UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Integrated computational and experimental methods for the characterization of the frequency and functional consequences of RNA editing in microRNAs

PhD DISSERTATION

Tastsoglou Spyridon

Supervisor: Artemis G. Hatzigeorgiou, Bioinformatics Professor

Volos, 2020

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY)

Ευρωπαϊκή Ένωση
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

PhD THESIS COMMITTEE

Artemis Hatzigeorgiou, Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly (Head of Doctoral Advisory Committee)

Gerasimos Potamianos, Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly (Member of Doctoral Advisory Committee)

Konstantinos Mathiopoulos, Professor, Department of Biochemistry and Biotechnology, University of Thessaly (Member of Doctoral Advisory Committee)

Dimitra Dafou, Assistant Professor, Department of Biology, Aristotle University of Thessaloniki

Dimitrios Katsaros, Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Antonios Giannakakis, Assistant Professor, Department of Molecular Biology and Genetics, Democritus, University of Thrace

Aristotelis Hatziioannou, Researcher A′, National Hellenic Research Foundation

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Συνδυασμένη υπολογιστική και πειραματική μέθοδος για τη χαρτογράφηση της συχνότητας και των λειτουργικών συνεπειών του RNA editing στα μικρά RNAs

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ταστσόγλου Σπυρίδων

Επιβλέπουσα: Άρτεμις Γ. Χατζηγεωργίου, Καθηγήτρια Βιοπληροφορικής

Βόλος, 2020

Το έργο συγχρηματοδοτείται από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση», στο πλαίσιο της Πράξης «Ενίσχυση του ανθρώπινου ερευνητικού δυναμικού μέσω της υλοποίησης διδακτορικής έρευνας» (MIS-5000432), που υλοποιεί το Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ)

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

Επιχειρησιακό Πρόγραμμα
Ανάπτυξη Ανθρώπινου Δυναμικού,
Εκπαίδευση και Διά Βίου Μάθηση
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Άρτεμις Χατζηγεωργίου, Καθηγήτρια, Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας (Επιβλέπουσα Καθηγήτρια)

Γεράσιμος Ποταμιάνος, Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας (Μέλος Τριμελούς Συμβουλευτικής Επιτροπής)

Κωνσταντίνος Ματθιόπουλος, Καθηγητής, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας (Μέλος Τριμελούς Συμβουλευτικής Επιτροπής)

Δήμητρα Ντάφου, Επίκουρη Καθηγήτρια, Τμήμα Βιολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Δημήτριος Κατσαρός, Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολογών Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Αντώνιος Γιαννακάκης, Επικουρος Καθηγητής, Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Αριστοτέλης Χατζηιωάννου, Ερευνητής Α', Εθνικό Ιδρυμα Ερευνών

ΕΥΧΑΡΙΣΤΙΕΣ

16/11/2020

Κλείνοντας την παρούσα διατριβή, συνειδητοποίησα πόσο λίγα συνώνυμα έχει η λέξη «ευχαριστώ», τουλάχιστον στην ελληνική. Επιτρέψτε μου να αφιερώσω λίγο χώρο για να εκφράσω την ευγνωμοσύνη μου σε μια σειρά ανθρώπων που με επηρέασαν σημαντικά μέσα σε αυτό το διάστημα.

Ευχαριστώ ιδιαίτερα την Καθηγήτρια Άρτεμη Χατζηγεωργίου που πριν 5 χρόνια με υποδέχτηκε στο DIANA-Lab, ένα χώρο – βασικά έχουμε αλλάξει τουλάχιστον 5 φορές «χώρο» -, ένα οικοσύστημα που λογίζω σα δεύτερο σπίτι μου. Υπό την επίβλεψη της Άρτεμης εκτίμησα ιδιαίτερα το πόσες ευκαιρίες είχα να εκτεθώ σε πληθώρα βιοπληροφορικών προβλημάτων, αλλά και από νωρίς σε διαδικασίες της ελλαδικής και παγκόσμιας ακαδημαϊκής πραγματικότητας. Πάνω από όλα την ευχαριστώ που με καθοδήγησε και καθοδηγεί στην έρευνα, τη βιοπληροφορική και την ερευνητική συγγραφή, αποδεικνύοντας, έμπρακτα και διαρκώς, πως «ο τολμών νικά».

Ακολούθως θέλω να εκφράσω τις ιδιαίτερες ευχαριστίες μου στον Καθηγητή κο. Κωνσταντίνο Ματθιόπουλο, επίσης μέλος της Τριμελούς Συμβουλευτικής Επιτροπής και της Επταμελούς Εξεταστικής Επιτροπής, για τις χρήσιμες υποδείξεις και συμβουλές του πριν και κατά τη διάρκεια της εκπόνησης του διδακτορικού, καθώς και γιατί συνετέλεσε στην επαφή μου με την ερευνητική ομάδα DIANA-Lab. Τον Αναπληρωτή Καθηγητή κο. Γεράσιμο Ποταμιάνο επειδή συμφώνησε να αποτελέσει μέλος της Τριμελούς Συμβουλευτικής Επιτροπής και της Επταμελούς Εξεταστικής Επιτροπής, καθώς και για τη διαθεσιμότητά του για συζήτηση κατά την κάλυψη των εξαμηνιαίων υποχρεώσεών μου απέναντι στο Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ). Ευχαριστώ επίσης πολύ τον Επίκουρο Καθηγητή Αντώνη Γιαννακάκη για εκτενείς και πολύ εποικοδομητικές συζητήσεις που είχαμε στο Ελληνικό Ινστιτούτο Παστέρ, και για τη συμμετοχή του στην Επταμελή Εξεταστική Επιτροπή. Ευχαριστώ πολύ επίσης την Επίκουρη Καθηγήτρια κα. Δήμητρα Ντάφου, τον Αναπληρωτή Καθηγητή κο. Δημήτριο Κατσαρό, και τον Ερευνητή Α΄ κο. Αριστοτέλη Χατζηιωάννου για την τιμή που μου κάνουν με τη συμμετοχή τους στην Επταμελή Εξεταστική Επιτροπή. Να αναφέρω συμπληρωματικά πως τα προαναφερθέντα άτομα συνέβαλαν στη σύσταση της Επταμελούς Εξεταστικής Επιτροπής και στην πραγματοποίηση της υποστήριξης σε ένα στενό χρονικό πλαίσιο και μάλιστα υπό καθεστώς καραντίνας.

Θέλω να εκφράσω τη μεγάλη ευγνωμοσύνη μου σε πρότερα μέλη της ερευνητικής ομάδας που με υποδέχθηκαν θερμότατα, με συμπεριέλαβαν σε εργασίες τους, με κατηύθυναν και θεωρώ ότι έπαιξαν καθοριστικό ρόλο στη διαμόρφωση της ερευνητικής μου αντίληψης και δράσης, ιδιαίτερα δε στην Ερευνήτρια Β΄ Μαρία Παρασκευοπούλου, τον Επίκουρο Καθηγητή Ιωάννη Βλάχο και τη Δρα. Δήμητρα Καραγκούνη. Με τη Δήμητρα πορευτήκαμε κοινά για μεγάλο διάστημα, μοιραστήκαμε συνεργασίες και άγρυπνα βράδυα πάνω από αναλύσεις και φόρμες υποβολών, με έχει συμβουλέψει και μού έχει σταθεί κάθε στιγμή σε ακαδημαϊκό και προσωπικό επίπεδο, ακούραστα και με μεγάλη υπομονή, και την εκτιμώ αφάνταστα για αυτό. Αντίστοιχα νοιώθω για τον υποψήφιο διδάκτορα Γιώργο Σκούφο, ο οποίος με έχει συντροφέψει σε τεράστιο μέρος της διαδρομής αυτής, αναβάθμισε την αντίληψή μου για πτυχές της υπολογιστικής επιστήμης και πηγαία αποτέλεσε *πολλές* φορές πηγή έμπνευσης, αλλά και ψυχικής ηρεμίας.

Είμαστε πια στα νυν μέλη του DIANA-Lab. Οφείλω ακόμη πολλά στον υποψήφιο διδάκτορα Μάριο Μηλιώτη, τον οποίο ευχαριστώ για τη βοήθειά του στις αναλύσεις δευτεροταγούς δομής των τροποποιημένων μορφών των microRNAs, αλλά και συνολικά για την ερευνητική επιμονή, το μεράκι του, την υποστήριξή του και την πάντα καλή διάθεση. Ευχαριστώ το Δρ. Νικόλαο Βακιρλή για τη διαθεσιμότητα, τη βοήθεια και τις υποδείξεις του στις αναλύσεις συντήρησης των τροποποιημένων microRNAs. Ευχαριστώ ιδιαίτερα την υποψήφια διδάκτορα Άννα Καραβαγγέλη για τη διάθεσή της για συζητήσεις βιοπληροφορικές, βιολογικές και άλλες, την υποψήφια διδάκτορα Βασιλική Κώτσιρα που συνέβαλε να βγαίνω πιο συχνά έξω από το πρώτο και το δεύτερο σπίτι μου και να είμαι καλά και το Δρ. Ιωάννη Καβακιώτη για τη στήριξή του, τις συνεργασίες και πολλές χρήσιμες συμβουλές. Χαίρομαι ιδιαίτερα που βλέπω όλα τα μέλη του εργαστηρίου μου να συμβάλλουν στη διατήρηση ενός πολύ φιλικού και ζεστού κλίματος.

Θέλω επίσης να ευχαριστήσω την υποψήφια διδάκτορα Joanna Elzbieta Handzlik για τη συνεργασία μας κατά την ανάπτυξη και αξιολόγηση του εργαλείου Manatee και την πολύ καλή μας επαφή. Εκφράζω τις ευχαριστίες μου για τις όμορφες συνεργασίες και την εμπιστοσύνη που μου επέδειξαν αναθέτοντάς μου την ανάλυση δεδομένων των εργαστηρίων τους, η Ερευνήτρια Ευδοκία Καραγκούνη και ο Ερευνητής Διονύσιος Σγούρας, καθώς και η Δρ. Εβίτα Αθανασίου, η Δρ. Μαρία Αγάλλου και ο υποψήφιος διδάκτωρ Ελευθέριος Κοντιζάς, από το Ελληνικό Ινστιτούτο Παστέρ.

Για το κατόρθωμα να ανακτήσουμε τοπικά στο Πανεπιστήμιο Θεσσαλίας ευαίσθητα δεδομένα αλληλούχησης από την υποδομή TCGA, οφείλω ευχαριστίες στο Δρα. Ιωάννη Καβακιώτη από το DIANA-Lab, τον κο. Δημήτρη Μεσαλούρη, την κα. Μαρία Παπαδοπούλου και τον κο. Χρόνη Βελέντζα από την Επιτροπή Ερευνών του Πανεπιστημίου, την Αντιπρύτανη Έρευνας και Δια Βίου Εκπαίδευσης του Πανεπιστημίου κα. Ιωάννα Λαλιώτου, καθώς και τον Επίκουρο Καθηγητή κο. Αντώνη Γιακουντή.

Ευχαριστώ το ΙΚΥ που επέτρεψε να υλοποιήσω την παρούσα διατριβή υπό χρηματοδότηση με τη μορφή υποτροφίας, καθώς και το προσωπικό του και ήταν συνεργάσιμο και κατανοητικό σε ζητήματα που ανέκυπταν.

Μια υπενθύμιση προς όλες και όλους· ιδιαίτερες ευχαριστίες και στήριξη πάντα στην Alexandra Elbakyan, που χωρίς την εφαρμογή της οι περισσότερες σύγχρονες διατριβές και εργασίες παγκοσμίως θα ήταν πολύ φτωχότερες.

Κλείνοντας, αισθάνομαι τυχερός που έχω δίπλα μου ανθρώπους που με υπομένουν και με στηρίζουν χωρίς να ξέρουν ακριβώς γιατί και μοιράστηκαν/μοιράζονται μαζί μου ένα σπίτι για πολύ – Μαριά-Θανάσης και τώρα Γεωργία –, ένα σπίτι για λιγότερο – Λουκία, Αγγελική, Αλέξανδρος, Στεφανία, Χρήστος, Αθηνά –, μια βόλτα και πέντε λέξεις – Κατερίνα, Ηλίας, Χάρης, Κωνσταντίνος, Κλεομένης, Νάντια, Λένα και σίγουρα κάποιον/α ξεχνώ! Να νικάμε!

ABSTRACT

The discovery of microRNAs (miRNAs) in the early 1990s, gave birth to a reactive new biomedical field, that of non-coding RNA (ncRNA) research. The emergence of functional ncRNA interactions with coding RNA (messenger RNA, mRNA) as well as with other ncRNAs revealed the importance of cataloguing the RNA interactome. Biotechnological leaps enabled the development of numerous distinctive high-throughput sequencing protocols, which yield immense amounts of biological data in each run. The translation of biological Big Data into meaningful information depends entirely on state-of-the-art algorithms and *in silico* methodologies. Bioinformatics analysis constitutes now the main workhorse for the large-scale characterization and comparison of coding and non-coding RNA abundance and interactions, in physiological and pathological states alike. Large-scale integration of Next Generation Sequencing (NGS) datasets offers the potential to develop models that accurately depict or capture ncRNA biogenesis and functions, as well as modifications these molecules are naturally subjected to, effectively boosting basic biological research and translational endeavors such as the identification of therapeutic targets and diagnostic biomarkers.

This dissertation focuses on the most actively studied ncRNA class, miRNAs. They are short RNA molecules (~22 nucleotides long) which exert their post-transcriptional regulatory function by targeting miRNA Recognition Elements (MREs) on the sequence of coding and non-coding RNAs (e.g. long non-coding RNAs, lncRNAs). miRNA-MRE binding follows complex rules of perfect or imperfect RNA complementarity and results in translational suppression and/or degradation of the targeted transcript(s). The vast majority of biological processes in higher mammals are under miRNA regulation.

RNA editing is a type of post-transcriptional modification that is catalyzed by specialized enzymes and occurs in specific miRNAs under physiological conditions. Since miRNA targeting is highly sequence-specific, the targeting repertoire of a miRNA can be affected substantially due to editing. NGS experiments enable the exploration of RNA editing in all transcribed molecules, however it’s a challenging task; mismatches on the reference RNA sequence could also be exhibited due to the presence of DNA variation on the corresponding genomic locus, or they could be sequencing errors.

During the course of this PhD thesis, a miRNA editing identification algorithm was developed. This algorithm combines small RNA Sequencing (sRNA-Seq) experimental datasets with variant information retrieved from reference resources or from matching Whole Genome Sequencing (WGS) datasets from the same individual, tissue or state, in order to filter out putative false positive RNA editing calls. Our algorithm was applied in datasets derived from numerous healthy/cancerous tissues and cell-lines and the most prominent editing events were catalogued. A downstream analysis of the effects of RNA editing in the targeting repertoire of edited miRNAs was conducted, by deploying state-of-the-art miRNA target prediction methods. Evolutionary conservation and thermodynamic stability characteristics of the observed edited and unedited miRNA forms were also studied.

During my doctoral dissertation, I participated in 7 published studies in high impact-factor journals (i.e. one presentation of a small RNA quantification application, three studies related to miRNA targeting, the showcase of a database of experimentally supported microbe-disease associations and two transcriptomic analyses of the efficacy of novel vaccines against leishmaniasis), as well as in the co-authoring of a book chapter describing computational methods in miRNA research. I presented research findings in 7 scientific conferences (4 national and 3 international), and have contributed to the organization of the European Conference of Computational Biology (ECCB'18) and the 5th Postgraduate and Postdoctoral Researchers' Meeting of the Hellenic Pasteur Institute (2019). My total publications have been cited 356 times, according to Google Scholar.

On 7/12/2016, the State Scholarship Foundation (I.K.Y.) accepted to fund my proposal on the computational and experimental study of RNA editing in microRNAs, in the form of a 36-month scholarship.

SUBJECT AREA: Bioinformatics

KEYWORDS: microRNA, small RNA-Seq, RNA editing, A-to-I editing, Inosine modification

ΠΕΡΙΛΗΨΗ

Η ανακάλυψη των μικρών RNAs (microRNAs, miRNAs) τη δεκαετία του 1990 και άλλων κλάσεων μη-κωδικοποιών RNAs (non-coding RNAs, ncRNAs), δημιούργησε ένα πεδίο εντατικής έρευνας. Στο χώρο αυτό αναδείχτηκε η σημασία της μελέτης των αλληλεπιδράσεων των ncRNAs, τόσο μεταξύ τους όσο και με κωδικοποιά μετάγραφα (messenger RNAs, mRNAs). Η βιοπληροφορική ανάλυση αποτελεί κύριο μέσο για τη χαρτογράφηση και σύγκριση της αφθονίας των κωδικοποιών και μη-κωδικοποιών RNAs και των αλληλεπιδράσεων μεταξύ τους, τόσο σε παθολογικές όσο και σε φυσιολογικές καταστάσεις. Η πρόοδος της βιοτεχνολογίας επέτρεψε την ανάπτυξη πληθώρας πειραμάτων αλληλούχησης υψηλής διεκπεραιωτικής ικανότητας (high-throughput experiments), τα οποία αποφέρουν τεράστιο όγκο βιολογικών δεδομένων. Η ανάλυση και αξιοποίηση των δεδομένων αυτών βασίζεται εξ' ολοκλήρου σε αλγορίθμους και υπολογιστικές μεθοδολογίες αιχμής. Η αξιοποίηση και ενσωμάτωση (integration) συνόλων δεδομένων από πειράματα Αλληλούχησης Επόμενης Γενιάς (Next Generation Sequencing, NGS) επιτρέπει την ανάπτυξη μεθόδων/μοντέλων που αναπαριστούν ή καταγράφουν με πιστότητα τη βιογένεση, το μηχανισμό δράσης, καθώς και τις τροποποιήσεις των ncRNAs και επιτρέπουν τη μελέτη βιολογικών μηχανισμών, την ανάδειξη θεραπευτικών στόχων και διαγνωστικών βιοδεικτών.

Η παρούσα διατριβή επικεντρώνεται στην πιο εντατικά μελετούμενη κατηγορία ncRNAs, τα miRNAs. Πρόκειται για RNAs μήκους περίπου 22 νουκλεοτιδίων που δρουν ως ισχυροί μετα-μεταγραφικοί ρυθμιστές της γονιδιακής έκφρασης, στοχεύοντας κωδικοποιά (mRNA) και μακρά μη-κωδικοποιά (π.χ. long non-coding RNA, lncRNA) μετάγραφα σε μικρές ακολουθίες τους που αποκαλούνται Στοιχεία Αναγνώρισης από miRNAs (miRNA Recognition Elements, MREs). Κατά τη στόχευση, τα miRNAs προσδένονται στα MREs με βάση σύνθετους κανόνες τέλειας ή ατελούς συμπληρωματικότητας του RNA, επάγοντας την καταστολή της μετάφρασης και/ή την αποικοδόμησή του. Η πλειοψηφία των βιολογικών διεργασιών στα ανώτερα θηλαστικά ρυθμίζεται από τη δράση των miRNAs.

Η μετα-μεταγραφική τροποποίηση (RNA editing) των miRNAs από εξειδικευμένα ένζυμα αποτελεί μια φυσιολογική βιολογική διεργασία που μεταβάλλει το ρεπερτόριο στόχευσής τους. Η μεταβολή αυτή δύναται να είναι ήπια ή δραστική, ανάλογα με τη θέση στην οποία λαμβάνει χώρα η τροποποίηση. Τα πειράματα NGS επιτρέπουν την ταυτόχρονη μελέτη του RNA editing σε όλα τα μεταγραφθέντα RNAs, όμως η αναγνώριση συμβάντων τροποποίησης μέσα από αυτά αποτελεί πρόκληση· οι παρατηρούμενες αλλαγές στη γνωστή ακολουθία των RNAs ενδέχεται εναλλακτικά να οφείλονται στην παρουσία μεταλλαγών στο επίπεδο του DNA ή σε σφάλματα κατά την αλληλούχηση.

Κατά τη διατριβή αυτή σχεδιάστηκε αλγόριθμος ταυτοποίησης περιστατικών RNA editing στα miRNAs, ο οποίος συνδυάζει σύνολα δεδομένων Αλληλούχησης των Μικρών RNAs (small RNA Sequencing) με πληροφορίες για την παρουσία μεταλλαγών, προερχόμενες από Βάσεις Δεδομένων αναφοράς ή πειράματα Αλληλούχησης του Γονιδιώματος (Whole Genome Sequencing) του ίδιου ατόμου, ιστού, ή της ίδιας πειραματικής συνθήκης. Ο αλγόριθμος αξιοποιήθηκε για την ταυτοποίηση συμβάντων RNA τροποποίησης στην ακολουθία των miRNAs σε υγιείς/καρκινικούς ιστούς και κυτταρικές σειρές. Πέραν της καταγραφής των πιο επιφανών περιπτώσεων τροποποίησης, τα miRNAs που βρέθηκαν τροποποιημένα υποβλήθηκαν σε ανάλυση για να αναδειχθεί η μεταβολή στο ρεπερτόριο στόχευσής τους, χρησιμοποιώντας αλγορίθμους αιχμής για την εύρεση των στόχων τους. Επίσης, μελετήθηκαν συγκριτικά

χαρακτηριστικά της εξελικτικής συντήρησης και της θερμοδυναμικής ευστάθειας των ακολουθιών των miRNAs στα οποία παρατηρούνται φαινόμενα RNA τροποποίησης.

Κατά την εκπόνηση του διδακτορικού μου, συμμετείχα σε 7 δημοσιευμένες εργασίες σε επιστημονικά περιοδικά υψηλού κύρους (μία παρουσίαση εργαλείου ποσοτικοποίησης των μικρών RNAs, τρεις εργασίες σχετικές με τη μελέτη της στόχευσης mRNAs/lncRNAs από miRNAs, μία παρουσίαση Βάσης Δεδομένων με συσχετίσεις της αφθονίας μικροβίων με ασθένειες και δύο μεταγραφωματικές μελέτες δράσης εμβολίων ενάντια στη λεϊσμάνια), καθώς και στη συγγραφή ενός κεφαλαίου βιβλίου αναφορικά με τις υπολογιστικές μεθόδους μελέτης των miRNAs. Έχω παρουσιάσει ερευνητικά ευρήματα συνολικά σε 7 επιστημονικά συνέδρια-ημερίδες (4 εθνικά, 3 διεθνή), και έχω συμβάλει στη διοργάνωση του Πανευρωπαϊκού Συνεδρίου Βιοπληροφορικής ECCB’18 και της 5ης Ημερίδας Μεταπτυχιακών και Μεταδιδακτόρων του Ελληνικού Ινστιτούτου Παστέρ, το 2019. Σύμφωνα με το Google Scholar οι εργασίες στις οποίες μετέχω έχουν μέχρι στιγμής λάβει 356 αναφορές.

Στις 7/12/2016 έγινε δεκτή η αίτησή μου στο Ίδρυμα Κρατικών Υποτροφιών για χρηματοδότηση διατριβής με αντικείμενο την υπολογιστική και πειραματική μελέτη του RNA editing φαινομένου στα microRNAs, μέσα από μια υποτροφία διάρκειας 36 μηνών.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ: Μικρά RNAs, Πειράματα αλληλούχησης των μικρών RNA, Τροποποιήσεις του RNA, Τροποποιήσεις Α-σε-Ι, Τροποποίηση σε Ινοσίνη

To optimizers of things collective and to conscious stargazers

Table of Contents

List of Figures

List of Tables

BACKGROUND

CHAPTER 1 – The Place of RNA in Molecular Biology

1.1 Central Dogma of Molecular Biology and the RNA Intermedium

In his monumental 1957 lecture called “Protein Synthesis”, Francis Crick integrated existing knowledge of the time to make two profoundly bold hypotheses of speculatory nature; the “Sequence Hypothesis” and the “Central Dogma of Molecular Biology”[1]. The former states that **(a)** the specificity of a piece of deoxyribonucleic acid (DNA) is dictated by its sequence alone and **(b)** that *that* sequence constitutes a code from which the amino acid sequence for a corresponding protein can be derived. The latter is the conjecture that “once information is passed into protein it cannot get out again”. In other words, this means that information may flow from nucleic acid to nucleic acid, or from nucleic acid to protein, but it cannot flow from protein to protein, or from protein “back” to nucleic acid (**Figure 1**). By “information” F. Crick refers precisely to the specification of each protein’s amino acid sequence, the “sequentialization” of amino acids in the correct order into a chain.

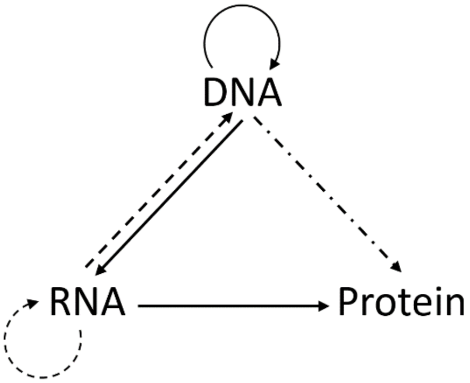


Figure 1. The Central Dogma of Molecular Biology. Information flows from DNA to DNA (replication), from DNA to RNA (transcription), from RNA to protein (translation). Special cases of transfer include RNA to RNA (viral replication), RNA to DNA (viral replication /reverse transcription) and DNA to protein (phenomenon only reported in cell-free experimental systems). Information cannot flow from protein to nucleic acids. Figure re- created from Crick FH (1970)[2] for the purpose of this thesis.

Flow (or “transfer”) between different molecule types is **(a)** the active transliteration (later on called *transcription*) of a DNA string of consecutive deoxyribonucleotides (i.e. adenine (A), thymine (T), guanine (G) and cytosine (C)) into a ribonucleic acid (RNA) string of ribonucleotides (i.e. adenine (A), uracil (U), guanine (G) and cytosine (C), according to the rules of complementary base pairing) and **(b)** the *translation* of consecutive non-overlapping triplets of nucleotides of an RNA string into a chain of amino acid residues, the protein sequence. Transfer between molecules of the same type is observed only for DNA (self-replication) and for the genomic RNA of some viruses.

At that time, protein was placed at the forefront of molecular biology, in the sense that cellular identity and distinct molecular functions inside cells were supposed to be realized and maintained almost exclusively through the production and enzymatic functions of distinct protein products. Under the simplified notion of information flow, DNA was considered as the carrier of genetically inherited pieces of information (*genes*), some of which could be transcribed into RNA molecules in the nucleus. RNA *transcripts* would then be exported to the cytoplasm and subjected to translation into protein on the ribosomes (which were called microsomal particles then).

Irrefutable experimental evidence supported the Sequence Hypothesis and the Central Dogma in the following years. The mechanisms behind transcription (DNA to RNA) and translation (RNA to protein) were identified in prokaryotic and eukaryotic organisms and their mechanisms of action and fine-tuning constitute active research fields until today.

Briefly, during transcription, the RNA polymerase and Transcription Factors bind to signal DNA sequences upstream of genes, called promoters. The DNA double helix unwinds locally and the hydrogen bonds between deoxyribonucleotides of the opposite DNA strands are broken to enable RNA polymerase to start adding ribonucleotides complementary to the one strand, creating a temporary DNA-RNA duplex (**Figure 2**). Upon encountering the transcriptional termination signal, RNA polymerase stops adding ribonucleotides, the hydrogen bonds of the created duplex are broken and the newly synthesized primary transcript is released[3].

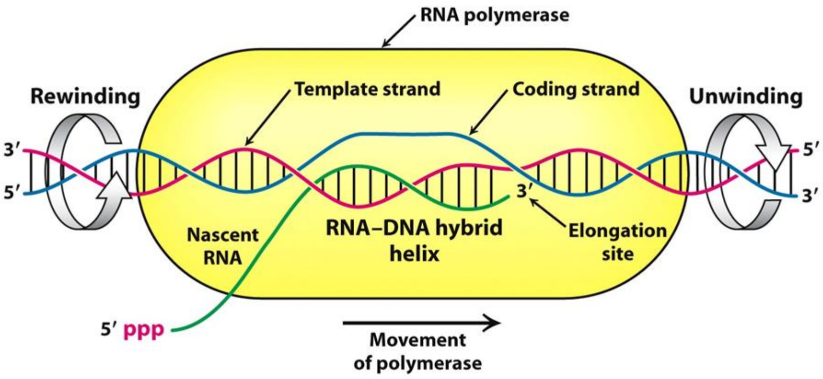


Figure 2. Eukaryotic transcription. Figure retrieved from Berg JM et al. (2010)[3].

In Eukaryotes, most nascent transcripts are subjected to maturation processes (**Figure 3**); they are protected by the non-standard linking of a guanine nucleotide at their 5′ (*five-prime*) end (5′-capping) and the addition of a stretch of adenines called poly(A)-tail at their 3′ (*three-prime*) end (polyadenylation). Most protein-coding transcripts (*messenger RNA, mRNA*) in Eukaryotes consist of expressed regions (*exons*) and intragenic regions (*introns*). In a process called splicing, introns are removed from transcripts prior to translation.

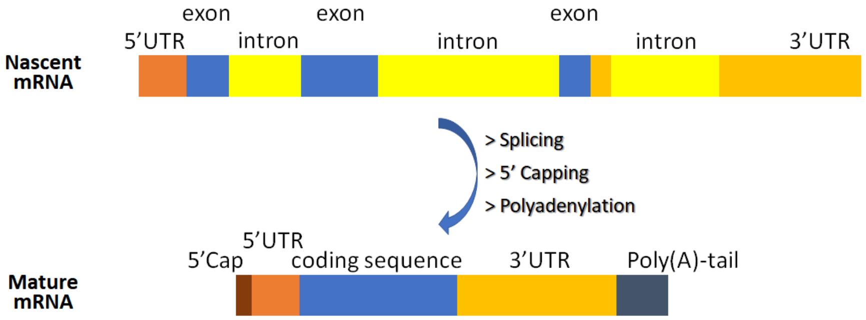


Figure 3. A eukaryotic messenger RNA in its nascent form and after undergoing maturation. Figure created for the purpose of this thesis.

Translation occurs on ribosomes, which are large multicomponent ribonucleoprotein complexes (i.e. consisting of both RNA and protein parts). The mature mRNA consists of the protected 5′ Untranslated Region (5′UTR), the Coding Sequence (CDS), the 3′ UTR and a poly(A)-tail. The ribosome "reads" the CDS serially in non-overlapping nucleotide triplets (codons) that correspond to one out of twenty amino acids or to termination signals (UGA, UAG, UAA), according to the genetic code. The first codon read during translation initiation is almost always an AUG codon, which corresponds to methionine amino acid. As shown in **Figure 4**, the target mRNA is bound on the ribosome and in a recursive step-wise fashion: **(a)** amino-acid-charged transfer RNAs (aminoacyl-tRNAs) that possess the correct anticodon sequence are incorporated in the complex, **(b)** a peptide bond occurs between the newly introduced amino acid and the forming protein (polypeptide) chain, **(c)** the discharged tRNA is released and **(d)** the ribosome moves one codon downstream on the mRNA sequence. When a termination signal is introduced, it cannot be recognized by any tRNA and the release process of the synthesized polypeptide sequence is induced[3].

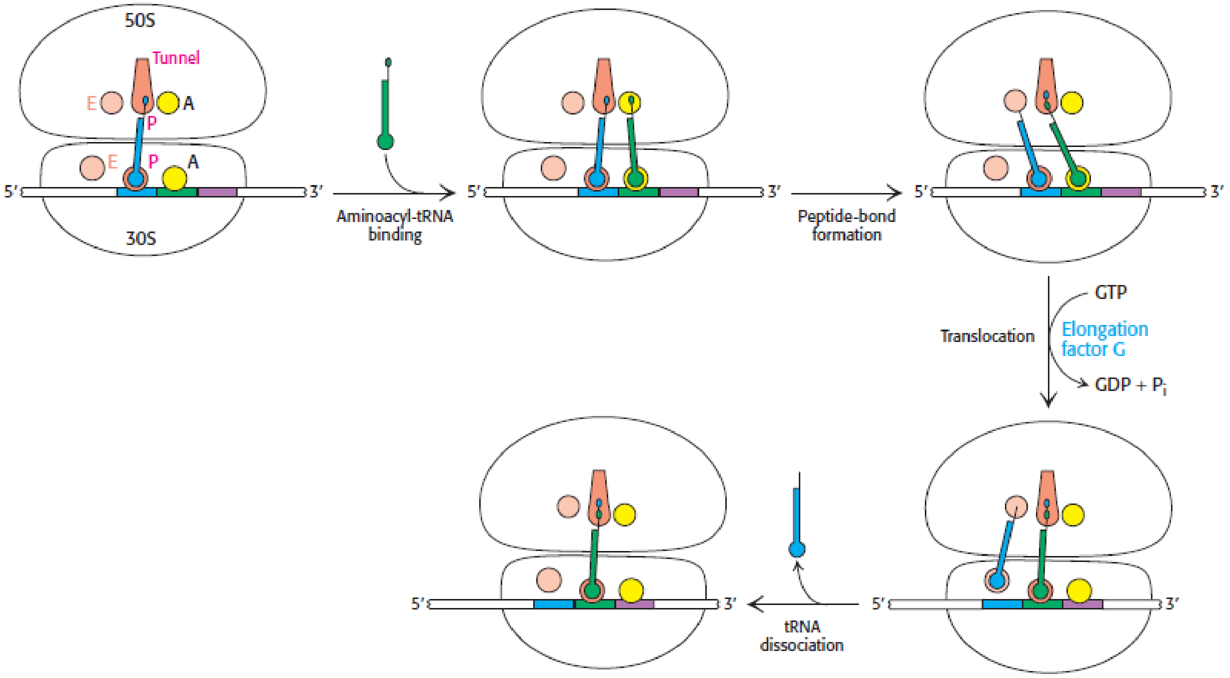


Figure 4. Translation. Figure retrieved from Berg JM et al. (2010)[3].

The Central Dogma is one of the most long-standing principles in Molecular Biology and lies at the heart of almost every scientific endeavor and application in modern biomedicine. What is astonishing is that, in an era of scarce knowledge and resources, F. Crick and his contemporaries managed to capture the multifaceted nature of what could now be called "the RNA Intermedium". Already from 1957, RNA was known to exist partly in the nucleus and partly in the cytoplasm, where it was found both in soluble form and bound to the microsomal particles (what we now call ribosomes)[1]. Notably, it was characterized as metabolically inhomogeneous, and as a molecular family that could contain more than one type, while the existence of, or ribosomal RNA (*rRNA*) as an exemplary other class of RNA beyond protein-coding RNA was only yet postulated.

Discoveries of the last 60 years are in line with these statements, deeming RNAs, among others:

- **a.** central message transmission intermediates of protein synthesis (mRNA)
- **b.** carriers of amino acid residues during translation (tRNA)
- **c.** active enzyme/ribozyme components of the ribosome (rRNA)
- **d.** post-transcriptional regulators of gene expression (microRNA, siRNA)
- **e.** whole-chromosomal silencers (lncRNA XIST)
- **f.** regulators of the degradation of other RNAs (lncRNA sponges)
- **g.** regulators of the modification of other RNAs (snoRNA, snRNA)
- **h.** effectors in nucleocytoplasmic transport (vtRNA)

1.2 Species of Non-Coding RNA

Apart from the widely studied protein-coding transcripts and their acknowledged significance in molecular biology, an exponentially increasing body of publications investigate the existence and functions of non-coding RNA (ncRNA). As already implied in **1.1 Central Dogma of Molecular Biology and the RNA Intermedium**, various classes of RNA bear catalytic or otherwise functional roles, beyond acting as messengers in protein synthesis. The major categories of ncRNA will be described in the following Subsections.

1.2.1 Transfer RNA

Members of the transfer RNA (tRNA) family are transcribed by RNA Polymerase III (Pol III), they are highly conserved and constitute the most abundant RNAs per cell in terms of sheer number of transcribed molecules[4]. Their size is approximately 76nt[5]. The main functions of tRNAs are two: **(a)** in a process called aminoacylation, a tRNA transcript is charged with and carries a specific amino acid and **(b)** during each translation elongation step, the mRNA triplet at place must be matched, or decoded, by the tRNA bearing the correct anticodon on its sequence. Thus, tRNAs are responsible for providing the correct amino acid residues to elongate the nascent polypeptide protein sequence.

tRNA functions are by and large dictated through the tRNA sequence (e.g. their anticodon sequence and their 3'-end which terminates in CCA almost universally[5]), secondary four-stem and tertiary L-shaped structure. Sequence conservation and the very specific cloverleaf-like structure of native tRNAs (**Figure 5**) are potent features that enable the genome-wide analysis and discovery of potential tRNA-encoding loci; bioinformatics applications that detect them primarily based on structural features are very widely accepted[6, 7]. Importantly, specific tRNA nucleotides are heavily modified; modification types (e.g. addition of methyl-, di-methyl-, carboxyl-groups, deamination of adenine) vary significantly among tRNAs and tissues or cell types and contribute to molecule stability, enhance decoder capacity or attribute extra-translational roles[4].

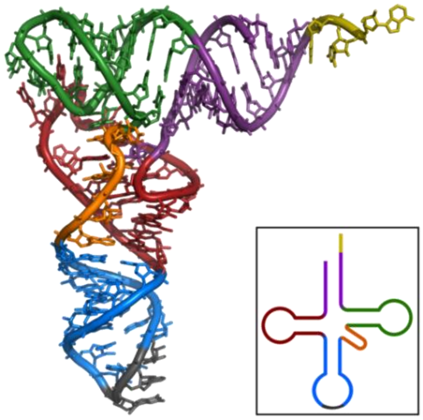


Figure 5. The characteristic tRNA structure. Secondary and tertiary structure of yeast tRNA-phe. The anticodon region is coloured light blue and the CCA-3′-end and acceptor sites in yellow and purple respectively. Tertiary structure derived from entry 1ehz in PDB. Figure under CC-Attribution-Share Alike License, created by Yikrazuul.

The total annotated tRNA genes reach 619 in human, according to the reference resource[8]. tRNAs carry anticodon sequences against 61 (i.e. 4^3 – 3 STOP codons) possible codons, thus an excess of tRNA loci exists in organisms' genomes. The application of tRNA-specific high-throughput sequencing assays has shown that the expression of tRNAs that match the same codon, isodecoders, varies among tissues, cell types and conditions, as does their amino acid charging capacity[9]. However, the high conservation and abundance of low charging capacity isodecoders is indicative of possible extra-translational functions that some tRNA might carry, such as the experimentally verified impact a $tRNA^{Asp}$ isodecoder has on the alternative polyadenylation and translation of its corresponding tRNA synthetase[10]. Recent analyses support the existence of tRNA-derived fragments (*tRFs*), which are small RNAs of functional potential, derived via specific tRNA cleavage by endogenous nucleases[11]. tRFs will be briefly described in **1.2.6** Other Small RNA Species.

1.2.2 Ribosomal RNA

Ribosomes are huge, highly complex multi-component ribonucleoprotein units. Bacterial ribosomes have a small subunit (SSU), consisting of 16S ribosomal RNA (rRNA) and 21 proteins, and the large subunit (LSU), consisting of 5S rRNA, 23S rRNA and 33 proteins. The interaction between RNA and protein in the ribosomal subunits confers specific structural arrangements; a body, a protein-rich head and a central domain on the SSU, and a solid extrusion called central protuberance and flexible ones called flexible stalks on the LSU. As exhibited in bacteria, eukaryotic ribosomes also consist of two subunits, (SSU: 18S rRNA, LSU: 5S, 5.8S, 25S rRNA), however 80 proteins interact with the rRNA and 21 additional RNA Expansion Segments and 7 Variable Regions exist[12]. The presence of supplemental components in eukaryotes reflects the higher order of regulation required in eukaryote translation, in order to coordinate the complex assembly, translation initiation, progress and localization[13, 14]. rRNA is transcribed by Pol I and III.

The binding sites for the charged tRNA (A site), peptidyl-tRNA (P site), and discharged tRNA (exit site) on the ribosome[15], as well as the ribosomal tunnel where primary folding of the nascent protein peptide is performed[16], are mainly composed of deeply conserved rRNA, signifying similarities of the underlying mechanisms across life.

rRNA is extensively modified, yet a limited set of different modification types is observed. rRNA modification is also conserved, it occurs on specific positions and clustering of modification sites on functionally important regions is observed (e.g. on tRNA binding sites and peptidyl-transferase domain, which creates peptide bonds among adjacent amino acid residues)[17, 18], indicating their functional or structural importance[17]. Modifications of rRNA can occur in a snoRNA-guided manner (e.g. the isomerization of uridine to pseudouridine (Ψ), further details on snoRNAs in **1.2.6** Other Small RNA Species)[19], or with stand-alone (i.e. not RNA-guided) modification enzymes. Pseudouridylation of U and ribose methylation on any base are among the most common rRNA modifications[20].

1.2.3 microRNA

microRNAs (miRNAs) are short endogenous single-stranded RNAs that post-transcriptionally regulate gene expression. miRNAs were discovered in 1993 by two teams (Victor Ambros's and Gary Ruvkun's – same day publications) that studied early post-embryonic developmental events in *Caenorhabditis elegans*. Lee, Feinbaum and Ambros reported the existence of two lin-4 transcripts (lengths 22 and 61nt) that are essential for the repression of LIN-14 protein. They indicated that the short transcript was part of the longer one, and that the latter could probably fold into a stem-/hairpin-like structure. Based on the presence of antisense complementarity between short lin-4 transcripts and the 3′UTR of lin-14, they proposed the negative regulation of lin-14 translation was due to an antisense RNA-RNA interaction[21]. Wightman, Ha and Ruvkun, utilized an X-gal 3′UTR reporter system to reach similar conclusions; lin-4 small RNAs formed multiple RNA duplexes on the lin-14 3′UTR, negatively affecting its translation[22]. Both groups commented on the conservation of lin-4 in relative species[21, 22]. The explosion in miRNA research however occurred in 2001, when the more conserved let-7 and other members of this RNA class were discovered in invertebrates and vertebrates using molecular cloning and bioinformatics analysis, revealing the breadth of miRNA suppressive role[23, 24]. A database cataloguing miRNA sequences in 6 species (*C. elegans*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*) was created[25] and now constitutes the reference resource of miRNA annotation, with more than 48860 mature miRNAs in 271 organisms covering animal, plant and viral species[26].

1.2.3.1 Canonical and non-canonical routes of microRNA biogenesis

miRNAs are transcribed predominantly from Pol II, and some from Pol III. Their genomic localization can be intergenic or intragenic (at a 1:1 ratio in human), with intragenic microRNAs existing mainly in intronic regions[27]. The phenomenon of polycistronic transcription of miRNAs at close genomic proximity with each other exists, in which case co-transcribed miRNAs may be called a (polycistronic) cluster[28]. The term "cluster" is not a synonym for "family", which usually refers to miRNAs sharing the same seed sequence (described below in miRNA Function)[29].

