 3' UnTranslated Region |
| 3Life | Luminescent Identification of Functional Elements in 3'UTRs |
| 4SU | 4-thiouridine |
| 5' UTR | 5' UnTranslated Region |
| 6SG | 6-thioguanosine |
| AGO | Argonaute |
| AGO-IP | AGO Immunoprecipitation |
| AUC | Area Under Curve |
| Biotin-Microarrays | Biotin miRNA tagging combined with microarrays |
| Biotin-qPCR | Biotin miRNA tagging combined with qPCR |
| Biotin-Seq | Biotin miRNA tagging combined with sequencing |
| BLAST | Basic Local Alignment Search Tool |
| C. elegans | Caenorhabditis elegans |
| CDF | Cumulative distribution Function |
| CDS | Coding Sequence |
| CLASH | Crosslinking, ligation, and sequencing of hybrids |
| CLEAR-CLIP | Covalent ligation of endogenous Argonaute-bound RNAs |
| CLIP-Seq | Cross-linking immunoprecipitation sequencing |
| DDBJ | DNA Data Bank of Japan |
| dG | Free energy |
| dH | Enthalpy |
| DL | Deep Learning |
| DNA | Deoxyribonucleic Acid |
| dS | Entropy |
| EBV | Epstein-Barr virus |
| ELISA | Enzyme-linked immunosorbent assay |
| EM | Expectation Maximization |
| ENCODE | Encyclopedia of DNA Elements Consortium |
| FastQC | Fast Quality Control tool |

| | |
|------------|--|
| FDR | False Discovery Rate |
| GBMs | Gradient Boosting Machines |
| GEEs | Generalized estimating equations |
| GEO | Gene Expression Omnibus |
| GFP | Green Fluorescent Protein |
| GLMMs | Generalized linear mixed models |
| GLMs | Generalized Linear Models |
| HEK293 | 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 |
| ICA | Independent component analysis |
| iCLIP | Individual-nucleotide resolution UV crosslinking and immunoprecipitation |
| ID | Input-dropout |
| ID3 | Iterative Dichotomiser 3 |
| IGV | Integrative Genomics Viewer |
| 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 |
| lncRNAs | long non-coding RNAs |
| MCF7 | Human Mammary Gland Cancer Cells / Michigan Cancer Foundation-7 |
| miRISC | miRNA-induced silencing complex |
| miRNA | microRNA |
| miTRAP | miRNA trapping by RNA in vitro affinity purification |
| MNase | Micrococcal Nuclease |
| MREs | miRNA Recognition Elements |
| mRMR | Minimum-redundancy-maximum-relevance |
| mRNA | messenger RNA |
| 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 |
| PARS | Parallel analysis of RNA structure |
| PCA | Principal component analysis |
| PDB | Protein Data Bank |
| PHP | Hypertext Preprocessor |
| piRNA | Piwi-interacting RNA |
| 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 |
| RBP | RNA-binding proteins |
| RF | Random Forest |
| RIP-Seq | RNA immunoprecipitation combined with sequencing |
| 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 |
| SDS-PAGE | Sodium dodecyl sulfate polyacrylamide gel electrophoresis |
| SILAC | Stable isotope labeling by amino acids in cell culture |
| siRNA | Short interfering RNA |

| | |
|----------|--|
| SL | Super Learning |
| sncRNA | Small non-coding RNA |
| SNR | Signal-to-noise ratios |
| SRA | Sequence Read Archive |
| sRNA | Small RNA |
| sRNA-Seq | Small RNA sequencing |
| SVC | Support-vector clustering |
| SVMs | Support Vector Machines |
| TCGA | The Cancer Genome Atlas |
| TF | Transcription Factor |
| Tm | Melting temperature |
| tRNA | Transfer RNA |
| TZMBL | Human Cervical Cancer Cells generated from JC.53 cells |
| UV | Ultraviolet |

