ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εντοπισμός RNA τροποποιήσεων σε δεδομένα μεταγραφώματος και εκτίμηση της επίδρασής τους στους στόχους των miRNA

Μάριος Σ. Μηλιώτης

Επιβλέπουσα: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER'S THESIS

RNA editing identification in transcriptomics data and assessment of impact in miRNA targeting

Marios S. Miliotis

Supervisor: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly

ATHENS

NOVEMBER 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εντοπισμός RNA τροποποιήσεων σε δεδομένα μεταγραφώματος και εκτίμηση της επίδρασής τους στους στόχους των miRNA

Μάριος Σ. Μηλιώτης

Α.Μ.: ΠΙΒ0183

ΕΠΙΒΛΕΠΟΥΣΑ: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας
Δρ. Martin Reczko, Ερευνητής Καθηγητής, Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών “Αλέξανδρος Φλέμινγκ”
Δρ. Αλέξανδρος Δημόπουλος, Μεταδιδακτορικός ερευνητής, Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών “Αλέξανδρος Φλέμινγκ”

ΝΟΕΜΒΡΙΟΣ 2019

MASTER'S THESIS

RNA editing identification in transcriptomics data and assessment of impact in miRNA targeting

Marios S. Miliotis

SRN: ΠΙΒ0183

SUPERVISOR: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly

EXAMINATION COMMITTEE:

Prof. Artemis Hatzigeorgiou, Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly
Dr. Martin Reczko, Staff research scientist professor level at Biomedical Sciences Research Center "Alexander Fleming"
Dr. Alexandros Dimopoulos, Postdoctoral Researcher at Biomedical Sciences Research Center "Alexander Fleming"

NOVEMBER 2019

ΠΕΡΙΛΗΨΗ

Η μεταγραφική τροποποίηση είναι μια μεταγραφική/μετα-μεταγραφική διαδικασία, κατά την οποία ένα μόριο RNA υπόκειται στη μεταλλαγή της ακολουθίας του μέσω της εισαγωγής, της απαλοιφής ή της μεταβολής των βάσεων της. Στα μετάζωα, η πλειονότητα των μεταγραφικών τροποποιήσεων που συμβαίνουν αφορά τη μετατροπή του νουκλεοτιδίου αδενοσίνη (A) σε ινοσίνη (I), φαινόμενο που καταλύεται από τα μέλη της οικογένειας των γονιδίων των απαμινάσεων της αδενοσίνης (ADAR) που δρουν σε RNA με διπλή έλικα (dsRNA). Το φαινόμενο της μεταγραφικής τροποποίησης εμφανίζει σχετικά αυξημένη συχνότητα σε μόρια που φέρουν περιοχές ρετροτρανσποζονίων Alu στην ακολουθία τους.

Η τροποποίηση της κωδικής περιοχής των pre-mRNA μπορεί να οδηγήσει στην ενσωμάτωση διαφορετικού αμινοξέος κατά τη μετάφραση και να συμβάλει έτσι στην ποικιλότητα των πρωτεϊνικών προϊόντων και λειτουργιών. Ωστόσο, οι περισσότερες A-σε-I τροποποιήσεις απαντώνται σε μη κωδικές περιοχές των pre-mRNA και των mRNA, καθώς και σε μη κωδικά RNA. Οι μετατροπές στην UTR (μη μεταφραζόμενη περιοχή) ενός mRNA μπορούν να ρυθμίσουν τη μετάφραση, το μάτισμα και την αποικοδόμησή τους. Επίσης, τροποποιήσεις σε ακολουθίες microRNA (miRNA) και long non-coding RNA (lncRNA), καθώς και τροποποιήσεις στις θέσεις πρόσδεσής τους, μπορούν να επηρεάσουν τη βιογένεσή τους, την αναγνώριση των στόχων τους, τη δομή και τη σταθερότητά τους.

Στόχος αυτής της μελέτης είναι να γίνει σύγκριση μιας ομάδας εργαλείων για τον εντοπισμό RNA τροποποιήσεων, να διαχωριστούν τα πραγματικά συμβάντα μεταλλαγής στις 3'UTR των mRNA και να εκτιμηθεί η επίδρασή τους στην ειδικότητα και στην αποτελεσματικότητα της πρόσδεσής των miRNA.

Αρχικά, χρησιμοποιήθηκαν ζευγάρια από σύνολα δεδομένων αλληλούχισης του RNA και του DNA του ίδιου δείγματος ώστε να εντοπιστούν A-σε-I RNA τροποποιήσεις σε 3'UTR περιοχές. Η χρήση ζευγών έγινε με σκοπό να εξεταστούν συμβάντα σε επίπεδο δείγματος, αυξάνοντας έτσι την ειδικότητα. Το σύνολο των δεδομένων που χρησιμοποιήθηκε αποτελούνταν από 2 δείγματα για δοκιμή, 1 ADAR enzyme knockdown δείγμα για έλεγχο και τη RADAR, μία περιεκτική συλλογή A-σε-I δεδομένων στα μεταγραφώματα του ανθρώπου, του ποντικιού και της μύγας, με τους δύο προαναφερθέντες πόρους να αποτελούν τα δεδομένα αντικειμενικής αλήθειας για τη μελέτη. Στη συνέχεια, αναζητήθηκε ο καλύτερος αλγόριθμος στον εντοπισμό RNA τροποποιήσεων. Το ADAR knockdown δείγμα χρησιμοποιήθηκε ώστε να επισημανθούν τα υψηλά ψευδώς θετικά ποσοστά. Η σύγκριση περιλάμβανε το RES-Scanner, που χρησιμοποιεί τον Burrows-Wheeler aligner (BWA), το REDItools που τρέχει με τον aligner GSNAP και το RNAEditor, το οποίο εκτελέστηκε τόσο με τον BWA (προκαθορισμένη επιλογή) όσο και με τον GSNAP. Οι δύο πρώτοι αλγόριθμοι υποστηρίζουν εκ κατασκευής ζευγάρια RNA-DNA συνόλων δεδομένων, ενώ ο τρίτος τροποποιήθηκε ώστε να λαμβάνει υπόψιν και την DNA πληροφορία. Πιο σταθερή συμπεριφορά σε σχέση με την ειδικότητα και την ευαισθησία στα αποτελέσματα αναδείχθηκε να έχει το RNAEditor αξιοποιώντας τον aligner BWA.

Μετέπειτα, 3'UTR που βρέθηκαν να φέρουν τροποποίηση δόθηκαν ως είσοδος στους αλγορίθμους πρόβλεψης στόχων των miRNA ώστε να εκτιμηθούν στατιστικά διαφορές στις υπολογισμένες περιοχές πρόσδεσης που δημιουργήθηκαν εξαιτίας των φαινομένων τροποποίησης. Σε αυτό το στάδιο η σύγκριση επεκτάθηκε προσθέτοντας ένα ακόμα δείγμα με φυσιολογική (wild-type) έκφραση του ADAR από το ίδιο πείραμα με το ADAR

knockdown δείγμα. Συμβάντα τα οποία καταγράφηκαν σε 3'UTR χρησιμοποιήθηκαν ώστε να παραχθούν 2 ισάριθμα σύνολα από ακολουθίες, από τις οποίες 2062 ανήκαν σε περιοχές με υψηλό αριθμό διαδοχικών επαναλήψεων Alu και 144 σε non-Alu. Επίσης, χρησιμοποιήθηκαν τα πρώτα 50 σε έκφραση miRNA για κάθε δείγμα, ώστε να περιοριστεί το εύρος των περιοχών πρόβλεψης στόχων miRNA στην ανάλυση που έγινε με τους αλγορίθμους TargetScan και MIRZA-G. Και οι 2 εκτελέστηκαν χωρίς να λαμβάνονται υπόψιν εξελικτικά χαρακτηριστικά, καθότι αυτά δεν είναι δυνατόν να υπολογιστούν για τις περιοχές που υφίστανται τροποποίηση.

Τα αποτελέσματα υποδηλώνουν ότι οι τροποποιήσεις κυρίως μεταβάλλουν τα χαρακτηριστικά των υφιστάμενων περιοχών πρόσδεσης, ενώ σε πολύ μικρότερο βαθμό τις καθιστούν εντελώς μη λειτουργικές ή δημιουργούν νέες περιοχές. Επιπροσθέτως, παρατηρήθηκε ήπια μεταβολή της κατασταλτικής δράσης των miRNA που στοχεύουν τροποποιημένες UTR. Ξεχωριστή ανάλυση των UTR που εμφανίζουν υψηλό αριθμό τροποποιήσεων δεν υπέδειξε σημαντική συσχέτιση με το βαθμό αυξομείωσης της καταστολής. Το γεγονός ότι η κατασταλτική δράση των miRNA δε φάνηκε να επηρεάζεται καθολικά προς μία κατεύθυνση, υποδηλώνει πως ο ρυθμιστικός ρόλος των RNA τροποποιήσεων δεν ακολουθεί ένα γενικό κανόνα, αντιθέτως, δρα ως μηχανισμός βελτιστοποίησης, κατά περίπτωση ισχυροποιώντας ή αποδυναμώνοντας την πρόσδεση.

Σε αυτή τη μελέτη συγκρίναμε εργαλεία για τον εντοπισμό RNA τροποποιήσεων, καταλήξαμε με ένα σύνολο τροποποιημένων 3'UTR και πραγματοποιήσαμε ανάλυση για την πρόβλεψη στόχων των miRNA σε αυτές, η οποία υπέδειξε άλλοτε ενίσχυση και άλλοτε εξασθένηση της κατασταλτικής δράσης των miRNA στους στόχους τους, με ένταση ανεξάρτητη του αριθμού των συμβάντων τροποποίησης στην περιοχή πρόσδεσης. Περαιτέρω αναλύσεις με περισσότερα δείγματα και καταστάσεις θα φανούν χρήσιμες ώστε να επιβεβαιωθούν και να καταστούν στατιστικά σημαντικότερα τα ευρήματα της παρούσας εργασίας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική, Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: RNA τροποποιήσεις, αλληλούχιση επόμενης γενιάς, μεταγράφωμα, A-σε-I μετατροπή, πρόβλεψη στόχων miRNA

ABSTRACT

RNA editing is a co/post-transcriptional process, during which an RNA molecule is undergone an alteration of its sequence by insertion, deletion or modification. The majority of such changes in metazoans is comprised by adenosine (A) to inosine (I) nucleotide transitions, which are catalyzed by members of the adenosine deaminase gene family (ADAR) acting on double-stranded RNA (dsRNA). RNA editing is relatively widespread in Alu-containing mRNA molecules.

Editing of the coding sequence in pre-mRNAs can modify codons and lead to the incorporation of different amino acids during translation, contributing to protein function diversity. However, most A-to-I editing events occur in non-coding regions of pre-mRNAs and mRNAs, as well as in non-coding RNAs. Editing in the UTR (untranslated region) of mRNAs can regulate their translation, splicing and degradation. Also, events in microRNA (miRNA) and long non-coding RNA (lncRNA) sequences, as well as their binding sites, can affect their biogenesis, target recognition, structure and stability.

The goal of this study was to compare a set of RNA editing identification tools, distinguish true substitution events in 3'UTR of mRNAs and assess their impact on miRNA specificity and binding efficacy.

Initially, we used matching RNA and DNA sequencing data to identify A-to-I RNA editing events in 3'UTR regions. This was done to investigate event calls in individual level, increasing specificity. Our dataset consisted of 2 test samples, 1 ADAR enzyme knockdown control sample and RADAR, a comprehensive collection of A-to-I editing events in human, mouse and fly transcripts, with the last two resources being used as ground truth. Then we went on to find the best algorithm to identify events. The ADAR knockdown dataset was useful to pinpoint high false positive rates. The comparison included RES-Scanner employing the Burrows-Wheeler Aligner (BWA), REDItools running with GSNAP aligner and RNAEditor which was run with BWA (default option) and GSNAP. The first two approaches natively support paired/matched RNA-DNA datasets, while the latter was modified to include DNA information. The most robust behaviour in terms of sensitivity and specificity was observed from RNAEditor with BWA aligner.

Edited and non-edited 3'UTR were subsequently used as input in miRNA target prediction algorithms to statistically assess differences in the computed binding sites that arose due to the editing phenomena. The wildtype counterpart of the ADAR knockdown experiment was employed here to further enhance the comparison. Events annotated in 3'UTR were used to generate 2 equally numbered sets of sequences, 2062 of which belonged to highly repetitive Alu regions and 144 in non-Alu. The top 50 expressed miRNA in each sample were used to confine the target prediction analysis that was performed using TargetScan and MIRZA-G algorithms. Both of them were run without incorporating evolutionary features, which cannot be effectively measured in the case of the edited sequences.

The results show a strong preference towards modification of binding site feature distributions, rather than generating new or depleting existing sites. Moreover, we observed a mild alteration of the repressive action of miRNA targeting edited UTR. A separate analysis of highly edited UTR, i.e. UTR subjected to multiple editing events, did not indicate any correlation with the degree of the change. The lack of a global trend in the alteration of miRNA repressive activity implies RNA editing can serve distinct roles in miRNA efficacy, fine-tuning their targeting action on a case-by-case basis.

In this study, we did a benchmark of RNA editing identification tools, we came up with a set of edited UTRs and performed miRNA target prediction on them. This analysis indicated alteration of the targeting efficacy by miRNA, irrespective of the number of editing events in the region. Further analyses of more samples and conditions will be useful to validate and empower our findings.

SUBJECT AREA: Bioinformatics, Computational Biology

KEYWORDS: RNA editing, NGS, transcriptomics, A-to-I substitution, miRNA target prediction

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας, Καθ. Άρτεμις Χατζηγεωργίου, για τη διαρκή καθοδήγηση από την έναρξη των μεταπτυχιακών μου σπουδών μέχρι την ολοκλήρωση της παρούσας διπλωματικής εργασίας και για την ευκαιρία που μου προσέφερε για να δουλέψω στο περιβάλλον του αναγνωρισμένου DIANA Lab.

Επίσης, θα ήθελα να ευχαριστήσω τους Δρ. Martin Reczko και Δρ. Αλέξανδρο Δημόπουλο για τις γνώσεις που μου μετέδωσαν κατά τη διάρκεια των μεταπτυχιακών μου σπουδών ως καθηγητές και για τη συμμετοχή τους στην τριμελή εξεταστική επιτροπή.

Τέλος, θα ήθελα να ευχαριστήσω τον Σπύρο Ταστσόγλου, υποψήφιο διδάκτορα στο Πανεπιστήμιο Θεσσαλίας, για την αμέριστη βοήθεια και τις συμβουλές καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας, χωρίς την οποία δεν θα ήταν εφικτή η ολοκλήρωσή της και τα υπόλοιπα παιδιά του DIANA Lab, για το εξαίρετο περιβάλλον και την προθυμία για βοήθεια, όποτε αυτή ζητήθηκε.

ΠΕΡΙΕΧΟΜΕΝΑ

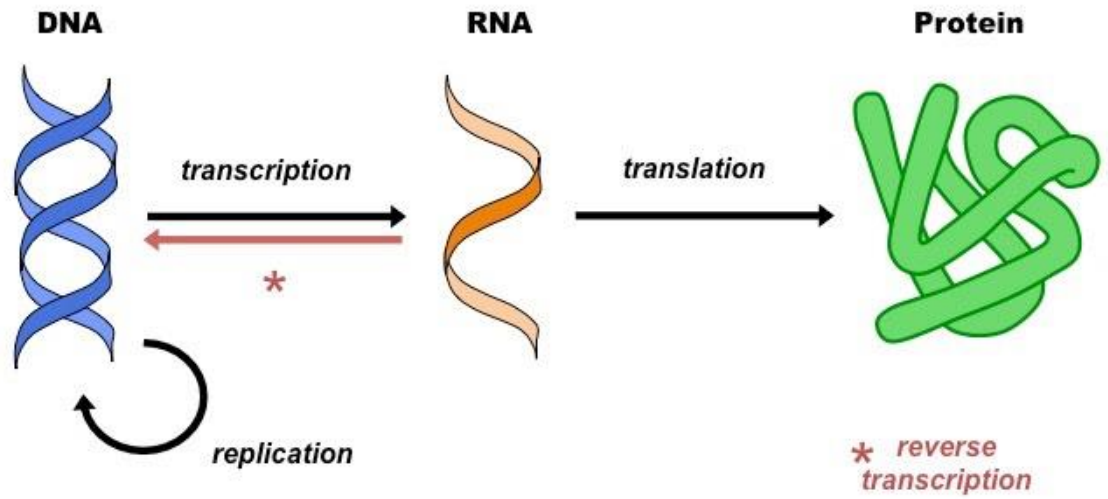
ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1. ΕΙΣΑΓΩΓΗ

Το υπερσύνολο λειτουργιών μέσω του οποίου κάθε βιολογικό σύστημα αξιοποιεί τη γενετική πληροφορία αποτυπώνεται στο Κεντρικό Δόγμα της Μοριακής Βιολογίας (εικόνα 1) [1], όπως αυτό διατυπώθηκε από τον Francis Crick το 1957. Η λέξη υπερσύνολο χρησιμοποιείται διότι δε χρησιμοποιούνται όλες οι λειτουργίες από όλα τα είδη οργανισμών π.χ. η διαδικασία κατά την οποία δημιουργείται ένα μόριο δεσοξυριβοζονουκλεϊκού οξέος (DNA) από ένα μόριο ριβοζονουκλεϊκού οξέος (RNA) και η οποία ονομάζεται αντίστροφη μεταγραφή, συμβαίνει μόνο σε ορισμένους ιούς. Επέκταση του συνόλου των λειτουργιών (απευθείας δημιουργία πρωτεΐνης από DNA) δεν έχει παρατηρηθεί στους μέχρι τώρα γνωστούς οργανισμούς, όμως μπορεί να επιτευχθεί εντός του εργαστηριακού περιβάλλοντος.



Εικόνα 1. Απεικόνιση του Κεντρικού Δόγματος της Βιολογίας

1.1 Η ροή της γενετικής πληροφορίας

Στη γενική περίπτωση, η γενετική πληροφορία μπορεί να μεταφερθεί μέσω της αντιγραφής ενός δίκλωνου μορίου DNA σε ένα καινούργιο δίκλωνο μόριο DNA (αντιγραφή). Επιπλέον, ένα κομμάτι ακολουθίας νουκλεοτιδίων DNA μπορεί να χρησιμοποιηθεί ως εκμαγείο ώστε να δημιουργηθεί, σε αντιστοιχία από αυτή, ένα μονόκλωνο μόριο RNA (μεταγραφή). Τέλος, ακολουθίες μονόκλωνων ώριμων μορίων RNA που ονομάζονται αγγελιοφόρα (messenger RNA - mRNA) μπορούν να «διαβαστούν» ανά τριάδες νουκλεοτιδίων (κωδικόνια/τριπλέτες) από τα ριβοσώματα και να μεταφραστούν σε αμινοξέα, με μια ακολουθία αμινοξέων να αποτελεί μία πρωτεΐνη.

Τα 4 βασικά νουκλεοτίδια που παρατηρούνται στο DNA είναι η Αδενίνη (A), η Θυμίνη (T), η Κυτοσίνη (C) και η Γουανίνη (G), ενώ στο RNA η Θυμίνη αντικαθίσταται από την Ουρακίλη (U). Κατά κανόνα, οι δεσμοί που αναπτύσσονται μεταξύ τους για να δώσουν την επιθυμητή δομή είναι μεταξύ A – T (ή U) και C – G. Ο βασικός αυτός δεσμός ονομάζεται συμπληρωματικός. Κατά ανάγκη δύναται να δημιουργηθούν και μη συμπληρωματικοί δεσμοί (π.χ. δεσμοί G – U (GU wobbles) στο RNA).

Μπορεί οι προαναφερθείσες διαδικασίες να αποτελούν το βασικό λίθο για τη δομή και τη λειτουργία των έμβιων οργανισμών, όμως οι ιδιαιτερότητες που παρουσιάζει ο κάθε ένας προκύπτουν μέσα από ένα τεράστιο δίκτυο που περιλαμβάνει πολλά περισσότερα μόρια, αλληλεπιδράσεις, ρυθμιστικά μονοπάτια και μηχανισμούς βελτιστοποίησης. Αυτοί οι

παράγοντες δεν οδηγούν απαραίτητα από ένα βήμα του κεντρικού δόγματος σε ένα άλλο, αλλά εμπλέκονται και σε ενδιάμεσα βήματα, αλλάζοντας τη ροή και το αποτέλεσμα της διαδικασίας, είτε σύμφωνα με τις ανάγκες του συστήματος, είτε λόγω λαθών στη λειτουργία του. Στοιχεία τα οποία μεταφράζονται σε πρωτεΐνες ονομάζονται κωδικά, ενώ στοιχεία τα οποία δεν μεταφράζονται ονομάζονται μη κωδικά.

1.2 (Μετα-)μεταγραφική ρύθμιση της σύνθεσης RNA και πρωτεΐνης

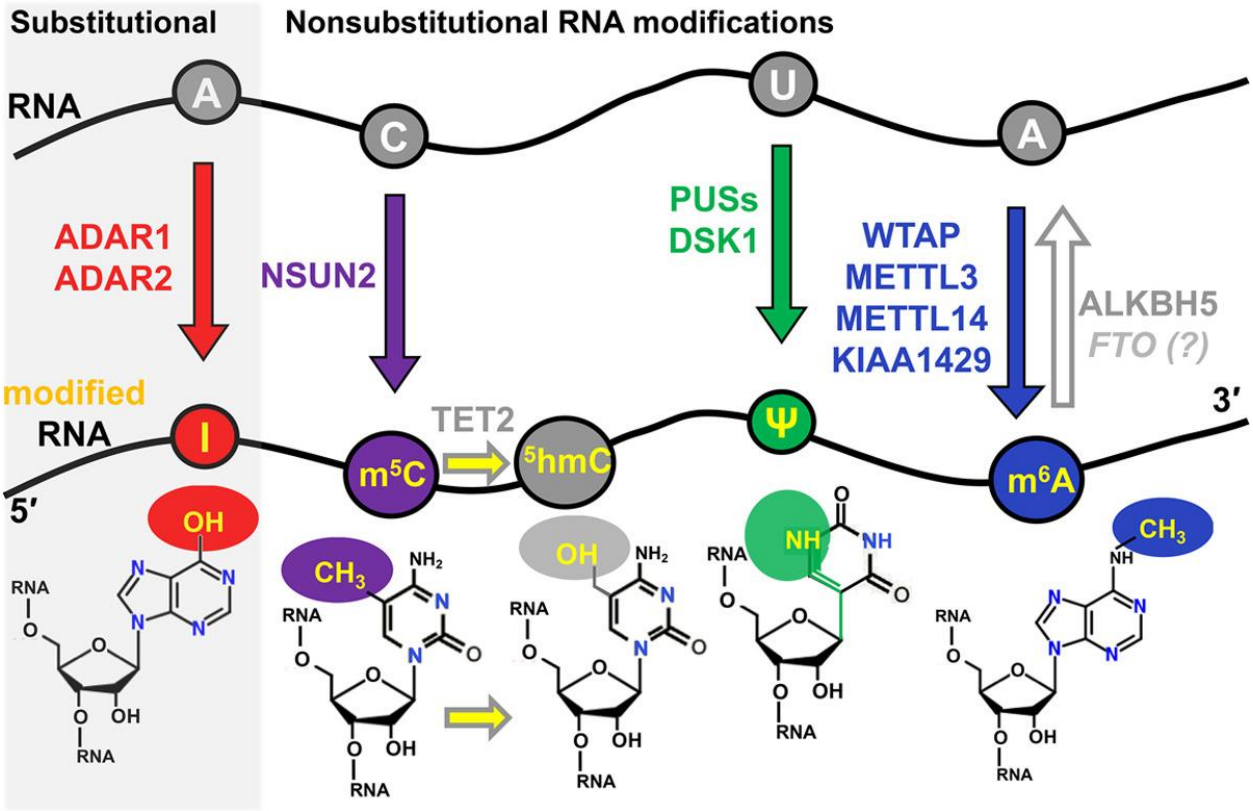
Η επίτευξη τεχνικών αλληλούχισης επόμενης γενιάς (Next Generation Sequencing - NGS) τα τελευταία χρόνια έχει επιτρέψει την αναλυτικότερη και πιο λεπτομερή παρατήρηση και εξερεύνηση των ενδιάμεσων βημάτων και των μη κωδικών παραγόντων. Σημαντικό ρόλο στη ρύθμιση της ροής της πληροφορίας διαδραματίζουν οι τροποποιήσεις νουκλεοτιδίων, που συμβαίνουν τόσο στο επίπεδο του DNA (π.χ. μέσω της μεθυλίωσης ή της μεταλλαγής ενός ή περισσοτέρων νουκλεοτιδίων), όσο και στο επίπεδο του RNA. Το DNA εμφανίζει υψηλότερη δομική σταθερότητα και ως επι το πλείστον βρίσκεται εντοπισμένο στο πυρήνα του κυττάρου, επομένως οι τροποποιήσεις του επηρεάζουν κυρίως την αλληλεπίδραση, πρόσβαση και αναγνώριση από διάφορα μόρια, με επιπτώσεις στην έκφραση κοντινών γονιδίων, αλλά και μόνιμες αλλαγές στα μόρια που κωδικοποιούνται [2]. Αντιθέτως, η φύση των λειτουργιών των RNA, που τα φέρει να αλλάζουν υποκυτταρικά διαμερίσματα, τους προσδίδει μια ελαστικότερη και πιο φορητή δομή, η οποία επηρεάζεται σε μεγαλύτερο βαθμό από αντίστοιχες τροποποιήσεις στα νουκλεοτίδια των ακολουθιών τους. Έχουν αναδειχθεί τροποποιήσεις στο RNA που επηρεάζουν την αναγνώριση μορίων με τα οποία αλληλεπιδρούν [3], όσο και ως προς τη στερεοδιαμόρφωση και τον υποκυτταρικό τους εντοπισμό, το χρόνο ημιζωής, την ανθεκτικότητα, την πρωτεΐνη που κωδικοποιούν κ.ά. [4][5][6][7][8]. Όπως και στην περίπτωση του DNA, η επεξεργασία μπορεί να μετέλθει μέσω της πρόσθεσης χημικών ομάδων στις ήδη υπάρχουσες βάσεις ή να τροποποιηθεί εντελώς η ακολουθία σε μία θέση, αντίστοιχα με τις DNA μεταλλάξεις. Έχουν εντοπιστεί πάνω από 170 διαφορετικά είδη τροποποιήσεων σε διάφορους οργανισμούς [9], ενώ η χαρτογράφηση του πλήρους φάσματος σχετικά με τα εμπλεκόμενα μόρια και τις λειτουργίες στις οποίες εμπλέκονται συνιστούν ένα σημαντικό θέμα που βρίσκεται ακόμη υπό μελέτη από την επιστημονική κοινότητα [10].

Κομβικοί ρυθμιστικοί παράγοντες της ροής της πληροφορίας θεωρούνται και τα μη κωδικά RNA (ncRNA). Αν και αρχικά θεωρούνταν παραπροϊόντα της μεταγραφικής λειτουργίας ή εξελικτικά απομεινάρια που δεν παίζουν κάποιο ρόλο, η ανακάλυψη εκατοντάδων χιλιάδων λειτουργικών ncRNA μέσω της χρήσης NGS έχει επιτρέψει την ανάλυσή τους σε βάθος, αναδεικνύοντας τη σημασία τους στον κυτταρικό κύκλο, την ανάπτυξη, τον κυτταρικό πολλαπλασιασμό, την απόπτωση αλλά και σε διάφορες παθογενείς καταστάσεις όπως σε πολλές μορφές καρκίνου κ.ά. [11][12]. Ειδικότερα, τα microRNA (miRNA) είναι ncRNA, τα οποία εκφράζονται σε αφθονία μέσα στα κύτταρα και είναι υπεύθυνα (υπό φυσιολογικές συνθήκες) για την διατήρηση των επιθυμητών επιπέδων έκφρασης των RNA μέσα σε αυτά. Στοχεύουν κωδικά αλλά και μη κωδικά RNA, ώστε να προκαλέσουν την αποικοδόμηση και/ή τη μεταφραστική καταστολή τους. Η εξερεύνηση και η κατανόηση των φυσιολογικών και των παθολογικών μηχανισμών και ρυθμιστικών μονοπατιών στα οποία εμπλέκονται επίσης αποτελούν σημαντική πρόκληση για τα επιστημονικά πεδία της Μοριακής Βιολογίας, της Βιοπληροφορικής, της Φαρμακευτικής και της Ιατρικής, καθότι οι κρίσιμες λειτουργίες στις οποίες μετέχουν τα καθιστούν στόχους-κλειδιά στη θεραπεία και την αποσαφήνιση διαφόρων ασθενειών και λειτουργιών [13].

1.3 RNA τροποποίηση

Ως τροποποίηση του RNA ορίζεται η διαδικασία κατά την οποία ένα ή περισσότερα νουκλεοτίδια της ακολουθίας του υπόκεινται σε κάποια αλλαγή. Η αλλαγή αυτή μπορεί να αφορά στη μεθυλίωση κάποιου νουκλεοτιδίου σε κάποια θέση, στην τροποποίηση της σύστασής του με αποτέλεσμα να προκύψει ένα νέο νουκλεοτίδιο που αναγνωρίζεται διαφορετικά από τους κυτταρικούς μηχανισμούς και στην αφαίρεση ή στην προσθήκη κάποιου/ων νουκλεοτιδίου/ων [14]. Στη διαδικασία αυτή συμμετέχουν σύμπλοκα μορίων που είτε εγγράφουν, είτε διαβάζουν είτε απαλείφουν μια τροποποίηση. Επειδή η τροποποίηση συμβαίνει από το μεταγραφικό στάδιο και ύστερα, οι αλλαγές που εφαρμόζονται διαφοροποιούνται ως προς τη λειτουργία που επιτελούν και τη συχνότητα που συμβαίνουν και δεν είναι μόνιμες/κληρονομήσιμες, προσφέροντας στο κύτταρο την αναγκαία βελτιστοποίηση και πολυπλοκότητα κατά περίπτωση [15].

Οι τροποποιήσεις μπορούν να διακριθούν σε 2 κατηγορίες σύμφωνα με το αποτέλεσμα τους: σε αυτές που προκαλούν μεταλλαγή του νουκλεοτιδίου σε κάποιο άλλο (RNA editing) και σε αυτές που τροποποιούν τη χημική σύσταση κάποιου υπάρχοντος νουκλεοτιδίου με προσθήκη χημικών ομάδων (εικόνα 2). Πρόκειται για εξαιρετικά συντηρημένες λειτουργίες που απαντώνται σε όλα τα είδη [16] και επιτελούν πολύ βασικές λειτουργίες, όπως είναι η πολυαδενυλίωση του 3’ άκρου του mRNA, ενώ εμπλέκονται και σε παθολογικές καταστάσεις [17].



Εικόνα 2. Κατηγορίες RNA τροποποιήσεων (πηγή: [18])

Ο εντοπισμός και η μελέτη τροποποιήσεων στα mRNA αποτελεί πρόκληση εξαιτίας μιας σειράς διαφόρων εμποδίων βιολογικής και τεχνικής φύσης. Αρχικά, η έλλειψη αρκετής ποσότητας στην αλληλούχιση των mRNA τα προηγούμενα χρόνια επέτρεπε την ανάδειξη στατιστικά σημαντικών φαινομένων σε μόρια που εμφανίζονται σε μεγαλύτερη περίσσεια μέσα στο κύτταρο, και άρα σε μεγαλύτερη συχνότητα στα αποτελέσματα των πειραμάτων, όπως τα μεταφορικά RNA (tRNA) και τα ριβοσωμικά RNA (rRNA) [19]. Στην εποχή του NGS, ένας τρόπος εντοπισμού είναι η χρήση αντισωμάτων τα οποία

προσδένονται στις θέσεις που παρουσιάζεται ένα συγκεκριμένο φαινόμενο και αποκαλύπτουν τη θέση του. Ωστόσο, απαιτείται μεγάλη αναγνωριστική εξειδίκευση του αντισώματος, ώστε να περιοριστούν τα ψευδώς θετικά σήματα (false positive). Άλλη τεχνική είναι η φασματομετρία μάζας, με την οποία ποσοτικοποιούνται διαφορές στη μάζα των νουκλεοτιδίων που οφείλονται στην τροποποίηση.

Με υπολογιστικό τρόπο μπορούν να εντοπιστούν τροποποιήσεις οι οποίες ανήκουν στην κατηγορία των μεταλλαγών. Η τεχνική που χρησιμοποιείται είναι η σύγκριση RNA ακολουθιών έναντι κάποιας άλλης πληροφορίας (π.χ. άλλες RNA ακολουθίες, αντίστοιχες DNA ακολουθίες, άλλα δεδομένα που δρουν ως αντικειμενική αλήθεια για το τι νουκλεοτίδιο αναμένεται να συναντήσουμε σε κάποια θέση). Εμπόδιο εδώ αποτελεί η υψηλή συχνότητα λανθασμένων σημάτων, η οποία προκύπτει λόγω σφαλμάτων κατά το διάβασμα της αλληλουχίας από τα μηχανήματα αλληλούχισης. Τα σφάλματα αυτά προκύπτουν από τη λανθασμένη στοίχιση των κομματιών αλληλουχίας (τα οποία ονομάζονται reads) στις αλληλουχίες αναφοράς. Το διάβασμα ακολουθιών μεγάλου μήκους χαρακτηρίζεται από μεγάλη συχνότητα σφαλμάτων, επομένως η απευθείας αλληλούχισή τους είναι αδύνατη. Για το λόγο αυτό, οι ακολουθίες κόβονται σε μικρότερα κομμάτια και αντιστοιχίζονται εκ των υστέρων στην αλληλουχία αναφοράς, με τη διαδικασία να ονομάζεται στοίχιση (alignment). Ακόμη, η συχνότητα του φαινομένου κατά περιπτώσεις είναι μικρή, με αποτέλεσμα να μην είναι στατιστικά σημαντική.

Η πιο συχνή μεταλλαγή που εμφανίζεται στα μετάζωα είναι από αδενοσίνη (το σύμπλοκο αδενίνης-δεοξυριβόζης/ριβόζης-φωσφορικών ομάδων) σε ινοσίνη (I). Η κατάλυσή της γίνεται από τα μέλη της οικογένειας των γονιδίων των απαμινάσεων της αδενοσίνης (adenosine deaminase acting on RNA - ADAR), τα οποία δρουν σε RNA με διπλή έλικα (dsRNA) και σε πρώιμα RNA (pre-RNA) [20]. Η απαμίνωση γίνεται μέσω υδρόλυσης, αφότου το νουκλεοτίδιο μεταφερθεί από την ακολουθία σε μια καταλυτική «τσέπη» του ενζύμου, και συμβαίνει στον 6^{ο} άνθρακα (C6) της αδενίνης, ο οποίος συμμετέχει στη δημιουργία του Watson-Crick δεσμού. Η Ινοσίνη επίσης είναι ικανή να συμμετέχει σε δεσμούς και αναγνωρίζεται ως Γουανοσίνη (G) από τη μεταφραστική μηχανή. Εκ των πρωτεϊνών ADAR, τα ένζυμα ADAR1 και ADAR2 συναντώνται στους περισσότερους ιστούς και είναι καταλυτικά ενεργά, ενώ το ADAR3 συναντάται αποκλειστικά στον εγκέφαλο και δεν παρουσιάζει καταλυτική δράση, για άγνωστους λόγους μέχρι σήμερα [21]. Οι ADAR δρουν στον πυρήνα αλλά και στο κυτταρόπλασμα. Η A-σε-I τροποποίηση είναι μια δυναμική λειτουργία, δεδομένου ότι το εύρος εφαρμογής της μπορεί να εκτείνεται από 0% έως 100% και να διαφέρει μεταξύ των ιστών και κυτταρικών τύπων. Άλλη σημαντική μετατροπή που έχει παρατηρηθεί στα θηλαστικά και τα φυτά είναι η κυτοσίνη σε ουρακίλη (C-σε-U), η οποία όμως συναντάται σε πολύ μικρότερη συχνότητα στα πρώτα [22].

Οι A-σε-I τροποποιήσεις συμβαίνουν με πολύ μεγαλύτερη συχνότητα στις Alu περιοχές από ότι στις non-Alu [23]. Τα Alu στοιχεία είναι μικρά σε μήκος DNA κομμάτια (short interspersed elements – SINEs) τα οποία περιέχουν πολλαπλές επαναλήψεις ενός ή περισσοτέρων νουκλεοτιδίων. Είναι ρετρομεταθετά στοιχεία, δηλαδή μπορούν να μεταπηδήσουν και να αντιγραφούν από μια θέση του DNA σε άλλη. Αποτελούν την πιο πολυπληθή ομάδα της κατηγορίας των μεταθετών στοιχείων, είναι καλά διατηρημένα μεταξύ των ειδών και υπάρχουν σε μεγάλη περίσσεια μέσα στο ανθρώπινο γονιδίωμα, καταλαμβάνοντας το 11% της συνολικής του ακολουθίας [24].

Σε κλινικό επίπεδο, οι φυσιολογικές και παθολογικές λειτουργίες στις οποίες συναντάται η A-σε-I τροποποίηση δεν είναι πλήρως γνωστές. Η απενεργοποίηση της οικογένειας των ενζύμων ADAR στη *Drosophila* έχει σημαντικές επιπτώσεις στη φυσιολογική λειτουργία του εγκεφάλου, στην ακεραιότητα της κίνησης και στην εκφύλιση του νευρολογικού συστήματος με την πάροδο του χρόνου [25]. Στα ποντίκια, η εισαγωγή ομόζυγης μετάλλαξης στο γονίδιο του ADAR2, η οποία αποτρέπει τη μεταγραφή του, προκαλεί το

θάνατο των δειγμάτων μερικές εβδομάδες μετά τη γέννηση τους, κατά τις οποίες παρατηρούνται επιληπτικά επεισόδια και εκφυλισμός των νευρώνων τους λόγω αυξημένης εισροής Ca^{2+} και μειωμένης συχνότητας RNA τροποποιήσεων στην περιοχή του μεταγράφου GluR-B Q/R [26]. Επιπλέον, η απενεργοποίηση του ADAR1 προκαλεί θάνατο των εμβρυικών στο εμβρυικό στάδιο, λόγω επιπλοκών στο αιμοποιητικό σύστημα και στην ανάπτυξη του ήπατος [27][28]. Στον *Homo sapiens,* οι RNA τροποποιήσεις αναμένονται να δρουν αντιστοίχως, παίζοντας σημαντικό ρόλο σε βασικές λειτουργίες του οργανισμού και στο νευρικό σύστημα. Πράγματι, δυσλειτουργίες στην έκφραση των ADAR ενζύμων και στα επίπεδα των τροποποιήσεων έχουν συνδεθεί με κυτταρικές λειτουργίες, ψυχικές ασθένειες, κατάθλιψη, την DSH (*dyschromatosis symmetrica hereditaria*) και διάφορες μορφές καρκίνου [29][30][31][32].

Ο βιολογικός ρόλος των RNA τροποποιήσεων επίσης δεν έχει αποσαφηνιστεί πλήρως. Οι δράσεις στις οποίες μπορεί να συμμετέχει σε σχέση με τα RNA συνοψίζονται στην εικόνα 3:

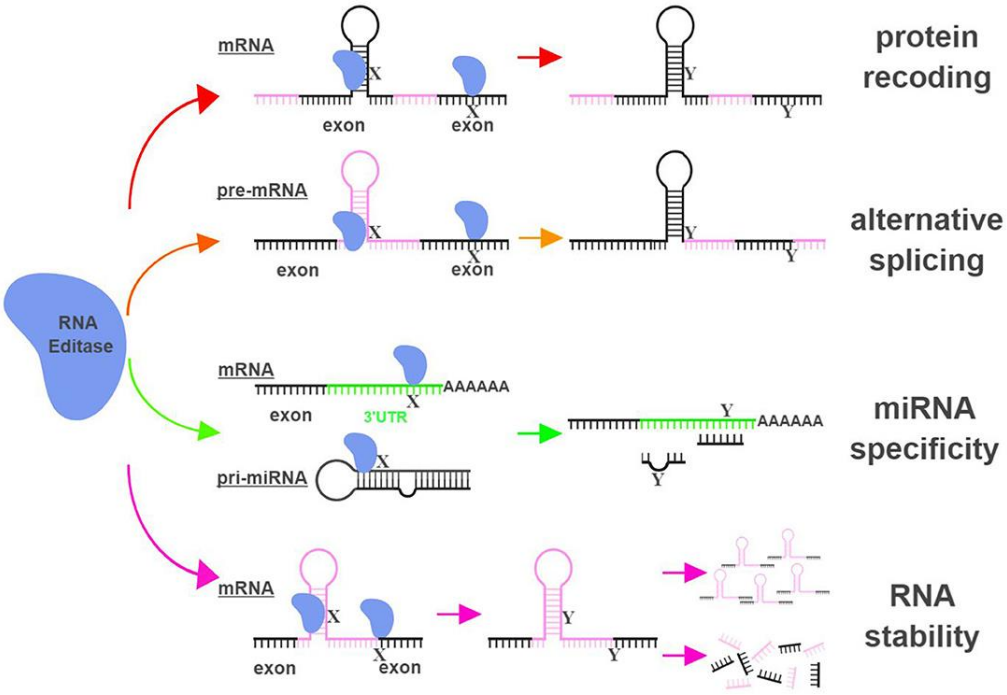
i) Στην περίπτωση που η τροποποίηση αλλάξει την κωδικοποίηση ενός αμινοξέος, τότε και η πρωτεΐνη που θα παραχθεί θα είναι διαφορετική, δίνοντας έτσι την ιδιότητα στις RNA τροποποιήσεις να συμβάλουν στην αύξηση της πρωτεϊνικής ποικιλομορφίας.

ii) Μια τροποποίηση μπορεί να αλλάξει το σημείο αναγνώρισης που οδηγεί στο μάτισμα του pre-RNA και να παραχθεί ένα νέο ώριμο RNA, προσφέροντας αντίστοιχα νέες επιλογές στην ποικιλομορφία του μεταγραφώματος.

iii) Η τροποποίηση δύναται να αλλάξει τα χαρακτηριστικά μιας υπάρχουσας περιοχής πρόσδεσης ενός miRNA, να προκαλέσει την αναγνώρισή της από ένα διαφορετικό miRNA, να καταργήσει ή να δημιουργήσει μια περιοχή.

iv) Η τροποποιημένη RNA ακολουθία μπορεί να έχει διαφορετικά βιοχημικά χαρακτηριστικά, καθιστώντας διαφορετική τη σταθερότητα του μορίου.

Οι RNA τροποποιήσεις στοχεύουν επίσης και ncRNA. Πιο συγκεκριμένα, έχουν παρατηρηθεί μεταλλαγές σε miRNA και σε long non-coding RNA (lncRNA) [33][34].



Εικόνα 3. Σύνοψη των λειτουργιών των RNA τροποποιήσεων σε βιολογικό επίπεδο. (πηγή: [35])

Αλλαγές στην ακολουθία τους, μπορεί να επηρεάσει τη βιογένεση, τη δομή και τη σταθερότητά τους. Επιπλέον, με βάση τους κανόνες που διέπουν την αναγνώριση των στόχων των miRNA και lncRNA, οι οποίοι καθορίζονται από την ακολουθία των εμπλεκόμενων μορίων, είναι δυνατό να επηρεάζεται το σύνολο των στόχων με τους οποίους αλληλεπιδρούν.

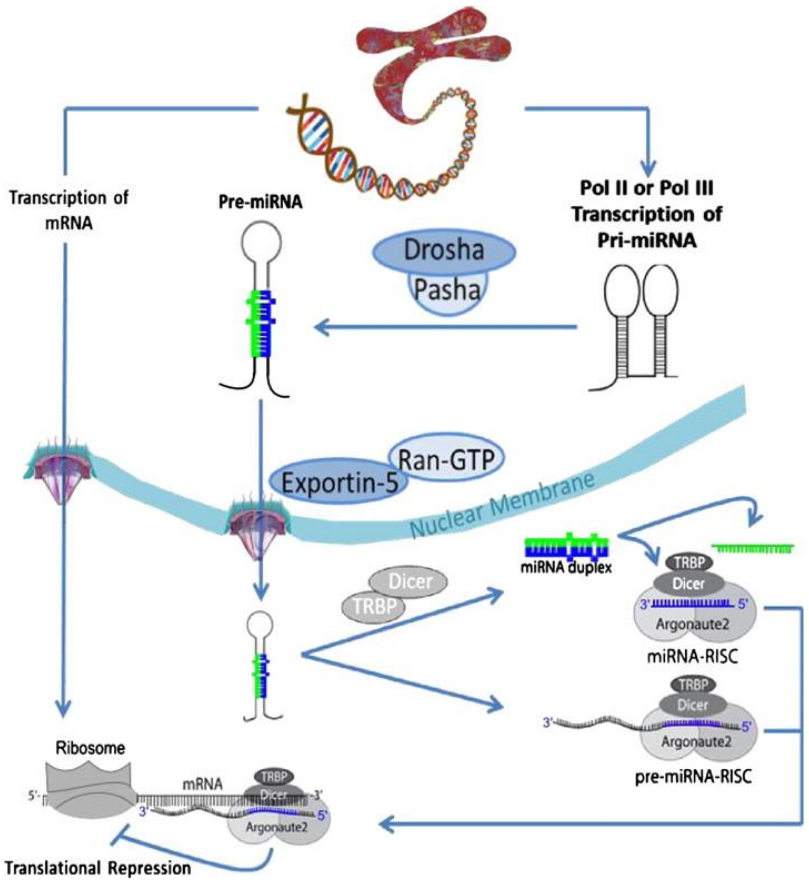
1.4 Βιογένεση και ρύθμιση των miRNA

Τα miRNA είναι μονόκλωνα ncRNA και είναι ενδογενή (παράγονται από το ίδιο βιολογικό σύστημα στο οποίο δρουν). Έχουν μήκος ~22 νουκλεοτίδια και ο κύριος ρόλος τους είναι η ενεργή ρύθμιση της γονιδιακής έκφρασης στους περισσότερους ευκαρυωτικούς οργανισμούς, μέσω πρόσδεσης στη μη μεταφραζόμενη περιοχή (untranslated region – UTR) που βρίσκεται στο 3’ άκρο των mRNA, την 3’UTR. Η ανακάλυψη των miRNA έγινε το 1993 στον *Caenorhabditis elegans*, όπου παρατήρησαν ότι το *lin-4* γονίδιο δεν κωδικοποιεί κάποια πρωτεΐνη, αλλά παράγει ένα ζευγάρι μικρών RNA [36]. Στη συνέχεια, είδαν ότι ο μηχανισμός των miRNA έχει εκτεταμένη δράση μέσα στα κύτταρα [37]. Έκτοτε, έχουν καταχωρηθεί 48.860 ώριμες ακολουθίες miRNA από 271 διαφορετικούς οργανισμούς. Οι πληροφορίες αυτές έχουν εξαχθεί από τη miRBase [38], ένα state-of-the-art αποθετήριο στο οποίο μπορεί κανείς να αντλήσει στοιχεία που αφορούν το όνομα, την ακολουθία αλλά και κυτταρικές λειτουργίες και γονίδια που έχουν συσχετιστεί με τα miRNA.

Η βιογένεση των miRNA συμβαίνει τόσο σε ιντρόνια (introns, περιοχές των μεταγράφων οι οποίες δε μεταφράζονται) όσο και σε εξώνια (περιοχές που μεταφράζονται) [39]. Στην περίπτωση που το miRNA προέρχεται από εξώνιο, είναι συχνό φαινόμενο ένα μέρος της ακολουθίας του να προέρχεται ταυτόχρονα και από ένα κομμάτι ιντρονίου [40]. Η μεταγραφή και η ρύθμισή τους γίνεται κατά κύριο λόγο παράλληλα με το γονίδιο που τα περιέχει (host gene), ενώ συχνά μοιράζονται και τον ίδιο εκκινητή (promoter, DNA περιοχή η οποία συμβάλλει στην εκκίνηση της μεταγραφής) [41]. Το host gene μπορεί είτε να κωδικοποιεί κάποια πρωτεΐνη είτε κάποιο lncRNA. Υπάρχουν ακόμα miRNA τα οποία δεν περιέχονται σε άλλα γονίδια, αλλά εδράζονται σε περιοχές μεταξύ των γονιδίων (intergenic regions) και μεταγράφονται ανεξάρτητα. Αξίζει να σημειωθεί ότι ένα γονίδιο μπορεί να κωδικοποιεί ένα miRNA (monocistronic transcript) ή μια ομάδα miRNA (polycistronic transcript).