The canonical miRNA biogenesis is described in **Figure 6A**. Inside the nucleus, primary miRNAs (pri-miRNAs) are transcribed from their own genes and are bound by the microprocessor complex. This complex consists of RNA-binding protein DGCR8 (DiGeorge Syndrome Critical Region 8), which identifies specific motifs on the pri-miRNA sequence and Drosha type III RNase, an enzyme that specifically cleaves the pri-miRNA. Drosha action results in distinctive hairpins, precursor miRNAs (pre-miRNAs). Pre-miRNAs processed by Drosha possess a 3′ overhang of a few nucleotides that is typically seen in RNase III cleavage products[30]. They are transported to the cytoplasm by XPO5/RanGTP complex, where Dicer, another RNase III processes them further. Dicer cleaves the stem-loop and an imperfect duplex consisting of the mature miRNA and the opposing strand remains. The functional mature miRNA can be derived from either the 5′ or the 3′ end of the pre-miRNA, but "5p" mature miRNAs are more prevalent[26]. The RNA-

Induced Silencing Complex (RISC), a ribonucleoprotein complex with endonuclease activity, selectively loads one strand of the duplex and cleaves the other[28].

Non-canonical miRNA biogenesis pathways that make use of alternative machinery have been reported. A prominent example of non-canonical Drosha/DGCR8-independent pathways is the processing of miRNA-containing introns during transcript splicing (**Figure 6B**) whose hairpin-like products can be canonically exported to the cytoplasm and follow the canonical downstream route[31]. 7-methylguanosine-capped pre-miRNAs are an instance that bypasses Drosha processing and is exported directly by XPO1 (**Figure 6C**), where a main RISC component, Argonaute (AGO) processes and loads it[31]. Dicer-independent biogenesis characterizes endogenous short-hairpin transcripts that are processed by the microprocessor complex in the nucleus, exported and then directly loaded and processed by AGO in the cytoplasm (**Figure 6D**)[32].

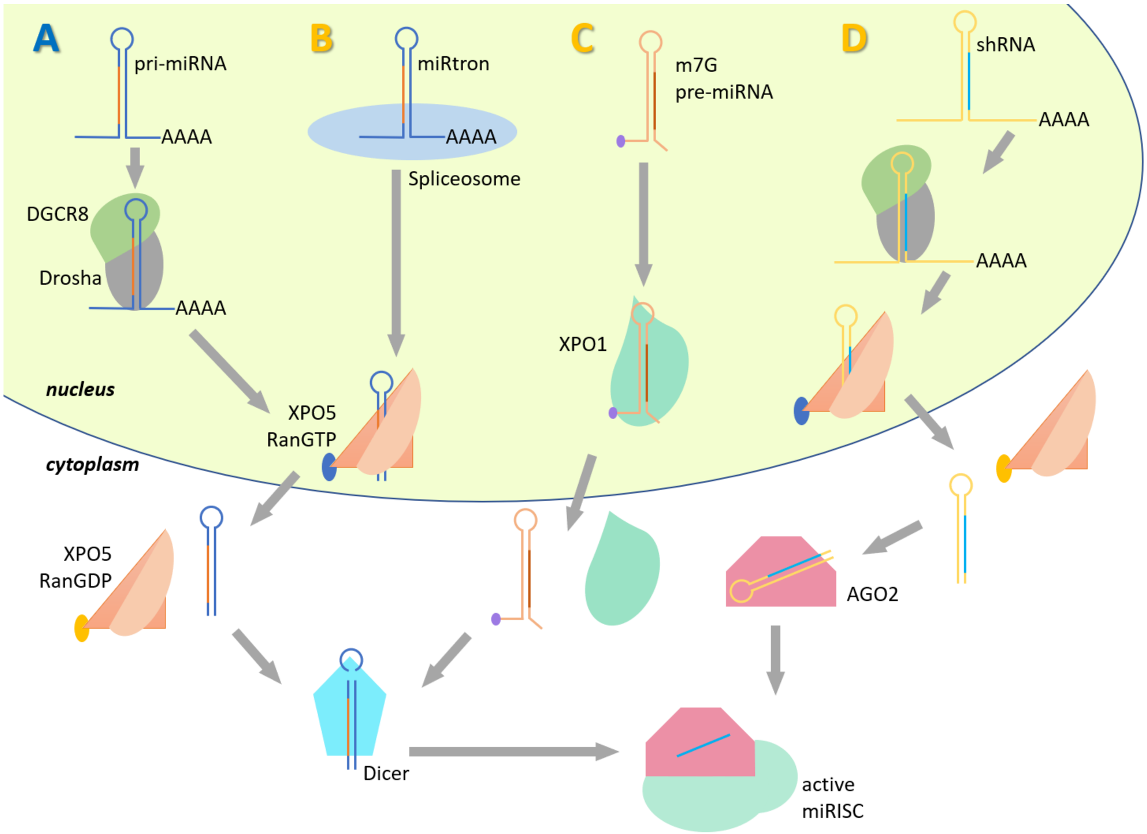


Figure 6. Canonical (A) and non-canonical miRNA biogenesis, from a miRtron (B), a modified precursor (C) and a small endogenous hairpin (D). Figure created for the purpose of this thesis.

1.2.3.2 microRNA function and roles: protein-coding transcript targets

The primary miRNA function displays commonalities with the main what is known as RNAi – RNA interference. During RNA interference, long double stranded RNAs occurring from sense and antisense transcription of genomic regions are processed by Dicer into 22nt small interfering RNAs (siRNAs). miRNAs are incorporated in RISC, like single stranded siRNA forms do, and guide it based on imperfect complementarity to target mRNAs. RISC's endonuclease activity can cleave the targeted mRNA or it can cause translational stall (**Figure 7**).

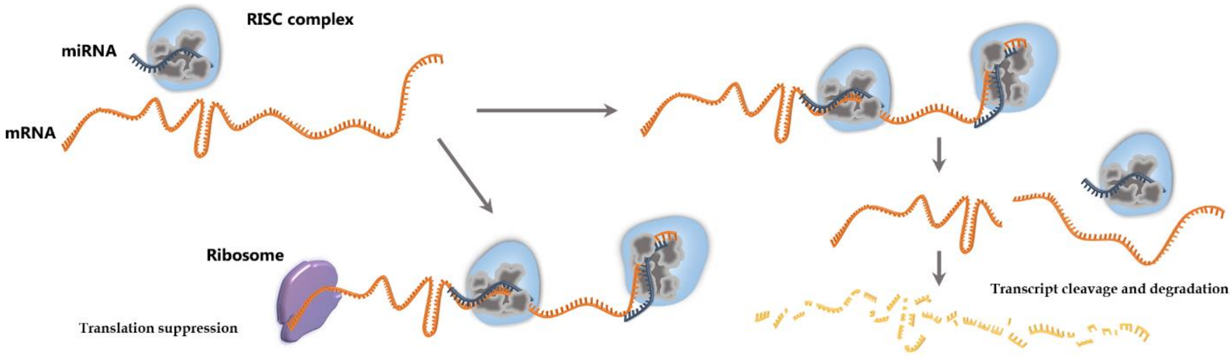


Figure 7. Canonical miRNA function. Targeting of protein-coding transcripts for translational suppression and/or cleavage and degradation. Figure adapted from Paraskevopoulou MD et al. (2018)[33] for the purpose of this thesis.

1.2.3.3 microRNAs as potential biomarkers

In recent years, the discovery that miRNAs exist in traceable and distinctive amounts in human tissues[34] has created increasing interest in their potential biomarker roles. Provided they do have discriminatory capacity, the short length of miRNAs makes them more robust biomarkers than longer RNA, since their degradation by nucleases is more limited. Especially their differential abundance in the circulatory system in response to various stimuli has brought forth the unique premise of using them as minimally-invasive biomarkers. A growing body of publications that estimate miRNA abundance in blood, blood derivatives (serum and plasma), extracellular vehicles and other body fluid samples (e.g. urine samples) have revealed miRNAs' potential to discriminate between healthy and disease states and to correlate significantly with a plethora of disease outcomes including metastasis, relapse, post-operative survival and the efficacy of therapeutic interventions[35-37].

1.2.4 Long non-coding RNA

Although coding RNA dominated our view of the transcriptome for decades, large multi-laboratory efforts such as the GENCODE Consortium[38] and the advent of Next Generation Sequencing (NGS) experiments have shaped differently our understanding of transcriptomic unit annotation and of RNA function as a whole. Long non-coding RNAs (lncRNAs) are a diverse ncRNA class, involved in numerous biological processes. They are defined loosely as transcripts that are longer than 200nt and bear no or very limited protein-coding potential. Under this definition, lncRNAs have been found in all branches of life and, depending on the choice of annotation source and terminology (e.g. "Should pseudogenes be considered as lncRNAs?"), their numbers can reach or even surpass those of protein-coding RNAs[39, 40].

GENCODE, the reference consortium harboring the most comprehensive annotation of non-coding transcripts, classifies lncRNAs according to their genomic locus of origin with respect to coding genes. As such, the main are the sense intronic, sense overlapping, antisense and intergenic, while the latest GENCODE version also integrates bidirectional promoter and macro lncRNA, as novel lncRNA types. Transcripts annotated as processed transcripts and 3’ prime overlapping ncRNAs are also specified as lncRNAs. Like mRNA, most lncRNAs are transcribed by Pol II, although Pol III and Pol I transcription of specific lncRNA members is reported[41].

The expression profiles exhibited by lncRNAs are being actively scrutinized. Unlike coding RNA, lncRNAs are thought to be more lowly expressed in general[42] and to display tissue-, cell-type- and even condition-specific [e.g. promoter-associated antisense lncRNAs (si-paancRNAs), induced upon oxidative stress[43]] expression patterns[44]. However, the idea that lncRNAs are lowly abundant is being challenged by more recent developments in RNA-Seq; single-cell sequencing techniques suggest that specific lncRNAs can be very highly expressed in very specific cell populations[45].

Members of the lncRNA class have been found to take part in diverse processes. Among the most prominent ones are: the role of H19 in epigenomic imprinting[46]; the capacity of XIST chromosome to confer X-chromosome inactivation in females, resulting in dosage compensation[47]; the nuclear retention and regulation of mRNAs by NEAT1[48]; the enhancer-like positive regulation of Snai1 and other lncRNAs in transcription of protein-coding genes[49]. Some lncRNAs act as sponges for miRNAs, forming interaction networks that are described as competing endogenous RNA (ceRNA) networks. This interaction competition can affect miRNA targeting, sequestering it from its protein-coding targets, while lncRNAs can also be degraded due to miRNA targeting. Examples of lncRNA sponge functions include intergenic-muscle differentiation 1 lncRNA (linc-MD1), which is implicated in myogenesis by sequestering miR-133 and miR-135[50], H19, which mediates muscle differentiation by acting as a sponge for let-7 miRNA[51] and cytoplasmic circular non-coding RNA CDR1as, which contains more than 70 MREs[52]. Pseudogenes can be regarded as lncRNAs and several studies confirm their interplay with miRNAs in ceRNA networks[53, 54].

1.2.5 Circular RNAs

Circular RNAs (circRNAs) constitute a newly discovered[55] class of RNAs that are produced via head-to-tail splicing (i.e. back-splicing). circRNAs have been found at least in *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*[56] and their high abundance has been reported in human and mouse neuronal tissue, independently of linear transcripts, while development- and spatiotemporal-specific expression patterns have been defined[57, 58]. circRNAs range from hundreds to thousands of nucleotides in length[56]. No clear consensus can be reached regarding circRNA function; some of them have been found to be translated[59], others to partake in RNA regulatory networks affecting miRNA targeting and turnover[60] and one instance to be involved in protein complex assembly, affecting the progression of the cell cycle[61]. On a higher level, a number of circRNAs have been found to exhibit crucial roles in development and plasticity[58, 62].

1.2.6 Other Small RNA Species

The discovery of functional short RNA classes such as miRNAs and siRNAs revealed their involvement in pervasive post-transcriptional regulation of gene expression, inaugurating the functional small RNA revolution. miRNAs have been the focal point of small non-coding RNA research, since they play a pivotal role in post-transcriptional regulation of gene expression[63], controlling pathways in health and disease[64, 65]. Recent studies have provided insight into the existence of novel biological roles of such sRNAs[66-68]. Using relevant approaches, new sRNA classes with debatable biological functions are still being discovered[69].

1.2.5.1 Small Nuclear RNAs

Small nuclear RNAs (snRNAs) are RNAs ~150nt in size that are bound to proteins, forming small nuclear ribonucleoprotein complexes (snRNPs) involved primarily in RNA splicing. snRNA biogenesis follows canonical Pol II/III transcription, 5′-capping and processing. The Initiator complex recognizes a 3′-signal sequence on immature snRNAs and cleaves the 3′ tail, prior to their Exportin1-dependent nuclear export[70]. In the cytoplasm, snRNAs undergo further maturation procedures and snRNP assembly. The survival motor neuron (SMN) protein complex and additional factors are employed in order to assemble Sm proteins and construct the snRNP complexes. Inside the snRNP, snRNAs are subjected to methylations of their 5′ cap, which functions as a nuclear localization signal[71]. Upon re-entry to the nucleus, some snRNPs are subjected to further modifications and re-assembly procedures in the Cajal bodies. Fully assembled snRNPs are directed to transcription sites where they can engage in spliceosome assembly[72]. snRNAs group into complexes that, via RNA-RNA interactions, identify splice sites and increase proximity of the exon$_1$ end, branch point and exon$_2$ start and form the catalytic spliceosome center [73], effectively acting as ribozymes.

1.2.5.2 Small Nucleolar RNAs

Small nucleolar RNAs (snoRNAs) are Pol II/III-transcribed RNAs acting as central regulators of post-transcriptional RNA modifications. Based on the presence of characteristic short motifs on their sequence, snoRNAs are divided into two main subfamilies: C/D-box snoRNAs and H/ACA-box snoRNAs[74]. The secondary structure of C/D-box snoRNAs enables protein recruitment towards the formation of snoRNP complexes, while a short antisense motif they carry, enables them to bind to target RNAs. Fibrillarin catalyzes the transfer of a methyl group to the 2′-hydroxyl group of a target RNA's ribose, in a site-specific manner, relevant to the D motif. H/ACA-box snoRNAs form two ACA-motif-containing hairpin-like structures connected via a hinge region harboring the H motif. This snoRNA group recruits, among others, uridine isomerase dyskerin, which is responsible for performing pseudouridylation on uridines of target RNAs, residing upstream of the H or on ACA motifs. Another snoRNA subclass is that of Cajal-body-specific RNAs, which predominantly act on the subnuclear Cajal body structures and direct the modification of snRNAs[75]. Orphan snoRNAs (i.e. snoRNAs not possessing significant sequence complementarity enabling canonical RNA targeting) have been also characterized and are under study[76]. Non-standard modifications have been found to be mediated by some orphan snoRNAs – found to guide the site-specific acetylation of rRNA cytidines[77] – and C/D-box snoRNAs (i.e. 2′-O-methylation of tRNA^{Met} C34 [78]), while their involvement in mRNA expression regulation has been proposed but constitutes an area under active study[74].

1.2.5.3 Piwi-interacting RNAs

Piwi-interacting RNAs (piRNAs) are short (24-30nt) RNAs that are present in animals in both the nucleus and the cytoplasm[79]. They are highly germline-specific (i.e. they are predominantly expressed in testes and ovaries) and are critically involved in germline development. piRNA sequences do not exhibit the well-defining characteristics observed in other sRNAs (e.g. the almost-universal hairpin forming capacity of their precursors, or the presence of a conserved seed region, in miRNAs); instead they are highly diverse among themselves, apart from a strong bias towards possessing a uridine in their first position[80] and a strong strand bias observed in some species[81]. The genomic distribution of piRNAs is clustered in transposon- and repeat-rich loci. During piRNA biogenesis, long single-stranded primary piRNA polycistronic RNAs are transcribed, out of which thousands of mature piRNAs can be derived in a non-coherent way (i.e. there is no definitive pattern in their production and even overlapping piRNAs are created). In *D. melanogaster*, the endoribonuclease Zuc can cleave primary piRNAs in a non-RNA-guided manner, without sequence preference[82]. An alternative pathway involves the combined action of Zuc with piRNA-loaded Piwi proteins, which are members of the Ago protein superfamily. Piwi proteins contain, Slicer, a catalytic cleavage unit that is RNA-guided. Mature piRNAs can be used to further enhance piRNA production in a positive feedback loop mechanism called ping-pong[83]. piRNA ping-pong and downstream piRNA activity is also observed in response to actively transcribed deleterious antisense transposons and constitute a process to downregulate their activity and maintain genomic integrity[84]. piRNA exerted silencing functions on transposons are exerted epigenetically, through DNA methylation[85], as well as at the post-transcriptional level. piRNA mediated decay of spermatogenic mRNAs and lncRNAs has also been observed[84].

1.2.5.4 tRNA-derived RNA fragments

tRNA-derived RNA fragments (tRFs), are a novel sRNA class of 13-48nt long molecules[86] that is derived from precise cleavage and processing at the 5′ or 3′ end of mature or precursor tRNAs, and studies indicate their possible involvement in miRNA-like RNA targeting as well as global translational suppression[87]. Based on the induced cleavage site(s), tRFs can be grouped into[88, 89]: **(a)** tRF-1s, derived from the 3′-end of tRNA precursors; **(b)** short tRF-3sa and **(c)** long tRF-3sb, derived from the 3′-end of mature tRNAs; **(d)** tRF-5sa, **(e)** tRF-5sb and **(f)** tRF-5sc, derived from the 5′-end of matures; **(g)** 5tiRs and **(h)** 3tiRs constituting the regions from the 5′/3′ end to the end of the anticodon stem; **(i)** tRF-2s which are centered around the anticodon stem. tRFs have been suggested to carry diverse roles, including regulation of gene expression, priming of viral reverse transcriptases and RNA processing[89], while some have been linked to pathogenic conditions[90].

1.2.5.5 Membrane-associated RNA fragments

Very recently, the presence of functional membrane-associated extracellular RNA (maxRNAs) fragments was explored via seamless integration of NGS approaches with low-yield molecular methods (i.e. FISH detection)[69]. Nuclear-encoded maxRNAs were found on cell types that involve cell-cell contacts, such as monocytes and their dendritic progeny cells, while their experimental perturbation led to attenuation of attachment between endothelial cells and monocytes, suggesting a potential functional role for maxRNAs in cell adhesion via interplay with membrane ribonucleoprotein complexes.

CHAPTER 2 – RNA Editing

2.1 A Plethora of RNA Modifications

Despite the apparent simplicity of the DNA and RNA four-letter nucleotide code, nucleic acids are subjected to a great number of chemical modifications. The first discovered modified RNA nucleotide was pseudouridine[91] (mentioned in **1.2.2** Ribosomal RNA) and numerous others followed as nucleic acid sequencing emerged. The contemporary "epitranscriptome" features more than 163 distinct RNA modification types[92]. Some are fairly simple modifications, such as the addition of a methyl or acetyl group, or the removal of an amino group (e.g. 1-methyladenosine, 4-acetylcytidine and inosine from adenine), while others can be more sophisticated additions of one or more larger chemical groups. Example cases of modifications are provided in **Figure 8**.

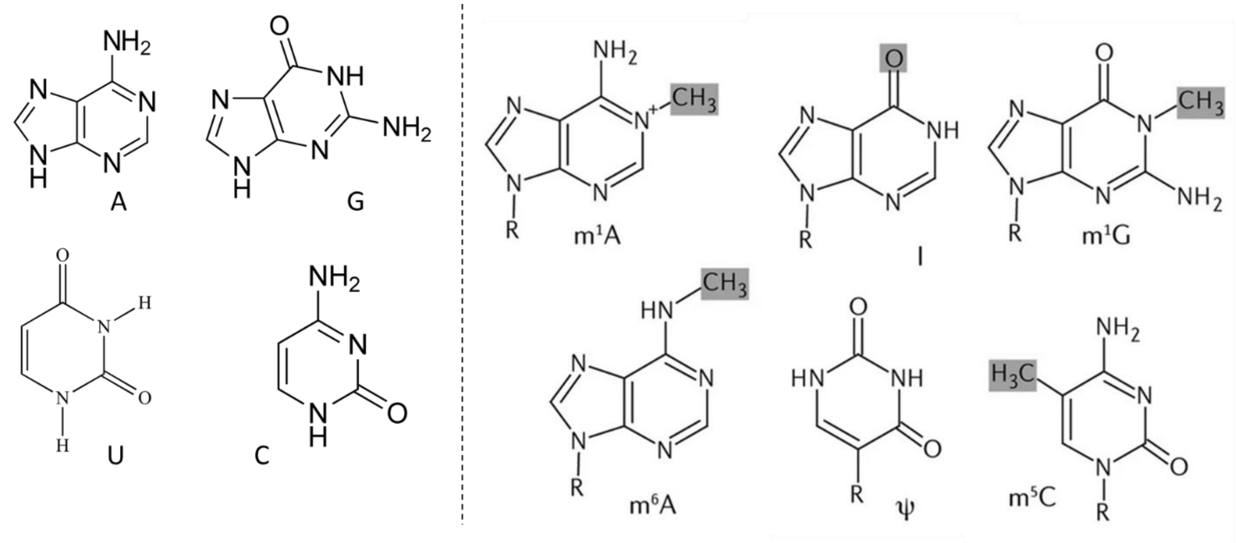


Figure 8. The un-modified RNA bases A, U, G and C and common, simple modifications: 1-methyladenosine (m^1A), N6-methyladenosine (m^6A), Inosine (I), pseudouridine (Ψ), 1-methylguanosine (m^1G) and 5-methylcytidine (m^5C). Figure adapted from Helm M & Motorin Y (2020)[93].

From early on, RNA modifications were being studied intensively on tRNAs and rRNAs which appeared to contain numerous extensively modified ribonucleosides (e.g. 13 modifications on average are harbored on each tRNA[94]). mRNA was also known from early studies to be methylated[95], while more recent studies on short non-coding RNAs revealed that snRNAs[96], snoRNAs[97] and miRNAs[98, 99] are also subjected to chemical modifications. The emergence of NGS technologies enabled the creation of global modification event maps[100], while the characterization of proteins involved in RNA de-methylation sparkled new interest in the RNA modification field, since it implied the possibility of even more flexible regulation of RNA products via the coordinated interplay of methyltransferases and demethylases[101].

2.2 Adenosine-to-Inosine Editing Catalyzed by ADAR Proteins

RNA editing has been characterized as substitutional since it is an RNA modification mechanism that does not result in a modified version of a common base. Instead, it chemically introduces a different common base in place of the existing one. However, the term "substitution" can be misleading, since no replacement or *transition* from one purine base to another occurs; RNA editing results in removal of a chemical group. RNA editing was first introduced as a process when de-amination was discovered to be exerted on an in-frame cytosine in human apolipoprotein-B transcript, transforming it to uridine and creating an early stop codon and thus a different protein isoform[102, 103]. APOBEC-1 RNA-specific cytidine deaminase and the Apobec-1 complementation factor (ACF) constitute the minimal components required to perform C-to-U transition. However, a number of cis-acting RNA elements (e.g. a sequence motif downstream of the edited site[104]), RNA secondary structure characteristics (preference for stem-loop structures[105, 106]) and the assembly of trans-acting co-factors in a multi-component Editosome complex[107] jointly regulate RNA editing extent and site-specificity. Importantly, **(a)** other APOBEC-targeted sites besides the apolipoprotein-B transcript have been found and **(b)** editing frequency can vary significantly among tissues and conditions (e.g. < 1% C-to-U editing in human liver and > 90% in small intestine)[108].

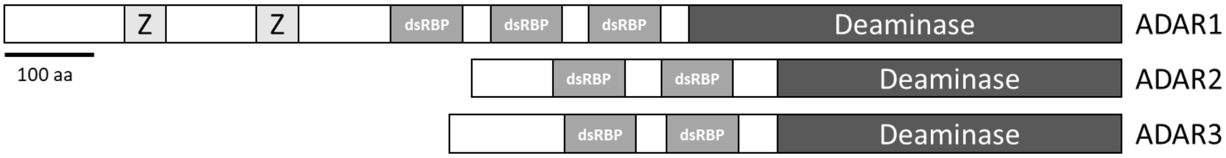


Figure 9. Graphic representation of the dsRNA binding domains and deaminase domain on the human ADAR 1, 2 and 3 aminoacid sequences. ADAR1 also containg Z-DNA binding domains. Figure adapted from Savva YA et al. (2012)[109] for the purposes of this thesis.

C-to-U editing discoveries were closely followed by observations of A-to-I conversions, taking place in double-stranded RNA (dsRNA) regions and being coupled with partial unwinding of the dsRNA duplexes[110]. A-to-I RNA editing was identified in *Xenopus laevis* oocytes and embryos but is known to affect the transcriptome of Metazoa[111]. This second and more widely exerted RNA editing mechanism is catalyzed by adenosine deaminase acting on dsRNA 1 (ADAR1) and ADAR2. A third member of the conserved ADAR family, ADAR3 is thought to be metabolically inactive; however it is speculated that it can compete with binding sites of ADARs 1 and 2, thus participating indirectly in regulation of A-to-I editing[109]. Catalytic ADARs are expressed in numerous tissues, while ADAR3 is only found expressed in the brain[112]. While the subcellular localization of ADARs is predominantly nuclear and nucleolar[113], amino-truncated isoforms of ADAR1 that still possess dsRNA binding domains are found to shuttle between nucleus and the cytoplasm[114]. Notably, dsRNA binding domains that confer the double-strand specificity are distinct from the de-aminase domain on ADARs (**Figure 9**). What is rarely clarified is the fact that ADAR RNA binding activity does require a double-stranded RNA substrate, while de-amination can also occur on single-stranded adenines along the dsRNA structure, such as RNA bulges and stem-loops.

Recent efforts to catalogue A-to-I editing events have resulted in human "inositome" maps spanning multiple tissues[115] and several analyses characterizing RNA editing in pathological conditions, mainly neurological and neoplastic disorders[116]. Findings in these studies indicate that A-to-I editing is a highly context-specific phenomenon.

2.3 Functional Impact of RNA Editing in Long RNA

The functional importance of RNA editing events has been established primarily in protein-coding genes. Editing events occurring in exonic regions can result in recoding events, that is, change in the translated amino acid sequence, inosine is treated like guanosine by the translation machinery. Additionally, bases to inosine have been found to cause delays to the ribosome function called translational stalling[117]. One of the most prominent and conserved[118] identified recoding events is the glutamine to arginine amino acid change induced by ADAR2 in GluA2, one of the subunits forming α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) tetrameric receptors in the Central Nervous System. AMPA receptors normally allow the influx of Na^+ and Ca^{+2} cations, upon activation by the neurotransmitter glutamate. The positively charged arginine on tetrameric receptors composed of recoded versions of GluA2 inhibits Ca^{+2} influx[119]. Hypo-editing of GluA2 in has been found to possibly contribute to AMPA-mediated excitotoxicity in Amyotrophic lateral sclerosis patients, leading to neuronal death[120] and to correlate significantly with increased malignancy in pediatric type IV astrocytoma [121]. *In vivo* studies in mice have revealed multiple editing events in serotonin GPCR receptor that can create 12 isoforms, which are responsible for various pathological metabolic phenotypes, such as hyperphagia, decreased fat mass and energy dissipation[122]. These events have been linked to Prader-Willi syndrome phenotypes[123]. cAMP signaling pathway components and neuronal transcripts have been found edited in early and clinical prion stages that could hold biomarker or translational premise[124].

A-to-I editing on splice-sites has been studied in insects, and a coordinated function between editing and alternative splicing has been identified in nicotinic acetylcholine receptors in Drosophila. Notably A-to-G mutations were found in related insect species that also exhibited the same splicing pattern[125]. ADAR knock-down experiments in human cancer cell lines has also resulted in global changes in splicing patterns, although the study's authors comment that the loss of specific events might not be enough to justify the extend of the observed phenomena and ADAR perturbation might affect splicing co-factors[126]. The implication of RNA editing in splicing, via recoded splicing factors has also been studied in the context of schizophrenia, using hundreds of cortical brain post-mortem samples[127].

Apart from exonic and splice-sites, A-to-I events happen even more frequently in non-protein-coding transcript regions (i.e. introns and the 5'/3'UTRs). Recent studies indicate that editing sites closer to computationally predicted miRNA Recognition Elements (MREs) in the 3'UTR had significantly higher A-to-I levels than distal sites and that RNA editing conferred a decrease in the targets' accessibility. Therefore, structural changes caused on target RNAs via RNA editing induce changes in their targeting by miRNAs[128].

A-to-I RNA editing has also been pinpointed as an important defense mechanism against the action of persevering deleterious transcription of Alu transposable repeat elements[129] and as a co-factor for nuclear retention of transcripts[130]. Recent single-molecule sequencing advances indicate the potential existence of edited adenines in poly(A)-tails, an observation which requires further investigation[131].

2.4 microRNA editing

The hypothesis that miRNAs could be edited in double-strand steps of their biogenesis has been formulated and supported in human and mouse brain from quite early using ADAR knock-out assays and fragments of the primary transcript of miR-22, which is abundantly expressed in a non-tissue-specific manner[132]. Consequently, editing on the transcript forms of miR-142, miR-143, miR-1-1 and miR-223 on specific adenosine positions was shown to occur in a site-specific manner. Importantly, specific editing events were noted to exert inhibitory function on Drosha processing, affecting miRNA biogenesis and edited pri-miRNAs to be degraded by Tudor-Sn[133], a nuclease identified as having potentially I-dsRNA specific substrates[134]. Apart from miRNA biogenesis intermediates, variable fractions of edited mature miRNAs have also been described in literature since and the effects of editing to their targeting repertoire have been speculated[135].

The extent of RNA editing in miRNAs begun to become explored extensively more recently with the use of small RNA-Seq[136, 137]. Major challenges arise in sRNA-Seq-based approaches, due to the existence of **(a)** reads that map to multiple genomic locations/features, **(b)** genomic variants (e.g. SNPs) in samples that reduce the specificity of editing identification, **(c)** sequencing errors. A sRNA-Seq study by de Hoon et al. utilized an expectation-maximization algorithm to efficiently guide short multimapping reads' placement and robustly identify editing events and suggested that miRNA editing is rarer than thought[136]. Alon S et al. set the state-of-the-art in systematic analysis of sRNA-Seq experiments for identification of ADAR-mediated editing by following a step-wise filtering approach which utilizes only uniquely mapping reads (**Figure 10**).

The Alon method has been adopted by numerous implementations. One issue not highlighted enough is the need to filter or minimally annotate potential RNA editing sites using SNP information from databases or, ideally, from matching DNA Sequencing experiments. Large-scale analyses of miRNA editing have been implemented on cancer datasets and editing events on miRNAs have been correlated with pathological phenotypes, such as cell invasion, as well as patient survival[138].

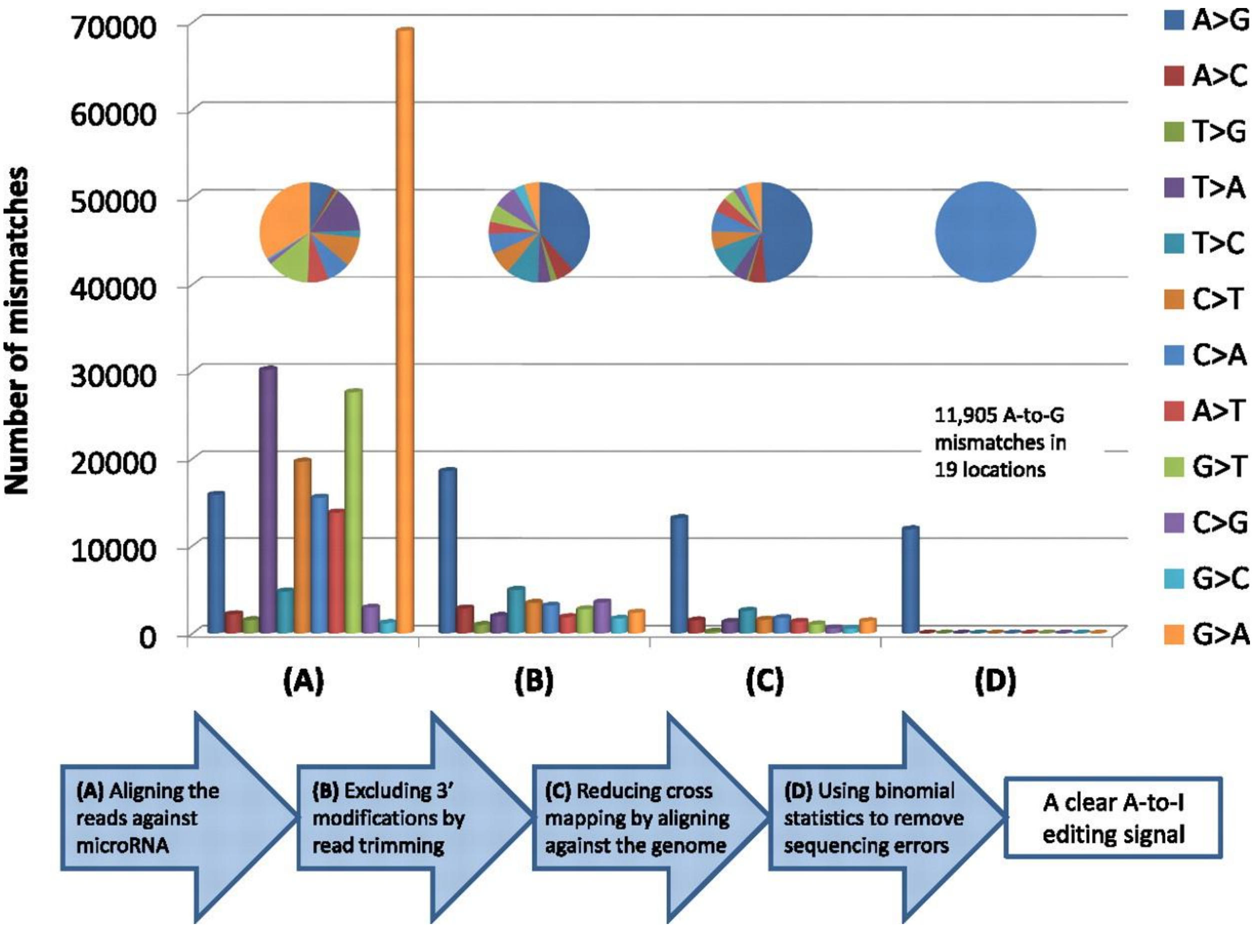


Figure 10. Bioinformatics analysis step-wise procedure to define robust editing events in small RNA-Seq experimental datasets. Figure retrieved from Alon S et al. (2012)[137].

CHAPTER 3 – Next Generation Sequencing

3.1 Reading DNA Bases in the Pre-NGS Era

Early sequencing attempts were focused on RNA molecules, because they were mainly single-stranded and shorter than DNA and depended on analytical chemistry methods defining di- and tri-nucleotide sequences after stepwise digestions from specific nucleases[139]. Using such methods, the first nucleic acid sequence – and structure – to be determined was that of a yeast alanine tRNA[140]. Sequencing endeavors took off in 1977, when A.M. Maxam and W. Gilbert reported a sequencing method that could enable reading up to 100nt fragments. It employed chemical degradation reactions capable of cleaving radio-labelled DNA at **(a)** preferentially G, **(b)** preferentially A, **(c)** C and T equally and **(d)** C. Resolving the products of these reactions by size via polyacrylamide gel electrophoresis permitted defining the sequence out of the pattern of radioactive bands (**Figure 11A**)[141]. The same year F. Sanger, S. Nicklen and A.R. Coulson proposed the chain-termination method, a synthesis-based method. They exploited enzyme DNA polymerase to begin copying the same DNA template sequence in four reaction tubes, however each tube would contain a different polymerase inhibitor nucleoside triphosphate (ddNTPs) out of ddGTP, ddCTP, ddTTP and ddATP, along with radio-labelled deoxyribonucleosides (dNTPs). Incorporation of an inhibitor would make the copying cease, resulting in a fragment of specific length. Size resolution in a gel would enable reading up to 200nt, in the same manner as in Maxam and Gilbert's method[142] (**Figure 11B**).

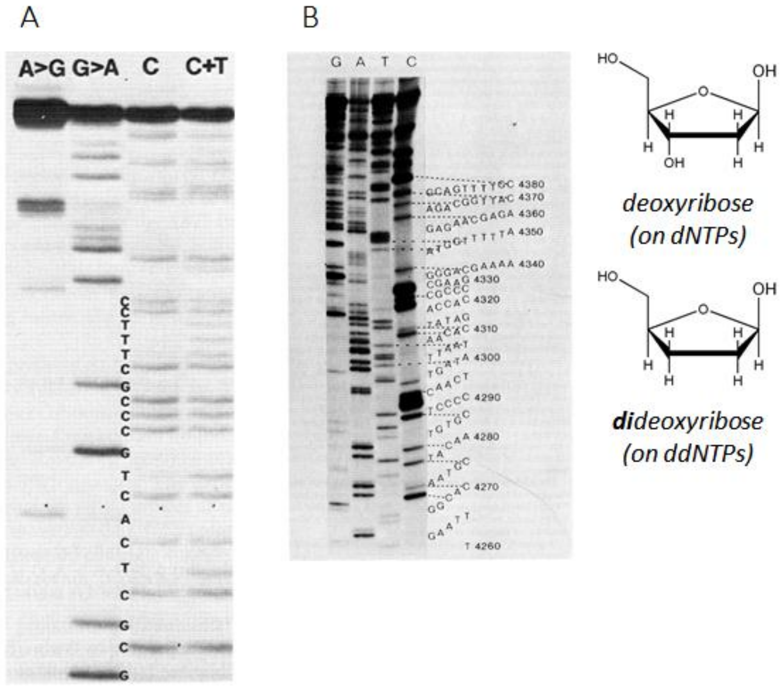


Figure 11. Primary gel-based methods of DNA Sequencing. (A) Maxam and Gilbert's chemical degradation method ran the radio-labelled DNA products of four different reactions in a polyacrylamide gel to separate them by size. (B) Sanger's method utilized DNA polymerase and a different ddNTP in each of four same reactions to inhibit the elongation process at A, T, C or G and again separated reaction products via electrophoresis in a polyacrylamide gel. Figure adapted from Maxam & Gilbert (1977)[141] and Sanger F. et al. (1977)[142] for the purpose of this thesis.

In the following decade, Sanger's "dideoxy" method became the gold standard in nucleotide sequencing. Important further developments aided the automation of sequencing. Most notably, fluorescent ddNTP dyes substituted radioactive labels and were combined with laser detection, enabling co-electrophoresis of all reaction products simultaneously and automated reading of fluorograms corresponding to sequence[143, 144]. The introduction of ~70μm inner diameter silica capillaries, in which samples would follow electroosmotic flow, enabled bypassing the laborious and time-consuming gel electrophoresis and paved the way towards commercialization and the sequencing of multiple samples in a parallel fashion[145]. The pace of the Human Genome Project was significantly boosted by these developments[146, 147].

3.2 Optimizing Sequencing

Solid phase sequencing, developed in the same time, permitted reactions to take place under optimal conditions (i.e. maintaining stoichiometry) and increased sequencing accuracy and yields. In the seminal solid phase sequencing protocol[148] the target DNA sequence (part of the *Staphylococcus aureus*'s protein A gene) was replicated exponentially via Polymerase Chain Reaction (PCR), employing biotin-labelled DNA primers. The double-stranded DNA PCR products were immobilized on magnetic streptavidin beads, by strong covalent bonds between streptavidin and biotin and converted into single stranded form. Sequencing reactions were performed on the beads with either radio-labelled or fluorescent primers and synthesized oligonucleotides eluted and subjected to electrophoresis.

The first "Next Generation Sequencing" application came with 454 Pyrosequencing technology. It employed bead-based amplification methods (emulsion PCR), by creating water-in-oil droplets where PCR of bead-fixed DNA occurred under optimal conditions (**Figure 12A**). Downstream, 454 Pyrosequencing depended on base-by-base detection of DNA polymerase activity via luminescence[149], in microwells containing one bead each; in each synthesis round a specific dNTP was allowed to react and pyrophosphate synthesis would indicate the extension of specific DNA chains bound to the solid phase[150]. 454 Pyrosequencing could yield approximately 1 million 700nt-long reads.

Flow-cell-based sequencing was introduced after the attachment of DNA on glass surfaces was made possible in 2006[151]. Solexa (and later, Illumina) designed glass flow-cells as a solid phase and proposed Sequencing-By-Synthesis (SBS). On the flow-cell, clusters of each starting DNA sequence are created via consecutive rounds of bridge amplification: single-strand DNA "bends" on the flow-cell, the adapters on both its ends being bound by complementary sequences on the flow-cell floor (**Figure 12B**). Polymerase extends a second strand and DNA annealing results in two single strands. After the initial DNA sequences have been bridge amplified, Sequencing-By-Synthesis occurs in a manner reminiscent of "dideoxy" method. Using fluorescent reversible termination dNTPs[152], all clusters on a flow-cell are extended by one nucleotide at a time. Fluorophores on the dNTPs must be cleaved before any other nucleotide can be added to the extending DNA chain. Rounds of extension and fluorophore cleavage determine the length of sequencing reads for the whole flow-cell. Illumina sequencing yields range from ~11 million to 40 billion reads, with read length varying from 50 to 500nt.

Importantly, via a supplemental step after sequencing of the one end of fragments, they can also be sequenced from the other end, producing paired-end reads. Depending on fragment size and read length, paired-end reads can be overlapping or simply distance-linked to each other and this greatly enhances their accurate placement during bioinformatic analysis.

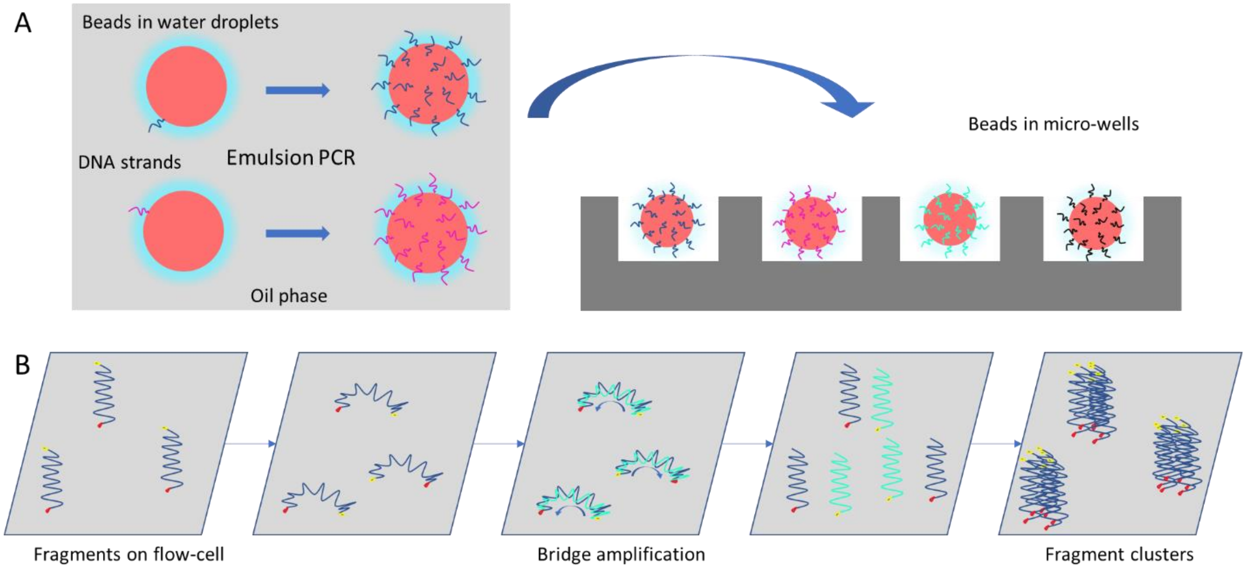


Figure 12. Amplification of starting DNA material with (A) emulsion PCR on magnetic beads, or (B) bridge amplification on flow-cells. Figure created for the purpose of this thesis.