REFERENCES

1. Cech, T.R. and J.A. Steitz, *The noncoding RNA revolution-trashing old rules to forge new ones*. Cell, 2014. **157**(1): p. 77-94.
2. Eddy, S.R., *Noncoding RNA genes*. Current opinion in genetics & development, 1999. **9**(6): p. 695-9.
3. Yue, F., et al., *A comparative encyclopedia of DNA elements in the mouse genome*. Nature, 2014. **515**(7527): p. 355-64.
4. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.
5. Vlachos, I.S. and A.G. Hatzigeorgiou, *Online resources for miRNA analysis*. Clin Biochem, 2013. **46**(10-11): p. 879-900.
6. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. Cell, 1993. **75**(5): p. 843-854.
7. Reinhart, B.J., et al., *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans*. Nature, 2000. **403**(6772): p. 901-6.
8. Pasquinelli, A.E., et al., *Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA*. Nature, 2000. **408**(6808): p. 86-9.
9. Hutvagner, G., et al., *A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA*. Science, 2001. **293**(5531): p. 834-8.
10. Catalanotto, C., C. Cogoni, and G. Zardo, *MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions*. Int J Mol Sci, 2016. **17**(10).
11. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*. Cell, 2010. **141**(1): p. 129-41.
12. Dai, R. and S.A. Ahmed, *MicroRNA, a new paradigm for understanding immunoregulation, inflammation, and autoimmune diseases*. Translational research : the journal of laboratory and clinical medicine, 2011. **157**(4): p. 163-79.
13. Wang, H. and D.Q. Peng, *New insights into the mechanism of low high-density lipoprotein cholesterol in obesity*. Lipids in health and disease, 2011. **10**: p. 176.
14. Erson, A.E. and E.M. Petty, *MicroRNAs in development and disease*. Clinical genetics, 2008. **74**(4): p. 296-306.
15. Guay, C., et al., *Diabetes mellitus, a microRNA-related disease?* Translational research : the journal of laboratory and clinical medicine, 2011. **157**(4): p. 253-64.
16. Ono, K., Y. Kuwabara, and J. Han, *MicroRNAs and cardiovascular diseases*. The FEBS journal, 2011. **278**(10): p. 1619-33.
17. Paraskevopoulou, M.D., et al., *microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions*. Nat Commun, 2018. **9**(1): p. 3601.
18. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies*. Nature Reviews Genetics, 2016. **17**(6): p. 333-351.
19. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets*. Cell, 2005. **120**(1): p. 15-20.
20. Agarwal, V., et al., *Predicting effective microRNA target sites in mammalian mRNAs*. eLife, 2015. **4**.

21. Garcia, D.M., et al., *Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs*. *Nature structural & molecular biology*, 2011. **18**(10): p. 1139-46.
22. Paraskevopoulou, M.D., et al., *DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows*. *Nucleic Acids Res*, 2013. **41**(Web Server issue): p. W169-73.
23. Reczko, M., et al., *Functional microRNA targets in protein coding sequences*. *Bioinformatics*, 2012. **28**(6): p. 771-6.
24. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps*. *Nature*, 2009. **460**(7254): p. 479-86.
25. Pla, A., X. Zhong, and S. Rayner, *miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts*. *PLoS Comput Biol*, 2018. **14**(7): p. e1006185.
26. Vlachos, I.S., et al., *DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions*. *Nucleic Acids Res*, 2015. **43**(Database issue): p. D153-9.
27. Chou, C.H., et al., *miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database*. *Nucleic acids research*, 2016. **44**(D1): p. D239-47.
28. Grosswendt, S., et al., *Unambiguous identification of miRNA:target site interactions by different types of ligation reactions*. *Molecular cell*, 2014. **54**(6): p. 1042-54.
29. Helwak, A., et al., *Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding*. *Cell*, 2013. **153**(3): p. 654-65.
30. Garcia, D.M., et al., *Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs*. *Nat Struct Mol Biol*, 2011. **18**(10): p. 1139-46.
31. Wen, M., et al., *DeepMirTar: a deep-learning approach for predicting human miRNA targets*. *Bioinformatics*, 2018. **34**(22): p. 3781-3787.
32. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions*. *Nucleic acids research*, 2008. **37**(suppl_1): p. D105-D110.
33. Lu, Y. and C.S. Leslie, *Learning to Predict miRNA-mRNA Interactions from AGO CLIP Sequencing and CLASH Data*. *PLoS Comput Biol*, 2016. **12**(7): p. e1005026.
34. Kishore, S., et al., *A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins*. *Nature methods*, 2011. **8**(7): p. 559-64.
35. Moore, M.J., et al., *miRNA-target chimeras reveal miRNA 3 [prime]-end pairing as a major determinant of Argonaute target specificity*. *Nature communications*, 2015. **6**.
36. Gumienny, R. and M. Zavolan, *Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G*. *Nucleic acids research*, 2015. **43**(18): p. 9095.
37. Khorshid, M., et al., *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets*. *Nat Methods*, 2013. **10**(3): p. 253-5.
38. Betel, D., et al., *Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites*. *Genome biology*, 2010. **11**(8): p. R90.
39. John, B., et al., *Human MicroRNA targets*. *PLoS biology*, 2004. **2**(11): p. e363.
40. Menor, M., et al., *mirMark: a site-level and UTR-level classifier for miRNA target prediction*. *Genome biology*, 2014. **15**(10): p. 500.