Η μεταγραφή των miRNA γίνεται συνήθως με τη βοήθεια της RNA πολυμεράσης II, η οποία συνθέτει πρώιμα μετάγραφα (pri-miRNA). Στη συνέχεια, τα pri-miRNA υπόκεινται σε μεταγραφικές επεξεργασίες (πολυαδενυλίωση, μάτισμα κτλ.). Στην ακολουθία των pri-miRNA σχηματίζονται μία ή περισσότερες δίκλωνες δομές φουρκέτας μήκους ~70 νουκλεοτιδίων που ονομάζονται πρόδρομα miRNA (pre-miRNA). Στα μετάζωα, τα pre-miRNA με δίκλωνη δομή αναγνωρίζονται από το σύμπλοκο του μικροεπεξεργαστή (Microprocessor), το οποίο αποτελείται από την πυρηνική πρωτεΐνη DiGeorge critical region 8 dimer (DGCR8 ή Pasha στα μη σπονδυλωτά) και το ένζυμο Drosha που προσδένεται σε αυτά και τα αποκόπτει από τα pri-miRNA. Στη συνέχεια, τα άκρα των pre-miRNA αντικαθίστανται από μια υδροξυλική ομάδα (OH) στο 3’ άκρο και μια φωσφορική (P) στο 5’ άκρο και μεταφέρονται από τον πυρήνα στο κυτταρόπλασμα με τη βοήθεια της πρωτεΐνης exportin 5 (Exp5). Εκεί κόβεται η φουρκέτα του pre-miRNA από το RNase III ένζυμο Dicer, αφήνοντας ένα ζευγάρι συνδεδεμένων miRNA μήκους ~22 νουκλεοτιδίων, που αποτελούν τα ώριμα miRNA. Σε ένα ή και στα 2 από αυτά θα καταλήξει προσδεμένο ένα σύμπλοκο πρωτεϊνών που ονομάζεται RNA induced silencing complex (RISC). Στο σύμπλοκο αυτό συμμετέχει και ένα μέλος της οικογένειας πρωτεϊνών Ago (Argonaute), το οποίο θα το κατευθύνει στο mRNA-στόχο [42]. Η Ago συναντάται τόσο στον πυρήνα

όσο και στο κυτταρόπλασμα και η καταλυτική της δράση παίζει κύριο ρόλο στο σύμπλοκο RISC [43]. Η AGO2 είναι από τα πιο σημαντικά μέλη της Ago οικογένειας στον ανθρώπινο οργανισμό [44]. Υπάρχουν περιπτώσεις όπου το στάδιο κατά το οποίο προσδένεται το ένζυμο Drosha ή το Dicer προσπερνιέται και προσαρμόζεται σε κάποιο από τα υπόλοιπα βήματα. Σε αυτές τις περιπτώσεις, η βιογένεση του miRNA χαρακτηρίζεται ως μη κανονική (non-canonical) [45][46].



Εικόνα 4. Βιογένεση και βιολογική λειτουργία των miRNA. (πηγή: [47])

Η σχέση miRNA:mRNA δεν είναι 1-1· ένα miRNA μπορεί να στοχεύει ένα σύνολο από mRNA, όπως επίσης ένα mRNA μπορεί να στοχεύεται από πολλαπλά miRNA. Στα φυτά, η συμπληρωματικότητα των βάσεων μεταξύ του miRNA και του mRNA είναι σχεδόν τέλεια, κάτι το οποίο οδηγεί στην υπόθεση ότι δρουν ως εμπόδια στη μετάφραση των μεταγράφων (ρόλος που επιτελείται από τα small interfering RNA – siRNA) και τα κατευθύνουν για καταστροφή [48]. Αντιθέτως, στα μετάζωα δεν τηρείται πλήρης συμπληρωματικότητα, με τον πιο συχνό δεσμό να βρίσκεται μεταξύ ενός τμήματος του miRNA στο 5’ άκρο του, μεταξύ των νουκλεοτιδίων 2-7, που έχει παρατηρηθεί ότι είναι η πιο σημαντική περιοχή στην αναγνώριση των στόχων του [49][50]. Η περιοχή αυτή ονομάζεται seed region. Μικρές αλλαγές στην ακολουθία της seed περιοχής μπορούν να επηρεάσουν την ποιότητα και το φάσμα των στόχων του miRNA.

Εκτός από την πρόσδεση στο 3’UTR άκρο των mRNA, τα miRNA μπορούν να στοχεύσουν και το 5’UTR άκρο, καθώς και περιοχές της κωδικής περιοχής τους (coding sequence – CDS) [51][52].

Οι φυσιολογικοί και παθολογικοί μηχανισμοί στους οποίους εμπλέκονται τα miRNA είναι πολλοί και βασικοί και για αυτό το λόγο μελετώνται εντατικά ως πιθανοί θεραπευτικοί

στόχοι. Κάποιες από αυτές τις λειτουργίες που επηρεάζουν είναι η ρύθμιση του πολλαπλασιασμού των κυττάρων, της απόπτωσής τους, της διαφοροποίησής τους (δηλαδή της διαδικασίας κατά της οποίας θα αποκτήσουν συγκεκριμένο ρόλο), της έκφρασης των host genes και του ρυθμού ανάπτυξης, αλλά και διαφόρων μορφών καρκίνου [11][12][13]. Επιπλέον, εξελικτικά είναι σε μεγάλο βαθμό συντηρημένα στα μετάζωα [53].

Επεξεργασίες στα διάφορα στάδια της βιογένεσης των miRNA μπορούν να προκαλέσουν αλλαγές στη λειτουργία τους. Τέτοιες τροποποιήσεις μπορεί να είναι μεταλλάξεις στο DNA (single nucleotide polymorphism – SNP) [54], αλλαγές στη μεθυλίωσή του [55] ή RNA τροποποιήσεις [3]. Πιο συγκεκριμένα, τα SNP μπορούν να προκαλέσουν τη μόνιμη δημιουργία ή κατάργηση λειτουργιών των παραγόμενων miRNA, τροποποιώντας την ακολουθία τους, το ρυθμό παραγωγής τους καθώς και την ακολουθία των ρυθμιστικών παραγόντων της μεταγραφής τους. Αντίστοιχα, τα ρυθμιστικά μονοπάτια στα οποία συμμετέχουν τα τροποποιημένα miRNA ενδέχεται να οδηγηθούν σε απορρύθμιση που μπορεί να λάβει παθογενείς προεκτάσεις. Η DNA μεθυλίωση επηρεάζει την έκφραση των γονιδίων των miRNA, ρυθμίζοντας έτσι τα επίπεδά τους μέσα στο κύτταρο. Οι RNA τροποποιήσεις μπορούν να προκαλέσουν αλλαγές στο μετα-μεταγραφικό επίπεδο, εισάγοντας μεγαλύτερη ετερογένεια στις λειτουργίες των miRNA. Έτσι, η ρύθμιση που προέρχεται από αυτά γίνεται πιο ευέλικτη, με τον κίνδυνο όμως η επέκταση αυτή να οδηγήσει σε ανεπιθύμητες καταστάσεις, όπως η κωδικοποίηση ανενεργών ή λιγότερο αποτελεσματικών πρωτεϊνών, η αναποτελεσματική καταστολή ογκογονιδίων ή η προώθηση της ογκογένεσης με άλλα μέσα (π.χ. επάγοντας τον ανεξέλεγκτο πολλαπλασιασμό).

1.5 Εργαλεία

Η εξάπλωση του NGS έχει επιτρέψει την ανάπτυξη υπολογιστικών μεθόδων για την πρόβλεψη RNA τροποποιήσεων αλλά και στόχων miRNA, οι οποίες επιστρατεύουν σε αρκετές περιπτώσεις διαφορετικά μοντέλα για την επίτευξη της επιθυμητής ακρίβειας και ευαισθησίας.

Στο πεδίο των RNA τροποποιήσεων, η μεγάλη ετερογένεια στην επιλογή των RNA στόχων προς τροποποίηση και της συχνότητας των τροποποιήσεων μεταξύ των διαφορετικών ιστών, η συνύπαρξη των ίδιων μορίων σε τροποποιημένη και μη-τροποποιημένη μορφή στον ίδιο χώρο και σε διαφορετικές ποσότητες, οι μεταλλάξεις στο DNA αλλά και τα ποιοτικά λάθη στην παρατήρηση των δειγμάτων, λόγω τεχνικών αδυναμιών, καθιστούν την πρόβλεψή τους πρόκληση για τους ερευνητές [56]. Τα εργαλεία που προτείνονται από την ερευνητική κοινότητα είναι αρκετά πρόσφατα στην πλειονότητά τους και εφαρμόζουν διαφορετικές στρατηγικές μεταξύ τους. Άλλα αξιοποιούν την πληροφορία που εξάγουν από ζευγάρια RNA και DNA δειγμάτων, ώστε να μειώσουν τα εσφαλμένα false positive σήματα στα αποτελέσματα που αφορούν στην πραγματικότητα SNP (π.χ. RES-Scanner [57]), άλλα υπολογίζουν πιο στοχευμένα τα συμβάντα μόνο στα RNA δείγματα, σε συνδυασμό με πληροφορία για γνωστά SNP (π.χ. RNAEditor [58][59], GIREMI [60], SPRINT [61], RED-ML [62]), ενώ άλλα δίνουν την επιλογή και για τα 2 (π.χ. REDItools [63], JACUSA [64]). Ο συνδυασμός RNA-DNA, αν και σπανιότερος στην εύρεση δειγμάτων, αποτελεί την πιο ορθή επιλογή καθότι οι μεταλλάξεις στο γονιδίωμα διαφέρουν από άτομο σε άτομο και απαιτείται εξατομίκευση. Ακόμα, σημαντικό ρόλο παίζει η επιλογή της μεθόδου με την οποία θα γίνει το alignment, με το κάθε εργαλείο να προτείνει τη δική του επιλογή και σε κάποιες φορές να την κρίνει αναγκαία (π.χ. το RNAEditor δέχεται μόνο αρχεία που δεν είναι pre-aligned). Η ακρίβεια σε αυτό το βήμα είναι καθοριστική για την εγκυρότητα των αποτελεσμάτων. Όπως γίνεται εύκολα αντιληπτό, η ποιότητα των δειγμάτων επηρεάζει επίσης άμεσα τον υπολογισμό,

λόγω των false positive, και άρα η σωστή επιλογή και η προεπεξεργασία τους αποτελεί κύριο κομμάτι της διαδικασίας. Τέλος, τα περισσότερα διαθέσιμα RNA sequencing δείγματα (αλληλούχιση του μεταγραφώματος των δειγμάτων, RNA-seq) δεν περιέχουν πληροφορία για τον κλώνο από όπου προήλθε το μετάγραφο που αλληλουχήθηκε (έχουν προκύψει από non strand-specific πειραματικά πρωτόκολλα). Η πληροφορία αυτή όμως είναι χρήσιμη για να αποφανθεί το εργαλείο αν το φαινόμενο που παρατηρεί είναι A-σε-I ή G-σε-A. Η πρώτη περίπτωση μπορεί να αποτελεί μια πραγματική RNA τροποποίηση ενώ η δεύτερη μπορεί να είναι τεχνικό σφάλμα κατά την αλληλούχιση, λάθος στοίχιση των διαβασμάτων στο γονιδίωμα ή περίπτωση μεταλλαγής του DNA. Για το λόγο αυτό, επιθυμητά είναι τα δείγματα που είναι strand-specific, τα οποία όμως είναι πιο σπάνια λόγω του αυξημένου κόστους. Σε αντίθετη περίπτωση, τα paired-end δείγματα αποτελούν την ορθότερη εναλλακτική. Στα non strand-specific, paired-end δείγματα, κάθε read διαβάζεται και από τα δύο άκρα, πετυχαίνοντας υψηλότερη ακρίβεια στο διάβασμά του και τη στοίχιση, άρα και μείωση των false positive.

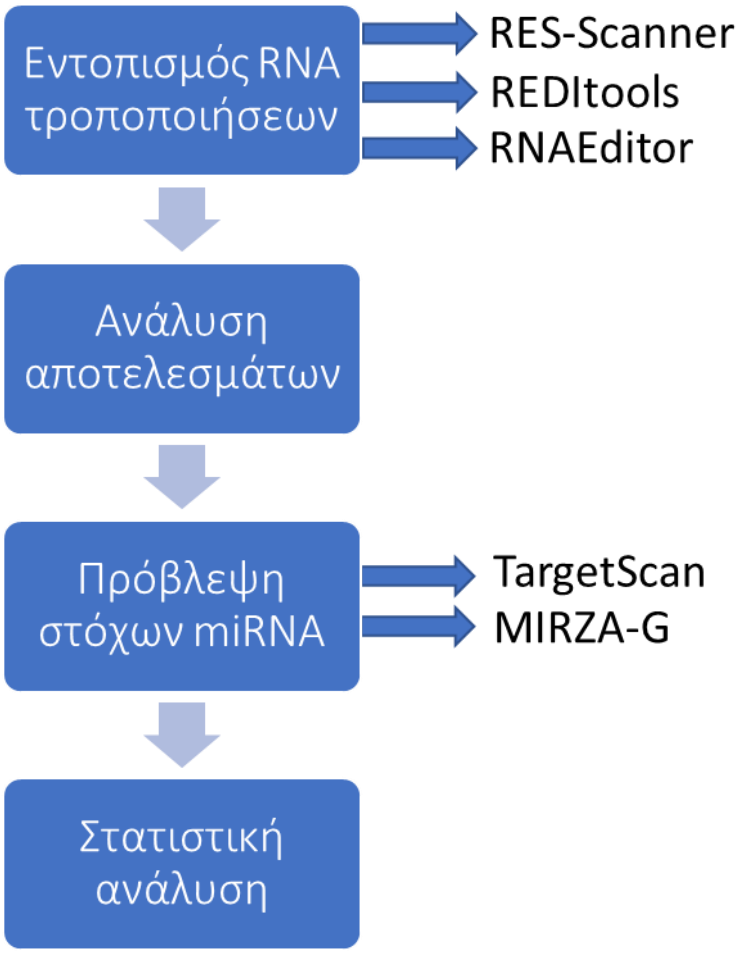
Στον τομέα της πρόβλεψης στόχων των miRNA, οι αλγόριθμοι που προτείνονται εμφανίζουν πιο σταθερή συμπεριφορά και πιο ακριβή αποτελέσματα. Οι περισσότεροι αλγόριθμοι χρησιμοποιούν ένα υπολογιστικό μοντέλο, το οποίο επιστρατεύει χαρακτηριστικά που αφορούν τις θερμοδυναμικές ιδιότητες, την προσβασιμότητα της δομής, τη συμπληρωματικότητα μεταξύ των αλληλεπιδρώντων μορίων RNA και το περιεχόμενο της ακολουθίας του miRNA, ώστε να καταλήξουν σε ένα τελικό score που αντιπροσωπεύει την προτίμηση μιας περιοχής ως σημείο πρόσδεσης. Αρκετοί αλγόριθμοι αξιοποιούν και το γεγονός ότι τα σημεία πρόσδεσης των miRNA είναι αρκετά συντηρημένα μεταξύ των ειδών, δίνοντας ακόμα μεγαλύτερη αξιοπιστία σε προβλέψεις που εμφανίζουν υψηλή συντήρηση. Η πρόσδεση στην 3’UTR είναι η πιο συχνή, με κάποιους αλγορίθμους να δέχονται μόνο τέτοιες περιοχές ως είσοδο και να περιορίζουν την αναζήτησή τους σε αυτές (π.χ. TargetScan [65], MIRZA-G [66], Miranda [67]). Ταυτόχρονα σε 3’UTR και CDS ψάχνει το DIANA-microT-CDS [68], ενώ υπάρχουν και άλλοι αλγόριθμοι που επεκτείνουν το εύρος της αναζήτησης και σε 5’UTR (π.χ. PITA [69], miRDB [70]).

1.6 Σύνοψη διπλωματικής εργασίας

Στα κεφάλαια που ακολουθούν αναπτύσσεται η διαδικασία που ακολουθήθηκε ώστε να γίνει σύγκριση τριών εργαλείων που εντοπίζουν RNA τροποποιήσεις (RES-Scanner, REDItools, RNAEditor) σε RNA-seq δείγματα, με σκοπό την επιλογή του πιο ακριβούς και ευαίσθητου μοντέλου. Στη συνέχεια, περιγράφεται η επιλογή των φαινομένων τροποποίησης που εντοπίστηκαν σε 3’UTR περιοχές και η χρήση των τροποποιημένων/μη-τροποποιημένων ακολουθιών για την πραγματοποίηση πρόβλεψης στόχων των miRNA με δύο αλγορίθμους (TargetScan, MIRZA-G). Τελικός στόχος της εργασίας είναι η στατιστική σύγκριση μεταξύ των χαρακτηριστικών που υπολογίστηκαν από τους 2 αλγορίθμους και η εκτίμηση των διαφορών μεταξύ των τροποποιημένων και των μη-τροποποιημένων συνόλων.

2. ΜΕΘΟΔΟΙ

Τα βήματα που ακολουθήθηκαν για τη διεξαγωγή της εργασίας απεικονίζονται συγκεντρωτικά στο σχήμα 1.



Σχήμα 1. Σύνοψη βημάτων της διπλωματικής εργασίας

Για το πρώτο βήμα της διαδικασίας, που αφορά τον εντοπισμό RNA τροποποιήσεων, επιλέχθηκε η εκτέλεση και η σύγκριση 3 εργαλείων: του RES-Scanner, του REDItools και του RNAEditor. Στο δεύτερο βήμα, η ανάλυση αποτελεσμάτων έγινε με τη χρήση των γλωσσών προγραμματισμού R και Python. Για την πρόβλεψη στόχων των miRNA στο τρίτο βήμα εκτελέστηκαν τα TargetScan και MIRZA-G, ενώ η στατιστική ανάλυση του τέταρτου βήματος έγινε επίσης με τη χρήση της R.

2.1 Δεδομένα

Για τη διεξαγωγή της ανάλυσης χρησιμοποιήθηκαν 4 ζευγάρια RNA-DNA δειγμάτων. Πληροφορίες για τα δείγματα αναφέρονται στους πίνακες 1 και 2. Τα ERR188182 και ERR188298 είναι δείγματα από τη μελέτη E-GEUV-1 του GEUVADIS project (θυγατρικό πρόγραμμα του 1000 Genomes) που βρίσκονται διαθέσιμα για δημόσια χρήση στο αποθετήριο ArrayExpress [71][72]. Χρησιμοποιήθηκαν ως test σύνολο, για να δοκιμαστεί η επίδοση των εργαλείων σε πραγματικές συνθήκες. Πρόκειται για paired-end, non-strand

specific mRNA-seq δείγματα, προερχόμενα από λεμφοβλαστοειδή κυτταρική σειρά. Τα RNA δείγματα λήφθηκαν σε fastq μορφή (μορφή στην οποία τα reads δεν έχουν αντιστοιχηθεί σε κάποιο γονιδίωμα), ενώ τα DNA NA18912 και NA18867 λήφθηκαν σε Binary Alignment Map (BAM) μορφή, τα οποία είναι συμπιεσμένα αρχεία σε δυαδική μορφή με προ-αντιστοιχισμένα reads. Στη δεύτερη στήλη του πίνακα 1 αναφέρεται το ID του DNA δείγματος που τους αντιστοιχεί.

Τα δείγματα ENCLB155EFP και ENCLB544CFT (από τα εργαστήρια Brenton Graveley, UConn και Alexander Urban, Stanford αντίστοιχα) [73] λήφθηκαν από το αποθετήριο ENCODE (https://www.encodeproject.org/) [75][76]. Το ENCLB155EFP είναι strand-specific mRNA-seq δείγμα, που προέρχεται από την κυτταρική σειρά χρόνιας μυελογενούς λευχαιμίας K562. Η μορφή στην οποία λήφθηκε είναι fastq. Χρησιμοποιήθηκε ως δείγμα ελέγχου (control), καθώς έχει κατασκευαστεί με ένα short hairpin RNA (shRNA), το οποίο στοχεύει τα γονίδια ADAR. Τα shRNA είναι τεχνητά RNA μόρια που εισάγονται σε δείγματα, προκειμένου να παρεμποδιστεί η έκφραση συγκεκριμένων πρωτεϊνών μέσω της αποσιώπησης των mRNA που τις παράγουν. Το ENCLB544CFT είναι το DNA δείγμα που του αντιστοιχεί, επίσης σε fastq μορφή.

Πίνακας 1. Πληροφορίες για τα RNA-seq δείγματα ERR188182 και ERR188298 από το E-GEUV-1 mRNA σύνολο δεδομένων στο ArrayExpress αποθετήριο του GEUVADIS project για το 1000 Genomes και των δειγμάτων ENCLB155EFP και ENCLB544CFT από το αποθετήριο ENCODE.

Sample ID	DNA sample ID	Sex/Age	Organ	Cell type	Cell line	Layout	Instrument	# of Bases	# of sequences	Read length
ERR188182	NA18912	Female/adult	Blood	B-Cell	**Lymphoblastoid cell line**	Paired/Non-strand-specific	Illumina HiSeq 2000	6.96 Gbp	~45.7M	75
ERR188298	NA18867	Female/adult	Blood	B-Cell	**Lymphoblastoid cell line**	Paired/Non-strand-specific	Illumina HiSeq 2000	4.2 Gbp	~28.3M	75

Sample ID	Assay	Biosample	Organ	Cell type	Cell line	Layout	Instrument	# of sequences	Read length
ENCLB155EFP	shRNA RNA-seq	*Homo sapiens* K562	Blood	Erythroleukemia	**Chronic myelogenous leukemia**	Paired/Strand-specific	Illumina HiSeq 2000	~17.3M	100
ENCLB544CFT	Genotyping	*Homo sapiens* K562	Blood	Erythroleukemia	**Chronic myelogenous leukemia**	Paired/Non-strand-specific	Illumina HiSeq X Ten	~890M	151

Σε όλα τα δείγματα έγινε ποιοτικός έλεγχος χρησιμοποιώντας το FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Επιπλέον, στα δείγματα ENCLB544CFT και ENCLB155EFP έγινε έλεγχος και για adapter αλληλουχίες με χρήση του Minion από τη σουίτα εργαλείων Kraken [77]. Οι αλληλουχίες adapter είναι συνθετικές, μικρές αλληλουχίες που χρησιμοποιούνται κατά τη διαδικασία της αλληλούχισης. Δεν έχουν βιολογική σημασία και η παρουσία τους ενδέχεται να επηρεάσει τη στοίχιση και τις επακόλουθες αναλύσεις, ως εκ τούτου πρέπει να εντοπίζονται και να αφαιρούνται από το εκάστοτε δείγμα.

Ο διαχωρισμός σε test και control δείγματα έγινε με τη λογική ότι στα δείγματα όπου τα ADAR εκφράζονται κανονικά, αναμένεται φυσιολογικός αριθμός φαινομένων RNA τροποποίησης ενώ στο δείγμα ελέγχου, η αποσιώπηση των ADAR αναμένεται να μειώσει σημαντικά τον αριθμό των φαινομένων. Επομένως, η χρήση του control δείγματος έχει αξία για την ανάδειξη του βαθμού των false positive που αναφέρει κάθε εργαλείο.

Στο στάδιο της πρόβλεψης στόχων των miRNA χρησιμοποιήθηκε επιπλέον το δείγμα ENCLB420RAA από το αποθετήριο ENCODE για να ενισχύσει περαιτέρω τη στατιστική σημαντικότητα των ευρημάτων αυτού το βήματος. Το πείραμα εκτελέστηκε από το ίδιο εργαστήριο που διεξήγαγε και το ENCLB155EFP και αποτελεί το πειραματικό control,

δεδομένου ότι έχει γίνει εισαγωγή ενός shRNA το οποίο δεν στοχεύει κάποιο γονίδιο και άρα τα ADAR γονίδια εκφράζονται φυσιολογικά (το δείγμα που εμφανίζει τη φυσιολογική συμπεριφορά που υπάρχει στη φύση ονομάζεται wild-type). Και αυτό το δείγμα προέρχεται από strand-specific mRNA-seq, στην ίδια κυτταρική σειρά (K562) και λήφθηκε σε fastq μορφή. Περισσότερες πληροφορίες για το ENCLB420RAA αναφέρονται στον πίνακα 2.

Πίνακας 2. Πληροφορίες για το mRNA-seq δείγμα ENCLB420RAA από το αποθετήριο ENCODE.

Sample ID	Assay	Biosample	Organ	Cell type	Cell line	Layout	Instrument	# of sequences	Read length
ENCLB420RAA	RNA-seq	*Homo sapiens* K562	Blood	Erythroleukemia	**Chronic myelogenous leukemia**	Paired/Strand-specific	Illumina HiSeq 2000	~14.75M	100

Επιπλέον, συγκεντρώθηκαν και λήφθηκαν όλα τα αρχεία που ήταν αναγκαία ως είσοδος για την εκτέλεση των εργαλείων, τα οποία επισημαίνονται στην ακόλουθη λίστα μαζί με τις εκδόσεις τους. Η χρήση τους περιγράφεται στο αντίστοιχο κεφάλαιο του βήματος στο οποίο εφαρμόστηκαν.

Για το στάδιο του εντοπισμού RNA τροποποιήσεων:

- GRCh38_full_analysis_set_plus_decoy_hla.fa (γονιδίωμα αναφοράς του *Homo sapiens*, πάνω στο οποίο θα αντιστοιχηθούν όλα τα reads)
- GTF έκδοσης Ensembl 94 (https://www.ensembl.org/index.html) [78] (χαρακτηρισμοί γονιδιακών περιοχών του γονιδιώματος του *Homo sapiens*)
- Rmsk από το UCSC goldenPath (http://genome.ucsc.edu/) [79] (χαρακτηρισμοί των περιοχών του γονιδιώματος του *Homo sapiens* που περιέχουν επαναλήψεις)

Αρχεία με καταγεγραμμένα SNP του DNA σε γενικότερους πληθυσμούς:

- dbSNP146 for hg 38
- ESP65*
- HAPMAP
- 1000G Omni 2.5

Για το στάδιο της πρόβλεψης στόχων των miRNA:

- Χαρακτηρισμοί 3'UTR του γονιδιώματος του *Homo sapiens* της έκδοσης GENCODE v19 από το UCSC
- Βοηθητικό υλικό που προσφέρεται για δημόσια χρήση από τους συγγραφείς του TargetScan

2.2 Εντοπισμός RNA τροποποιήσεων

Στο πρώτο στάδιο της διπλωματικής εργασίας, αφότου συλλέχθηκαν τα απαραίτητα δεδομένα, χρησιμοποιήθηκαν 3 εργαλεία τα οποία εντοπίζουν RNA τροποποιήσεις σε RNA-seq δεδομένα και συγκρίθηκαν με στόχο την εύρεση της καλύτερης επιλογής με κριτήριο την ακρίβεια (precision) και την ευαισθησία (sensitivity). Η ακρίβεια δίνεται από την ακόλουθη εξίσωση:

$$ακρίβεια = \frac{true\ positive}{true\ positive + false\ positive}$$

, όπου τα true positive αποτελέσματα είναι οι πραγματικές RNA τροποποιήσεις που επεστράφησαν και false positive αυτά που λανθασμένα θεωρήθηκαν RNA τροποποιήσεις. Διαισθητικά, η ακρίβεια υποδεικνύει το ποσοστό των πραγματικών γεγονότων που κατάφερε να επιστρέψει ως αποτέλεσμα το κάθε εργαλείο. Για τον υπολογισμό της ακρίβειας χρησιμοποιήθηκε η βάση δεδομένων RADAR [80], η οποία περιλαμβάνει A-σε-I RNA τροποποιήσεις που έχουν επιμεληθεί είτε με πειραματικό είτε με αυτόματο τρόπο. Στην παρούσα εργασία, η ευαισθησία υπονοείται με την απλοποιημένη έννοια του συνολικού πλήθους των αποτελεσμάτων, εφόσον δεν είναι γνωστό το πλήθος των πραγματικών γεγονότων στα δείγματα.

Τα 3 εργαλεία μοιράζονται την ίδια φιλοσοφία ως προς τα βήματα που ακολουθούν. Αρχικά, όταν τα αρχεία εισόδου που περιέχουν τα reads είναι σε μορφή fastq, τότε καλείται ο αλγόριθμος που θα εκτελέσει το alignment (aligner). Η τακτική και οι παράμετροι που θα εφαρμοστούν σε αυτό το βήμα είναι κρίσιμης σημασίας, καθώς θα καθορίσουν τον αριθμό των false positive που προέρχονται από λάθος στοίχιση. Για το λόγο αυτό, τα κριτήρια που χρησιμοποιούνται κατά τη διάρκεια και μετά το πέρας της διαδικασίας είναι αυστηρά και επιλέγονται μόνο τα reads για τα οποία έχει υπολογιστεί αξιόπιστη στοίχιση. Στη συνέχεια, διάφορα φίλτρα ποιότητας εφαρμόζονται, ώστε να μειωθούν τα λάθη που έχουν προκύψει κατά την ανάγνωση της αλληλουχίας και με τη χρήση διαφόρων στατιστικών μετρικών (π.χ. συχνότητα εμφάνισης, αριθμός reads που υποστηρίζουν την τροποποίηση κ.ά.) εντοπίζονται πιθανές θέσεις τροποποίησης. Από το σύνολο που προκύπτει, φιλτράρονται τα DNA SNP και εξετάζονται πολύ αυστηρά ή αφαιρούνται θέσεις που βρίσκονται σε σημεία που εμφανίζουν μεγάλη πιθανότητα λάθους στο διάβασμα και την αντιστοίχιση, όπως περιοχές κοντά σε σημεία ματίσματος και γονιδιωματικές περιοχές που φέρουν επαναλαμβανόμενο DNA, στις οποίες έχουν παρατηρηθεί τεχνικές δυσκολίες στην αξιόπιστη ανάγνωση, ανάλογα με την τεχνολογία αλληλούχισης που χρησιμοποιείται.

Οι υποψήφιες RNA τροποποιήσεις του τελικού συνόλου φιλτράρονται στο τέλος μέσω της διωνυμικής κατανομής, ώστε να διαχωρισθούν από τα λάθη αλληλούχισης. Χρησιμοποιώντας την εξίσωση της κατανομής

$$P\ \text{value} = \sum_{\mathrm{m=k}}^{\mathrm{n}} \binom{\mathrm{n}}{\mathrm{m}} \mathrm{p^m q^{n-m}}, \mathrm{m} \leq \mathrm{n}$$

υπολογίζεται ένα p-value, το οποίο εκφράζει την πιθανότητα η τροποποίηση που έχει γίνει να είναι πραγματική, δεδομένων: του πλήθους τροποποιήσεων k, του συνολικού αριθμού βάσεων m που παρατηρήθηκαν σε αυτή τη θέση και πέρασαν τον έλεγχο ποιότητας αλληλούχισης και της πιθανότητας λάθους διαβάσματος p (που είναι μια σταθερή τιμή που σχετίζεται με την τεχνολογία αλληλούχισης). Υπενθυμίζεται ότι $q = 1 - p$. Το p-value που δίνεται ως αποτέλεσμα διορθώνεται έπειτα μέσω της στατιστικής

δοκιμασίας του False Discovery Rate (FDR), ώστε να ληφθεί υπόψιν το μέγεθος του πλήθους που δοκιμάζεται με τη διωνυμική κατανομή, το οποίο αυξάνει την τάση για λάθος υπολογισμό του p-value (π.χ. αν υπολογίζεται στο διάστημα 95%, τότε στους 100 υπολογισμούς οι 5 είναι στατιστικά πιθανό να είναι εσφαλμένοι, στους 1000 αυξάνονται στους 50 κ.ο.κ.). Οι τιμές συγκρίνονται με ένα κατώφλι και μόνο οι στατιστικά σημαντικές θεωρούνται πραγματικές A-σε-I τροποποιήσεις.

Εξαιτίας της διαφοράς στη συχνότητα του φαινομένου στις Alu και non-Alu περιοχές οι παράμετροι εντοπισμού στις μεταξύ τους διαφέρουν, ώστε να επιτυγχάνεται η επιθυμητή ακρίβεια. Για το λόγο αυτό, εκτελούνται 2 διαφορετικοί έλεγχοι για τις Alu και τις non-Alu περιοχές, με τις δεύτερες να εξετάζονται με πιο αυστηρές παραμέτρους, εφόσον περιέχουν λιγότερες τροποποιήσεις που είναι πραγματικές A-σε-I μεταλλαγές.

Όλα τα εργαλεία της σύγκρισης χρησιμοποιούν συμπληρωματικά τον BLAT [81], έναν aligner ο οποίος είναι πιο ακριβής, δίνει περισσότερα σημεία αντιστοίχισης και βελτιστοποιεί τα κενά που προκύπτουν σε κάθε αντιστοίχιση, με κόστος τον αυξημένο χρόνο εκτέλεσης. Για να αντισταθμίσουν αυτό το μειονέκτημα, τα εργαλεία χρησιμοποιούν τον BLAT για τοπικές μόνο αντιστοιχίσεις, σε σημεία μειωμένης αξιοπιστίας.

Οι εκδόσεις των εργαλείων που χρησιμοποιήθηκαν επισημαίνονται στον πίνακα 4.

Πίνακας 3. Τα εργαλεία εντοπισμού RNA τροποποιήσεων και οι εκδόσεις που χρησιμοποιήθηκαν στη διπλωματική εργασία.

Εργαλείο	Έκδοση
RES-Scanner	**1.0**
REDItools	**1.0.4**
RNAEditor	**1.0**

2.2.1 RES-Scanner

Το RES-Scanner δέχεται RNA και DNA ζευγάρια αρχείων είτε σε fastq είτε σε BAM μορφή. Είναι γραμμένο σε Perl. Ο χρήστης παροτρύνεται να εισάγει πολλαπλά δείγματα για ανάλυση, τα οποία βελτιώνουν την ακρίβεια του εντοπισμού πραγματικών RNA τροποποιήσεων (δίνεται η δυνατότητα εντοπισμού θέσεων που έχουν μικρή συχνότητα ή μικρό αριθμό υποστήριξης από reads και που κανονικά δεν θα τηρούσαν τα κριτήρια πραγματικής RNA τροποποίησης) αλλά και SNP (περισσότερα δείγματα DNA δίνουν περισσότερη πληροφορία για τις βάσεις που υπάρχουν σε κάθε εξεταζόμενη θέση). Η σύνοψη των βημάτων που ακολουθεί το εργαλείο φαίνεται παρακάτω στο σχήμα 2.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, ο αλγόριθμος εκτελέστηκε ξεχωριστά για τα δείγματα ERR188182, ERR188298 και ENCLB155EFP, ώστε να εξεταστεί η ορθότητα της συμπεριφοράς του. Στις περιπτώσεις που υπήρχε fastq (RNA-seq όλων των δειγμάτων, exome-seq/DNA για το K562) εκτελέστηκαν τα script που δίνει το RES-Scanner για alignment, ενώ σε κάθε άλλη περίπτωση η διαδικασία εκτελέστηκε από το βήμα του identification.

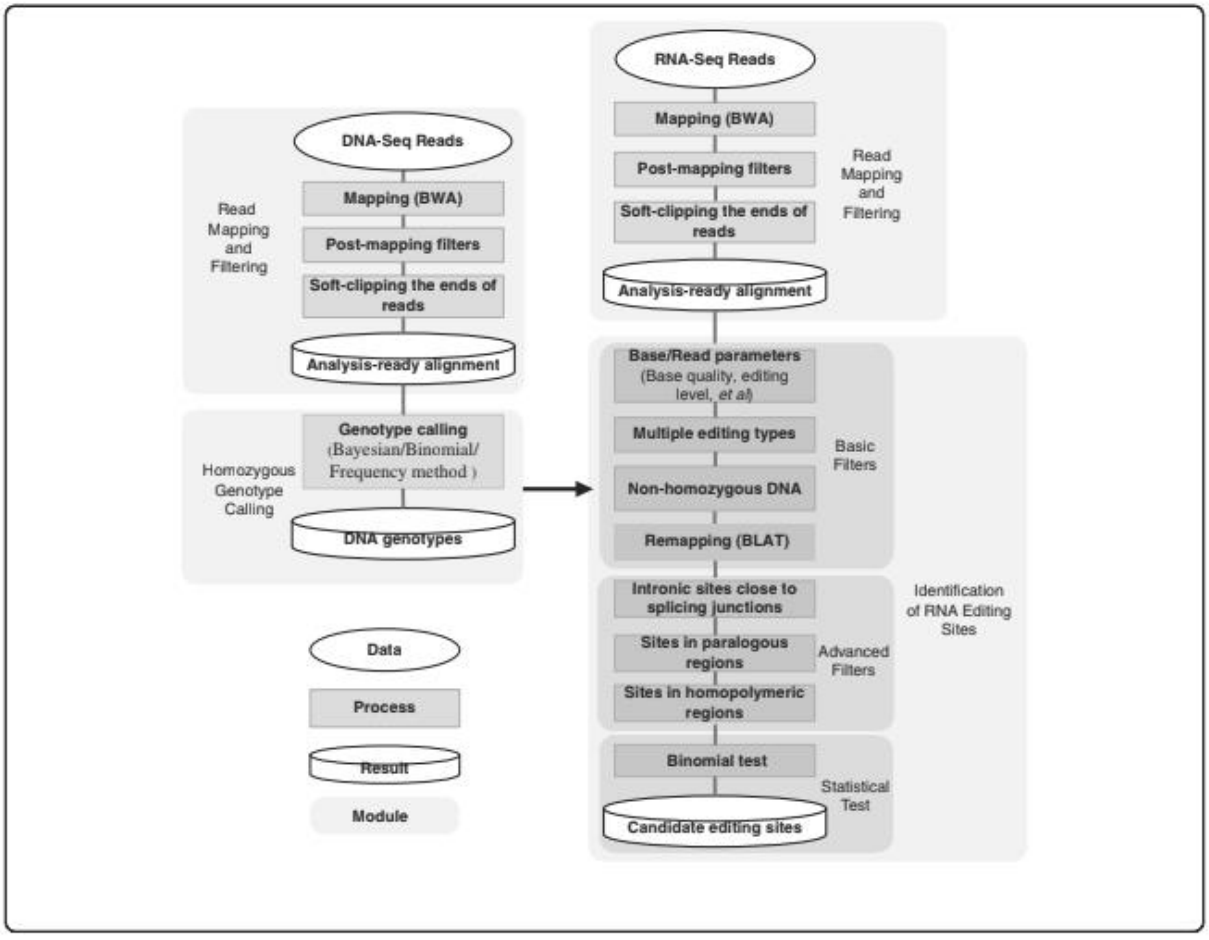
Πριν το alignment, απαιτείται η δημιουργία ενός φακέλου που περιέχει τους χαρακτηρισμούς των περιοχών του γονιδιώματος για κάθε θέση, ώστε να αναφέρονται

για κάθε αποτέλεσμα τα στοιχεία που υπάρχουν στην περιοχή που εντοπίστηκε. Οι χαρακτηρισμοί αφορούν τη θέση των στοιχείων 3'UTR, 5'UTR, CDS, intron και repeat. Η μορφή στην οποία συντάσσονται τα αρχεία είναι στήλες, χωρισμένες με τον ειδικό χαρακτήρα TAB (TAB-delimited) και περιλαμβάνουν τα εξής:

ID (το αναγνωριστικό που θα αναφέρεται)	χρωμόσωμα	έλικα (+ ή -)	αρχή	τέλος

π.χ.

ENSG00000223972@@ENST00000456328_intron-N1	chr1	+	12228	12612



Σχήμα 2. Σύνοψη των βημάτων εκτέλεσης του RES-Scanner

Το διάστημα των συντεταγμένων είναι 1-based inclusive, δηλαδή της κλειστής μορφής [αρχή, τέλος].

Για τη δημιουργία των αρχείων συντεταγμένων συντάχτηκαν 2 scripts στη γλώσσα προγραμματισμού Python, τα οποία δέχονται ως είσοδο ένα αρχείο που περιέχει όλα τα annotations (*Homo sapiens* GTF Ensembl 94 για UTR, CDS, intron και rmsk από το UCSC golden path για τα repeats του ίδιου οργανισμού) και παράγουν ως έξοδο αρχεία

στην προαναφερθείσα μορφή. Ιδιαίτερη προσοχή δόθηκε στη δημιουργία των χαρακτηρισμών για τα introns, καθώς δεν περιέχεται ο χαρακτηρισμός τους στο GTF. Θεωρήθηκε ως intron κάθε περιοχή μεταξύ 2 CDS, ξεκινώντας από το τέλος της προηγούμενης (CDS + 1) και καταλήγοντας μία βάση πριν την αρχή της επόμενης (CDS – 1). Στον αντίθετο κλώνο (του οποίου τα άκρα είναι αντεστραμμένα στο DNA, δηλαδή η αρχή θεωρείται το τέλος του άλλου κλώνου), τα αρχεία χαρακτηρισμού ξεκινούν τη μέτρηση από το 3’ άκρο, δηλαδή από το βιολογικό τέλος του. Για το λόγο αυτό, οι θέσεις που σημειώνονται είναι και αυτές αντεστραμμένες, με την αρχή να θεωρείται στην επόμενη (CDS – 1) και το τέλος στην προηγούμενη (CDS + 1).

Επιπλέον, απαιτείται από τον χρήστη η δημιουργία δύο TAB-delimited configuration αρχείων, ενός για το alignment βήμα και ενός για το βήμα του εντοπισμού των τροποποιήσεων. Σε αυτά δηλώνονται οι ετικέτες όλων των δειγμάτων που πρόκειται να συμπεριληφθούν στην ανάλυση, τα μονοπάτια στα οποία βρίσκονται στο σύστημα αρχείων του υπολογιστή και κάποια συμπληρωματικά μετα-δεδομένα. Για το βήμα του alignment, τα αρχεία που συντάχθηκαν ήταν της παρακάτω μορφής:

RNA ή DNA	ID δείγματος	Max insertion size	Μονοπάτι αρχείων στο σύστημα (2 μονοπάτια αν πρόκειται για paired-end δείγματα)

Το Max insertion size αναφέρεται στην μέγιστη απόσταση μεταξύ 2 paired reads ώστε να θεωρηθεί σωστό το mapping από τον aligner που χρησιμοποιεί το RES-Scanner. Η τιμή λαμβάνεται υπόψιν μόνο σε περιπτώσεις κακού alignment, όταν δεν μπορεί να υπολογιστεί αυτόματα κατά την εκτέλεση. Οι συγγραφείς του RES-Scanner προτείνουν για τα RNA δείγματα τιμή ίση με το μεγαλύτερο καταγεγραμμένο intron στο γονιδίωμα αναφοράς, εφόσον το μήκος του αντιστοιχεί στη μέγιστη απόσταση 2 εξωνίων, ενώ για το DNA προτείνουν μια αρκετά μικρότερη τιμή. Στην παρούσα μελέτη, για τα DNA δείγματα χρησιμοποιήθηκε η τιμή 373, ενώ για τα RNA δείγματα χρησιμοποιήθηκε η τιμή 1.500.000 (όπως προτείνεται στο συμπληρωματικό υλικό του αλγορίθμου).

Ο aligner που χρησιμοποιείται από το RES-Scanner είναι ο Burrows-Wheeler Aligner (BWA) [82], ο οποίος εξειδικεύεται σε short sequencing reads, όπως είναι αυτά που προκύπτουν κατά κανόνα από το sequencing για RNA. Τρέχει σε aln mode και όχι σε mem, διότι έτσι λαμβάνεται υπόψιν ολόκληρο το read κατά το alignment, ενώ το mem προσπαθεί να υπολογίσει το alignment και σε τοπικά σημεία εντός του read, οδηγώντας σε μεγαλύτερο ποσοστό συνολικής αντιστοίχισης αλλά με περισσότερα false positive σε κάποιες περιπτώσεις. Επειδή ο BWA δε λαμβάνει υπόψιν τα σημεία ματίσματος (splice sites ή junctions), στα οποία όπως έχει προαναφερθεί γίνονται συχνά λάθη αντιστοίχισης, το alignment του RES-Scanner είναι χωρισμένο σε 2 βήματα: στο πρώτο υπολογίζονται τα junctions και προαιρετικά γίνεται indexing και τμηματοποίηση του γονιδιώματος αναφοράς, ώστε να χωριστεί σε κομμάτια στα οποία θα μπορεί να γίνεται παράλληλη αναζήτηση για μεγαλύτερη απόδοση και ταχύτητα, ενώ στο δεύτερο υπολογίζεται το alignment. Τα scripts που εκτελούν αυτή τη διαδικασία είναι τα RES-Scanner_alignment.part1.pl και RES-Scanner_alignment.part2.pl, με την ενδιάμεση εκτέλεση είτε από τον χρήστη είτε με αυτόματη κλήση με την παράμετρο --run, κάποιων bash script που εκτελούν τη δημιουργία, τη μετατροπή και την εκτέλεση ενδιάμεσων αρχείων, φακέλων και λειτουργιών. Η εκτέλεση των βημάτων έγινε με τις τιμές που προτείνουν οι συγγραφείς, δηλαδή με τις default τιμές για το μέγεθος του seed sequence (32) με το οποίο αναζητά σημεία στα οποία θα ξεκινήσει να αντιστοιχίζει τις βάσεις ο αλγόριθμος, μέγιστο όριο λαθών στο seed sequence (3) και ποσοστό μη αντιστοίχισης στην τελική στοίχιση του read (4%). Μετά την επιτυχή εκτέλεση των 2 βημάτων του

alignment, προκύπτει ένα τελικό αρχείο BAM για κάθε δείγμα, που περιέχει τα αντιστοιχισμένα reads, μαζί με κάποια μετα-δεδομένα. Οι εντολές που χρησιμοποιήθηκαν για το ENCLB155EFP δίνονται παρακάτω, ενώ αντίστοιχες χρησιμοποιήθηκαν και για τα RNA δείγματα των ERR188182 και ERR188298.

```
perl RES-Scanner_alignment.part1.pl –outDir /BWA_aligned_files
--ref GRCh38_full_analysis_set_plus_decoy_hla.fa --bwa /bwa.kit/bwa --index 1
--config K562_ADARkd_alignment.configuration.txt --junction /posdir/intron.pos
--split --run
```

```
perl RES-Scanner_alignment.part2.pl --outDir /BWA_aligned_files
--ref GRCh38_full_analysis_set_plus_decoy_hla.fa --bwa /bwa.kit/bwa --index 1
--config /K562_ADARkd_alignment.configuration.txt
--samtools /samtools-0.1.18/samtools --n 0.04 --t 8 --run
```

Το επόμενο βήμα είναι αυτό του εντοπισμού των RNA τροποποιήσεων. Το script που εκτελεί τη διαδικασία είναι το RES-Scanner_identification.pl. Η κύρια είσοδος του είναι το configuration αρχείο που περιέχει τις πληροφορίες για τα BAM αρχεία στα οποία θα τρέξει η ανάλυση, το οποίο έχει αντίστοιχη μορφή με αυτό που χρησιμοποιήθηκε στο βήμα του alignment.

ID δείγματος	Μονοπάτι του RNA BAM	Μονοπάτι του DNA BAM

Προσοχή πρέπει να δοθεί στην εκτέλεση αυτού του βήματος, καθότι τα BAM αρχεία περιέχουν ετικέτες (flags), που χαρακτηρίζουν το alignment από κάθε read. Οι ετικέτες αυτές αξιοποιούνται σε μετέπειτα στάδια κατά τον εντοπισμό των RNA τροποποιήσεων και πρέπει να είναι συμβατές με την έκδοση του BWA που χρησιμοποιεί το RES-Scanner, ειδάλλως ενδέχεται να τερματιστεί κάποια εσωτερική λειτουργία κατά την εκτέλεση και να προκύψουν κενά αποτελέσματα. Για το λόγο αυτό, τα pre-aligned DNA BAM των ERR188182 και ERR188298 φιλτραρίστηκαν για flags με τιμές >256 χρησιμοποιώντας την εργαλειοθήκη samtools [83], αφαιρώντας έτσι reads τα οποία δεν έχουν μοναδική στοίχιση, δεν έχει γίνει ποιοτική αλληλούχιση ή είναι διπλότυπα (η ανάλυση του ορισμού γίνεται παρακάτω).

Η ποιοτική προ-επεξεργασία καθώς και όλα τα βήματα για το φιλτράρισμα των λαθών αλληλούχισης, των SNP και των λαθών στην αντιστοίχιση είναι ενσωματωμένα στο script και ελέγχονται με παραμέτρους από το χρήστη, ανάλογα με τα δείγματα προς ανάλυση. Το script είναι ήδη παραμετροποιημένο ώστε να έχει μια μέση, ορθή συμπεριφορά. Οι παράμετροι που μπορούν να οριστούν από το χρήστη, όπως αναφέρονται στο βοηθητικό υλικό που δίνεται από τους συγγραφείς του RES-Scanner, δίνονται παρακάτω.

--config	*FILE*	Tab-delimited configuration file with three columns (see details in Input above).
--out	*STR*	The output directory.
--genome	*FILE*	Reference genome.
--ss	*NUM*	Strand-specific RNA-seq data; '1' for yes, '0' for no . Note: Only strand-specific RNA-seq library generated by the dUTP protocol is currently supported [1].
--trim	*INT*	The number of bases self-clipped at 5' and 3' ends of a read, respectively [6,6].
--mis	*NUM*	The maximum number of mismatches allowed in a read alignment [5].
--q	*NUM*	Phred-scaled base quality score cutoff [30].