Other high-throughput technologies have emerged that enable single molecule sequencing, bypassing amplification stages that were required by most methods in order to amplify the emitted fluorescence signal. Nanopore sequencing is a very prominent method in which native DNA/RNA molecules are transported from one chamber to another, by an artificial bioengineered enzyme, the nanopore[153]. The transport of single stranded DNA/RNA through the nanopore channel is performed in a base-by-base manner, causing characteristic drops in electric current which enable nucleotide detection[154]. Oxford Nanopore's applications are highly customizable, thus yields and read lengths can vary substantially. Notably, extremely long reads (i.e. up to 2 million bases long) have been reported[155]. These ultra-long-read, label-free technologies show premise even in detection of nucleic acid modifications from native sequence alone, however a number of obstacles including their relatively low-yield, high error rates and limitations in mass production of consumables hinder their widespread use. Currently, Illumina instruments and Illumina-generated datasets dominate the sequencing space and public repositories respectively.

3.3 Sequencing Applications

Principles presented in **3.2 Optimizing Sequencing**, largely refer to instrumentation details and sequencing per se; based on these principles, hundreds of library-preparation protocols have been developed. These have enabled, among others, the sample-specific **(a)** identification of genomic variants, **(b)** abundance estimation of coding and non-coding RNAs, characterization of **(c)** accessible DNA, and **(d)** of DNA and RNA regions bound by DNA-binding-/RNA-binding-proteins (DBPs/RBPs respectively). Some of the most widely utilized protocols are briefly presented below, with a focus on Illumina sequencing.

3.3.1 DNA Sequencing

The protocols upstream of genomic DNA sequencing include DNA extraction, fragmentation adapter ligation and amplification. Fragmentation, required primarily due to limitations in read length exhibited by sequencing technologies, can be performed mechanically (i.e. sonification) or chemically via transposase enzymatic action (e.g. "tagmentation" in Nextera DNA Sample Preparation Kit), which randomly cleaves DNA and also ligates adaptor molecules on both ends of cleavage products[156]. Amplification via PCR is performed on adapter-ligated DNA fragments, using primers that additionally introduce sample-specific barcode indices.

Alternative protocols enable the enrichment of distinct genomic regions of interest, in order to achieve their increased sequencing depth at the expense of sequencing breadth. Targeted sequencing employs hundred-base-long biotin-tagged DNA probes derived from RNA, and streptavidin beads, in order to capture relevant fragments from a typical DNA Sequencing library[157]. Whole Exome Seq (WXS) constitutes a critically acclaimed version of targeted DNA Sequencing that enables very deep sequencing (e.g. 500X) of the exonic parts of the genome and consequently robust and relatively inexpensive identification of germline and pervasive somatic mutations in patient samples.

3.3.2 RNA Sequencing

The unbiased extraction of total cellular RNA yields a lysate dominated quantitatively by the most abundant RNA species in a cell, rRNA. Thus, a number of techniques have been developed in order to enrich the RNA fractions of interest during RNA-Seq library preparation. The most common practice is poly-A selection, a process that utilizes oligo-dT sequences attached to magnetic beads to capture mRNAs and other poly(A)-tailed RNAs. Alternatively, rRNA depletion processes can be followed using beads coated with probes against rRNA sequences to sequester them, or specifically digesting them with RNase H, using antisense DNA oligos.

After target RNA species enrichment, remaining RNAs are subjected to fragmentation and reverse-transcribed into double-stranded complementary DNA (cDNA). Likewise, RNA fragmentation occurs via chemical or mechanical means, or it can be conducted post-cDNA-construction[158].

In classic RNA-Seq, DNA adapters are ligated in a non-specific manner on cDNAs and the directionality of RNA is lost upon sequencing. This hinders the accurate resolution of bi-directionally transcribed genomic loci. However, approaches for stranded RNA-Seq have been developed and mainly depend on adapter incorporation in steps prior to conversion to cDNA[159].

Light PCR amplification of adapter-containing cDNAs is performed prior to sequencing in order to overcome the sequencer's detection limits, although this can result in misrepresentation of RNA populations in the generated reads relative to the starting material and a number of in silico and bench methodologies have been developed to overcome this[158].

3.3.3 Small RNA-Sequencing

Unlike typical RNA-Seq, where the most relevant transcripts can be retrieved via poly-A selection and/or unwanted ribosomal RNAs can be depleted, small RNAs do not possess characteristics to enable their sequence specific discrimination during library construction. Therefore, 5′ and 3′ adapters are indiscriminately ligated on the whole RNA lysate and synthesis of cDNA is performed (**Figure 13**). During PCR amplification, sample-specific indices are added to enable parallel sequencing of more than one sample and de-multiplexing of generated reads afterwards. The small RNA fraction is usually extracted from the library via electrophoresis and size selection. Single-end sequencing by synthesis is performed only from the side of the 5′ adapter, producing typically 50nt reads. Since adapters are ligated in 5′- and 3′-specific manner on the native RNA, sRNA-Seq libraries are stranded, i.e. no reads deriving from the opposite strand of starting RNAs are produced. The dominant RNA species in sRNA-Seq datasets is that of miRNAs, although tRNA fragments, tRFs, piRNAs, snRNAs, snoRNAs and fragments of longer coding and non-coding RNAs are also frequently observed[160]. Very recently, sRNA-Seq has been reported to also carry RNA evidence on microbial species abundance in tissue samples[161].

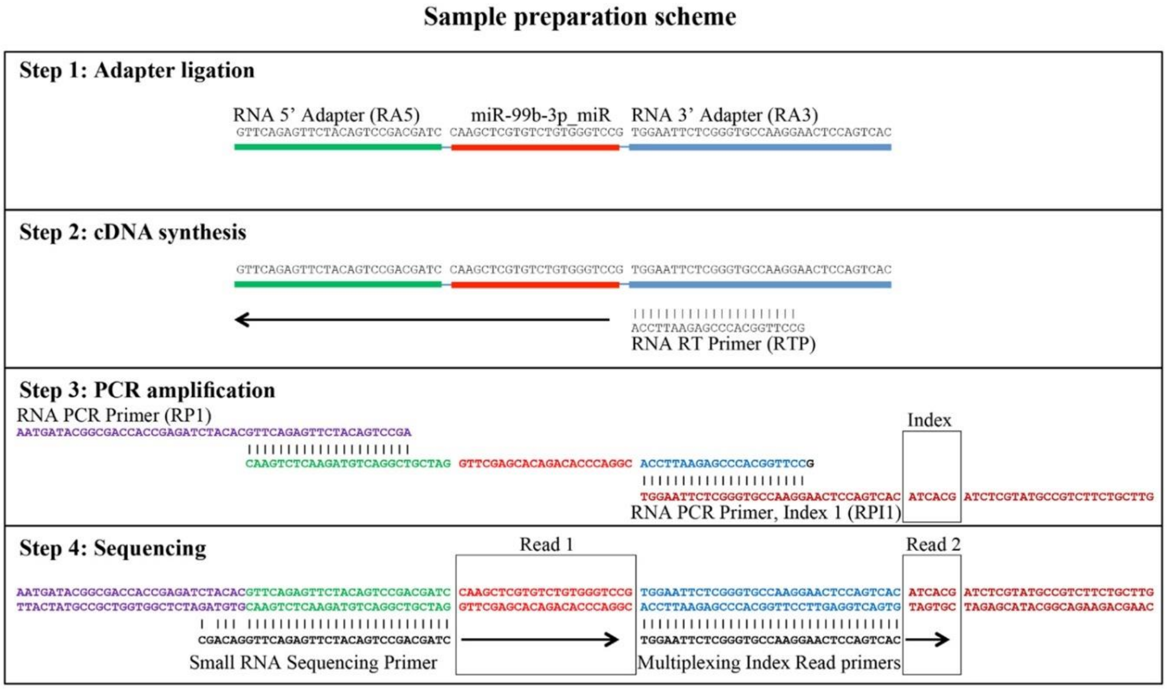

Sample preparation scheme
Step 1: Adapter ligation
RNA 5' Adapter (RA5)
miR-99b-3p_miR
RNA 3' Adapter (RA3)
GTTCAGAGTTCTACAGTCCGACGATC CAAGCTCGTGTCTGTGGGTCCG TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC
Step 2: cDNA synthesis
GTTCAGAGTTCTACAGTCCGACGATC CAAGCTCGTGTCTGTGGGTCCG TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC
ACCTTAAGAGCCCACGGTTCCG
RNA RT Primer (RTP)
Step 3: PCR amplification
RNA PCR Primer (RP1)
AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA
CAAGTCTCAAGATGTCAGGCTGCTAG GTTCGAGCACAGACACCCAGGC ACCTTAAGAGCCCACGGTTCCG
Index
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC ATCACG ATCTCGTATGCCGTCTTCTGCTTG
RNA PCR Primer, Index 1 (RPI1)
Step 4: Sequencing
Read 1
Read 2
AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATC CAAGCTCGTGTCTGTGGGTCCG TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC ATCACG ATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGCAAGTCTCAAGATGTCAGGCTGCTAG GTTCGAGCACAGACACCCAGGC ACCTTAAGAGCCCACGGTTCCTTGAGGTCAGTG TAGTGC TAGAGCATACGGCAGAAGACGAAC
CGACAGGTTCAGAGTTCTACAGTCCGACGATC
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC
Small RNA Sequencing Primer
Multiplexing Index Read primers

Figure 13. Illumina TruSeq small RNA Sequencing library preparation protocol. Figure retrieved from Muller et al. (2014)[162].

3.3.4 Immunoprecipitation Followed by Sequencing

Apart from targeted/non-targeted techniques for sequencing of DNA and RNA fractions, numerous more sophisticated approaches have been developed to enable the characterization of epigenetic and epitranscriptomic phenomena that underlie physiological and pathological conditions. The experimental antibody-based targeting of DNA- and RNA-binding proteins and their co-precipitation along with bound DNA/RNA fragments has contributed significantly to this. Chromatin Immunoprecipitation and Sequencing (ChIP-Seq) experiments have resulted in comprehensive maps of actively transcribed genomic loci and of active Transcription Factor Binding Sites (TFBSs) against specific Transcription Factors (TFs) in a genome-wide scale, leading to our better understanding of the epigenomic regulatory mechanisms of transcription[163].

In a similar fashion, high-throughput methodologies have been developed for the pull-down of RBPs. The protocol variants of Crosslinking and Immunoprecipitation followed by Sequencing (CLIP-Seq) enabled the high-throughput mapping of RNA-binding protein marks[164]. The RBPs that have been most widely studied via CLIP-Seq are RISC-component proteins AGO. Combined with sRNA-Seq datasets from the same cell line/type, AGO-CLIP datasets provide direct evidence on miRNA-RNA interactions in high-fidelity and on a transcriptome-wide scale. HITS-CLIP, PAR-CLIP CLEAR-CLIP and CLASH are prominent CLIP-Seq examples (**Figure 14**)[165].

In HITS-CLIP-Seq, cultured cells or live tissues are subjected to 254nm ultraviolet (UV) irradiation, enabling the creation of covalent bonds, crosslinks, between RBPs and their bound RNA[164, 166]. Cell lysis, RNase digestion and immunoprecipitation steps remove unbound RNAs and leave only RBP-bound fragments (e.g. fragments of miRNA-targeted transcripts containing the targeted region in the case of AGO-HITS-CLIP), which can be then subjected to library preparation and high-throughput sequencing.

PAR-CLIP[167] constitutes a CLIP variant displaying increased performance in protein-RNA crosslinking. This is achieved by culturing cells in medium supplemented with 4-thiouridine (4-SU) or 6-thioguanosine (6-SG), which are photoactivatable ribonucleoside analogues of U and G respectively, that enhance UV-crosslinks at 365nm. Additionally, T-to-C and G-to-A crosslinking-induced mutations occur during library preparation of PAR-CLIP immunoprecipitates (developed with 4-SU and 6-SG respectively). Mutated PAR-CLIP-derived regions have been proposed to carry diagnostic potential over non-mutated regions regarding miRNA targeting status. However, we have shown that non-T-to-C-containing read clusters from PAR-CLIP experiments also harbor functional miRNA binding events and can be confidently incorporated in PAR-CLIP-guided miRNA target identification algorithms[33].

CLEAR-CLIP and CLASH are CLIP variants that critically ligate miRNA-MRE duplexes found during immunoprecipitation. This step invoked the generation of "chimeric" miRNA-MRE reads, resolving by and large the problem of matching miRNAs to their experiment-derived targeted regions.

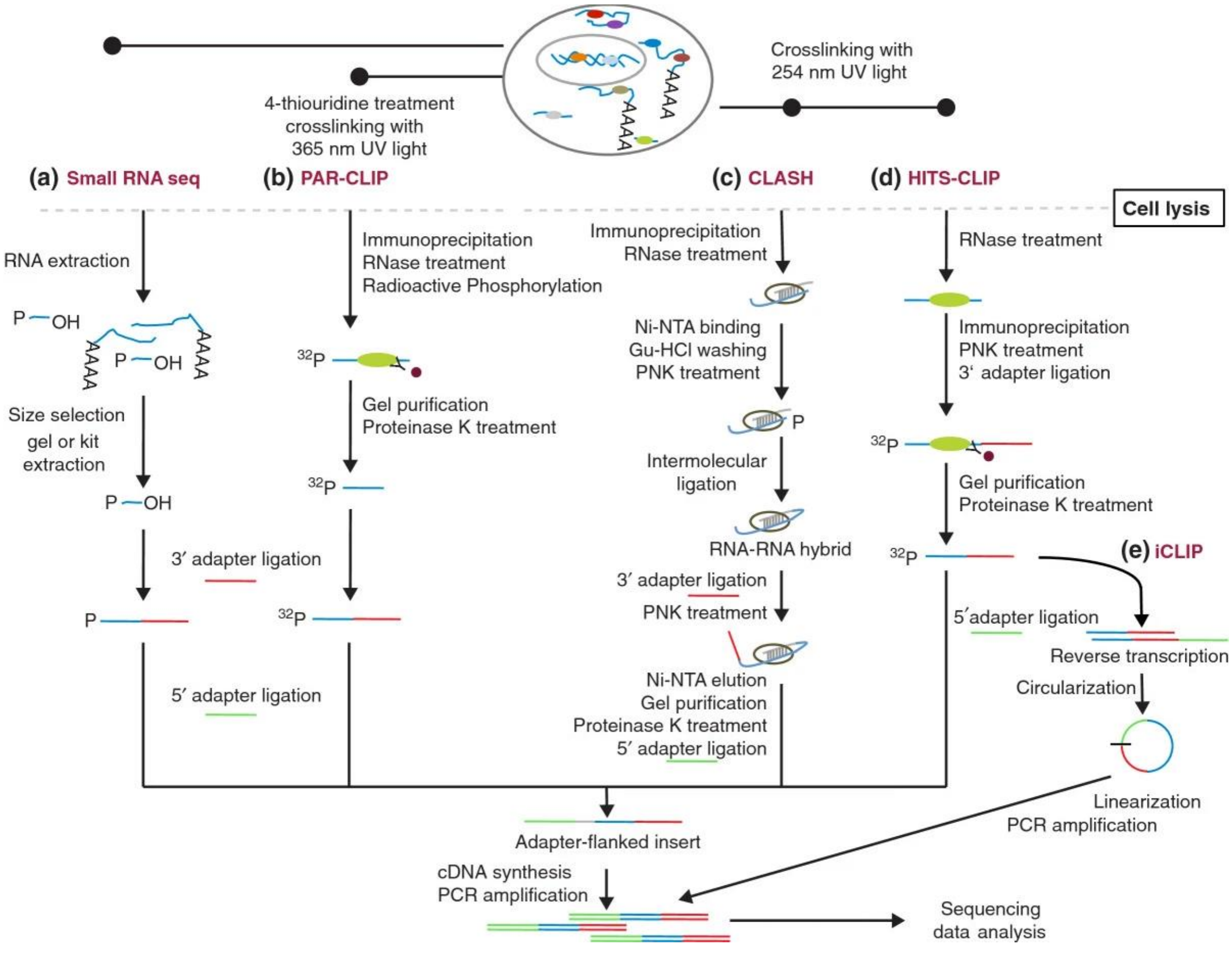


Figure 14. Brief presentation of sRNA-Seq, PAR-CLIP, CLASH and HITS-CLIP library preparation protocols.Figure retrieved from Mittal N & Zavolan M (2014)[165].

SCOPE

Chapters 4 to 7 describe RNA bioinformatics methods developed during the course of this thesis, and corresponding results.

Small RNA-Seq bioinformatics has yet to reach maturity, as a number of unique challenges arise from the protocol's unique characteristics, compared to classic RNA-Seq, and from new discoveries in RNA biology that require specialized implementations. **METHODS & RESULTS**

CHAPTER 4 – Quantification of Small RNAs describes collaborative sRNA-Seq analyses that led to the creation of Manatee[160], a tool for the quantification of small RNA species from sRNA-Seq experiments that combines annotation information with robustly aligned reads to guide the placement of multi-mapping reads.

MicroRNAs are estimated to participate in the majority of molecular processes and defining the individual and combined implications of miRNAs goes beyond and above miRNA quantification. The cataloguing of their experimentally verified interactions in as many organisms, tissues, cell types and experimental conditions as possible is the ultimate keystone for accurate incorporation of miRNAs in functional investigations. **CHAPTER 5 – An Enhanced microRNA Interactome** presents the collaborative efforts that led to the development of DIANA-TarBase v8.0[168] and DIANA-LncBase v3.0[169], which constitute worldwide reference databases for the retrieval of coding and non-coding miRNA targets.

The identification of RNA interplay and its functional consequences if often investigated from a reference-biased viewpoint (i.e. we rely on reference transcript/sRNA/genome sequences, therefore underlying sample-specific genomic, epigenomic and even epitranscriptomic phenomena of importance may be ignored). On this note, an effort was placed to provide a customizable RNA editing identification pipeline that can incorporate sample-specific DNA information on the editing calls. **CHAPTER 6 – RNA Editing Bioinformatics** presents **(a)** the creation of this R-based *in silico* pipeline for identification of RNA editing events from sRNA-Seq datasets, **(b)** its case-study application on brain and non-brain samples and **(c)** meta-analyses regarding the prominence and potential effects of RNA editing events on pre-miRNA hairpin formation, miRNA abundance and targeting of transcripts and molecular pathways.

CHAPTER 7 – Participation in Research Projects briefly describes my participation in additional research projects and collaborations during the course of my PhD.

METHODS & RESULTS

CHAPTER 4 – Quantification of Small RNAs

In subsections **4.1** and **4.2** sRNA-Seq analyses of multimapping reads and the development and performance evaluation of Manatee small RNA quantification tool are presented. During this study, I performed algorithmic optimizations on Manatee code. I also created artificial datasets and utilized them to test the behavior of Manatee under different parameter options. I executed Manatee and sRNAbench against the created simulated dataset and created scripts to derive performance metrics and plots for the benchmarked tools.

This collaborative work was published in Scientific Reports (I.F. 3.998): JE Handzlik, **S Tastsoglou**, IS Vlachos and AG Hatzigeorgiou. *Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data*. 2020.

4.1 Multimapping sRNA-Seq Reads Withhold Useful Information

The problem of correctly assigning multimapping sequencing reads (i.e. reads that map equally well to more than one genomic/transcriptomic position) differentiates RNA-Seq from sRNA-Seq experiments. The larger size and paired-end nature of typical RNA-Seq reads, diminishes multimapping reads. In sRNA-Seq experiments it is a more broadly occurring phenomenon. Still, a number of sRNA-Seq quantification applications and studies adopt strategies derived from RNA-Seq analysis for the placement of multimapping reads (e.g. retaining only uniquely mapping reads, assigning them wholly, or splitting them equally among positions, as well as even randomly assigning them to one out of multimapping positions). A primary analysis of the putative genomic origins of multimapping reads, with regards to the existence of gene annotation and of read clusters composed of Uniquely Aligned Reads (UARs) seemed appropriate.

4.1.1 Overlapping Multimapping Reads, Uniquely Aligned Clusters and Annotation

Quality-check and adapter/contaminant removal was performed on 30 sRNA-Seq libraries using FastQC[170] and Cutadapt[171], respectively. Reads were mapped against the GRCh38 human genome assembly using bowtie. Multimaps with up to 50 multimapping positions were retained for further analysis, along with UARs (i.e. reads mapping uniquely to the genome with one allowed mismatch and Bowtie "--best --strata"[172]). Clusters of UARs were created genome-wide for each sample, retaining details on the chromosome, strand, start and end position of clusters consisting of at least one read. Non-coding annotation available in Ensembl v85[173] and miRBase v21[174] was used to construct a reference for genomic features. Specifically, lincRNA, Mt-rRNA, Mt-tRNA, processed transcript, rRNA, scRNA, snoRNA, snRNA, sRNA and vtRNA gene types were derived from Ensembl, while miRNA precursor and mature forms were derived from miRBase. A minimum 1nt overlap between an aligned read and any annotation feature was required to assign that read to cases of annotated regions (containing/lacking UAR clusters). Annotation features and UARs were extended by 50nt at each end to allow flexibility in read assignment without adding bias towards one or the other.

4.1.2 Defining Distinct Multimapper Groups

Analysis of 30 distinct human sRNA-Seq libraries derived from hepatoblastoma, liver, brain, gallbladder, colon, lung, pancreas, skin, tongue, thyroid, and heart tissue, embryonic stem cells, MCF7 and HepG2 cell-lines was performed. The extent of multimapping reads (also, "multimaps") and uniquely mapping reads in sRNA-Seq datasets is presented in **Figure 15a**. Five examined cases of positioning multimaps were based on reads with 2 to 17 multimapping regions (**Figure 15b**). According to the analyzed cases, a multimap may fall into:

unannotated regions of UAR clusters (denoted as blue)
annotated regions lacking UAR clusters (red)
annotated regions that also contain UAR clusters (green)
unannotated regions that also lack UAR clusters (orange)
annotated regions and regions with UAR clusters with no concordance (pink).

Case 3, which includes multimaps falling into regions with both existing annotation and UAR clusters, was further analyzed and examined for the number of such regions per multimap (**Figure 15c**). For example, the majority of multimaps with two possible mapping loci had UARs and annotation for both mapping positions. The majority of reads with four possible mapping loci presented one out of four regions with existing UARs and annotation.

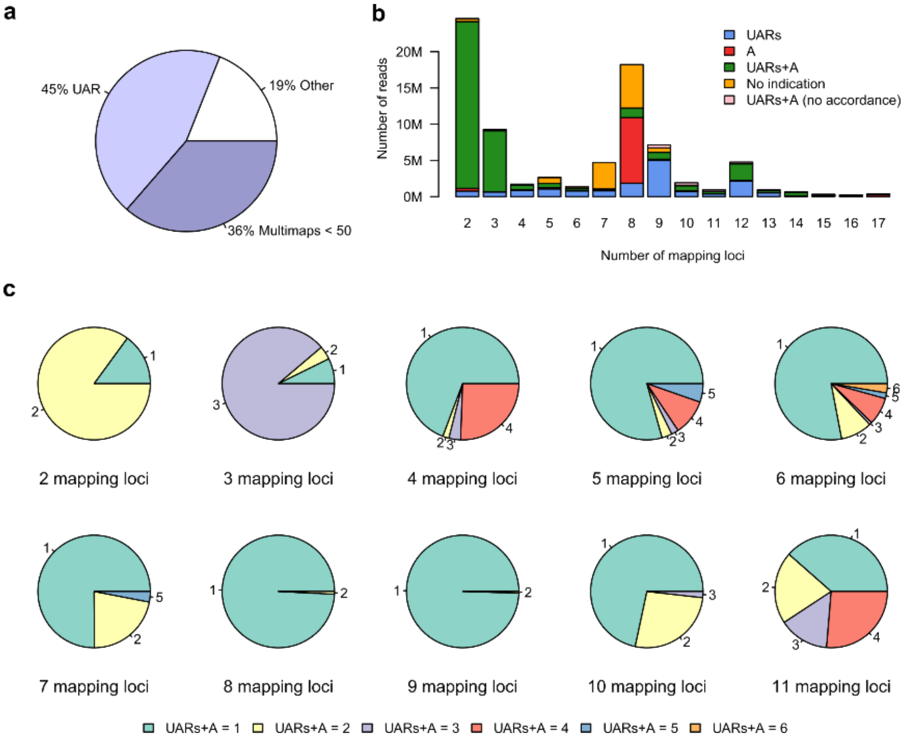


Figure 15. Frequency, proportions and characteristics of multimaps in sRNA-Seq libraries. (a) The average number of UARs, multimaps, and other reads (i.e. unaligned/multimaps exceeding the defined threshold) across all samples. (b) Multimap read categories based on available annotation and UARs. Colors mark five examined cases where each multimap is screened for available annotation and UARs. (c) Proportion of multimaps and the number of their mapping regions with both UAR clusters and available annotation. Figure retrieved from Handzlik E et al. (2020)[160].

A large portion of sRNA-Seq reads (36%) in the analyzed datasets mapped to more than one genomic location (**Figure 15a**). 19.7% of total multimaps fell into regions with UARs lacking annotation, while for 15.2% no straightforward information of positioning or annotation was available (**Figure 15b**). Algorithms based on genomic alignment which rely entirely on UAR information, would fail to account for cases of multimaps that could otherwise be assigned to existing annotation (red in **Figure 15b,** 13.3% of total multimaps). On the other hand, tools dedicated entirely to a specific RNA biotype would be biased towards that type and disregard the existence of other mapping/multimapping locations.

4.2 Small RNA Transcriptome Quantification Employing Annotation and Uniquely Aligned Reads

The conclusions yielded by the multimaps analysis constituted the basis for the Manatee algorithm which simultaneously incorporates information from UARs and existing annotation to tackle the multimap issue, without prioritizing any particular sRNA type (**Figure 16**). Manatee additionally attempts to rescue highly multimapping and unaligned reads, by transcriptomic alignment rounds and outputs information about expressed unannotated genomic loci that could harbor still unknown small RNA products.

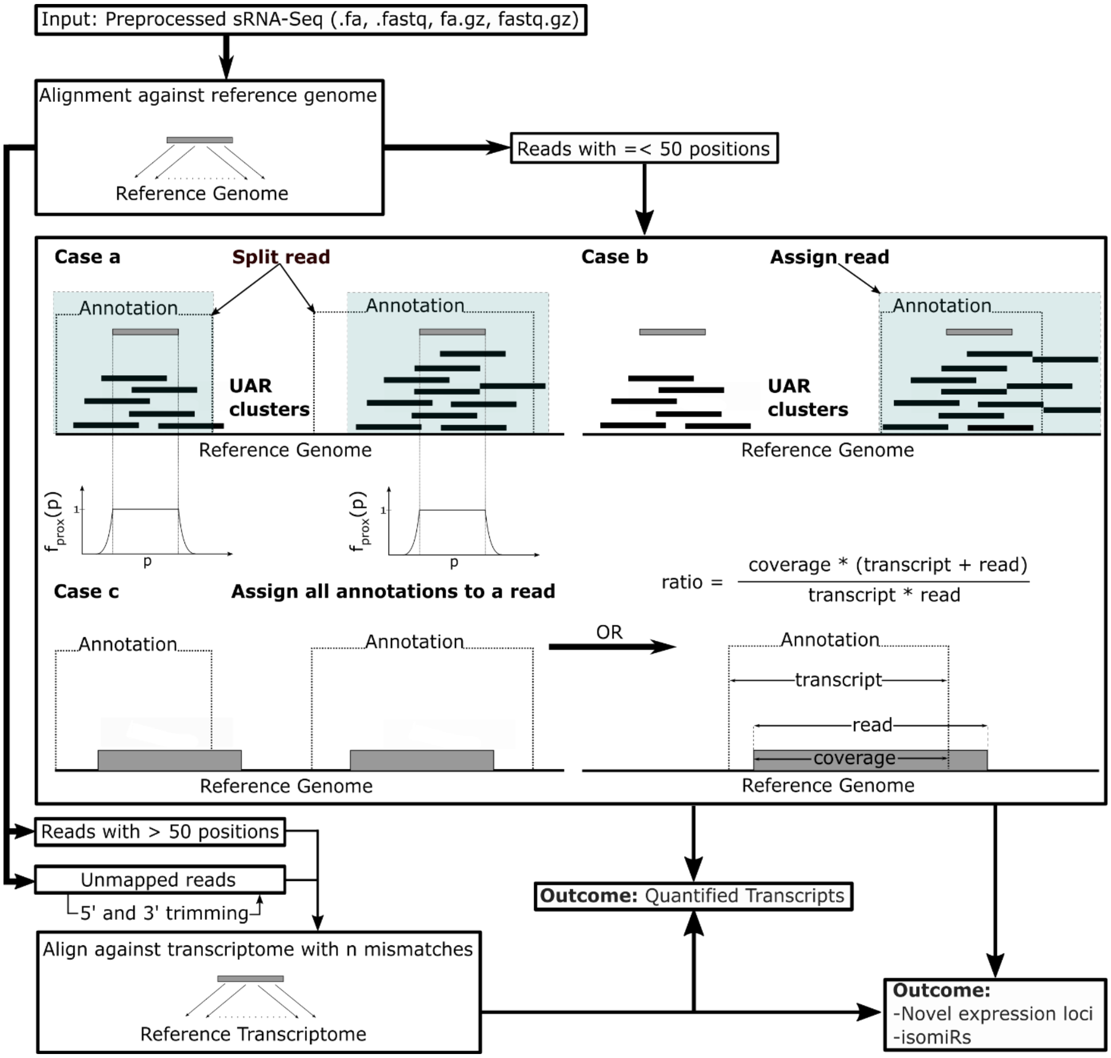

Input: Preprocessed sRNA-Seq (.fa, .fastq, fa.gz, fastq.gz)
Alignment against reference genome
Reference Genome
Reads with =< 50 positions
Case a
Split read
Annotation
UAR clusters
Reference Genome
$f_{prox}(p)$
1
p
Case b
Assign read
Case c
Assign all annotations to a read
OR
ratio = coverage * (transcript + read) / transcript * read
transcript
read
coverage
Reads with > 50 positions
Unmapped reads
5' and 3' trimming
Align against transcriptome with n mismatches
Reference Transcriptome
Outcome: Quantified Transcripts
Outcome:
-Novel expression loci
-isomiRs

Figure 16. Manatee analysis workflow. Reads with up to 50 multimapping positions are either (a) split among their annotated and uniquely aligned read (UAR)-containing loci according to Equation 1, (b) assigned to regions containing both annotation and UARs or (c) assigned to loci with existing annotation. In case of (c), if an annotated miRNA is within the annotated loci, a ratio for selecting the best fitted transcript is used to prioritize mature miRNAs over precursors. Reads with more than 50 mapping positions, reads which could not be mapped to the genome, and reads that could not be assigned to regions with existing annotation and UARs are aligned against the transcriptome with gradual increment of allowed mismatches. The output results contain quantified transcripts, putative novel expression loci and isomiR sequences. Figure retrieved from Handzlik E et al. (2020)[160].

4.2.1 MANATEE quantification platform

Manatee requires preprocessed FASTQ/FASTA sRNA-Seq data files as input. ncRNA annotation in GTF format with the following tags in the attributes field: gene_name, gene_id, gene_biotype. The outline of abundance estimation adopted by Manatee is provided in **Figure 16**. Mapping of sequencing reads is carried out using bowtie aligner[172]. In the primary phase, reads aligned uniquely to the genome are used to form the UAR clusters across the genome. Multimaps are assigned to loci based on the following approach:

$$f_{split}(x_i, y_i) = \frac{f_{score}(x_i, y_i)}{\sum_{i=1}^{MML} f_{score}(x_i, y_i)} \quad (1)$$

$$f_{score}(x_i, y_i) = \sum_{p=x_i-r}^{y_i+r} f_{cov}(p) \cdot f_{prox}(p) \quad (2)$$

$$f_{prox}(p) = \begin{cases} 1, & x_i \leq p \leq y_i \\ \frac{1}{e^{(x_i-p)/n}}, & x_i - r < p < x_i \\ \frac{1}{e^{(p-y_i)/n}}, & y_i < p < y_i + r \end{cases} \quad (3)$$

where x_i and y_i are the start and end placement positions of the multimap i and r is the range in the close proximity of the read (default 50). Function f_{cov} denotes the UAR density at genomic position p and f_{prox} assigns weights to f_{cov} based on the position p within the genomic region [x_i - r, y_i + r]. The multimap is split across its valid multimapped locations (*MML)* according to the score calculated using function f_{split}. n denotes the relevance of approximate density distribution and is set by default to 10. For multimaps with non-matching annotation and positioning of UAR clusters, annotation is preferred and used to guide the final placement of the reads. If a multimap falls into regions which are annotated completely or partially, all relevant transcripts are noted in the output file in the form of alternative transcripts. In case at least one annotated miRNA is present among those features, the read is assigned to the transcript which exhibits the highest coverage score (ratio):

$$ratio = \frac{coverage \cdot (transcript\ length\ + read\ length)}{transcript\ length \cdot read\ length} \quad (4)$$

Coverage is the number of overlapping nucleotides between the annotated feature (transcript) and the read length. The ratio heuristic prioritizes the annotation with the highest coverage, while considering read and transcript lengths (e.g. reads mapping wholly to mature miRNAs are assigned there instead of on the respective precursor).

A secondary transcriptomic alignment step attempts to rescue reads that exceed the multimapping threshold or that could not be mapped to the genome. In the latter case, the number of allowed mismatches is gradually augmented (maximum default 3). In both cases, reads that can be assigned to transcripts with existing mapping densities calculated in previous steps are assigned to those transcripts. If no expression densities exist, reads are assigned to up to five transcripts exhibiting the highest mapping quality.

Apart from the counts file, a supplementary isomiR output file is provided. It contains each detected putative isomiR sequence along with its estimated read counts and can be used to identify miRNA isoforms, derived from post transcriptional modifications, 5′ and 3′ templated additions or single nucleotide variations. Additionally, a file providing counts for unannotated regions harboring reads is offered. In this file, UARs mapping to loci lacking genomic features are organized into read clusters based on their genomic positions (default: regions with minimum 5 reads and gaps less than 50nt between consecutive reads).

4.2.3 Manatee Benchmarking

4.2.3.1 sRNA-Seq simulation

A simulated short read dataset was created using sample-guided random sampling. Human annotation was derived from Ensembl v85[173], GtRNAdb[8] and miRBase v21[174]. Three sRNA-Seq libraries (Liver tissue: SRR2061810, Gallbladder tissue: ERR842903 and Breast tissue: SRR191548) obtained from Gene Expression Omnibus[175] were employed in the process. Samples were aligned against GRCh38 human reference assembly after 3'-adapter sequences were removed using Cutadapt[171]. Since processed sRNA fragments/features are derived from their precursors by biogenesis/cleavage mechanisms that are distinct to each biotype, simulated reads were designed to follow this rationale. Based on uniquely aligned reads observed in the real data, probability mass functions (PMFs) were created for each biotype describing the read start positions. Eight different PMFs were created for the following types: miRNA, tRNA, Mt-tRNA, rRNA, snRNA, snoRNA, lincRNA and non-coding transcript. Likewise, SNPs and read lengths for each sRNA type were also estimated based on PMFs of UARs. The counts of each read were calculated by drawing from the negative binomial distribution with mean and variance depending on the UARs of the input datasets.

4.2.3.2 Manatee Benchmarks against a Simulated Dataset

The accuracy of Manatee was initially evaluated using a simulated short read dataset. Bowtie v1[172] alignment coupled with HTSeq-Count[176] quantification was used as baseline, while miRge[177], ShortStack[178], and sRNAbench[179] were employed as state-of-the-art approaches in the evaluation. Those sRNA quantification approaches constituted attractive candidates for direct comparison, since miRge performs read alignment entirely against sRNA annotation, ShortStack emphasizes extensively on the genomic alignment of short reads and sRNAbench[179] applies genomic alignment and quantification of each sRNA type in a hierarchical step-wise manner. State-of-the art tools and Manatee were executed under their default settings. Bowtie was executed with 1 allowed mismatch and up to 5 multimaps, while HTSeq-Count transcript quantification was performed using parameters "intersection-nonempty" and "nonunique all" parameter.

Estimated sRNA counts for HTSeq-Count, Manatee, miRge, ShortStack and sRNAbench were contrasted to the ground truth (i.e. simulated counts). Tools tend to over-estimate zero-abundant transcripts (**Figure 17a**, Sim.=0 & Est.>=5). However, the opposite behavior was observed at the other end of the spectrum; expressed, and highly expressed transcripts were not assigned any reads (**Figure 17a**, Sim.>5 & Est.=0). Among the tested tools, counts estimated by Manatee appeared closest to the simulated abundances (**Figure 17b**). Manatee also fared favorably compared to miRDeep2[180], the gold-standard method for miRNAs. miRDeep2 uses Bowtie to map sequencing reads against precursors and discards or assigns multimaps equally to their valid loci and was executed with default settings. Therefore, Manatee runs can quantify still underexplored small RNAs, while sustaining accurate results for miRNAs.

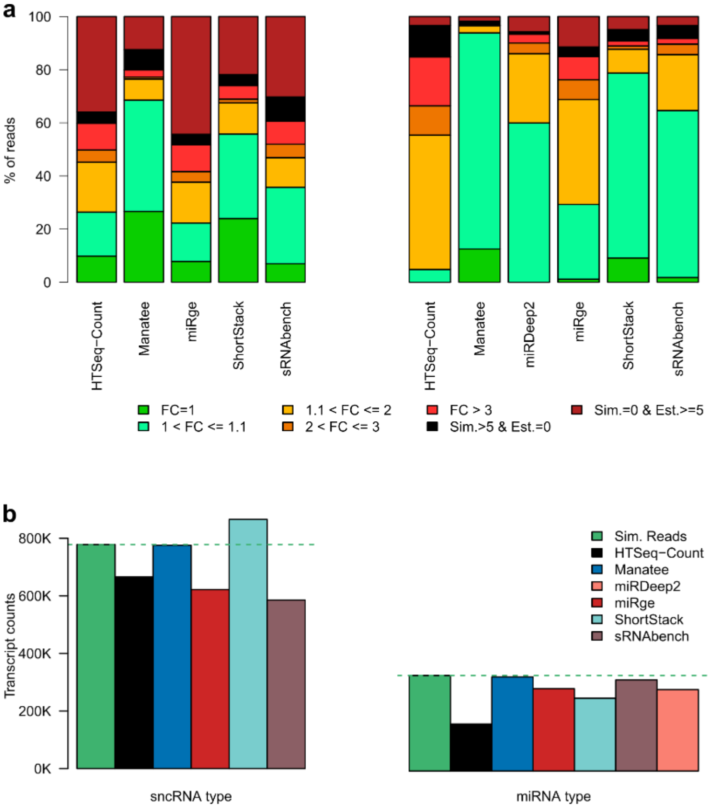


Figure 17. Tools evaluation statistics for simulated sRNA-Seq data. (a) Fold changes for simulated vs. estimated transcript counts for evaluated tools utilizing all small ncRNA species or only miRNAs. Fold change of 1 denotes no difference between the simulated and the calculated counts. Sim.>5 & Est.=0 denotes percentage of reads where the simulated transcript counts > 5 were estimated as zeros by the examined tools. Sim.=0 & Est.>=5 relates with proportion of estimated transcript counts > 5 for which the true simulated count was zero. (b) Comparison between the ground truth count sum of simulated reads and the total estimated transcript counts across implementations. Figure retrieved from Handzlik E et al. (2020)[160].

The sum of simulated transcript counts was contrasted against the estimated counts by the six implementations. ShortStack displayed tendency for count inflation, while miRge, miRDeep2 and sRNAbench underestimated overall transcript counts (**Figure 17b**). Precision metrics were also calculated to assess the performance of the examined algorithms by comparing simulated to estimated read counts for the entire pool of small ncRNA transcripts, as well as for microRNAs only. Root-mean-squared-deviation (RMSD), distance metrics and correlation coefficient values computed for estimated counts versus the ground truth, indicate that Manatee outperforms the other implementations by providing less inflated/deflated transcript counts that are more closely associated with the simulated counts (**Table 1**). A major driving force for this increase in accuracy is the rescue of multimapping reads. Manatee rescue steps enable multimapping assignment to the most probable loci. In contrast, the commonly employed approach of utilizing uniquely aligned reads (e.g. Bowtie run) is the lowest performer with 0.1 and 0.2 Pearson's and Spearman correlation, respectively (**Table 1**).

Table 1. Performance metrics for the accuracy of evaluated sRNA-Seq implementations using simulated data.

Tool	RNA type	RMSD	Jaccard distance	Euclidean distance	Pearson correlation	Spearman correlation
HTSeq-Count	small ncRNAs	372.614	0.298	20439.478	0.798	0.577
Manatee		**341.494**	**0.173**	**15730.981**	**0.879**	**0.796**
miRge		408.08	0.503	26553.417	0.641	0.581
ShortStack		456.499	0.271	20939.313	0.801	0.655
sRNAbench		361.399	0.395	22805.388	0.744	0.556

HTSeq-Count	miRNAs	369.164	0.442	8813.660	0.529	0.392
Manatee		**107**	**0.031**	**2556.831**	**0.929**	**0.954**
miRge		249.67	0.216	6419.008	0.731	0.684
ShortStack		236.657	0.151	5738.626	0.683	0.737
sRNAbench		215.574	0.138	5218.504	0.752	0.725
miRDeep2		155.313	0.078	3867.274	0.893	0.827

4.2.3.3 Manatee Benchmarks against a Real Dataset

Simulated datasets offer the advantage of knowing the ground truth beforehand, still they can lack the complexity observed in real datasets. Lack of a ground truth is recompensated by the capacity to inspect the agreement among implementations in a richer and more realistic set. For this reason, sRNA-Seq data from MCF7 cells (SRR2084358) were used to cross-correlate the compared sRNA/miRNA quantification methods using Pearson correlation (**Figure 18**).