41. Bandyopadhyay, S., et al., *MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets*. Scientific reports, 2015. **5**: p. 8004.
42. Vergoulis, T., et al., *TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support*. Nucleic Acids Research, 2012. **40**(Database issue): p. D222-9.
43. Li, J.-H., et al., *starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data*. Nucleic Acids Research, 2014. **42**(D1): p. D92-D97.
44. Sulc, M., et al., *PACCMIT/PACCMIT-CDS: identifying microRNA targets in 3' UTRs and coding sequences*. Nucleic Acids Res, 2015. **43**(W1): p. W474-9.
45. Baek, D., et al., *The impact of microRNAs on protein output*. Nature, 2008. **455**(7209): p. 64-71.
46. Selbach, M., et al., *Widespread changes in protein synthesis induced by microRNAs*. Nature, 2008. **455**(7209): p. 58-63.
47. Thomson, D.W., C.P. Bracken, and G.J. Goodall, *Experimental strategies for microRNA target identification*. Nucleic acids research, 2011. **39**(16): p. 6845-6853.
48. Cloonan, N., et al., *Stem cell transcriptome profiling via massive-scale mRNA sequencing*. Nature methods, 2008. **5**(7): p. 613-619.
49. Eichhorn, S.W., et al., *mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues*. Molecular cell, 2014. **56**(1): p. 104-115.
50. Vlachos, I.S., et al., *DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions*. Nucleic acids research, 2015. **43**(Database issue): p. D153-9.
51. Helwak, A., et al., *Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding*. Cell, 2013. **153**(3): p. 654-665.
52. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain*. Science, 2003. **302**(5648): p. 1212-5.
53. Skalsky, R.L., et al., *The viral and cellular microRNA targetome in lymphoblastoid cell lines*. PLoS pathogens, 2012. **8**(1): p. e1002484.
54. Farazi, T.A., et al., *Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets*. Genome biology, 2014. **15**(1): p. R9.
55. Khorshid, M., et al., *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets*. Nature methods, 2013. **10**(3): p. 253-5.
56. Majoros, W.H., et al., *MicroRNA target site identification by integrating sequence and binding information*. Nature methods, 2013. **10**(7): p. 630-3.
57. Corcoran, D.L., et al., *PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data*. Genome biology, 2011. **12**(8): p. R79.
58. Erhard, F., et al., *PARma: identification of microRNA target sites in AGO-PAR-CLIP data*. Genome biology, 2013. **14**(7): p. R79.
59. Chou, C.H., et al., *miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions*. Nucleic Acids Res, 2018. **46**(D1): p. D296-D302.
60. Weinstein, J.N., et al., *The cancer genome atlas pan-cancer analysis project*. Nature genetics, 2013. **45**(10): p. 1113-1120.