--mq	*NUM*	Mapping quality score cutoff [20].
--phred	*NUM*	Encoding methods of Phred quality score for reads in DNA.bam and RNA.bam files,
--posdir	*STR*	The directory for genomic feature position files. The files in the directory should be named as *FeatureName*.pos (e.g. 5UTR.pos, CDS.pos, intron.pos, 3UTR.pos, ncRNA.pos, repeat.pos, etc.). If a file with the name of CDS.pos is provided, the function of inferring the codon and amino acid change after RNA editing is activated [null].
--editLevel	*Float*	The minimum editing level required by a candidate editing site; range from 0 to 1 [0.05].
--editDepth	*INT*	The minimum number of RNA reads supporting editing for a candidate editing site [3].
--extremeLevel	*NUM*	Exclude editing sites with extreme editing levels (100%); '1' for yes, '0' for no [0].
--refined	*NUM*	Whether refined the number of RNA reads supporting candidate editing sites; '1' for yes, '0' for no [1].
--refinedDepth	*INT*	The minimum number of RNA reads in the middle of its length supporting editing for a candidate editing site. (e.g. from positions 23-68 of a 90-bp read) [1].
--readType	*INT*	The minimum number of *unique* RNA reads supporting editing for a candidate editing site [3].
--junctionCoordinate	*FILE*	The file named junctionFlankSequenceRegion.txt created by the RES-Scanner alignment --junction option, applicable only for the input reference genome including exonic sequences surrounding splicing junctions [null].
--editPvalue	*Float*	The cutoff of binomial test FDR for candidate editing sites [0.05].
--ploidy	*INT*	Ploidy level of the samples: 1 for haploid, 2 for diploid, 3 for triploid, 4 for tetraploid, etc. [2].
--paralogous_R	*NUM*	Remove candidate editing sites from those regions that are similar to other parts of the genome by BLAT alignment; '1' for yes, '0' for no [1].
--paralogous_D	*NUM*	Discard candidate editing sites with DNA reads depth of more than twice the genome-wide peak or mean depth; '1' for yes, '0' for no [1].
--homopolymer	*NUM*	Remove candidate editing sites in homopolymer runs of ≥ 5 bp; '1' for yes, '0' for no [1].
--intronic	*NUM*	Remove intronic candidate editing sites occurring within n bases of a splice site [6].
--knownSNP	*FILE*	The file of known SNPs in GFF3 format [null].
--rmdup	*NUM*	Remove PCR duplicates for BAM file; '1' for yes, '0' for no [1].
--bestHitRatio	*Float*	The proportion of qualifying reads relative to all BLAT-realigned reads. Note: force --paralogous_R [0.5].
--uniqTag	*NUM*	Identify the unique mapping, without suboptimal hits, reads in BAM file with the tags 'XT:A:U', 'X0:i:1' and 'X1:i:0' or base on flag value; '1' for tags, '0' for flag [1]. Note: If the BAM file was generated by the RES-Scanner alignment pipeline, please set --uniqTag 1 to infer unique alignment.
--samtools	*FILE*	The absolute path of the SAMtools package pre-installed on the local machine (mandatory).
--blat	*FILE*	The absolute path of the BLAT software pre-installed on the local machine (mandatory).
--run		Run the jobs in a serial working mode (i.e. run the jobs one by one automatically).
--help		Show the help information.

Parameters for homozygous genotype calling:

--method	*STR*	Method for calling homozygous genotypes: Bayesian, Binomial or Frequency [Bayesian].
--HomoPrior	*Float*	The prior probability for a genomic position to be homozygous (force --method Bayesian) [0.99].
--rate	*NUM*	The rate of transitions over transversions of the genome (force --method Bayesian) [2].
--Bayesian_P	*Float*	The minimum Bayesian posterior probability cutoff for calling a homozygous genotype; range from 0 to 1, the bigger the better (force --method Bayesian) [0.95].
--Binomial_P	*Float*	The maximum p-value cutoff of the binomial test for calling a homozygous genotype; range from 0 to 1, the smaller the better (force --method Binomial) [0.05].
--Binomial_FDR	*Float*	The maximum FDR cutoff of the binomial test for calling a homozygous genotype; range from 0 to 1, the smaller the better (force --method Binomial) [0.05].
--Frequency_N	*NUM*	The maximum count of the alternative allele present in the DNA-seq data for a candidate editing site (force --method Frequency) [0].
--Frequency_R	*Float*	The maximum frequency of the alternative allele present in the DNA-seq data for a candidate editing site; range from 0 to 1 (force --method Frequency) [0].

Οι συγγραφείς του RES-Scanner έχουν υλοποιήσει μια σειρά από προσεγγίσεις ώστε να μειώσουν τα false positive αποτελέσματα. Με την παράμετρο --readType, απαιτούνται n μοναδικά reads τα οποία να υποστηρίζουν την τροποποίηση κοντά στο κέντρο τους, καθότι οι άκρες των reads είναι επιρρεπείς σε λάθη αλληλούχισης, ενώ με το --refinedDepth ορίζεται ο ελάχιστος αριθμός των reads που θα υποστηρίζουν την ίδια ιδιότητα, ανεξαρτήτως μοναδικότητας. Επίσης, ο BLAT παραμετροποιείται για την αντιμετώπιση της αντιστοίχισης σε ομοπολυμερικές περιοχές, οι οποίες είναι περιοχές σε διαφορετικό σημεία του γονιδιώματος οι οποίες μοιάζουν σε μεγάλο βαθμό μεταξύ τους και στις οποίες συνήθως υπολογίζονται αναξιόπιστες αντιστοιχίσεις. Ορίζοντας τιμή στο --bestHitRatio, επιλέγεται το ελάχιστο ποσοστό των reads που περνούν τον BLAT έλεγχο σε σχέση με όλα τα reads που επαναντιστοιχήθηκαν, εκτιμώντας έτσι τον βαθμό εμπιστοσύνης στις περιοχές αναξιοπιστίας. Με το --paralogous_R, τροποποιήσεις που βρίσκονται σε περιοχές που ομοιάζουν πολύ μεταξύ τους, σύμφωνα με το BLAT, απορρίπτονται. Με την παράμετρο --ss ο αλγόριθμος μπορεί να συμπεράνει την έλικα προέλευσης της τροποποίησης που παρατηρεί, δεδομένου ότι το πρωτόκολλο που έχει χρησιμοποιηθεί ώστε να παραχθούν τα strand-specific δείγματα είναι το dUTP. Τέλος, για την πρόβλεψη των SNP στο DNA δείγμα έχουν υλοποιηθεί 3 μοντέλα που υπολογίζουν την πιθανότητα ένας γονότυπος να είναι ομόζυγος (δηλαδή να εμφανίζεται η ίδια DNA βάση σε όλες τις αντιγραφές της στα χρωμοσώματα που υπάρχουν στον οργανισμό).

Στο default μοντέλο, χρησιμοποιείται ένα Bayesian μοντέλο με τις πιθανότητες εμφάνισης κάθε πιθανού συνδυασμού με βάση τα δεδομένα εισόδου (π.χ. για ένα δείγμα του *Homo sapiens* ισχύουν οι εξής συνδυασμοί: AA, TT, CC, GG, AT, AC, AG, CT, CG and GT). Από αυτούς, μόνο αυτοί που ξεπερνάνε κάποιο ελάχιστο κατώφλι στην posterior πιθανότητα εξετάζονται για RNA τροποποιήσεις. Η εξίσωση που περιγράφει το Bayesian μοντέλο δοθέντων G (γενότυπος) και D (τα αντιστοιχισμένα δεδομένα με τις ποιότητες αλληλούχισης) είναι η ακόλουθη:

$$\mathrm{p(G|D)} = \frac{\mathrm{p(G)p(D|G)}}{\mathrm{p(D)}}$$

Το $\mathrm{p(G)}$ είναι η prior πιθανότητα ομοζυγοτίας για κάθε γονότυπο. Εξαρτάται από 3 παραμέτρους: το επίπεδο ομοζυγοτίας του εκάστοτε γονιδιώματος (--HomoPrior), το ποσοστό GC και το ποσοστό transition/transversion (παράμετρος --rate), καθώς έχει παρατηρηθεί ότι τα transitions (A ⇔ G και C ⇔ T) έχουν διπλάσια συχνότητα εμφάνισης στο ανθρώπινο γονιδίωμα σε σχέση με τα transversions (A ⇔ C, A ⇔ T, G ⇔ C ή G ⇔ T). Το ποσοστό εξαρτάται από το κομμάτι του γονιδιώματος που μελετάται π.χ. τα γονίδια του πυρήνα εμφανίζουν διαφορετικά ποσοστά. Το $\mathrm{p(D)}$ εκφράζεται από το άθροισμα των επιμέρους πιθανοτήτων εμφάνισης του κάθε νουκλεοτιδίου, εξαιτίας του θεωρήματος της ολικής πιθανότητας. Κάθε όρος έχει ειδικό βάρος, εξαιτίας της τεχνολογίας που χρησιμοποιείται για την αλληλούχιση π.χ. το μηχάνημα αλληλούχισης Illumina χρησιμοποιεί 2 laser για να ξεχωρίσει τα 4 νουκλεοτίδια, εκ των οποίων ένα πράσινο laser διαβάζει τα G και T και ένα κόκκινο τα A και C. Σαν αποτέλεσμα, τα περισσότερα λάθη στο διάβασμα της αλληλουχίας συμβαίνουν μεταξύ G/T και A/C. Τέλος, το $\mathrm{p(D|G)}$ υπολογίζεται από το γινόμενο των πιθανοτήτων εμφάνισης όλων των νουκλεοτιδίων σε ένα σημείο, δοθέντος ενός γενοτύπου. Διαισθητικά, υπολογίζεται η πιθανότητα να εμφανίζεται κάποια βάση αξιοποιώντας τις συχνότητες που παρατηρούνται στα δεδομένα και τις βάσεις που υπάρχουν στο DNA. Η εξίσωση που χρησιμοποιείται για τον

υπολογισμό της, δοθέντων b για τη βάση που υπάρχει στα δεδομένα, $pileup$ για το σύνολο βάσεων, n για τον αριθμό αντιγραφών κάθε χρωμοσώματος στον οργανισμό (ploidy level), $A1$ και $A2$ για τις αλληλομορφές βάσεων που συναντώνται στο γονιδίωμα και m το πλήθος της αλληλομορφής $A1$, είναι η εξής:

$$\mathrm{p(b|G)} = \mathrm{p(b|\{A1, A2\})} = \frac{\mathrm{m}}{\mathrm{n}}\mathrm{p(b|A1)} + \frac{\mathrm{n-m}}{\mathrm{n}}\mathrm{p(b|A2)}$$

Με περισσότερα δείγματα, οι συνδυασμοί των βάσεων είναι περισσότεροι, δίνοντας υψηλότερο βαθμό εμπιστοσύνης στην posterior πιθανότητα, όπως φαίνεται και από τη μαθηματική σχέση της εικόνας 5.

$$\binom{r+n-1}{r} = \frac{(r+n-1)!}{r!(n-1)!}$$

where n is the number of things to choose from,
and we choose r of them
repetition allowed,
order doesn't matter.

Εικόνα 5. Συνδυασμοί με επανάληψη (https://www.mathsisfun.com/combinatorics/combinations-permutations.html).

Επίσης, αντί για το Bayesian μοντέλο δίνεται στο χρήστη η επιλογή της Binomial κατανομής, με βάση την εξίσωση της κατανομής.

$$P\ \mathrm{value} = \sum_{\mathrm{m=0}}^{\mathrm{k}} \binom{\mathrm{n}}{\mathrm{m}} \mathrm{p^m q^{n-m}}\ , \mathrm{m} \leq \mathrm{n}$$

Με p ορίζεται η κατώτατη συχνότητα εμφάνιση μιας αλληλομορφής (π.χ. 0,5 για ploidy level = 2), με n ο αριθμός των DNA reads που έχουν περάσει τον έλεγχο της ποιότητας αλληλούχισης και k ο αριθμός των DNA reads που εμφανίζουν διαφορετική αλληλομορφή από αυτή του γονιδιώματος αναφοράς. Το p-value που προκύπτει εκφράζει την πιθανότητα να παρατηρείται εναλλακτική αλληλομορφή σε reads ≤ k, και άρα μικρές τιμές υποδεικνύουν ομοζυγωτία (κατώφλι ορισμένο από --Binomial_P). Γίνεται διόρθωση του p-value με FDR και μόνο τα σημεία με τιμή μεγαλύτερη ενός κατωφλίου (--Binomial_FDR) θεωρούνται αξιόπιστες ομόζυγες περιοχές.

Τέλος, με το Frequency μοντέλο γίνεται απλή εκτίμηση της πιθανότητας με βάση το read depth (κατώφλι με --Frequency_N) και την συχνότητα εμφάνισης (κατώφλι με --Frequency_R) της κάθε αλληλομορφής που εμφανίζεται.

Τα διαφορετικά μοντέλα έχουν αντίστροφη σχέση μεταξύ ακρίβειας και χρόνου εκτέλεσης, με το Bayesian μοντέλο να είναι το πιο ακριβές και με το μεγαλύτερο χρόνο εκτέλεσης.

Οι εντολές που εκτελέστηκαν στο πλαίσιο της διπλωματικής, με τις παραμέτρους που προτείνονται από τους συγγραφείς στις εκτελέσεις με *Homo sapiens* δείγματα, ήταν οι παρακάτω:

- ERR188182 Alu περιοχές

```
perl RES-Scanner_identification.pl --config ERR188182.config.txt
--out /res_scanner_alu_bwa_ERR188182/
--genome GRCh38_full_analysis_set_plus_decoy_hla.fa --ss 0
--samtools /samtools-0.1.18/samtools --trim 6,0 --q 25 --mq 20 --DNAdepth 10
--RNAdepth 2 --editLevel 0 --editDepth 2 --refined 0 --paralogous_R 0 --
paralogous_D 0 --homopolymer 0 --posdir /posdir/ --intronic 0 --editPvalue 1 --
rmdup 0 --uniqTag 1
--run
```

- ERR188182 non-Alu περιοχές

```
perl RES-Scanner_identification.pl --config ERR188182.config.txt
--out /res_scanner_non_alu_bwa_ERR188182/
--genome GRCh38/GRCh38_full_analysis_set_plus_decoy_hla.fa --ss 0
--samtools /samtools-0.1.18/samtools --blat /blatSuite.36/blat --trim 6,0 --q 25
--mq 20 --DNAdepth 10 --RNAdepth 3 --editLevel 0.1 --editDepth 3 --refined 1
--refinedDepth 1 -readType 3 --paralogous_R 1 --paralogous_D 1 --homopolymer 1
--posdir /posdir/ --intronic 4 --editPvalue 0.05 --rmdup 1 --bestHitRatio 0.6
--uniqTag 1 --run
```

Αντίστοιχα για το ERR188298.

- ENCLB155EFP Alu περιοχές

```
perl RES-Scanner_identification.pl --config ENCLB155EFP.config.txt
--out /res_scanner_alu_bwa_ENCLB155EFP/
--genome GRCh38_full_analysis_set_plus_decoy_hla.fa --ss 1
--samtools /samtools-0.1.18/samtools --trim 6,0 --q 25 --mq 20 --DNAdepth 10
--RNAdepth 2 --editLevel 0 --editDepth 2 --refined 0 --paralogous_R 0
--paralogous_D 0 --homopolymer 0 --posdir /posdir/ --intronic 0 --editPvalue 1
--rmdup 0
--uniqTag 1 --run
```

- ENCLB155EFP non-Alu περιοχές

```
perl RES-Scanner_identification.pl --config ENCLB155EFP.config.txt
--out /res_scanner_non_alu_bwa_ENCLB155EFP/
--genome GRCh38_full_analysis_set_plus_decoy_hla.fa --ss 1
--samtools /samtools-0.1.18/samtools --blat /blatSuite.36/blat --trim 6,0 --q 25
--mq 20 --DNAdepth 10 --RNAdepth 3 --editLevel 0.1 --editDepth 3 --refined 1
--refinedDepth 1 -readType 3 --paralogous_R 1 --paralogous_D 1 --homopolymer 1
--posdir /posdir/ --intronic 4 --editPvalue 0.05 --rmdup 1 --bestHitRatio 0.6
--uniqTag 1 --run
```

Τα αποτελέσματα δίνονται στη μορφή ενός TAB-delimited αρχείου με τις εξής στήλες:

Chromosome	χρωμόσωμα
Coordinate	συντεταγμένη περιοχής
Strand	έλικα
Gbase	βάση στο reference genome
EditType	βάση τροποποίησης
DNA_baseCount[A,C,G,T]	Πλήθος βάσεων στο DNA δείγμα
RNA_baseCount[A,C,G,T];P_value	Πλήθος βάσεων στο RNA δείγμα και p-value binomial

Στην περίπτωση που έχουν δοθεί και αρχεία με τις θέσεις των διαφόρων στοιχείων στο γονιδίωμα, τότε ο αλγόριθμος μπορεί να προχωρήσει σε χαρακτηρισμό των περιοχών στις οποίες βρίσκονται οι RNA τροποποιήσεις, προσθέτοντας τις παρακάτω επιπλέον στήλες:

TargetedGenomicFeature	χαρακτηρισμός περιοχής
TargetedFeatureID	ID χαρακτηρισμένου στοιχείου
CodonChange	όταν η τροποποίηση συμβαίνει στην τριπλέτα νουκλεοτιδίων ενός κωδικονίου και συμβάλλει στην αλλαγή του
AminoAcidChange	όταν η αλλαγή του κωδικονίου συμβάλλει στην μεταλλαγή ενός ολόκληρου αμινοξέος που αποτελείται από κωδικόνια

2.2.2 REDItools

Το REDItools αποτελεί μια συλλογή από scripts γραμμένα στη γλώσσα προγραμματισμού Python. Προσφέρει 3 scripts τα οποία μπορούν να χρησιμοποιηθούν για να εντοπίσουν RNA τροποποιήσεις σε RNA-seq δεδομένα, καθένα με διαφορετική προσέγγιση.

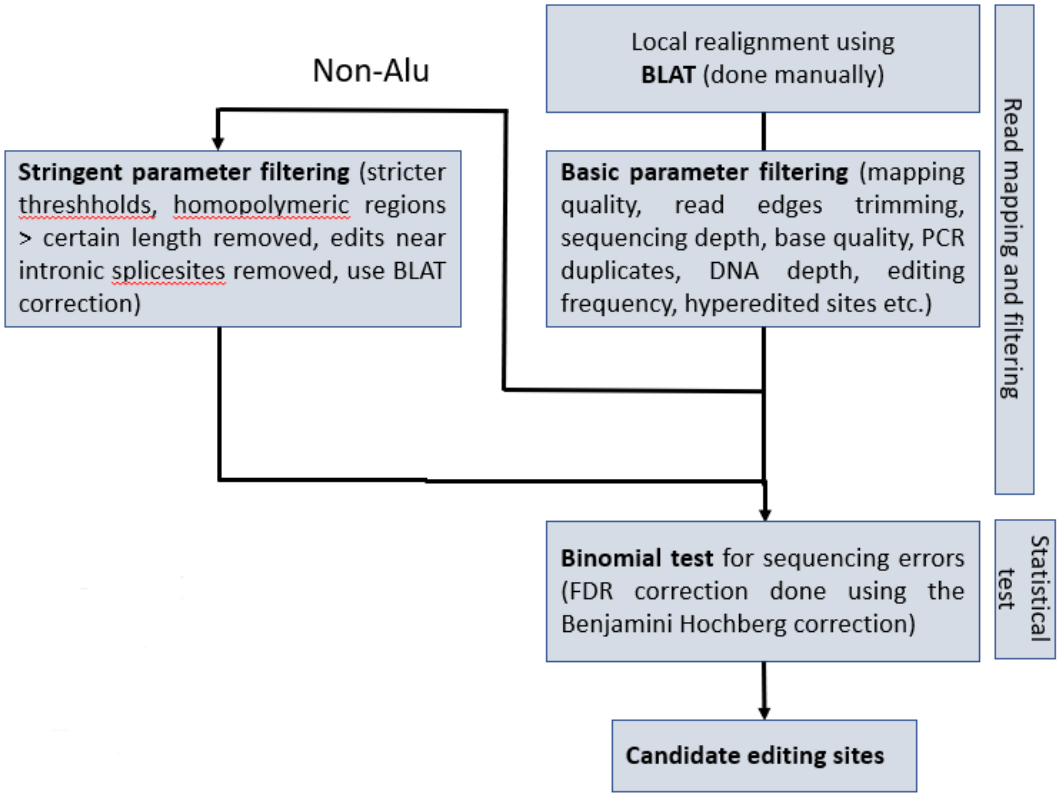
Το REDItoolDnaRna.py ανιχνεύει και αναγνωρίζει RNA τροποποιήσεις σε ένα prealigned δείγμα RNA. Αν δοθεί μόνο το RNA δείγμα, πραγματοποιεί εκτίμηση για πιθανές RNA τροποποιήσεις, ενώ αν δοθεί και το αντίστοιχο DNA δείγμα, τότε η πληροφορία που αντλείται από αυτό συνδυάζεται ώστε να φιλτραριστούν πιθανά SNP και να μειωθούν τα false positive αποτελέσματα. Η σύνοψη της διαδικασίας που ακολουθεί απεικονίζεται στο σχήμα 3.

Το REDItoolKnown.py βρίσκει και επιστρέφει RNA τροποποιήσεις σε ένα δείγμα RNA, τις οποίες ταυτοποιεί μέσω ενός αρχείου με γνωστές RNA τροποποιήσεις. Είναι χρήσιμο σε περιπτώσεις που η μελέτη αφορά συγκεκριμένες τροποποιήσεις σε κάποια δείγματα.

Το REDItoolDenovo.py ανιχνεύει και αναγνωρίζει RNA τροποποιήσεις σε ένα δείγμα RNA χωρίς a priori γνώση για αυτές, χρησιμοποιώντας την εμπειρική κατανομή που αναμένεται να παρουσιάζουν οι μεταλλαγές των νουκλεοτιδίων, καθώς και τα αντίστοιχα

φίλτρα και τις διαδικασίες που εφαρμόζονται στο στάδιο του εντοπισμού από το REDItoolDnaRna.py.

Το REDItools εστιάζει στη χρήση πολλαπλών παραμέτρων, οι οποίες μπορούν να ρυθμιστούν από τον χρήστη, ώστε να μειωθούν στο μέγιστο βαθμό οι false positive τροποποιήσεις. Οι default τιμές και σε αυτό το εργαλείο είναι ρυθμισμένες ώστε να παρουσιάζει μια μέση, ορθή συμπεριφορά. Οι παράμετροι που μπορούν να οριστούν στο REDItoolDnaRna.py αναφέρονται παρακάτω.



Σχήμα 3. Σύνοψη βημάτων του REDItoolDnaRna.py που περιλαμβάνεται στο REDItools.

Παράμετροι που περιλαμβάνονται στο REDItoolDnaRna.py:

-i RNA-Seq BAM file

-j DNA-Seq BAM files separated by comma or folder containing BAM files. Note that each chromosome/region must be present in a single BAM file only.

-I Sort input RNA-Seq BAM file

-J Sort input DNA-Seq BAM file

-f Reference file in fasta format. Note that chromosome/region names in the reference must match chromosome/region names in BAMs files.

-C Base interval to explore [100000]. It indicates how many bases have to be loaded during the run.

-k List of chromosomes to skip separated by comma or file (each line must contain a chromosome/region name).

-t Number of threads [1]. It indicates how many processes should be launched. Each process will work on an individual chromosome/region.

-o Output folder [rediFolder_XXXX] in which all results will be stored. XXXX is a random number generated at each run.

-F Internal folder name [null] is the main folder containing output tables.

-M Save a list of columns with quality scores. It produces at most two files in the pileup-like format.

-c Minimum read coverage (dna,rna) [10,10]

-q Minimum quality score (dna,rna) [25,25]

-m Minimum mapping quality score (dna,rna) [25,25]

-O Minimum homoplymeric length (dna,rna) [5,5]

-s Infer strand (for strand oriented reads) [1]. It indicates which read is in line with RNA. Available values are: 1:read1 as RNA,read2 not as RNA; 2:read1 not as RNA,read2 as RNA; 12:read1 as RNA,read2 as RNA; 0:read1 not as RNA,read2 not as RNA.

-g Strand inference type 1:maxValue 2:useConfidence [1]; maxValue: the most prominent strand count will be used; useConfidence: strand is assigned if over a prefixed frequency confidence (-x option)

-x Strand confidence [0.70]

-S Strand correction. Once the strand has been inferred, only bases according to this strand will be selected.

-G Infer strand by GFF annotation (must be GFF and sorted, otherwise use -X). Sorting requires grep and sort unix executables.

-K GFF File with positions to exclude (must be GFF and sorted, otherwise use -X). Sorting requires grep and sort unix executables.

-T Work only on given GFF positions (must be GFF and sorted, otherwise use -X). Sorting requires grep and sort unix executables.

-X Sort annotation files. It requires grep and sort unix executables.

-e Exclude multi hits in RNA-Seq

-E Exclude multi hits in DNA-Seq

-d Exclude duplicates in RNA-Seq

-D Exclude duplicates in DNA-Seq

-p Use paired concardant reads only in RNA-Seq

-P Use paired concardant reads only in DNA-Seq

-u Consider mapping quality in RNA-Seq

-U Consider mapping quality in DNA-Seq

-a Trim x bases up and y bases down per read [0-0] in RNA-Seq

-A Trim x bases up and y bases down per read [0-0] in DNA-Seq

-b Blat folder for correction in RNA-Seq

-B Blat folder for correction in DNA-Seq

-l Remove substitutions in homopolymeric regions in RNA-Seq

-L Remove substitutions in homopolymeric regions in DNA-Seq

-v Minimum number of reads supporting the variation [3] for RNA-Seq

-n Minimum editing frequency [0.1] for RNA-Seq

-N Minimum variation frequency [0.1] for DNA-Seq

-z Exclude positions with multiple changes in RNA-Seq

-Z Exclude positions with multiple changes in DNA-Seq

-W Select RNA-Seq positions with defined changes (separated by comma ex: AG,TC) [default all]

-R Exclude invariant RNA-Seq positions

-V Exclude sites not supported by DNA-Seq

-w File containing splice sites annotations (SpliceSite file format see above for details)

-r Num. of bases near splice sites to explore [4]

--gzip Gzip output files

-h, --help Print the help

Συγκριτικά με το RES-Scanner, το REDItools χρησιμοποιεί περισσότερες παραμέτρους. Ο BLAT εκτελείται πριν το βήμα του εντοπισμού RNA τροποποιήσεων, ώστε να βελτιωθεί το ποσοστό ακριβείας του alignment, με το κόστος του μεγαλύτερου χρόνου εκτέλεσης. Για το φιλτράρισμα των DNA SNP εφαρμόζονται στατιστικά φίλτρα, αντίστοιχα με το Frequency μοντέλο του RES-Scanner. Με την παράμετρο -s, ο αλγόριθμος μπορεί να συμπεράνει την έλικα προέλευσης της τροποποίησης, στην περίπτωση strand-specific δειγμάτων.

Πριν την εκτέλεση του script εντοπισμού, οι συγγραφείς προτείνουν μια διαδικασία προετοιμασίας των δειγμάτων, η οποία περιλαμβάνει το alignment, την κατασκευή ενός αρχείου με τα splice sites, καθώς και το BLAT realignment (επαναστοίχιση). Για το βήμα του alignment, οι συγγραφείς του REDItools προτείνουν το GSNAP (Genomic Short-read Nucleotide Alignment Program) [84]. Αντίστοιχα με τον BWA, το GSNAP εξειδικεύεται στα short sequencing reads, είναι όμως splice-aware, δηλαδή λαμβάνει υπόψιν τα splice sites. Η προετοιμασία περιλάμβανε τα εξής βήματα:

1. GSNAP alignment. Οι παράμετροι που χρησιμοποιήθηκαν είναι ίδιες με αυτές που προτείνονται στο manual των συγγραφέων. Δίνονται οι εντολές για τα προαπαιτούμενα της εκτέλεσης του GSNAP, καθώς και για το alignment των RNA και DNA δειγμάτων του ENCLB155EFP. Αντίστοιχες εντολές χρησιμοποιήθηκαν και για τα RNA δείγματα των ERR188182 και ERR188298.

```
# GR38.12 ⇔ Ensembl94 (GTF) ⇔ GENCODEv29
gmap_build -d GRCh38_full_analysis_set_plus_decoy_hla
-D GRCh38 GRCh38_full_analysis_set_plus_decoy_hla.fa

mkdir /GRCh38/spliceGMAP
cat GENCODEv29_knownGenes_annotation.txt |
psl_splicesites > /GRCh38/spliceGMAP/GENCODEv29_knownGenes_annotation.splicesites
cat /GRCh38/spliceGMAP/GENCODEv29_knownGenes_annotation.splicesites |
iit_store -o /GRCh38/spliceGMAP/GENCODEv29_knownGenes_annotation.splicesitesFile

cat /GENCODEv29_knownGenes_annotation.txt |
psl_introns > /GRCh38/spliceGMAP/GENCODEv29_knownGenes_annotation.introns
cat /GRCh38/spliceGMAP/GENCODEv29_knownGenes_annotation.introns |
iit_store -o /GRCh38/spliceGMAP/GENCODEv29_knownGenes_annotation.intronsFile


gsnap -d GRCh38_full_analysis_set_plus_decoy_hla -D /GRCh38/GMAP -B 5 -t 12
--use-splicing=GENCODEv29_knownGenes_annotation.intronsFile.iit
--splicingdir=/GRCh38/spliceGMAP -E 1000 -N1 -n1 -Q -O --nofails -A sam --gunzip
--split-output=/gsnap_aligned_files/rnaK562_gsnap -a paired ENCFF093ZYA.fastq.gz
ENCFF085DKT.fastq.gz
```

Από τα αρχεία που παράγονται, χρησιμοποιείται μόνο το αρχείο με κατάληξη concordant_uniq.

2. Μετατροπή των SAM αρχείων σε BAM και sort με τη χρήση των samtools.

- RNA ENCLB155EFP

```
samtools view -@ 8 -bS /gsnap_aligned_files/rnaK562_gsnap.concordant_uniq >
/gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.bam

samtools sort -@ 8 -o /gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.sorted.bam
/gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.bam
```

- DNA ENCLB155EFP

```
samtools view -@ 8 -bS /gsnap_aligned_files/dnaK562_gsnap.concordant_uniq >
/gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.bam

samtools sort -m 15000000000
-o /gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.sorted.bam
/gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.bam
```

3. Μαρκάρισμα των διπλότυπων reads που έχουν προκύψει από τη χρήση της polymerase chain reaction (PCR) τεχνικής, μέσω της οποίας πολλαπλασιάζονται τα reads του δείγματος ώστε να γίνεται πιο ακριβής η αλληλούχιση. Τα reads αυτά δεν εκφράζονται πραγματικά στα δείγματα, προσθέτοντας λανθασμένη πληροφορία στα δεδομένα. Επίσης, γίνεται δημιουργία των index των BAM αρχείων. Το εργαλείο που χρησιμοποιήθηκε είναι το MarkDuplicates από την εργαλειοθήκη Picard [85].

- RNA ENCLB155EFP

```
java -Xmx16G -jar /picard-tools/MarkDuplicates.jar INPUT=
/gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.sorted.bam OUTPUT=
/gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.sorted.nodup.bam METRICS_FILE=
/gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.sorted.metrics.txt
REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
MAX_RECORDS_IN_RAM=1000000 CREATE_INDEX=true
```

- DNA ENCLB155EFP

```
java -Xmx16G -jar /picard-tools/MarkDuplicates.jar INPUT=
/gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.sorted.bam OUTPUT=
/gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.sorted.nodup.bam METRICS_FILE=
/gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.sorted.metrics.txt
REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
MAX_RECORDS_IN_RAM=1000000 CREATE_INDEX=true TMP_DIR=/gsnap_aligned_files/tmp/
```

4. Κλήση του BLAT για τη δημιουργία φακέλου με τα realignments

Πριν κληθεί το REDItoolBlatCorrecion.py, απαιτείται ένα binary indexing του γονιδιώματος αναφοράς σε .2bit μορφή που χρησιμοποιείται από το gfClient, το οποίο έγινε με το εργαλείο faToTwoBit που περιέχεται στο φάκελο του BLAT.

- Binary indexing

```
/blatSuite.36/faToTwoBit GRCh38_full_analysis_set_plus_decoy_hla.fa
GRCh38_full_analysis_set_plus_decoy_hla.2bit
```

- RNA BLAT correction

```
python REDItoolBlatCorrection.py
-i /gsnap_aligned_files/rnaK562_gsnap.concordant_uniq.sorted.nodup.bam
-f GRCh38_full_analysis_set_plus_decoy_hla.fa
-F GRCh38_full_analysis_set_plus_decoy_hla.2bit
-o /gsnap_aligned_files/rnaK562BlatCorrectionRNA -V -T -t 4
```

- DNA BLAT correction

```
python REDItoolBlatCorrection.py
-i /gsnap_aligned_files/dnaK562_gsnap.concordant_uniq.sorted.nodup.bam
-f GRCh38_full_analysis_set_plus_decoy_hla.fa
-F GRCh38_full_analysis_set_plus_decoy_hla.2bit
-o /gsnap_aligned_files/dnaK562BlatCorrectionDNA -V -T -t 16
```

Αφού ολοκληρωθεί η προετοιμασία των BAM αρχείων, καλείται το REDItoolDnaRna.py για να εκτελέσει τον εντοπισμό των RNA τροποποιήσεων σε αυτά. Οι εντολές για την εκτέλεση της διαδικασίας ήταν οι εξής:

- ERR188182 Alu περιοχές

```
python REDItoolDnaRna.py -i ERR188182.concordant_uniq.sorted.bam
-j NA18912.sort.rmdup.bam -f GRCh38_full_analysis_set_plus_decoy_hla.fa
-o alu_ERR188182 -t 8 -F alu -w reditools_splicesites.ss
-G GTF_ENS94_without_genes.gtf.gz -a 6-0 -m 20,20 -c 10,2 -n 0.0 -N 0.0 -v 2 -u -V
-R -e
```

- ERR188182 non-Alu περιοχές

```
python REDItoolDnaRna.py -i ERR188182.concordant_uniq.sorted.bam
-j NA18912.sort.rmdup.bam -f GRCh38_full_analysis_set_plus_decoy_hla.fa
-o non_alu_ERR188182 -t 8 -F non_alu -w reditools_splicesites.ss
-G GTF_ENS94_without_genes.gtf.gz -a 6-0 -m 20,20 -c 10,2 -n 0.1 -N 0.0 -v 3 -u -l
-r 4 -O 0,5 -V -R -e -b /gsnap_aligned_files/ERR188182BlatCorrectionRNA
-B /gsnap_aligned_files/ERR188182BlatCorrectionDNA
```

Αντίστοιχα για το ERR188298.

- ENCLB155EFP Alu περιοχές

```
python REDItoolDnaRna.py -i rnaK562_gsnap.concordant_uniq.sorted.nodup.bam
-j dnaK562_gsnap.concordant_uniq.sorted.nodup.bam
-f GRCh38_full_analysis_set_plus_decoy_hla.fa
-o alu_ ENCLB155EFP -t 8 -F alu -w reditools_splicesites.ss
-G GTF_ENS94_without_genes.gtf.gz -a 6-0 -m 20,20 -c 10,2 -n 0.0 -N 0.0 -v 2 -u -V
-R -s # -e
```

- ENCLB155EFP non-Alu περιοχές

```
python REDItoolDnaRna.py -i rnaK562_gsnap.concordant_uniq.sorted.nodup.bam
-j dnaK562_gsnap.concordant_uniq.sorted.nodup.bam
-f GRCh38_full_analysis_set_plus_decoy_hla.fa
-o non_alu_ENCLB155EFP -t 8 -F non_alu -w reditools_splicesites.ss
-G GTF_ENS94_without_genes.gtf.gz -a 6-0 -m 20,20 -c 10,2 -n 0.1 -N 0.0 -v 3 -u -l
-r 4 -O 0,5 -V -R -s # -e -b /gsnap_aligned_files/rnaK562BlatCorrectionDNA
-B /gsnap_aligned_files/dnaK562BlatCorrectionDNA
```

Από το GTF αρχείο με τους χαρακτηρισμούς της Ensembl 94 αφαιρέθηκαν οι χαρακτηρισμοί των γονιδίων, διότι δημιουργούσαν σφάλμα στο script με το οποίο αναθέτει τους χαρακτηρισμούς το REDItools. Αυτό δεν επηρέασε τη διαδικασία, καθώς ο χαρακτηρισμός των γονιδίων αναφέρεται ως μετα-δεδομένο σε κάθε χαρακτηρισμό μεταγράφων και άρα δεν παραλείπεται στο χαρακτηρισμό των τροποποιήσεων στα αποτελέσματα.

Τα αποτελέσματα δίνονται σε μορφή TAB-delimited αρχείου, συνοδευόμενο από ένα πρόσθετο αρχείο που περιέχει την εντολή εκτέλεσης, το χρόνο που κλήθηκε η εντολή, τη διάρκεια και αναλυτικά τις τιμές των παραμέτρων. Οι στήλες των αποτελεσμάτων είναι οι εξής:

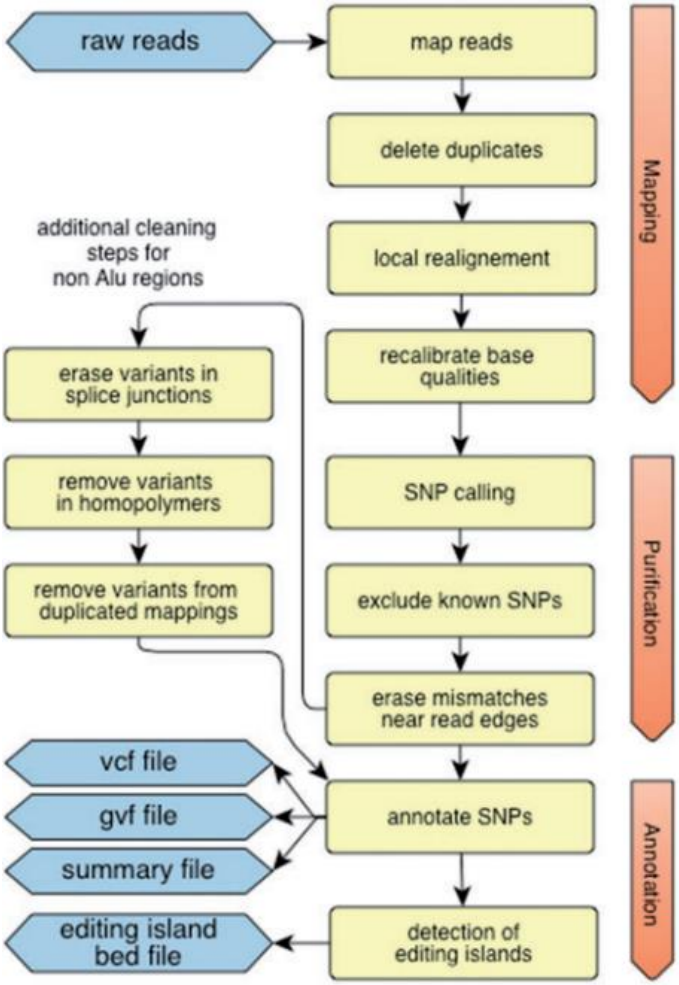
Region	χρωμόσωμα
Position	συντεταγμένη περιοχής
Reference	βάση στο γονιδίωμα αναφοράς
Strand	έλικα τροποποίησης, 1 για (+) έλικα, 0 για (-) έλικα και 2 για άγνωστη ή μη ορισμένη έλικα
Coverage-qxx	πλήθος reads με ποιότητα αλληλούχισης xx (ελάχιστη τιμή)
MeanQ	μέση τιμή ποιότητας αλληλούχισης
BaseCount[A,C,G,T]	κατανομή πλήθους νουκλεοτιδίων στη σειρά A, C, G και T
AllSubs	λίστα τροποποιήσεων, χωρισμένες με κενό. Ο χαρακτήρας “-” δηλώνει μη τροποποιημένες περιοχές.
Frequency	συχνότητα τροποποίησης. Σε περίπτωση πολλαπλών τροποποιήσεων, αναφέρεται στην πρώτη τροποποίηση.

Το REDItoolDnaRna.py σημειώνει 5 επιπλέον στήλες για την πληροφορία από DNA-Seq reads:

gCoverage-qxx	πλήθος reads με ποιότητα αλληλούχισης xx (ελάχιστη τιμή) στο DNA-Seq
gMeanQ	μέση τιμή ποιότητας αλληλούχισης στο DNA-seq
gBaseCount[A,C,G,T]	κατανομή πλήθους νουκλεοτιδίων στη σειρά A, C, G και T
gAllSubs	λίστα DNA τροποποιήσεων, χωρισμένες με κενό. Ο χαρακτήρας “-” δηλώνει μη τροποποιημένες περιοχές.
gFrequency	συχνότητα DNA τροποποίησης. Σε περίπτωση πολλαπλών τροποποιήσεων, αναφέρεται στην πρώτη τροποποίηση.

2.2.3 RNAEditor

Το RNAEditor είναι το τελευταίο εργαλείο που συμπεριλήφθηκε στο σύνολο σύγκρισης. Είναι γραμμένο στη γλώσσα προγραμματισμού Python. Εκτελείται είτε μέσω γραμμής εντολών είτε μέσω γραφικού περιβάλλοντος. Είναι κατασκευασμένο ώστε να δέχεται μόνο RNA-seq δείγματα σε fastq μορφή και να εκτελεί την ανάλυση που απεικονίζεται στο σχήμα 4.



Σχήμα 4. Σύνοψη βημάτων που εκτελεί το RNAEditor.

Από τη σύνοψη παρατηρείται ότι η προεπεξεργασία των δειγμάτων δεν είναι ενσωματωμένη στον αλγόριθμο και επομένως πρέπει να γίνει από τον χρήστη σε πρότερο χρόνο. Ο aligner που χρησιμοποιείται είναι και εδώ ο BWA σε aln mode. Η προσέγγιση που επιστρατεύεται στον εντοπισμό των RNA τροποποιήσεων είναι εντελώς διαφορετική από τα 2 προηγούμενα εργαλεία, καθώς γίνεται χρήση μιας εργαλειοθήκης αιχμής στην εύρεση μεταλλαγών στις ακολουθίες δειγμάτων, του GATK (Genomic Analysis Toolkit – GATK) [86]. Παρόλο που το GATK είναι βελτιστοποιημένο για εύρεση μεταλλαγών σε DNA δείγματα, προσφέρει τα θετικά της χρήσης ενός εδραιωμένου, στο χώρο του εντοπισμού μεταλλαγών, εργαλείου όσον αφορά τη συμπεριφορά του, το οποίο λειτουργεί ανεξαρτήτως της τεχνολογίας αλληλούχισης των δειγμάτων και εντοπίζει ακόμη και μικρής συχνότητας τροποποιήσεις (που είναι πιο δύσκολα αναγνωρίσιμες). Το μαρκάρισμα των διπλότυπων reads γίνεται με το MarkDuplicates του Picard, ενώ η τοπική αντιστοίχιση σε reads που δεν έχουν αντιστοιχηθεί αξιόπιστα και ο

επανυπολογισμός της ποιότητας αλληλούχισης των βάσεων, ώστε να λαμβάνονται υπόψιν λάθη που προκύπτουν από τις διάφορες τεχνολογίες αλληλούχισης, είναι ενσωματωμένες στο GATK. Ο εντοπισμός των τροποποιήσεων γίνεται με το UnifiedGenotyper του GATK, το οποίο κατασκευάζει ένα Bayesian μοντέλο που υπολογίζει τις πιθανότητες εμφάνισης των βάσεων στο γενότυπο. Σε αυτό το στάδιο, προεπιλεγμένα φιλτράρονται reads με ποιότητα αντιστοίχισης μικρότερη ή ίσης του 20 και βάσεις με επαν-υπολογισμένη ποιότητα αλληλούχισης μικρότερη ή ίσης του 25. Στη συνέχεια εφαρμόζονται αντίστοιχα φίλτρα με τα υπόλοιπα εργαλεία, όπως η απόρριψη τροποποιήσεων κοντά στις άκρες των reads (default: 3), η απόρριψη στο τέλος ομοπολυμερικών περιοχών με εκτεταμένες επαναλήψεις βάσεων με μήκος άνω των 4 (π.χ. AAAAAA) και η απόρριψη κοντά σε splice sites (που υπολογίζονται μέσω των συντεταγμένων των introns από το GTF αρχείο). Τα SNP αφαιρούνται βάση χρωμοσώματος και θέσης από ένα πλήθος βάσεων δεδομένων με SNP, που δίνονται στην είσοδο του script. Ο BLAT χρησιμοποιείται για το φιλτράρισμα των reads που δεν έχουν αντιστοιχηθεί μοναδικά σε κάποια περιοχή, εφόσον έχει προηγηθεί η επαν-αντιστοίχιση από το GATK. Οι διαφορετικές εκτελέσεις για Alu και non-Alu περιοχές γίνονται εσωτερικά κατά τη διάρκεια του script. Τέλος, το RNAEditor υπολογίζει και κάποιες περιοχές που ονομάζει RNA editing islands, οι οποίες είναι περιοχές με πολυπληθείς ομάδες RNA τροποποιήσεων, μέσω του clustering αλγορίθμου DBSCAN. Η εξαγωγή πληροφορίας για ταυτοποίηση της έλικας προέλευσης των διαβασμάτων και τροποποιήσεων σε strand-specific δείγματα δεν έχει υλοποιηθεί.

Οι παράμετροι που δέχεται το RNAEditor σε μορφή configuration αρχείου είναι οι παρακάτω.

refGenome	γονιδίωμα αναφοράς
gtfFile	αρχείο χαρακτηρισμών σε GTF μορφή
dbSNP	αρχείο με SNP (dbSNP) σε VCF μορφή
hapmap	αρχείο με SNP (HapMap) σε VCF μορφή
omni	αρχείο με SNP (Omni) σε VCF μορφή
esp	αρχείο με SNP (ESP) σε VCF μορφή
dna	αρχείο με SNP (DNA-seq match) σε VCF μορφή
aluRegions	αρχείο με repeat χαρακτηρισμούς σε BED μορφή
output	μονοπάτι συστήματος για αποθήκευση αποτελεσμάτων
sourceDir	μονοπάτι συστήματος με τα εκτελέσιμα πακέτα που χρειάζεται το RNAEditor
maxDiff	ποσοστό μέγιστου ορίου μη αντιστοιχίας στο BWA alignment
seedDiff	μέγιστο πλήθος μη αντιστοιχίας βάσεων στη seed ακολουθία του BWA
standCall και standEmit	κατώφλια ποιότητας αλληλούχισης των βάσεων που θα φιλτράρει το GATK στα αποτελέσματά του

edgeDistance	αριθμός βάσεων που θα απορριφθούν στις άκρες των reads
paired	true ή false τιμή για το αν τα δείγματα είναι paired-end
keepTemp	true ή false τιμή για το αν θα διατηρηθούν τα ενδιάμεσα αρχεία της ανάλυσης
overwrite	true ή false τιμή για το αν πρέπει ο αλγόριθμος να ξαναγράψει πάνω από ήδη υπάρχοντα αρχεία
threads	αριθμός παράλληλων threads

Τα αποτελέσματα είναι σε μορφή Variant Call Format (VCF) και εμφανίζουν τις εξής TAB-delimited στήλες:

CHROM	χρωμόσωμα
POS ID	θέση
REF	νουκλεοτίδιο στο γονιδίωμα αναφοράς
ALT	νουκλεοτίδιο τροποποίησης
QUAL	ποιότητα αλληλούχισης
FILTER	ένδειξη περάσματος βήματος φιλτραρίσματος
INFO	μετα-δεδομένα (χαρακτηρισμός, πλήθος reads κ.ά.)

Επίσης, παράγονται γραφήματα με αριθμητικά στατιστικά (barplots), με τις συχνότητες εμφάνισης τροποποιήσεων συνολικά ή ανά χαρακτηρισμό περιοχής (5'UTR, 3'UTR, CDS, intron), καθώς και αρχεία με τα πρώτα 20 γονίδια ανά χαρακτηρισμό σε αριθμό τροποποιήσεων.

Στο πλαίσιο της διπλωματικής υλοποιήθηκαν κάποιες πρόσθετες λειτουργίες για το RNAEditor, τις οποίες έχουν ήδη υλοποιημένες τα προηγούμενα εργαλεία, ώστε να εκτιμηθεί η επίδρασή τους στα αποτελέσματα και να γίνει σύγκριση επί ίσοις όροις. Για το λόγο αυτό, στο στάδιο που φιλτράρονται οι υποψήφιες RNA τροποποιήσεις, προστέθηκε η λειτουργία φιλτραρίσματος με τη χρήση της διωνυμικής κατανομής και το FDR correction, με τον ίδιο τρόπο που υλοποιείται και στα άλλα εργαλεία (ίδια είσοδος και τιμές κατωφλίων). Επίσης, προστέθηκε η δυνατότητα για χρήση συμπληρωματικού VCF αρχείου με SNP από το αντίστοιχο δείγμα DNA, με τον τρόπο που δίνονται και οι υπόλοιπες βάσεις. Επομένως, ο χρήστης μπορεί να τρέξει το GATK για το DNA δείγμα που εξετάζει και να αντλήσει πρόσθετη πληροφορία για το RNA δείγμα του, εξειδικεύοντας το φιλτράρισμα των SNP για μεγαλύτερο ποσοστό ακρίβειας. Τέλος, όπως προτείνεται και από τους συγγραφείς του RNAEditor, ο αλγόριθμος αναγνωρίζει το βήμα της ανάλυσης στο οποίο βρίσκεται μέσω των ονομάτων των αρχείων των δειγμάτων. Επομένως δεν έχει γίνει υλοποίηση για την απευθείας είσοδο prealigned BAM αρχείων, όμως δύναται με τη μετονομασία των αρχείων εισόδου να εκκινηθεί η διαδικασία από το σημείο του εντοπισμού των RNA τροποποιήσεων, αρκεί να είναι συμβατά με την έκδοση του GATK που χρησιμοποιείται.

Το RNAEditor εκτελέστηκε 2 φορές, μία με τον BWA ως aligner και μία με το GSNAP, ώστε να εκτιμηθεί η διαφορά στα αποτελέσματα από την επιλογή του aligner. Για τον BWA διεξάχθηκε η παρακάτω διαδικασία.