In the total small RNA space, the highest concordance (r = 0.92) was observed for the performance of Bowtie+HTSeq-Count and ShortStack, followed by the Manatee-sRNAbench pair-wise comparison (r = 0.77). For miRNAs, Manatee exhibited > 0.8 correlation coefficient with ShortStack, sRNAbench and miRDeep2, and exhibited the highest correlation (r = 0.94) with miRDeep2, the reference tool in miRNA quantification. miRge and Bowtie+HTSeq-Count displayed similar patterns of underestimating transcript abundance, as occurred in comparisons using simulated data. ShortStack results also appeared on par with the findings of the previous section, where a high number of false positive and negative counts was present. When comparing the total sRNA transcriptome results, a substantial divergence between the estimated counts was observed across executions. These findings may indicate the existence of intrinsic properties in each tool that, in some cases, drive to misclassification and erroneous quantification of sRNAs.

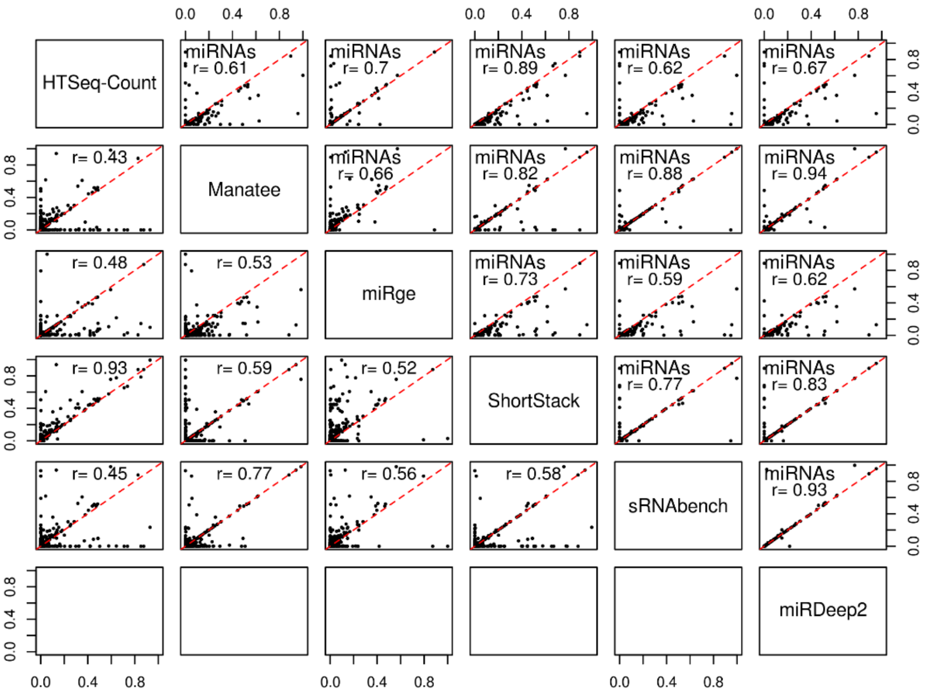


Figure 18. The analyzed sRNA-Seq sample was compared across 5 methods for all sRNA types (lower left panel) and across 6 methods for miRNAs (upper right panel). Pearson correlation was calculated for each pair of compared tools and denoted on each plot with the red line indicating the perfect correlation. Figure retrieved from Handzlik E et al. (2020)[160]*.*

CHAPTER 5 – An Enhanced microRNA Interactome

In subsections **5.1** to **5.4**, the development of DIANA-TarBase and DIANA-LncBase, two reference databases of miRNA targets, is presented.

DIANA-TarBase is devoted to the indexing of experimentally supported protein-coding targets of miRNAs[168]. For the purposes of this work, I performed pre-processing and differential expression analysis of more than 100 microarray datasets. These datasets were derived from cell-lines where experimental perturbation of specific miRNAs (i.e. overexpression or knock-down) was performed to characterize their putative targets. Additionally, I collected pre-analyzed RNA-Seq datasets derived from similar experimental scenarios. I annotated these indirect miRNA-mRNA interactions with publication- and experiment-specific metadata.

The 8th version of DIANA-TarBase was published in the 2018 Database Issue of Nucleic Acids Research (I.F. 11.501) and has until now received 234 citations: D Karagkouni, MD Paraskevopoulou, S Chatzopoulos, IS Vlachos, **S Tastsoglou**, I Kanellos, D Papadimitriou, I Kavakiotis, S Maniou, G Skoufos, T Vergoulis, T Dalamagas and AG Hatzigeorgiou. *DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions*. 2018.

DIANA-LncBase is a database catering miRNA-lncRNA interactions to the scientific community[169]. I performed re-annotation of 7 different microarray platforms to make them compliant to our adopted LncBase v3.0 genome assemblies and transcriptomic annotations, significantly enhancing the representation of non-coding RNAs. I utilized these unique lncRNA-aware annotations during pre-processing and differential expression analysis of more than 80 microarray datasets of miRNA perturbations and derived indirect miRNA-lncRNA interactions. To increase sensitivity of these indirectly-defined interactions, I employed a naïve filter requiring de-regulated lncRNAs to harbor at least one canonical miRNA Recognition Element in their UTR sequence. Additionally, I performed manual annotation of publications assessing miRNA-lncRNA interactions with low-yield techniques. Notably, LncBase interactions derived from AGO-HITS-CLIP and AGO-PAR-CLIP experiments were produced using microCLIP framework. I participated in the development of microCLIP (MD Paraskevopoulou, D Karagkouni, IS Vlachos, **S Tastsoglou** and AG Hatzigeorgiou), by analyzing microarray datasets after microRNA perturbation, which were utilized during its training, testing and evaluation. microCLIP was published in 2018 in Nature Communications and is described in detail in other works[33, 181]. Lastly, I retrieved known variants (dbSNP), clinical variants (ClinVar) and somatic mutations (COSMIC) and used them to annotate all lncRNA MREs in LncBase v3.0 that have been derived from direct experimental methods (e.g. AGO-PAR-CLIP-Seq).

The 3rd version of DIANA-LncBase was published in the 2020 Database Issue of Nucleic Acids Research (I.F. 11.501) and has been cited 11 times until now: D Karagkouni, MD Paraskevopoulou, **S Tastsoglou**, G Skoufos, A Karavangeli, V Pierros, E Zacharopoulou and AG Hatzigeorgiou. *DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts*. 2020.

5.1 TarBase v8.0 Database

Collected data: Approximately 370 publications were manually curated and added. More than 230 high-throughput datasets harboring (in-)direct interactions were collected and/or analyzed. Emphasis was placed on extracting extensive metadata to accompany indexed entries. miRNA-target interactions are coupled with information regarding their publication, applied methodologies, tissues, cell types, experimental conditions and the positive/negative type of regulation. In direct techniques, the exact miRNA binding locations and supplemental information (e.g. cloning primers, targeted transcript region, genomic location) is included. Interactions supported from high-throughput experiments, were extracted either from relevant publications or from the analysis of raw libraries retrieved from GEO[182] and DDBJ[183] repositories.

High-throughput data analysis: High-throughput experiments were analyzed to retrieve gene expression alterations upon specific miRNA perturbation. Raw microarray datasets were processed with a standardized *in-silico* pipeline developed in R[184]. In Affymetrix arrays, Robust Multi-Array Average (RMA) from Bioconductor packages affy[185] or oligo[186] was utilized to perform probe set summarization. Agilent and Illumina microarray data sets were background corrected using normexp method and quantile normalization[187]. Probe sets were mapped to Ensembl gene IDs[188] utilizing chip-specific Bioconductor R packages[189]. Differential expression was assessed with limma[187], using moderated t-statistics and adjusting p-values with Benjamini-Hochberg method. The $\log_2$ fold-change values of probe sets matching the same gene were averaged to calculate its expression alteration. Positive and negative interactions from each set were inferred using a +/-0.5 $\log_2$ fold-change (i.e. 1.585 fold-change) threshold, according to perturbation type. Processed RPF-Seq, RNA-Seq and RIP-Seq libraries, submitted to specific miRNA perturbation were collected from the respective publications. Positive/negative miRNA interactions were formed from genes presenting >10 RPKM and >50% expression change. AGO-CLIP-Seq methodologies were analyzed as described in TarBase v7.0[190].

Database interface development: A new relational schema was designed to host TarBase v8.0 data. Indices were created to guarantee the efficient execution of the system and foreign keys were added to avoid integrity violations. PostgreSQL was utilized to implement the hosting database. The interface was developed using the Yii 2.0 PHP framework. Interactive charts were implemented using the D3.js JavaScript library.

Ranking system: A novel ranking system was incorporated in the interface. miRNA targets are by default sorted according to descending robustness of the respective experimental techniques. In brief, miRNA-gene interactions determined from low-yield experiments are reported first, followed by those derived from high-throughput techniques. Specifically, miRNA-binding events retrieved from reporter gene assays, the gold standard method in miRNA target recognition, are prioritized, pursued 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. Entries 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.

Database interconnections: DIANA-TarBase is integrated in RNAcentral since 2015 [191] and in ENSEMBL[188] since its 6th version. Interactions bearing exact binding locations 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[192] for *in silico* identification of miRNA targets, LncBase v2.0[193] for display of miRNA-lncRNA interactions and DIANA-miRPath v3.0[194] for miRNA functional characterization. miRNA targets from other relevant databases, including miRTarBase[195] and miRecords[196] are also catered.

5.2 TarBase v8.0 Content and Interface

DIANA-TarBase v8.0 caters more than one million entries of experimentally supported miRNA targets. This collection has been derived from more than 33 distinct low-yield and high-throughput techniques, spanning 82 tissues, 510 cell types and ~430 experimental conditions from 18 species (**Figure 19**). Approximately 1,160 publications have been manually curated and more than 330 high-throughput datasets have been analyzed. Version 8.0 incorporates an assortment of positive and negative direct miRNA interactions. It comprises more than 10,000 interactions derived from low-yield techniques. Approximately 5,000 of these miRNA targets are verified by reporter gene assays, extracted from ~900 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 recently emerged CLASH and CLIP-Seq variants, as well as from a previous meta-analysis of published AGO-CLIP datasets[197], have been integrated to the repository. This collection of interactions corresponds to an 8-fold increase compared to the previous release. Approximately 90,000 entries derived from the analysis of AGO CLIP-Seq libraries, have been incorporated in the updated repository. More than 230,000 interactions have been extracted from miRNA-specific transfection/knockdown microarray, RPF-Seq, RIP-Seq and RNA-Seq experiments, performed in 25 tissues and 83 cell types under 220 experimental conditions. Brief metrics of the TarBase v8.0 content are also presented in **Table 2** at the end of **CHAPTER 5 – An Enhanced microRNA Interactome**.

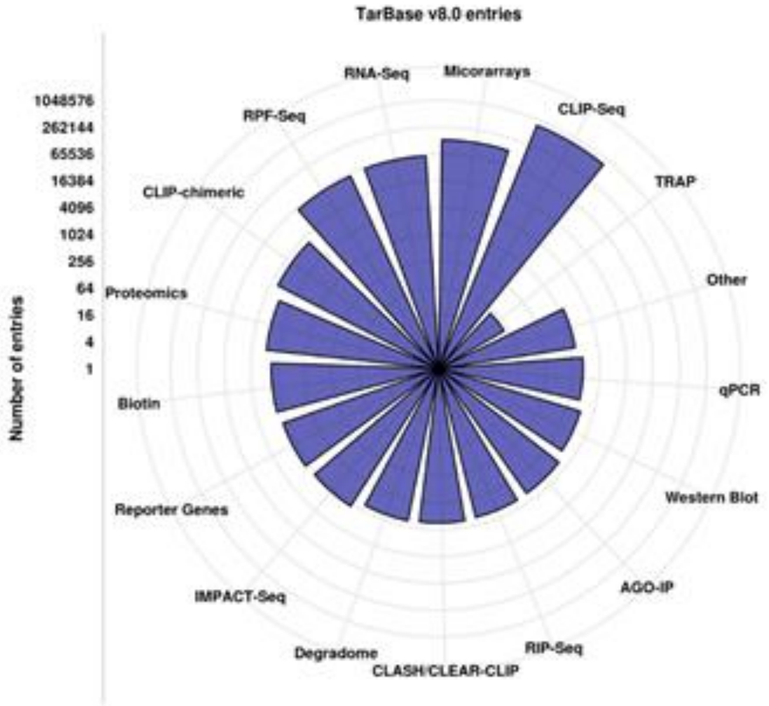
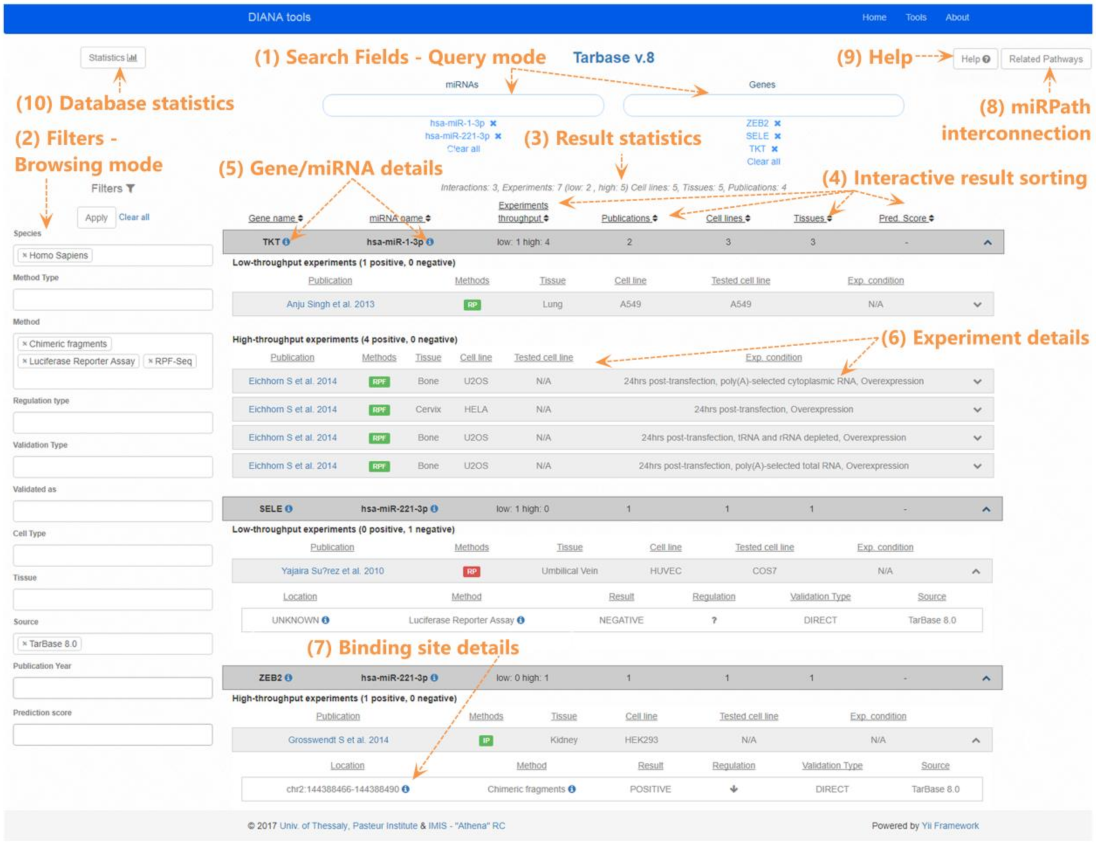


Figure 19. Total miRNA:gene entries incorporated in TarBase v8.0, divided per methodology. Values are plotted in log2 scale. Each grid line corresponds to quadrupling of indexed miRNA interactions. Figure adapted from Karagkouni D et al. (2018)[168].

The interface of TarBase was redesigned to provide an intuitive user-friendly application and flexible querying options (**Figure 20**). Users can retrieve interactions by performing a query with miRNA and/or gene names. Identifiers from Ensembl[188] and miRBase[198] are supported. Positive and/or negative miRNA targets can be retrieved via smart filter combinations. Detailed meta-data including the binding location and experimental conditions are displayed in the relevant sections. A novel aspect in the new interface is the browsing mode, in which users may use filters to retrieve top-quality targets (up to a maximum of 3,000) without applying any specific query.



miRNAs: hsa-miR-1-3p ✖, hsa-miR-221-3p ✖, Clear all

Genes: ZEB2 ✖, SELE ✖, TKT ✖, Clear all

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:

Prediction score:

Interactions: 3, Experiments: 7 (low: 2 , high: 5) Cell lines: 5, Tissues: 5, Publications: 4

Gene name	miRNA name	Experiments throughput	Publications	Cell lines	Tissues	Pred. Score
TKT	hsa-miR-1-3p	low: 1 high: 4	2	3	3	-

Low-throughput experiments (1 positive, 0 negative)

Publication	Methods	Tissue	Cell line	Tested cell line	Exp. condition
Anju Singh et al. 2013	RP	Lung	A549	A549	N/A

High-throughput experiments (4 positive, 0 negative)

Publication	Methods	Tissue	Cell line	Tested cell line	Exp. condition
Eichhorn S et al. 2014	RPF	Bone	U2OS	N/A	24hrs post-transfection, poly(A)-selected cytoplasmic RNA, Overexpression
Eichhorn S et al. 2014	RPF	Cervix	HELA	N/A	24hrs post-transfection, Overexpression
Eichhorn S et al. 2014	RPF	Bone	U2OS	N/A	24hrs post-transfection, tRNA and rRNA depleted, Overexpression
Eichhorn S et al. 2014	RPF	Bone	U2OS	N/A	24hrs post-transfection, poly(A)-selected total RNA, Overexpression

Gene name	miRNA name	Experiments throughput	Publications	Cell lines	Tissues	Pred. Score
SELE	hsa-miR-221-3p	low: 1 high: 0	1	1	1	-

Low-throughput experiments (0 positive, 1 negative)

Publication	Methods	Tissue	Cell line	Tested cell line	Exp. condition
Yajaira Su?rez et al. 2010	RP	Umbilical Vein	HUVEC	COS7	N/A

Location	Method	Result	Regulation	Validation Type	Source
UNKNOWN	Luciferase Reporter Assay	NEGATIVE	?	DIRECT	TarBase 8.0

Gene name	miRNA name	Experiments throughput	Publications	Cell lines	Tissues	Pred. Score
ZEB2	hsa-miR-221-3p	low: 0 high: 1	1	1	1	-

High-throughput experiments (1 positive, 0 negative)

Publication	Methods	Tissue	Cell line	Tested cell line	Exp. condition
Grosswendt S et al. 2014	IP	Kidney	HEK293	N/A	N/A

Location	Method	Result	Regulation	Validation Type	Source
chr2:144388466-144388490	Chimeric fragments	POSITIVE	↓	DIRECT	TarBase 8.0

© 2017 Univ. of Thessaly, Pasteur Institute & IMIS - "Athena" RC

Powered by Yii Framework

Figure 20. Snapshot of 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]. Figure retrieved from Karagkouni D et al. (2018)[168].

5.3 LncBase v3.0 Database

Collected data: miRNA-lncRNA interactions from manual curation of 159 publications and the analysis of >300 high-throughput datasets were incorporated. An auxiliary in-house developed text mining pipeline with full-text capacity was employed to retrieve publications comprising miRNA-lncRNA pairs and terms conveying ceRNA activity and sponge/decoy function. Sorting of publications was performed based on miRNA target existence. Sentences predicted to contain miRNA-lncRNA associations were retained for manual curation. miRNA-lncRNA interactions supported by high-throughput methodologies were extracted from relevant publications and the analysis of raw libraries. Raw datasets were retrieved from publicly available sources; GEO[175], the Encyclopedia of DNA Elements (ENCODE)[163, 199] and DDBJ[200].

CLIP-Seq data analysis: Raw AGO-CLIP-Seq libraries were quality checked using FastQC[201], while adapters/contaminants were detected utilizing in-house developed algorithms and the Kraken suite[202]. Upon pre-processing[203], CLIP-Seq libraries were aligned against the reference genomes (GRCh38 and mm10 assemblies for human and mouse respectively) with GMAP/GSNAP[204] spliced aligner. microCLIP[33] CLIP-Seq-guided algorithm was utilized to identify binding events for the expressed miRNAs. In datasets with more than one replicates, an event had to be present in at least two replicates[205]. Top expressed miRNAs were retrieved either from the relevant publications or from relevant sRNA-Seq library analyses. 88 AGO-PAR-CLIP and 148 AGO-HITS-CLIP libraries were analyzed, yielding ~370,000 binding events, corresponding to 37 cell types, 14 tissues and 90 experimental conditions. ~25,000 lncRNA transcripts had at least one miRNA interaction site. ~90,000 intronic miRNA binding events, partially attributed to the ambiguous splicing of lncRNA transcripts, are appropriately labeled and provided to users. LncBase also indexes 2,220 viral miRNA binding events on host lncRNA transcripts, retrieved from the analysis of 16 virus-infected AGO-PAR-CLIP libraries (from Epstein-Barr virus (EBV) and Kaposi's sarcoma-associated herpesvirus (KSHV)).

Analysis of miRNA perturbation experiments: For the analysis of microarray miRNA perturbation experiments, we adopted procedures from Mercer *et al.* and Liao *et al.*[206, 207] to create custom CDF files for Affymetrix chips Human Genome U133A 2.0, Human Genome U133 Plus 2.0, Human Gene 1.0 ST, Human Exon 1.0 ST, Mouse Genome 430 2.0 and Mouse Gene 1.0 ST. Briefly, probes were aligned to the GRCh38 and mm10 assemblies with Bowtie[172], allowing zero mismatches and multimaps. HTSeq-Count[176] was used to find protein-coding/non-coding genes that each overlap with ≥3 aligned probe sequences. Probes that presented partial overlap with a gene or that overlapped with more than one gene were excluded. Probe set summarization packages were built with makecdfenv package[208]. RMA from packages affy[185] or oligo[186] was employed for probe set summarization and normalization. In cases with replicates, differential expression analysis was performed using moderated t-statistics and FDR correction by limma[187]. A threshold of 1.585 fold-change (FDR < 0.05 where applicable) and the existence of at least one putative canonical binding site for the perturbed miRNA were used as filters to retrieve positive interactions. Eighty-six microarray perturbation experiments were analyzed, corresponding to 70 cell types and tissues. This process enabled the formation of 1,740 and 415 positive miRNA-lncRNA pairs for human and mouse species respectively.

Tissue/cell type lncRNA expression: Raw RNA-Seq datasets were retrieved from ENCODE[163, 199] and GEO[175] repositories, corresponding to 34 distinct cell types and tissues for human and mouse species. RNA-Seq data corresponding to similar cell types and tissues with AGO-CLIP-Seq samples were preferentially selected. Raw datasets were quality checked and preprocessed using FASTQC[201] and Cutadapt[203]. Quantification was conducted at the transcript level, using Salmon[209] version 0.14.1 on quasi-mapping mode and Transcripts Per Million (TPM) values were extracted. 48 whole transcriptome libraries, corresponding to 22 cell types/tissues were analyzed. Transcripts with TPM > 1 were retained, while median TPM values were estimated in case of more than one biological replicates. For characterization of the subcellular localization of transcripts, 55 libraries from RNA-Seq experiments, conducted separately in nucleus and cytoplasm in 15 distinct cell types/tissues, were pre-processed. We retained transcripts with TPM > 1 in at least one of the two subcellular compartments. We adopted the Relative Concentration Index (RCI)[210], estimated by transforming the cytoplasmic-to-nuclear TPM fraction into $\log_2$ scale, to define the trend of lncRNA transcripts localization towards the two different cellular compartments. Human transcriptomes were compiled from ENSEMBL 96[211], RefSeq 109[212] and Cabili *et al.*[213], as well as mouse transcriptomes derived from ENSEMBL 96[211] and RefSeq 106[212].

Annotating MREs with known variants: LncBase v3.0 integrates short variant information on experimentally supported miRNA binding events on non-coding transcripts. Variants located on MREs may induce both disruption of these sites and loss of the respective miRNA-lncRNA interactions. Binding sites retrieved from AGO-CLIP-Seq experiments and miRNA-lncRNA chimeric fragments, were intersected with **(a)** ~37 million common variations from dbSNP build 151[214], **(b)** 498,490 variants with clinical annotation from ClinVar[215] and **(c)** ~26 million somatic mutations from COSMIC v90[216]. The 9,318 MREs that were annotated as variant-related, are located on 4,220 human lncRNA transcripts and were associated with 10,968 unique variants, composing a set of 13,831 variant-MRE pairs. Specifically, 49% (6,731) of those pairs are associated with common variants, 47% (6,574) with somatic mutations and 4% (526) with ClinVar variants.

Advanced visualizations: DIANA-LncBase v3.0 also provides interactive visualization plots, implemented using the D3.js JavaScript library. The user can explore **(a)** the clustering of cell types and tissues, based on CLIP-Seq-derived miRNA-lncRNA interactions, **(b)** bar-plots portraying the expression profiles of lncRNAs within the cell and **(c)** in different subcellular compartments, among distinct cell types.

Database interface development: In the advanced relational schema of the database new indices were created in PostgreSQL to ensure quick query execution. A new backend was developed using Java Spring framework and .NET Core 2.2. The database interface was redesigned using Angular v.8 and enhanced to provide an intuitive user-friendly application.

Database inter-connections: Since 2015, DIANA-LncBase is integrated in RNAcentral[191]. Interactions per miRNA can be viewed in a dedicated page provided by the repository. LncBase v3.0 is interconnected with other DIANA-tools, including TarBase[168], microT-CDS and miRPath v3.0[217]. Predicted miRNA-lncRNA pairs are also supported by LncBase v2.0[205].

5.4 LncBase v3.0 Content and Interface

DIANA-LncBase v3.0 indexes approximately half a million entries, corresponding to the largest collection of experimentally supported cell type and tissue specific miRNA-lncRNA interactions. Incorporated interactions were defined by 15 distinct low-yield and high-throughput methodologies, corresponding to 192 cell types, 52 tissues and 162 experimental conditions. >730 miRNA-lncRNA entries were manually curated, while 2,094 interactions were extracted from the re-analysis of miRNA-specific transfection/knockdown microarray experiments. Most entries correspond to AGO-CLIP-derived miRNA-binding events. LncBase v3.0 incorporates 2,924 miRNA-lncRNA chimeric fragments, while >235,000 interactions have been retrieved from the re-analysis of 236 AGO-CLIP-Seq datasets with a robust CLIP-Seq-guided algorithm. The number of miRNA-lncRNA interactions per tissue and miRNA species, retrieved from direct high-throughput techniques, accompanied by the distribution of interactions in the different lncRNA categories, is depicted in **Figure 21**. Eighty-five percent (±10%) of miRNA targets is classified to the main categories of lncRNAs (sense, antisense, intergenic) and to pseudogenes.

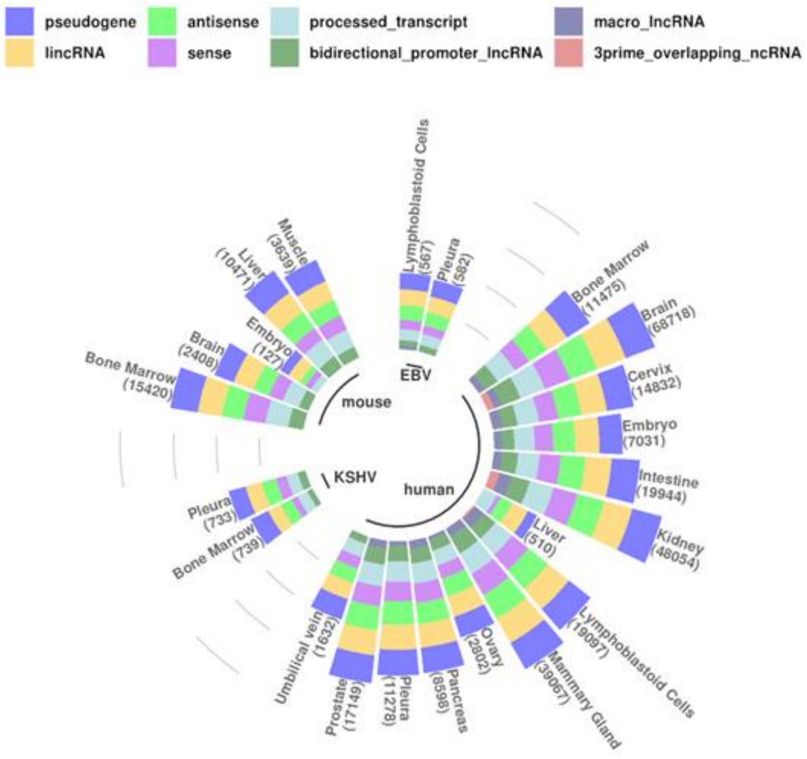
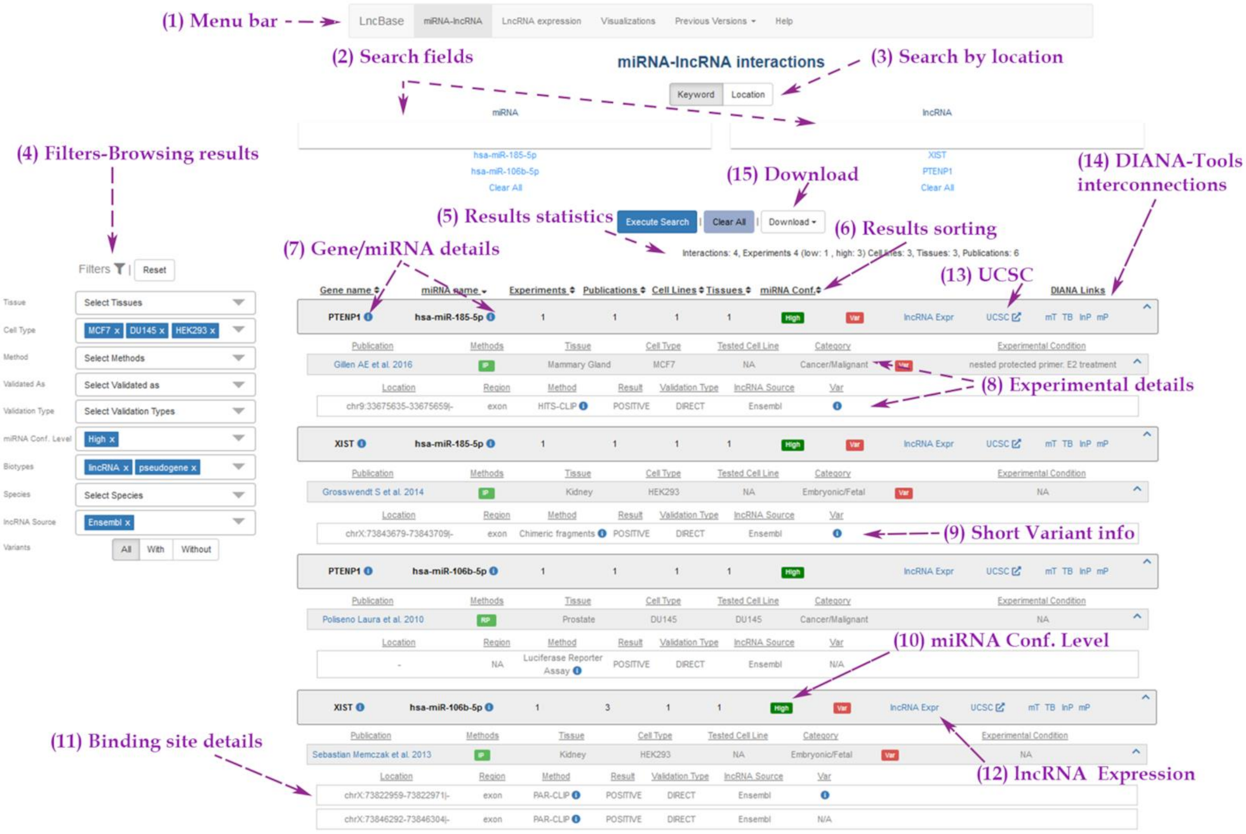


Figure 21. miRNA-lncRNA interactions derived from direct high-throughput techniques per tissue and miRNA species. 85 ± 10% of interactions is spatially classified to sense, antisense, lincRNAs and pseudogenes. Values are plotted in $\log_2$ scale. Figure adapted from Karagkouni et al. (2020)[169] for the purpose of this thesis.

DIANA-LncBase v3.0 provides two modules. The main module presents the experimentally supported interactions and a supplemental module enables tissue-specific queries of lncRNA expression. The two modules are inter-connected to easily direct users querying interactions to inspect the expression of lncRNAs under study and vice versa.

Users can retrieve interactions by **(a)** performing queries with miRNA and/or gene names – identifiers from Ensembl[211], miRBase[218], RefSeq[212] and the Cabili *et al.* reference lncRNA study[213] are also supported -, **(b)** applying different combinations of the filtering criteria including species, cell types/tissues and methodologies, **(c)** searching a specific genomic location for the presence of MREs on lncRNA transcripts (**Figure 22**). Filtering combinations include among others cell type/tissue, experimental methodology, transcript category, species and lncRNA annotation source, as well as "miRNA confidence level" indication, a latest miRBase feature, and information about known short variants on MRE regions. Detailed meta-data escort the interactions (e.g. exact miRNA binding location, experimental conditions and detailed variant information where applicable, i.e. known variant identifier, reference and alternative allele, the exact overlapping position with the MRE region and link to the original variant source). miRNA binding events per miRNA-lncRNA interaction, coupled with the MRE-overlapping variant genomic locations, can be visualized in an interactive UCSC genome browser[219] graphic, where users can combine them with all the UCSC-amassed functionalities.

LncRNAs' expression can be explored either via an inter-connected link in the module of experimentally supported interactions, or by applying a query with one or more ncRNA transcripts in the dedicated "lncRNA expression" page (**Figure 23**). Importantly, **(a)** the overall expression profiles of lncRNAs ("Expression" mode) and **(b)** a comparison between the nuclear and cytoplasmic subcellular compartment fractions ("Localization" mode), can be queried for a wide range of human and mouse cell types. Transcript-per-million (TPM) values describing the expression of lncRNAs are provided (in case of more than one biological replicates, median TPM is specified). Specifically, in the "Expression" mode the user can also retrieve results by selecting a particular range of TPM values, described as "Low" (range:1-10), "Medium" (range:11-600) and "High" (range:>600). In "Localization" mode TPM values, estimated separately in nucleus and cytoplasm, are provided, followed by the RCI metric and the apparent inclination of the sub-localization of lncRNAs, either towards the nucleus or the cytoplasm. At any time, users can navigate to the module of experimentally supported targets via a dedicated link.



miRNA: hsa-miR-185-5p, hsa-miR-106b-5p (Clear All)

lncRNA: XIST, PTENP1 (Clear All)

Tissue	Select Tissues
Cell Type	MCF7 x DU145 x HEK293 x
Method	Select Methods
Validated As	Select Validated as
Validation Type	Select Validation Types
miRNA Conf. Level	High x
Biotypes	lincRNA x pseudogene x
Species	Select Species
lncRNA Source	Ensembl x
Variants	[✓ All] With Without

Interactions: 4, Experiments 4 (low: 1 , high: 3) Cell lines: 3, Tissues: 3, Publications: 6

Gene name	miRNA name	Experiments	Publications	Cell Lines	Tissues	miRNA Conf.				DIANA Links
PTENP1	hsa-miR-185-5p	1	1	1	1	High	Var	lncRNA Expr	UCSC	mT TB lnP mP

Publication	Methods	Tissue	Cell Type	Tested Cell Line	Category		Experimental Condition
Gillen AE et al. 2016	IP	Mammary Gland	MCF7	NA	Cancer/Malignant	Var	nested protected primer. E2 treatment

Location	Region	Method	Result	Validation Type	lncRNA Source	Var
chr9:33675635-33675659|-	exon	HITS-CLIP	POSITIVE	DIRECT	Ensembl	●

Gene name	miRNA name	Experiments	Publications	Cell Lines	Tissues	miRNA Conf.				DIANA Links
XIST	hsa-miR-185-5p	1	1	1	1	High	Var	lncRNA Expr	UCSC	mT TB lnP mP

Publication	Methods	Tissue	Cell Type	Tested Cell Line	Category		Experimental Condition
Grosswendt S et al. 2014	IP	Kidney	HEK293	NA	Embryonic/Fetal	Var	NA

Location	Region	Method	Result	Validation Type	lncRNA Source	Var
chrX:73843679-73843709|-	exon	Chimeric fragments	POSITIVE	DIRECT	Ensembl	●

Gene name	miRNA name	Experiments	Publications	Cell Lines	Tissues	miRNA Conf.				DIANA Links
PTENP1	hsa-miR-106b-5p	1	1	1	1	High		lncRNA Expr	UCSC	mT TB lnP mP

Publication	Methods	Tissue	Cell Type	Tested Cell Line	Category		Experimental Condition
Poliseno Laura et al. 2010	RP	Prostate	DU145	DU145	Cancer/Malignant		NA

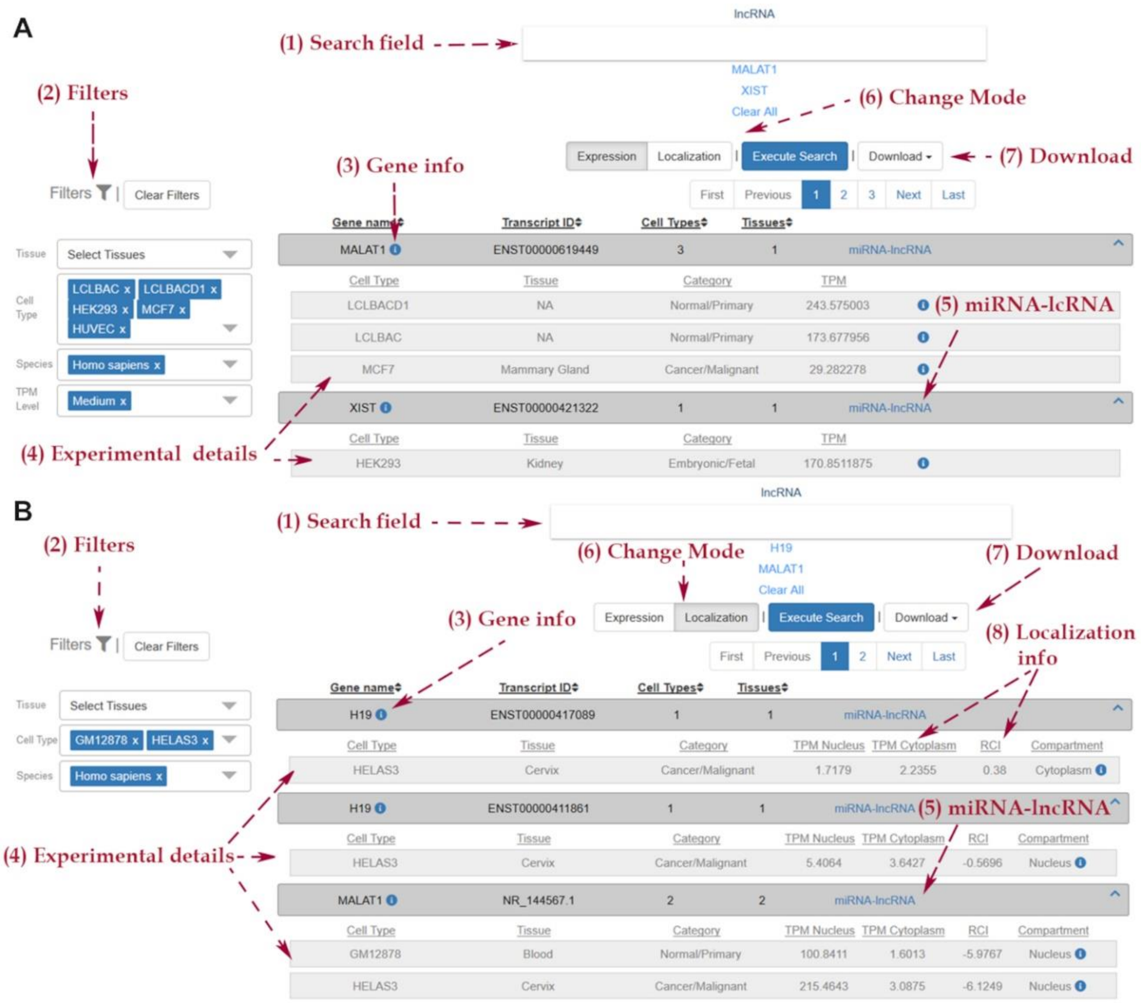
Location	Region	Method	Result	Validation Type	lncRNA Source	Var
-	NA	Luciferase Reporter Assay	POSITIVE	DIRECT	Ensembl	N/A

Gene name	miRNA name	Experiments	Publications	Cell Lines	Tissues	miRNA Conf.				DIANA Links
XIST	hsa-miR-106b-5p	1	3	1	1	High	Var	lncRNA Expr	UCSC	mT TB lnP mP

Publication	Methods	Tissue	Cell Type	Tested Cell Line	Category		Experimental Condition
Sebastian Memczak et al. 2013	IP	Kidney	HEK293	NA	Embryonic/Fetal	Var	NA

Location	Region	Method	Result	Validation Type	lncRNA Source	Var
chrX:73822959-73822971|-	exon	PAR-CLIP	POSITIVE	DIRECT	Ensembl	●
chrX:73846292-73846304|-	exon	PAR-CLIP	POSITIVE	DIRECT	Ensembl	N/A

Figure 22. Snapshot depicting the DIANA-LncBase v3.0 interface. Users can explore different database modules and interactive visualizations through a dedicated menu bar (1). They can retrieve interactions by querying with miRNA and/or gene names (2), genomic location (3), and/or by applying different filtering combinations (4). Interactions can be refined with a series of filtering options including cell type/tissue, experimental methodology, transcript category, species and lncRNA annotation source1 (4). Result statistics are promptly calculated (5). Interactions can be also sorted in ascending or descending order (6). Gene/miRNA details are complemented with active links to Ensembl, RefSeq, miRBase and RNAcentral(7). Additional details regarding the experimental procedures(8), variant information, where applicable (9), as well as miRNA confidence level indication are provided(10). Interactions are accompanied by miRNA-binding site details (11). Inter-connection with the lncRNA expression module is provided (12). miRNA binding events and MRE-overlapping variant genomic locations can be visualized in an interactive UCSC genome browser (13). Links to other DIANA-Tools are also available (14). Figure retrieved from Karagkouni et al. (2020)[169]*.*



Gene name	Transcript ID	Cell Types	Tissues	
MALAT1	ENST00000619449	3	1	miRNA-lncRNA

Cell Type	Tissue	Category	TPM
LCLBACD1	NA	Normal/Primary	243.575003
LCLBAC	NA	Normal/Primary	173.677956
MCF7	Mammary Gland	Cancer/Malignant	29.282278

Gene name	Transcript ID	Cell Types	Tissues	
XIST	ENST00000421322	1	1	miRNA-lncRNA

Cell Type	Tissue	Category	TPM
HEK293	Kidney	Embryonic/Fetal	170.8511875

Gene name	Transcript ID	Cell Types	Tissues	
H19	ENST00000417089	1	1	miRNA-lncRNA

Cell Type	Tissue	Category	TPM Nucleus	TPM Cytoplasm	RCI	Compartment
HELAS3	Cervix	Cancer/Malignant	1.7179	2.2355	0.38	Cytoplasm

Gene name	Transcript ID	Cell Types	Tissues	
H19	ENST00000411861	1	1	miRNA-lncRNA

Cell Type	Tissue	Category	TPM Nucleus	TPM Cytoplasm	RCI	Compartment
HELAS3	Cervix	Cancer/Malignant	5.4064	3.6427	-0.5696	Nucleus

Gene name	Transcript ID	Cell Types	Tissues	
MALAT1	NR_144567.1	2	2	miRNA-lncRNA

Cell Type	Tissue	Category	TPM Nucleus	TPM Cytoplasm	RCI	Compartment
GM12878	Blood	Normal/Primary	100.8411	1.6013	-5.9767	Nucleus
HELAS3	Cervix	Cancer/Malignant	215.4643	3.0875	-6.1249	Nucleus

Figure 23. Snapshot depicting the interface of lncRNA expression profile dedicated page. Users can retrieve expression profiles of lncRNAs within the cell (a) and comparatively between the nuclear and cytoplasmic subcellular compartments (b). They can explore lncRNA abundance by performing queries with gene and/or transcript names (a-1, b-1), as well as combination of tissues and cell (a-2, b-2). Gene details, complemented with active links to Ensembl and RefSeq, are provided (a-3, b-3). TPM values describing the expression of lncRNAs, accompanied with experimental details are catered to users (a-4). Links directing to the experimentally supported targets module are provided (a-5, b-5). In "Localization" mode, expression TPM values, followed by the RCI value (b-4), are provided separately in nucleus and cytoplasm. Apparent inclination of the sub-localization of lncRNAs is indicated (b-7). Users can easily swap between "Localization" and "Expression" modes and retrieve lncRNA abundance without performing new queries (a-6, b-6). Figure retrieved from Karagkouni et al. (2020)[169].