61. Jiang, Q., et al., *miR2Disease: a manually curated database for microRNA deregulation in human disease*. Nucleic acids research, 2008. **37**(suppl_1): p. D98-D104.
62. Li, J.-H., et al., *starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data*. Nucleic acids research, 2013. **42**(D1): p. D92-D97.
63. Khorshid, M., C. Rodak, and M. Zavolan, *CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins*. Nucleic acids research, 2010. **39**(suppl_1): p. D245-D252.
64. Karagkouni, D., et al., *DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions*. Nucleic acids research, 2017. **46**(D1): p. D239-D245.
65. Bishop, C.M., *Pattern recognition and machine learning*, 2006. 대한토목학회지, 2012. **60**(1): p. 78-78.
66. Trevor, H., T. Robert, and F. JH, *The elements of statistical learning: data mining, inference, and prediction*, 2009, New York, NY: Springer.
67. Jolliffe, I., *Principal component analysis*, in *International encyclopedia of statistical science*. 2011, Springer. p. 1094-1096.
68. Abe, S., *Feature selection and extraction*, in *Support Vector Machines for Pattern Classification*. 2010, Springer. p. 331-341.
69. Hand, D.J. and K. Yu, *Idiot's Bayes – not so stupid after all?* International statistical review, 2001. **69**(3): p. 385-398.
70. Vapnik, V.N., *An overview of statistical learning theory*. IEEE transactions on neural networks, 1999. **10**(5): p. 988-999.
71. Quinlan, J.R., *Induction of decision trees*. Machine learning, 1986. **1**(1): p. 81-106.
72. Ho, T.K. *Random decision forests*. in *Document analysis and recognition, 1995., proceedings of the third international conference on*. 1995. IEEE.
73. Breiman, L., *Random forests*. Machine learning, 2001. **45**(1): p. 5-32.
74. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning*. nature, 2015. **521**(7553): p. 436.
75. Breiman, L., *Arcing the edge*, 1997, Technical Report 486, Statistics Department, University of California at
76. Mason, L., et al. *Boosting algorithms as gradient descent*. in *Advances in neural information processing systems*. 2000.
77. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets – update*. Nucleic acids research, 2012. **41**(D1): p. D991-D995.
78. Kodama, Y., M. Shumway, and R. Leinonen, *The Sequence Read Archive: explosive growth of sequencing data*. Nucleic acids research, 2011. **40**(D1): p. D54-D56.
79. R Development Core Team, *R: A language and environment for statistical computing.*, 2016, R Foundation for Statistical Computing: Vienna, Austria.
80. Gautier, L., et al., *affy--analysis of Affymetrix GeneChip data at the probe level*. Bioinformatics, 2004. **20**(3): p. 307-15.
81. Carvalho, B.S. and R.A. Irizarry, *A framework for oligonucleotide microarray preprocessing*. Bioinformatics, 2010. **26**(19): p. 2363-7.
82. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Res, 2015. **43**(7): p. e47.
83. Flicek, P., et al., *Ensembl 2012*. Nucleic acids research, 2011. **40**(D1): p. D84-D90.

84. Huber, W., et al., *Orchestrating high-throughput genomic analysis with Bioconductor*. Nat Methods, 2015. **12**(2): p. 115-21.
85. Paraskevopoulou, M.D., et al., *DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows*. Nucleic acids research, 2013. **41**(W1): p. W169-W173.
86. Robinson, J.T., et al., *Integrative genomics viewer*. Nature biotechnology, 2011. **29**(1): p. 24.
87. van der Laan Mark, J., C. Polley Eric, and E. Hubbard Alan, *Super Learner*, in *Statistical Applications in Genetics and Molecular Biology* 2007.
88. Moore, M.J., et al., *miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity*. Nature communications, 2015. **6**: p. 8864.
89. Karagkouni, D., et al., *DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions*. Nucleic Acids Res, 2017.
90. Friedersdorf, M.B. and J.D. Keene, *Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs*. Genome biology, 2014. **15**(1): p. R2.
91. Eichhorn, S.W., et al., *mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues*. Molecular cell, 2014. **56**(1): p. 104-15.
92. Nam, J.W., et al., *Global analyses of the effect of different cellular contexts on microRNA targeting*. Molecular cell, 2014. **53**(6): p. 1031-43.
93. Pellegrino, L., et al., *miR-23b regulates cytoskeletal remodeling, motility and metastasis by directly targeting multiple transcripts*. Nucleic acids research, 2013. **41**(10): p. 5400-12.
94. Zhang, C., et al., *Primate-specific miR-603 is implicated in the risk and pathogenesis of Alzheimer's disease*. Aging (Albany NY), 2016. **8**(2): p. 272-90.
95. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update*. Nucleic acids research, 2013. **41**(Database issue): p. D991-5.
96. Kodama, Y., M. Shumway, and R. Leinonen, *The Sequence Read Archive: explosive growth of sequencing data*. Nucleic acids research, 2012. **40**(Database issue): p. D54-6.
97. Davis, M.P., et al., *Kraken: a set of tools for quality control and analysis of high-throughput sequence data*. Methods, 2013. **63**(1): p. 41-9.
98. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. EMBnet. journal, 2011. **17**(1): p. pp. 10-12.
99. Wu, T.D. and S. Nacu, *Fast and SNP-tolerant detection of complex variants and splicing in short reads*. Bioinformatics, 2010. **26**(7): p. 873-881.
100. Friedländer, M.R., et al., *miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades*. Nucleic acids research, 2011. **40**(1): p. 37-52.
101. Cunningham, F., et al., *Ensembl 2015*. Nucleic acids research, 2014. **43**(D1): p. D662-D669.
102. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data*. Nucleic Acids Research, 2014. **42**(D1): p. D68-D73.
103. Wan, Y., et al., *Landscape and variation of RNA secondary structure across the human transcriptome*. Nature, 2014. **505**(7485): p. 706-9.