ERR188182 με BWA

- Προεπεξεργασία δειγμάτων

```
cutadapt -u 6 -q 25 -m 1 -o ERR188182_QC_1.fastq -p ERR188182_QC_2.fastq
ERR188182_1.fastq ERR188182_2.fastq
```

- Δημιουργία DNA VCF αρχείου με SNP

```
# Picard marking
java -Xmx16G -jar /miniconda3/bin/picard-tools/MarkDuplicates.jar
INPUT=NA18912.bam OUTPUT=NA18912.dedup_reads.bam
METRICS_FILE=NA18912.pcr.metrics VALIDATION_STRINGENCY=LENIENT
CREATE_INDEX=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
MAX_RECORDS_IN_RAM=1000000 TMP_DIR=/home/marios/tmp

# The file is indexed, so it's sorted

# RealignerTargetCreator
# No dbSNP given for known indels, as the RNAEditor authors didn't give either
for the RNA sample (the BAM header of the sample says that this step has been
executed anyway with a known indels file)
java -Xmx16G -jar /miniconda3/bin/GATK/GenomeAnalysisTK.jar -nt 8 -T
RealignerTargetCreator -R GRCh38_full_analysis_set_plus_decoy_hla.fa
-I NA18912.dedup_reads.bam -o NA18912.indels.intervals -l ERROR

java -Xmx16G -jar /miniconda3/bin/GATK/GenomeAnalysisTK.jar
-T IndelRealigner -R GRCh38_full_analysis_set_plus_decoy_hla.fa
-I NA18912.dedup_reads.bam -targetIntervals NA18912.indels.intervals
-o NA18912.noDup.realigned.bam

java -Xmx16G -jar /miniconda3/bin/GATK/GenomeAnalysisTK.jar -nct 8 -T
BaseRecalibrator -R GRCh38_full_analysis_set_plus_decoy_hla.fa
-knownSites dbSNP.vcf -I NA18912.noDup.realigned.bam -cov CycleCovariate
-cov ContextCovariate -o NA18912.noDup.realigned.recal_data.table

java -Xmx16G -jar /miniconda3/bin/GATK/GenomeAnalysisTK.jar -nct 8 -T
PrintReads -R GRCh38_full_analysis_set_plus_decoy_hla.fa
-I NA18912.noDup.realigned.bam -BQSR NA18912.noDup.realigned.recal_data.table
-o NA18912.noDup.realigned.recal_reads.bam

java -Xmx16G -jar /miniconda3/bin/GATK/GenomeAnalysisTK.jar -nt 8 -T
UnifiedGenotyper -R GRCh38_full_analysis_set_plus_decoy_hla.fa -glm SNP
-I NA18912.noDup.realigned.recal_reads.bam -D dbSNP.vcf -o NA18912.snp.vcf
-metrics NA18912.snp.metrics -stand_call_conf 25 -stand_emit_conf  25
-A Coverage -A AlleleBalance -A BaseCounts
```

- Configuration αρχείο

```
refGenome       GRCh38_full_analysis_set_plus_decoy_hla.fa
gtfFile         GTF_ENS94_without_genes.gtf
dbSNP           dbSNP.vcf
hapmap          HAPMAP.vcf
omni            1000G_omni2.5.hg38.vcf
esp             ESP.vcf
dna             NA18912.snp.vcf
aluRegions      rmsk.bed
output          /RNAeditor_ERR188182_BWA/ERR188182_BWA
```

```
sourceDir       /miniconda3/bin/
maxDiff         0.04
seedDiff        2
standCall       0
standEmit       0
edgeDistance    3
paired          True
keepTemp        True
overwrite       False
threads         1
```

- Εντολή εκτέλεσης

```
python /RNAEditor/RNAEditor.py -i ERR188182_QC_1.fastq ERR188182_QC_2.fastq
-c /RNAEditor/ERR188182_configuration.txt
```

Αντίστοιχα για ERR188298 και ENCLB155EFP με BWA.

ERR188182 με GSNAP

Για την εκτέλεση του RNAEditor με τον aligner GNSNAP χρησιμοποιήθηκε η ίδια διαδικασία που χρησιμοποιήθηκε κατά την εκτέλεση του REDItools, εφόσον έγινε ποιοτικός έλεγχος στα fastq αρχεία. Η εκτέλεση του RNAEditor, δοσμένου BAM αρχείου, ξεκινάει από τη διαδικασία του εντοπισμού με τον UnifiedGenotyper. Οι εντολές για τα προηγούμενα βήματα προεπεξεργασίας παρουσιάζονται ακολούθως.

- Ποιοτικός έλεγχος

```
cutadapt -u 6 -q 25 -m 1 -o ERR188182_QC_1.fastq -p ERR188182_QC_2.fastq
ERR188182_1.fastq ERR188182_2.fastq
```

- GSNAP alignment (ακολούθως όπως στην εκτέλεση του REDItools)

- Προεπεξεργασία δειγμάτων

```
samtools sort -@ 12 -o ENCLB155EFP_GSNAP.sorted.bam ENCLB155EFP_GSNAP.qc.bam

java -Xmx16G -jar /miniconda3/bin/picard-tools/MarkDuplicates.jar
INPUT=ERR188182_QC_GSNAP.sorted.bam OUTPUT=ERR188182_QC_GSNAP.sorted.noDup.bam
METRICS_FILE= ERR188182_QC_GSNAP.metrics.txt REMOVE_DUPLICATES=false
ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
MAX_RECORDS_IN_RAM=1000000 CREATE_INDEX=true TMP_DIR=/reditools/gsnap_aligned_files

# Skip the realignment part, because GSNAP is splice-aware, do not compromise
integrity
java -Xmx16G -jar /miniconda3/bin/picard-tools/AddOrReplaceReadGroups.jar
I=ERR188182_QC_GSNAP.sorted.noDup.bam O=ERR188182_QC_GSNAP.sorted.noDup.fixedRG.bam
RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=K562_kd CREATE_INDEX=TRUE

java -Xmx16G -jar /miniconda3/bin/picard-tools/ReorderSam.jar
I=ERR188182_QC_GSNAP.sorted.noDup.fixedRG.bam O=ERR188182_QC_GSNAP.reordered.bam
R=/RNAEditor/GRCh38_full_analysis_set_plus_decoy_hla.fa CREATE_INDEX=TRUE
ALLOW_INCOMPLETE_DICT_CONCORDANCE=TRUE
```

- Δημιουργία DNA VCF αρχείου με SNP

- Configuration αρχείο

```
refGenome       GRCh38_full_analysis_set_plus_decoy_hla.fa
gtfFile         GTF_ENS94_without_genes.gtf
dbSNP           dbSNP.vcf
hapmap          HAPMAP.vcf
omni            1000G_omni2.5.hg38.vcf
esp             ESP.vcf
dna             NA18912.snp.vcf
aluRegions      rmsk.bed
output          /RNAeditor_ERR188182_GSNAP/ERR188182_QC_GSNAP.reordered.bam
sourceDir       /miniconda3/bin/
maxDiff         0.04
seedDiff        2
standCall       0
standEmit       0
edgeDistance    3
paired          True
keepTemp        True
overwrite       False
threads         1
```

- Εντολή εκτέλεσης

```
python /RNAEditor/RNAEditor.py -i ERR188182_QC_1.fastq ERR188182_QC_2.fastq
-c /RNAEditor/ERR188182_configuration.txt
```

Αντίστοιχα για ERR188298 και ENCLB155EFP με GSNAP.

2.3 Επεξεργασία αποτελεσμάτων

Στο επόμενο βήμα, έγινε προεπεξεργασία των αποτελεσμάτων από τις εκτελέσεις των εργαλείων και διεξάχθηκε στατιστική ανάλυση ώστε να γίνει η επιλογή του βέλτιστου αλγορίθμου με κριτήρια το πλήθος συμβάντων τροποποίησης και την παρατηρούμενη ακρίβεια.

2.3.1 Χαρακτηρισμός αποτελεσμάτων και διόρθωση έλικας προέλευσης

Και τα 3 εργαλεία προσφέρουν ως παράμετρο την επιλογή να πραγματοποιηθεί χαρακτηρισμός κάθε RNA τροποποίησης χρησιμοποιώντας τα GTF και repeat αρχεία εισόδου. Στο REDItools έχει υλοποιηθεί ξεχωριστό Python script που δέχεται ένα TAB-delimited αρχείο στη μορφή των αποτελεσμάτων του REDItools και προσθέτει τους χαρακτηρισμούς. Εξαιτίας της καθοριστικής σημασίας, από βιολογική σκοπιά, της έλικας προέλευσης των RNA τροποποιήσεων (π.χ. τα φαινόμενα A-σε-I τροποποιήσεων εμφανίζουν πολύ υψηλή συχνότητα στα θηλαστικά, δε συμβαίνει όμως το ίδιο για τις τροποποιήσεις G-σε-A), τροποποιήθηκε ο κώδικας του script ώστε κατά τη διάρκεια της εκτέλεσης να λαμβάνονται υπόψιν οι εξής 2 περιπτώσεις (εικόνα 6):

1. Εντοπίστηκε μοναδικός χαρακτηρισμός για την παρουσία γονιδίου ή επαναλαμβανόμενου μεταγραφόμενου στοιχείου στη μία μόνο έλικα, χωρίς επικάλυψη άλλου χαρακτηρισμού στην απέναντι πλευρά. Στην περίπτωση αυτή, θεωρείται ότι η τροποποίηση προήλθε από την έλικα του χαρακτηρισμού, εφόσον μόνο αυτή παράγει γνωστά μετάγραφα που δύνανται να τροποποιούνται.

2. Εντοπίστηκαν χαρακτηρισμοί στην περιοχή της RNA τροποποίησης που βρίσκονται σε αντίθετες έλικες και επικαλύπτονται στο σημείο τροποποίησης. Στην περίπτωση αυτή, δε δύναται η επιλογή κάποιας από τις 2 έλικες χωρίς τη βοήθεια κάποιου strand-specific πρωτοκόλλου αλληλούχισης και άρα η τροποποίηση απορρίπτεται, θυσιάζοντας ευαισθησία για μεγαλύτερη ακρίβεια.

Εικόνα 6. Περιπτώσεις χαρακτηρισμών για εξαγωγή πληροφορίας για την έλικα προέλευσης.

Δημιουργήθηκαν 3 εκδόσεις του REDItools AnnotateTable.py, ώστε να είναι συμβατό με τη δομή των αποτελεσμάτων κάθε εργαλείου. Επειδή είναι αναγκαίες 2 ξεχωριστές εκτελέσεις, μία χρησιμοποιώντας τους χαρακτηρισμούς του GTF και μία του repeat, δημιουργήθηκαν wrapper shell scripts. Η διόρθωση της έλικας προέλευσης εφαρμόστηκε στα non strand-specific δείγματα, δηλαδή τα ERR188182 και ERR188298. Οι εντολές εκτέλεσης για το ERR188182 και το RES-Scanner ήταν οι παρακάτω, ενώ αντίστοιχες εντολές εκτελέστηκαν και για τα υπόλοιπα δείγματα και εργαλεία.

- ERR188182 Alu περιοχές

```
sh RES_Scanner_scripts/scanner_output_table_annotation.sh -t alu
-i RES_Scanner_ERR188182_BWA.alu
```

- ERR188182 non-Alu περιοχές

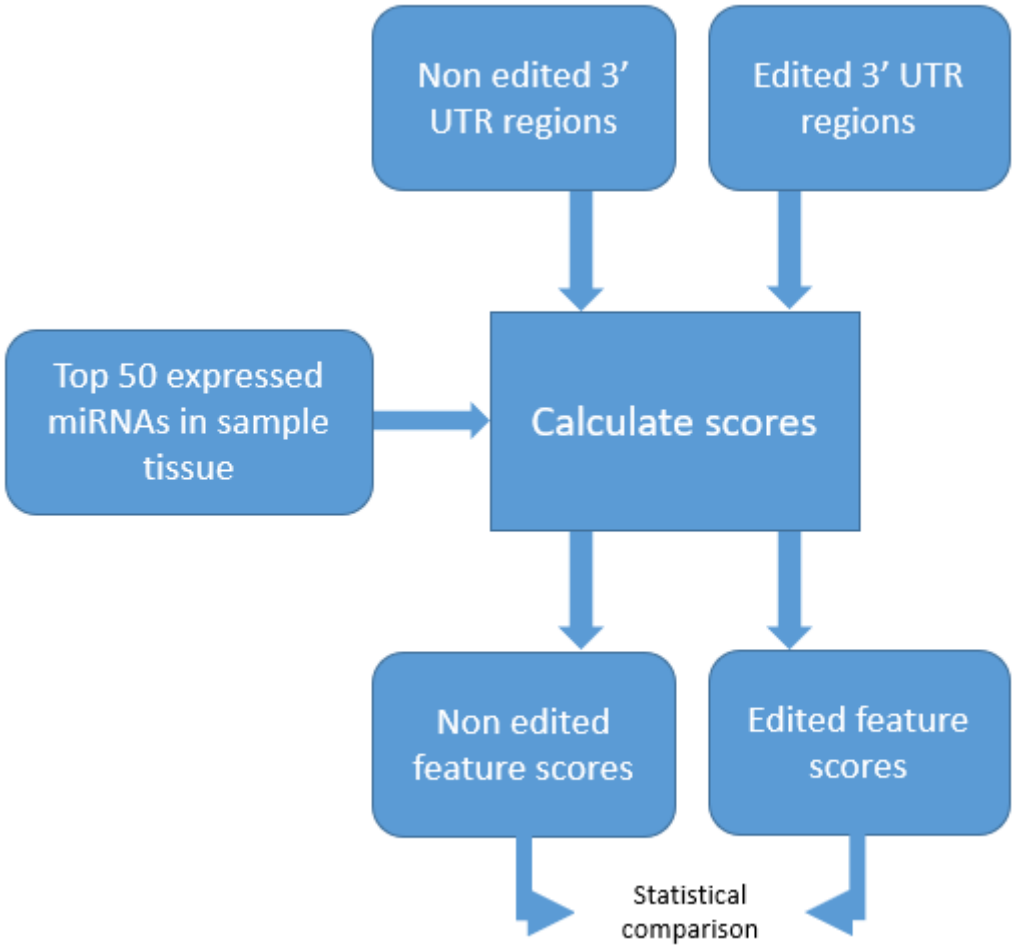
```
sh RES_Scanner_scripts/scanner_output_table_annotation.sh -t non_alu
-i RES_Scanner_ERR188182_BWA.non_alu
```

2.3.2 Στατιστική ανάλυση

Για τη σύγκριση των εργαλείων, λήφθηκε υπόψιν η συνολική τους απόδοση σε μια σειρά από κριτήρια. Για την καλύτερη εκτίμηση των διαφορών, τα κριτήρια απεικονίστηκαν σε διαγράμματα με τη χρήση της R.
Η συχνότητα των διαφόρων RNA τροποποιήσεων που εντοπίστηκαν από τα εργαλεία για κάθε δείγμα, απεικονίστηκε σε ιστογράμματα, όπου τα A-σε-G φαινόμενα (υπενθυμίζεται ότι η ινοσίνη μοιάζει με γουανοσίνη και διαβάζεται ως G από το μηχάνημα αλληλούχισης) αναμένονται να παρουσιάζουν μεγαλύτερο πλήθος, ενώ τα υπόλοιπα, εκτός της C-σε-T, αναμένονται να είναι κοντά στο 0. Η σχέση μεταξύ τους (τομή μεταξύ των διαφορετικών συνδυασμών των εργαλείων) αποτυπώθηκε με τετραπλά διαγράμματα Venn, ώστε να εξαχθούν συμπεράσματα σχετικά με την ομοιογένεια της συμπεριφοράς των εργαλείων. Επίσης, σχεδιάστηκε σε καρτεσιανό σύστημα η συνάρτηση μεταξύ του αριθμού των εντοπισμένων τροποποιήσεων και της RADAR, ώστε να υπολογιστεί το ποσοστό των αποτελεσμάτων που είναι καταχωρημένα σε αυτή. Σε όλα αυτά, αναμένεται ο αριθμός των εντοπισμών να είναι αισθητά μειωμένος στο δείγμα ENCLB155EFP, που έχει πραγματοποιηθεί αποσιώπηση των ADAR.

2.4 Πρόβλεψη στόχων miRNA

Το δεύτερο βήμα της εργασίας αφορούσε την ανάλυση των στόχων των miRNA στις περιοχές με A-σε-I τροποποιήσεις. Ο κορμός της ανάλυσης απεικονίζεται στο σχήμα 5. Από τα αποτελέσματα που προέκυψαν στο πρώτο βήμα της εργασίας επιλέχθηκαν εκείνα που βρίσκονταν σε 3'UTR στοιχεία. Η πληροφορία αυτή συνδυάστηκε με τις ακολουθίες των 3'UTR ώστε να δημιουργηθούν δύο σύνολα από τις ίδιες ακολουθίες, με την κανονική και την τροποποιημένη μορφή τους. Στη συνέχεια, κάθε σύνολο δόθηκε ως είσοδος στους αλγορίθμους πρόβλεψης, μαζί με τα πρώτα 50 σε έκφραση miRNA για τους ιστούς των δειγμάτων, ώστε να υπολογιστούν τα score αλληλεπίδρασης ανά ζευγάρι mRNA-miRNA. Τέλος, έγινε σύγκριση των αποτελεσμάτων των δύο συνόλων για κάθε δείγμα, ώστε να εκτιμηθούν στατιστικά οι διαφορές μεταξύ τους, σχετικά με τα χαρακτηριστικά της πρόσδεσης mRNA-miRNA.



Σχήμα 5. Σύνοψη βημάτων πρόβλεψης και ανάλυσης στόχων των miRNA στις περιοχές με A-σε-I τροποποιήσεις.

Για την πρόβλεψη των στόχων των miRNA χρησιμοποιήθηκαν 2 αλγόριθμοι που ακολουθούν παρόμοια στρατηγική με κοινά χαρακτηριστικά μεταξύ τους, ώστε να εξασφαλιστεί η εγκυρότητα των αποτελεσμάτων, τα TargetScan (έκδοση 7.2) και MIRZA-G (έκδοση 1.0). Η πλήρης περιγραφή των μοντέλων που επιστρατεύονται από τους αλγορίθμους είναι εκτός του πλαισίου της παρούσας διπλωματικής εργασίας, όμως περισσότερες πληροφορίες μπορούν να αναζητηθούν στις αντίστοιχες δημοσιεύσεις

(TargetScan [65], MIRZA-G [66]). Συνοπτικά, δοθέντων ακολουθιών 3’UTR και miRNA, γίνεται αναζήτηση των seed region των miRNA στις ακολουθίες ενδιαφέροντος. Βάση της αντιστοίχισης, η οποία εξαρτάται από τις θέσεις συμπληρωματικότητας μεταξύ της seed ακολουθίας και του mRNA οι πιθανές θέσεις πρόσδεσης του miRNA χαρακτηρίζονται ως προς την ισχύ τους με σειρά κατάταξης ισχύος 8mer (2-8 θέσεις στο seed, 1 βάση A) > 7mer-m8 (2-8 θέσεις) > 7mer-A1 (2-7 θέσεις, 1 βάση A) > 6mer. Συμπληρωματικά υπολογίζονται άλλα χαρακτηριστικά που αφορούν τη σύσταση και τη δομή της miRNA ακολουθίας και της ακολουθίας γύρω από την περιοχή πρόσδεσης. Ακολούθως, από το γραμμικό συνδυασμό των χαρακτηριστικών με βάρη, υπολογίζεται ένα τελικό score που αντικατοπτρίζει την ισχύ της προβλεπόμενης θέσης πρόσδεσης ως προς την κατασταλτική της δράση.

Το TargetScan χρησιμοποιεί ως τελικό score το context++ score, το οποίο είναι ο συνδυασμός των παρακάτω 14 χαρακτηριστικών:

- site type — Ο τύπος της περιοχής πρόσδεσης (8mer, 7mer κτλ.).
- 3’pairing contribution — Αντιπροσωπεύει την συνεχόμενη συμπληρωματικότητα βάσεων εκτός της seed περιοχής, η οποία βελτιώνει την αποτελεσματικότητα της πρόσδεσης.
- local AU — Περιεχόμενο της περιβάλλουσας ακολουθίας βάσεων έως 30 νουκλεοτίδια από την περιοχή πρόσδεσης σε αδενίνες και ουρακίλες. Μεγαλύτερο ποσοστό αυτών των βάσεων οδηγεί σε πιο αποτελεσματικές περιοχές πρόσδεσης.
- minimum distance — Ελάχιστη απόσταση από τα άκρα της 3’UTR. Ακραίες περιοχές προσδίδουν αποτελεσματικότερες περιοχές πρόσδεσης
- sRNA1A, sRNA1C, sRNA1G, sRNA8A, sRNA8C, sRNA8G — Η ύπαρξη της βάσης που αναγράφεται στο χαρακτηριστικό, στην αντίστοιχη θέση (sRNA#νουκλεοτίδιο). Η ύπαρξη των συγκεκριμένων βάσεων σε αυτές τις θέσεις, σε συγκριμένες ακολουθίες περιοχών πρόσδεσης, μπορεί να αυξήσουν ή να μειώσουν την αποτελεσματικότητα της περιοχής κατά περίπτωση.
- site8A, site8C, site8G — Το παραπάνω ισχύει και για αυτά τα χαρακτηριστικά.
- 3' UTR length — Το μήκος της 3’UTR. Μακρύτερες περιοχές πρόσδεσης προσδίδουν λιγότερο κατασταλτικές περιοχές πρόσδεσης.
- SA (structural accessibility) — Η προσβασιμότητα της δευτεροταγούς δομής της περιοχής πρόσδεσης, ώστε να

calculated with RNAplfold [87]	προσδεθεί το miRNA. Περισσότερο προσβάσιμες περιοχές προσδίδουν μεγαλύτερη κατασταλτική ικανότητα στις περιοχές πρόσδεσης.
• ORF length	Μήκος open reading frame (ORF) περιοχής μέσα στη 3'UTR. Μεγαλύτερες ORF περιοχές προσδίδουν λιγότερο κατασταλτικές περιοχές πρόσδεσης.
• ORF 8mer count	Η ύπαρξη 8μερών περιοχών πρόσδεσης σε ORF προσδίδει περισσότερο αποτελεσματικές περιοχές πρόσδεσης.
• Offset 6mer count	Η ύπαρξη συνεχούς συμπληρωματικού δεσμού στις θέσεις 3-8 της seed περιοχής δημιουργεί περιοχές πρόσδεσης που επάγουν την καταστολή του mRNA.
• TA (target site abundance)	Το πλήθος των στόχων μιας οικογένειας miRNA στο σύνολο των 3'UTR. Μεγαλύτερο πλήθος προσδίδει λιγότερο αποτελεσματική κατασταλτική δράση στα miRNA της οικογένειας.
• SPS (seed-pairing stability)	Η σταθερότητα του δεσμού miRNA-mRNA στην περιοχή πρόσδεσης. Λιγότερο σταθεροί δεσμοί μειώνουν την κατασταλτική ικανότητα του miRNA.
• PCT (probability of conserved targeting)	Η εξελικτική συντήρηση της περιοχής πρόσδεσης. Καλά συντηρημένες περιοχές πρόσδεσης προσδίδουν καλύτερες προβλέψεις και τείνουν να είναι πιο αποτελεσματικές.

Όλα τα χαρακτηριστικά είναι τροποποιημένα ώστε πιο αρνητικές τιμές να εκφράζουν αποτελεσματικότερες περιοχές πρόσδεσης.

Επίσης, ο αλγόριθμος του TargetScan υπολογίζει, εκτός από το context++ score, και το weighted context++ score. Η εξίσωση του υπολογισμού του περιλαμβάνει το Affected Isoform Ratio (AIR), το οποίο ορίζεται ως το ποσοστό των ισομορφών ενός μεταγράφου που περιέχουν την περιοχή πρόσδεσης προς το σύνολο των ισομορφών του που υπάρχουν στο δείγμα. Λαμβάνοντας αυτό υπόψιν, το weighted context++ score αντιπροσωπεύει τη συνολική καταστολή που αναμένεται να υποστεί η έκφραση ενός γονιδίου λόγω της συνολικής καταστολής των mRNA του, συνυπολογίζοντας την ύπαρξη πολλαπλών, κοντινών περιοχών πρόσδεσης ανά miRNA.

Τέλος, στα αποτελέσματα δίνονται και τα percentiles του κάθε score, τα οποία ορίζονται ως το ποσοστό των περιοχών πρόσδεσης ως προς μία περιοχή ενός miRNA, που έχουν χειρότερο score από αυτή.

Αντίστοιχα, το MIRZA-G συνδυάζει τα ακόλουθα χαρακτηριστικά:

• Distance to boundary	Η απόσταση από τα άκρα της 3'UTR. Ακραίες περιοχές προσδίνουν αποτελεσματικότερες περιοχές πρόσδεσης.

• Accessibility calculated with CONTRAfold [88]	Η προσβασιμότητα της δευτεροταγούς δομής της περιοχής πρόσδεσης, ώστε να προσδεθεί το miRNA. Η τιμή συμβολίζει την πιθανότητα η δομή να είναι μονόκλωνη, τροποποιημένη με τη συνάρτηση log. Περισσότερο προσβάσιμες περιοχές προσδίδουν μεγαλύτερη κατασταλτική ικανότητα στις περιοχές πρόσδεσης.
• Flanks G • Flanks U	Περιεχόμενο της περιβάλλουσας ακολουθίας βάσεων έως 50 νουκλεοτίδια από την περιοχή πρόσδεσης σε γουανίνες και ουρακίλες. Μεγαλύτερο ποσοστό αυτών των βάσεων οδηγεί σε πιο αποτελεσματικές περιοχές πρόσδεσης.
• Branch length score (BLS, conservation score)	Η εξελικτική συντήρηση της περιοχής πρόσδεσης. Καλά συντηρημένες περιοχές πρόσδεσης προσδίδουν καλύτερες προβλέψεις και τείνουν να είναι πιο αποτελεσματικές.
• MIRZAscore (MIRZA target quality score) [89]	Scoring μοντέλο που υλοποιείται από τον αλγόριθμο του MIRZA που χρησιμοποιεί βιοφυσικά χαρακτηριστικά για την πρόβλεψη υποψήφιων στόχων miRNA.

Για όλα τα χαρακτηριστικά, θετικότερες τιμές υποδεικνύουν μεγαλύτερη αποτελεσματικότητα της περιοχής πρόσδεσης. Τα επιμέρους χαρακτηριστικά συνδυάζονται συνολικά στα χαρακτηριστικά probability without conservation και probability with conservation, τα οποία υπολογίζουν την πιθανότητα η περιοχή να είναι πραγματικά λειτουργική ως περιοχή πρόσδεσης, με ή χωρίς τα εξελικτικά κριτήρια αντιστοίχως. Ο συνδυασμός γίνεται με τη χρήση εσωτερικού γινομένου μεταξύ των χαρακτηριστικών και τη μετατροπή του αποτελέσματος x στο διάστημα [0,1] με τη χρήση της αντίστροφης συνάρτησης logit

$$\frac{scaling\ factor \cdot e^{x}}{scaling\ factor \cdot e^{x} + 1}$$

όπου ο scaling factor είναι σταθερά κανονικοποίησης (0,24 από προεπιλογή).

Σύμφωνα και με τους 2 αλγορίθμους, το χαρακτηριστικό που προσδίδει τη μεγαλύτερη ακρίβεια στην πρόβλεψη των στόχων των miRNA είναι η εξελικτική συντήρηση (conservation score). Υπενθυμίζεται ότι τα miRNA είναι καλά συντηρημένα μεταξύ των διαφόρων ειδών, επομένως η ύπαρξη μιας θέσης στις ακολουθίες πολλαπλών ειδών αποτελεί ένδειξη σωστής πρόβλεψης. Ο τρόπος με τον οποίο υπολογίζουν αυτήν την πληροφορία οι 2 αλγόριθμοι είναι μέσω της πολλαπλής αντιστοίχισης ακολουθιών μιας περιοχής μεταξύ πολλών ειδών, η οποία καθορίζει και το τελικό score διατήρησης. Το score συντήρησης βοηθάει στον έμμεσο υπολογισμό μοριακών χαρακτηριστικών ή των συνδυασμών τους, που μέχρι τώρα δεν έχουν ενσωματωθεί στα υπολογιστικά μοντέλα πρόβλεψης και που ευθύνονται για τη συντήρηση των περιοχών πρόσδεσης στην πορεία της εξέλιξης [65].

Είναι σαφές ότι η δημιουργία ενός τροποποιημένου 3'UTR του *Homo sapiens* θα δημιουργούσε ασυμβατότητες στην αντιστοίχιση της ακολουθίας του με αυτές των άλλων ειδών και θα μείωνε το score συντήρησης, χωρίς όμως αυτό να περιγράφει το πραγματικό βιολογικό φαινόμενο. Επομένως, οι 2 αλγόριθμοι εκτελέστηκαν χωρίς τη χρήση των εξελικτικών κριτηρίων.

Ακολουθεί η περιγραφή των βημάτων που ακολουθήθηκαν για την ανάλυση, με τη σειρά που εκτελέστηκαν κατά την εργασία.

2.4.1 Συλλογή συμπληρωματικών δεδομένων και χαρακτηρισμός περιοχών τροποποίησης

Η βασική είσοδος των αλγορίθμων είναι οι ακολουθίες των 3'UTR και οι ακολουθίες των miRNA. Το TargetScan προσφέρει δικά του συμπληρωματικά αρχεία με προεπεξεργασμένες 3'UTR ακολουθίες, οι οποίες έχουν επιλεχθεί ως οι εκπρόσωποι του εκάστοτε μεταγράφου. Οι ακολουθίες-εκπρόσωποι έχουν επεκταθεί κατάλληλα για μεγαλύτερη απόδοση του αλγορίθμου και έχουν σημειωθεί σε αυτές τα Open Reading Frames, κωδικές περιοχές, στις οποίες δεν αναζητούνται θέσεις πρόσδεσης miRNA. Το MIRZA-G δεν πρότεινε συγκεκριμένα αρχεία εισόδου, επομένως για λόγους συμβατότητας των αποτελεσμάτων, επιλέχθηκε η χρήση των αρχείων του TargetScan και για τους 2 αλγορίθμους.

Η προετοιμασία των αρχείων εισόδου περιλάμβανε τα ακόλουθα βήματα:

1. Τα συμπληρωματικά αρχεία του TargetScan χρησιμοποιούν την έκδοση hg19 του ανθρώπινου γονιδιώματος. Για το λόγο αυτό, χρησιμοποιήθηκε το εργαλείο LiftOver [79], ώστε να μετατραπούν από την έκδοση hg38 οι συντεταγμένες των A-σε-I τροποποιήσεων που προέκυψαν από το πρώτο βήμα του εντοπισμού. Οι ανεπιτυχείς μετατροπές αφαιρέθηκαν από τα αρχεία.
2. Προστέθηκαν οι χαρακτηρισμοί των 3'UTR χρησιμοποιώντας τη GENCODE v19 [90] από τη UCSC, η οποία χρησιμοποιείται και από το TargetScan, ώστε να φιλτραριστούν οι περιοχές μη ενδιαφέροντος.
3. Εξάχθηκαν οι 3'UTR από το συμπληρωματικό αρχείο του TargetScan στις οποίες είχε εντοπιστεί τροποποίηση, αντιστοιχίζοντας τα ID των μεταγράφων.
4. Δημιουργήθηκε Python script που δέχεται ως είσοδο ακολουθίες και περιοχές τροποποίησης και δημιουργεί ένα νέο αρχείο με τις τροποποιημένες εκδόσεις των ακολουθιών. Εσωτερικά του script γίνεται έλεγχος ύπαρξης αδενίνης στο σημείο της τροποποίησης, την οποία πρόκειται να αλλάξει σε γουανίνη. Σε αντίθετη περίπτωση, ο χρήστης ενημερώνεται με προειδοποιητικό μήνυμα και η ακολουθία αφαιρείται.

Τα ID επιλέχθηκαν να έχουν τη μορφή Ensembl_transcript_ID@Ensembl_gene_ID. Η μετατροπή εφαρμόστηκε σε όλα τα συμπληρωματικά αρχεία του TargetScan (UTR, ORF, AIR).

Για κάθε δείγμα, επιλέχθηκαν τα 50 πρώτα miRNA σε έκφραση για την εκτέλεση των αλγορίθμων. Αφενός, δεν υπάρχει νόημα στη μελέτη miRNA τα οποία δεν εκφράζονται καθόλου σε έναν ιστό, αφετέρου η χρήση miRNA που δεν έχουν βιολογική σημασία προσθέτει θόρυβο στον αλγόριθμο, με αποτέλεσμα να αντιστοιχίζονται miRNA σε σημεία που στην πραγματικότητα προσδένονται άλλα. Για τα ERR188182 και ERR188298, χρησιμοποιήθηκαν τα αντίστοιχα small RNA-seq δείγματα (ERR187902 και ERR187891),

στα οποία έγινε προεπεξεργασία (αναζήτηση για adapters, ποιοτικός έλεγχος) με μικτή χρήση των εργαλείων FastQC, Cutadapt και Minion και ποσοτικοποίηση της έκφρασης των μεταγράφων με το miRDeep2 που αξιοποιεί τον aligner αναφοράς για μικρού μεγέθους reads, bowtie [91]. Από την ανάλυση του miRDeep2 εξάχθηκαν τα πρώτα 50 σε αφθονία microRNA [92]. Για τα δείγματα της κυτταρικής σειράς K562 δεν υπήρχαν αντίστοιχα small RNA-seq πειράματα. Για το λόγο αυτό, χρησιμοποιήθηκαν έτοιμα αρχεία με ποσοτικοποιημένες εκφράσεις των miRNA, που συλλέχθηκαν από το πείραμα με ID GSE78037 [93] από τη βάση δεδομένων Gene Expression Omnibus (GEO) του NCBI (National Center for Biotechnology Information) [94]. Η τεχνική που έχει χρησιμοποιηθεί για την ανίχνευση των μορίων στο πείραμα είναι microarray. Το πείραμα περιλαμβάνει 6 δείγματα, 3 σε φυσιολογική κατάσταση και 3 σε πειραματική. Στα αποτελέσματα, οι συγγραφείς δίνουν κανονικοποιημένες τιμές των εκφράσεων, έχοντας χρησιμοποιήσει quantile normalization μεταξύ όλων των δειγμάτων, όπου η τιμή κάθε δείγματος από τις μικρότερες στις μεγαλύτερες γίνεται ο μέσος όρος των τιμών όλων των δειγμάτων στην αντίστοιχη θέση. Για την ελαχιστοποίηση του bias που έχει εισαχθεί από την κανονικοποίηση, για κάθε miRNA επιλέχθηκε η διάμεσος των τιμών μεταξύ των 3 φυσιολογικών δειγμάτων. Το kit των microarray που χρησιμοποιήθηκε στο πείραμα χρησιμοποιεί ως ετικέτες τα ονόματα των miRNA της miRBase 12, όμως οι συγγραφείς έχουν εισάγει ονόματα από νεότερες εκδόσεις και επομένως χρησιμοποιήθηκε η miRBase 16.

Ο αριθμός των 3’UTR που δόθηκε ως είσοδος στους 2 αλγορίθμους φαίνεται στον παρακάτω πίνακα.

Πίνακας 4. Στατιστικά εισόδου 3’UTR περιοχών στους αλγορίθμους πρόβλεψης στόχων miRNA

	Alu 3’UTR περιοχές	non-Alu 3’UTR περιοχές
ERR188182	776	70
ERR188298	617	45
ENCLB155EFP (ADAR knockdown)	237	10
ENCLB420RAA (K562 wildtype)	432	19

Κάθε εργαλείο εκτελέστηκε 2 φορές, μία για τις μη τροποποιημένες 3’UTR και μία για τις τροποποιημένες. Το TargetScan εκτελείται με τη μορφή σειριακών εντολών, ενώ το MIRZA-G χρησιμοποιεί ένα ξεχωριστό αρχείο που ορίζει όλες τις παραμέτρους και τα αρχεία εισόδου και εκτελείται μέσω ενός python script. Η εκτέλεση του MIRZA-G παρουσίασε αρκετές προκλήσεις. Στη γραμμή 217 του script rg-merge-results-and-add-probability.py τοποθετήθηκε η εντολή

```
data.to_csv(os.getcwd() + '/features_and_probabilities.csv', sep='\t',
index=None, na_rep="NaN", header=data.columns)
```

ώστε να αποθηκευτούν τα score των επιμέρους χαρακτηριστικών, εκτός από το τελικό score.Οι εντολές εκτέλεσης για το ERR188182, και αναλόγως για τα υπόλοιπα δείγματα, παρατίθενται ακολούθως.

TargetScan

```
# Identify miRNA binding sites
targetscan_70.pl ts_ERR188182_top50expressed__mirnas.txt
ERR188182_UTR_no_editing_TS.txt ERR188182_ts_pred_unedited.txt
targetscan_70.pl ts_ERR188182_top50expressed__mirnas.txt
ERR188182_UTR_with_edit_TS.txt ERR188182_ts_pred_edited.txt

# Calculate branch length (BL) and probability of conserved targeting (PCT).
Both are needed for evolutionary features. The files are needed for the
context score calculation, even though they are all zeros for this thesis,
without introducing any bias to the final results
targetscan_70_BL_bins.pl ERR188182_ts_pred_unedited.txt >
ERR188182_unedited_median_BLs_bins.txt
targetscan_70_BL_bins.pl ERR188182_ts_pred_edited.txt >
ERR188182_edited_median_BLs_bins.txt
targetscan_70_BL_PCT.pl ts_ERR188182_top50expressed__mirnas.txt
ERR188182_ts_pred_unedited.txt ERR188182_unedited_median_BLs_bins.txt >
ERR188182_unedited.BL_PCT.txt
targetscan_70_BL_PCT.pl ts_ERR188182_top50expressed__mirnas.txt
ERR188182_ts_pred_edited.txt ERR188182_edited_median_BLs_bins.txt >
ERR188182_edited.BL_PCT.txt

# The transcripts@genes between edited and unedited are the same, so run the
next 3 steps only for edited files and use them for unedited too
# Subsetting from the supplementary files only the transcripts of interest
awk -F "\@|\t" '{print $1 "\t" $1 "@" $2}' ERR188182_UTR_with_edit_TS.txt >
ERR188182_edited_trascripts_and_genes_list.txt
awk -F'\t' 'NR==FNR{a[$1]=$2} NR>FNR{if (a[$1]) {print a[$1] "\t" $2 "\t"
$3}}' ERR188182_edited_trascripts_and_genes_list.txt
ORF_Sequences_only_human.txt > ERR188182_ORF_Sequences_only_edited.txt
targetscan_count_8mers.pl ts_ERR188182_top50expressed__mirnas.txt
ERR188182_ORF_Sequences_only_edited.txt >|
ERR188182_alu_ORF_8mer_only_edited_counts.txt

# Calculating the final context++ score
targetscan_70_context_scores.pl ts_ERR188182_top50expressed_mature_mirnas.txt
ERR188182_UTR_no_editing_TS.txt ERR188182_alu_unedited.BL_PCT.txt
ERR188182_ORF_Sequences_only_edited.lengths.txt
ERR188182_ORF_8mer_only_edited_counts.txt
ERR188182_context_scores_unedited.txt
targetscan_70_context_scores.pl ts_ERR188182_top50expressed_mature_mirnas.txt
ERR188182_UTR_with_edit_TS.txt ERR188182_edited.BL_PCT.txt
ERR188182_ORF_Sequences_only_edited.lengths.txt
ERR188182_ORF_8mer_only_edited_counts.txt ERR188182_context_scores_edited.txt
```

MIRZA-G

Παράμετροι (για τροποποιημένες 3'UTR, παρομοίως για τις μη τροποποιημένες)

```
general:
    scripts_dir: "/pipeline_MIRZA/scripts" # Absolute path to the directory
where Pipeline scripts resides
    pipeline_dir: "/pipeline_MIRZA/" # Absolute path to the directory where
Pipeline script (pipeline_MIRZA.py) resides
    UTRs: "/pipeline_MIRZA/data/ERR188182_UTR_with_edit_TS.txt" # file with
the UTR sequences from which the coordinate file will be generated (This
should be in the pipeline/data directory)
    motifs: "/pipeline_MIRZA/data/ERR188182_top50expressed__mirnas.fa" # file
with miRNA/siRNA sequences that was used to generate coordinate file
    mirza_binary: "/mirza/MIRZA" # path to MIRZA binary (or how you invoke in
the bash)
    contrafold_binary: "/contrafold/src/contrafold" # path to CONTRAfold
binary (or how you invoke in the bash)
    model_with_bls: "/pipeline_MIRZA/data/glm-with-bls.bin" # abs path to the
model with BLS (you can find it in the pipeline/data directory)
    model_without_bls: "/pipeline_MIRZA/data/glm-without-bls.bin"
    job_id: "test_mirzag" # id of the job - useful in case when two pipelines
would be run on one cluster in the same time
```

Εντολή εκτέλεσης (για τροποποιημένες 3'UTR, παρομοίως για τις μη τροποποιημένες)

```
python pipeline_MIRZA/pipeline_MIRZA.py --config
/pipeline_MIRZA/ERR188182_config_edited.yaml -T calculate_per_gene_scores -v 4
-L ERR188182_edited.log
```

2.5 Στατιστική ανάλυση πρόβλεψης στόχων miRNA

Σκοπός του βήματος της στατιστικής ανάλυσης των αποτελεσμάτων από την πρόβλεψη στόχων των miRNA στα τροποποιημένα και μη-τροποποιημένα σύνολα 3'UTR ήταν η επίδραση του φαινομένου των A-σε-I τροποποιήσεων στην πρόσδεση των miRNA.

Με τη χρήση bar plots, έγινε σύγκριση του αριθμού περιοχών πρόσδεσης που πρόβλεψαν οι 2 αλγόριθμοι. Οι περιοχές πρόσδεσης χωρίστηκαν σε 3 σύνολα: περιοχές που υπήρχαν τόσο στο σύνολο των μη τροποποιημένων όσο και των τροποποιημένων 3'UTR (persistent sites), περιοχές που υπήρχαν μόνο στο σύνολο των μη τροποποιημένων 3'UTR και επομένως χάθηκαν λόγω του φαινομένου της τροποποίησης (lost sites) και περιοχές που υπήρχαν μόνο στο σύνολο των τροποποιημένων 3'UTR και άρα δημιουργήθηκαν λόγω της τροποποίησης (created sites).

Επίσης, χρησιμοποιήθηκαν split violin plots, για να απεικονιστούν οι διαφορές στις κατανομές των χαρακτηριστικών των αλγορίθμων μεταξύ των μη τροποποιημένων και των τροποποιημένων persistent sites. Με τη χρήση normal probability plots, QQ-plots και των Shapiro-Wilk και Lilliefors στατιστικών τεστ κανονικότητας παρατηρήθηκε ότι κανένα χαρακτηριστικό δεν ικανοποιεί τα κριτήρια της κανονικής κατανομής. Επομένως, όπου ο αριθμός των δειγμάτων το επέτρεπε (αριθμός δειγμάτων τροποποιημένων και μη 3'UTR > 500), έγινε τυχαία δειγματοληψία με τη χρήση της συνάρτησης sample στην R, που δεν ξεπερνούσε σε αριθμό το 10% του συνολικού πλήθους των δειγμάτων, και εφαρμόστηκε το Κεντρικό Οριακό Θεώρημα. Όπου επιτεύχθηκε η μετατροπή των κατανομών σε κανονικές, η εκτίμηση της διαφοράς μεταξύ τους έγινε με τη χρήση του paired t-test, εξαιτίας της μη ανεξαρτησίας των δύο συνόλων (ίδιες περιοχές πρόσδεσης, διαφορά στην κατάστασή τους). Στις υπόλοιπες περιπτώσεις, χρησιμοποιήθηκε το Wilcoxon Signed-Rank test. Το κριτήριο της συμμετρίας στην κατανομή της διαφοράς των 2 συνόλων για αυτή την περίπτωση ικανοποιείται εφόσον πηγάζουν από την ίδια κατανομή των 3'UTR. Πιο αναλυτικά, αν α και b τα 2 τυχαία δείγματα από την ίδια κατανομή και δ η διαφορά τους, ισχύει

$$p(a_i = x, b_i = y) = p(a_i = y, b_i = x)$$

$$\delta_{i,1} = x - y$$

$$\delta_{i,2} = y - x = -\delta_{i,1}$$

$$p(\delta_i) = p(-\delta_i)$$

Σε κάθε περίπτωση, εξαιτίας του γεγονότος ότι τα σύνολα αποτελούνται από μεγάλο αριθμό δειγμάτων (>14000) και προστίθεται bias στον υπολογισμό του p-value, αναφέρεται και ο συντελεστής συσχέτισης ως effect size. Η εξίσωση που χρησιμοποιείται για τον υπολογισμό του είναι η ακόλουθη

$$r = \frac{z}{\sqrt{N}}$$

ενώ τα διαστήματα στα οποία κυμαίνεται ορίζουν την επίδραση του effect size ως εξής:

0.1	small effect size
0.3	medium effect size
0.5	large effect size

Μεγαλύτερες τιμές του effect size υποδεικνύουν μεγαλύτερη στατιστική σημαντικότητα του p-value [95][96][97].

Αντίστοιχη διαδικασία ακολουθήθηκε και για τη σύγκριση μεταξύ 2 διαφορετικών κατηγοριών περιοχών πρόσδεσης: περιοχές οι οποίες βρίσκονται σε 3'UTR με υψηλό αριθμό τροποποιήσεων (ως 3'UTR με υψηλό αριθμό τροποποιήσεων ορίστηκαν αυτά που βρίσκονταν στα πρώτα 20 γονίδια με τις περισσότερες τροποποιήσεις ανά δείγμα) και περιοχές που βρίσκονται στα υπόλοιπα τροποποιημένα 3'UTR. Τα διαγράμματα που επιλέχθηκαν σε αυτή την περίπτωση είναι box-and-whiskers plots. Επίσης, συγκρίθηκαν οι 3'UTR με υψηλό αριθμό τροποποιήσεων στην τροποποιημένη και τη μη τροποποιημένη μορφή τους.

Τέλος, ανά αλγόριθμο και δείγμα, συλλέχθηκαν στατιστικά στοιχεία για τη διαφορά στο τελικό score περιοχών πρόσδεσης πριν και μετά την τροποποίηση. Ακόμη, ελέγχθηκε η δημιουργία ή κατάργηση στοιχείων εξαιτίας του φαινομένου της τροποποίησης σε επίπεδο μεταγράφου και γονιδίου.

2.6 Ανάλυση έκφρασης

Επιπρόσθετα της ανάλυσης στόχων των miRNA, διεξάχθηκε και ανάλυση της έκφρασης σε γονιδιακό επίπεδο, με σκοπό την παρατήρηση συσχέτισης μεταξύ έκφρασης και εντοπισμού τροποποιήσεων στα διάφορα δείγματα.

Για τη διεξαγωγή της ανάλυσης χρησιμοποιήθηκε το Salmon [96]. Χρησιμοποιήθηκε η λειτουργία decoy alignment, η οποία απαιτεί ως βήμα προεπεξεργασίας τη δημιουργία ενός αρχείου μορφής fasta με το γονιδίωμα αναφοράς που περιέχει τις ακολουθίες decoy (ακολουθίες οι οποίες είναι γνωστές ότι υπάρχουν στο ανθρώπινο γονιδίωμα, όμως δεν έχει επιτευχθεί η τοποθέτησή τους σε κάποια περιοχή του γονιδιώματος) και τις ακολουθίες των μεταγράφων, καθώς και το indexing αυτών. Οι εντολές με τις οποίες εκτελέστηκαν οι παραπάνω λειτουργίες ήταν οι εξής:

```
grep "^>" <(cat GRCh38/GRCh38_full_analysis_set_plus_decoy_hla.fa) | cut -d " " -f
1 > decoys.txt
sed -i -e 's/>//g' decoys.txt
cat Homo_sapiens.GRCh38.cdna.all.fa.gz
GRCh38_full_analysis_set_plus_decoy_hla.fa.gz > gentrome.fa.gz

salmon-latest_linux_x86_64/bin/salmon index -t gentrome.fa.gz -d decoys.txt -p 12
-i gentrome_index
```

Η ποσοτικοποίηση της έκφρασης του ERR188182 (και αναλόγως των υπολοίπων) έγινε με την παρακάτω εντολή:

```
salmon-latest_linux_x86_64/bin/salmon quant -i gentrome_index -l A
-1 ERR188182_1.fastq.gz -2 ERR188182_2.fastq.gz -p 8 --validateMappings
-o ERR188182_quant
```

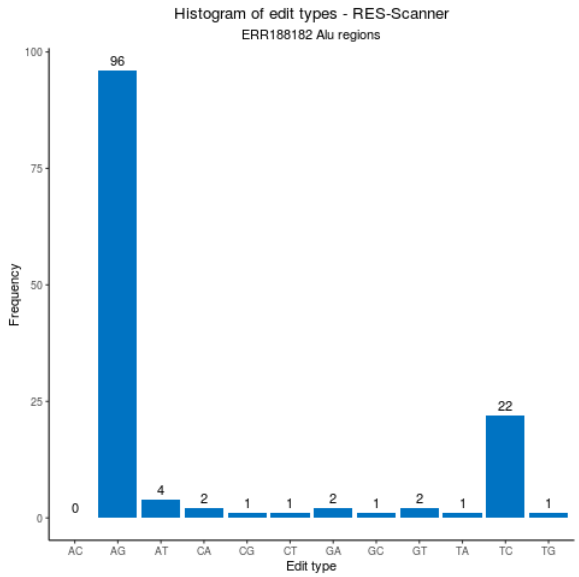
Τα αποτελέσματα του Salmon δίνονται με τη μετρική Transcripts Per Million (TPM), η οποία αποτελεί μια μορφή κανονικοποίησης, για να διορθώσει το bias που εισάγεται από το διαφορετικό βάθος αλληλούχισης ανά μετάγραφο (μεγαλύτερο βάθος θα δώσει εσφαλμένα μεγαλύτερο πλήθος μεταγράφων, χωρίς αυτό να οφείλεται απαραίτητα στη βιολογική έκφραση) και για τη διαφορά μήκους των γονιδίων (μεγαλύτερα γονίδια θα εμφανίζουν περισσότερα reads). Για τη μετατροπή των αποτελεσμάτων από το μεταγραφικό επίπεδο στο γονιδιακό, χρησιμοποιήθηκε το R πακέτο tximport. Η απεικόνιση των αποτελεσμάτων έγινε με τη μορφή ιστογράμματος της έκφρασης ανά γονίδιο.