Table 2. Summary of TarBase v8.0 and LncBase content.

	Metric – Characteristic	TarBase v8.0	LncBase v3.0
Database	**Total entries**	>1,070,000	>500,000
	Entries (low-yield methods)	10,161	242
	Entries (high-throughput methods)	~1,060,000	~239,000
	Cell types	510	192
	Tissues	82	51
	Publications	1,165	236
Analyzed high-throughput datasets	**Datasets**	335	322
	Conditions	212	150
	Publications	98	79
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	Reporter genes, northern blot, qPCR, **RIP-qPCR**, biotin miRNA tagging, CLIP-Seq, **CLEAR-CLIP**, **CLIP-chimeric**, **miR-CLIP**, AGO-IP, RNA-Seq, microarrays
Target expression information	Datasets (cell)	MeSH-derived expression information instead	48
	Datasets (nucleus/cytoplasm)		55
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	**Re-designed interface**, support of specific queries, **browsing results by different cell type/tissue combinations**, search by location, enhanced filtering options (**transcript biotype, miRNA confidence level, short variant information on MREs**), **customizable sorting of results, statistics**, detailed meta-data, **a dedicated module for lncRNA expression profiles,** DIANA-tools interconnection, UCSC graphical support, **RNAcentral integration**

CHAPTER 6 – RNA Editing Bioinformatics

Subsections **6.1** and **6.2** present my progress in developing an R pipeline for miRNA editing identification from sRNA-Seq datasets and some primary findings regarding miRNA editing events on brain tissues.

Preliminary results of this work have been presented in the 2019 Hellenic Computational Biology and Bioinformatics Conference (Patras, Greece): **S Tastsoglou** and AG Hatzigeorgiou. *Detection and functional analysis of edited microRNAs in small RNA sequencing experiments*. Poster presentation.

A relevant manuscript is currently under preparation for submission: **S Tastsoglou**, M Miliotis, N Vakirlis, D Karagkouni, G Skoufos and AG Hatzigeorgiou. *Commonalities and differences in miRNA editing events in the human brain*.

6.1 RNA Editing Identification in small RNA Sequencing Datasets

The developed miRNA editing identification pipeline is presented in **Figure 24**. The main identification operations were written in R employing efficient genomics R packages, such as GenomicRanges[220] and Biostrings[221]. System calls to external applications are also made via R. Each pipeline component is developed as an RNA-editing-tailored script unit that can either be executed and parameterized separately, or as part of the pipeline.

As stated in the seminal miRNA editing publication by Alon *et al.*[137], pre-processing steps are paramount for the elimination of potential false positives. Raw reads are quality checked with FASTQC[201] and search for adapters and contaminants with Minion[222]. Adapter and contaminant removal is performed with trim-galore tool[223], allowing reads between 15 and 28 bases, a minimum 20 Phred quality score in the 3′-end. Additionally, ShortRead[224] operations are employed to filter out reads carrying more than 3 bases with Phred score below 20. Bowtie[172] alignment is performed on the genome in order to avoid mapping biases introduced by straight transcriptomic alignment (i.e. the alignment of reads on miRNA locations while they could be mapped with equal or better scores elsewhere). Reads are aligned at "--best --strata -trim3 2" allowing up to 1 mismatch and only one mapping position. The trimming parameter is set so that spurious mismatches at the 3′end of reads will not be taken into account, since 3′ non-template additions are common in mature miRNA sequences. BAM alignment files are used with miRBase annotation to perform HTSeq-Count-based quantification. Although this option is not optimal for generic miRNA quantification[160], it enables miRNA counts and the total counts of bases used for miRNA editing identification to not be discordant.

During the main identification steps, Rsamtools and GenomicAlignments R packages[225] are used to import per-chromosome slices of the BAM file in a parallel fashion. Each process carries out calculation of the nucleotide frequencies of all miRBase v22[218] positions that carry at least one mismatch. Mismatch counts include only bases with Phred score at least 30, signifying a 10^{-3} sequencing error rate. Results from all processes are collected and combined in a "data.frame" and pre-miRNA annotation of the mismatch positions is performed with GenomicRanges. This enables **(a)** providing richer metadata in the editing results and, more importantly, **(b)** correcting for strand: unlike non-stranded RNA-Seq ambiguity, we are confident that T-to-C mismatches found in a region where a pre-miRNA is annotated on the minus strand are actually A-to-G mismatches (evidence of potential A-to-I editing events), since sRNA-Seq is a stranded experimental protocol. Once strand correction is made, the per-base frequencies are calculated and binomial tests are performed to filter out events that could be sequencing errors. Conservative, Bonferroni adjustment is applied to the tests' p-values. Metainformation regarding the relative pre-miRNA and mature miRNA coordinates of RNA editing positions, as well as secondary structure information (i.e. whether edited base was single- or double-stranded) are calculated and the Primary Editing Events are exported.

Importantly, variant filtering is incorporated in the pipeline; if matching sets of variants for the sRNA-Seq dataset are available, they can be either collapsed with known variants (here, dbSNP 151[214] and COSMIC 90[216] were included for known variants and somatic mutations respectively) or used solitarily, enabling sample-specific (e.g. patient- or tissue-specific) variant filtering.

Primary editing events and variant information are combined enabling either the provision of editing events with annotated events for more in-depth analysis, or of editing events with the variant positions filtered out. A coverage-frequency filter is applied, which keeps editing events with at least 5% frequency, in mature miRNAs that exhibit at least 1 RPM (Reads Per Million mapped reads), or in precursor (i.e. non-mature regions) that are covered by at least 10 reads (e.g. equals 1 RPM in a sRNA-Seq dataset with 10 million mapped reads).

Depending on the experimental setup and research question, multi-sample experiments might require the combination of results derived from more than one sample. An auxiliary script, tailored to the pipeline's default output, has been created for this purpose. It annotates editing events derived from multiple samples and outputs filtered/annotated results.

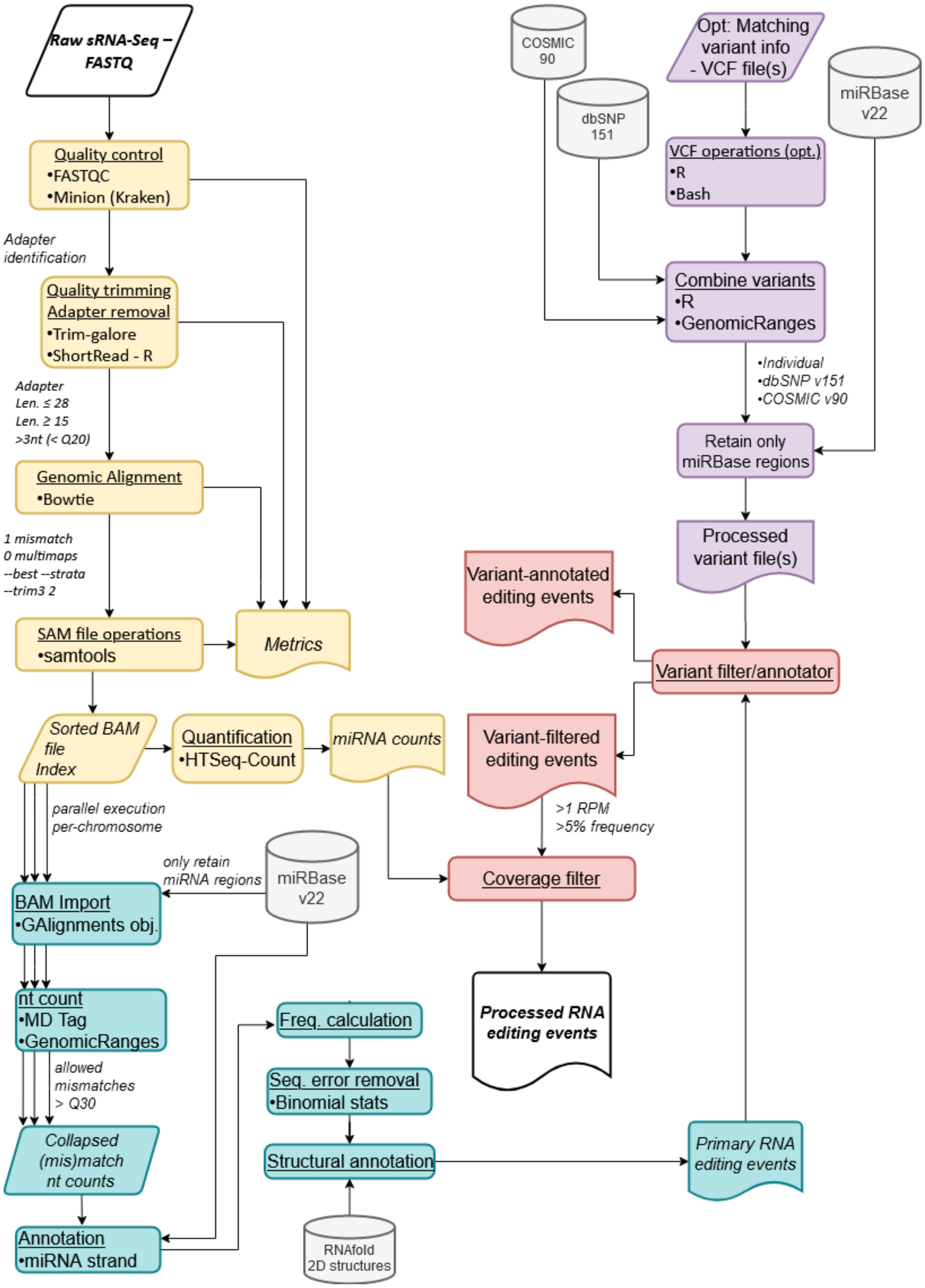


Figure 24. The developed RNA editing identification pipeline. Analysis starts with sRNA-Seq pre-processing steps that are tailored to RNA editing analysis and miRNA quantification (yellow flowchart components). Operations to combine variant information that matches the sRNA-Seq sample, if available, with variants from existing repositories result in a merged file containing variants in microRNA genomic loci (purple components). The core RNA editing components (turquoise) are developed in R and perform parallel per-chromosome operations starting from the genomic alignments of sRNA-Seq reads. Primary RNA editing events are contrasted with matching variants and overlapping positions may be annotated or filtered out, while sequencing coverage and editing frequency filters are applied (red

components). Files of processed RNA editing events (A-to-I, C-to-U and total events) are given as the final output. Figure created for the purpose of this thesis.

6.2 Aspects of the microRNA Editome in the Human Brain

6.2.1 Case Studies of miRNA Editing Analysis

For the purposes of this thesis, 6 distinct datasets were employed for miRNA editing analysis.

Brain samples were chosen based on numerous reports of increased RNA editing in the neural tissue. EBV-transformed lymphocyte sRNA-Seq datasets (n = 452) derived from the GEUVADIS Project were also incorporated, because they were coupled with WGS-derived (Whole Genome Sequencing) variants depicting phased genotypes of each individual[226]. An auxiliary script was created to separate individuals' variants from a collapsed VCF file containing phased-level variants from all 1000 Genomes WGS samples. Brief summarized metrics of the analyzed datasets are provided in **Table 3**. The relatively low mapping rates are due to the extremely stringent alignment parameters required for RNA editing identification and correspond to allowed alignments; the median percent of reads mapping to any genomic location (i.e. including all multimapper reads) exceeded 98% in the analyzed sets.

Studies of healthy brain samples were not complemented with matching WGS datasets, therefore only known variants available in dbSNP and somatic mutations from COSMIC were filtered out. Each GEUVADIS sample was filtered using variants from its matching WGS experiment, in combination with dbSNP and COSMIC positions. TCGA-LGG datasets were also coupled with somatic variants derived from Whole Exome Sequencing experiments. These variants were incorporated in a sample-specific fashion in the filtering step with dbSNP and COSMIC, however WXS coverage in pre-miRNA regions is very suboptimal and resulted in very few sample-specific variants.

Table 3. Summarized metrics of the sRNA-Seq datasets incorporated in the RNA editing analysis.

Study ID	Tissue	Subregion	Condition	Sample number	Raw reads (million, median)	Processed reads (million, median)	Mapped reads (million, median)	Mapping rate
SRP221185	Brain	Cerebellum	Healthy	1	20.05	15.64	5.90	37.71%
SRP052236	Brain	Prefrontal cortex	Healthy	30	15.22	9.51	3.70	38.88%
SRP063627	Brain	Prefrontal cortex	Healthy	16	10.96	7.06	2.97	42.10%
SRP174906	Brain	Hippocampus	Healthy	5	16.49	9.71	4.31	44.34%
GEUVADIS	Blood	Lymphocytes	Healthy	452	8.78	3.63	1.33	36.63%
TCGA-LGG	Brain	Glial tissue	Lower Grade Glioma	530	9.17	8.59	2.20	25.61%

6.2.2 Total mismatch counts reflect RNA editing activity

The global miRNA editing landscape of each dataset was inspected by calculating total numbers of **(a)** mismatch bases **(b)** and mismatch events. A-to-G and C-to-T mismatches (i.e. those corresponding to potential A-to-I and C-to-U editing events respectively were calculated and compared among tissues and conditions and editing commonalities/differences between studied conditions and samples were highlighted using bar plots and editing percentage heat maps. Prominent events occurring in at least 30% of a dataset's samples were pinpointed.

Analysis of identified editing events was performed discriminating ADAR-mediated (A-to-I) from APOBEC-mediated (C-to-U) event cases. In **Table 4**, **(a)** total and **(b)** consistent A-to-I and C-to-U event counts are presented. Events were defined as consistent when they were exhibited in at least 30% of a dataset's samples. Despite the measures taken to account for sequencing errors and the existence of DNA-level variation in potential editing sites, the possibility of yielding false positives still exists; requiring consistency of RNA events across samples increases confidence against both aspects and is also indicative of a functional role for the remaining events. The robustness of this threshold is quite evident in same-tissue samples from different studies (i.e. in prefrontal cortex and glioma samples), where, despite the difference in sample size (30 *vs.* 16 and 512 *vs.* 18 respectively), consistent event sums remain the same.

Among healthy samples, consistent A-to-I editing positions were the fewest in GEUVADIS Lymphocyte Cell-Lines (LCL), indicating a limited catalytic activity of ADAR in this cell-line. They were also less pronounced in glioma samples, compared to healthy neural tissue and this has been documented for glioblastoma multiforme[227], a glioma subtype that is more aggressive than lower grade glioma. Importantly, equal numbers of consistent editing events were found in prefrontal cortex samples from two independent studies and in hippocampus samples (n = 19). Consistent A-to-I events were also close between primary (n = 8) and recurrent (n = 7) LGG samples.

Table 4. Counts of total and consistent (i.e. found in at least 30% of samples) A-to-I and C-to-U miRNA editing events. The percent of consistent-over-total events is also provided.

					A-to-I Events			**C-to-U Events**		
Study ID	**Tissue**	**Subregion**	**Condition**	**Samples (n)**	**Total**	**Consistent**	**Cons. %**	**Total**	**Consistent**	**Cons. %**
SRP221185	Brain	Cerebellum	Healthy	1	31	-	-	47	-	-
SRP052236	Brain	Prefrontal cortex	Healthy	30	50	19	38.0%	63	25	39.7%
SRP063627	Brain	Prefrontal cortex	Healthy	16	35	19	54.3%	37	22	59.5%
SRP174906	Brain	Hippocampus	Healthy	5	32	19	59.4%	56	46	82.1%
GEUVADIS	Blood	Lymphocyte cell-lines	Healthy	452	82	4	4.9%	148	13	8.8%
TCGA-LGG	Brain	Glial tissue	Primary tumor	512	60	8	13.3%	121	10	8.3%
			Recurrent tumor	18	22	7	31.8%	37	12	32.4%

C-to-U events were more abundant than A-to-I events in all datasets and particularly consistent in hippocampus samples (82.1% events consistent). However, this abundance reflects the number of positions in which RNA editing occurs (at frequency greater than 5% and RPM > 1). It indicates neither a high frequency nor a high expression of edited miRNAs. Therefore, the extent of RNA editing was evaluated by contrasting evaluating raw numbers of mismatches for each study (**Figure 25**). Using this approach, prominence of A-to-I over C-to-U editing was depicted clearly in healthy brain samples. When the same method was employed per-sample, the same signature was presented. It should be noted, however, that use of raw mismatch numbers does not permit absolute cross-study comparisons, due to intra-study differences in sample numbers (i.e. biological replicates), as well as in sRNA-Seq depth.

In the absence of robust sample-specific variants in pre-miRNA regions, primary and recurrent lower grade glioma samples did not present the same prominence in A-to-I editing (**Figure 26**). However, even LGG datasets are discernible from the GEUVADIS lymphocyte dataset (n = 452), regarding A-to-G and C-to-T mismatch numbers (**Figure 27**).

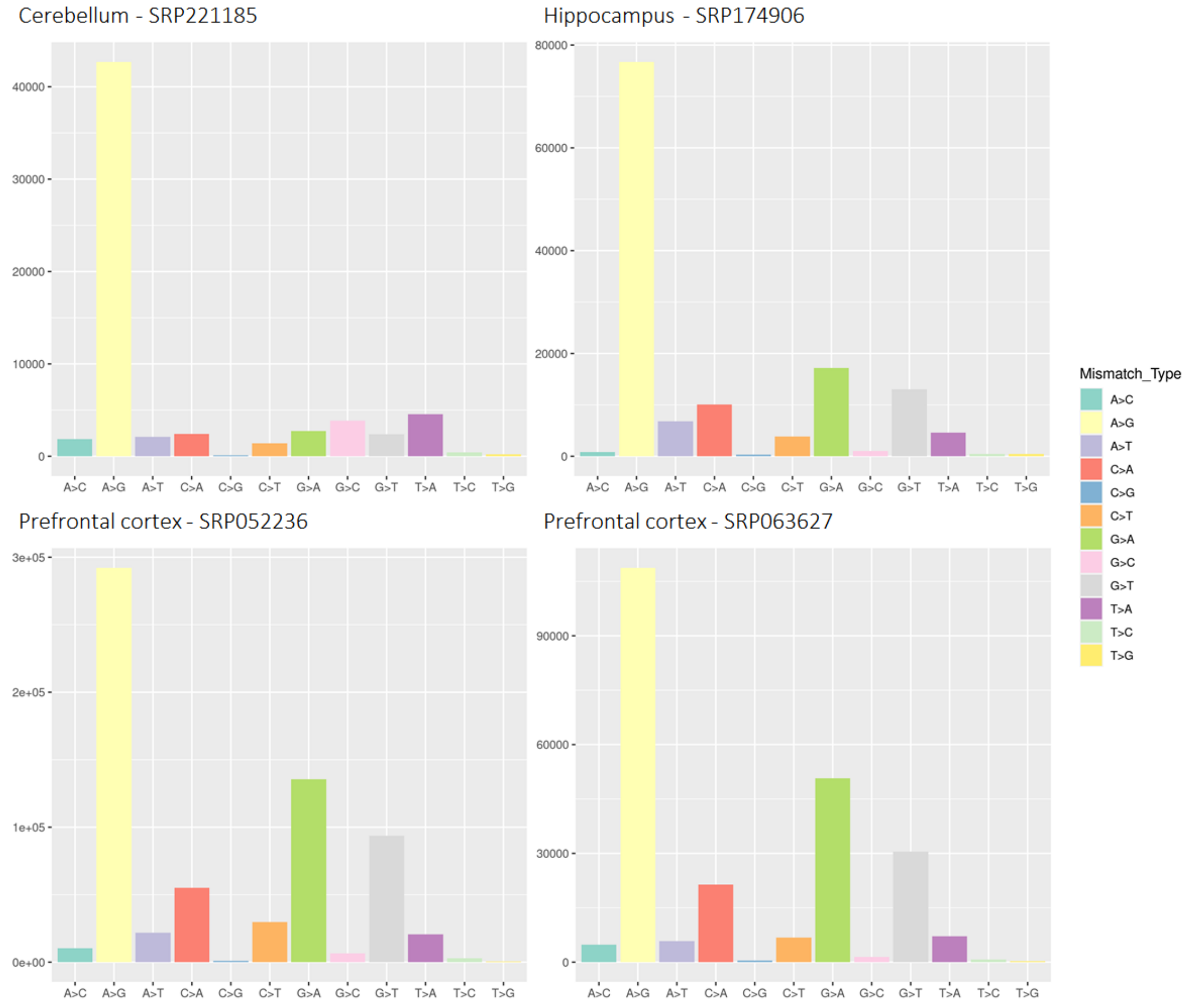


Figure 25. The RNA editing signature. Sums of each mismatch type from all datasets in four sRNA-Seq studies of healthy brain tissue. Figure created for the purpose of this thesis.

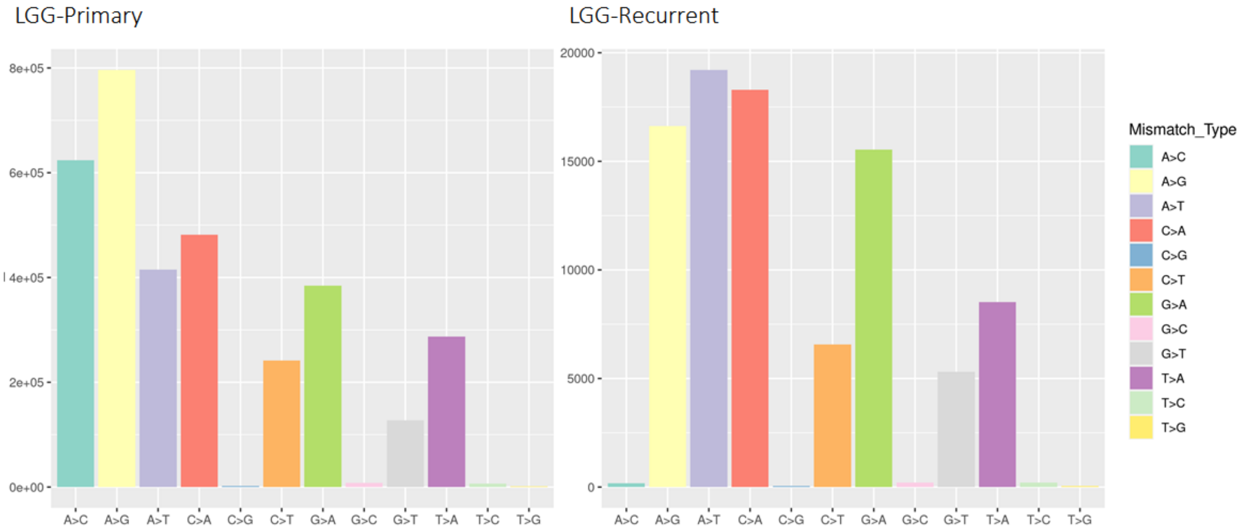


Figure 26. Sum of mismatches per mismatch type for TCGA Lower Grade Glioma samples. Primary (n = 512) and recurrent (n = 18) tumor metrics are presented separately. Figure created for the purpose of this thesis.

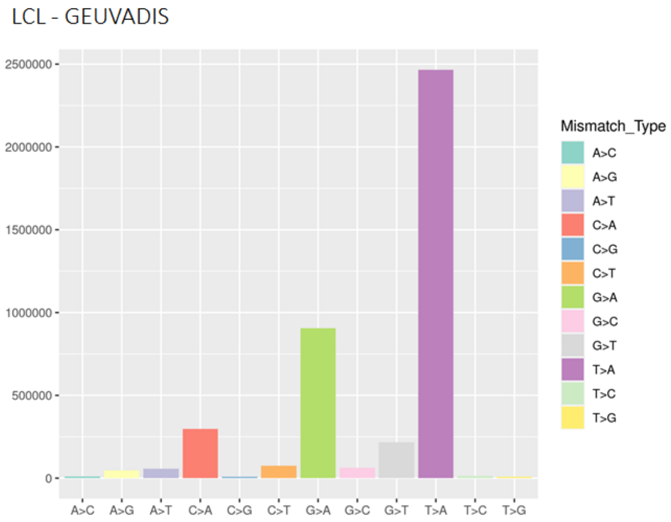


Figure 27. Sum of mismatches pre mismatch type for GEUVADIS LCL cell line dataset (n = 452). Minimal evidence of RNA editing activity is exhibited in lymphocyte miRNA regions. Figure created for the purpose of this thesis.

6.2.3 miRNA Editing Commonalities and Differences in the Human Brain

All RNA editing events that were identified on miRNAs in more than 30% of each study's samples are presented in the **Appendices** section (**Table 16-Table 29**). In total, 47 A-to-I and 67 C-to-U events were consistently detected in brain samples. A significant difference ($p_{fisher's}$ = 0.037) was observed for the preference of A-to-I events to occur inside mature regions, relative to the occurrence of C-to-U events on regions flanking mature regions (**Table 5**). Focusing on events inside mature miRNA sequences, a clear preference for A-to-I editing is observed in the miRNA seed region (positions 2-8), while C-to-U events mainly occur at the end of mature sequences (p_{MWU} = 10^{-7}, **Figure 28**).

Table 5. Total miRNA events consistently occurring in mature/non-mature positions in brain samples.

Events	A-to-I	C-to-U
Total	47	67
In mature regions	28	26
In non-mature regions	19	41

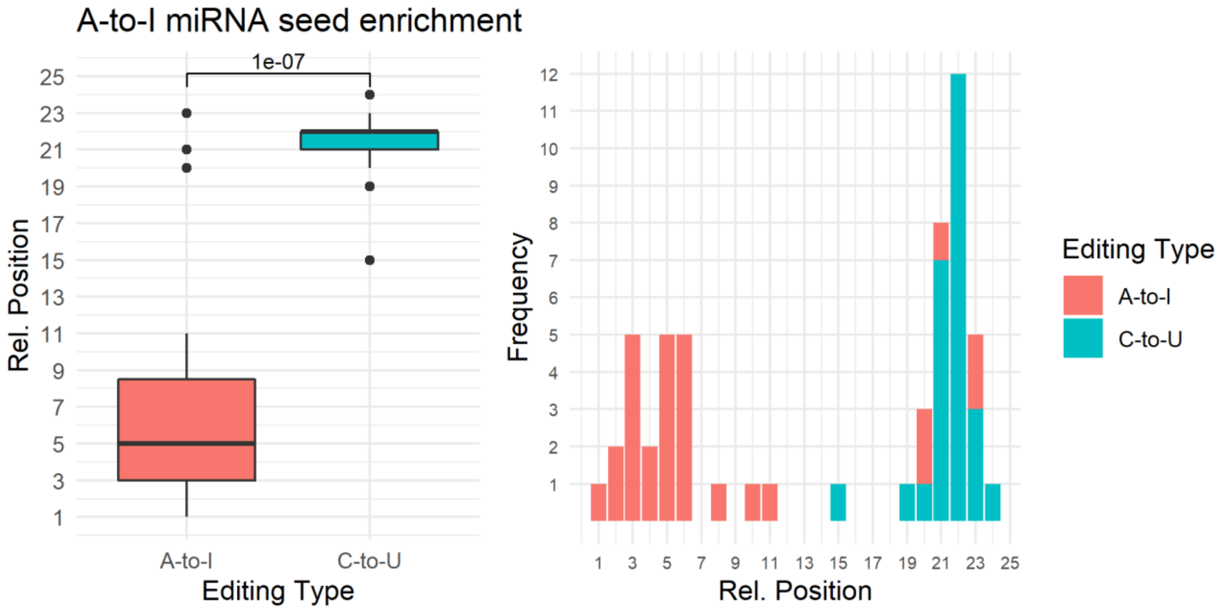


Figure 28. Box-and-whiskers plots and frequency bar-plots, denoting the differential occurrence of A-to-I and C-to-U miRNA editing inside mature miRNA sequence.

In the following subsections, the notation "miRNA-name **:** relative-position" (e.g. miR-379-5p:5) is used to denote RNA editing positions. Events found uniquely in one tissue are marked in **bold font**. Events that are common among tissues are presented concisely in **Table 6-Table 7**.

Consistent A-to-I RNA Editing Events

In total, 17 mature miRNA A-to-I events were identified in both prefrontal cortex studies, 4 of them being unique to this tissue (i.e. miRs *379-5p:5, 589-3p:6, 411-5p:5, 1301-3p:5, 301b-3p:20, 497-3p:20, 381-3p:4, 337-3p:6, 1251-5p:6, 3622a-3p:3, 7977:6, 497-5p:2* and *3681-5p:2*, and unique miRs ***200b-3p:5, 664a-5p:8, 376c-3p:6 and 203b-3p:11***, **Table 16-Table 17**). The edited form of *miR-641:3* was observed only in the SRP052236 study, in 97% of its samples. Regarding events occurring outside of mature miRNA sequences, low prominence was observed; ***mir-125b-2:78*** (38%), ***mir-26a-2:13*** (38%), ***mir-338:64*** (31%) events were exhibited in SRP063627, while *mir-320a:65* (30%) was exhibited in SRP052236.

The hippocampus datasets (SRP174906, **Table 18**) exhibited 13 mature miRNA editing events, one of them being hippocampus-specific (miRs 376a-5p:3, 379-5p:5, 337-3p:6, 411-5p:5, 3622a-3p:3, 381-3p:4, 1301-3p:5, 301b-3p:20, 7977:6, 589-3p:6, 641:3, 3681-5p:2 and **3144-3p:3**). Six events occurred in regions flanking mature miRNAs at varying consistency: ***mir-221:88*** (100%), ***mir-140:85*** (100%), *mir-320a:65* (80%), *mir-103a-2:71* (60%), ***let-7b:30*** (40%), ***mir-27b:82*** (40%).

The cerebellum study (SRP221185, **Table 19**) only contained one sample and unfortunately prominence of events could not be properly evaluated. Eleven mature miRNAs were found edited in cerebellum and other tissues (i.e. miRs *376a-5p:3, 379-5p:5, 381-3p:4, 497-5p:2, 497-3p:20, 589-3p:6, 411-5p:5, 301b-3p:20, 1251-5p:6, 3622a-3p:3* and *7977:6*), while 7 were cerebellum-specific (miRs ***19b-3p:23, 33a-5p:21, 99a-5p:1, 136-5p:23, 539-5p:10, 376a-2-5p:4 and 624-3p:5***). Twelve non-mature miRNA editing events were reported in cerebellum, 2 also occurring in other tissues (i.e. *mir-320a:65* and *mir-103a-2:71*) and 10 presenting specificity for cerebellum: ***mir-23a:67***, ***mir-26b:10***, ***mir-107:73***, ***mir-204:56***, ***mir-27b:13***, ***mir-30b:16***, ***mir-101-2:72***, ***mir-340:15***, ***mir-324:39*** and ***mir-338:65***.

In primary and recurrent Lower Grade Glioma samples (TCGA, **Table 20-Table 21**), 7 common editing events were found (i.e. miRs *381-3p:4, 589-3p:6, 411-5p:5, 3622a-3p:3, 379-5p:5, 1251-5p:6* and *376a-5p:3*), while ***miR-151a-3p:3*** was only consistently (39%) detected in primary tumors and also not found in other tissues. No A-to-I events were reported outside mature miRNA regions.

Three consistent A-to-I events were reported on mature miRNAs (i.e. *miR-7977:6, miR-3681-5p:2* and *miR-589-3p:6*), as well as one event on a non-mature pre-miRNA region (*miR-320a:65*), in the GEUVADIS dataset of Lymphoblastoid cells (**Table 22**). These 4 events were also found in other brain tissues.

Table 6. Commonly shared active A-to-I editing positions.

Mature miRNA regions

All brain tissues (i.e. prefrontal cortex, hippocampus, cerebellum and LGG)	5	miR-379-5p:5, miR-381-3p:4, miR-411-5p:5, miR-589-3p:6 (also shared by Lymphocytes) and miR-3622a-3p:3
Healthy brain tissues (i.e. prefrontal cortex, hippocampus and cerebellum)	2	miR-301b-3p:20, miR-7977:6 (also shared by Lymphocytes)
Hippocampus, cerebellum and LGG	1	miR-376a-5p:3
Prefrontal cortex, cerebellum and LGG	1	miR-1251-5p:6
Hippocampus and prefrontal cortex	4	miR-337-3p:6, miR-641:3, miR-1301-3p:5 and miR-3681-5p:2 (also shared with Lymphocytes)
Prefrontal cortex and cerebellum	2	miR-497-5p:2, miR-497-3p:20

Flanking mature miRNA regions

Healthy brain tissues (non-mature precursor region)	1	mir-320a:65 (also shared by Lymphocytes)
Hippocampus and cerebellum (non-mature precursor region)	1	mir-103a-2:71

Despite the significant overlap between A-to-I editing events observed in brain samples, tissues samples can be separated quite well solely on the basis on miRNA editing percentage of each site (**Figure 29**). Heatmaps were created by performing hierarchical complete-linkage clustering, employing Euclidean distances between miRNAs found edited in more than one tissue, and between samples. For sake of presentation the 512 primary LGG samples were reduced to 30 by random sampling, enabling a clearer evaluation of the sample clusters. Disease samples form a distinct cluster, together with hippocampus samples, which could be further discerned by the editing profile of specific miRNAs (e.g. miR-589-3p and miR-381-3p). Cerebellum and the two prefrontal cortex samples form a separate clade, while one prefrontal cortex sample is mis-placed within the disease-hippocampus group. Interestingly, both miR-497-5p and miR-497-3p displayed editing events in the prefrontal cortex, the former exhibiting a low frequency seed-based event, the latter exhibiting a very prominent signal at position 20.

Assessing the correlation between editing percentage and the abundance of edited miRNAs, as well as of their respective counterparts, i.e. the 3p mature form of 5p edited miRNAs and vice versa was very tempting; selection and loading in RISC of a "dominant" miRNA form has been proposed from very early to be primarily based on nucleotide-directed characteristics. Significant ($p < 2.2e-16$) negative (Pearson's coefficient = -0.22) and positive (Pearson's coefficient = 0.68) correlations were found for edited forms and their opposite forms respectively when testing all editing events collectively (**Figure 30**). However, this observation was not found to be generalizable at the individual edited miRNA level. One unique exception was event miR-497-3p:20, which correlated strongly and significantly (Pearson's correlation = 0.68, $R^2 = 0.46$, $p = 1.15e-6$) with the abundance of the opposite miRNA form, miR-497-5p. The editing event on miR-497-5p (position 2) also correlated significantly ($p = 0.0024$) with the abundance levels of miR-497-3p, which however were particularly lower (**Figure 30**).

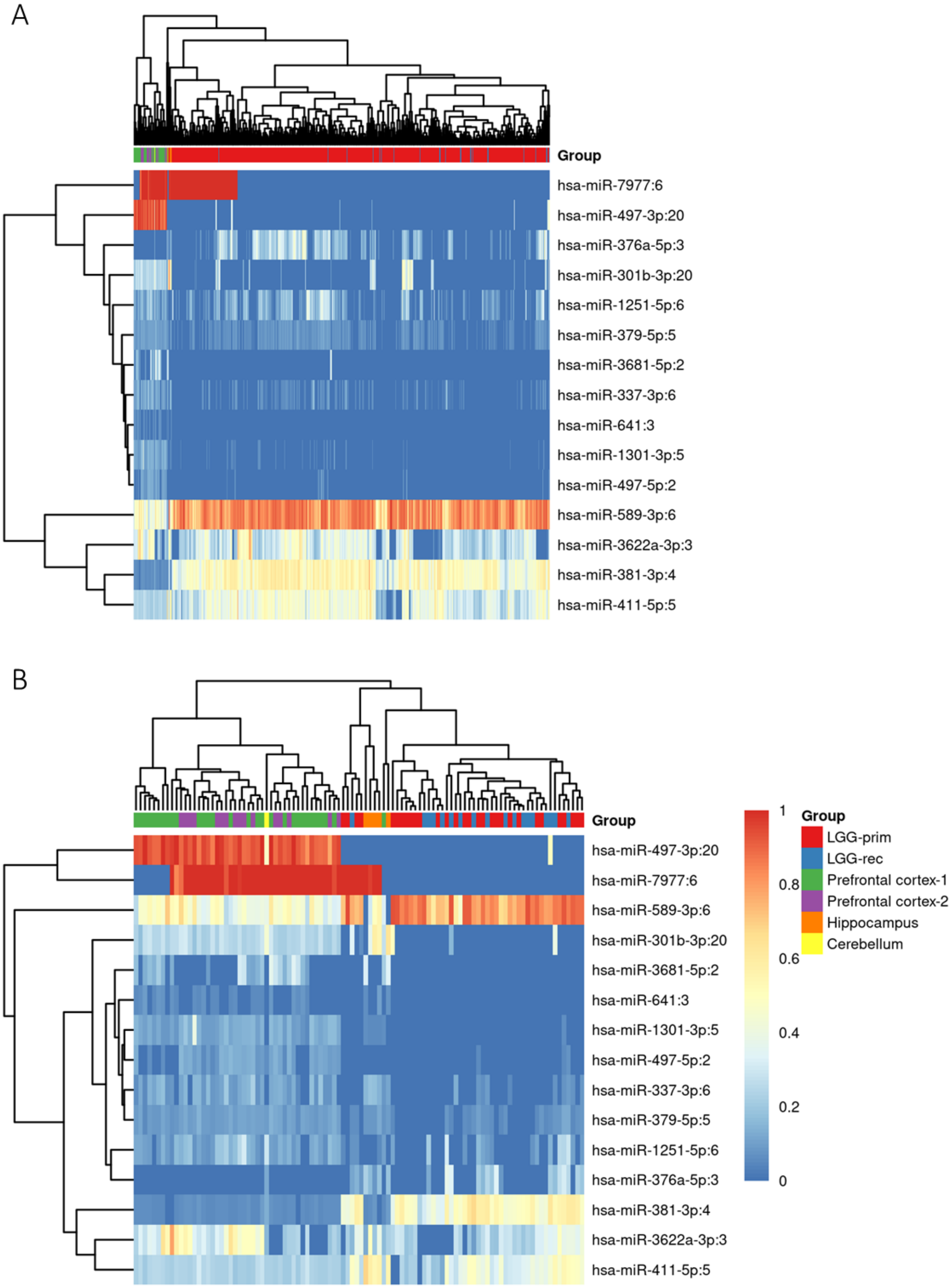


Figure 29. Heatmaps depicting A-to-I editing frequency for miRNAs found edited in more than one tissue. (A) The total number of samples is presented. (B) Only thirty out of 512 Primary LGG samples were randomly sampled to

enable clearer presentation of the underlying clusters. Primary and recurrent LGG samples, as well as Hippocampus samples are hierarchically clustered together and Cerebellum, Prefrontal Cortex samples form their own cluster.

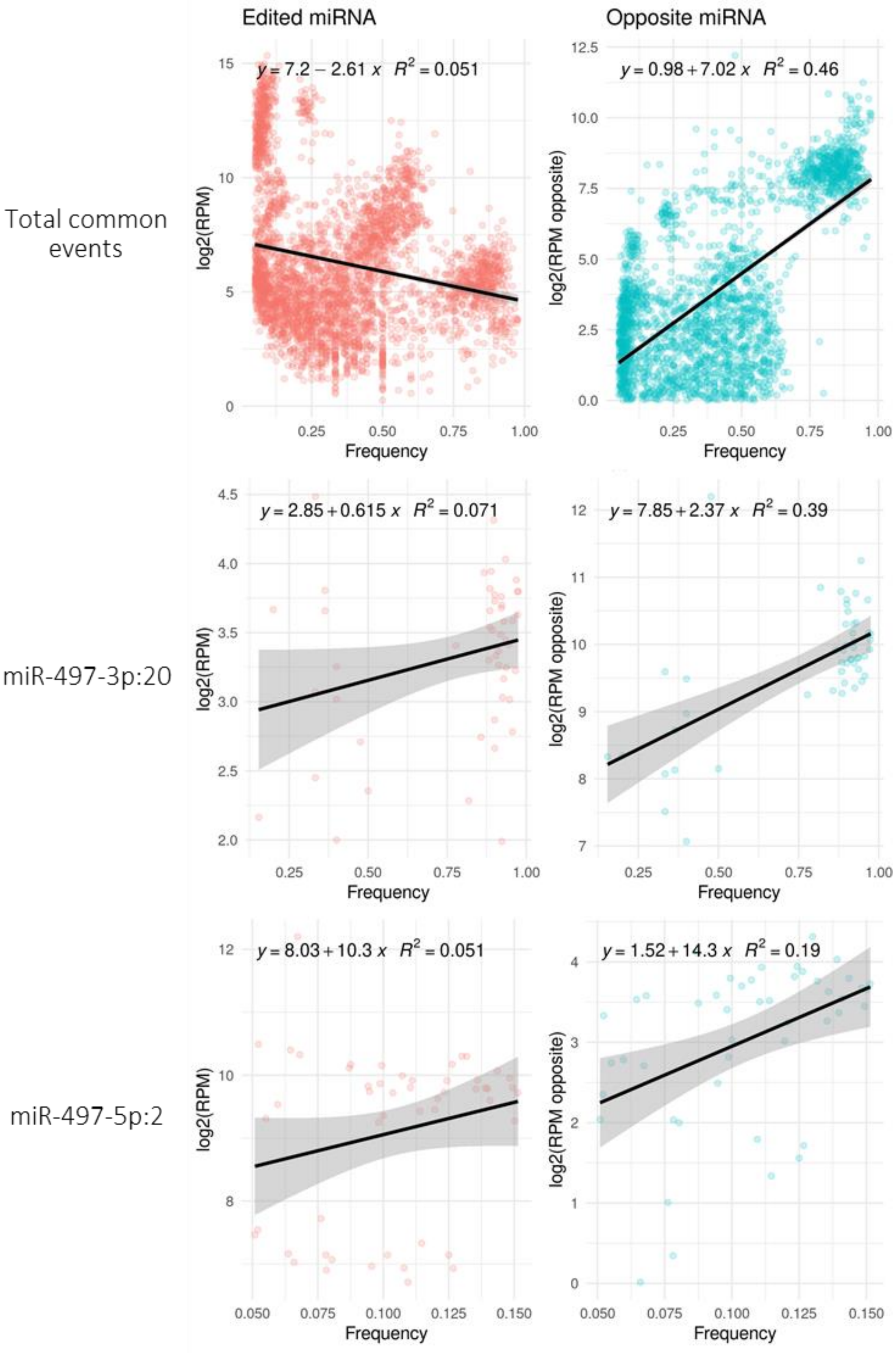


Figure 30. Pearson's correlation analysis between abundance levels of edited and each opposite miRNA forms and the editing frequency. Trends for the total A-to-I events that were common between tissues and for specific editing event on miR-497-5p/-3p are shown.