104. Morgulis, A., et al., *A fast and symmetric DUST implementation to mask low-complexity DNA sequences*. Journal of computational biology : a journal of computational molecular cell biology, 2006. **13**(5): p. 1028-40.
105. Schmieder, R. and R. Edwards, *Quality control and preprocessing of metagenomic datasets*. Bioinformatics, 2011. **27**(6): p. 863-4.
106. Lorenz, R., et al., *ViennaRNA Package 2.0*. Algorithms for molecular biology : AMB, 2011. **6**: p. 26.
107. Rosenbloom, K.R., et al., *The UCSC Genome Browser database: 2015 update*. Nucleic acids research, 2015. **43**(Database issue): p. D670-81.
108. Candel, A., et al., *Deep Learning with H2O*, 2015, H2O.
109. Kuhn, M., *Caret package*. Journal of Statistical Software, 2008. **28**(5).
110. Memczak, S., et al., *Circular RNAs are a large class of animal RNAs with regulatory potency*. Nature, 2013. **495**(7441): p. 333-8.
111. Hamilton, M.P., et al., *The Landscape of microRNA Targeting in Prostate Cancer Defined by AGO-PAR-CLIP*. Neoplasia, 2016. **18**(6): p. 356-70.
112. Lorenz, R., et al., *ViennaRNA Package 2.0*. Algorithms for Molecular Biology, 2011. **6**(1): p. 26.
113. Grosswendt, S., et al., *Unambiguous identification of miRNA: target site interactions by different types of ligation reactions*. Molecular cell, 2014. **54**(6): p. 1042-1054.
114. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data*. Nucleic acids research, 2013. **42**(D1): p. D68-D73.
115. The, R.C., *RNAcentral: a hub of information for non-coding RNA sequences*. Nucleic Acids Res, 2018.
116. Paraskevopoulou, M.D., et al., *DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts*. Nucleic acids research, 2016. **44**(D1): p. D231-D238.
117. Vlachos, I.S., et al., *DIANA-miRPath v3. 0: deciphering microRNA function with experimental support*. Nucleic acids research, 2015. **43**(W1): p. W460-W466.
118. Chou, C.-H., et al., *miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database*. Nucleic acids research, 2015. **44**(D1): p. D239-D247.
119. Marin, R.M., et al., *Analysis of the accessibility of CLIP bound sites reveals that nucleation of the miRNA:mRNA pairing occurs preferentially at the 3'-end of the seed match*. RNA, 2012. **18**(10): p. 1760-70.
120. Whisnant, A.W., et al., *In-depth analysis of the interaction of HIV-1 with cellular microRNA biogenesis and effector mechanisms*. mBio, 2013. **4**(2): p. e000193.
121. Hoell, J.I., et al., *RNA targets of wild-type and mutant FET family proteins*. Nat Struct Mol Biol, 2011. **18**(12): p. 1428-31.
122. Lipchina, I., et al., *Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response*. Genes Dev, 2011. **25**(20): p. 2173-86.
123. Chen, B., et al., *PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis*. Genome Biol, 2014. **15**(1): p. R18.
124. Sievers, C., et al., *Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data*. Nucleic Acids Res, 2012. **40**(20): p. e160.

125. Anders, S. and W. Huber, *Differential expression analysis for sequence count data*. Genome Biol, 2010. **11**(10): p. R106.