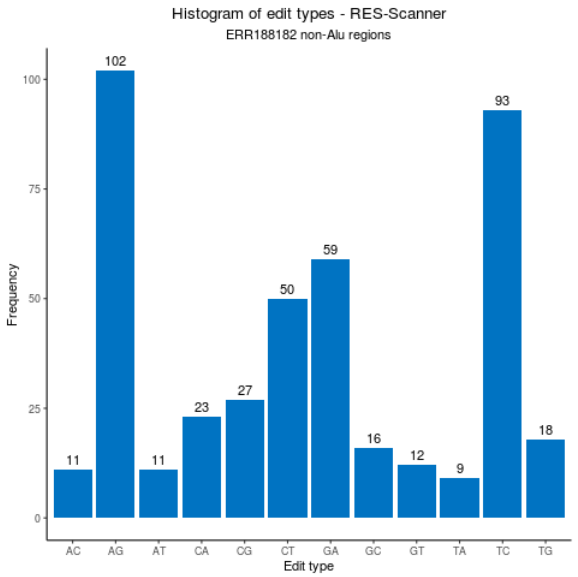
3. ΑΠΟΤΕΛΕΣΜΑΤΑ

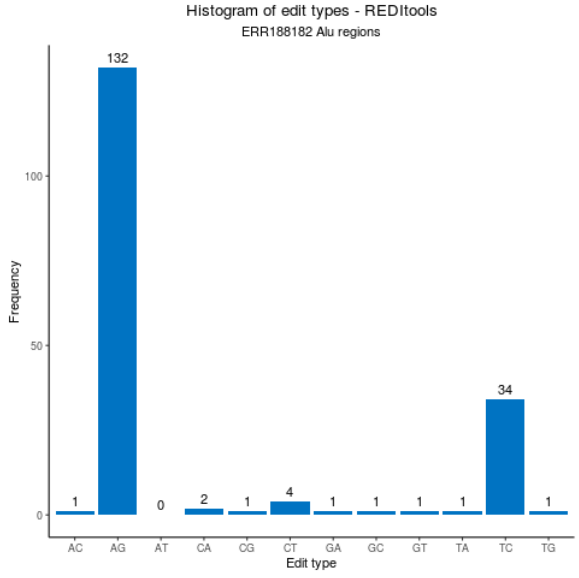
3.1 Σύγκριση εργαλείων εντοπισμού RNA τροποποιήσεων

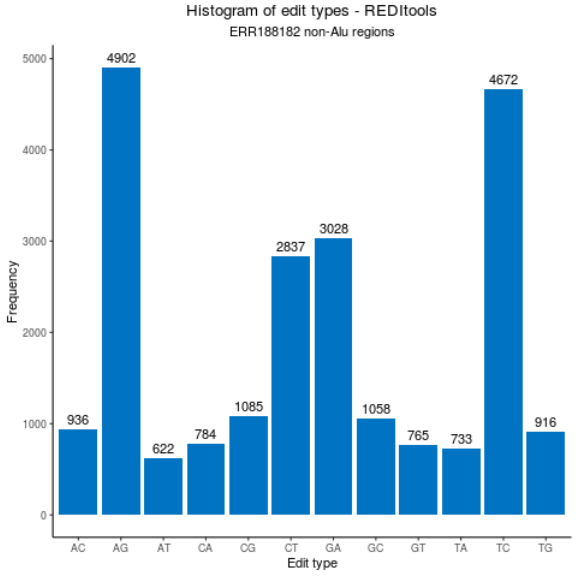
Τα αποτελέσματα από τις εκτελέσεις των εργαλείων εντοπισμού RNA τροποποιήσεων εμφανίζουν ποικιλία ως προς το εργαλείο που εφαρμόστηκε, αλλά και ως προς τον aligner σε μικρότερο βαθμό. Αρχικά, σχεδιάστηκαν τα ιστογράμματα όλων των εντοπισμένων τροποποιήσεων για κάθε εργαλείο και δείγμα. Η αναμενόμενη κατανομή πρέπει να χαρακτηρίζεται από μεγάλη συχνότητα στις A-σε-G τροποποιήσεις και λιγότερες T-σε-C, όταν τα RNA-seq δείγματα δεν είναι strand-specific (ERR188182 και ERR188298) και η χειρωνακτική διόρθωση της έλικας στη συνέχεια είναι ανεπιτυχής. Τα εργαλεία εστιάζουν στον εντοπισμό A-σε-G τροποποιήσεων, όμως μια πολύ μικρή συχνότητα C-σε-T (και αντίστοιχα G-σε-A) είναι αναμενόμενη και επιτρεπτή. Επιπλέον, το πλήθος των τροποποιήσεων αναμένεται να είναι μειωμένο στο ENCLB155EFP της κυτταρικής σειράς K562, καθώς υπενθυμίζεται ότι έχει γίνει knockdown των ADAR σε αυτό το δείγμα.

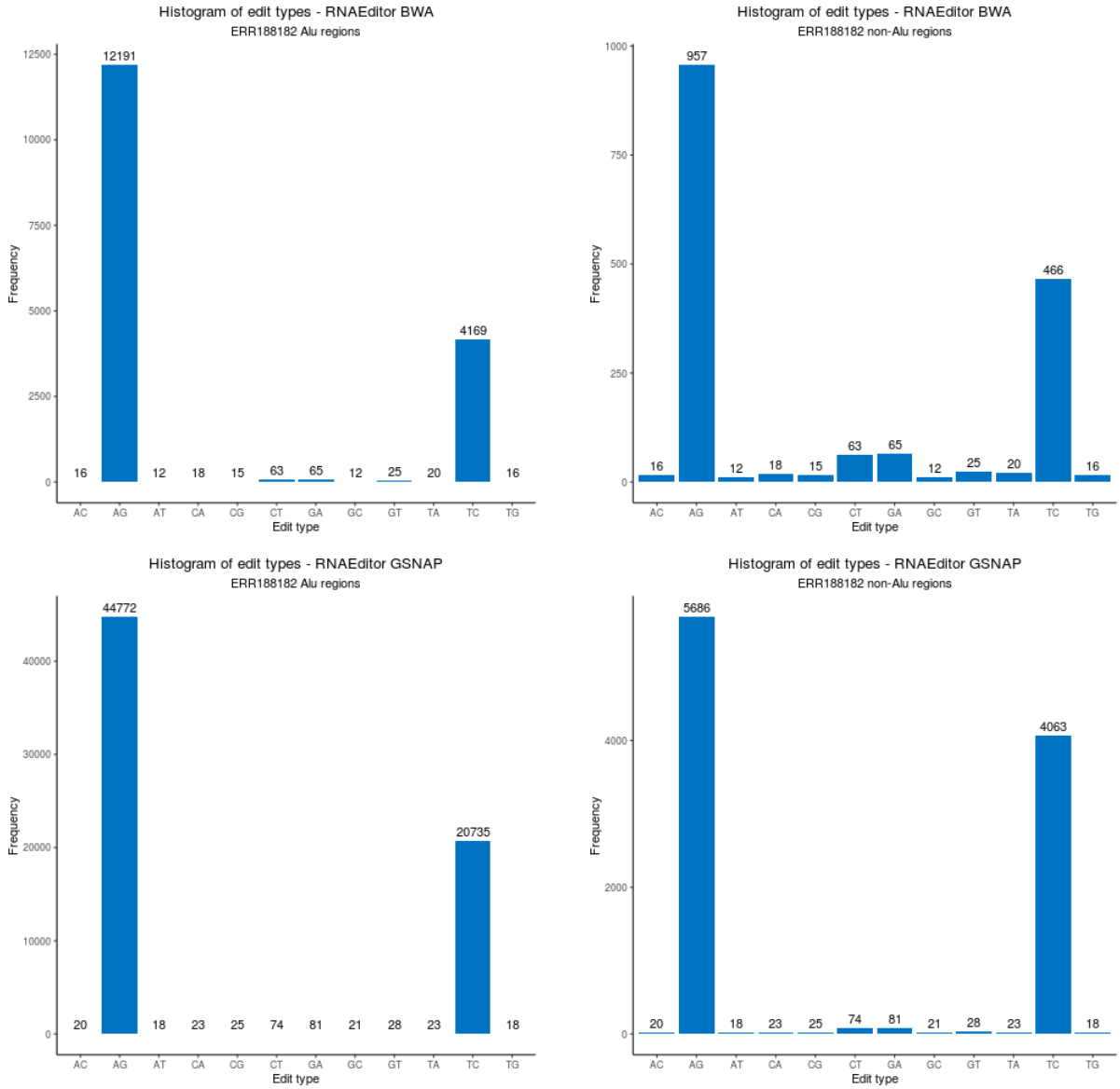
Παρουσιάζονται 3 σύνολα ιστογραμμάτων, ένα για κάθε δείγμα. Κάθε γραμμή απεικονίζει τα αποτελέσματα ενός εργαλείου, ενώ οι 2 στήλες διαχωρίζουν τα αποτελέσματα στις Alu περιοχές (αριστερή στήλη) και στις non-Alu (δεξιά στήλη)



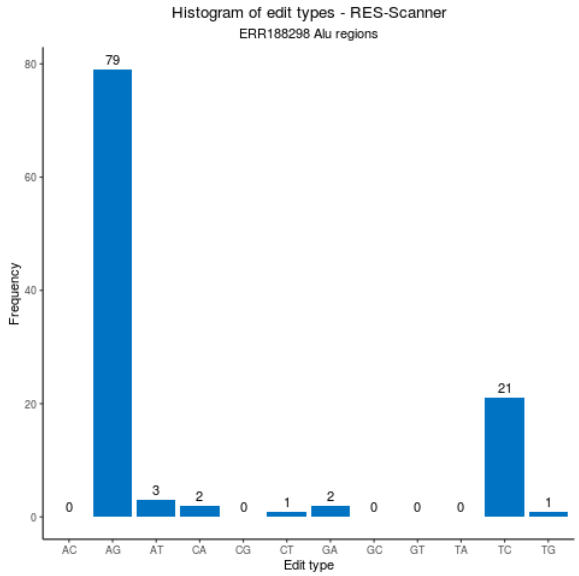


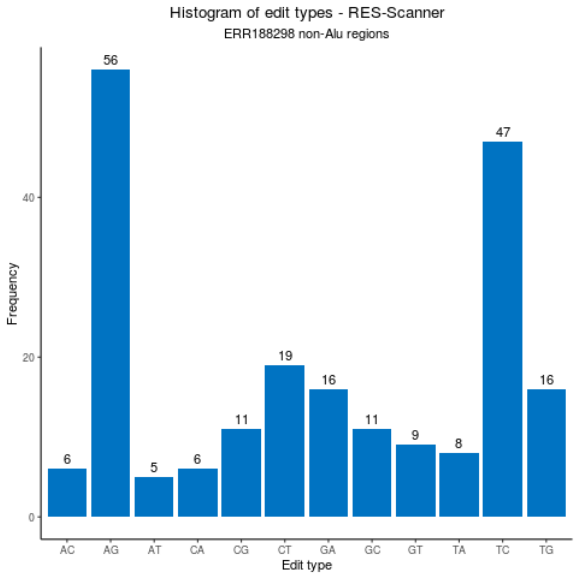


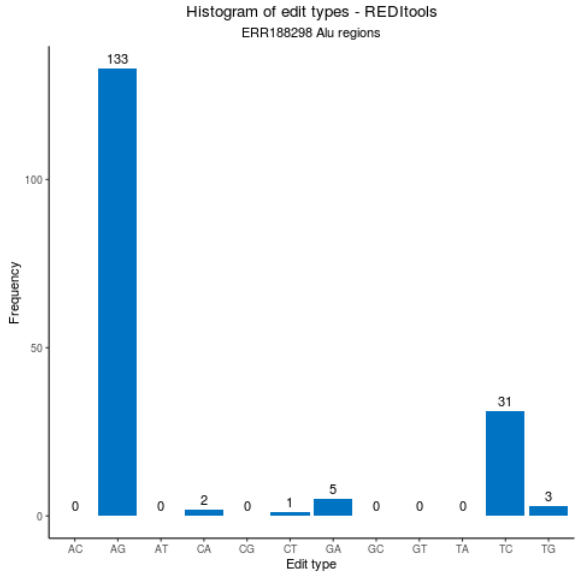


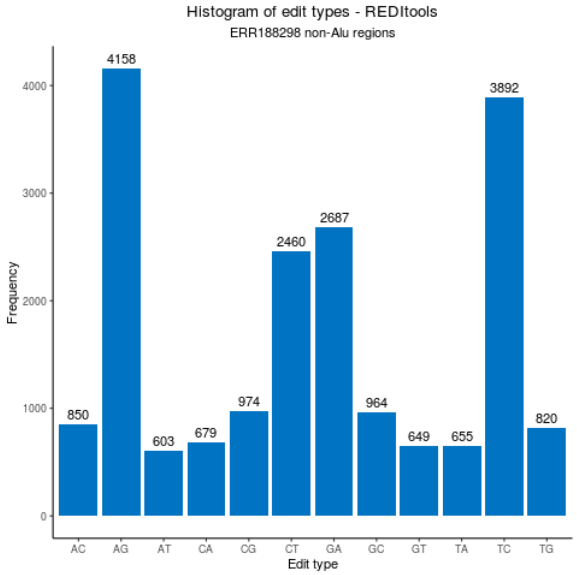


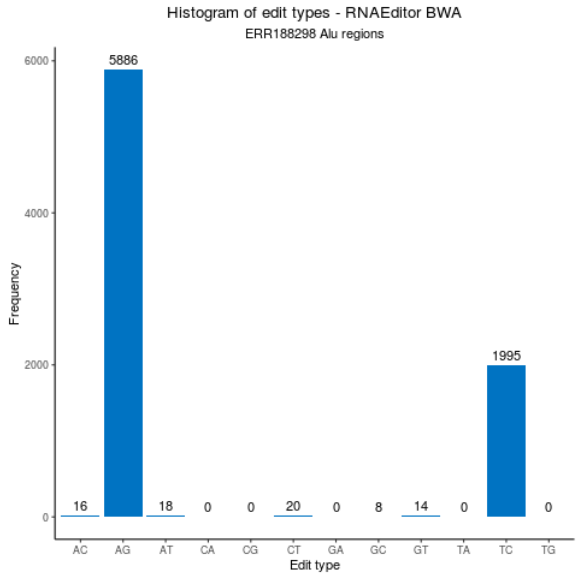
Σχήμα 6. Ιστογράμματα RNA τροποποιήσεων για το ERR188182.

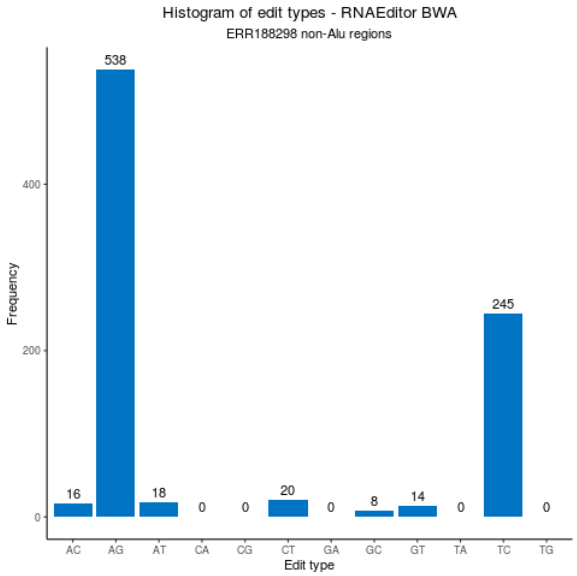

Histogram of edit types - RES-Scanner
ERR188298 Alu regions
Frequency
Edit type
AC AG AT CA CG CT GA GC GT TA TC TG
0 79 3 2 0 1 2 0 0 0 21 1

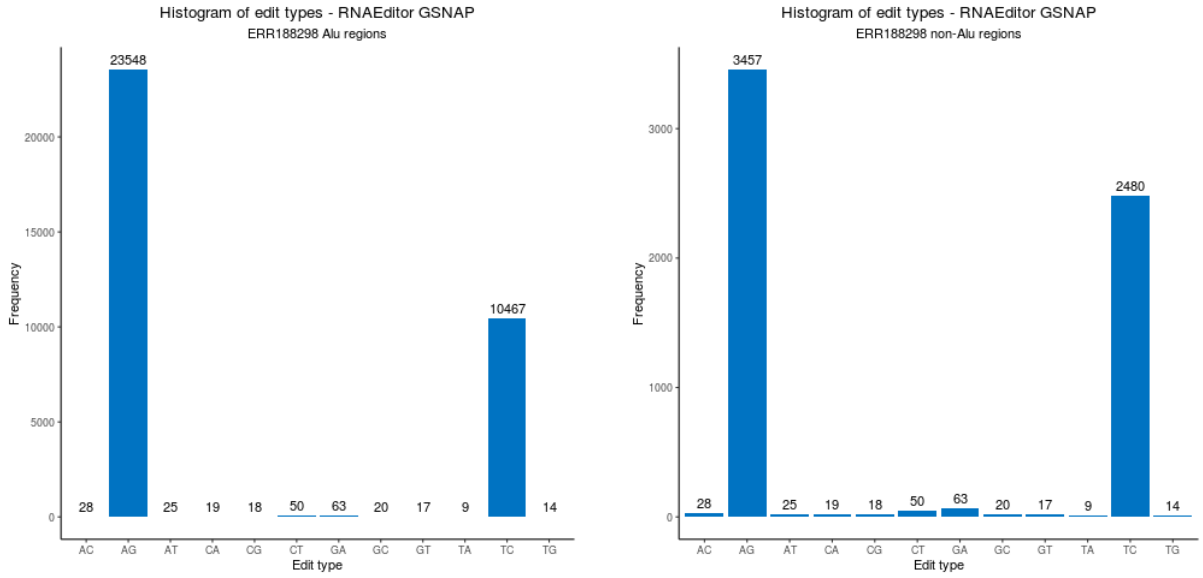

Histogram of edit types - RES-Scanner
ERR188298 non-Alu regions
Frequency
Edit type
AC AG AT CA CG CT GA GC GT TA TC TG
6 56 5 6 11 19 16 11 9 8 47 16


Histogram of edit types - REDItools
ERR188298 Alu regions
Frequency
Edit type
AC AG AT CA CG CT GA GC GT TA TC TG
0 133 0 2 0 1 5 0 0 0 31 3

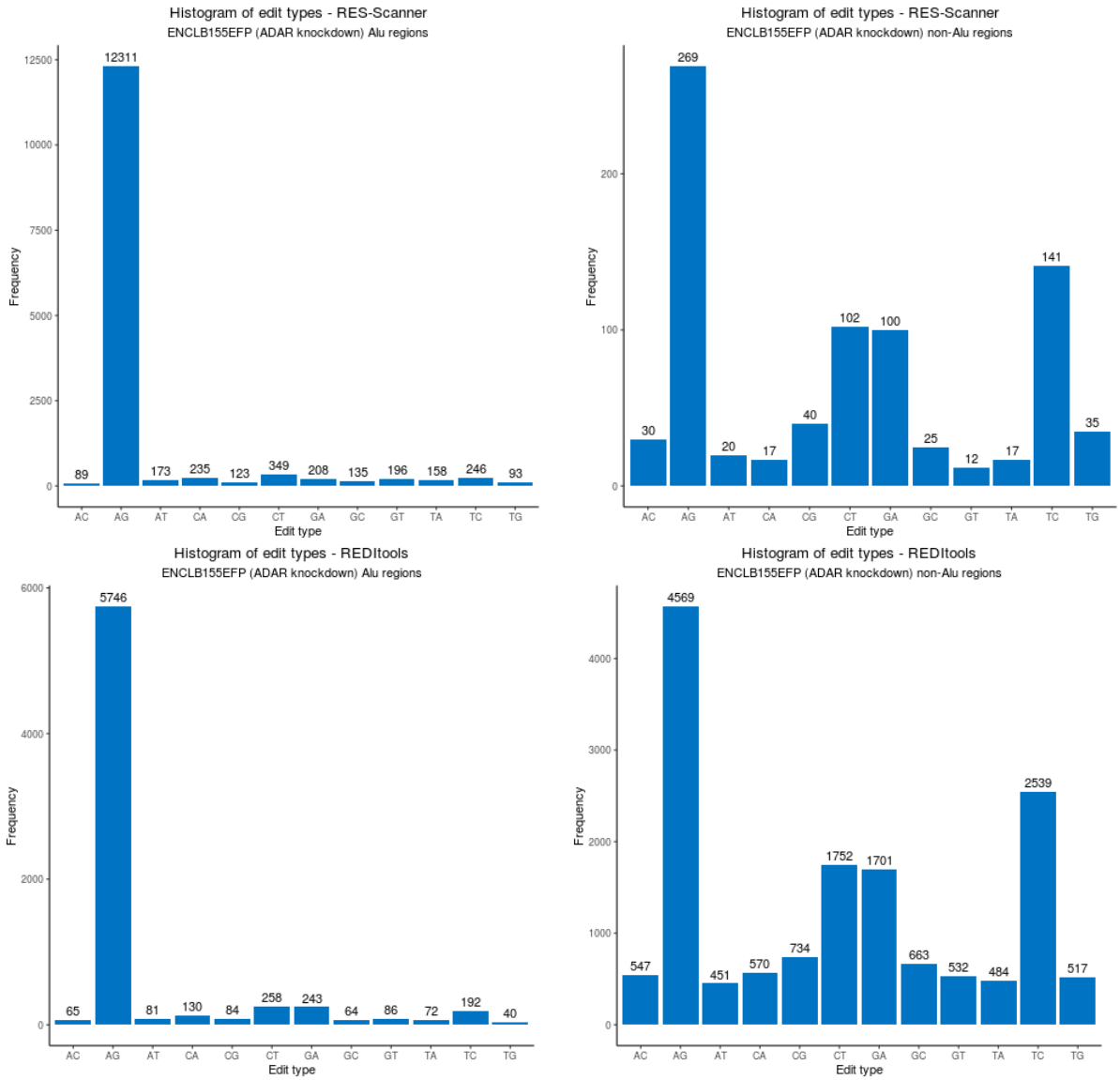

Histogram of edit types - REDItools
ERR188298 non-Alu regions
Frequency
Edit type
AC AG AT CA CG CT GA GC GT TA TC TG
850 4158 603 679 974 2460 2687 964 649 655 3892 820

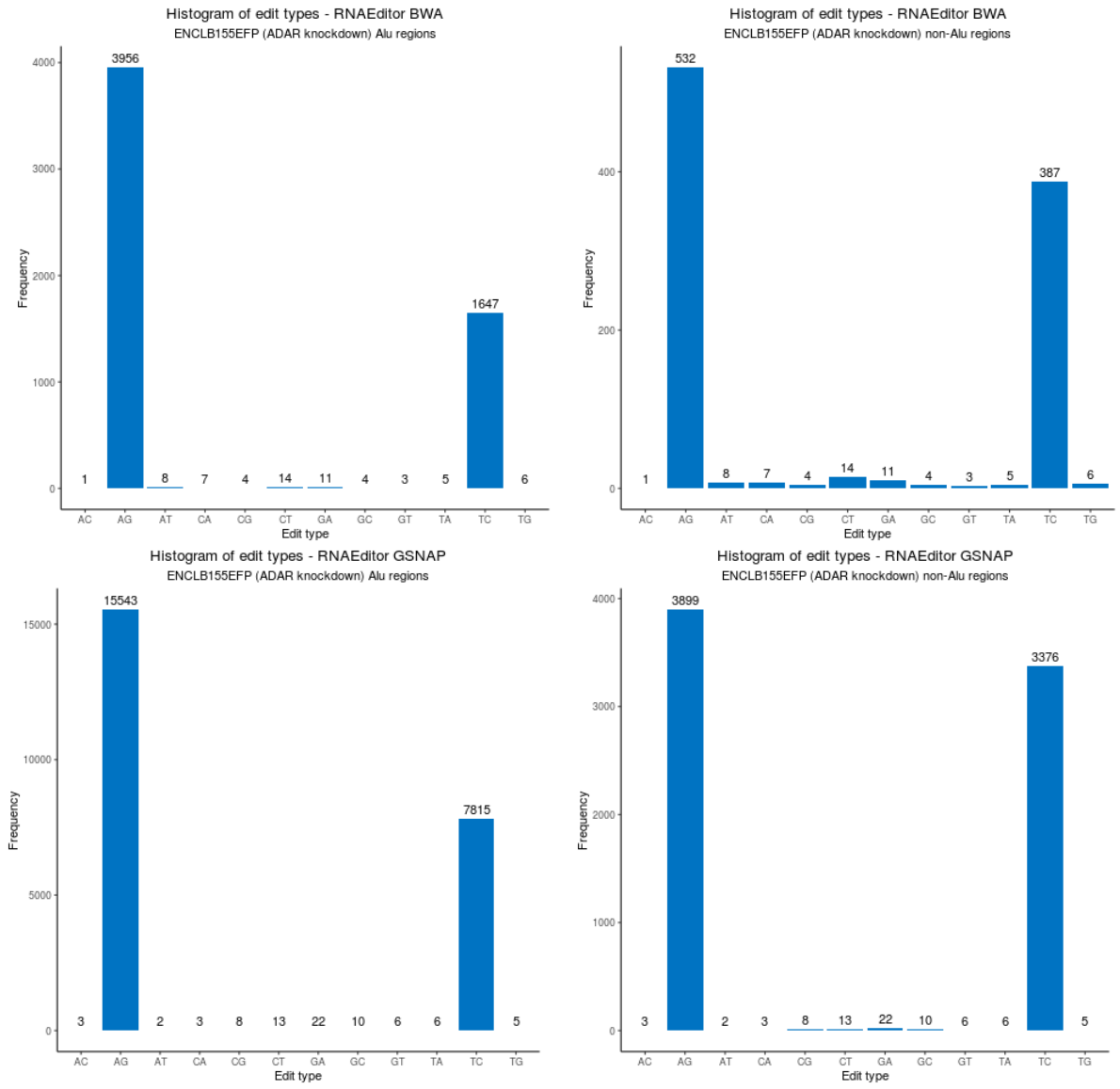

Histogram of edit types - RNAEditor BWA
ERR188298 Alu regions
Frequency
Edit type
AC AG AT CA CG CT GA GC GT TA TC TG
16 5886 18 0 0 20 0 8 14 0 1995 0


Histogram of edit types - RNAEditor BWA
ERR188298 non-Alu regions
Frequency
Edit type
AC AG AT CA CG CT GA GC GT TA TC TG
16 538 18 0 0 20 0 8 14 0 245 0



Σχήμα 7. Ιστογράμματα RNA τροποποιήσεων για το ERR188298.

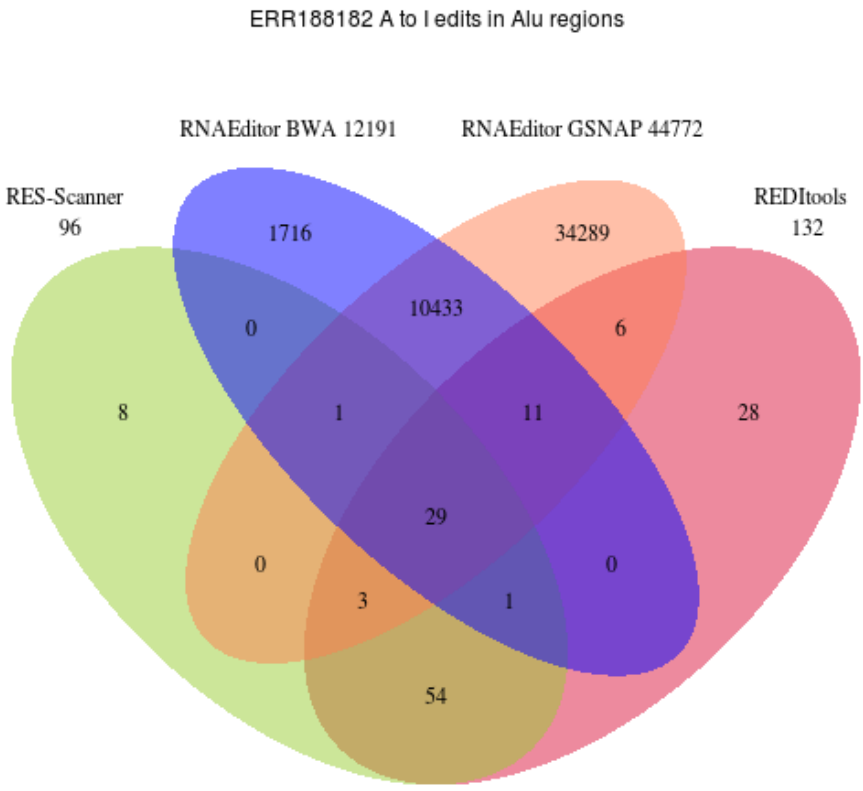




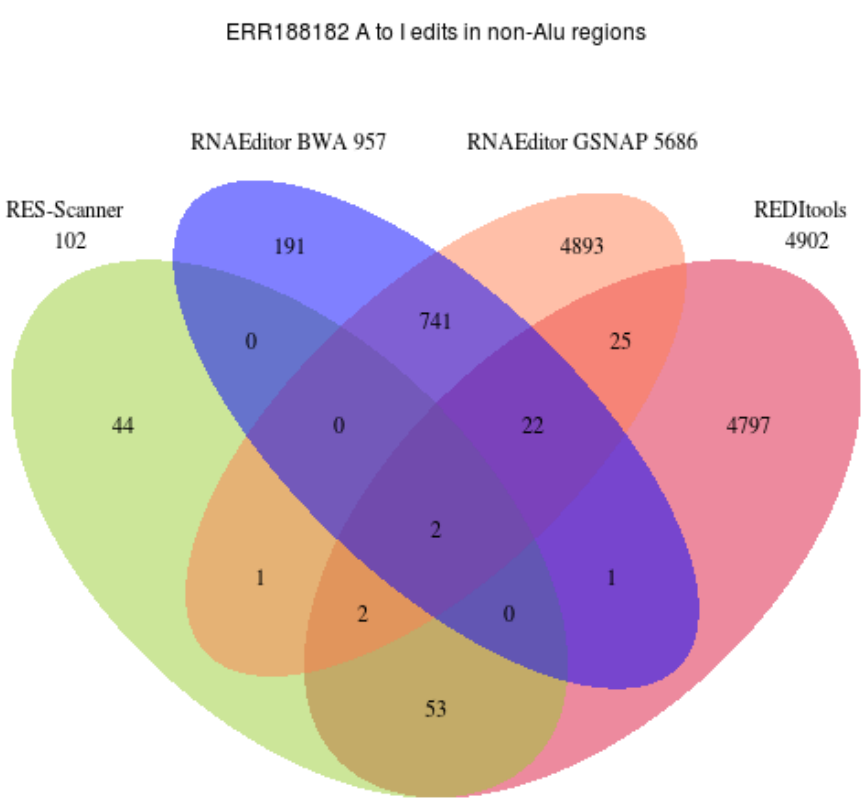
Σχήμα 8. Ιστογράμματα RNA τροποποιήσεων για το ENCLB155EFP.

Όπως μπορεί να παρατηρήσει κανείς στα ιστογράμματα, τα εργαλεία RES-Scanner και REDItools δεν παρουσιάζουν την επιθυμητή κατανομή στις non-Alu περιοχές. Ο θόρυβος που παρουσιάζεται υπονοεί την ανάγκη για αυστηρότερα κριτήρια στον εντοπισμό τροποποιήσεων σε αυτή την περιοχή, ενώ ο μειωμένος αριθμός τροποποιήσεων στα δείγματα ERR188182 και ERR188298 και ο αυξημένος αριθμός τροποποιήσεων στο ENCLB155EFP υπονοεί τον εντοπισμό περισσότερων false positive φαινομένων. Η συμπεριφορά αυτή μπορεί να οφείλεται σε κάποιο βαθμό και στη μικρότερη ανοχή στο βάθος της αλληλούχισης (το ENCLB155EFP έχει μικρότερο βάθος από τα άλλα δύο δείγματα). Το RNAEditor παρουσιάζει πιο σταθερές κατανομές ιστογραμμάτων.

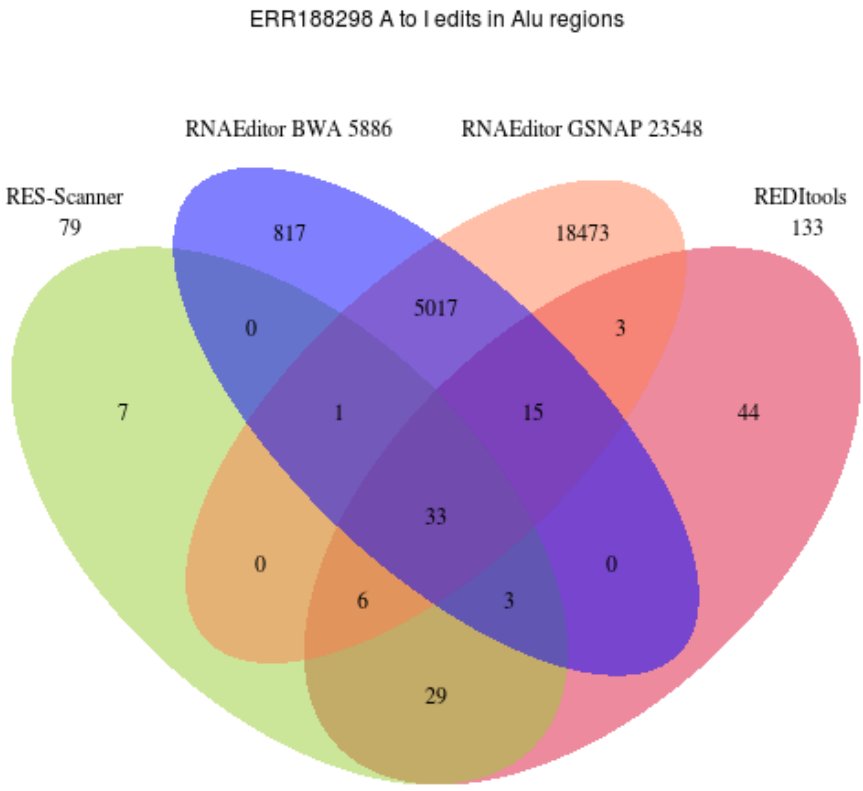
Για καλύτερη απεικόνιση του πλήθους εντοπισμένων τροποποιήσεων, αλλά και τη σχέση των εντοπισμών μεταξύ των διαφόρων εργαλείων, σχεδιάστηκαν τα ακόλουθα διαγράμματα Venn. Από το σύνολο των τροποποιήσεων έχουν χρησιμοποιηθεί μόνο τα A-σε-I φαινόμενα.



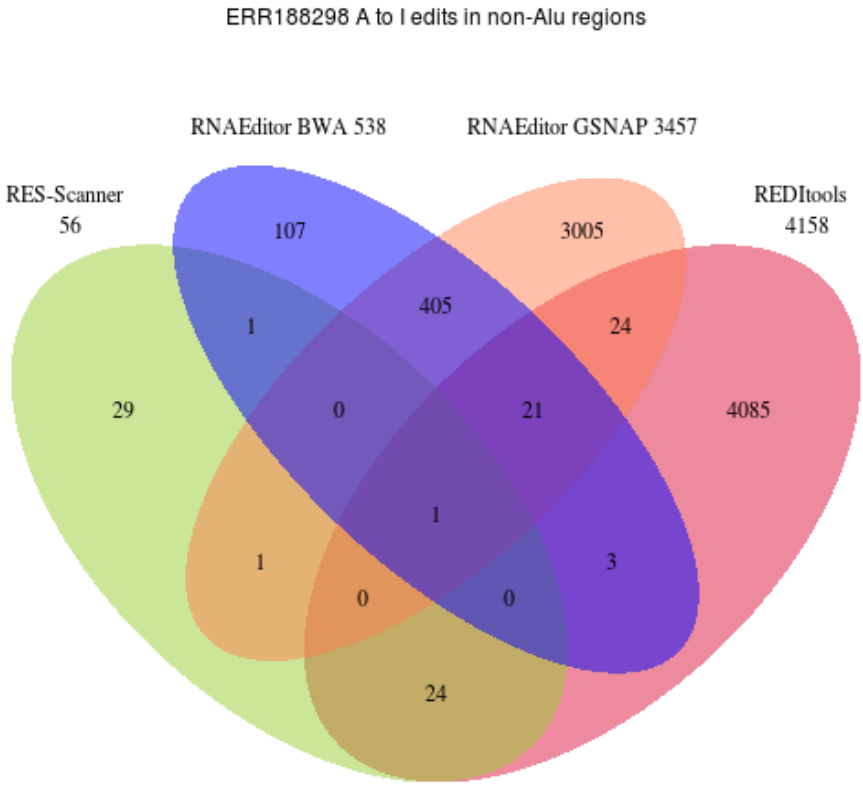
Σχήμα 9. Διάγραμμα Venn για τις Alu περιοχές του ERR188182.



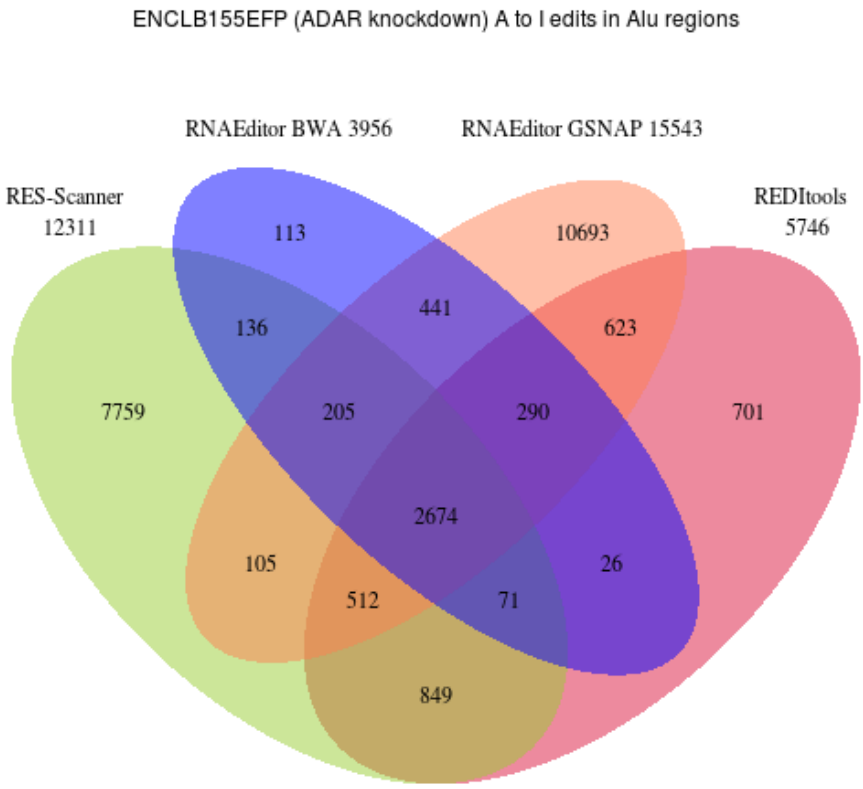
Σχήμα 10. Διάγραμμα Venn για τις non-Alu περιοχές του ERR188182.



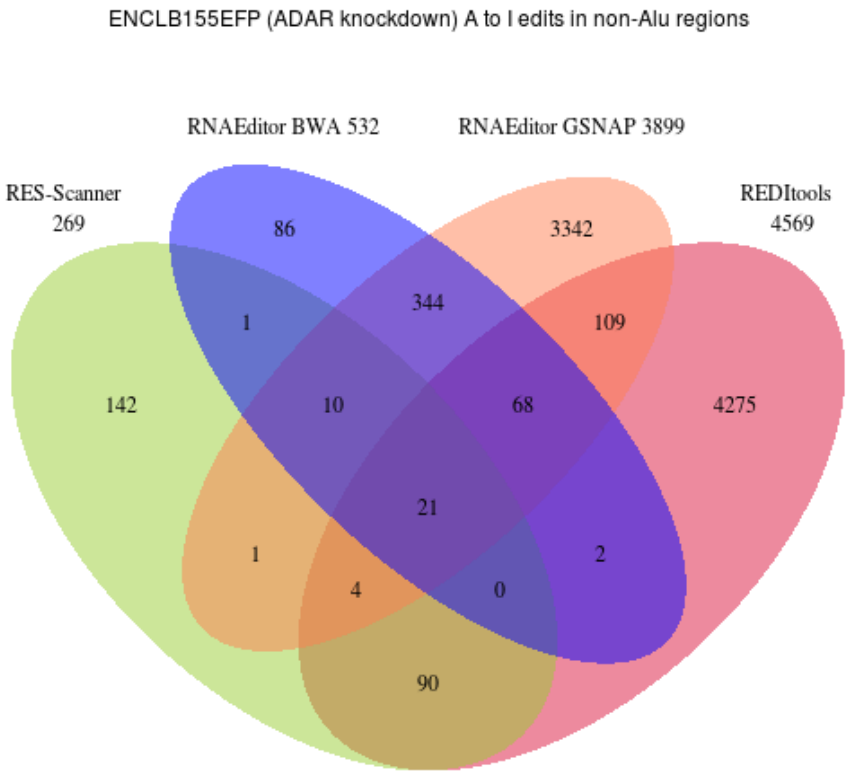
Σχήμα 11. Διάγραμμα Venn για τις Alu περιοχές του ERR188298.



Σχήμα 12. Διάγραμμα Venn για τις non-Alu περιοχές του ERR188298.



Σχήμα 13. Διάγραμμα Venn για τις Alu περιοχές του ENCLB155EFP.



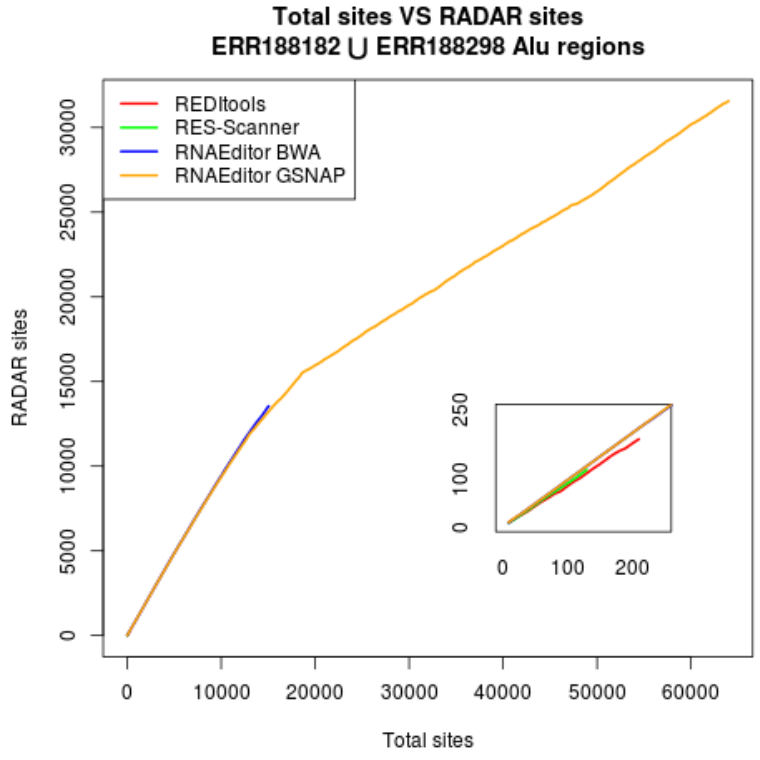
Σχήμα 14. Διάγραμμα Venn για τις non-Alu περιοχές του ENCLB155EFP.

Το RES-Scanner εντοπίζει μικρό αριθμό τροποποιήσεων, ενώ στις Alu περιοχές του ENCLB155EFP εντοπίζει το μεγαλύτερο αριθμό τροποποιήσεων, φαινόμενο το οποίο είναι μη αναμενόμενο και λανθασμένο (εφόσον έχει γίνει αποσιώπηση της έκφρασης των

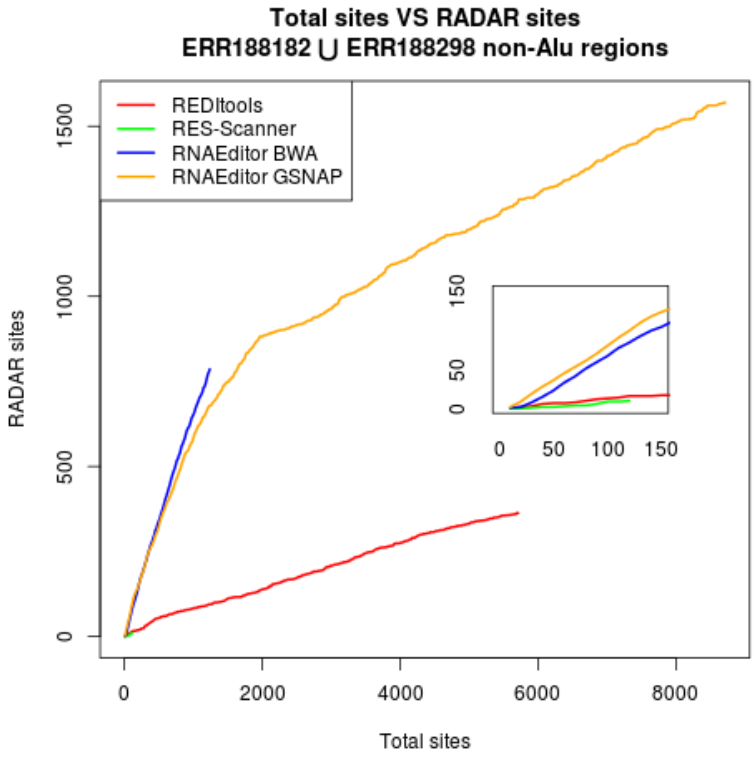
ADAR στο ENCLB155EFP, που είναι υπεύθυνα για την κατάλυση των τροποποιήσεων A-σε-I). Το REDItools παρουσιάζει αυξημένη ευαισθησία, όμως στις non-Alu περιοχές του ENCLB155EFP εντοπίζει επίσης μεγάλο πλήθος τροποποιήσεων, το οποίο υποδεικνύει λανθασμένη συμπεριφορά. Αντίστοιχη συμπεριφορά παρουσιάζει και το RNAEditor (GSNAP), ενώ το RNAEditor (BWA) φαίνεται να παρουσιάζει την πιο ισορροπημένη συμπεριφορά.

Αξιοσημείωτες παρατηρήσεις παρουσιάζονται στις αριθμητικές τιμές των τομών, όπου μεγάλο ρόλο φαίνεται να έχει η επιλογή των κριτηρίων που εφαρμόζονται κατά τον εντοπισμό. Το RES-Scanner και το REDItools, τα οποία ακολουθούν παρόμοια στρατηγική στον εντοπισμό, έχουν σημαντική τομή μεταξύ τους, με το RES-Scanner να παρουσιάζει συνολικά μικρότερο αριθμό εντοπισμών λόγω των αυστηρότερων κριτηρίων που εφαρμόζει. Σημαντική τομή δείχνει και το RNAEditor με τους ξεχωριστούς aligners. Ακόμη, τομή παρουσιάζεται και στις τροποποιήσεις που έχουν εντοπιστεί με το RNAEditor (GSNAP) και το REDItools (που επίσης χρησιμοποιεί τον GSNAP), υποδεικνύοντας τη σημασία της επιλογής στο στάδιο του alignment.

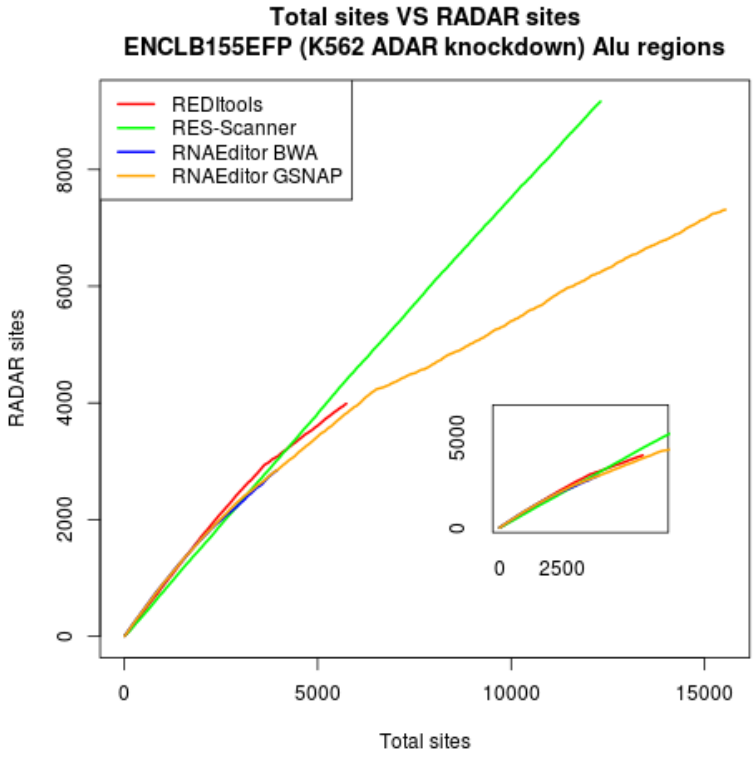
Τέλος, για κάθε δείγμα σχεδιάστηκε σε καρτεσιανό σύστημα η τομή του κάθε εργαλείου με τη βάση δεδομένων RADAR, για τις Alu και τις non-Alu περιοχές. Ο άξονας των x αναπαριστά τις τροποποιήσεις A-σε-I που εντοπίστηκαν από το εργαλείο, ενώ ο άξονας των y αναπαριστά το υποσύνολο αυτών που βρίσκονται στη RADAR. Οι καμπύλες είναι χρωματισμένες με διαφορετικό τρόπο για να ξεχωρίζουν τα εργαλεία. Τα δείγματα ERR188182 και ERR188298 έχουν σχεδιαστεί στο ίδιο καρτεσιανό σύστημα και αναμένονται να δώσουν καμπύλη η οποία θα έχει μεγάλο μήκος και κλίση που προσεγγίζει όσο το δυνατόν τις 90°, που σημαίνει μεγάλο πλήθος φαινομένων (μεγάλη ευαισθησία) που βρίσκονται στη RADAR και άρα παρουσιάζουν μεγαλύτερο βαθμό εμπιστοσύνης (μεγάλη ακρίβεια). Αντιθέτως, για το δείγμα ENCLB155EFP οι καμπύλες αναμένονται να είναι αισθητά μικρότερου μήκους, με κλίση επίσης κοντά στις 90°.