Consistent C-to-U RNA Editing Events

In total, 8 mature miRNA C-to-U events were identified in both prefrontal cortex studies (i.e. miRs *30a-3p:22, 221-3p:23, 27a-3p:21, 1301-3p:24, let-7b-3p:21, 488-3p:21* and prefrontal-cortex-specific **143-3p:21** and ***125b-1-3p:21***, **Table 23-Table 24**). Edited form *miR-23b-3p:23* was observed only in SRP063627 study (50% samples), while edited forms *miR-92b-3p:22* and *miR-30e-3p:22* were observed in SRP052236 (67% and 37% samples respectively). Twelve editing evens were found in both prefrontal cortex studies on non-mature miRNA regions: *let-7b:29, mir-134:30, mir-326:80, mir-148b:85, mir-152:76, mir-3200:76, mir-423:76, mir-328:70, mir-760:70, mir-874:69, mir-652:82, mir-370:70*, while *mir-222:92* (50%) was found only on SRP063627 and *mir-148a:66* (43%), *mir-151a:69* (30%), *mir-486-2:23* (30%) only on SRP052236.

In the hippocampus study (**Table 25),** 13 common events were detected on mature miRNA sequences (i.e. miRs *let-7b-3p:21, 27a-3p:21, 30a-3p:22, 221-3p:23, 23b-3p:23, 744-3p:20, 30e-3p:22, let-7b-3p:22, let-7f-1-3p:22, 361-5p:22, 1301-3p:24, 98-3p:22, 488-3p:21*) and 4 hippocampus-specific (*33a-5p:15, 106b-3p:22, 425-3p:22 and 421:23*). On non-mature miRNA sequences, 29 events were identified, 4 of them being hippocampus-specific (*let-7b:29, mir-21:30, mir-148a:66, mir-27b:83, mir-152:76, mir-126:74, mir-134:30, mir-370:70, mir-151a:69, mir-148b:85, mir-423:76, mir-652:82, mir-421:71, mir-3200:76, mir-132:81, mir-126:36, mir-222:92, mir-128-1:73, mir-330:82, mir-328:70, mir-3615:73, mir-326:80, mir-874:69, mir-760:70,* ***mir-374b:34***, ***mir-212:93***, ***mir-127:46***, ***mir-770:43*** and ***mir-25:74****)*.

Analysis of the one dataset of the cerebellum study (**Table 26)** yielded 16 mature miRNA editing events (miRs *let-7b-3p:21, let-7b-3p:22, let-7f-1-3p:22, 27a-3p:21, 30a-3p:22, 98-3p:22, 30e-3p:22, 221-3p:23, 23b-3p:23, 488-3p:21, 1301-3p:24, 744-3p:20,* ***185-3p:22, let-7d-3p:21, 345-5p:22 and 483-3p:19***), the four latter being cerebellum-specific. Thirty-one events (9 unique to cerebellum) were detected on precursor sequences (*let-7b:29, mir-21:30, mir-222:92, mir-27b:83, mir-128-1:73, mir-132:81, mir-152:76, mir-126:36, mir-126:74, mir-134:30, mir-370:70, mir-328:70, mir-326:80, mir-148b:85, mir-331:82, mir-423:41, mir-423:76, mir-652:82, mir-421:71, mir-874:69, mir-760:70, mir-3615:73,* ***mir-27a:72***, ***let-7g:83***, ***mir-195:37***, ***mir-99b:5***, ***mir-330:41***, ***mir-330:82***, ***mir-346:43***, ***mir-497:23*** and ***mir-3085:75***).

The primary and recurrent LGG samples (**Table 27-Table 28)** yielded 8 consistent events on mature sequences (i.e. *27a-3p:21, 30a-3p:22, 30e-3p:22, let-7b-3p:22, let-7b-3p:21, 744-3p:20, 92b-3p:22* and *361-5p:22*, the latter only in primary tumors) and 4 precursor events (*mir-326:80, let-7b:29, mir-134:30* and *mir-132:81*, the latter only in primary tumors).

GEUVADIS Lymphocytes (**Table 29)** were found to exhibit events on 6 mature miRNA positions (i.e. on *92b-3p:22, 27a-3p:21, 221-3p:23, 30e-3p:22,* ***92b-3p:21*** and ***BART2-5p:22***), the latter two being Lymphocyte-specific, and on 7 precursor miRNA positions (*mir-148a:66, mir-148b:85, mir-423:76, mir-423:41, mir-222:92, mir-486-2:23* and ***mir-BHRF1-1:27***).

Table 7. Commonly shared active C-to-U editing positions.

Mature miRNA regions

All brain tissues (i.e. prefrontal cortex, hippocampus, cerebellum and LGG)	4	miR-30a-3p:22, let-7b-3p:21, miR-27a-3p:21 (also shared with Lymphocytes), miR-30e-3p:22 (also shared with Lymphocytes)
Healthy brain tissues (i.e. prefrontal cortex, hippocampus and cerebellum)	4	miR-1301-3p:24, miR-488-3p:21, miR-23b-3p:23, miR-221-3p:23 (also shared with Lymphocytes)
Hippocampus, cerebellum and LGG	3	miR-744-3p:20, let-7b-3p:22, miR-98-3p:22
Hippocampus and cerebellum	2	let-7f-1-3p:22, miR-1301-3p:24
Prefrontal cortex and LGG	1	miR-92b-3p:22
Hippocampus and LGG	1	miR-361-5p:22

Flanking mature miRNA regions

All brain tissues (i.e. prefrontal cortex, hippocampus, cerebellum and LGG)	3	let-7b:29, mir-134:30, mir-326:80
Healthy brain tissues (i.e. prefrontal cortex, hippocampus and cerebellum)	9	mir-152:76, mir-328:70, mir-760:70, mir-874:69, mir-652:82, mir-370:70 and mir-148b:85, mir-423:76, mir-222:92 the latter 3 also shared with Lymphocytes)
Hippocampus, cerebellum and LGG	1	mir-132:81
Hippocampus and cerebellum	8	mir-21:30, mir-27b:83, mir-126:74, mir-421:71, mir-126:36, mir-128-1:73, mir-330:82, mir-3615:73

Clusters that more weakly correspond to the studied tissues were observed in hierarchical clustering of C-to-U events (**Figure 31**). The majority of common events among tissues were shown to occur at particularly low percentage per sample. Prefrontal cortex samples mainly clustered in a group, due to lack of editing activity on let-7b-3p:22, which was observed predominantly in the remaining samples. A very weak positive correlation was observed between editing percentage and the abundance of both mature miRNA forms (**Figure 32**). However, miRNA abundance was particularly low in the case of C-to-U edited miRNAs and R^2 metrics were always indicative of no linear relationship at place.

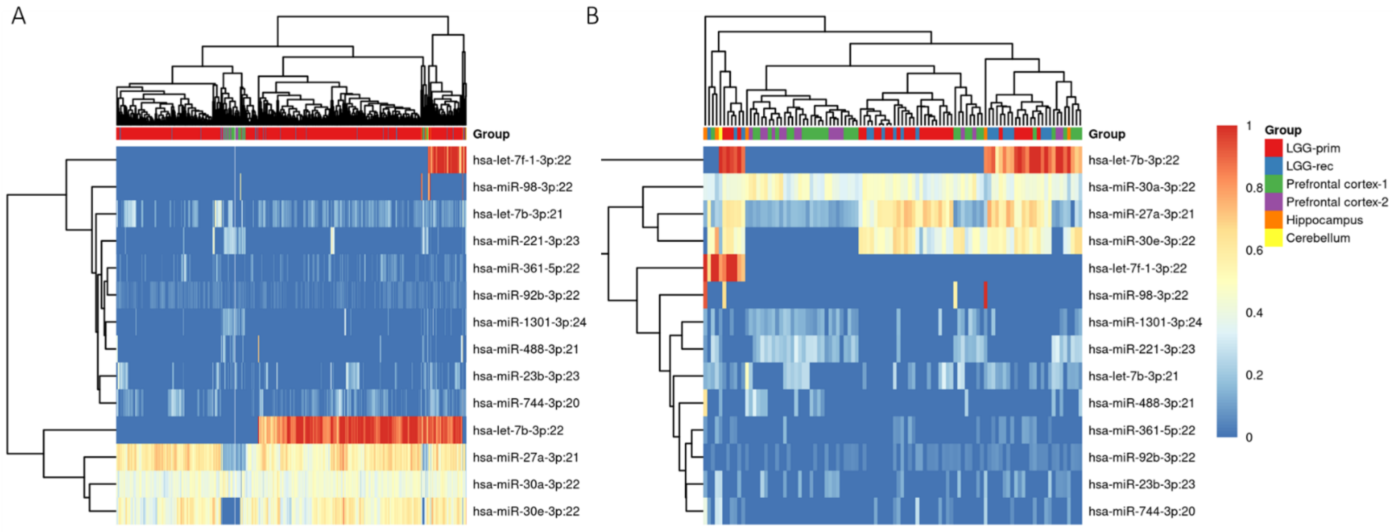


Figure 31. Heatmaps depicting C-to-U editing frequency for miRNAs found edited in more than one tissue. (A) The total number of samples is presented. (B) Only thirty out of 512 Primary LGG samples were randomly sampled.

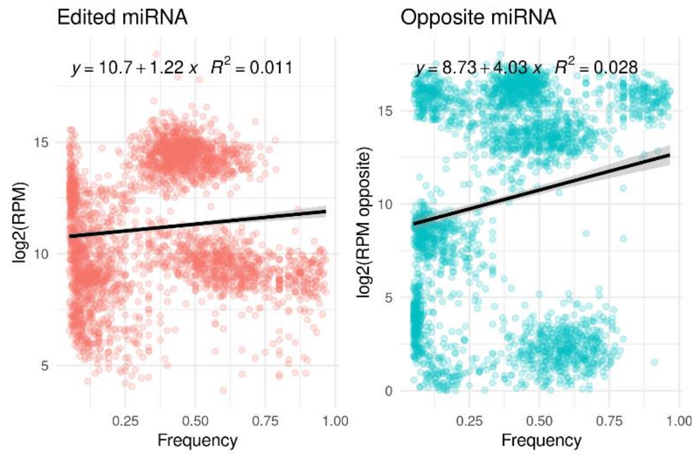


Figure 32. Pearson's correlation analysis between abundance levels of C-to-U-edited and each opposite miRNA form and the editing frequency. Trends for the total events that were common between tissues are shown.

6.2.4 Investigating Conservation and Structural Effects of miRNA Editing

Analysis of the structure impact of miRNA editing was performed by inspecting the effects of editing events on the capacity of pre-miRNAs for hairpin formation. RNAfold (Vienna package)[228] was employed at default settings to calculate the secondary structure and free energy of edited and non-edited pre-miRNAs derived from the sRNA-Seq analyses. The median differences in hairpin free energy were calculated and contrasted. Shapiro-Wilk normality tests, visual inspection of Q-Q plots and histograms were utilized to define the type of statistical test required. In case normality was assumed, paired t-tests were employed for edited *vs.* non-edited contrasts (i.e. cases involving the same pre-miRNAs), while unpaired t-tests for edited *vs.* edited and non-edited *vs.* non-edited contrasts (which included different sets of pre-miRNAs). Otherwise, Wilcoxon signed rank tests and Mann Whitney U tests were used respectively.

Conservation analysis of the identified editing positions was based on phastCons and phyloP conservation metrics which were retrieved from UCSC Genome Browser[219]. Both scores constitute per-base conservation metrics resulting from multiple alignments of 100 vertebrate genomes (100-way versions). phastCons is a hidden Markov model-based method that calculates the probability that each nucleotide belongs to a conserved element, weighing the contribution of its flanking regions as well. On contrast, phyloP ignores neighboring nucleotides, exhibiting greater granularity. Conservation metrics of A-to-I and C-to-U editing positions were contrasted, along with these of neighboring non-edited positions belonging to the same sub-sequence of the pre-miRNA structure (i.e. 5p-mature, 3p-mature or other).

6.2.4.1 Evolutionary Conservation Analysis of edited miRNAs

In the conservation analysis, an attempt to derive information on the potential evolutionary history of miRNA regions harboring editing events was made. Events utilized here were derived from miRNA editing analysis of healthy prefrontal cortex (SRP052236 and SRP063627), cerebellum (SRP221185), and hippocampus (SRP174906) sRNA-Seq datasets, as well as from analysis of TCGA-LGG primary and recurrent lower grade glioma sRNA-Seq samples.

Initially, we contrasted A-to-I to C-to-U events among the examined datasets (**Figure 33**). Apparently (bottom panel), phastCons scores, which incorporate information from bases in the immediate vicinity, tend to produce quite flattened scores. All medians equal 1 (i.e. 100% probability the specific base belongs to a conserved region), reflecting the conserved and functional nature of miRNAs. Interestingly, the boxplots' 1st quartile extends downwards only for A-to-I-edited positions, weakly suggesting that C-to-U positions are more conserved. phyloP scores are negative log p-values of the null hypothesis of neutral evolution, and do not depend on flanking bases. This enables a more granular per-base depiction of conservation *vs.* evolutionary acceleration. The mildly higher conservation of C-to-U positions in contrast to A-to-I positions is also evident here.

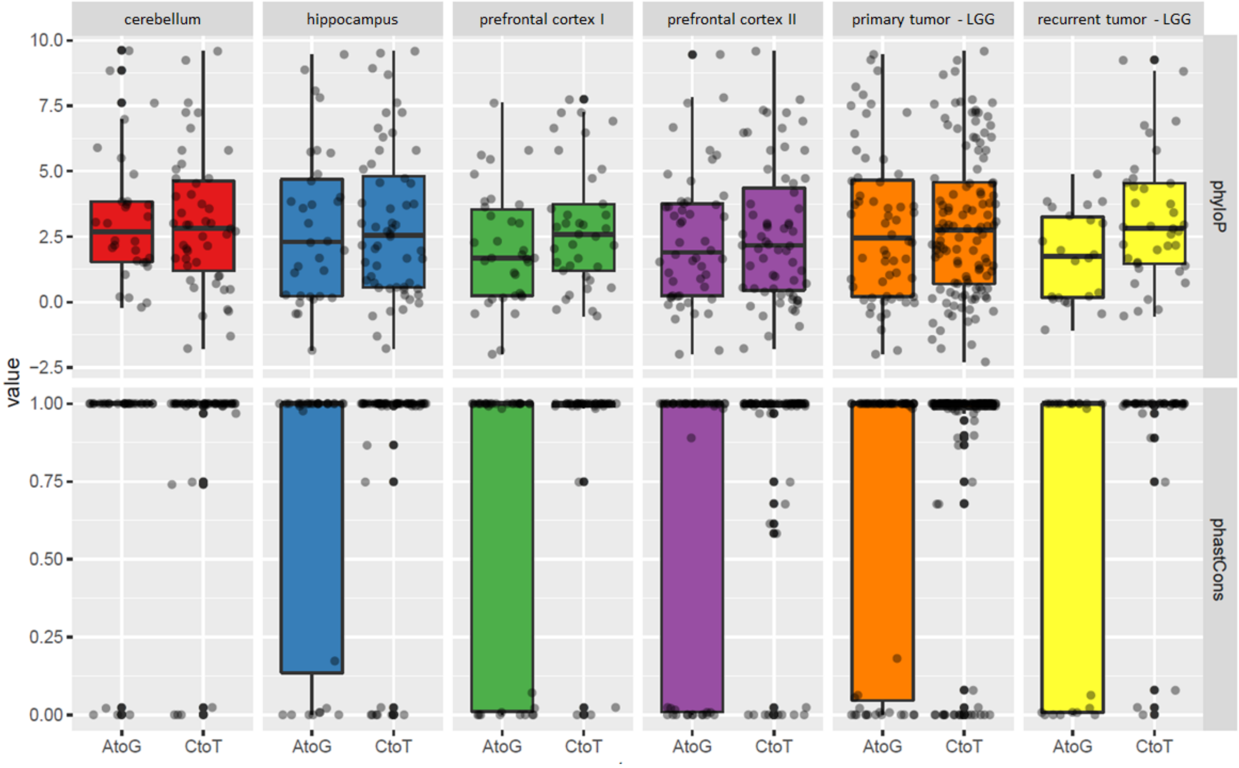


Figure 33. Conservation metrics contrasting A-to-I and C-to-U editing positions in brain-derived sRNA-Seq samples. Figure created for the purpose of this thesis.

The same analysis was used to examine pre-miRNA subregions; edited pre-miRNA positions were grouped according to whether they belong to the 5p/3p mature miRNA or not (**Figure 34**). For each position a neighboring non-edited base belonging to the same subregion was selected to form one negative set for each subregion-tissue combination (denoted in blue).

The results are not conclusive; Five prime regions are on average more conserved than the other regions, with C-to-U events in some cases (i.e. recurrent tumors, hippocampus) exhibiting high phyloP scores. A-to-I events on 3p regions are mildly less conserved than their neighboring base group and definitely less conserved than C-to-U events. Scores in positions outside mature regions are more ambiguous. This view might indicate either limited functional potential of the identified editing events, or the fact that their function is highly species-specific, requiring multi-species editing analysis to identify potentially conserved events instead of genomic positions.

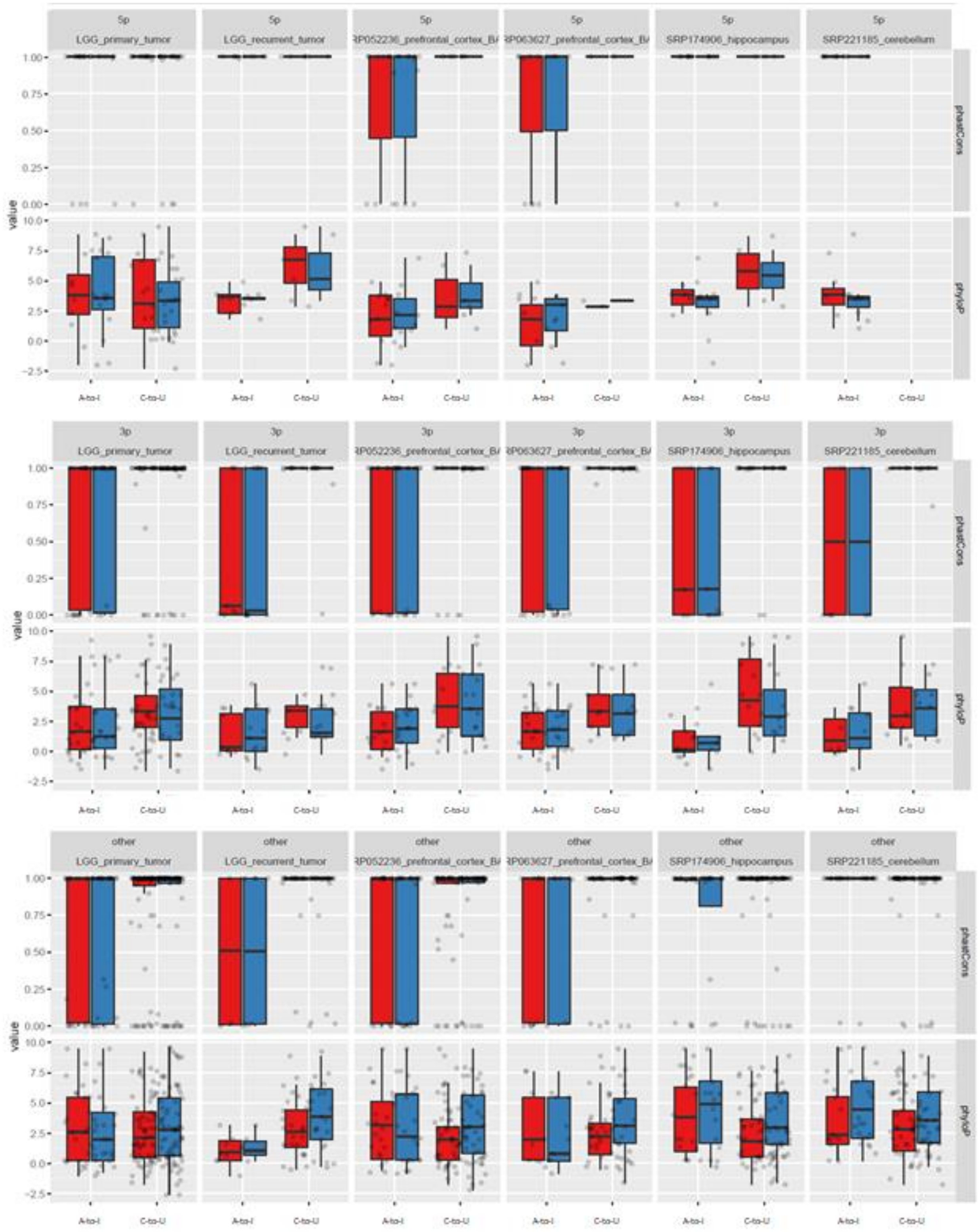


Figure 34. Conservation metrics for A-to-I, C-to-U editing positions and neighboring non-edited positions. Results are provided separately for the mature miRNA regions (5/3p) and for positions on non-mature pre-miRNA regions. Figure created for the purpose of this thesis.

6.2.4.2 Hairpin Formation Capacity Change due to miRNA Editing

The impact of revealed RNA editing events on the secondary structure of precursors was studied. Both A-to-I and C-to-U events were analyzed, **(a)** in whole precursor sequences, as well as in subregions **(b)** harboring and **(c)** not harboring mature miRNAs. Contrasts were computed for total and consistent (defined as the prevalence of a specific editing event in >30% of the study's samples) editing events found in each study. Brain RNA editing events were employed in the analysis, specifically those derived from healthy prefrontal cortex (SRP052236 and SRP063627), cerebellum (SRP221185), and hippocampus (SRP174906), and primary and recurrent lower grade glioma samples from the TCGA-LGG cohort.

Contrasts included:

1. Total (i.e. Healthy and Disease states) edited *vs.* non-edited free energies
2. Healthy edited *vs.* non-edited free energies
3. Disease edited *vs.* non-edited free energies
4. Healthy edited *vs.* Disease edited free energies
5. Healthy non-edited *vs.* Disease non-edited free energies

Initial comparison of RNA editing impact using all detected targets revealed that on average A-to-I and C-to-U RNA editing events affect hairpin formation in opposite ways: A-to-I (ADAR-mediated) editing results in a significant (p < 0.05) yet mediocre increase in pre-miRNA stability (i.e. decrease of free energy) (**Table 8**), while C-to-U (APOBEC-mediated) editing significantly decreases stability (i.e. increase of free energy). This phenomenon also holds true in separate, healthy-only and disease-only comparisons of A-to-I-/C-to-U-edited *vs.* their respective non-edited hairpin forms. The difference in editing events found in disease samples *vs.* healthy ones was insignificant.

Table 8. RNA editing effects in free energy of pre-miRNAs. All identified events were utilized.

Set	**Contrast type A *vs.* B**	**Free Energy (kcal/mol, median)**		**n**		**p-value**	**Stability in A**
		Subset A	**Subset B**	**Subset A**	**Subset B**		
Healthy and Disease	A-to-I *vs.* non-edited	-41.15	-38.35	120		1.69E-13	↑
	C-to-U *vs.* non-edited	-40.6	-42	181		1.50E-06	↓
Healthy	A-to-I *vs.* non-edited	-39	-37.2	67		2.13E-07	↑
	C-to-U *vs.* non-edited	-40.6	-42.3	77		1.12E-06	↓
Disease	A-to-I *vs.* non-edited	-44.9	-39.5	53		2.07E-07	↑
	C-to-U *vs.* non-edited	-40.5	-41.25	104		5.41E-13	↓
A-to-I sites	Disease *vs.* Healthy	-44.9	-39	53	67	**>0.05**	-
C-to-U sites	Disease *vs.* Healthy	-40.5	-40.6	104	77	**>0.05**	-
Non-edited sites (A-to-I)	Disease *vs.* Healthy	-39.5	-37.2	53	67	**>0.05**	-
Non-edited sites (C-to-U)	Disease *vs.* Healthy	-41.25	-42.3	104	77	**>0.05**	-

Performing the same analysis using only consistent events (i.e. only editing events found in >30% of each study's samples), yielded similar results (**Table 9**); the decrease of RNA stability of C-to-U edited pre-miRNAs was insignificant for disease-only events, but this could be attributed to small sample size (n = 10).

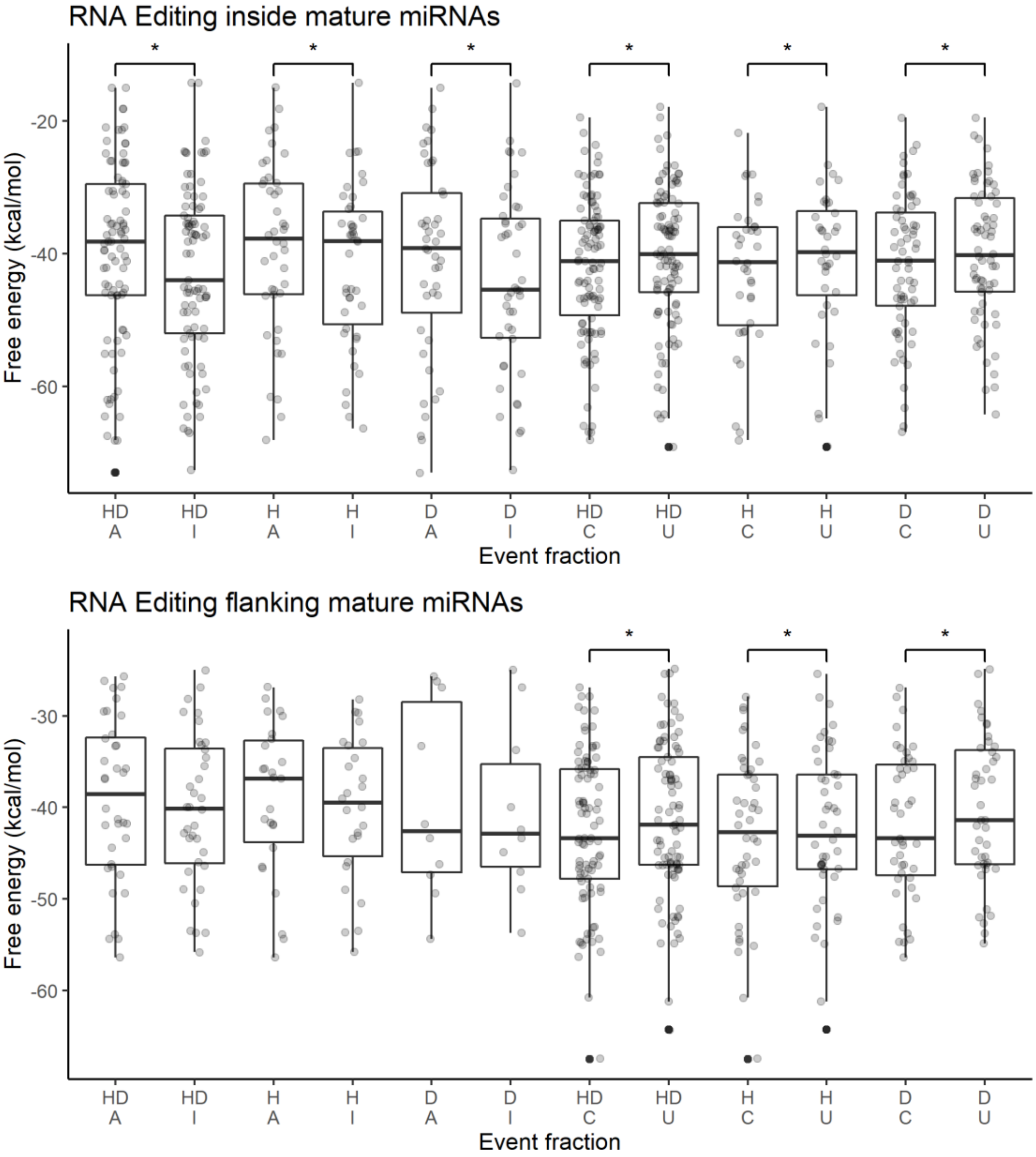
Table 9. RNA editing effects in free energy of pre-miRNAs. Events consistently appearing in >30% of study samples were utilized.

Set	**Contrast type A *vs.* B**	**Free Energy (kcal/mol, median)**		**n**		**p-value**	**Stability in A**
		Subset A	**Subset B**	**Subset A**	**Subset B**		
Healthy and Disease	A-to-I *vs.* non-edited	-37.9	-36.85	50		2.03E-07	↑
	C-to-U *vs.* non-edited	-41.05	-43.05	66		3.97E-06	↓
Healthy	A-to-I *vs.* non-edited	-38.5	-37.05	42		6.57E-06	↑
	C-to-U *vs.* non-edited	-41.05	-43.05	56		1.50E-06	↓
Disease	A-to-I *vs.* non-edited	-36.35	-33.05	8		1.45E-02	↑
	C-to-U *vs.* non-edited	-41.25	-42.05	10		**>0.05**	-
A-to-I sites	Disease *vs.* Healthy	-36.35	-38.5	8	42	**>0.05**	-
C-to-U sites	Disease *vs.* Healthy	-41.25	-41.05	10	56	**>0.05**	-
Non-edited sites (A-to-I)	Disease *vs.* Healthy	-33.05	-37.05	8	42	**>0.05**	-
Non-edited sites (C-to-U)	Disease *vs.* Healthy	-42.05	-43.05	10	56	**>0.05**	-

Notably, distinguishing pre-miRNA editing events residing on mature miRNA regions from those outside of it (i.e. either on the 5′/3′ of 5p/3p matures or on the stem loop), yielded different results (**Figure 35**). An increase of the pre-miRNA stability was observed for A-to-I events inside mature miRNA sequences and significance was lost for A-to-I events outside mature miRNA sequences. Regarding A-to-I events on the mature sequence, difference in pre-miRNA stability was more prominent in the disease-specific samples, compared to the healthy ones. C-to-U editing events appeared more robust to disease/healthy states. Statistical significance of the decrease in stability due to C-to-U editing was maintained both inside (**Table 10**) and outside (**Table 11**) of the mature miRNA sequence, in total events (Healthy and Disease) and in disease events, while a discrepancy was observed between non-mature and mature C-to-U miRNA events in healthy samples (i.e. decrease and mediocre (p = 0.013) increase respectively).

Table 10. RNA editing effects in free energy of pre-miRNAs. Events appearing inside mature miRNA sequences were utilized.

Set	**Contrast type A *vs.* B**	**Free Energy (kcal/mol, median)**		**n**		**p-value**	**Stability in A**
		Subset A	**Subset B**	**Subset A**	**Subset B**		
Healthy and Disease	A-to-I *vs.* non-edited	-44	-38.2	85		2.55E-08	↑
	C-to-U *vs.* non-edited	-40.1	-41.1	102		2.88E-11	↓
Healthy	A-to-I *vs.* non-edited	-38.1	-37.7	42		1.86E-06	↑
	C-to-U *vs.* non-edited	-39.75	-41.25	36		4.80E-05	↓
Disease	A-to-I *vs.* non-edited	-45.4	-39.2	43		2.31E-07	↑
	C-to-U *vs.* non-edited	-40.25	-41.05	66		1.90E-07	↓
A-to-I sites	Disease *vs.* Healthy	-45.4	-38.1	43	42	**>0.05**	-
C-to-U sites	Disease *vs.* Healthy	-40.25	-39.75	36	66	**>0.05**	-
Non-edited sites (A-to-I)	Disease *vs.* Healthy	-39.2	-37.7	43	42	**>0.05**	-
Non-edited sites (C-to-U)	Disease *vs.* Healthy	-41.05	-41.25	36	66	**>0.05**	-


RNA Editing inside mature miRNAs

RNA Editing flanking mature miRNAs

Figure 35. Change in miRNA hairpin stability due to RNA editing events. Boxplots depicting free energy (kcal/mol) of hairpin structure for non-edited (A or C) and edited (I or U, respectively) forms of pre-miRNA sequences. The top panel presents events occurring inside mature miRNA sequences, while the bottom panel shows events occurring on pre-miRNA regions flanking mature miRNAs. HD: Events from Healthy and Disease samples together, H: Events from Healthy samples, D: Events from Disease samples. Asterisks denote significance ($p < 0.05$). Figure created for the purpose of this thesis.

Table 11. RNA editing effects in free energy of pre-miRNAs. Events appearing outside mature miRNA sequences were utilized.

Set	Contrast type A *vs.* B	Free Energy (kcal/mol, median)		n		p-value	Stability in A
		Subset A	Subset B	Subset A	Subset B		
Healthy and Disease	A-to-I *vs.* non-edited	-41.15	-38.35	36		**>0.05**	-
	C-to-U *vs.* non-edited	-41.9	-43.4	89		2.48E-08	↓
Healthy	A-to-I *vs.* non-edited	-39.5	-36.85	26		**>0.05**	-
	C-to-U *vs.* non-edited	-43.1	-42.7	46		1.33E-02	↑
Disease	A-to-I *vs.* non-edited	-42.9	-42.6	10		**>0.05**	-
	C-to-U *vs.* non-edited	-41.4	-43.4	43		4.62E-06	↓
A-to-I sites	Disease *vs.* Healthy	-42.9	-39.5	10	26	**>0.05**	-
C-to-U sites	Disease *vs.* Healthy	-41.4	-43.1	43	46	**>0.05**	-
Non-edited sites (A-to-I)	Disease *vs.* Healthy	-42.6	-36.85	10	26	**>0.05**	-
Non-edited sites (C-to-U)	Disease *vs.* Healthy	-43.4	-42.7	43	46	**>0.05**	-

We conclude that A-to-I and C-to-U editing types seem to affect the stability of pre-miRNAs in a different manner in our studied samples. ADAR-mediated editing (A-to-I) confers a mediocre increase in pre-miRNA stability, which is exerted predominantly in the fraction of events happening inside mature miRNA sequences. On the other hand, APOBEC-mediated editing (C-to-U) significantly alters pre-miRNA stability irrespective of the event position. Intriguingly, stability change due to A-to-I events found on mature sequences of lower grade glioma samples was more prominent (ΔG = 6.2kcal/mol), compared to that of healthy brain samples (ΔG = 0.4kcal/mol).

6.2.5 miRNA Editing Impact of Targeting Repertoire and Regulation

The analysis of RNA editing effects on miRNA targeting was realized using DIANA-microT-CDS[192]. The edited and non-edited forms of mature miRNAs were subjected to target prediction employing transcript annotation from Ensembl v84.

microT prediction scores were compared between total predicted targets of non-edited and edited miRNA forms (without score threshold and applying one relaxed (i.e. 0.7) and one more stringent (0.9) minimum threshold). Mann Whitney U (MWU) non-parametric tests were utilized for the purpose of comparison. Additionally, Cumulative Distribution Functions (CDFs) were created for the score distributions and the two-sample Kolmogorov-Smirnov test (KS) was employed. When setting thresholds, total scores were unbalanced between the two conditions, therefore testing was also performed after sampling equal amounts of edited miRNA targeting scores, without replacement. Tests yielded significance (**Table 12**). The sets of targets yielded by non-edited and edited forms were compared by applying Jaccard Index (i.e. intersection over union) and calculating transcript- and gene-level numbers of targets each miRNA.

Finally, KEGG pathway enrichment analysis was performed for the edited and non-edited miRNA targets to identify molecular procedures that are differentially affected by editing events. Enrichment analyses were performed with Fisher's exact tests and FDR correction to identify significantly regulated pathways by each miRNA separately, as well as the by the union of the edited/non-edited miRNA targets, following miRPath rationale[194].

Table 12. Statistical tests utilized to determine change in microT scores distributions due to RNA editing.

Threshold	**p.value**	**n non-edited**	**n edited**	**Test type**
no threshold	1.39E-238	61064	61689	MWU
0.7 minimum	4.58E-12	5818	3893	MWU
0.9 minimum	6.54E-06	1253	600	MWU
no threshold	0	61064	61689	KS

A-to-I events that consistently occur in both primary (**Table 20**) and recurrent (**Table 21**) lower grade glioma were employed in a primary analysis of RNA editing effects in miRNA repertoire. The parallel Map-Reduce version of DIANA-microT-CDS target prediction algorithm was executed online and transcript-level targets were retrieved. microT scores range from zero to one, with values close to one being indicators of robust targeting efficiency. Scores were compared between the sets of non-edited and edited miRNA targets. Comparison was performed **(a)** without setting threshold ($n_{non\text{-}edited}$ = 61064, n_{edited} = 61689), **(b)** with a minimum 0.7 ($n_{non\text{-}edited}$ = 5818, n_{edited} = 3893) and **(c)** a 0.9 threshold ($n_{non\text{-}edited}$ = 1253, n_{edited} = 600). As shown in **Figure 36A**, a modest decreasing shift is observed for interactions of edited miRNAs ($p < 0.05$). This decrease is more discernible with the use of CDFs ($p < 0.05$) in **Figure 36B**.

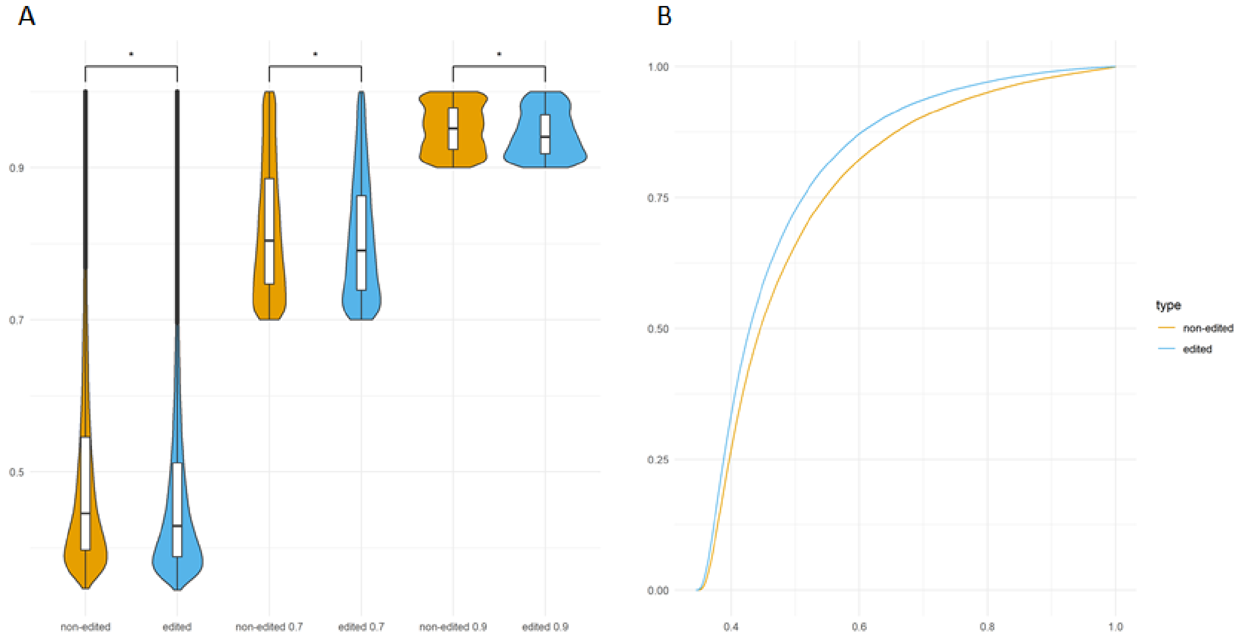


Figure 36. Distribution of microT target prediction scores for non-edited and A-to-I edited miRNAs common in primary and recurrent glioma. Panel A depicts score distributions for varying minimum thresholds and panel B shows the CDF of unfiltered scores. A mild score decrease is exhibited by edited miRNA forms, compared to non-edited, in all cases. Figure created for the purpose of this thesis.

miRNA targeting repertoire, as defined by microT-CDS target prediction algorithm, was significantly affected by A-to-I editing (**Table 13**). The numbers of genes/transcripts targeted with score ≥ 0.7 differed by the hundreds in 6 out of 7 cases (86%, with the exception of miR-376a-5p) and increase (43% of cases), as well as decrease of targets-per-miRNA was revealed. More importantly, gene/transcript targets were substantially different between the non-edited and the edited forms, as indicated by the especially low Jaccard Indices, which had a value range of 0.001-0.139 in both cases.

Table 13. Numbers of predicted miRNA targets for non-edited and edited forms and pair-wise similarity metrics (Jaccard Index).

miRNA	**Edit. Pos.**	**Mature Sequence (edited)**	**Gene Targets**			**Transcript Targets**		
			Reference	**Edited**	**Jaccard Index**	**Reference**	**Edited**	**Jaccard Index**
miR-1251-5p	6	acucu***I***gcugccaaaggcgcu	257	832	0.025	257	847	0.024
miR-3622a-3p	3	uc***I***ccugaccucccaugccugu	778	24	0.001	789	24	0.001
miR-376a-5p	3	gu***I***gauucuccuucuaugagua	516	597	0.039	519	607	0.038
miR-379-5p	5	uggu***I***gacuauggaacguagg	414	584	0.028	420	599	0.027
miR-381-3p	4	uau***I***caagggcaagcucucugu	2366	981	0.140	2378	988	0.139
miR-411-5p	5	uagu***I***gaccguauagcguacg	589	270	0.031	599	273	0.031
miR-589-3p	6	ugaga***I***ccacgucugcucugag	848	549	0.046	856	555	0.045

Pathway enrichment analysis was performed to investigate the potential involvement of edited miRNAs in regulation of the same processes as their corresponding non-edited forms. Initially, targets of individual non-edited/edited miRNAs were tested separately to inspect if they could significantly enrich KEGG pathways. Edited miR-381-3p and miR-1251-5p were the only edited forms that individually enriched pathways, yet statistical significance was marginal. Based on the non-binary nature of RNA editing (i.e. the fact that RNA editing occurs in portions of the existing RNAs and rarely 100%), an attempt to test for significant enrichment was made using combined sets of targets: the non-edited miRNA targets were contrasted to the target superset from both edited and unedited miRNAs.