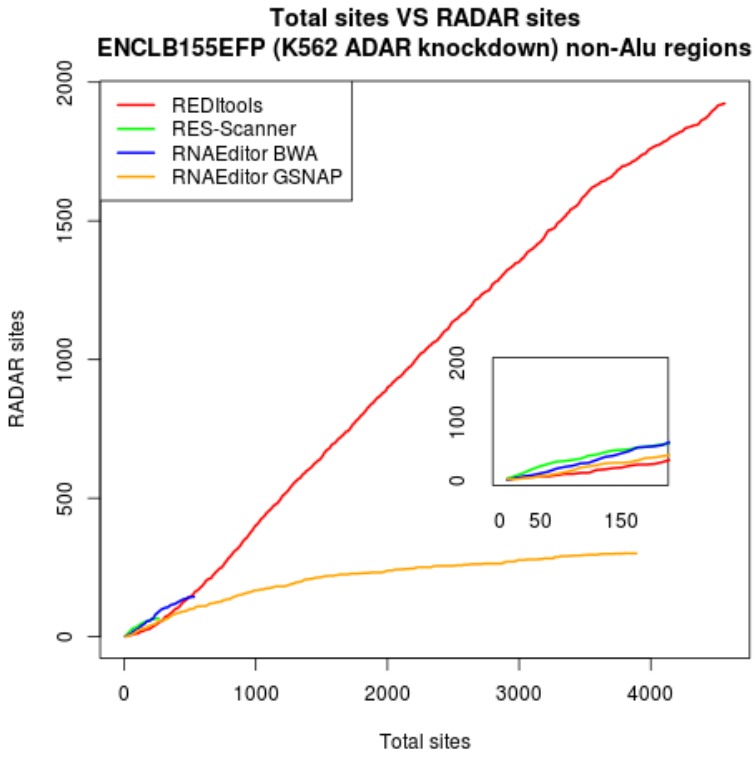
Σχήμα 15. RNA τροποποιήσεις κάθε εργαλείου σε συνάρτηση με τις τροποποιήσεις που βρίσκονται στη RADAR για τις Alu περιοχές των ERR188182 και ERR188298.



Σχήμα 16. RNA τροποποιήσεις κάθε εργαλείου σε συνάρτηση με τις τροποποιήσεις που βρίσκονται στη RADAR για τις non-Alu περιοχές των ERR188182 και ERR188298.



Σχήμα 17. RNA τροποποιήσεις κάθε εργαλείου σε συνάρτηση με τις τροποποιήσεις που βρίσκονται στη RADAR για τις Alu περιοχές του ENCLB155EFP.



Σχήμα 18. RNA τροποποιήσεις κάθε εργαλείου σε συνάρτηση με τις τροποποιήσεις που βρίσκονται στη RADAR για τις non-Alu περιοχές του ENCLB155EFP.

Σε συμφωνία με τα προηγούμενα διαγράμματα, στις παραπάνω καμπύλες παρατηρούμε ορθότερη συμπεριφορά από τις 2 εκτελέσεις του RNAEditor στα ERR188182 και ERR188298. Ειδικότερα, το RES-Scanner εντοπίζει μικρό αριθμό τροποποιήσεων, ενώ το REDItools παρουσιάζει μικρή τομή με τη RADAR στις non-Alu περιοχές. Το RNAEditor (BWA) ξεχωρίζει με τη συνέπεια σχετικά με τη βάση δεδομένων. Στο ENCLB155EFP, το RES-Scanner και το RNAEditor (GSNAP) εντοπίζουν μεγάλο πλήθος από τροποποιήσεις, το οποίο δε συμβαδίζει με τον ADAR knockdown τύπο του δείγματος. Επίσης, το ποσοστό τομής με τη RADAR είναι μικρό. Μη αναμενόμενο είναι και το μεγάλο πλήθος εντοπισμών στις non-Alu περιοχές από το REDItools.

Παρακάτω δίνονται οι τιμές της ακρίβειας που επιτεύχθηκε με κάθε εργαλείο στις Alu και τις non-Alu περιοχές.

Πίνακας 5. Ποσοστά ακρίβειας των εργαλείων εντοπισμού RNA τροποποιήσεων για τις Alu και τις non-Alu περιοχές κάθε δείγματος.

	ERR188182 & ERR188298		ENCLB155EFP (K562 ADAR knockdown)	
	Alu	non-Alu	Alu	non-Alu
RES-Scanner	92%	8%	74%	24%
REDItools	88%	6%	69%	42%
RNAEditor (BWA)	90%	63%	72%	27%
RNAEditor (GSNAP)	49%	18%	47%	7%

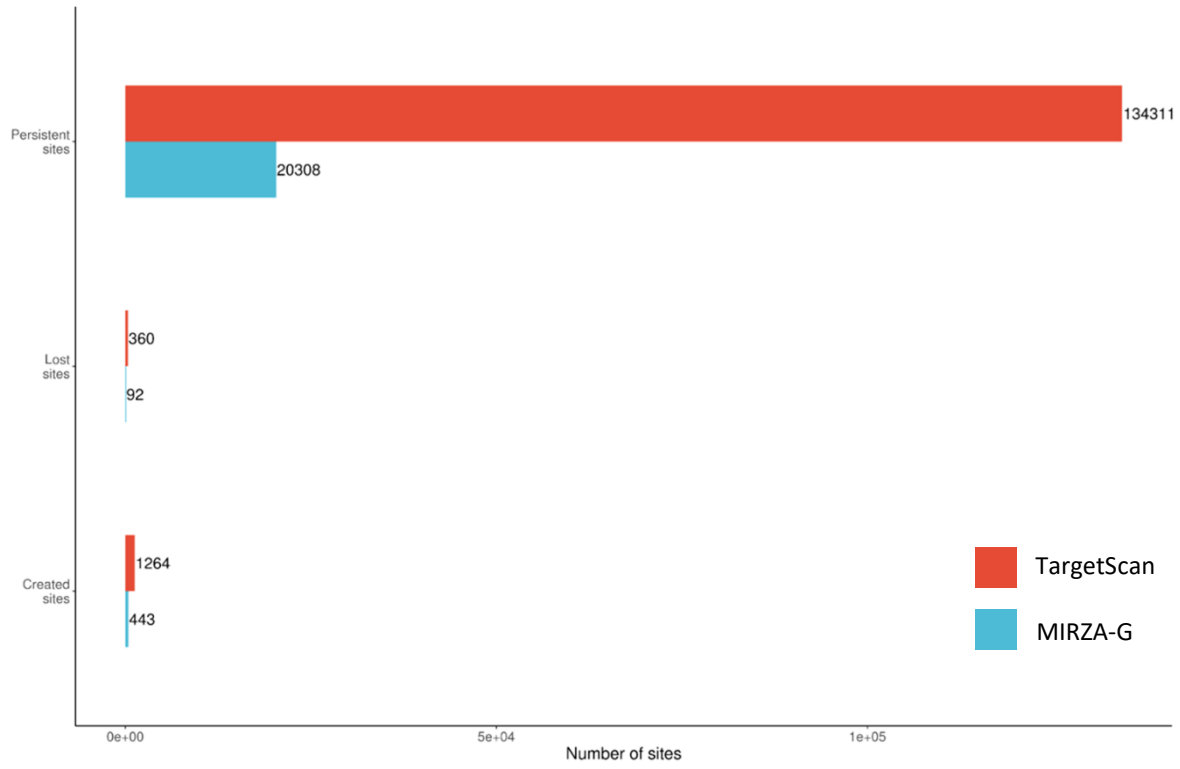
Το RNAEditor BWA επιτυγχάνει υψηλότερη ακρίβεια στο σύνολο των δειγμάτων. Τα RES-Scanner και REDItools επιτυγχάνουν καλά ποσοστά ακρίβειας στις Alu και τις non-Alu περιοχές του ENCLB155EFP.

Συνολικά, το RNAEditor (BWA) δείχνει την καλύτερη συμπεριφορά σε σχέση με τη ζητούμενη ευαισθησία και την ακρίβεια. Για το λόγο αυτό, επιλέχτηκε να είναι το εργαλείο εντοπισμού RNA τροποποιήσεων με το οποίο προχώρησε η εργασία στο βήμα της πρόβλεψης των στόχων των miRNA.

3.2 Επίδραση RNA τροποποιήσεων στη στόχευση των miRNA

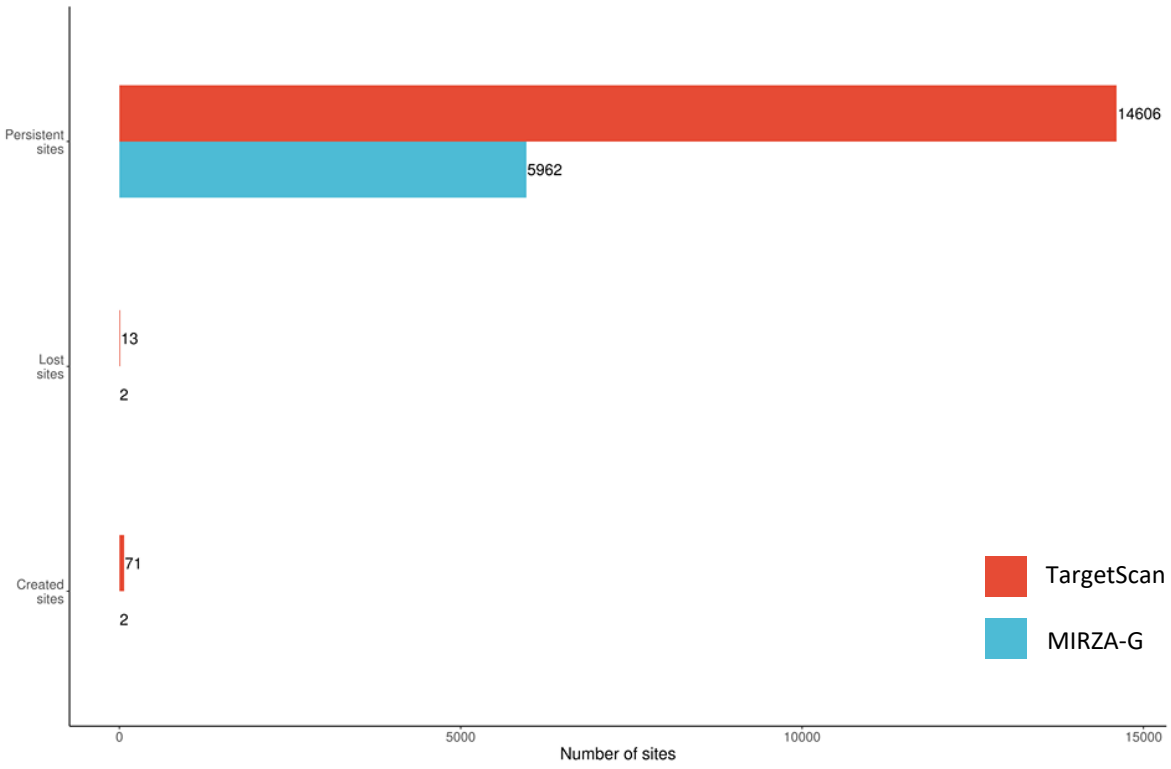
Στο στάδιο της πρόβλεψης στόχων των miRNA, με μια πρώτη ματιά παρατηρείται μεγάλη διαφορά στο πλήθος των περιοχών πρόσδεσης που επέστρεψε ο κάθε αλγόριθμος, με το TargetScan να προβλέπει υπερδιπλάσιο αριθμό. Το φαινόμενο οφείλεται πιθανόν στη μεγαλύτερη βαρύτητα που δίνει το MIRZA-G στα εξελικτικά κριτήρια, το οποίο συμπεραίνεται επιπροσθέτως από τη χρήση μικρότερου πλήθους από χαρακτηριστικά που έχουν να κάνουν με τη δομή και τη θερμοδυναμική. Στα barplots που ακολουθούν απεικονίζεται το πλήθος των περιοχών πρόσδεσης που προβλέφθηκαν από τους 2 αλγορίθμους για τις τρεις κατηγορίες περιοχών πρόσδεσης. Τα ERR188182 και ERR188298 έχουν συγκεντρωθεί σε ένα γράφημα, ενώ τα δείγματα της κυτταρικής σειράς K562 (ADAR knockdown και μη) διαχωρίστηκαν, ώστε να τονιστούν οι διαφορές μεταξύ τους.

- ERR188182 και ERR188298



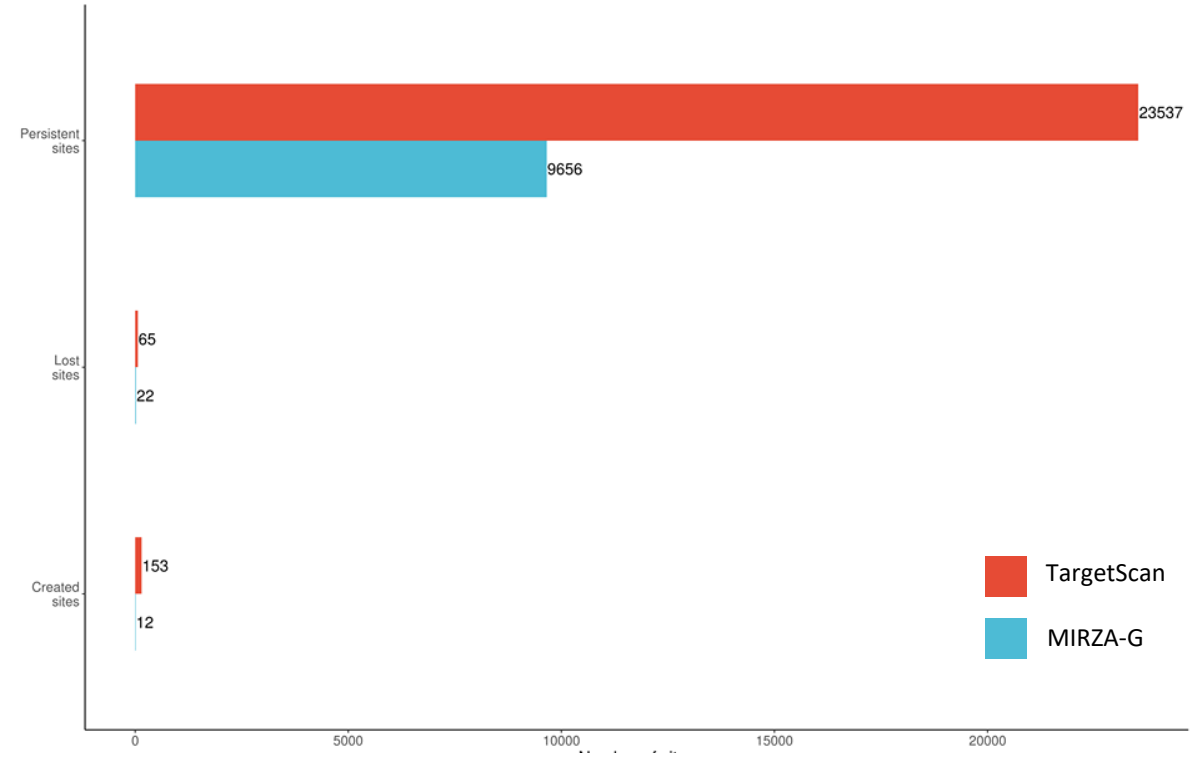
Σχήμα 19. Barplot περιοχών πρόσδεσης miRNA τις οποίες πρόβλεψε το TargetScan (κόκκινο) και το MIRZA-G (μπλε) στα δείγματα ERR188182 και ERR188298.

- ENCLB155EFP (K562 ADAR knockdown)



Σχήμα 20. Barplot περιοχών πρόσδεσης miRNA τις οποίες πρόβλεψε το TargetScan (κόκκινο) και το MIRZA-G (μπλε) στο δείγμα ENCLB155EFP.

- ENCLB420RAA (K562 ADAR with wildtype expression)



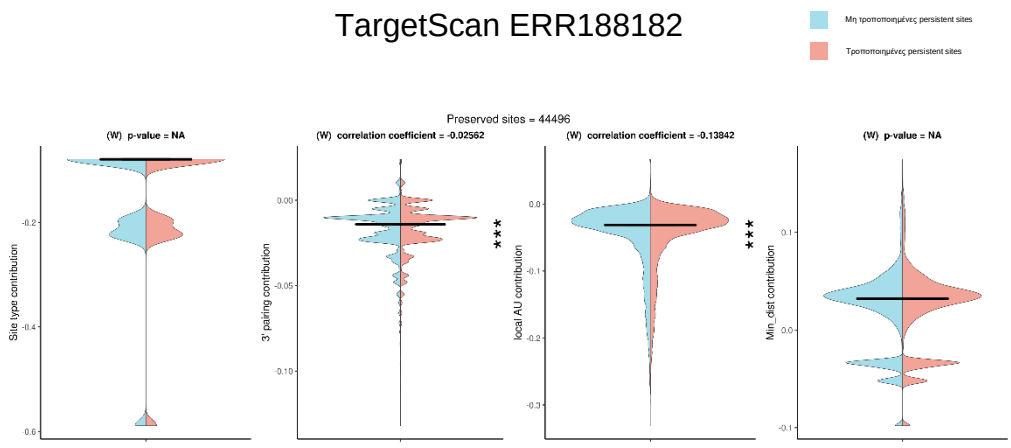
Σχήμα 21. Barplot περιοχών πρόσδεσης miRNA τις οποίες πρόβλεψε το TargetScan (κόκκινο) και το MIRZA-G (μπλε) στο δείγμα ENCLB420RAA.

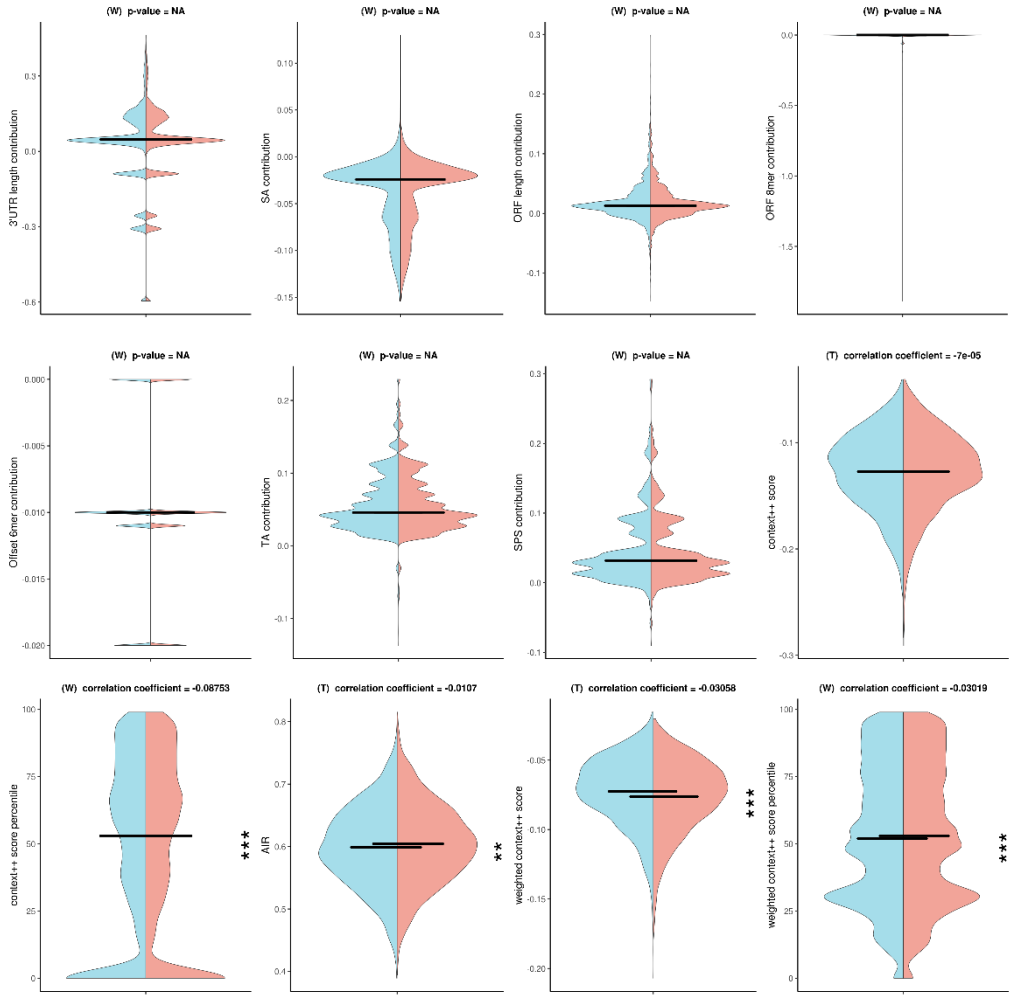
Παρατηρούμε ότι ~99% των περιοχών πρόσδεσης που προβλέφθηκαν από τους αλγορίθμους ανήκουν στο σύνολο των persistent περιοχών, με το αντίστοιχο ποσοστό των περιοχών που δημιουργούνται ή καταργούνται εξαιτίας των τροποποιήσεων να αποτελεί αθροιστικά μόλις ~1%. Το φαινόμενο αυτό υποδεικνύει μια γενικότερη τάση προς ρύθμιση των περιοχών πρόσδεσης λόγω των τροποποιήσεων, έναντι δραστικών αλλαγών όπως η κατάργηση ή η δημιουργία μιας ολόκληρης περιοχής (χωρίς αυτό να εκμηδενίζει τη συχνότητα τέτοιων φαινομένων). Ο μικρότερος αριθμός προβλέψεων στο ENCLB155EFP δικαιολογείται από τα μικρότερα σύνολα 3'UTR που δόθηκαν ως είσοδος στους αλγορίθμους, εξαιτίας των μειωμένων εντοπισμών A-σε-I τροποποιήσεων από το RNAEditor (BWA) λόγω του ADAR knockdown.

Στη συνέχεια, παρουσιάζονται τα διαγράμματα των χαρακτηριστικών των αλγορίθμων πρόβλεψης στόχων των miRNA.

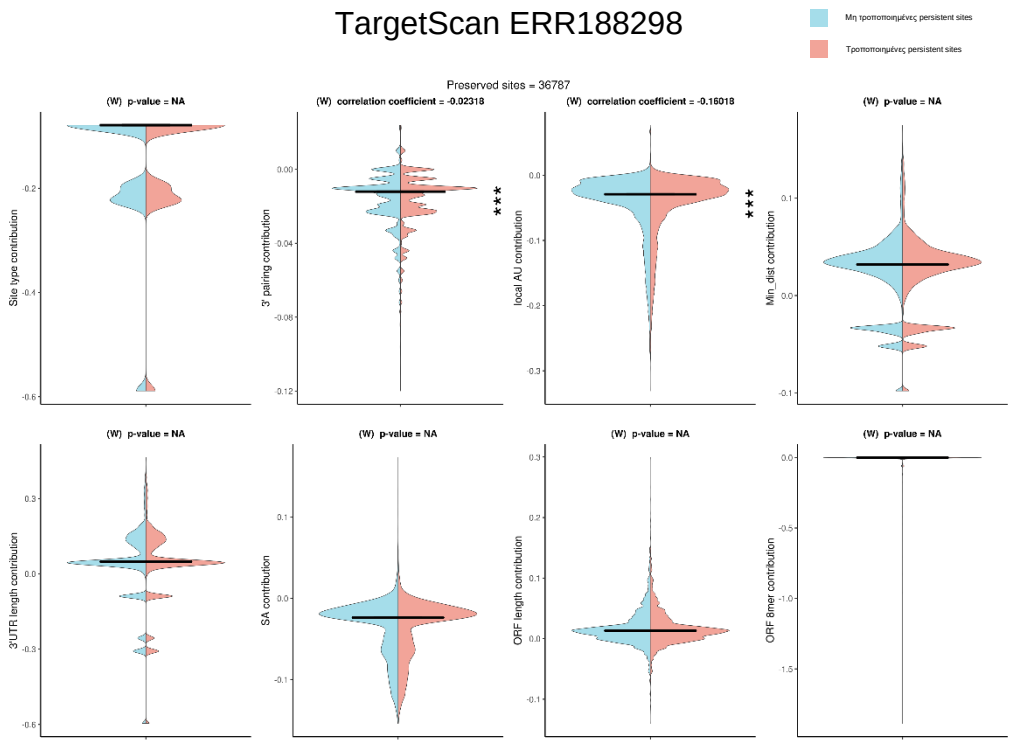
3.2.1 TargetScan

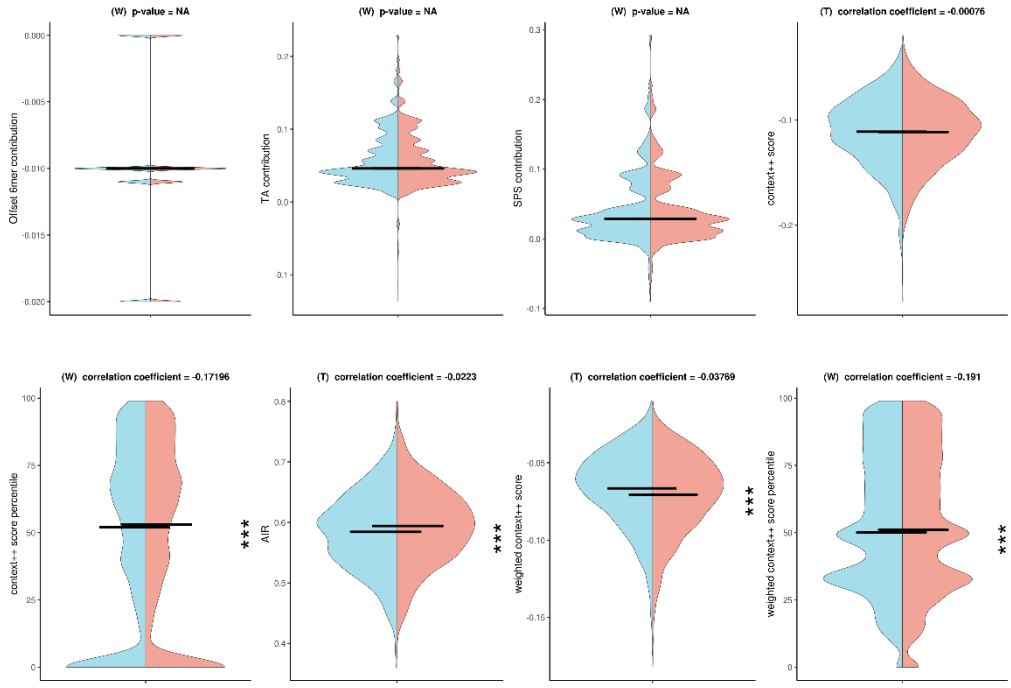
Για το TargetScan επιλέχθηκε να απεικονιστούν 16 από τα 23 χαρακτηριστικά, τα οποία εξετάστηκαν κατά τη στατιστική μελέτη, καθώς οι τιμές των υπολοίπων περιορίζονται σε πλήθος λίγων, διακριτών τιμών και η κατανομή τους δεν ωφελεί στην ερμηνεία της σύγκρισης. Παρατίθενται τα split violin plots των κατανομών των persistent sites, για τα τις μη τροποποιημένες 3'UTR (γαλάζιο) και για τις τροποποιημένες 3'UTR (κοραλλί). Οι τιμές των στατιστικών δοκιμασιών φαίνεται στους πίνακες κάτω από τα διαγράμματα. Επίσης, δίνονται τα box-and-whiskers plots της σύγκρισης των περιοχών πρόσδεσης στις 3'UTR με υψηλό αριθμό τροποποιήσεων, στη μη τροποποιημένη (μπλε) έναντι της τροποποιημένης μορφής τους (κόκκινη) και τα box-and-whiskers plots των ίδιων περιοχών στην τροποποιημένη μορφή τους (κόκκινο) έναντι των υπολοίπων περιοχών πρόσδεσης με τροποποιήσεις (μπλε). Σε κάθε διάγραμμα αναγράφεται η στατιστική σημασία, αν και όπως προέκυψε από τη στατιστική δοκιμασία ($*** = 0{,}01\%$, $** = 0{,}05\%$, $* = 0{,}1\%$), ενώ στην επικεφαλίδα αναγράφεται η στατιστική δοκιμασία που διεξήχθη για τη σύγκριση (T για paired t-test, W για Wilcoxon Signed-Rank), μαζί με την τιμή του correlation coefficient r ως effect size. Paired t-test έχει χρησιμοποιηθεί στις περιπτώσεις που εφαρμόστηκε επιτυχημένα το Κεντρικό Οριακό Θεώρημα, μετατρέποντας τις κατανομές σε κανονικές, όπως φαίνεται και στα αντίστοιχα σχήματα.



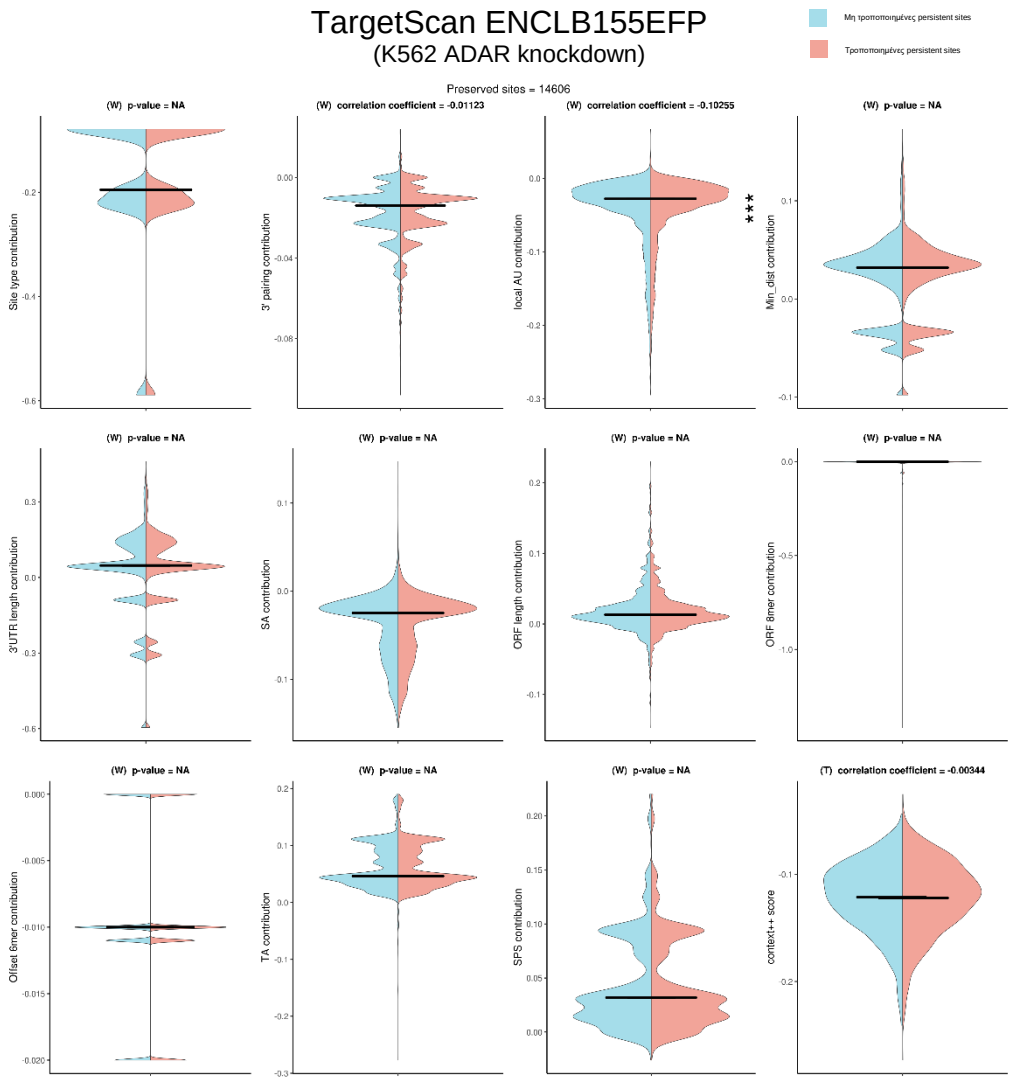


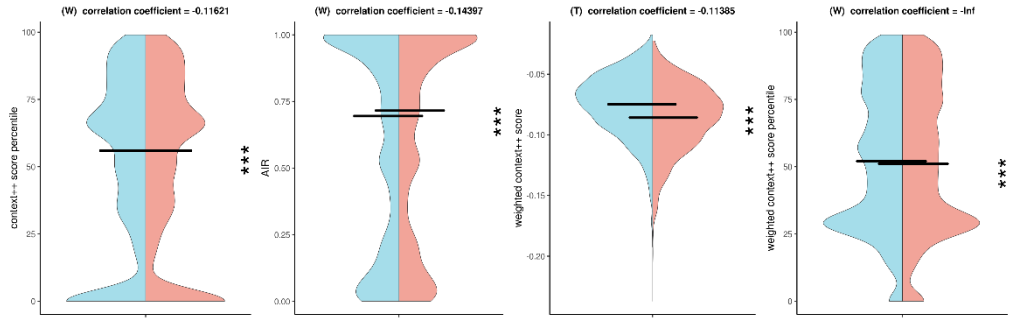
Σχήμα 22. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το TargetScan για το δείγμα ERR188182.



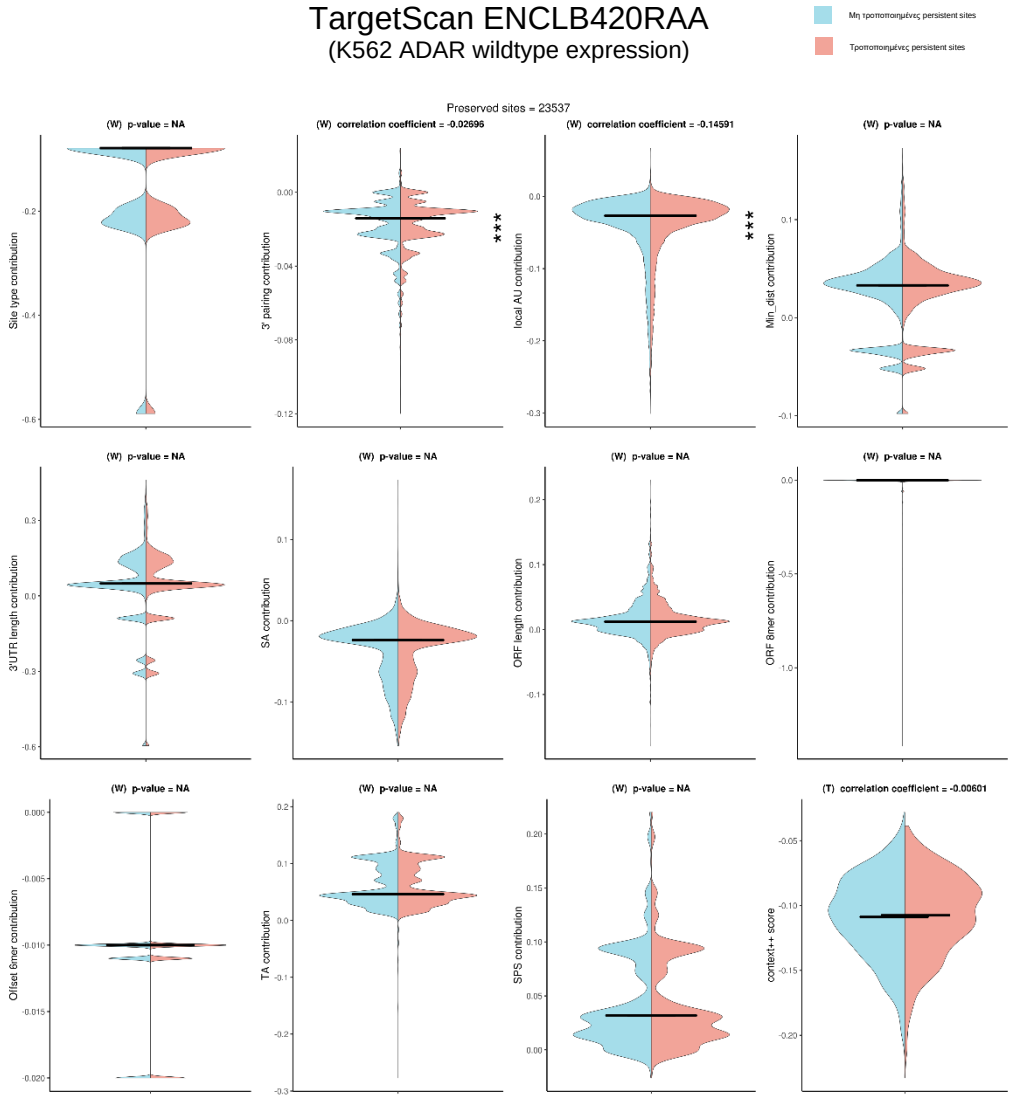


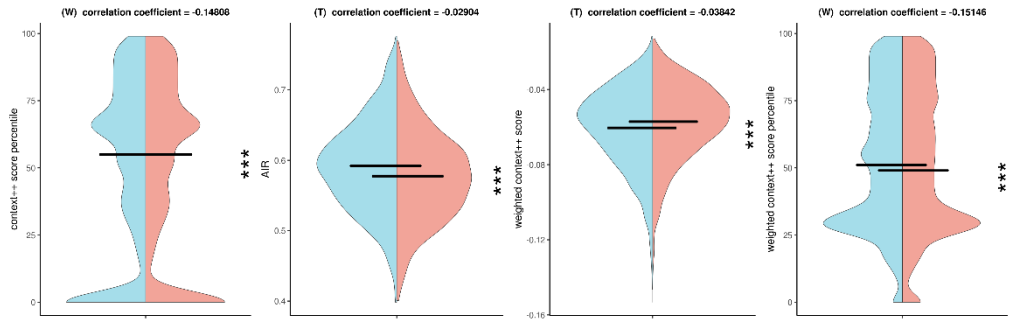
Σχήμα 23. **Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το TargetScan για το δείγμα ERR188**298.





Σχήμα 24. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το TargetScan για το δείγμα ENCLB155EFP.





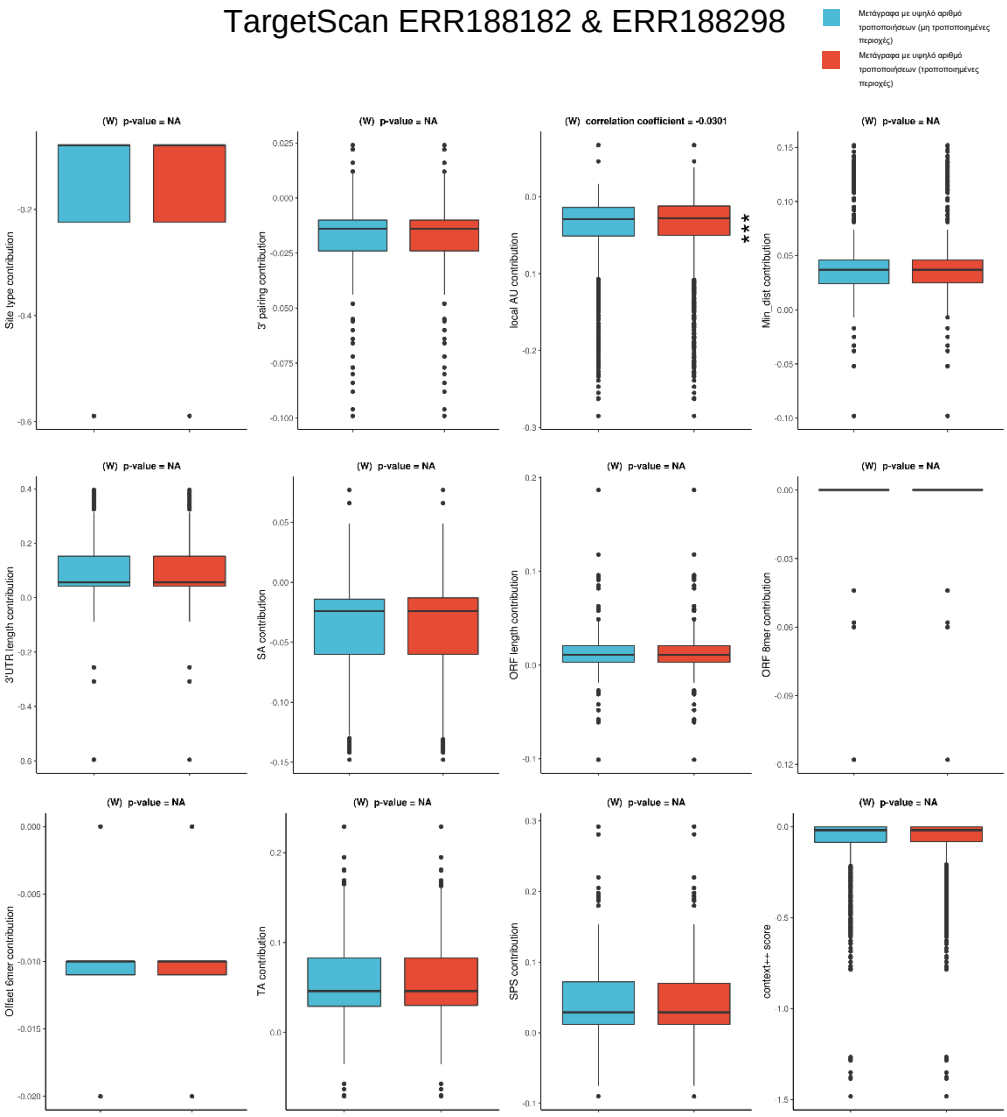
Σχήμα 25. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το TargetScan για το δείγμα ENCLB420RAA.

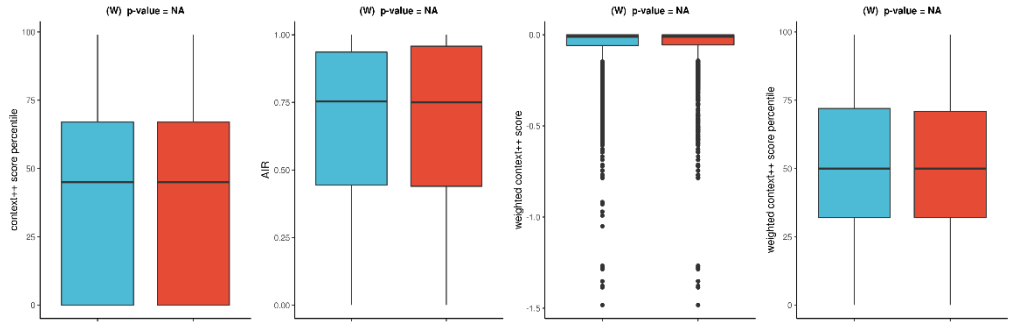
Πίνακας 6. Τιμή του p-value και του effect size (correlation coefficient) κάθε χαρακτηριστικού που προέκυψε από τη στατιστική δοκιμασία μεταξύ των τροποποιημένων/μη-τροποποιημένων persistent sites (TargetScan) των split violin plots κάθε δείγματος.

Targetscan	ERR188182		ERR188298		ENCLB155EFP (K562 ADAR knockdown)		ENCLB420RAA (K562 ADAR wildtype expression)	
	p-value	Effect size	p-value	Effect size	p-value	Effect size	p-value	Effect size
Site type	NA	NA	NA	NA	NA	NA	NA	NA
3' pairing	0	-0.02562	0.00001	-0.02318	0.17486	-0.01123	0.00004	-0.02696
local AU	0	-0.13842	0	-0.16018	0	-0.10255	0	-0.14591
Min_dist	NA	NA	NA	NA	NA	NA	NA	NA
sRNA1A	NA	NA	NA	NA	NA	NA	NA	NA
sRNA1C	NA	NA	NA	NA	NA	NA	NA	NA
sRNA1G	NA	NA	NA	NA	NA	NA	NA	NA
sRNA8A	NA	NA	NA	NA	NA	NA	NA	NA
sRNA8C	NA	NA	NA	NA	NA	NA	NA	NA
sRNA8G	NA	NA	NA	NA	NA	NA	NA	NA
site8A	0.00596	-0.01304	0.00036	-0.01859	NA	NA	NA	NA
site8C	NA	NA	NA	NA	NA	NA	NA	NA
site8G	0.00596	-0.01304	0.00036	-0.01859	NA	NA	0.34578	-0.00615
3'UTR length	NA	NA	NA	NA	NA	NA	NA	NA
SA	NA	NA	NA	NA	NA	NA	NA	NA
ORF length	NA	NA	NA	NA	NA	NA	NA	NA
ORF 8mer	NA	NA	NA	NA	NA	NA	NA	NA
Offset 6mer	NA	NA	NA	NA	NA	NA	NA	NA
TA	NA	NA	NA	NA	NA	NA	NA	NA
SPS	NA	NA	NA	NA	NA	NA	NA	NA
context++ score	0.98838	-0.00007	0.88359	-0.00076	0.67717	-0.00344	0.35681	-0.00601
context++ score percentile	0	-0.08753	0	-0.17196	0	-0.11621	0	-0.14808
AIR	0.02407	-0.0107	0.00002	-0.0223	0	-0.14397	0.00001	-0.02904
weighted context++ score	0	-0.03058	0	-0.03769	0	-0.11385	0	-0.03842
weighted context++ score percentile	0	-0.03019	0	-0.191	0	-∞	0	-0.15146

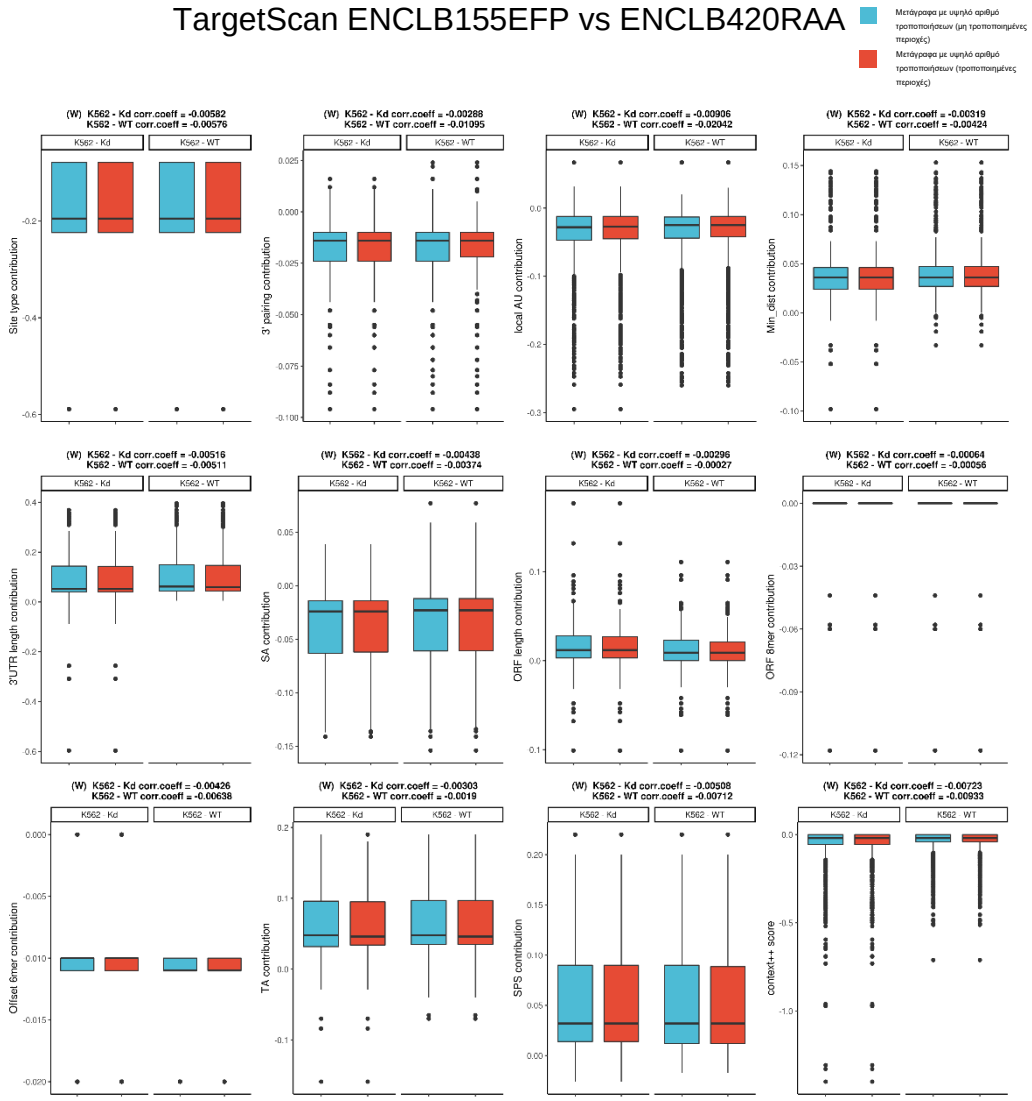
Οι διαφορές στις κατανομές είναι πολύ μικρές, με λίγα χαρακτηριστικά να διαφέρουν στατιστικώς σημαντικά, όπως φαίνεται από το συνδυασμό των τιμών του p-value και του

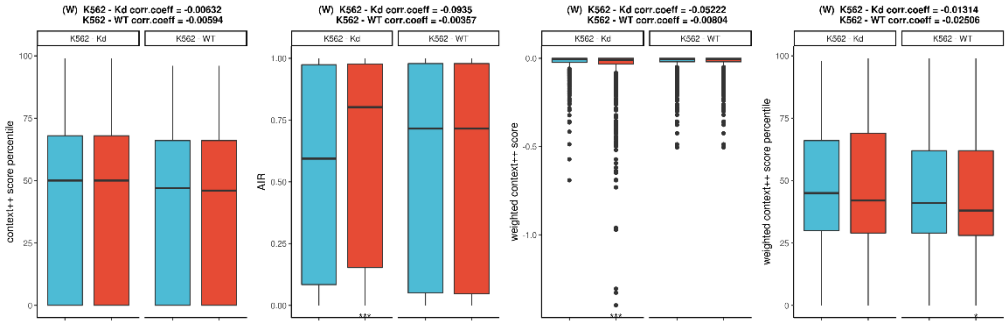
effect size. Πιο συγκεκριμένα, σε όλα τα δείγματα παρατηρείται ήπια διαφορά στην περίπτωση των τροποποιημένων 3'UTR για το χαρακτηριστικό local AU contribution, το οποίο είναι αναμενόμενο, καθώς με τις τροποποιήσεις A-σε-I, και την αναγνώριση της I σαν G, μειώνονται οι A και αυξάνονται οι G, μειώνοντας την αποτελεσματικότητα της περιοχής σαν σημείο πρόσδεσης. Στα δείγματα ERR188182, ERR188298 και ENCLB155EFP (ADAR knockdown) διακρίνεται μια πολύ ήπια μείωση των τιμών του context++ score (και άρα αύξηση της κατασταλτικής ικανότητας), ενώ το ENCLB420RAA (ADAR wildtype) παρουσιάζει αντιθέτως μια ήπια αύξηση. Αντίστοιχα, το ίδιο φαινόμενο, με μεγαλύτερη ένταση, παρατηρείται στο χαρακτηριστικό weighted context++ score. Στατιστικά σημαντικές διαφορές διακρίνονται και στα context++ score percentile και weighted context++ score percentile για τα ERR188298, όπου παρατηρείται αύξηση και των 2 χαρακτηριστικών στα τροποποιημένα σύνολα και για τα ENCLB155EFP και ENCLB420RAA, στα οποία παρατηρείται μείωση στο weighted context++ score percentile.