Out of the seven analyzed miRNAs, only two non-edited forms, those of miR-381-3p and miR-589-3p, yielded significant results, including cancer-related pathways, glioma amongst them (**Table 14**). The contribution of edited forms is highlighted by the "ΔTargets" notation. Targets of the remaining non-edited miRNAs did not significantly enrich any pathways.

Table 14. Pathways enriched by non-edited forms of miR-381-3p (top 10 out of 69 shown) and miR-589-3p. Enrichment results of combined edited and non-edited forms are also depicted.

miRNA	**Pathway**	**Non-edited**		**Combined**		
		Targets	**FDR**	**Targets**	**FDR**	**Δtargets**
miR-381-3p	Signaling pathways regulating pluripotency of stem cells	45	4.21E-07	50	2.23E-06	5
	Pathways in cancer	112	6.60E-07	126	2.01E-05	14
	Axon guidance	49	6.50E-06	58	2.82E-06	9
	Rap1 signaling pathway	51	1.04E-04	57	4.37E-04	6
	TGF-beta signaling pathway	29	1.18E-04	33	1.50E-04	4
	Adherens junction	24	1.27E-04	29	2.01E-05	5
	EGFR tyrosine kinase inhibitor resistance	25	2.61E-04	26	2.21E-03	1
	Proteoglycans in cancer	47	6.32E-04	55	6.53E-04	8
	Breast cancer	37	6.32E-04	40	4.17E-03	3
	mTOR signaling pathway	38	7.52E-04	41	5.24E-03	3
miR-589-3p	Neurotrophin signaling pathway	16	5.04E-03	22	8.97E-04	6
	ErbB signaling pathway	12	1.67E-02	14	3.28E-02	2
	Glioma	11	1.67E-02	14	1.78E-02	3
	Rap1 signaling pathway	20	1.72E-02	-	-	-
	Axon guidance	18	1.72E-02	23	3.28E-02	5
	Choline metabolism in cancer	12	2.40E-02	16	2.37E-02	4
	Long-term potentiation	9	4.26E-02	12	3.28E-02	3
	Insulin signaling pathway	14	4.26E-02	NA	-	-
	Oxytocin signaling pathway	15	4.26E-02	NA	-	-
	Mitophagy - animal	-	-	14	8.71E-03	-
	Longevity regulating pathway	-	-	14	4.11E-02	-

The union of miRNA targets by either non-edited or combined (i.e. non-edited and edited) miRNAs was also subjected to enrichment analysis (**Table 15**). Enrichment of 85 pathways was observed. Increased ΔTargets values signify contribution of targets of the edited miRNAs. However, based on the analysis presented in **Table 13**, it should be noted that regulatory effects of the edited miRNAs might seemingly complement the regulation via targeting more pathway-related transcripts and not the same ones.

Table 15. Pathways enriched by non-edited forms of all miRNAs under study and by the combination of non-edited and edited forms' targets. The top 20 out of 85 enriched pathways are shown.

Pathway	Non-edited		Combined		
	Targets	FDR	Targets	FDR	Δtargets
Signaling pathways regulating pluripotency of stem cells	65	3.38E-07	76	7.92E-06	11
Pathways in cancer	175	1.83E-06	233	1.71E-07	58
Axon guidance	73	4.52E-06	100	1.61E-08	27
Adherens junction	37	4.52E-06	48	1.07E-07	11
Wnt signaling pathway	62	1.82E-04	82	1.33E-05	20
Hippo signaling pathway	60	3.67E-04	78	6.43E-05	18
Rap1 signaling pathway	75	4.34E-04	103	8.52E-06	28
Regulation of actin cytoskeleton	77	4.59E-04	105	1.45E-05	28
Proteoglycans in cancer	73	4.98E-04	99	2.37E-05	26
EGFR tyrosine kinase inhibitor resistance	35	4.99E-04	46	4.32E-05	11
Circadian rhythm	18	6.55E-04	20	1.77E-03	2
AMPK signaling pathway	47	7.65E-04	61	2.02E-04	14
TGF-beta signaling pathway	39	7.65E-04	48	8.96E-04	9
Breast cancer	55	7.65E-04	71	3.42E-04	16
FoxO signaling pathway	50	8.80E-04	67	7.81E-05	17
ErbB signaling pathway	35	2.01E-03	47	1.45E-04	12
Parathyroid hormone synthesis, secretion and action	41	2.67E-03	53	8.64E-04	12
Bacterial invasion of epithelial cells	32	2.71E-03	39	4.09E-03	7
Arrhythmogenic right ventricular cardiomyopathy	32	2.71E-03	43	2.02E-04	11

CHAPTER 7 – Participation in Research Projects

During my thesis, I have also participated in a number of distinct projects, under major or more supportive roles, in collaboration with members of DIANA-Lab and other research teams. In this chapter, a brief presentation of the main analyses and derived findings is provided for each project.

7.1 sRNA-Seq Analysis of Nasopharyngeal Swabs from Symptomatic and Asymptomatic COVID-19 Patients and Healthy Individuals

In light of the ongoing global pandemic, we initiated the first, to our knowledge, analysis of the small RNA content of nasopharyngeal tissue from symptomatic and asymptomatic COVID-19 patients, as well as healthy individuals. Our study included individuals from the University General Hospital of Larissa, whose on-admission and follow-up nasopharyngeal swabs were subjected to sRNA-Seq (**Figure 37**).

Briefly, in this study we described small RNA profiles of 38 nasopharyngeal swab samples, that revealed infection-induced changes in the post-transcriptional regulation landscape. We identified a pronounced decrease in total miRNA levels upon infection, a phenomenon not previously observed in infectious diseases. This was accompanied with a drop of tRNA levels and an increase of bacterial mapping rates (**Figure 38**). Analysis of follow-up samples demonstrated a replenishment of miRNA abundance, only in patients that progressed better. A COVID-19 signature consisting of 12 miRNAs was constructed on the basis of significant differential abundance and their linear separability among groups. Analysis of the signature's targetome reveals its potential immunopathogenic roles. miRNAs presenting hardly detectable expression levels in 3 deceased patients' late samples carry prognostic potential as markers for COVID-19 severity.

This study has been co-led by me and post-doctoral researcher Dr. Karagkouni Dimitra. I have participated in the experiment design, sRNA-Seq pre-processing, quantification, differential abundance analysis and identification of signature small RNAs found in the nasopharyngeal tissue of recruited individuals. I also performed the statistical analyses of patient demographic and clinical meta-information. A second sRNA analysis of serum samples from SARS-CoV-2 positive and negative individuals is under way.

Our study is currently under review in Lancet Infectious Diseases (I.F. 24.446): **S Tastsoglou***, D Karagkouni*, A Karavangeli, FS Kardaras, M Miliotis, G Vatsellas, G Skoufos, N Perdikopanis, G Papadamou, T Karamitros, D Thanos, C Koumenis, E Petinaki and AG Hatzigeorgiou. *Sustained decrease of microRNA levels detected in COVID-19 patients is linked to disease severity: Insights from small RNA-Seq of nasopharyngeal swabs*. 2020.

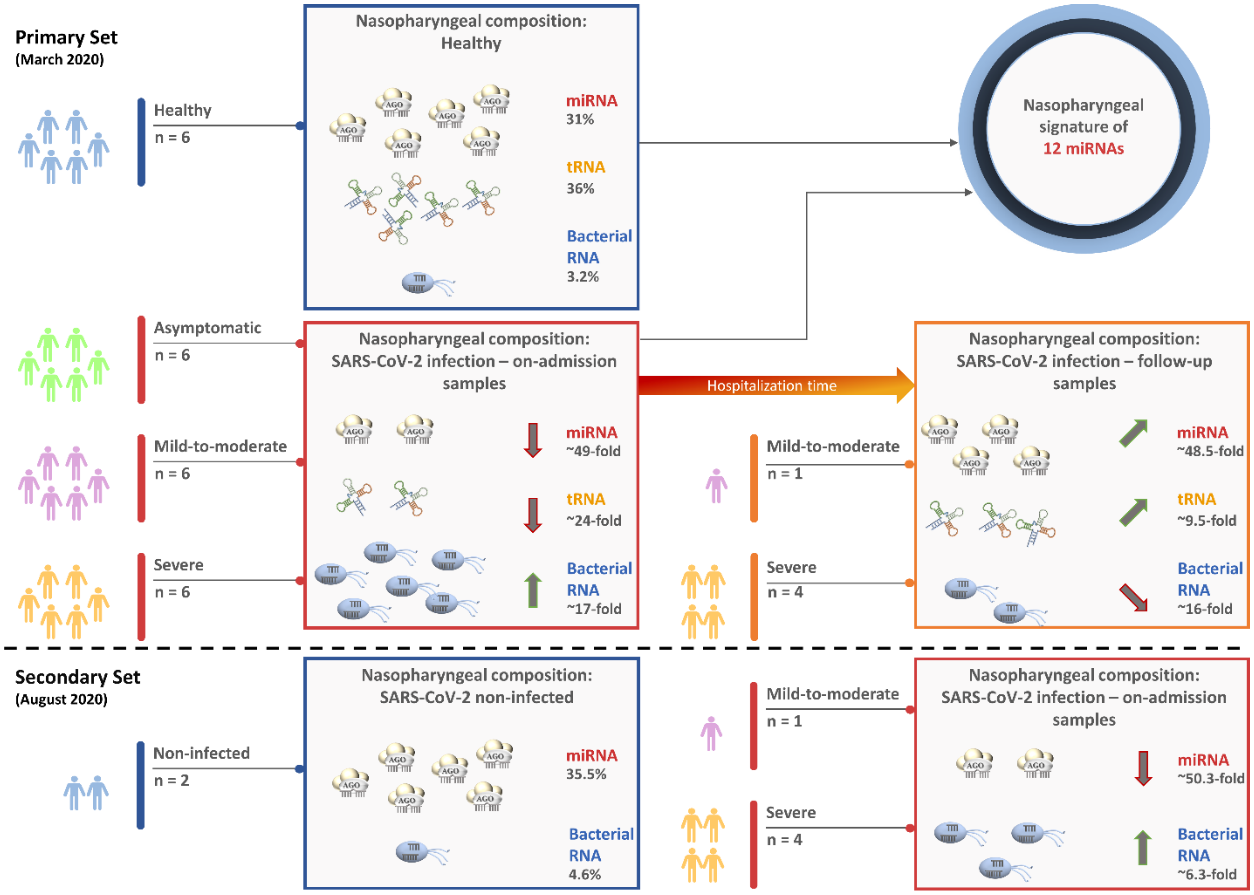


Figure 37. Overview of the nasopharyngeal swab sRNA-Seq analysis. Our cohort included Healthy volunteers, Asymptomatic, Mild-to-moderate symptomatic and Severe symptomatic COVID-19 patients. Analysis of total miRNA and tRNA levels revealed a dramatic decrease of more than 95% on early on-admission swab samples (median 49- and 24-fold respectively), as compared to Healthy samples, and an increasing trend on follow-up measurements of recovering patients. An opposite trend was identified for bacterial mapping rate; initially it was relatively high compared to Healthy and it gradually dropped during the later stages of hospitalization. The pronounced deviation of miRNA and bacterial mapping rates was further validated in a second, independent set of SARS-CoV-2 (non-)infected individuals. A COVID-19 disease progression signature consisting of specific miRNAs that were commonly dysregulated in all infected groups and were functionally characterized as active regulators of innate immune response, inflammation, cell stress and cell cycle pathways, was determined.

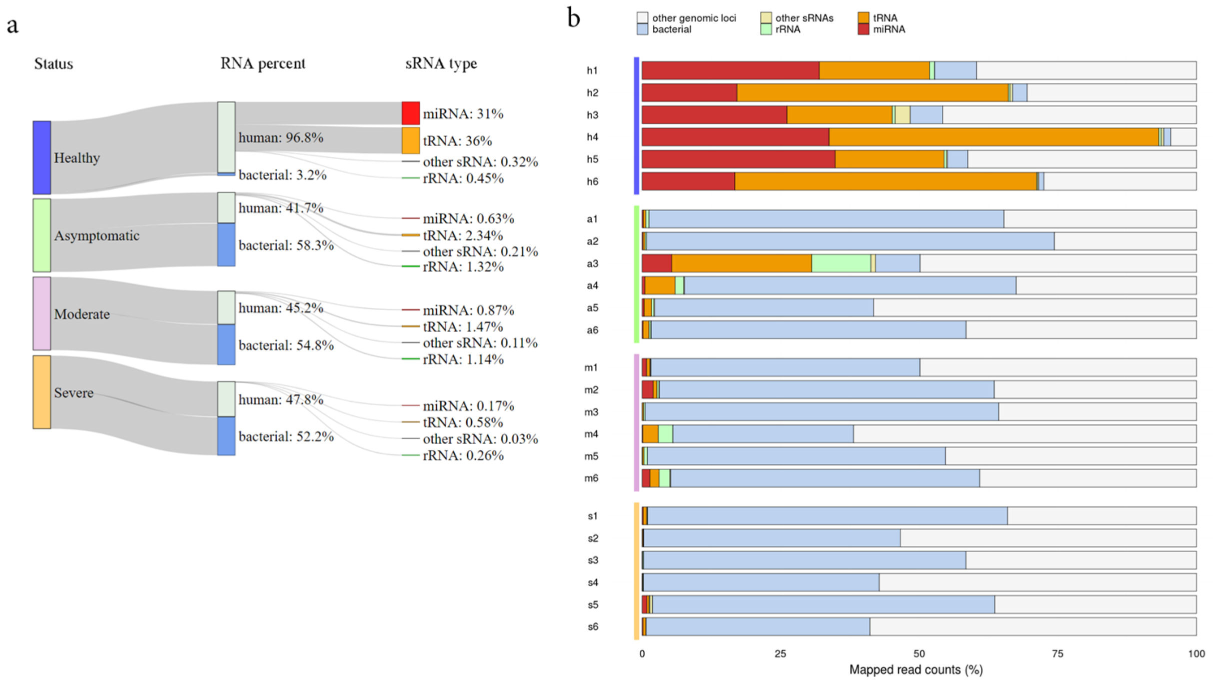


Figure 38. SARS-CoV-2 infection disrupts total abundance of miRNA, tRNA and bacterial RNA. Mapping rates of on-admission and follow-up nasopharyngeal sRNA-Seq datasets from SARS-CoV-2 patients, as well as of SARS-CoV-2-negative individuals are depicted. (a) Sankey diagram presenting the distribution of all reads that were attributed to host RNA species and microorganisms in our primary cohort. The portion of miRNA- and tRNA-assigned reads is significantly reduced in all SARS-CoV-2-infected groups. Bacterial species exhibit increased read alignment in all patient groups compared to Healthy individuals. Percentages correspond to medians across samples per group. (b) Per sample percentage of reads mapped to host RNA species and bacteria in the primary cohort. The portion of miRNA- and tRNA-assigned reads drops in all SARS-CoV-2-infected groups, while bacterial species exhibit increased read alignment compared to Healthy individuals.

7.2 Collection of Circulating Prognostic and Diagnostic miRNA Biomarkers

During the past 2 years, I have designed and coordinated the manual curation of articles harboring robust information on the diagnostic and prognostic potential of circulating microRNAs. We have extracted biomarker candidates from more than 200 publications that employ explicit methods used for biomarker validation (e.g. ROC analysis for diagnostic entries, Cox Regression analysis for prognostic modeling). Extensive experimental metadata support the more than 610 diagnostic and more than 160 prognostic biomarker entries that we have collected in a standardized manner.

The manuscript accompanying this database is currently under preparation: **S Tastsoglou**, M Miliotis, I Kavakiotis, A Alexiou, E Gkotsi, V Lygnos, V Maroulis, D Zisis & AG Hatzigeorgiou. plasmiR: a manual collection of circulating microRNAs of prognostic and diagnostic value. 2020.

7.3 Impact of *Helicobacter pylori* Infection and its Major Virulence Factor CagA on DNA Damage Repair

I have participated in a study led by PhD candidate Kontizas E and Sgouras D, regarding the transcript and protein level changes induced by *H. pylori* infection on AGS gastric epithelial cell-lines. In this study, I performed transcriptomic meta-analysis of RNA-Seq experimental datasets from AGS cell lines that were infected with 3 strains of *Helicobacter pylori*. These analyses led to the identification of DNA damage response pathways and genes that are affected by *H. pylori* infection and explicitly by the presence of CagA, a major virulence factor of *H. pylori*. The effects of infection on DNA damage response pathways were further supported with downstream western blot experiments.

This study is currently under review in the Special Issue on *Helicobacter pylori* and Gastric Carcinogenesis of the MDPI journal Microorganisms (I.F. 4.167): E Kontizas, **S Tastsoglou**, T Karamitros, Y Karayiannis, P Kollia, AG Hatzigeorgiou, D Sgouras. *Impact of Helicobacter pylori Infection and its Major Virulence Factor CagA on DNA Damage Repair*. 2020.

7.4 Dendritic Cell Transcriptional Profiling upon *Leishmania* Infection and Vaccination with Novel Putative Nano-formulations

In a collaborative study with Athanasiou E and Karagouni E, the maturation status of mouse dendritic cells (DCs) upon *Leishmania* infection and in response to vaccination with novel nano-vaccine formulations was studied from a transcriptomic standpoint. I performed differential expression analysis of microarray datasets of DCs after stimulation with different mixes of nano-formulations and utilized hypergeometric tests to evaluate the enrichment of KEGG pathways and Gene Ontology terms relevant to cytokine production and inflammatory response. Formulations containing MPLA adjuvant were found to result in the most potent transcriptional responses, deregulating genes that could induce protective $CD8^+$ T cell activation and $CD4^+$ T_{H1} polarization.

This study was published in Frontiers in Immunology (I.F. 5.085) and has received 19 citations: E Athanasiou, M Agallou, **S Tastsoglou**, O Kammona, AG Hatzigeorgiou, C Kiparissides and E Karagouni. *A Poly(Lactic-co-Glycolic) Acid Nanovaccine Based on Chimeric Peptides from Different Leishmania infantum Proteins Induces Dendritic Cells Maturation and Promotes Peptide-Specific IFNγ-Producing $CD8^+$ T Cells Essential for the Protection against Experimental Visceral Leishmaniasis*. 2017.

7.5 Short Time Series Analysis of Gene Expression in *Leishmania*-infected Mice

In the study titled "Transcriptome analysis identifies immune markers related to visceral leishmaniasis establishment in the experimental model of BALB/c mice", we delineated the transcriptional changes underlying *Leishmania*-infected mouse spleens during vaccination with a prominent novel nano-vaccine candidate[229]. A set of genes was identified that was indicative of the association of enhanced interferon-mediated inflammation, neutrophil infiltration, and T cell exhaustion with *L. infantum* infection establishment in spleen of BALB/c mice.

In this study, I performed pre-processing and differential expression analysis on generated microarray datasets and utilized CDF distributions to depict the trends of differentially regulated immune-related genes at 4 weeks post-infection/-vaccination (**Figure 39A**). Also, by employing the Short Time-Series Expression Miner tool I identified clusters of co-expressed genes that are common or distinct between vaccinated and non-vaccinated (PBS) groups (**Figure 39B**).

This study was published in Frontiers in Immunology: M Agallou, E Athanasiou, O Kammona, **S Tastsoglou**, AG Hatzigeorgiou, C Kiparissides and E Karagouni. *Transcriptome analysis identifies immune markers related to visceral leishmaniasis establishment in the experimental model of BALB/c mice*. 2019.

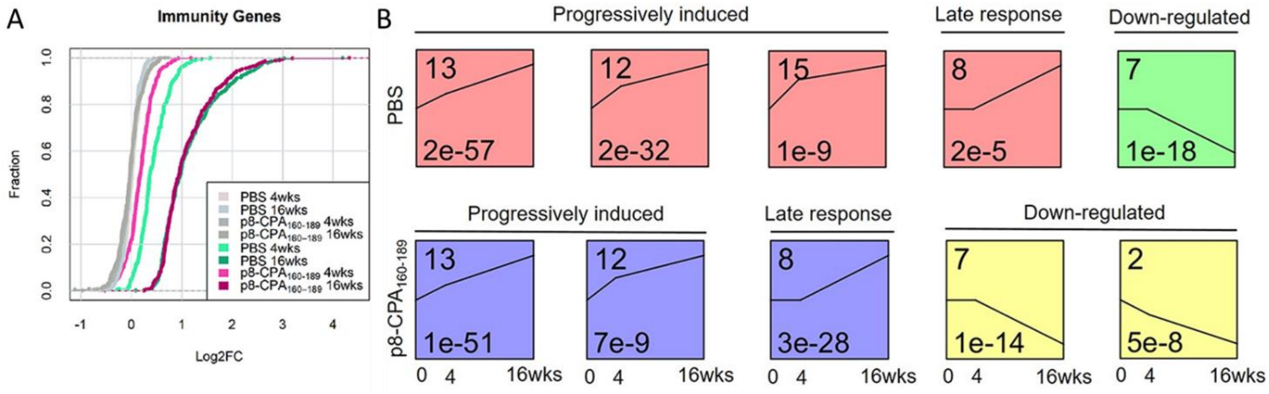


Figure 39. (A) Cumulative Distribution Functions (CDFs) of immunity-related genes across contrasts. CDFs in gray where obtained by random sampling of equal numbers from each contrast among all genes. Distribution shifts were assessed using Wilcoxon rank-sum tests. (B) Microarray time course analysis of DEGs classified 629 DEGs of spleens from non-vaccinated (PBS) and vaccinated (p8-CPA160−189) mouse groups at 0 (non-vaccinated non-infected), 4 and 16 weeks into 6 significant patterns of gene expression ($p < 0.05$). DEGs were classified into five main profiles according to the temporal gene expression pattern. Figure adapted from Agallou M et al. (2020)[229].

7.6 State-of-the-art PAR-CLIP-guided Identification of miRNA Targets

In a comprehensive analysis of publicly available PAR-CLIP datasets, we demonstrated for the first time that read clusters lacking the PAR-CLIP-specific diagnostic T-to-C mutations should not be neglected from analysis, since they are also indicative of functional and miRNA-binding events (**Figure 40**) and exhibit strong RNA accessibility signal. These findings formed the basis for the creation of an avant-garde super-learning scheme for CLIP-guided identification of miRNA targets, named microCLIP. The training and testing of microCLIP was based on extensive integration of evidence for the existence/absence of miRNA targeting events from various experimental sources. microCLIP adopts a Super Learning scheme, which constitutes a weighted combination of many machine learning models, and was found to outperform leading implementations in PAR-CLIP analysis.

In this study, I conducted differential expression analysis in numerous microarray datasets assessing the effects of miRNA overexpression or knock-down. Sets of MRE-containing genes that did or did not significantly change upon miRNA perturbation were utilized as positive or negative instances during microCLIP training, validation and testing. I also conducted a preliminary pathway enrichment analysis of microCLIP-derived T-to-C and non-T-to-C-targets to evaluate the importance of including non-T-to-C results in downstream functional analyses.

The findings of this study were published in Nature Communications (I.F. 12.121) and have received 6 citations since: MD Paraskevopoulou, D Karagkouni, IS Vlachos, **S Tastsoglou** and AG Hatzigeorgiou. *microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions*. 2018.

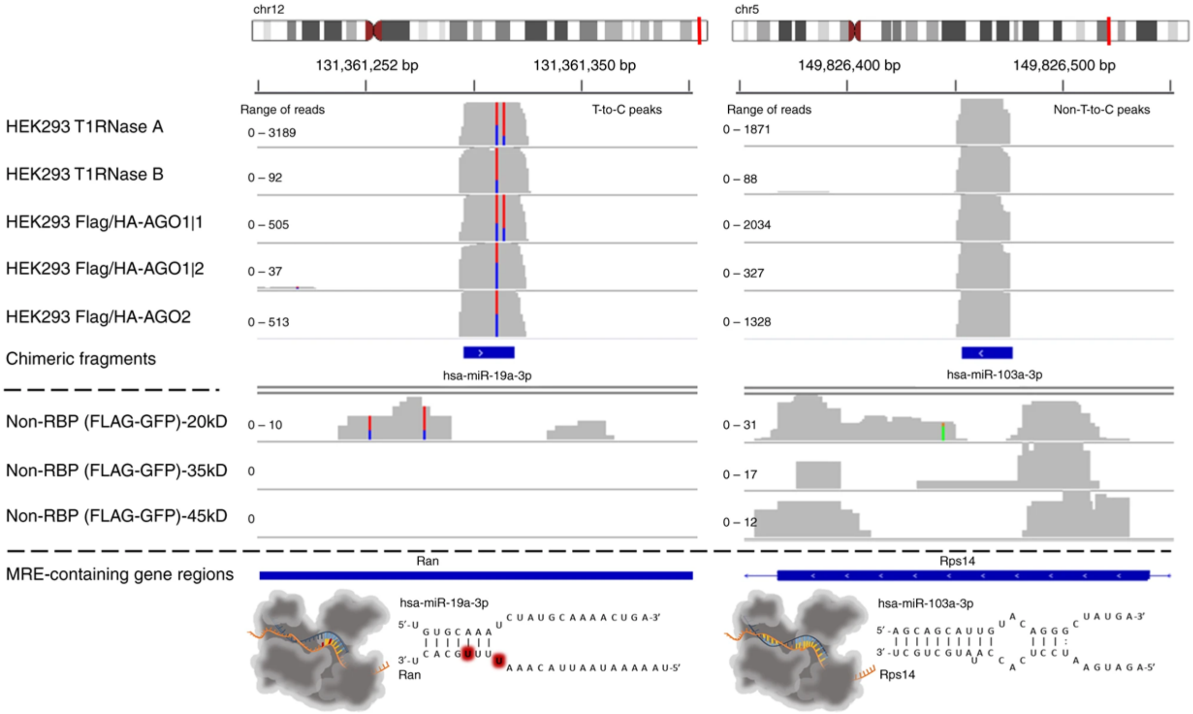


Figure 40. Peaks derived from 5 AGO-PAR-CLIP libraries on HEK293 cells and from 3 non-RBP background libraries are presented for T-to-C and non-T-to-C AGO-bound regions. The red-and-blue vertical lines represent T-to-C transition sites. Both types of AGO-enriched clusters are clearly distinguished from 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. Figure retrieved from Paraskevopoulou MD et al. (2018)[33].

7.7 Collection of bacterial-disease associations with experimental validation

This constitutes one of the latest publications of DIANA-Lab in another direction, that of the identification of associative relationships between bacteria found in the host (mainly human) microbiota and underlying diseases. Peryton is a database collecting such relationships from publications harboring robust experimental evidence. It hosts more than 7900 entries, linking 43 diseases with 1396 microorganisms. Peryton's content is exclusively sustained by manual curation of biomedical articles. Diseases and microorganisms are provided in a systematic, standardized manner using reference resources to create database dictionaries. Information about the experimental design, study cohorts and the applied high- or low-throughput techniques is meticulously annotated and catered to users[230]. Extensive visualization options are available to enable exploration in the up-to-date Peryton content (**Figure 41**).

This study was accomplished by and was just published in the 2021 Database Issue of the Nucleic Acids Research. For the purposes of this database, I performed systematic annotation of the collected diseases to avoid the occurrence of differences in naming conventions introduced by each publication. I also created a list of known common contaminants derived from available literature, in order to annotate the relevant entries, informing users to handle them more cautiously. Finally, the Peryton's Help Section, the publication's snapshots and legends, were created by me.

This database was published in the 2021 Special Database Issue of the Nucleic Acids Research (I.F. 11.501): G Skoufos, FS Kardaras, A Alexiou, I Kavakiotis, A Lambropoulou, V Kotsira, **S Tastsoglou** and AG Hatzigeorgiou. *Peryton: a manual collection of experimentally supported microbe-disease associations*. 2021 (already available doi.org/10.1093/nar/gkaa902).

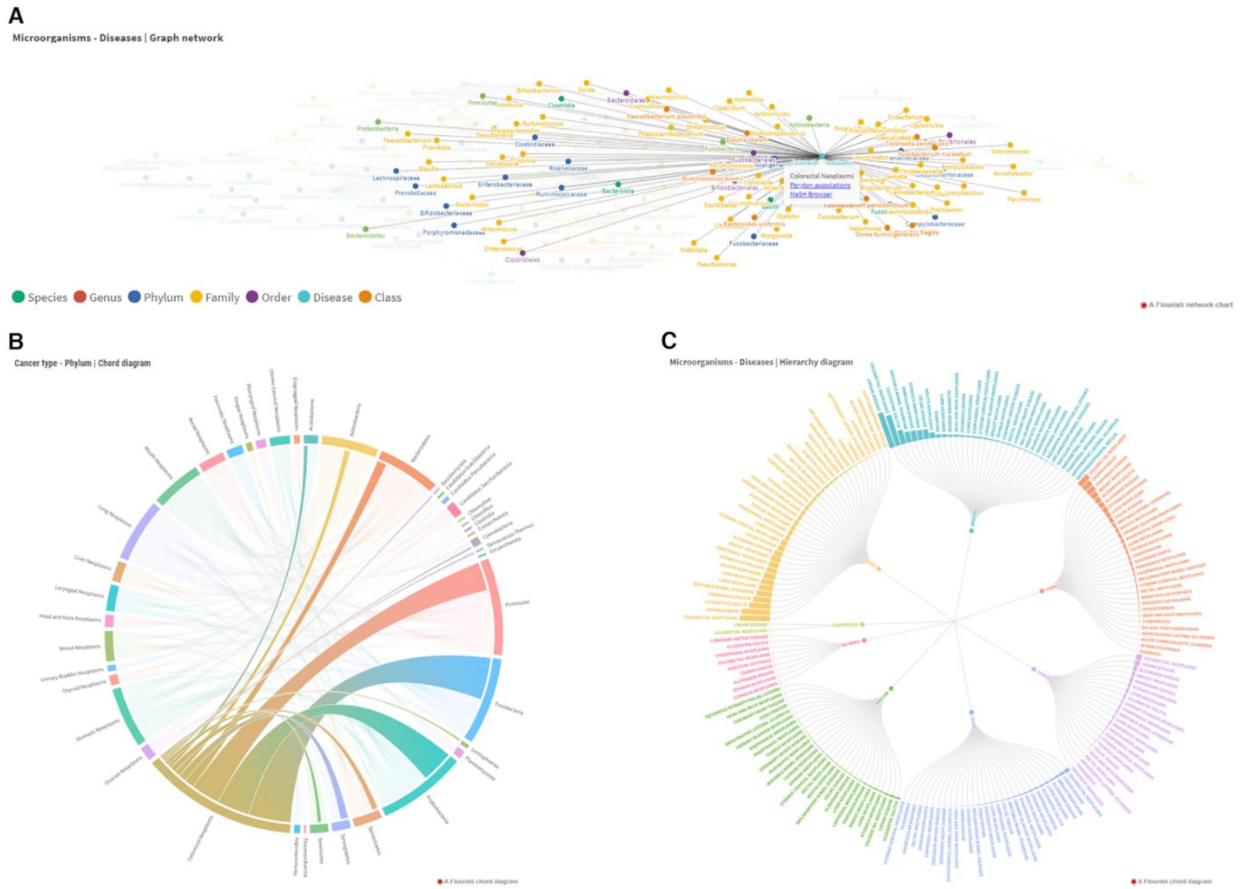


Figure 41. Visualization options offered in Peryton. (A) In Graph network, available associations are depicted in an interactive network. Users can explore the graph, highlight nodes of interest and filter-in and -out taxonomic ranks, according to taste. Importantly, by moving a selected node, connected nodes will move as well, with velocity depending on their connectivity status, allowing fast identification of hub or unitary nodes. Via pop-up boxes, each node in Graph network directs to its corresponding query results on the main 'Associations' page and NCBI Taxonomy or MeSH Browser, accordingly. (B) Chord diagram provides an interactive view of available cancer-related associations. Diseases and microorganisms are deployed along the circle's arcs, and chords of width relative to the number of existing associations depict connections. Users may select on or more components (i.e. arcs and/or chords) to highlight them permanently, or hover over them to highlight them temporarily. (C) Users can utilize Hierarchy diagram to browse Peryton's content in a hierarchically structured manner. For each taxonomic rank, numbers of available associations are provided as bars surrounding the circle. By selecting on bars and/or taxonomic ranks, a zoomed-in depiction of the relevant content is offered, enabling focused examination on associations of interest. Deepest layers in the Hierarchy diagram are also inter-connected with the Associations page via hyperlinks in microbe-disease-specific pop-up boxes.

CHAPTER 8 – DISCUSSION & CONCLUSION

The non-coding RNA landscape constitutes a revolutionary, rapidly evolving research field. The massive production of enormous amounts of data from Next Generation Sequencing experiments contributes to our gradually deeper understanding of non-coding RNA biology and at the same time deems bioinformatic analyses indispensable in almost every investigation. During this dissertation, effort was placed on three active non-coding RNA topics: sRNA-Seq quantification, miRNA interactomics and the study of miRNA editing.

Small RNA-Seq experimental datasets require caution during alignment and quantification due to technical obstacles arising from small read and transcript lengths, variations and post-transcriptional modifications. Multimaps analysis of sRNA-Seq datasets revealed a substantial proportion of multimapping reads in annotated (same and different sRNA species) and unannotated loci. Manatee sRNA-Seq quantification algorithm simultaneously utilizes available annotation and density information from uniquely aligned reads towards optimized multimapping read placement, improving the accuracy of small RNA quantification. Furthermore, its results facilitate downstream discovery of novel expressed small RNAs from loci that lack annotation, deeming it a tool of choice for various sRNA-Seq-based research efforts.

The comprehensive cataloguing of the miRNA interactome is considered pivotal to miRNA research efforts. DIANA-Tarbase v8.0 is the latest version of a reference miRNA-mRNA resource with a more-than-a-decade-long history, and comprises approximately one million entries. On the same lines, LncBase v3.0 showcases incremental improvements in the cataloguing of tissue-specific, experimentally supported miRNA-lncRNA interactions, by indexing ~240,000 interactions. The development and deployment of microCLIP, a cutting-edge AGO-CLIP-Seq data analysis algorithm, yields a plethora of *bona fide* miRNA binding events. LncRNA expression profiles and subcellular localization information enable studying of lncRNA sponge-functions and tissue-specific analyses of the competing endogenous RNA interactome. The extensive interconnection of DIANA-TarBase v8.0 and DIANA-LncBase v3.0 with DIANA-tools and external resources facilitates integration of multiple sources and a user-tailored experience. Both databases are important tools for both the experimentalists and the theorists of the biological community; they empower specific experimental investigations, *in silico* explorations and context-specific studies of the RNA interactome.

RNA editing is a crucial post-transcriptional modification that dynamically enhances organismic plasticity. Challenges is the identification of RNA editing events include the discrimination from sequencing errors and from genomic variants at the individual- or even the tissue-level (i.e. somatic mutations). A number of miRNA RNA editing identification investigations utilize only variant information from existing databases, potentially filtering out strong and relevant events. An A-to-Z pipeline for RNA editing identification was developed and utilized to identify editing effects in healthy and disease brain regions.

In post-identification analysis, a sample-based filter (>30% of study's samples) was employed to demarcate consistent editing events and the stable numbers of resulting events that were obtained in same tissue studies with low small or large difference in replicates (i.e. 16 *vs.* 30 and 18 *vs.* 512, respectively), indicate its robustness. A prominence in the number of A-to-I editing events and total number of A-to-I edited bases was observed in brain tissues, it was however less profound in glioma cancer samples and was followed by minimal events in lymphocytes. Enrichment of the miRNA seed region in A-to-I events was observed.

Consistent miRNA editing events appear to be shared by a number of tissues and between healthy and disease condition in the utilized, limited set of studies. Yet, variability in event frequency appeared to hold discriminatory capacity in the case of A-to-I events. Notably, unsupervised clustering based on the frequency of A-to-I events occurring on mature miRNAs resulted in two large clusters: one containing samples from cerebellum tissue and prefrontal cortex tissue from two studies and one that consisted of primary and recurrent lower grade glioma samples and hippocampus healthy tissue. Hippocampus samples formed a distinct group inside the "disease" cluster. The same analyses conducted using consistent C-to-U events did not yield similar outcomes; no coherence was observed between sample groups. Correlation analyses between RNA editing percentage and the abundance metrics (i.e. RPM values) of the edited miRNA and of its opposite-strand mature form highlighted one A-to-I case (miR-497-3p:20) where editing frequency correlated positively, strongly and significantly with the abundance of the opposite mature form, indicating possible implication of miRNA editing with the selection of the mature form during miRNA biogenesis.

Characteristics of the identified A-to-I and C-to-U events were contrasted under structural and functional scopes. A-to-I editing in mature miRNA sites appeared to mildly increase the stability of precursor miRNA hairpins and this effect was more prominent in glioma samples relative to healthy brain tissue. On the other hand, C-to-U events appeared to mildly increase the hairpin free energy (i.e. decrease its stability). miRNA targeting capacity analyses were deployed for A-to-I events with use of target prediction software. A remarkable difference in targeting repertoires of edited miRNAs and a mild (yet significant) drop in targeting efficacy were measured. Pathway analyses utilizing combined un-edited and edited miRNA targets hinted that edited miRNA targets further enriched pathways that seemed targeted by un-edited miRNA forms, however a more comprehensive analysis of this issue is required, employing direct interaction data and a wider array of conditions.

This ongoing effort will be expanded by employing thousands of available sRNA-Seq samples. Integration of datasets from other types of high-throughput experiments might empower these findings and yield fruitful results of basic or translational value. RNA editing and the rest of RNA modifications constitute a still understudied field. Extensive and innovative efforts will be required to integrate these phenomena soundly in future quests to better comprehend the complex nature of the RNA Intermedium.

CHAPTER 9 – PUBLICATIONS AND ACADEMIC ACTIVITY

9.1 Scientific Publications

During my doctoral dissertation, I participated in 7 studies which were published in scientific journals. The main research topics I have contributed in are: sRNA-Seq analysis, analysis and non-coding-RNA-aware re-annotation of microarrays, non-coding and coding RNA quantification, differential abundance analysis and statistical interrogation of biological pathway enrichment.

The published studies are presented below grouped per research sub-field:

Small RNA quantification from sRNA-Seq experimental datasets

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

Databases cataloguing interactions of miRNAs with coding and non-coding RNAs

- D Karagkouni, MD Paraskevopoulou, **S Tastsoglou**, G Skoufos, A Karavangeli, V Pierros, E Zacharopoulou, AG Hatzigeorgiou. *DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts*. Nucleic Acids Research, Volume 48, Issue D1 (2020). https://doi.org/10.1093/nar/gkz1036
- D Karagkouni, MD Paraskevopoulou, S Chatzopoulos, IS Vlachos, **S Tastsoglou**, I Kanellos, D Papadimitriou, I Kavakiotis, S Maniou, G Skoufos, T Vergoulis, T Dalamagas, AG Hatzigeorgiou. *DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions*. Nucleic Acids Research, Volume 46, Issue D1 (2018). https://doi.org/10.1093/nar/gkx1141

CLIP-guided miRNA target prediction

- MD Paraskevopoulou, D Karagkouni, IS Vlachos, **S Tastsoglou** & AG Hatzigeorgiou. *microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions*. Nat Commun 9, 3601 (2018). https://doi.org/10.1038/s41467-018-06046-y

Database of microbe-disease associations

- G Skoufos, F Kardaras, A Alexiou, I Kavakiotis, A Lambropoulou, V Kotsira, **S Tastsoglou**, *AG Hatzigeorgiou. Peryton: a manual collection of experimentally supported microbe-disease associations*. Nucleic Acids Research, Accepted for publication (Sept. 29, 2020).

Analysis of nano-vaccine efficiency against *Leishmania infantum*

- M Agallou, E Athanasiou, O Kammona, **S Tastsoglou**, AG Hatzigeorgiou, C Kiparissides, E Karagouni. *Transcriptome Analysis Identifies Immune Markers Related to Visceral Leishmaniasis Establishment in the Experimental Model of BALB/c Mice.* Front. Immunol. (2019). https://doi.org/10.3389/fimmu.2019.02749
- E Athanasiou, M Agallou, **S Tastsoglou**, O Kammona, AG Hatzigeorgiou, C Kiparissides, E Karagouni. *A Poly(Lactic-co-Glycolic) Acid Nanovaccine Based on Chimeric Peptides from Different Leishmania infantum Proteins Induces Dendritic Cells Maturation and Promotes Peptide-Specific IFNγ-Producing CD8+ T Cells Essential for the Protection against Experimental Visceral Leishmaniasis.* Front. Immunol. (2017). https://doi.org/10.3389/fimmu.2017.00684

Book Chapters

- IS Vlachos, G Georgakilas, **S Tastsoglou**, MD Paraskevopoulou, D Karagkouni, AG Hatzigeorgiou. *Computational Challenges and -omics Approaches for the Identification of microRNAs and Targets*. Book chapter in "*Essentials of Noncoding RNA in Neuroscience*", p39-59 (2017). https://doi.org/10.1016/B978-0-12-804402-5.00003-0

9.2 Conferences and Seminars

I have presented research findings in 7 scientific conferences (4 national and 3 international), and contributed to the organization of the European Conference of Computational Biology (ECCB'18), as well as the 5th Postgraduate and Postdoctoral Researchers' Meeting of the Hellenic Pasteur Institute (2019). Also, I have presented (with and without other co-tutors) 4 seminary courses regarding *in silico* investigations of miRNA functions.

Appendices

Appendix A. Identified A-to-I RNA Editing Events

Lists of consistent A-to-I editing events on (pre-)miRNAs are provided per dataset below.