Σχήμα 26. Box-and-whiskers plots της ένωσης των περιοχών πρόσδεσης (TargetScan) των δειγμάτων ERR188182 και ERR188298 που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων. Για κάθε χαρακτηριστικό γίνεται σύγκριση μεταξύ των μη τροποποιημένων (μπλε) έναντι των τροποποιημένων (κόκκινο) περιοχών.





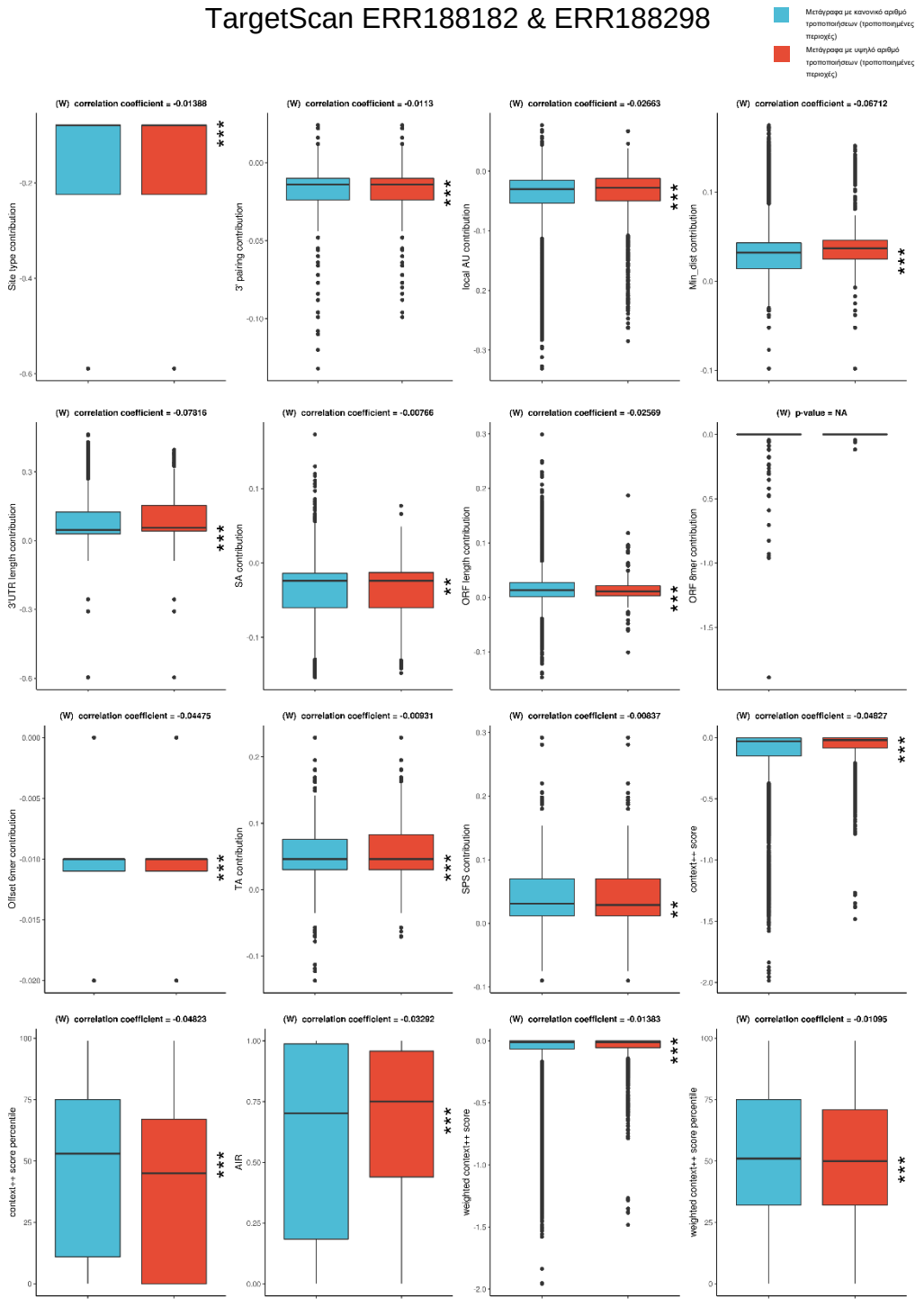
Σχήμα 27. Box-and-whiskers plots των περιοχών πρόσδεσης (TargetScan) των δειγμάτων ENCLB155EFP και ENCLB420RAA ξεχωριστά, που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων. Για κάθε χαρακτηριστικό γίνεται σύγκριση μεταξύ των μη τροποποιημένων (μπλε) έναντι των τροποποιημένων (κόκκινο) περιοχών κάθε δείγματος.

Πίνακας 7. Τιμή του p-value και του effect size (correlation coefficient) κάθε χαρακτηριστικού που προέκυψε από τη στατιστική δοκιμασία μεταξύ των τροποποιημένων/μη-τροποποιημένων περιοχών (TargetScan) των box-and-whiskers plots κάθε δείγματος.

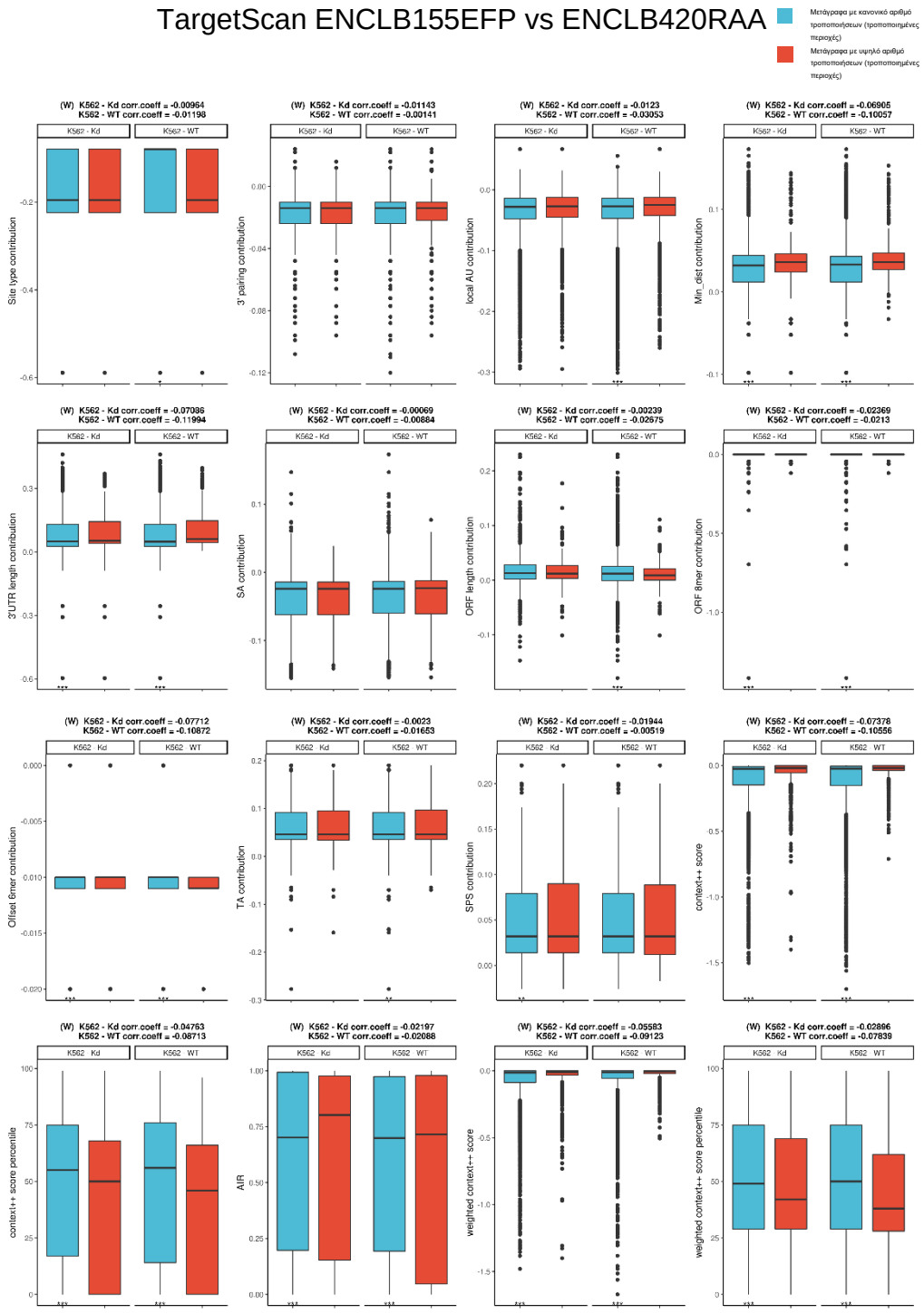
Targetscan	ERR188182 & ERR188298		ENCLB155EFP (K562 ADAR knockdown)		ENCLB420RAA (K562 ADAR wildtype expression)	
	p-value	Effect size	p-value	Effect size	p-value	Effect size
Site type	0.8177	-0.00225	0.76839	-0.00582	0.70533	-0.00576
3' pairing	0.97744	-0.00028	0.88423	-0.00288	0.47191	-0.01095
local AU	0.00204	-0.0301	0.64642	-0.00906	0.17976	-0.02042
Min_dist	0.69751	-0.00379	0.87157	-0.00319	0.78054	-0.00424
sRNA1A	0.86409	-0.00167	0.96311	-0.00091	0.9599	-0.00077
sRNA1C	0.34749	-0.00917	0.86174	-0.00344	0.75614	-0.00473
sRNA1G	NA	NA	0.98125	-0.00046	0.97687	-0.00044
sRNA8A	0.86142	-0.0017	0.78367	-0.00542	0.63854	-0.00715
sRNA8C	0.99907	-0.00001	0.95493	-0.00112	0.86546	-0.00258
sRNA8G	0.0913	-0.01648	0.78893	-0.00529	0.96528	-0.00066
site8A	0.42466	-0.00779	0.93581	-0.00159	0.96067	-0.00075
site8C	0.32775	-0.00955	0.536	-0.01222	0.32644	-0.01494
site8G	0.69804	-0.00379	0.92195	-0.00194	0.71454	-0.00557
3'UTR length	0.90885	-0.00112	0.79385	-0.00516	0.73707	-0.00511
SA	0.68762	-0.00392	0.82436	-0.00438	0.8057	-0.00374
ORF length	0.75527	-0.00304	0.88085	-0.00296	0.98586	-0.00027
ORF 8mer	0.83645	-0.00201	0.97418	-0.00064	0.97077	-0.00056
Offset 6mer	0.85836	-0.00174	0.82931	-0.00426	0.67536	-0.00638
TA	0.35153	-0.00909	0.87809	-0.00303	0.90053	-0.0019
SPS	0.59662	-0.00516	0.79697	-0.00508	0.63981	-0.00712
context++ score	0.3932	-0.00833	0.71436	-0.00723	0.54021	-0.00933
context++ score percentile	0.73507	-0.0033	0.74918	-0.00632	0.69622	-0.00594
AIR	0.35935	-0.00894	0	-0.0935	0.81451	-0.00357
weighted context++ score	0.39294	-0.00834	0.0082	-0.05222	0.5973	-0.00804
weighted context++ score percentile	0.19562	-0.01263	0.50605	-0.01314	0.09978	-0.02506

Τα box-and-whiskers plots της σύγκρισης των 3'UTR με υψηλό αριθμό τροποποιήσεων στην τροποποιημένη και μη μορφή τους δεν παρουσιάζουν στατιστικά σημαντικές διαφορές, με εξαίρεση τη διαφορά στο χαρακτηριστικό AIR στο δείγμα ENCLB155EFP,

που οδηγεί σε μια στατιστικά ασήμαντη μείωση του weighted context++ score στις τροποποιημένες περιοχές.



Σχήμα 28. Box-and-whiskers plots της ένωσης των περιοχών πρόσδεσης (TargetScan) των δειγμάτων ERR188182 και ERR188298 μεταξύ των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων (κόκκινο) έναντι των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται στα υπόλοιπα τροποποιημένα γονίδια (μπλε).



Σχήμα 29. **Box-and-whiskers plots** της ένωσης των περιοχών πρόσδεσης (**TargetScan**) των δειγμάτων **ENCLB**155**EFP** και **ENCLB**420**RAA** ξεχωριστά, μεταξύ των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων (κόκκινο) έναντι των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται στα υπόλοιπα τροποποιημένα γονίδια (μπλε).

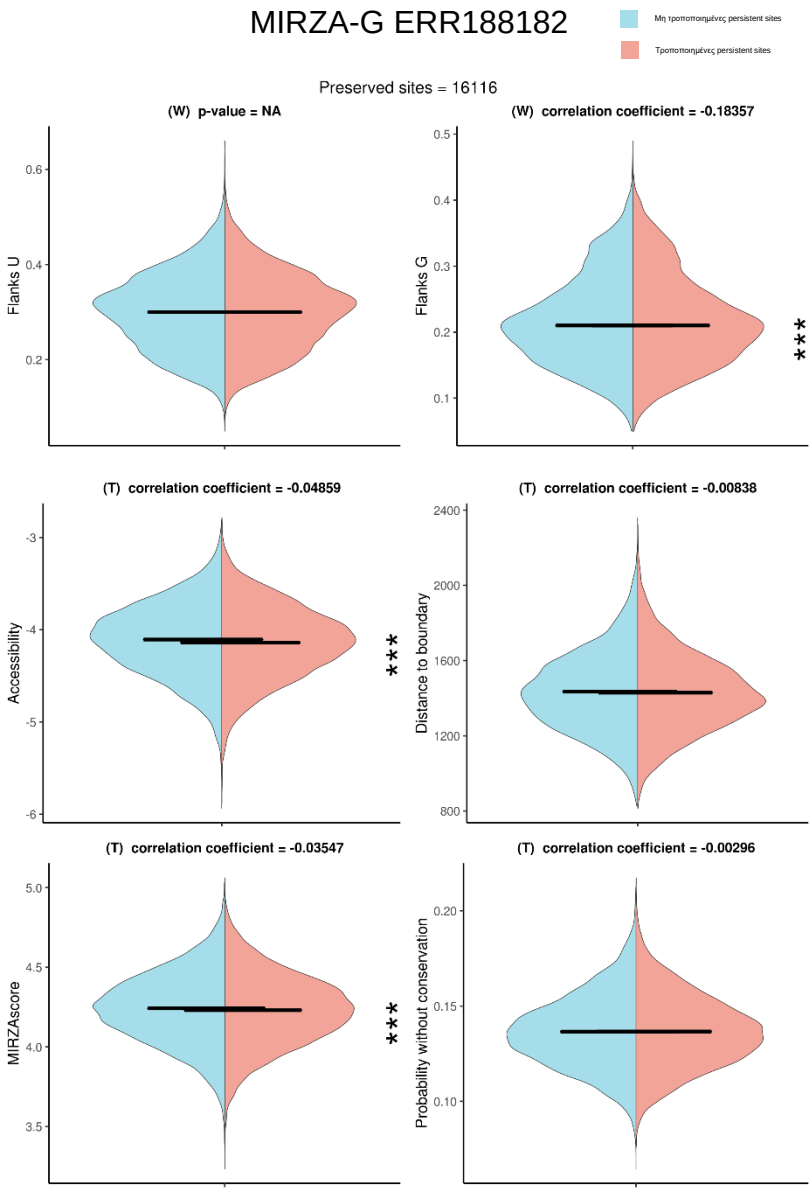
Πίνακας 8. Τιμή του p-value και του effect size (correlation coefficient) κάθε χαρακτηριστικού που προέκυψε από τη στατιστική δοκιμασία μεταξύ των τροποποιημένων περιοχών πρόσδεσης (TargetScan) που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων έναντι των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται στα υπόλοιπα τροποποιημένα γονίδια.

Targetscan	ERR188182 & ERR188298		ENCLB155EFP (K562 ADAR knockdown)		ENCLB420RAA (K562 ADAR wildtype expression)	
	p-value	Effect size	p-value	Effect size	p-value	Effect size
Site type	0.0001	-0.01388	0.25118	-0.00964	0.06509	-0.01198
3' pairing	0.00159	-0.0113	0.17351	-0.01143	0.82761	-0.00141
local AU	0	-0.02663	0.14286	-0.0123	0	-0.03053
Min_dist	0	-0.06712	0	-0.06905	0	-0.10057
sRNA1A	0	-0.02433	0.4956	-0.00572	0.00614	-0.0178
sRNA1C	0.69099	-0.00142	0.65087	-0.0038	0.00048	-0.0227
sRNA1G	NA	NA	0.47125	-0.00605	0.84574	-0.00126
sRNA8A	0.03056	-0.00774	0.27893	-0.00909	0.61644	-0.00325
sRNA8C	0	-0.01836	0.97423	-0.00027	0	-0.03658
sRNA8G	0.62971	-0.00172	0.43681	-0.00653	0.07194	-0.01169
site8A	0.73727	-0.0012	0.81248	-0.00199	0.44109	-0.00501
site8C	0.09598	-0.00596	0.56639	-0.00481	0.91905	-0.00066
site8G	0.51813	-0.00231	0.34759	-0.00789	0.74834	-0.00208
3'UTR length	0	-0.07816	0	-0.07086	0	-0.11994
SA	0.0323	-0.00766	0.93466	-0.00069	0.17357	-0.00884
ORF length	0	-0.02569	0.77618	-0.00239	0.00004	-0.02675
ORF 8mer	0.93056	-0.00031	0.00479	-0.02369	0.00104	-0.0213
Offset 6mer	0	-0.04475	0	-0.07712	0	-0.10872
TA	0.00929	-0.00931	0.78445	-0.0023	0.01095	-0.01653
SPS	0.01926	-0.00837	0.02059	-0.01944	0.42402	-0.00519
context++ score	0	-0.04827	0	-0.07378	0	-0.10556
context++ score percentile	0	-0.04823	0	-0.04763	0	-0.08713
AIR	0	-0.03292	0.0089	-0.02197	0.00131	-0.02088
weighted context++ score	0.00011	-0.01383	0	-0.05583	0	-0.09123
weighted context++ score percentile	0.0022	-0.01095	0.00056	-0.02896	0	-0.07839

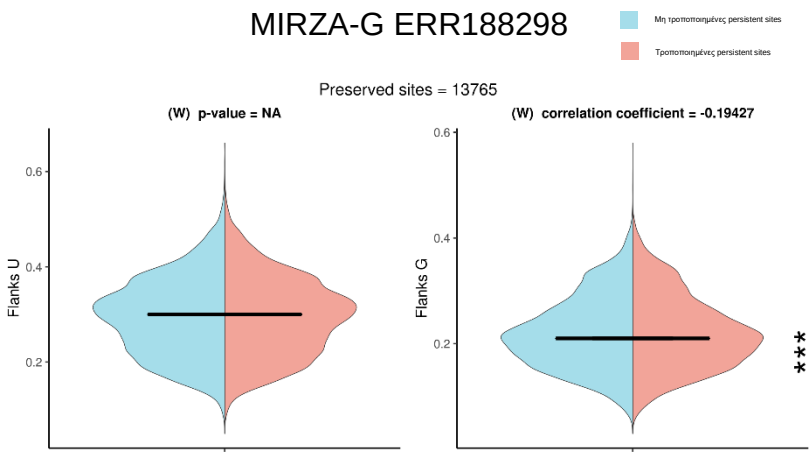
Στη σύγκριση της κατηγορίας των 3'UTR με πολλές τροποποιήσεις έναντι της κατηγορίας με κανονικό αριθμό τροποποιήσεων, εμφανίζονται περισσότερες στατιστικά σημαντικές διαφορές. Η ένταση των διαφορών είναι μεγαλύτερη στα δείγματα εκτός του ADAR knockdown. Το min_dist παρουσιάζει στατιστικά σημαντική αύξηση στην κατηγορία των τροποποιημένων περιοχών για το ENCLB420RAA, ενώ και τα υπόλοιπα δείγματα παρουσιάζουν σε πολύ ήπια ένταση την ίδια τάση. Ίδια αποτελέσματα παρουσιάζει και η σύγκριση του 3'UTR length. Στατιστικά σημαντική διαφορά παρατηρείται επίσης στο χαρακτηριστικό offset 6mer. Τα context++ score και weighted context++ score παρουσιάζουν μια τάση αύξησης σε όλα τα δείγματα, η οποία όμως είναι στατιστικά σημαντική μόνο για το ENCLB420RAA, ενώ το ENCLB155EFP παρουσιάζει ήπια αυξημένο effect size μόνο για το context++ score.

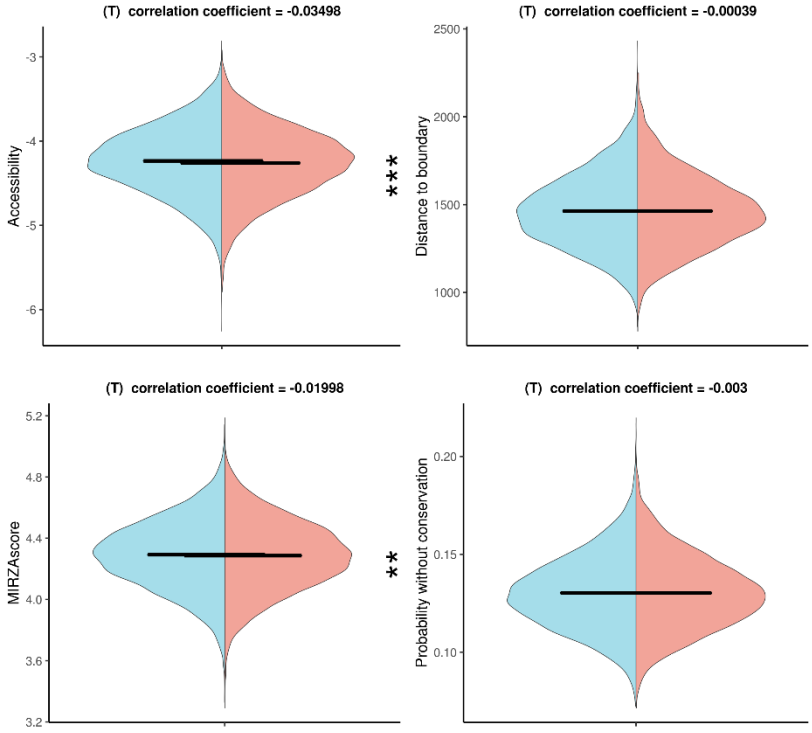
3.2.2 MIRZA-G

Η ίδια διαδικασία ακολουθήθηκε και για το MIRZA-G. Τα διαγράμματα παρουσιάζονται ακολούθως.

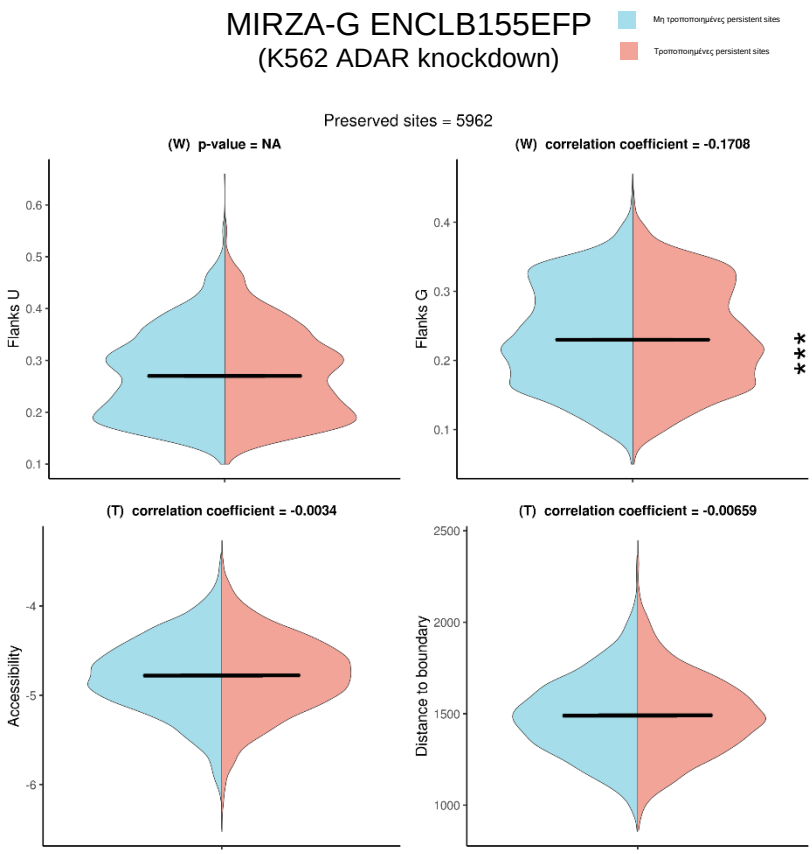


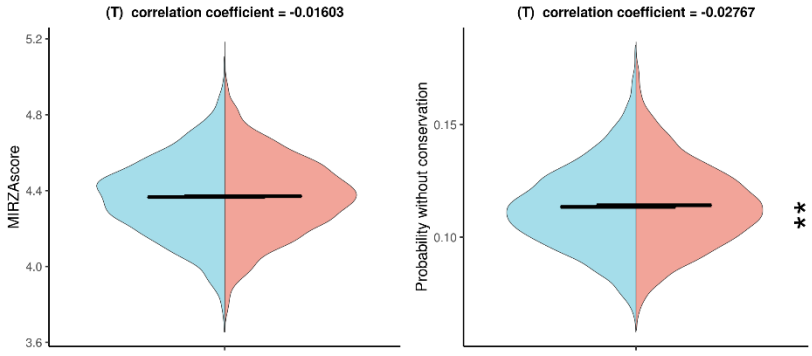
Σχήμα 30. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το MIRZA-G για το δείγμα ERR188182.



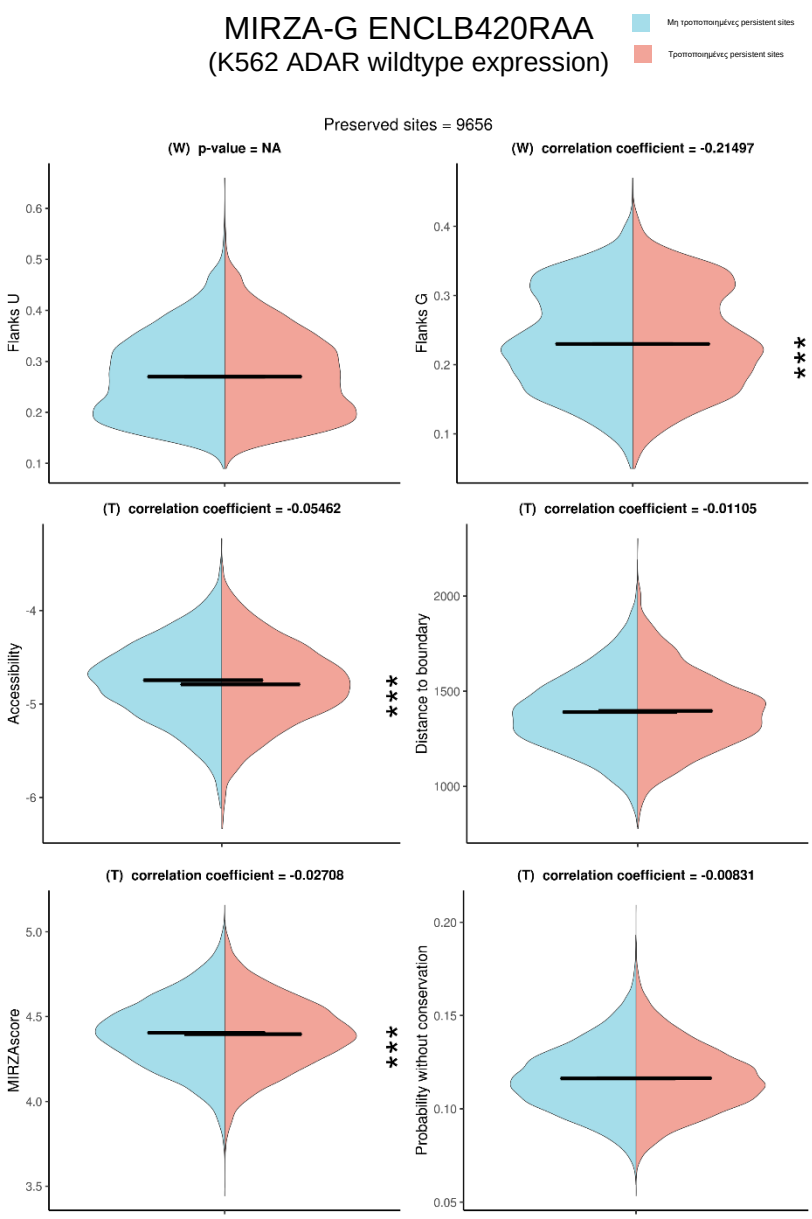


Σχήμα 31. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το MIRZA-G για το δείγμα ERR188298.





Σχήμα 32. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το MIRZA-G για το δείγμα ENCLB155EFP.

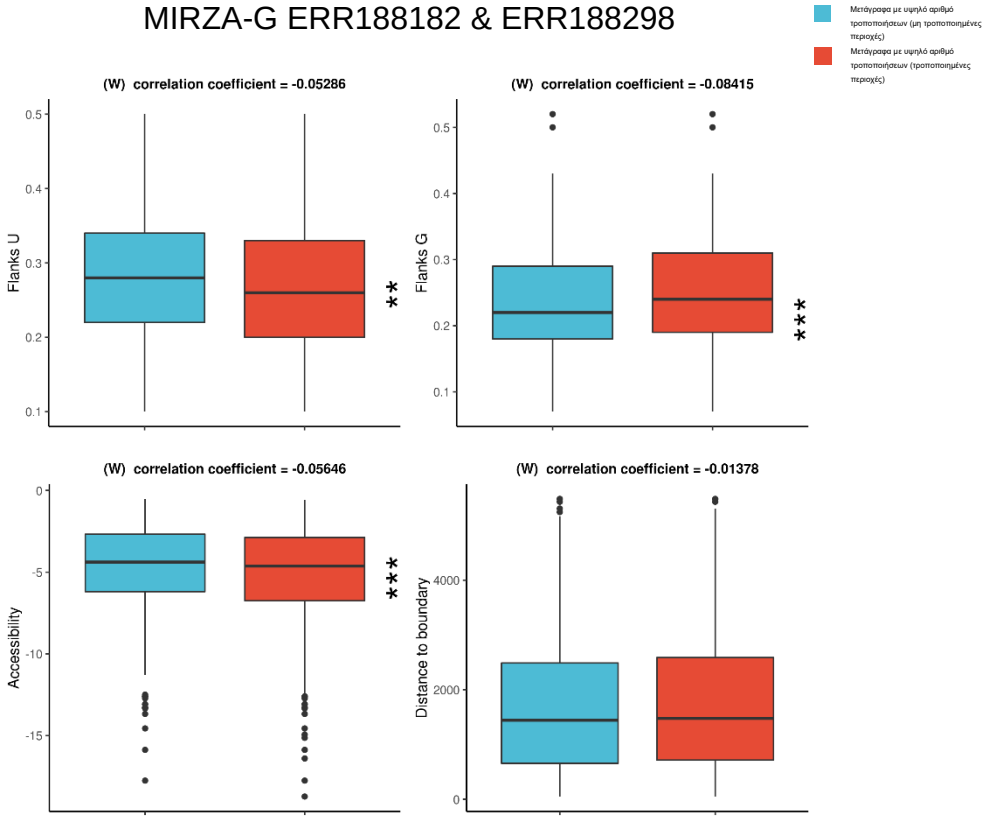


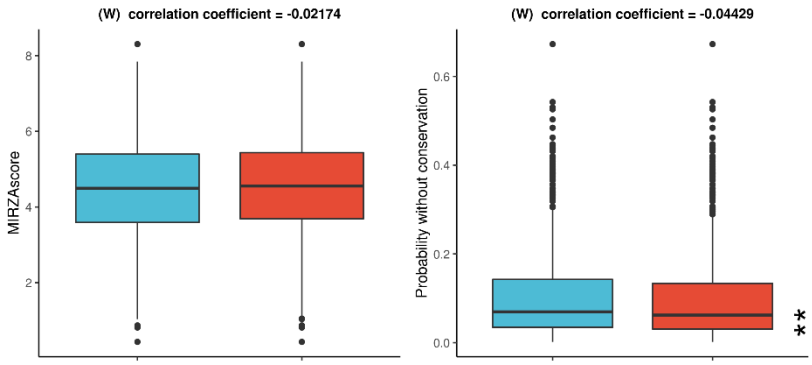
Σχήμα 33. Split violin plot για τα χαρακτηριστικά στις τροποποιημένες (μπλε)/μη-τροποποιημένες (κόκκινο) persistent sites που πρόβλεψε το MIRZA-G για το δείγμα ENCLB420RAA.

Πίνακας 9. Τιμή του p-value και του effect size (correlation coefficient) κάθε χαρακτηριστικού που προέκυψε από τη στατιστική δοκιμασία μεταξύ των τροποποιημένων/μη-τροποποιημένων περιοχών (MIRZA-G) των box-and-whiskers plots κάθε δείγματος.

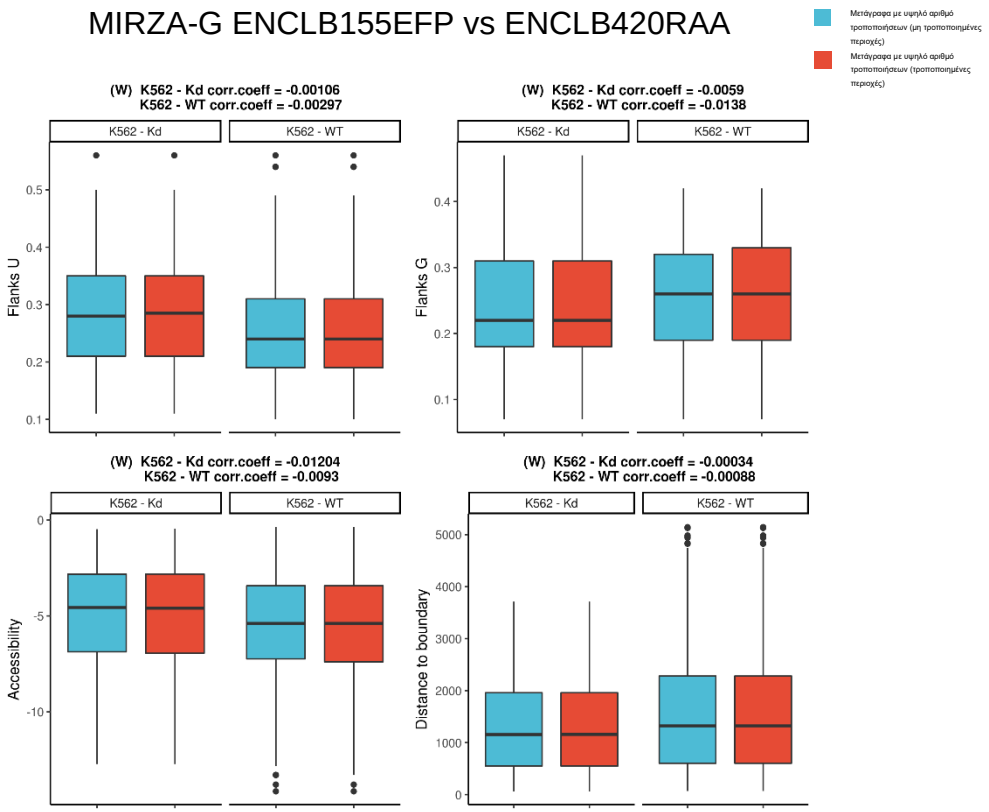
MIRZA-G	ERR188182		ERR188298		ENCLB155EFP (K562 ADAR knockdown)		ENCLB420RAA (K562 ADAR wildtype expression)	
	p-value	Effect size	p-value	Effect size	p-value	Effect size	p-value	Effect size
Flanks U	NA	NA	NA	NA	NA	NA	NA	NA
Flanks G	0	-0.18357	0	-0.19427	0	-0.1708	0	-0.21497
Accessibility	0	-0.04859	0.00004	-0.03498	0.79296	-0.0034	0	-0.05462
Distance to boundary	0.28712	-0.00838	0.96322	-0.00039	0.61071	-0.00659	0.27761	-0.01105
MIRZAscore	0.00001	-0.03547	0.01904	-0.01998	0.21582	-0.01603	0.00779	-0.02708
Probability without conservation	0.70698	-0.00296	0.72475	-0.003	0.03266	-0.02767	0.41393	-0.00831

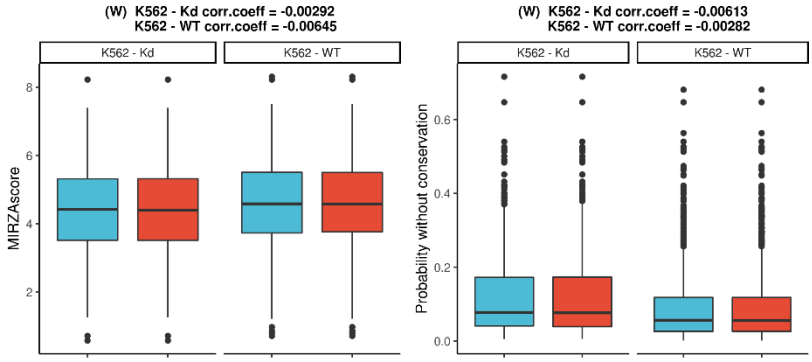
Στη σύγκριση για το MIRZA-G των persisting sites, το μοναδικό στατιστικά σημαντικό φαινόμενο παρατηρείται στη διαφορά σε όλα τα δείγματα για το χαρακτηριστικό Flanks G. Ήπια μείωση παρατηρείται στο χαρακτηριστικό Accessibility όλων των δειγμάτων, πλην του ENCLB155EFP, του οποίου η μείωση είναι στατιστικά ασήμαντη. Πολύ ήπια μείωση παρατηρείται και στο MIRZAscore των τροποποιημένων 3'UTR για τα ίδια δείγματα, ενώ το ENCLB155EFP παρουσιάζει το αντίθετο φαινόμενο, το οποίο αποτυπώνεται στο ίδιο δείγμα με παρόμοια ένταση και στο χαρακτηριστικό probability without conservation.





Σχήμα 34. Box-and-whiskers plots της ένωσης των περιοχών πρόσδεσης (MIRZA-G) των δειγμάτων ERR188182 και ERR188298 που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων. Για κάθε χαρακτηριστικό γίνεται σύγκριση μεταξύ των μη τροποποιημένων (μπλε) έναντι των τροποποιημένων (κόκκινο) περιοχών.



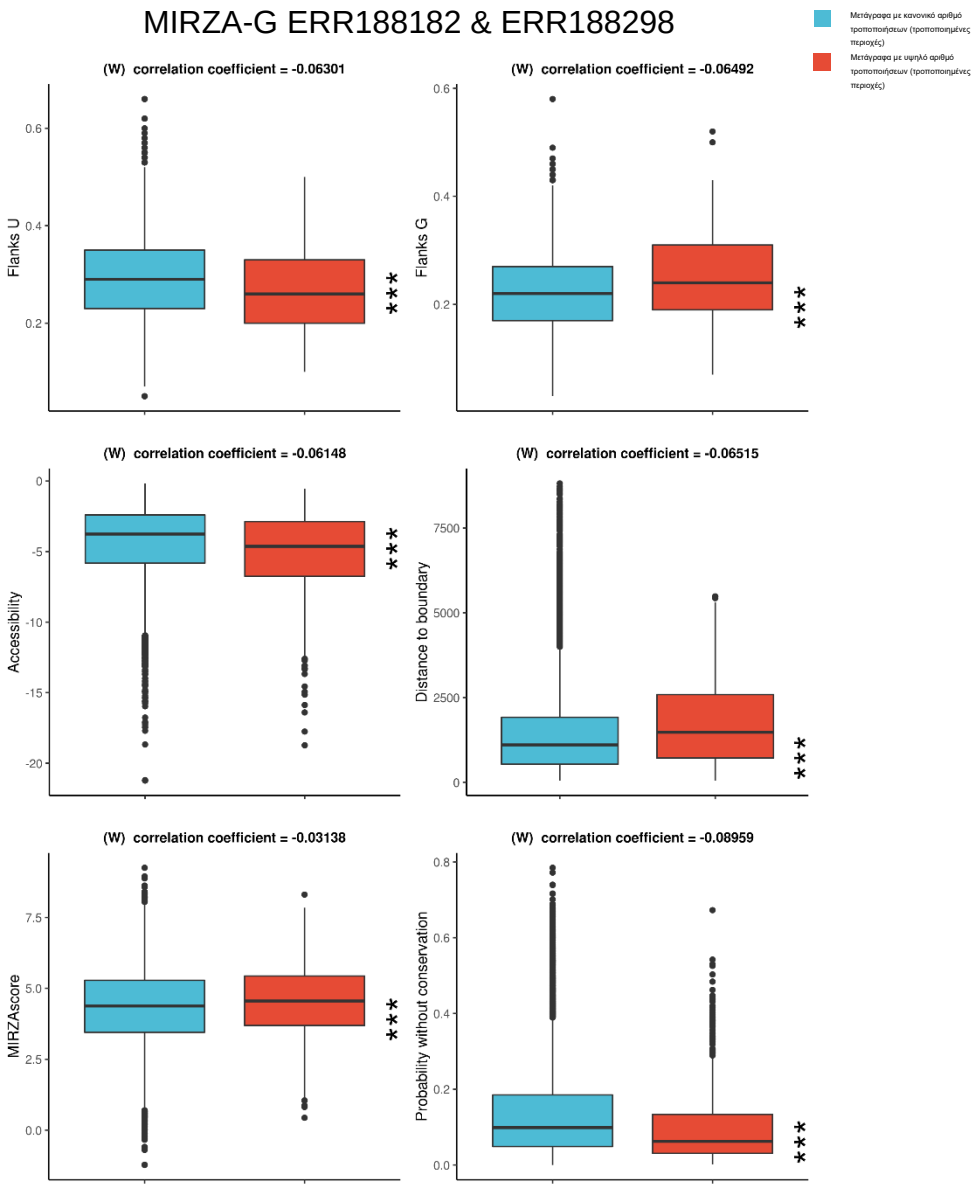


Σχήμα 35. Box-and-whiskers plots των περιοχών πρόσδεσης (MIRZA-G) των δειγμάτων ENCLB155EFP και ENCLB420RAA ξεχωριστά, που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων. Για κάθε χαρακτηριστικό γίνεται σύγκριση μεταξύ των μη τροποποιημένων (μπλε) έναντι των τροποποιημένων (κόκκινο) περιοχών κάθε δείγματος.

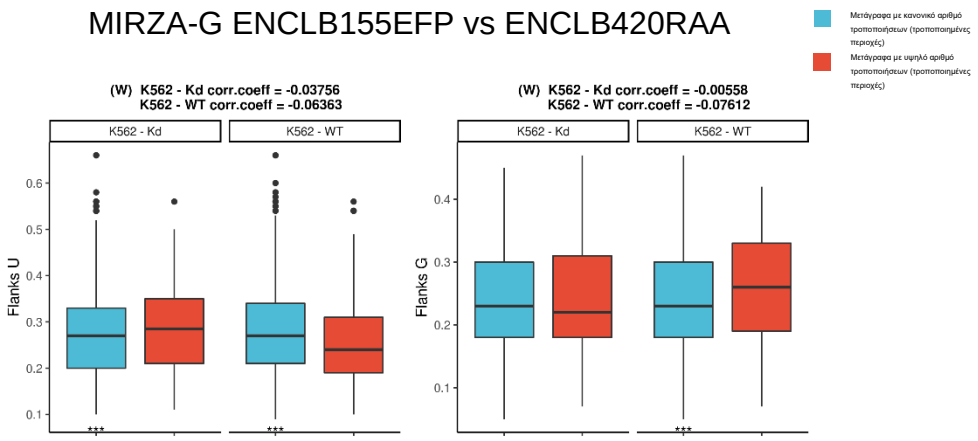
Πίνακας 10. Τιμή του p-value και του effect size (correlation coefficient) κάθε χαρακτηριστικού που προέκυψε από τη στατιστική δοκιμασία μεταξύ των τροποποιημένων/μη-τροποποιημένων περιοχών (MIRZA-G) των box-and-whiskers plots κάθε δείγματος.

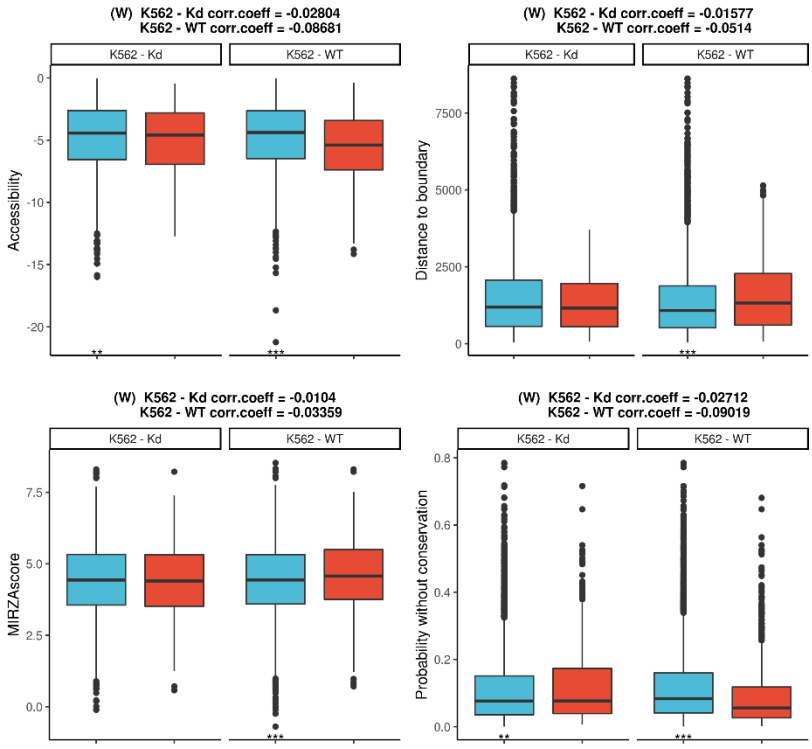
MIRZA-G	ERR188182 & ERR188298		ENCLB155EFP (K562 ADAR knockdown)		ENCLB420RAA (K562 ADAR wildtype expression)	
	p-value	Effect size	p-value	Effect size	p-value	Effect size
Flanks U	0.01288	-0.05286	0.97306	-0.00106	0.90083	-0.00297
Flanks G	0,00008	-0.08415	0.85035	-0.0059	0.56284	-0.0138
Accessibility	0.00789	-0.05646	0.70052	-0.01204	0.69663	-0.0093
Distance to boundary	0.51683	-0.01378	0.99145	-0.00034	0.97043	-0.00088
MIRZAscore	0.30637	-0.02174	0.92559	-0.00292	0.78668	-0.00645
Probability without conservation	0.03714	-0.04429	0.84468	-0.00613	0.90583	-0.00282

Όπως και στην περίπτωση του TargetScan, από τις συγκρίσεις των 3'UTR με υψηλό αριθμό τροποποιήσεων στην τροποποιημένη και μη μορφή τους δεν προέκυψαν στατιστικά σημαντικές διαφορές. Όπως φαίνεται από τα σχήματα, οι κατηγορίες για τα δείγματα ENCLB155EFP και ENCLB420RAA κινούνται στο ίδιο εύρος τιμών, ενώ τα ERR παρουσιάζουν στατιστικά σημαντική αύξηση για τις τροποποιημένες 3'UTR στο Flanks G. Στα ίδια δείγματα, παρατηρείται ήπια μείωση στα Flanks U και Accessibility.