Table 16. Consistent A-to-I editing events in dataset SRP052236 (prefrontal cortex).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-379	10	miR-379-5p	5	100%
mir-589	66	miR-589-3p	6	100%
mir-411	20	miR-411-5p	5	100%
mir-1301	52	miR-1301-3p	5	100%
mir-301b	64	miR-301b-3p	20	100%
mir-497	83	miR-497-3p	20	97%
mir-381	52	miR-381-3p	4	90%
mir-337	66	miR-337-3p	6	90%
mir-1251	10	miR-1251-5p	6	90%
mir-3622a	52	miR-3622a-3p	3	80%
mir-200b	61	miR-200b-3p	5	70%
mir-664a	18	miR-664a-5p	8	70%
mir-7977	6	miR-7977	6	70%
mir-497	25	miR-497-5p	2	63%
mir-641	18	miR-641	3	63%
mir-3681	10	miR-3681-5p	2	60%
mir-376c	48	miR-376c-3p	6	53%
mir-320a	65	-	-	30%
mir-203b	64	miR-203b-3p	11	30%

Table 17. Consistent A-to-I editing events in dataset SRP063627 (prefrontal cortex).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-200b	61	miR-200b-3p	5	100%
mir-379	10	miR-379-5p	5	100%
mir-381	52	miR-381-3p	4	100%
mir-497	25	miR-497-5p	2	100%
mir-497	83	miR-497-3p	20	100%
mir-589	66	miR-589-3p	6	100%
mir-411	20	miR-411-5p	5	100%
mir-1301	52	miR-1301-3p	5	100%
mir-301b	64	miR-301b-3p	20	100%
mir-1251	10	miR-1251-5p	6	100%
mir-664a	18	miR-664a-5p	8	100%
mir-7977	6	miR-7977	6	100%
mir-376c	48	miR-376c-3p	6	88%
mir-337	66	miR-337-3p	6	88%
mir-3622a	52	miR-3622a-3p	3	88%
mir-125b-2	78	-	-	38%
mir-26a-2	13	-	-	38%
mir-3681	10	miR-3681-5p	2	38%
mir-338	64	-	-	31%

Table 18. Consistent A-to-I editing events in dataset SRP174906 (hippocampus).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-221	88	-	-	100%
mir-140	85	-	-	100%
mir-376a-1	9	miR-376a-5p	3	100%
mir-379	10	miR-379-5p	5	100%
mir-337	66	miR-337-3p	6	100%
mir-411	20	miR-411-5p	5	100%
mir-3622a	52	miR-3622a-3p	3	100%
mir-320a	65	-	-	80%
mir-381	52	miR-381-3p	4	80%
mir-1301	52	miR-1301-3p	5	80%
mir-301b	64	miR-301b-3p	20	80%
mir-7977	6	miR-7977	6	80%
mir-103a-2	71	-	-	60%
mir-589	66	miR-589-3p	6	60%
mir-641	18	miR-641	3	60%
mir-3144	50	miR-3144-3p	3	60%
let-7b	30	-	-	40%
mir-27b	82	-	-	40%
mir-3681	10	miR-3681-5p	2	40%

Table 19. A-to-I editing events in dataset SRP221185 (cerebellum). Consistency could not be evaluated due to lack of replicates.

Hairpin Name	Rel. Position	Mature Name	Rel. Position
mir-19b-1	76	miR-19b-3p	23
mir-23a	67	-	-
mir-26b	10	-	-
mir-33a	26	miR-33a-5p	21
mir-99a	13	miR-99a-5p	1
mir-103a-2	71	-	-
mir-107	73	-	-
mir-204	56	-	-
mir-27b	13	-	-
mir-30b	16	-	-
mir-136	37	miR-136-5p	23
mir-320a	65	-	-
mir-101-2	72	-	-
mir-376a-1	9	miR-376a-5p	3
mir-379	10	miR-379-5p	5
mir-381	52	miR-381-3p	4
mir-340	15	-	-
mir-324	39	-	-
mir-338	65	-	-
mir-497	25	miR-497-5p	2
mir-497	83	miR-497-3p	20
mir-539	18	miR-539-5p	10
mir-376a-2	15	miR-376a-2-5p	4
mir-589	66	miR-589-3p	6
mir-624	58	miR-624-3p	5
mir-411	20	miR-411-5p	5
mir-301b	64	miR-301b-3p	20
mir-1251	10	miR-1251-5p	6
mir-3622a	52	miR-3622a-3p	3
mir-7977	6	miR-7977	6

Table 20. Consistent A-to-I editing events in LGG primary tumor dataset.

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-381	52	miR-381-3p	4	100%
mir-589	66	miR-589-3p	6	100%
mir-411	20	miR-411-5p	5	96%
mir-3622a	52	miR-3622a-3p	3	85%
mir-379	10	miR-379-5p	5	59%
mir-1251	10	miR-1251-5p	6	50%
mir-376a-1	9	miR-376a-5p	3	46%
mir-151a	49	miR-151a-3p	3	39%

Table 21. Consistent A-to-I editing events in LGG recurrent tumor dataset.

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-381	52	miR-381-3p	4	100%
mir-589	66	miR-589-3p	6	100%
mir-411	20	miR-411-5p	5	100%
mir-3622a	52	miR-3622a-3p	3	72%
mir-379	10	miR-379-5p	5	50%
mir-376a-1	9	miR-376a-5p	3	39%
mir-1251	10	miR-1251-5p	6	33%

Table 22. Consistent A-to-I editing events in LCL dataset (GEUVADIS).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-7977	6	miR-7977	6	97%
mir-320a	65	-	-	49%
mir-3681	10	miR-3681-5p	2	35%
mir-589	66	miR-589-3p	6	31%

Appendix B. Identified C-to-U RNA Editing Events

Lists of consistent C-to-U editing events on (pre-)miRNAs are provided per dataset below.

Table 23. Consistent C-to-U editing events in dataset SRP052236 (prefrontal cortex).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
let-7b	29	-	-	100%
mir-30a	68	miR-30a-3p	22	100%
mir-221	87	miR-221-3p	23	100%
mir-134	30	-	-	100%
mir-326	80	-	-	100%
mir-148b	85	-	-	100%
mir-27a	71	miR-27a-3p	21	97%
mir-152	76	-	-	97%
mir-3200	76	-	-	97%
mir-423	76	-	-	93%
mir-328	70	-	-	90%
mir-760	70	-	-	90%
mir-874	69	-	-	87%
mir-1301	71	miR-1301-3p	24	73%
mir-143	81	miR-143-3p	21	67%
mir-92b	82	miR-92b-3p	22	67%
mir-652	82	-	-	63%
mir-148a	66	-	-	43%
mir-125b-1	75	miR-125b-1-3p	21	43%
let-7b	80	let-7b-3p	21	40%
mir-30e	80	miR-30e-3p	22	37%
mir-370	70	-	-	30%
mir-151a	69	-	-	30%
mir-488	72	miR-488-3p	21	30%
mir-486-2	23	-	-	30%

Table 24. Consistent C-to-U editing events in dataset SRP063627 (prefrontal cortex).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
let-7b	29	-	-	100%
mir-27a	71	miR-27a-3p	21	100%
mir-30a	68	miR-30a-3p	22	100%
mir-134	30	-	-	100%
mir-326	80	-	-	100%
mir-148b	85	-	-	100%
mir-1301	71	miR-1301-3p	24	100%
mir-3200	76	-	-	100%
mir-423	76	-	-	94%
mir-874	69	-	-	88%
mir-221	87	miR-221-3p	23	81%
mir-760	70	-	-	81%
mir-152	76	-	-	69%
mir-328	70	-	-	63%
mir-222	92	-	-	50%
mir-23b	80	miR-23b-3p	23	50%
mir-143	81	miR-143-3p	21	50%
let-7b	80	let-7b-3p	21	44%
mir-125b-1	75	miR-125b-1-3p	21	44%
mir-370	70	-	-	44%
mir-488	72	miR-488-3p	21	38%
mir-652	82	-	-	38%

Table 25. Consistent C-to-U editing events in dataset SRP174906 (hippocampus).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
let-7b	29	-	-	100%
let-7b	80	let-7b-3p	21	100%
mir-21	30	-	-	100%
mir-25	74	-	-	100%
mir-27a	71	miR-27a-3p	21	100%
mir-30a	68	miR-30a-3p	22	100%
mir-148a	66	-	-	100%
mir-221	87	miR-221-3p	23	100%
mir-23b	80	miR-23b-3p	23	100%
mir-27b	83	-	-	100%
mir-152	76	-	-	100%
mir-126	74	-	-	100%
mir-134	30	-	-	100%
mir-370	70	-	-	100%
mir-151a	69	-	-	100%
mir-148b	85	-	-	100%
mir-423	76	-	-	100%
mir-425	76	miR-425-3p	22	100%
mir-652	82	-	-	100%
mir-421	71	-	-	100%
mir-744	87	miR-744-3p	20	100%
mir-3200	76	-	-	100%
mir-132	81	-	-	80%
mir-126	36	-	-	80%
mir-127	46	-	-	80%
mir-106b	73	miR-106b-3p	22	80%
mir-30e	80	miR-30e-3p	22	80%
let-7b	81	let-7b-3p	22	60%
let-7f-1	84	let-7f-1-3p	22	60%
mir-212	93	-	-	60%
mir-222	92	-	-	60%
mir-128-1	73	-	-	60%
mir-361	27	miR-361-5p	22	60%
mir-330	82	-	-	60%
mir-328	70	-	-	60%
mir-1301	71	miR-1301-3p	24	60%
mir-3615	73	-	-	60%
mir-33a	20	miR-33a-5p	15	40%
mir-98	101	miR-98-3p	22	40%

mir-326	80	-	-	40%
mir-488	72	miR-488-3p	21	40%
mir-421	70	miR-421	23	40%
mir-770	43	-	-	40%
mir-874	69	-	-	40%
mir-374b	34	-	-	40%
mir-760	70	-	-	40%

Table 26. C-to-U editing events in dataset SRP221185 (cerebellum). Consistency could not be evaluated, due to lack of replicates.

Hairpin Name	Rel. Position	Mature Name	Rel. Position
let-7b	29	-	-
let-7b	80	let-7b-3p	21
let-7b	81	let-7b-3p	22
let-7d	82	let-7d-3p	21
let-7f-1	84	let-7f-1-3p	22
mir-21	30	-	-
mir-27a	72	-	-
mir-27a	71	miR-27a-3p	21
mir-30a	68	miR-30a-3p	22
mir-98	101	miR-98-3p	22
mir-221	87	miR-221-3p	23
mir-222	92	-	-
let-7g	83	-	-
mir-23b	80	miR-23b-3p	23
mir-27b	83	-	-
mir-128-1	73	-	-
mir-132	81	-	-
mir-152	76	-	-
mir-126	36	-	-
mir-126	74	-	-
mir-134	30	-	-
mir-185	71	miR-185-3p	22
mir-195	37	-	-
mir-99b	5	-	-
mir-30e	80	miR-30e-3p	22
mir-370	70	-	-
mir-330	41	-	-
mir-330	82	-	-
mir-328	70	-	-
mir-326	80	-	-
mir-148b	85	-	-
mir-331	82	-	-
mir-345	39	miR-345-5p	22
mir-346	43	-	-
mir-423	41	-	-
mir-423	76	-	-
mir-483	66	miR-483-3p	19
mir-488	72	miR-488-3p	21

mir-497	23	-	-
mir-652	82	-	-
mir-421	71	-	-
mir-1301	71	miR-1301-3p	24
mir-874	69	-	-
mir-744	87	miR-744-3p	20
mir-760	70	-	-
mir-3615	73	-	-
mir-3085	75	-	-

Table 27. Consistent C-to-U editing events in LGG primary tumor dataset.

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-27a	71	miR-27a-3p	21	100%
mir-30a	68	miR-30a-3p	22	100%
mir-30e	80	miR-30e-3p	22	100%
mir-326	80	-	-	88%
let-7b	29	-	-	74%
let-7b	81	let-7b-3p	22	62%
let-7b	80	let-7b-3p	21	52%
mir-134	30	-	-	48%
mir-744	87	miR-744-3p	20	45%
mir-92b	82	miR-92b-3p	22	45%

Table 28. Consistent C-to-U editing events in LGG recurrent tumor dataset.

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-27a	71	miR-27a-3p	21	100%
mir-30a	68	miR-30a-3p	22	100%
mir-30e	80	miR-30e-3p	22	100%
mir-326	80	-	-	78%
let-7b	29	-	-	61%
let-7b	81	let-7b-3p	22	61%
let-7b	80	let-7b-3p	21	61%
mir-134	30	-	-	50%
mir-92b	82	miR-92b-3p	22	50%
mir-744	87	miR-744-3p	20	50%
mir-361	27	miR-361-5p	22	44%
mir-132	81	-	-	39%

Table 29. Consistent C-to-U editing events in LCL dataset (GEUVADIS).

Hairpin Name	Rel. Position	Mature Name	Rel. Position	Sample Fraction
mir-148a	66	-	-	99%
mir-148b	85	-	-	89%
mir-92b	82	miR-92b-3p	22	87%
mir-27a	71	miR-27a-3p	21	83%
mir-486-2	23	-	-	67%
mir-423	76	-	-	58%
mir-BHRF1-1	27	-	-	47%
mir-BART2	24	miR-BART2-5p	22	43%
mir-423	41	-	-	41%
mir-221	87	miR-221-3p	23	41%
mir-92b	81	miR-92b-3p	21	35%
mir-222	92	-	-	33%
mir-30e	80	miR-30e-3p	22	31%

Acronyms - Abbreviations

5′	five-prime
3′	three-prime
A	adenine/adenosine
A-to-I	Adenosine-to-Inosine (RNA editing type)
ADAR	Adenine Deaminase Acting on double-stranded RNA, enzymes catalyzing A-to-I RNA editing
AGO	Argonaute protein
APOBEC	Apolipoprotein B mRNA Editing Catalytic Polypeptide-like, enzyme catalyzing C-to-U RNA editing
BAM	binary representation format of Sequence Alignment Map (SAM) files, containing alignments of sequencing reads against reference sequences (e.g. a genome)
C	cytosine/cytidine
C-to-U	Cytosine-to-Uracil (RNA editing type)
CDF	Cumulative Distribution Function
cDNA	complementary DNA
ceRNA	competing endogenous RNA (as in ceRNA interactions/network)
ChIP-Seq	chromatin immunoprecipitation and sequencing
circRNA	circular RNA
CLASH	crosslinking, ligation, and sequencing of hybrids
CLEAR-CLIP	covalent ligation of endogenous Argonaute-bound RNAs-CLIP
CLIP	cross linking and immunoprecipitation (e.g. CLIP-Seq)
COSMIC	reference database of somatic mutations (i.e. variants that occur in somatic cells and are not heritable)
DBP	DNA binding protein
dbSNP	reference database of SNPs
ddNTP	di-deoxyribonucleoside, a modified nucleoside that lacks a hydroxyl-group in the third carbon of the ribose and inhibits DNA elongation (i.e. ddATP, ddTTP, ddCTP, ddGTP)
dNTP	deoxyribonucleoside, used in elongation of a nascent DNA strand (i.e. dATP, dTTP, dCTP and dGTP)
DDBJ	DNA Databank of Japan
DGCR8	DiGeorge Syndrome Critical Region 8
DNA	deoxyribonucleic acid
EBV	Epstein-Barr virus
ENCODE	Encyclopaedia Of DNA Elements, consortium
Exon	expressed regions *(portmanteau)*
FASTA	text-based format for storing biological sequences
FASTQ	text-based format for storing biological sequences and their per-base sequencing qualities
G	guanine/guanosine
GEO	Gene Expression Omnibus, repository

GEUVADIS	Genetic EUropean VAriation in DISease consortium, which generated expression sequencing data for participants in 1000 Genomes Project
GTF	Gene Transfer Format, 1-based file format holding chromosome coordinates of genomic features (e.g. genes)
HITS-CLIP	high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
I	inosine
Intron	intragenic region *(portmanteau)*
IsomiR	miRNA isoform *(portmanteau)*
KS-test	Kolmogorov-Smirnov test
KSHV	Kaposi's sarcoma-associated herpesvirus
LGG	Lower Grade Glioma
LSU	large ribosomal subunit
lncRNA	long non-coding RNA
miRBase	reference database of miRNA sequences
miRNA	microRNA
miRtron	intronic miRNA *(portmanteau)*
MRE	miRNA Recognition Element
mRNA	messenger RNA
Mt	mitochondrion, mitochondrial
MWU	Mann Whitney U (statistical test)
ncRNA	non-coding RNA
NGS	Next Generation Sequencing
nt	nucleotides (DNA or RNA), measure of nucleic acid length in monomers
PAR-CLIP	photoactivatable ribonucleoside-assisted crosslinking and immunoprecipitation
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
phastCons	per-base conservation metric derived from multi-species alignment, utilizing neighboring nucleotide information
Phred	quality score of sequenced nucleobases, corresponding to an error probability P, $P = 10^{(-Q/10)}$
phyloP	per-base conservation metric derived from multi-species alignment
piRNA	PIWI-Interacting RNA
Pol	polymerase
PMF	probability mass function
RBP	RNA binding protein
RCI	Relative Concentration Index, metric of subcellular localization
RISC	RNA-Induced Silencing Complex
RIP	RNA immunoprecipitation, as in RIP-Seq
RMA	robust multi-array average, microarray probe set summarization technique
RMSD	Root-Mean-Squared-deviation metric
RNA	ribonucleic acid

RNAcentral	reference knowledgebase integrating ncRNA sources
RNA-Seq	RNA Sequencing
RPF	Ribosome-Protected Footprint, as in RPF-Seq
RPM	Reads Per Million, NGS sequencing normalized quantification metric
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism, typically a variant observed with at least 1% prevalence in a population
snRNA	small nuclear RNA
snRNPs	small nuclear ribonucleoprotein complexes
SRA	Sequence Read Archive
sRNA	small RNA
sRNA-Seq	small RNA Sequencing
SSU	small ribosomal subunit
T	thymine/thymidine
TF	transcription factor
TFBS	transcription factor binding site
TPM	Transcripts Per kilobase Million, NGS sequencing normalized quantification metric
tRF	tRNA-derived Fragment
tRNA	transfer RNA
TCGA	The Cancer Genome Atlas, consortium
U	uracil/uridine
UAR	uniquely aligned reads
UCSC	university of California Santa Cruz
UV	ultra violet (radiation)
VCF	Variant Call Format, file type used to store variant information
vtRNA	vault RNA
WGS	Whole Genome Sequencing
WXS	Whole Exome Sequencing

References

1. Crick, F.H. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138-163 (1958).
2. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).
3. Berg, J.M., Tymoczko, J.L. & Stryer, L. Biochemistry. (W. H. Freeman, 2010).
4. Pan, T. Modifications and functional genomics of human transfer RNA. *Cell research* **28**, 395-404 (2018).
5. Juhling, F. et al. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic acids research* **37**, D159-162 (2009).
6. Lowe, T.M. & Chan, P.P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic acids research* **44**, W54-57 (2016).
7. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955-964 (1997).
8. Chan, P.P. & Lowe, T.M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic acids research* **44**, D184-189 (2016).
9. Evans, M.E., Clark, W.C., Zheng, G. & Pan, T. Determination of tRNA aminoacylation levels by high-throughput sequencing. *Nucleic acids research* **45**, e133 (2017).
10. Rudinger-Thirion, J., Lescure, A., Paulus, C. & Frugier, M. Misfolded human tRNA isodecoder binds and neutralizes a 3' UTR-embedded Alu element. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E794-802 (2011).
11. Shen, Y. et al. Transfer RNA-derived fragments and tRNA halves: biogenesis, biological functions and their roles in diseases. *Journal of molecular medicine* **96**, 1167-1176 (2018).
12. Wilson, D.N. & Doudna Cate, J.H. in Cold Spring Harb Perspect Biol, Vol. 4 (2012).
13. Freed, E.F., Bleichert, F., Dutca, L.M. & Baserga, S.J. When ribosomes go bad: diseases of ribosome biogenesis. *Mol Biosyst* **6**, 481-493 (2010).
14. Sonenberg, N. & Hinnebusch, A.G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731-745 (2009).
15. Selmer, M. et al. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science (New York, N.Y.)* **313**, 1935-1942 (2006).
16. Nissen, P., Hansen, J., Ban, N., Moore, P.B. & Steitz, T.A. The structural basis of ribosome activity in peptide bond synthesis. *Science (New York, N.Y.)* **289**, 920-930 (2000).
17. Polikanov, Y.S., Melnikov, S.V., Soll, D. & Steitz, T.A. Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nature structural & molecular biology* **22**, 342-344 (2015).
18. Decatur, W.A. & Fournier, M.J. rRNA modifications and ribosome function. *Trends in biochemical sciences* **27**, 344-351 (2002).
19. Taoka, M. et al. The complete chemical structure of Saccharomyces cerevisiae rRNA: partial pseudouridylation of U2345 in 25S rRNA by snoRNA snR9. *Nucleic acids research* **44**, 8951-8961 (2016).
20. Birkedal, U. et al. Profiling of ribose methylations in RNA by high-throughput sequencing. *Angewandte Chemie* **54**, 451-455 (2015).
21. Lee, R.C., Feinbaum, R.L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).
22. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855-862 (1993).
23. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)* **294**, 853-858 (2001).

24. Lee, R.C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)* **294**, 862-864 (2001).

25. Griffiths-Jones, S. The microRNA Registry. *Nucleic acids research* **32**, D109-D111 (2004).

26. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic acids research* **47**, D155-D162 (2018).

27. Ha, M. & Kim, V.N. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology* **15**, 509-524 (2014).

28. O'Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology* **9** (2018).

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

30. Basyuk, E., Suavet, F., Doglio, A., Bordonné, R. & Bertrand, E. Human let-7 stem–loop precursors harbor features of RNase III cleavage products. *Nucleic acids research* **31**, 6593-6597 (2003).

31. Xie, M. et al. Mammalian 5'-capped microRNA precursors that generate a single microRNA. *Cell* **155**, 1568-1580 (2013).

32. Yang, J.S. et al. Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15163-15168 (2010).

33. Paraskevopoulou, M.D., Karagkouni, D., Vlachos, I.S., Tastsoglou, S. & Hatzigeorgiou, A.G. microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions. *Nat Commun* **9**, 3601 (2018).

34. Xi, Y. et al. Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. *Rna* **13**, 1668-1674 (2007).

35. Condrat, C.E. et al. miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. *Cells* **9** (2020).

36. Corsten, M.F. et al. Circulating MicroRNA-208b and MicroRNA-499 reflect myocardial damage in cardiovascular disease. *Circulation. Cardiovascular genetics* **3**, 499-506 (2010).

37. Wiedrick, J.T. et al. Validation of MicroRNA Biomarkers for Alzheimer's Disease in Human Cerebrospinal Fluid. *Journal of Alzheimer's disease : JAD* **67**, 875-891 (2019).

38. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766-D773 (2019).

39. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766-D773 (2018).

40. Yates, A.D. et al. Ensembl 2020. *Nucleic acids research* **48**, D682-D688 (2019).

41. Sun, Q., Hao, Q. & Prasanth, K.V. Nuclear long noncoding RNAs: key regulators of gene expression. *Trends in Genetics* **34**, 142-157 (2018).

42. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22**, 1775-1789 (2012).

43. Giannakakis, A. et al. Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. *Scientific reports* **5**, 9737 (2015).

44. Carlevaro-Fita, J. & Johnson, R. Global positioning system: understanding long noncoding RNAs through subcellular localization. *Molecular cell* **73**, 869-883 (2019).

45. Liu, S.J. et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome biology* **17**, 67 (2016).

46. Zhang, Y. & Tycko, B. Monoallelic expression of the human H19 gene. *Nature genetics* **1**, 40-44 (1992).

47. Heard, E. & Disteche, C.M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes & development* **20**, 1848-1867 (2006).

48. Chen, L.L. & Carmichael, G.G. Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Molecular cell* **35**, 467-478 (2009).

49. Orom, U.A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

50. Cesana, M. et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358-369 (2011).

51. Kallen, A.N. et al. The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Molecular cell* **52**, 101-112 (2013).

52. Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333-338 (2013).

53. An, Y., Furber, K.L. & Ji, S. Pseudogenes regulate parental gene expression via ceRNA network. *Journal of cellular and molecular medicine* **21**, 185-192 (2017).

54. Glenfield, C. & McLysaght, A. Pseudogenes Provide Evolutionary Evidence for the Competitive Endogenous RNA Hypothesis. *Molecular biology and evolution* **35**, 2886-2899 (2018).

55. Salzman, J., Gawad, C., Wang, P.L., Lacayo, N. & Brown, P.O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PloS one* **7**, e30733 (2012).

56. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *Rna* **20**, 1666-1670 (2014).

57. Rybak-Wolf, A. et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Molecular cell* **58**, 870-885 (2015).

58. You, X. et al. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nature neuroscience* **18**, 603-610 (2015).

59. Pamudurti, N.R. et al. Translation of circRNAs. *Molecular cell* **66**, 9-21. e27 (2017).

60. Kleaveland, B., Shi, C.Y., Stefano, J. & Bartel, D.P. A network of noncoding regulatory RNAs acts in the mammalian brain. *Cell* **174**, 350-362. e317 (2018).

61. Du, W.W. et al. Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nucleic acids research* **44**, 2846-2858 (2016).

62. Suenkel, C., Cavalli, D., Massalini, S., Calegari, F. & Rajewsky, N. A highly conserved circular RNA is required to keep neural cells in a progenitor state in the mammalian brain. *Cell Reports* **30**, 2170-2179. e2175 (2020).

63. Baek, D. et al. The impact of microRNAs on protein output. *Nature* **455**, 64-71 (2008).

64. Vlachos, I.S. & Hatzigeorgiou, A.G. Online resources for miRNA analysis. *Clinical biochemistry* **46**, 879-900 (2013).

65. Giza, D.E., Vasilescu, C. & Calin, G.A. Key principles of miRNA involvement in human diseases. *Discoveries* **2**, e34 (2014).

66. Lee, Y.S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development* **23**, 2639-2649 (2009).

67. Langenberger, D., Bermudez-Santana, C.I., Stadler, P.F. & Hoffmann, S. Identification and classification of small RNAs in transcriptome sequence data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 80-87 (2010).

68. Stepanov, G.A. et al. Regulatory role of small nucleolar RNAs in human diseases. *BioMed research international* **2015**, 206849 (2015).

69. Huang, N. et al. Natural display of nuclear-encoded RNA on the cell surface and its impact on cell interaction. *Genome biology* **21**, 1-23 (2020).

70. Baillat, D. et al. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* **123**, 265-276 (2005).

71. Fischer, U. & Luhrmann, R. An essential signaling role for the m3G cap in the transport of U1 snRNP to the nucleus. *Science* **249**, 786-790 (1990).

72. Matera, A.G. & Wang, Z. A day in the life of the spliceosome. *Nature reviews Molecular cell biology* **15**, 108-121 (2014).

73. Valadkhan, S., Mohammadi, A., Jaladat, Y. & Geisler, S. Protein-free small nuclear RNAs catalyze a two-step splicing reaction. *Proceedings of the National Academy of Sciences* **106**, 11901-11906 (2009).

74. Bratkovič, T., Božič, J. & Rogelj, B. Functional diversity of small nucleolar RNAs. *Nucleic acids research* **48**, 1627-1651 (2020).

75. Bratkovič, T. & Rogelj, B. Biology and applications of small nucleolar RNAs. *Cellular and molecular life sciences : CMLS* **68**, 3843-3851 (2011).

76. Jorjani, H. et al. An updated human snoRNAome. *Nucleic acids research* **44**, 5068-5082 (2016).

77. Dudnakova, T., Dunn-Davies, H. & Peters, R. Mapping targets for small nucleolar RNAs in yeast. **3**, 120 (2018).

78. Vitali, P. & Kiss, T. Cooperative 2'-O-methylation of the wobble cytidine of human elongator tRNA(Met)(CAT) by a nucleolar and a Cajal body-specific box C/D RNP. *Genes & development* **33**, 741-746 (2019).

79. Klattenhoff, C. & Theurkauf, W. Biogenesis and germline functions of piRNAs. *Development* **135**, 3-9 (2008).

80. Aravin, A. et al. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203-207 (2006).

81. Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103 (2007).

82. Nishimasu, H. et al. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature* **491**, 284-287 (2012).

83. Huang, X., Tóth, K.F. & Aravin, A.A. piRNA biogenesis in Drosophila melanogaster. *Trends in Genetics* **33**, 882-894 (2017).

84. Ernst, C., Odom, D.T. & Kutter, C. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nature communications* **8**, 1-10 (2017).

85. Molaro, A. et al. Two waves of de novo methylation during mouse germ cell development. *Genes & development* **28**, 1544-1549 (2014).

86. Lee, Y.S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development* **23**, 2639-2649 (2009).

87. Keam, S.P. & Hutvagner, G. tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. *Life* **5**, 1638-1651 (2015).

88. Kumar, P., Mudunuri, S.B., Anaya, J. & Dutta, A. tRFdb: a database for transfer RNA fragments. *Nucleic acids research* **43**, D141-D145 (2015).

89. Kumar, P., Kuscu, C. & Dutta, A. Biogenesis and function of transfer RNA-related fragments (tRFs). *Trends in biochemical sciences* **41**, 679-689 (2016).

90. Sun, C. et al. Roles of tRNA-derived fragments in human cancers. *Cancer Letters* **414**, 16-25 (2018).

91. Cohn, W.E. Pseudouridine, a carbon-carbon linked ribonucleoside in ribonucleic acids: isolation, structure, and chemical characteristics. *The Journal of biological chemistry* **235**, 1488-1498 (1960).

92. Boccaletto, P. et al. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic acids research* **46**, D303-D307 (2018).

93. Helm, M. & Motorin, Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature Reviews Genetics* **18**, 275-291 (2017).

94. Roundtree, I.A., Evans, M.E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **169**, 1187-1200 (2017).

95. Rottman, F.M., Desrosiers, R.C. & Friderici, K. in Progress in nucleic acid research and molecular biology, Vol. 19 21-38 (Elsevier, 1977).

96. Bohnsack, M.T. & Sloan, K.E. Modifications in small nuclear RNAs and their roles in spliceosome assembly and function. *Biological chemistry* **399**, 1265-1276 (2018).

97. Kishore, S. et al. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome biology* **14**, R45 (2013).

98. Kawahara, Y. et al. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315**, 1137-1140 (2007).

99. Zinshteyn, B. & Nishikura, K. Adenosine-to-inosine RNA editing. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **1**, 202-209 (2009).

100. Limbach, P.A. & Paulines, M.J. Going global: the new era of mapping modifications in RNA. *Wiley Interdisciplinary Reviews: RNA* **8**, e1367 (2017).

101. Jia, G. et al. N 6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature chemical biology* **7**, 885-887 (2011).

102. Chen, S.-H. et al. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* **238**, 363-366 (1987).

103. Powell, L.M. et al. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, 831-840 (1987).

104. Backus, J.W. & Smith, H.C. Apolipoprotein B mRNA sequences 3′ of the editing site are necessary and sufficient for editing and editosome assembly. *Nucleic acids research* **19**, 6781-6786 (1991).

105. Richardson, N., Navaratnam, N. & Scott, J. Secondary Structure for the Apolipoprotein B mRNA Editing Site AU-BINDING PROTEINS INTERACT WITH A STEM LOOP. *Journal of Biological Chemistry* **273**, 31707-31717 (1998).

106. Sharma, S. & Baysal, B.E. Stem-loop structure preference for site-specific RNA editing by APOBEC3A and APOBEC3G. *PeerJ* **5**, e4136 (2017).

107. Blanc, V. & Davidson, N.O. C-to-U RNA editing: mechanisms leading to genetic diversity. *Journal of Biological Chemistry* **278**, 1395-1398 (2003).

108. Blanc, V. & Davidson, N.O. APOBEC-1-mediated RNA editing. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **2**, 594-602 (2010).

109. Savva, Y.A., Rieder, L.E. & Reenan, R.A. The ADAR protein family. *Genome biology* **13**, 1-10 (2012).

110. Bass, B.L. & Weintraub, H. An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell* **55**, 1089-1098 (1988).

111. Porath, H.T., Knisbacher, B.A., Eisenberg, E. & Levanon, E.Y. Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. *Genome biology* **18**, 1-12 (2017).

112. Melcher, T. et al. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *Journal of Biological Chemistry* **271**, 31795-31798 (1996).

113. Desterro, J.M. et al. Dynamic association of RNA-editing enzymes with the nucleolus. *Journal of cell science* **116**, 1805-1818 (2003).

114. Strehblow, A., Hallegger, M. & Jantsch, M.F. Nucleocytoplasmic distribution of human RNA-editing enzyme ADAR1 is modulated by double-stranded RNA-binding domains, a leucine-rich export signal, and a putative dimerization domain. *Molecular biology of the cell* **13**, 3822-3835 (2002).

115. Picardi, E. et al. Profiling RNA editing in human tissues: towards the inosinome Atlas. *Scientific reports* **5**, 14941 (2015).

116. Gallo, A., Vukic, D., Michalik, D., O'Connell, M.A. & Keegan, L.P. ADAR RNA editing in human disease; more to it than meets the I. *Human genetics* **136**, 1265-1278 (2017).
117. Licht, K. et al. Inosine induces context-dependent recoding and translational stalling. *Nucleic acids research* **47**, 3-14 (2019).
118. Kung, S.-S., Chen, Y.-C., Lin, W.-H., Chen, C.-C. & Chow, W.-Y. Q/R RNA editing of the AMPA receptor subunit 2 (GRIA2) transcript evolves no later than the appearance of cartilaginous fishes. *FEBS letters* **509**, 277-281 (2001).
119. Higuchi, M. et al. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78-81 (2000).
120. Hideyama, T. et al. Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiology of disease* **45**, 1121-1128 (2012).
121. Cenci, C. et al. Down-regulation of RNA editing in pediatric astrocytomas ADAR2 editing activity inhibits cell migration and proliferation. *Journal of Biological Chemistry* **283**, 7251-7260 (2008).
122. Kawahara, Y. et al. Dysregulated editing of serotonin 2C receptor mRNAs results in energy dissipation and loss of fat mass. *Journal of Neuroscience* **28**, 12834-12844 (2008).
123. Morabito, M.V. et al. Mice with altered serotonin 2C receptor RNA editing display characteristics of Prader–Willi syndrome. *Neurobiology of disease* **39**, 169-180 (2010).
124. Kanata, E. et al. RNA editing alterations define manifestation of prion diseases. *Proceedings of the National Academy of Sciences* **116**, 19727-19735 (2019).
125. Jin, Y. et al. RNA editing and alternative splicing of the insect nAChR subunit alpha6 transcript: evolutionary conservation, divergence and regulation. *BMC evolutionary biology* **7**, 98 (2007).
126. Solomon, O. et al. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR). *Rna* **19**, 591-604 (2013).
127. Breen, M.S. et al. Global landscape and genetic regulation of RNA editing in cortical samples from individuals with schizophrenia. *Nature neuroscience* **22**, 1402-1412 (2019).
128. Brümmer, A., Yang, Y., Chan, T.W. & Xiao, X. Structure-mediated modulation of mRNA abundance by A-to-I editing. *Nature communications* **8**, 1-13 (2017).
129. Herbert, A. ADAR and immune silencing in cancer. *Trends in cancer* **5**, 272-282 (2019).
130. Palazzo, A.F. & Lee, E.S. Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. *Frontiers in genetics* **9**, 440 (2018).
131. Liu, Y., Nie, H., Liu, H. & Lu, F. Poly (A) inclusive RNA isoform sequencing (PAIso− seq) reveals wide-spread non-adenosine residues within RNA poly (A) tails. *Nature communications* **10**, 1-13 (2019).
132. LUCIANO, D.J., MIRSKY, H., VENDETTI, N.J. & MAAS, S. RNA editing of a miRNA precursor. *Rna* **10**, 1174-1177 (2004).
133. Yang, W. et al. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nature structural & molecular biology* **13**, 13-21 (2006).
134. Scadden, A.D. The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage. *Nature structural & molecular biology* **12**, 489-496 (2005).
135. Kawahara, Y. et al. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315**, 1137-1140 (2007).
136. de Hoon, M.J. et al. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome research* **20**, 257-264 (2010).
137. Alon, S. et al. Systematic identification of edited microRNAs in the human brain. *Genome research* **22**, 1533-1540 (2012).
138. Wang, Y. et al. Systematic characterization of A-to-I RNA editing hotspots in microRNAs across human cancers. *Genome research* **27**, 1112-1125 (2017).

139. Holley, R.W., Madison, J.T. & Zamir, A. A new method for sequence determination of large oligonucleotides. *Biochemical and Biophysical Research Communications* **17**, 389-394 (1964).

140. Holley, R.W. et al. Structure of a Ribonucleic Acid. *Science* **147**, 1462-1465 (1965).

141. Maxam, A.M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 560-564 (1977).

142. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).

143. Smith, L.M. et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).

144. Prober, J.M. et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336-341 (1987).

145. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic acids research* **18**, 1415-1419 (1990).

146. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

147. Lander, E.S. et al. Initial sequencing and analysis of the human genome. (2001).

148. Hultman, T., Stahl, S., Hornes, E. & Uhlen, M. Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. *Nucleic acids research* **17**, 4937-4946 (1989).

149. Nyrén, P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical biochemistry* **167**, 235-238 (1987).

150. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363-365 (1998).

151. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research* **34**, e22-e22 (2006).

152. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research* **36**, e25-e25 (2008).

153. Chen, P. et al. Probing single DNA molecule transport using fabricated nanopores. *Nano letters* **4**, 2293-2298 (2004).

154. Ding, T. et al. DNA nanotechnology assisted nanopore-based analysis. *Nucleic acids research* **48**, 2791-2806 (2020).

155. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193-2198 (2019).

156. Marine, R. et al. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Applied and environmental microbiology* **77**, 8071-8079 (2011).

157. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology* **27**, 182-189 (2009).

158. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA* **8**, e1364 (2017).

159. Levin, J.Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* **7**, 709-715 (2010).

160. Handzlik, J.E., Tastsoglou, S., Vlachos, I.S. & Hatzigeorgiou, A.G. Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data. *Scientific reports* **10**, 705 (2020).

161. Mjelle, R., Aass, K.R., Sjursen, W., Hofsli, E. & Sætrom, P. sMETASeq: combined profiling of microbiota and host small RNAs. *iScience*, 101131 (2020).
162. Muller, H., Marzi, M.J. & Nicassio, F. IsomiRage: from functional classification to differential expression of miRNA isoforms. *Frontiers in bioengineering and biotechnology* **2**, 38 (2014).
163. Davis, C.A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research* **46**, D794-D801 (2018).
164. Licatalosi, D.D. et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469 (2008).
165. Mittal, N. & Zavolan, M. Seq and CLIP through the miRNA world. *Genome biology* **15**, 202 (2014).
166. Chi, S.W., Zang, J.B., Mele, A. & Darnell, R.B. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* **460**, 479-486 (2009).
167. Hafner, M. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141 (2010).
168. Karagkouni, D. et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic acids research* **46**, D239-D245 (2018).
169. Karagkouni, D. et al. DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic acids research* **48**, D101-D110 (2020).
170. Andrews S., Vol. 2019 (2010).
171. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
172. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
173. Zerbino, D.R. et al. Ensembl 2018. *Nucleic acids research* **46**, D754-D761 (2018).
174. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42**, D68-73 (2014).
175. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210 (2002).
176. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
177. Baras, A.S. et al. miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy. *PloS one* **10**, e0143066 (2015).
178. Johnson, N.R., Yeoh, J.M., Coruh, C. & Axtell, M.J. Improved Placement of Multi-mapping Small RNAs. *G3* **6**, 2103-2111 (2016).
179. Barturen, G. et al. sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods in Next Generation Sequencing* **1** (2014).
180. Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* **40**, 37-52 (2012).
181. Καραγκούνη, Δ.Ε. Υπολογιστική ανάλυση των λειτουργιών των μη κωδικών μεταγράφων στη γονιδιωματική ρύθμιση. (2019).
182. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991-D995 (2012).
183. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research* **40**, D54-D56 (2011).
184. R Development Core Team (R Foundation for Statistical Computing, Vienna, Austria; 2016).
185. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315 (2004).

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

187. Ritchie, M.E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

188. Flicek, P. et al. Ensembl 2012. *Nucleic acids research* **40**, D84-D90 (2011).

189. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115-121 (2015).

190. Vlachos, I.S. et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res* **43**, D153-159 (2015).

191. The, R.C. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic acids research* **47**, D221-D229 (2019).

192. Paraskevopoulou, M.D. et al. DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows. *Nucleic acids research* **41**, W169-W173 (2013).

193. Paraskevopoulou, M.D. et al. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic acids research* **44**, D231-D238 (2016).

194. Vlachos, I.S. et al. DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. *Nucleic acids research* **43**, W460-W466 (2015).

195. Chou, C.-H. et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research* **44**, D239-D247 (2015).

196. Xiao, F. et al. miRecords: an integrated resource for microRNA–target interactions. *Nucleic acids research* **37**, D105-D110 (2008).

197. Grosswendt, S. et al. Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Molecular cell* **54**, 1042-1054 (2014).

198. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42**, D68-D73 (2013).

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

200. Tateno, Y. et al. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic acids research* **30**, 27-30 (2002).

201. Andrews, S. & FastQC, A. A quality control tool for high throughput sequence data. 2010. *URL: http://www. bioinformatics. babraham. ac. uk/projects/fastqc* (2015).

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

203. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, pp. 10-12 (2011).

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

205. Paraskevopoulou, M.D. et al. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic acids research* **44**, D231-238 (2016).

206. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F. & Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 716-721 (2008).

207. Liao, Q. et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic acids research* **39**, 3864-3878 (2011).

208. Irizarry, R.A., Gautier, L., Huber, W. & Bolstad, B., Edn. 1.60.0 (2019).

209. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14**, 417-419 (2017).

210. Mas-Ponte, D. et al. LncATLAS database for subcellular localization of long noncoding RNAs. *Rna* **23**, 1080-1087 (2017).

211. Cunningham, F. et al. Ensembl 2019. *Nucleic acids research* **47**, D745-D751 (2019).

212. O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733-745 (2016).

213. Cabili, M.N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25**, 1915-1927 (2011).

214. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).

215. Landrum, M.J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **46**, D1062-D1067 (2018).

216. Tate, J.G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).

217. Vlachos, I.S. et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic acids research* **43**, W460-466 (2015).

218. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic acids research* **47**, D155-D162 (2019).

219. Haeussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic acids research* **47**, D853-D858 (2019).

220. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118 (2013).

221. Pages, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: String objects representing biological sequences, and matching algorithms. *R package version* **2**, 10.18129 (2016).

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

223. Krueger, F. Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* **516**, 517 (2015).

224. Morgan, M. et al. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607-2608 (2009).

225. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* **12**, 115-121 (2015).

226. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).

227. Paul, D. et al. A-to-I editing in human miRNAs is enriched in seed sequence, influenced by sequence contexts and significantly hypoedited in glioblastoma multiforme. *Scientific reports* **7**, 2466 (2017).

228. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB* **6**, 26 (2011).

229. Agallou, M. et al. Transcriptome analysis identifies immune markers related to visceral leishmaniasis establishment in the experimental model of BALB/c mice. *Frontiers in immunology* **10**, 2749 (2019).

230. Skoufos, G. et al. Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic acids research* (2020).