Σχήμα 36. Box-and-whiskers plots της ένωσης των περιοχών πρόσδεσης (MIRZA-G) των δειγμάτων ERR188182 και ERR188298 μεταξύ των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων (κόκκινο) έναντι των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται στα υπόλοιπα τροποποιημένα γονίδια (μπλε).





Σχήμα 37. Box-and-whiskers plots της ένωσης των περιοχών πρόσδεσης (MIRZA-G) των δειγμάτων ENCLB155EFP και ENCLB420RAA ξεχωριστά, μεταξύ των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων (κόκκινο) έναντι των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται στα υπόλοιπα τροποποιημένα γονίδια (μπλε).

Πίνακας 11. Τιμή του p-value και του effect size (correlation coefficient) κάθε χαρακτηριστικού που προέκυψε από τη στατιστική δοκιμασία μεταξύ των τροποποιημένων περιοχών πρόσδεσης που βρίσκονται σε γονίδια με υψηλό αριθμό τροποποιήσεων έναντι των τροποποιημένων περιοχών πρόσδεσης (MIRZA-G) που βρίσκονται στα υπόλοιπα τροποποιημένα γονίδια.

MIRZA-G	ERR188182 & ERR188298		ENCLB155EFP (K562 ADAR knockdown)		ENCLB420RAA (K562 ADAR wildtype expression)	
	p-value	Effect size	p-value	Effect size	p-value	Effect size
Flanks U	0	-0.06301	0.00373	-0.03756	0	-0.06363
Flanks G	0	-0.06492	0.6667	-0.00558	0	-0.07612
Accessibility	0	-0.06148	0.03036	-0.02804	0	-0.08681
Distance to boundary	0	-0.06515	0.22328	-0.01577	0	-0.0514
MIRZAscore	0.00001	-0.03138	0.42203	-0.0104	0.00096	-0.03359
Probability without conservation	0	-0.08959	0.03623	-0.02712	0	-0.09019

Και στην περίπτωση των συγκρίσεων της κατηγορίας των 3'UTR με πολλές τροποποιήσεις έναντι της κατηγορίας με κανονικό αριθμό τροποποιήσεων για τα τέσσερα δείγματα με το MIRZA-G δεν εμφανίζονται στατιστικά σημαντικές διαφορές. Μείωση παρατηρείται στο χαρακτηριστικό probability without conservation για τις τροποποιημένες περιοχές των δειγμάτων ERR και ENCLB420RAA, με τις τιμές των p-value και effect size

να προσεγγίζουν τη στατιστική σημαντικότητα. Παρατηρώντας τα σχήματα και τις τιμές του effect size, διακρίνεται επίσης μια τάση των ίδιων δειγμάτων για μείωση στα χαρακτηριστικά Flanks U και Accessibility και αύξηση για τα χαρακτηριστικά Flanks G, Distance to boundary και MIRZAscore. Το δείγμα ENCLB155EFP με πολύ ήπιες διαφορές, στατιστικά αμελητέες, παρουσιάζει αύξηση στο Flanks U και μείωση στα Flanks G, Accessibility και Distance to boundary.

3.3 Σύνοψη

Συνοψίζοντας, στην παρούσα μελέτη δεν εντοπίστηκε κάποια στατιστικά σημαντική επίδραση των τροποποιήσεων A-σε-I σε 3'UTR ως προς την αποτελεσματικότητα του συμπλόκου miRNA-mRNA. Πιο συγκεκριμένα, τα δείγματα ERR και το ADAR knockdown δείχνουν μια τάση προς μεγαλύτερη κατασταλτική δράση με το TargetScan, ενώ με το MIRZA-G μόνο το ADAR knockdown δείχνει την ίδια τάση, με τα υπόλοιπα να παραμένουν σταθερά στο συνολικό τους score. Το ADAR wildtype δείχνει αντίθετα μείωση της κατασταλτικής δράσης με το TargetScan. Συγκρίνοντας της 3'UTR με υψηλό αριθμό τροποποιήσεων στη μορφή τους με και χωρίς την τροποποίηση, δεν εντοπίστηκε κάποια διαφορά, ενώ στη σύγκριση των ίδιων περιοχών με τις υπόλοιπες τροποποιημένες περιοχές και οι 2 αλγόριθμοι υποδεικνύουν μια τάση για μείωση της κατασταλτικής δράσης των υψηλά τροποποιημένων περιοχών (το TargetScan σε όλα τα δείγματα, με διαφορετικές εντάσεις, το MIRZA-G σε όλα τα δείγματα εκτός του ADAR knockdown), η οποία όμως δεν επικυρώνεται στατιστικά.

Στους παρακάτω πίνακες παρουσιάζονται οι διαφορές σε επίπεδο μεταγράφου και γονιδίου που προέκυψαν μετά τις τροποποιήσεις. Για το TargetScan, ο υπολογισμός έγινε δύο φορές, μία για το context++ score και μία το weighted context++ score, ενώ για το MIRZA-G χρησιμοποιήθηκε το probability without conservation. Επίσης, στις 4 τελευταίες στήλες εξετάστηκε η ύπαρξη νέων μεταγράφων και γονιδίων που στοχεύονται μετά την τροποποίηση και αντίστοιχα μεταγράφων και γονιδίων που καταργήθηκαν λόγω της τροποποιήσεις, χωρίς όμως να προκύψει κάποια τέτοια περίπτωση.

Πίνακας 12. Αριθμητικά στοιχεία για την επίδραση των RNA τροποποιήσεων στις περιοχές πρόσδεσης των miRNA που πρόβλεψε το TargetScan. Κάθε κελί αναφέρει το πλήθος των περιοχών που ανήκουν στην κατηγορία της στήλης μετά την τροποποίηση. Η αποτελεσματικότητα υπολογίζεται ανά 3'UTR ως ο μέσος όρος όλων των περιοχών πρόσδεσης που έχουν προβλεφθεί σε αυτό, πριν και μετά τις τροποποιήσεις.

	TargetScan					
	Context++ score			Weighted context++ score		
	Πιο αποτελεσματικές 3'UTR περιοχές	Λιγότερο αποτελεσματικές 3'UTR περιοχές	Ανεπηρέαστες 3'UTR περιοχές	Πιο αποτελεσματικές 3'UTR περιοχές	Λιγότερο αποτελεσματικές 3'UTR περιοχές	Ανεπηρέαστες 3'UTR περιοχές
ERR188182	6	237	545	52	252	484
ERR188298	13	223	394	50	218	362
K562 ADAR knockdown	0	30	207	15	35	187
K562 ADAR wildtype	6	77	346	37	104	288

	Νέα μετάγραφα που στοχεύονται μετά τις τροποποιήσεις	Κατηργημένα μετάγραφα που δεν στοχεύονται μετά τις τροποποιήσεις	Νέα γονίδια που στοχεύονται μετά τις τροποποιήσεις	Κατηργημένα γονίδια που δεν στοχεύονται μετά τις τροποποιήσεις
ERR188182	0	0	0	0
ERR188298	0	0	0	0
K562 ADAR knockdown	0	0	0	0
K562 ADAR wildtype	0	0	0	0

Πίνακας 13. Αριθμητικά στοιχεία για την επίδραση των RNA τροποποιήσεων στις περιοχές πρόσδεσης των miRNA που πρόβλεψε το MIRZA-G. Κάθε κελί αναφέρει το πλήθος των περιοχών που ανήκουν στην κατηγορία της στήλης μετά την τροποποίηση. Η αποτελεσματικότητα υπολογίζεται ανά 3'UTR ως ο μέσος όρος όλων των περιοχών πρόσδεσης που έχουν προβλεφθεί σε αυτό, πριν και μετά τις τροποποιήσεις.

	MIRZA-G		
	Probability without conservation		
	Πιο αποτελεσματικές 3'UTR περιοχές	Λιγότερο αποτελεσματικές 3'UTR περιοχές	Ανεπηρέαστες 3'UTR περιοχές
ERR188182	126	152	386
ERR188298	118	142	268
K562 ADAR knockdown	35	26	130
K562 ADAR wildtype	65	59	218

	Νέα μετάγραφα που στοχεύονται μετά τις τροποποιήσεις	Κατηργημένα μετάγραφα που δεν στοχεύονται μετά τις τροποποιήσεις	Νέα γονίδια που στοχεύονται μετά τις τροποποιήσεις	Κατηργημένα γονίδια που δεν στοχεύονται μετά τις τροποποιήσεις
ERR188182	0	0	0	0
ERR188298	0	0	0	0
K562 ADAR knockdown	0	0	0	0
K562 ADAR wildtype	0	0	0	0

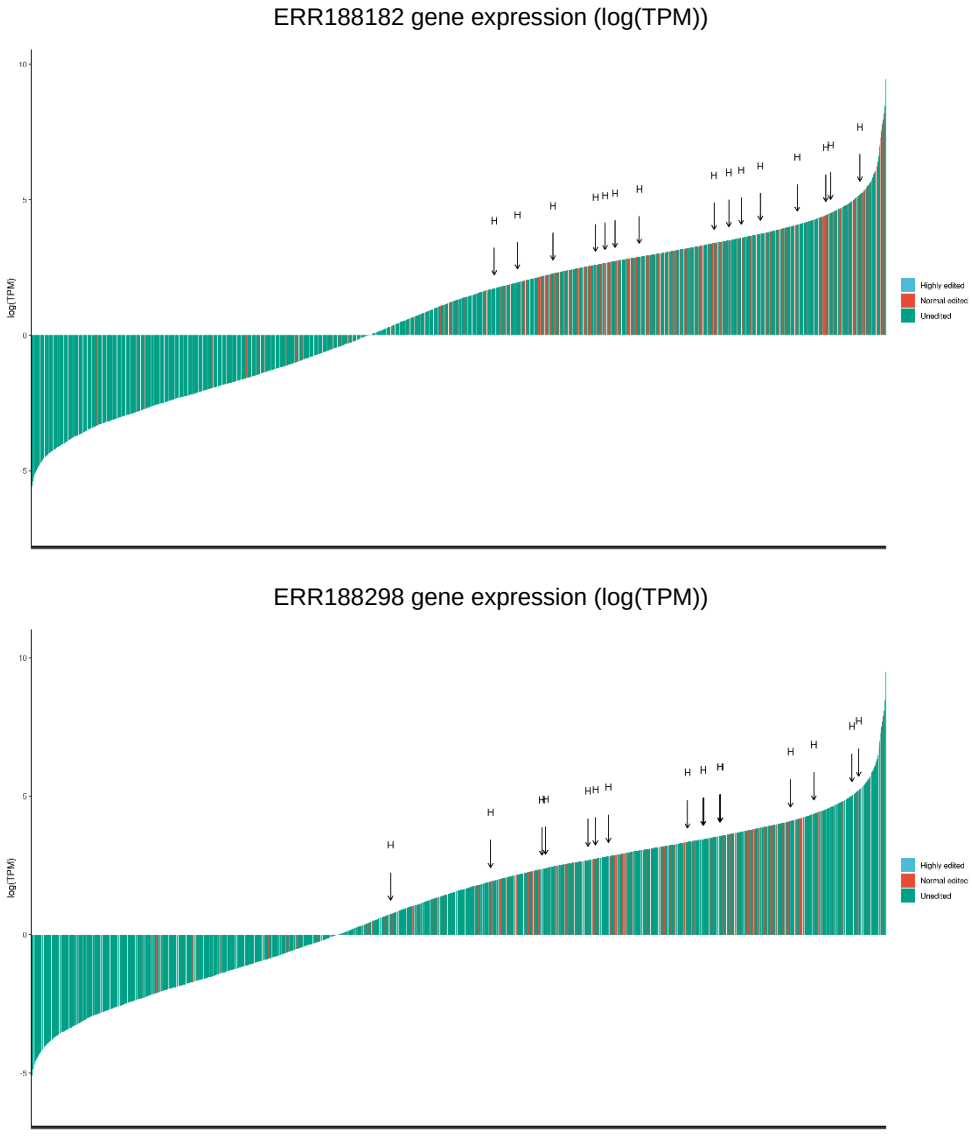
Τα στατιστικά του TargetScan δείχνουν υπερίσχυση της μείωσης της κατασταλτικής δράσης στις 3'UTR συγκριτικά με την αύξηση, χωρίς όμως να εκμηδενίζει το αντίθετο φαινόμενο. Στο MIRZA-G, τα δείγματα είναι χωρισμένα σε 2 σύνολα, τα οποία παρουσιάζουν αντίθετα φαινόμενα, με τις κατηγορίες όμως και στα 2 να βρίσκονται αριθμητικά κοντά.

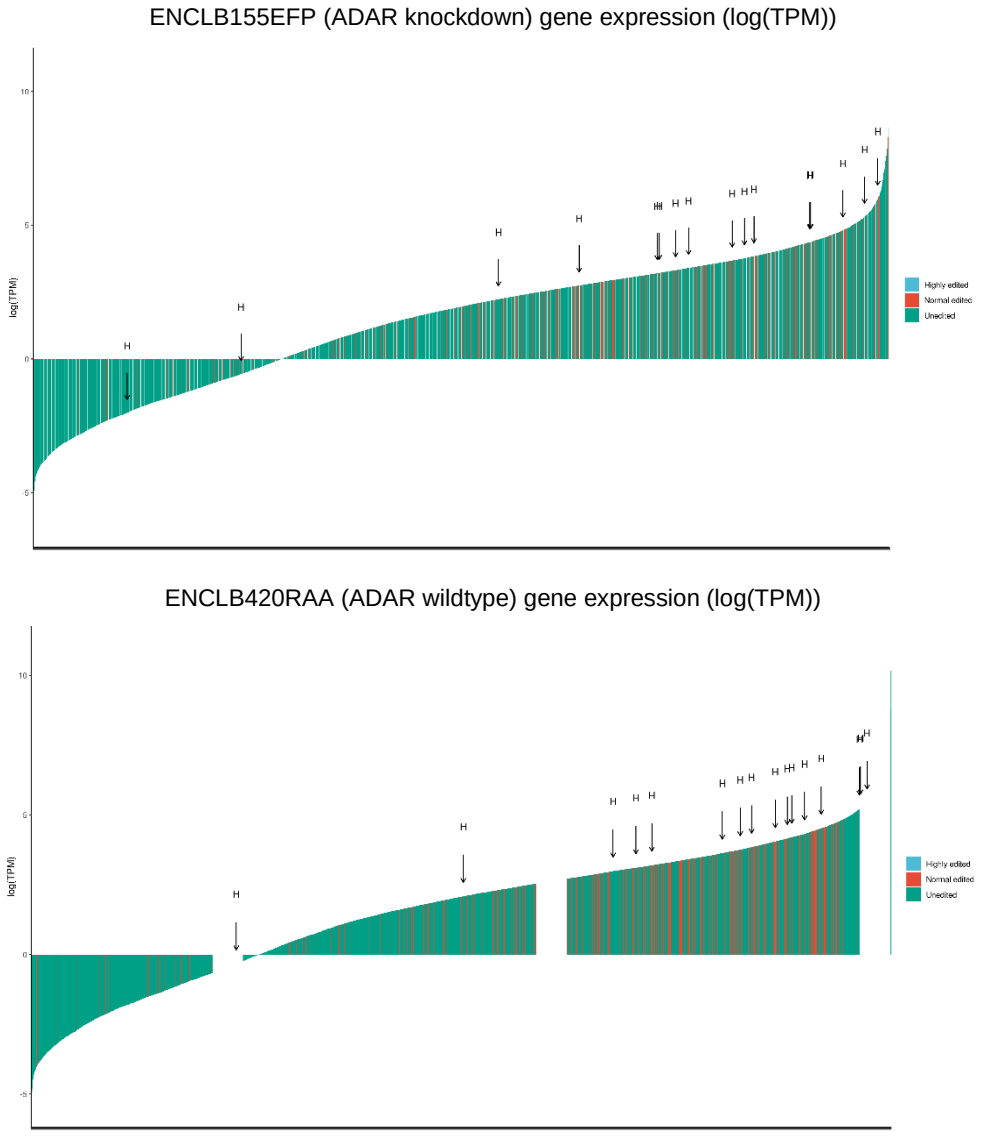
Λαμβάνοντας υπόψιν τις τιμές των πινάκων και τα σχήματα που παρουσιάστηκαν, διαφαίνεται πως η τροποποίηση ασκεί έναν ήπιο ρυθμιστικό ρόλο στις περιοχές πρόσδεσης των miRNA. Κατά περίπτωση, ισχυροποιείται ο μηχανισμός της καταστολής των miRNA, όμως δεν παρατηρείται συνολικά τάση προς την αύξηση ή μείωση της

καταστολής και δεν αναδεικνύονται έντονες και άμεσες αλλαγές στο φαινόμενο της στόχευσης που να οφείλονται στην τροποποίηση. Η επέκταση της μελέτης με εισαγωγή περισσότερων δειγμάτων θα φανεί χρήσιμη, επιφέροντας μεγαλύτερη στατιστική δύναμη στα αποτελέσματα και ενδεχομένως οδηγώντας την υπόθεση για το ρόλο της τροποποίησης στη σωστή κατεύθυνση.

3.4 Ανάλυση γονιδιακής έκφρασης

Τέλος, παρατίθενται τα σχήματα από την ποσοτικοποίηση της έκφρασης των τροποποιημένων γονιδίων που επιτεύχθηκε με το Salmon. Έχουν χρησιμοποιηθεί 3 κατηγορίες γονιδίων, με το ένα σύνολο να περιέχει τα πρώτα 20 γονίδια σε αριθμό τροποποιήσεων (γαλάζιο), το δεύτερο να περιέχει τα υπόλοιπα τροποποιημένα γονίδια (κόκκινο) και το τρίτο να περιέχει αυτά στα οποία δεν εντοπίστηκε καμία τροποποίηση (πράσινο). Οι τιμές ανά γονίδιο έχουν ταξινομηθεί από τη μικρότερη στη μεγαλύτερη και έχουν χρωματιστεί διαφορετικά, αναλόγως με την κατηγορία τοι πλήθους των τροποποιήσεων που περιέχουν. Τα βελάκια υποδεικνύουν τη θέση των γονιδίων με υψηλό αριθμό τροποποιήσεων.





Σχήμα 38. Ιστογράμματα ανά δείγμα της γονιδιακής έκφρασης τριών κατηγοριών: γονίδια με υψηλό αριθμό τροποποιήσεων (γαλάζιο), υπόλοιπα τροποποιημένα γονίδια (κόκκινο) και γονίδια χωρίς τροποποίηση (πράσινα). Η ποσοτικοποίηση έγινε με το εργαλείο Salmon.

Από τα σχήματα φαίνεται ότι η πλειονότητα των τροποποιημένων γονιδίων βρίσκεται στην περιοχή της υψηλότερης έκφρασης, με τα φαινόμενα στις χαμηλότερες εκφράσεις να παρουσιάζονται πιο αραιά. Επίσης, παρατηρούμε ότι τα γονίδια με υψηλό αριθμό τροποποιήσεων κατανέμονται σε ένα μεγάλο εύρος τιμών, το οποίο φαίνεται να είναι ανεξάρτητο της συχνότητας των τροποποιήσεων. Επομένως, δεν μπορεί να εξαχθεί κάποιος συσχετισμός του αριθμού τροποποιήσεων με την έκφραση, ενώ το μικρό πλήθος στις χαμηλότερες εκφράσεις θα μπορούσε να οφείλεται στην αδυναμία των αλγορίθμων να εντοπίσουν τροποποιήσεις υπό τέτοιες συνθήκες, αναδεικνύοντας την ανάγκη για περαιτέρω βελτιώσεις σε αυτό το πεδίο. Περισσότερα δείγματα και εκτελέσεις θα επιτρέψουν να αποσαφηνιστεί αυτός ο ισχυρισμός.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία πραγματεύτηκε το φαινόμενο των τροποποιήσεων σε RNA ακολουθίες και πιο συγκεκριμένα των τροποποιήσεων A-σε-I (αδενοσίνη σε ινοσίνη). Τροποποιήσεις A-σε-I ταυτοποιήθηκαν υπολογιστικά και μελετήθηκε η επίδρασή τους στη στόχευση των microRNA στις μη μεταγραφόμενες περιοχές στο 3’ άκρο των messenger RNA. Για την επίτευξη αυτού του στόχου εκτελέστηκε η εξής σειρά βημάτων:

1. Σύγκριση 3 εργαλείων που εντοπίζουν RNA τροποποιήσεις και επιλογή του καλύτερου σε σχέση με την ακρίβεια και την ευαισθησία
2. Εκτέλεση 2 αλγορίθμων πρόβλεψης στόχων miRNA στις 3’UTR που εντοπίστηκαν να περιέχουν τροποποιήσεις
3. Στατιστική ανάλυση των αποτελεσμάτων

Η σύγκριση που έγινε μεταξύ των εργαλείων εντοπισμού RNA τροποποιήσεων ανέδειξε ως καλύτερο εργαλείο το RNAEditor (BWA), σε σχέση με την ακρίβεια και την ευαισθησία. Από τα αποτελέσματα αυτού του σταδίου διαφαίνονται οι αδυναμίες που χαρακτηρίζουν τα σημερινά εργαλεία αυτού του πεδίου και οι προκλήσεις που αντιμετωπίζουν. Οι περιοχές τροποποίησης δεν έχουν χαρακτηριστικά που τις ξεχωρίζουν με σαφήνεια από τις διάφορες αλληλομορφές και τα λάθη αλληλούχισης. Ως αποτέλεσμα, εισάγεται πλήθος από παραμέτρους και φίλτρα για τους αλγορίθμους, χωρίς όμως μέχρι στιγμής να έχει δημιουργηθεί μια gold standard στρατηγική στην οποία να συναινεί η επιστημονική κοινότητα. Αυτό φαίνεται και από τις διαφορετικές πρακτικές που εφαρμόζουν στο επίπεδο του alignment, του εντοπισμού αλλά και του φιλτραρίσματος τα 3 εργαλεία που συγκρίθηκαν. Μεγάλη προσοχή πρέπει να δίνεται στην επιλογή ποιοτικών δειγμάτων, αλλά και στο στάδιο της προεπεξεργασίας τους [56].

Οι διαφορές στις τροποποιήσεις που εντοπίστηκαν είναι μεγάλες, τόσο ως προς το πλήθος όσο και ως προς τις περιοχές που προβλέπονται, με τις τομές των εργαλείων να είναι μικρές μεταξύ τους. Χαρακτηριστική είναι η συμπεριφορά στην παρούσα μελέτη των εργαλείων ως προς τον χειρισμό των non-Alu περιοχών, αλλά και του ADAR knockdown δείγματος, στο οποίο οι μετρήσεις διαφέρουν σημαντικά μεταξύ τους. Επίσης, διαφοροποίηση υπήρξε και ως προς τις τροποποιήσεις που περιλαμβάνονται στη RADAR, προβάλλοντας την ανάγκη για την εύρεση καλύτερης λύσης ως ground truth. Τέλος, διαφορές παρατηρήθηκαν και μεταξύ των 2 aligners που χρησιμοποιήθηκαν για το RNAEditor, αναδεικνύοντας τη μεγάλη προσοχή που πρέπει να δίνεται και σε αυτό το βήμα.

Στο στάδιο της πρόβλεψης στόχων των miRNA, χρησιμοποιήθηκαν 2 εργαλεία που επιτελούν αυτόν τον σκοπό, το TargetScan και το MIRZA-G, ώστε να συνδυαστούν τα συμπεράσματα των αποτελεσμάτων τους. Και τα 2 εκτελέστηκαν χωρίς να ληφθούν υπόψιν τα εξελικτικά κριτήρια, καθώς η πληροφορία που εισάγουν δε μπορεί να χρησιμοποιηθεί σε συνδυασμό με την πληροφορία της τροποποίησης. Ως είσοδος στους αλγορίθμους δόθηκαν οι τροποποιήσεις που προέκυψαν από το προηγούμενο βήμα και που βρίσκονταν σε 3’UTR. Από αυτά, δημιουργήθηκαν 2 σύνολα ακολουθιών, ένα χωρίς τις τροποποιήσεις και ένα με αυτές, ώστε να γίνει σύγκριση μεταξύ των διαφορών στις περιοχές που θα προβλεφθούν για το καθένα.

Από τις παρατηρήσεις των σχημάτων και τις στατιστικές συγκρίσεις και μετρικές που εξάχθηκαν από τις εκτελέσεις, παρατηρείται μια ήπια μεταβολή της κατασταλτικής δράσης των miRNA που προσδένονται στις τροποποιημένες περιοχές, η οποία δεν παρουσιάζει συγκεκριμένη τάση αύξησης ή μείωσης της αποτελεσματικότητάς της και φαίνεται να δρα ως μηχανισμός βελτιστοποίησης κατά περίπτωση. Ο αριθμός των τροποποιήσεων δεν φαίνεται να σχετίζεται με το βαθμό αυξομείωσης.

Τέλος, εξετάστηκε η επίδραση των τροποποιήσεων A-σε-I στην έκφραση των γονιδίων, συγκρίνοντας τρείς διαφορετικές κατηγορίες γονιδίων: με υψηλό αριθμό τροποποιήσεων, με κανονικό αριθμό τροποποιήσεων και χωρίς καθόλου τροποποιήσεις. Η κατανομή των κατηγοριών φαίνεται να διαχωρίζεται τυχαία στο πεδίο των τιμών της έκφρασης, χωρίς να υποδεικνύεται κάποια συσχέτιση των τροποποιήσεων με αυτή, υπάρχει δε μια τάση για τροποποίηση σε υψηλά εκφραζόμενα μετάγραφα.

4.1 Μελλοντικοί στόχοι

Το πεδίο της μελέτης των RNA τροποποιήσεων είναι ανοιχτό, καθώς μόλις τα τελευταία χρόνια η βελτίωση των τεχνολογιών αλληλούχισης και των πειραματικών τεχνικών έχει επιτρέψει την ενδελεχή διερεύνησή του, οδηγώντας την επιστημονική κοινότητα σε νέα συμπεράσματα και υποθέσεις.

Πρωταρχικός στόχος για την εργασία στο άμεσο μέλλον αποτελεί η επέκταση του συνόλου των δειγμάτων, ώστε να επιτευχθεί η επιθυμητή στατιστική δύναμη για εξαγωγή ασφαλέστερων συμπερασμάτων. Ιδανικά πρέπει να συλλεχθούν strand-specific δείγματα, βαθιάς αλληλούχισης και από περισσότερους ιστούς, ώστε να γίνει ορθότερη η σύγκριση μεταξύ των εργαλείων. Επιπλέον, όπως ειπώθηκε και παραπάνω, δεν έχει βρεθεί ακόμα η gold standard υπολογιστική διαδικασία για εντοπισμό RNA τροποποιήσεων. Νέα εργαλεία δημοσιεύονται συνεχώς, ενώ παράλληλα βελτιώνονται τα ήδη υπάρχοντα, κάνοντας χρήση διαφορετικών aligners που φαίνεται να πετυχαίνουν μεγαλύτερη ευαισθησία και ακρίβεια στον εντοπισμό και εισάγοντας νέα χαρακτηριστικά. Για το λόγο αυτό, αναγκαία είναι η συνεχής αξιολόγησή τους, ώστε να επιλέγονται εργαλεία που βελτιώνουν τις στατιστικές μελέτες, όπως αυτή της παρούσας εργασίας.

Καίριο ζήτημα στην αξιολόγηση των εργαλείων εντοπισμού RNA τροποποιήσεων αποτελεί η εύρεση αξιόπιστης ground truth. Απαραίτητη είναι η χρήση βάσεων δεδομένων που περιέχουν εγγραφές τροποποιήσεων που έχουν προκύψει από πειραματικές τεχνικές, οι οποίες συνεχώς θα εμπλουτίζονται τα επόμενα χρόνια, προσδίδοντας μεγαλύτερη αξιοπιστία στα αποτελέσματα, καθώς και καλύτερη κατεύθυνση στην ανάπτυξη των αλγορίθμων εντοπισμού. Η χρήση control δειγμάτων μέσω προσομοίωσης των δεδομένων, στα οποία θα είναι γνωστές οι τροποποιήσεις, αποτελεί επίσης μια καλή εναλλακτική προσέγγιση.

Τέλος, πολύ σημαντικά είναι και δεδομένα που προέρχονται από τεχνικές οι οποίες περιορίζονται στον ακριβή, πειραματικό εντοπισμό RNA τροποποιήσεων, όπως τα inosine chemical erasing (ICE-seq) και τα crosslinking immunoprecipitation (CLIP-seq) πειράματα που στοχεύουν τις ADAR, και τα οποία μπορούν να αποτελέσουν τη βάση για τον καλύτερο σχεδιασμό αλγορίθμων πρόβλεψης και για τον εμπλουτισμό των βάσεων δεδομένων τροποποιήσεων με περαιτέρω αξιόπιστα δεδομένα.

Με τη συλλογή συνόλων δεδομένων από τις παραπάνω μεθοδολογίες, είναι δυνατή η δημιουργία ενός στατιστικά μεγάλου συνόλου, το οποίο θα επιτρέψει την εξαγωγή ασφαλέστερων συμπερασμάτων για την επίδραση των τροποποιήσεων στη στόχευση των miRNA. Η δημιουργία μιας υψηλής αξιοπιστίας ροής εργασιών για εντοπισμό των RNA τροποποιήσεων θα επιτρέψει την επέκταση της μελέτης σε άλλα πεδία, όπως η σύγκριση φυσιολογικών και παθολογικών καταστάσεων.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
1-based inclusive	μέτρηση της μορφής [1,…n]
a priori	εκ των προτέρων
adapter	Τεχνητή ακολουθία που προσδένεται στα θραύσματα DNA για να επιτευχθεί η αλληλούχιση
Aligner	πρόγραμμα στοίχισης ακολουθιών
alignment	στοίχιση
bias	μεροληψία, πληροφορία που λανθασμένα μεταβάλλει το ποσοστό εμφάνισης ενός φαινομένου σε βάρος του πραγματικού φαινομένου
configuration file	αρχείο που ορίζει τις παραμέτρους εκτέλεσης ενός προγράμματος
control	σύνολο δεδομένων ελέγχου
created sites	περιοχές πρόσδεσης οι οποίες δημιουργήθηκαν λόγω του φαινομένου τροποποίησης
default	τιμή αρχικοποίησης
editing island	χαρακτηρισμός που χρησιμοποιείται από το RNAEditor για γονιδιακές περιοχές με πολλαπλά, συνεχόμενα φαινόμενα τροποποίησης
effect size	μέγεθος που μετρά την επίδραση της διαφοράς ενός άλλου μεγέθους σε 2 σύνολα, ανεξαρτήτως του πλήθους μετρήσεων
false positive	ψευδώς/λανθασμένα θετικά
flags	ετικέτες
gold standard	ο χρυσός κανόνας στην εκτέλεση μιας διαδικασίας
ground truth	η αντικειμενική αλήθεια ως μέτρο σύγκρισης
Homo sapiens	άνθρωπος
host gene	γονίδιο ξενιστής
indexing	διαδικασία ταξινόμησης ενός συνόλου σε ευρύτερες κατηγορίες για αποδοτικότερη αναζήτηση
intron	ιντρόνιο
kit	σύνεργα για εκτέλεση μιας τεχνικής π.χ. αλληλούχιση
knockdown	αποσιώπηση
lost sites	περιοχές πρόσδεσης οι οποίες καταργήθηκαν λόγω του φαινομένου τροποποίησης
manual	εγχειρίδιο χρήσης
Max insertion size	η μεγαλύτερη απόσταση μεταξύ 2 reads
messenger	αγγελιοφόρο

microarray	μικροσυστοιχία, πλακίδιο που αποτελείται από μικρά πηγαδάκια και χρησιμοποιείται για μέτρηση της έκφρασης συγκεκριμένων μορίων
non-canonical	μη κανονικός
paired-end	δείγματα των οποίων τα reads έχουν αλληλουχηθεί 2 φορές με αντίθετο και αντίστροφο τρόπο, αποτελώντας ένα ζευγάρι
percentiles	ποσοστιαία κομμάτια ενός συνόλου π.χ. τεταρτημόρια 0-100% = 0-25,25-50,50-75,75-100
persistent sites	περιοχές πρόσδεσης οι οποίες υπήρχαν τόσο πριν όσο και μετά το φαινόμενο τροποποίησης
posterior probability	η πιθανότητα να συμβεί ένα γεγονός εκ των υστέρων
pre-aligned	στοιχισμένος εκ των προτέρων
precision	ακρίβεια
prior probability	η πιθανότητα να συμβεί ένα γεγονός εκ των προτέρων
reads	κομμάτια ακολουθίας από αλληλούχιση
repeat	στοιχείο που περιέχει επαναλήψεις μιας ακολουθίας
RNA editing	RNA τροποποίηση
RNA sequencing	αλληλούχιση του RNA
scaling factor	παράγοντας κανονικοποίησης
script	σύντομο πρόγραμμα που εκτελεί διαδικασίες περιορισμένου μεγέθους
seed region	περιοχή «σπόρου» (κρίσιμη περιοχή για τη στόχευση των miRNA)
sensitivity	ευαισθησία
short sequencing reads	κομμάτια αλληλούχισης με μικρό μέγεθος
splice/junction sites	περιοχές μεταγράφου στις οποίες θα γίνει μάτισμα
splice-aware	που λαμβάνει υπόψιν την πληροφορία για τις περιοχές ματίσματος
state-of-the-art	εργαλείο αιχμής σε κάποιο πεδίο
strand-specific	που φέρει την πληροφορία της έλικας προέλευσης
TAB-delimited	χωρισμένος με τον ειδικό χαρακτήρα TAB
test	δεδομένο δοκιμής
true positive	αληθινά θετικά
wild-type	φυσιολογική μορφή που συναντάται στη φύση
wobbles	μη κανονικοί νουκλεοτιδικοί δεσμοί
wrapper	πρόγραμμα που καλεί εσωτερικά άλλα προγράμματα

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

A	Αδενίνη/Αδενοσίνη
ADAR	Απαμινάσες της Αδενοσίνης που δρουν σε RNA
AIR	Ποσοστό Επηρεασμένων Ισομορφών
Alu	Arthrobacter luteus restriction endonuclease element
BAM	Binary Alignment Map
BWA	Burrows-Wheeler Aligner
C	Κυτοσίνη/Κυτιδίνη
C^6	άνθρακας στην έκτη θέση
CDS	Κωδική Περιοχή
CLIP-seq	Crosslinking Immunoprecipitation sequencing
DNA	δεσοξυριβοζονουκλεϊκό οξύ
DSH	Dyschromatosis Symmetrica Hereditaria
dsRNA	δίκλωνο RNA
FDR	False Discovery Rate
G	Γουανίνη/Γουανοσίνη
GATK	Genome Analysis Toolkit
GEO	Gene Expression Omnibus
GSNAP	Genomic Short-read Nucleotide Alignment Program
I	Ινοσίνη
ICE-seq	αλληλούχιση με Χημική Απαλοιφή της Ινοσίνης
lncRNA	μακρύ μη-κωδικό RNA
miRNA	microRNA
mRNA	αγγελιοφόρο RNA
NCBI	National Center for Biotechnology Information
ncRNA	μη-κωδικό RNA
NGS	Αλληλούχιση Επομένης Γενιάς
ORF	Open Reading Frame
PCR	Αλυσιδωτή Αντίδραση της Πολυμεράσης
preRNA	precursor RNA
priRNA	primary RNA
RISC	RNA Induced Silencing Complex
RNA	Ριβοζονουκλεϊκό οξύ
RNA-seq	αλληλούχιση RNA

rRNA	ριβοσωμικό RNA
shRNA	short hairpin RNA
SINEs	Short Interspersed Nuclear Elements
siRNA	small interfering RNA
SNP	Πολυμορφισμός Μεμονωμένου νουκλεοτιδίου
sRNA	short RNA
T	Θυμίνη/Θυμιδίνη
TPM	Transcripts Per Million
tRNA	μεταφορικό RNA
U	Ουρακίλη/Ουριδίνη
UTR	Μη Μεταφραζόμενη Περιοχή
VCF	Variant Call Format

ΑΝΑΦΟΡΕΣ

[1] F. Crick, “Central Dogma of Molecular Biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, doi: 10.1038/227561a0.

[2] L. D. Moore, T. Le, and G. Fan, “DNA Methylation and Its Basic Function,” *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, Jan. 2013, doi: 10.1038/npp.2012.112.

[3] K. Nishikura, “A-to-I editing of coding and non-coding RNAs by ADARs,” *Nat. Rev. Mol. Cell Biol.*, vol. 17, pp. 83–96, 2016, doi: 10.1038/nrm.2015.4.

[4] I. X. Wang, E. So, J. L. Devlin, Y. Zhao, M. Wu, and V. G. Cheung, “ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression,” *Cell Rep.*, vol. 5, no. 3, pp. 849–860, Nov. 2013, doi: 10.1016/j.celrep.2013.10.002.

[5] O. Solomon *et al.*, “Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR),” *RNA*, vol. 19, no. 5, pp. 591–604, May 2013, doi: 10.1261/rna.038042.112.

[6] S. M. Rueter, T. R. Dawson, and R. B. Emeson, “Regulation of alternative splicing by RNA editing,” *Nature*, vol. 399, no. 6731, pp. 75–80, May 1999, doi: 10.1038/19992.

[7] O. Solomon *et al.*, “RNA editing by ADAR1 leads to context-dependent transcriptome-wide changes in RNA secondary structure,” *Nat. Commun.*, vol. 8, no. 1, pp. 1–14, Nov. 2017, doi: 10.1038/s41467-017-01458-8.

[8] I. A. Roundtree and C. He, “RNA epigenetics—chemical messages for posttranscriptional gene regulation,” *Curr. Opin. Chem. Biol.*, vol. 30, pp. 46–51, Feb. 2016, doi: 10.1016/j.cbpa.2015.10.024.

[9] P. Boccaletto *et al.*, “MODOMICS: a database of RNA modification pathways. 2017 update,” *Nucleic Acids Res.*, vol. 46, no. Database issue, pp. D303–D307, Jan. 2018, doi: 10.1093/nar/gkx1030.

[10] I. A. Roundtree, M. E. Evans, T. Pan, and C. He, “Dynamic RNA modifications in gene expression regulation,” *Cell*, vol. 169, no. 7, pp. 1187–1200, Jun. 2017, doi: 10.1016/j.cell.2017.05.045.

[11] P. Alexiou, M. Maragkakis, G. L. Papadopoulos, M. Reczko, and A. G. Hatzigeorgiou, “Lost in translation: an assessment and perspective for computational microRNA target identification,” *Bioinformatics*, vol. 25, no. 23, pp. 3049–3055, Dec. 2009, doi: 10.1093/bioinformatics/btp565.

[12] Y. Cai, X. Yu, S. Hu, and J. Yu, “A Brief Review on the Mechanisms of miRNA Regulation,” *Genomics Proteomics Bioinformatics*, vol. 7, no. 4, pp. 147–154, Dec. 2009, doi: 10.1016/S1672-0229(08)60044-3.

[13] I. S. Vlachos and A. G. Hatzigeorgiou, “Online resources for miRNA analysis,” *Clin. Biochem.*, vol. 46, no. 10, pp. 879–900, Jul. 2013, doi: 10.1016/j.clinbiochem.2013.03.006.

[14] M. Schaefer, U. Kapoor, and M. F. Jantsch, “Understanding RNA modifications: the promises and technological bottlenecks of the ‘epitranscriptome,’” *Open Biol.*, vol. 7, no. 5, p. 170077, doi: 10.1098/rsob.170077.

[15] N. Jonkhout, J. Tran, M. A. Smith, N. Schonrock, J. S. Mattick, and E. M. Novoa, “The RNA modification landscape in human disease,” *RNA N. Y. N*, vol. 23, no. 12, pp. 1754–1769, Dec. 2017, doi: 10.1261/rna.063503.117.

[16] S. Li and C. E. Mason, “The Pivotal Regulatory Landscape of RNA Modifications,” *Annu. Rev. Genomics Hum. Genet.*, vol. 15, no. 1, pp. 127–150, 2014, doi: 10.1146/annurev-genom-090413-025405.

[17] S. Delaunay and M. Frye, “RNA modifications regulating cell fate in cancer,” *Nat. Cell Biol.*, vol. 21, no. 5, pp. 552–559, May 2019, doi: 10.1038/s41556-019-0319-0.

[18] A. Gatsiou and K. Stellos, “Dawn of Epitranscriptomic Medicine,” *Circ. Genomic Precis. Med.*, vol. 11, no. 9, p. e001927, 2018, doi: 10.1161/CIRCGEN.118.001927.

[19] M. Frye, S. R. Jaffrey, T. Pan, G. Rechavi, and T. Suzuki, “RNA modifications: what have we learned and where are we headed?,” *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 365–372, Jun. 2016, doi: 10.1038/nrg.2016.47.

[20] B. Zinshteyn and K. Nishikura, “Adenosine-to-inosine RNA editing,” *WIREs Syst. Biol. Med.*, vol. 1, no. 2, pp. 202–209, 2009, doi: 10.1002/wsbm.10.

[21] Y. Wang, Y. Zheng, and P. A. Beal, “Adenosine Deaminases That Act on RNA (ADARs),” *The Enzymes*, vol. 41, pp. 215–268, 2017, doi: 10.1016/bs.enz.2017.03.006.

[22] L. T. Vu and T. Tsukahara, “C-to-U editing and site-directed RNA editing for the correction of genetic mutations,” *Biosci. Trends*, vol. 11, no. 3, pp. 243–253, 2017, doi: 10.5582/bst.2017.01049.

[23] G. Ramaswami, W. Lin, R. Piskol, M. H. Tan, C. Davis, and J. B. Li, “Accurate identification of human Alu and non-Alu RNA editing sites,” *Nat. Methods*, vol. 9, no. 6, pp. 579–581, Jun. 2012, doi: 10.1038/nmeth.1982.

[24] P. Deininger, “Alu elements: know the SINEs,” *Genome Biol.*, vol. 12, no. 12, p. 236, Dec. 2011, doi: 10.1186/gb-2011-12-12-236.

[25] M. J. Palladino, L. P. Keegan, M. A. O’Connell, and R. A. Reenan, “A-to-I Pre-mRNA Editing in Drosophila Is Primarily Involved in Adult Nervous System Function and Integrity,” *Cell*, vol. 102, no. 4, pp. 437–449, Aug. 2000, doi: 10.1016/S0092-8674(00)00049-0.

[26] M. Higuchi *et al.*, “Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2,” *Nature*, vol. 406, no. 6791, pp. 78–81, Jul. 2000, doi: 10.1038/35017558.
[27] J. C. Hartner, C. Schmittwolf, A. Kispert, A. M. Müller, M. Higuchi, and P. H. Seeburg, “Liver Disintegration in the Mouse Embryo Caused by Deficiency in the RNA-editing Enzyme ADAR1,” *J. Biol. Chem.*, vol. 279, no. 6, pp. 4894–4902, Feb. 2004, doi: 10.1074/jbc.M311347200.
[28] Q. Wang, J. Khillan, P. Gadue, and K. Nishikura, “Requirement of the RNA Editing Deaminase ADAR1 Gene for Embryonic Erythropoiesis,” *Science*, vol. 290, no. 5497, pp. 1765–1768, Dec. 2000, doi: 10.1126/science.290.5497.1765.
[29] C. Cenci *et al.*, “Down-regulation of RNA Editing in Pediatric Astrocytomas ADAR2 EDITING ACTIVITY INHIBITS CELL MIGRATION AND PROLIFERATION,” *J. Biol. Chem.*, vol. 283, no. 11, pp. 7251–7260, Mar. 2008, doi: 10.1074/jbc.M708316200.
[30] L. Yang *et al.*, “Deficiency in RNA editing enzyme ADAR2 impairs regulated exocytosis,” *FASEB J.*, vol. 24, no. 10, pp. 3720–3732, May 2010, doi: 10.1096/fj.09-152363.
[31] S. Maas, Y. Kawahara, K. M. Tamburro, and K. Nishikura, “A-to-I RNA Editing and Human Disease,” *RNA Biol.*, vol. 3, no. 1, pp. 1–9, Jan. 2006, doi: 10.4161/rna.3.1.2495.
[32] A. Gallo and F. Locatelli, “ADARs: allies or enemies? The importance of A-to-I RNA editing in human disease: from cancer to HIV-1,” *Biol. Rev.*, vol. 87, no. 1, pp. 95–110, 2012, doi: 10.1111/j.1469-185X.2011.00186.x.
[33] E. Picardi, C. Manzari, F. Mastropasqua, I. Aiello, A. M. D’Erchia, and G. Pesole, “Profiling RNA editing in human tissues: towards the inosinome Atlas,” *Sci. Rep.*, vol. 5, Oct. 2015, doi: 10.1038/srep14941.
[34] Y. Kawahara, B. Zinshteyn, P. Sethupathy, H. Iizasa, A. G. Hatzigeorgiou, and K. Nishikura, “Redirection of Silencing Targets by Adenosine-to-Inosine Editing of miRNAs,” *Science*, vol. 315, no. 5815, pp. 1137–1140, Feb. 2007, doi: 10.1126/science.1138050.
[35] C.-P. Kung, L. B. J. Maggi, and J. D. Weber, “The Role of RNA Editing in Cancer Development and Metabolic Disorders,” *Front. Endocrinol.*, vol. 9, 2018, doi: 10.3389/fendo.2018.00762.
[36] R. C. Lee, R. L. Feinbaum, and V. Ambros, “The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14,” *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993, doi: 10.1016/0092-8674(93)90529-Y.
[37] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, “Identification of Novel Genes Coding for Small Expressed RNAs,” *Science*, vol. 294, no. 5543, pp. 853–858, Oct. 2001, doi: 10.1126/science.1064921.
[38] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, “miRBase: from microRNA sequences to function,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D155–D162, Jan. 2019, doi: 10.1093/nar/gky1141.
[39] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley, “Identification of mammalian microRNA host genes and transcription units,” *Genome Res.*, vol. 14, no. 10A, pp. 1902–1910, Oct. 2004, doi: 10.1101/gr.2722704.
[40] A. F. Olena and J. G. Patton, “Genomic organization of microRNAs,” *J. Cell. Physiol.*, vol. 222, no. 3, pp. 540–545, 2010, doi: 10.1002/jcp.21993.
[41] A. Steiman-Shimony, O. Shtrikman, and H. Margalit, “Assessing the functional association of intronic miRNAs with their host genes,” *RNA N. Y. N*, vol. 24, no. 8, pp. 991–1004, Aug. 2018, doi: 10.1261/rna.064386.117.
[42] T. Treiber, N. Treiber, and G. Meister, “Regulation of microRNA biogenesis and its crosstalk with other cellular pathways,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 1, pp. 5–20, Jan. 2019, doi: 10.1038/s41580-018-0059-1.
[43] J. Höck and G. Meister, “The Argonaute protein family,” *Genome Biol.*, vol. 9, no. 2, p. 210, Feb. 2008, doi: 10.1186/gb-2008-9-2-210.
[44] Z. Ye, H. Jin, and Q. Qian, “Argonaute 2: A Novel Rising Star in Cancer Research,” *J. Cancer*, vol. 6, no. 9, pp. 877–882, 2015, doi: 10.7150/jca.11735.
[45] J. G. Ruby, C. H. Jan, and D. P. Bartel, “Intronic microRNA precursors that bypass Drosha processing,” *Nature*, vol. 448, no. 7149, pp. 83–86, Jul. 2007, doi: 10.1038/nature05983.
[46] S. Cheloufi, C. O. Dos Santos, M. M. W. Chong, and G. J. Hannon, “A dicer-independent miRNA biogenesis pathway that requires Ago catalysis,” *Nature*, vol. 465, no. 7298, pp. 584–589, Jun. 2010, doi: 10.1038/nature09092.
[47] B. C. Schanen and X. Li, “Transcriptional regulation of mammalian miRNA genes,” *Genomics*, vol. 97, no. 1, pp. 1–6, Jan. 2011, doi: 10.1016/j.ygeno.2010.10.005.
[48] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel, “Prediction of Plant MicroRNA Targets,” *Cell*, vol. 110, no. 4, pp. 513–520, Aug. 2002, doi: 10.1016/S0092-8674(02)00863-2.
[49] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets,” *Cell*, vol. 120, no. 1, pp. 15–20, Jan. 2005, doi: 10.1016/j.cell.2004.12.035.

[50] D. P. Bartel, “MicroRNAs: Target Recognition and Regulatory Functions,” *Cell*, vol. 136, no. 2, pp. 215–233, Jan. 2009, doi: 10.1016/j.cell.2009.01.002.
[51] W. Gu, Y. Xu, X. Xie, T. Wang, J.-H. Ko, and T. Zhou, “The role of RNA structure at 5’ untranslated region in microRNA-mediated gene regulation,” *RNA N. Y. N*, vol. 20, no. 9, pp. 1369–1375, Sep. 2014, doi: 10.1261/rna.044792.114.
[52] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel, “The impact of microRNAs on protein output,” *Nature*, vol. 455, no. 7209, pp. 64–71, Sep. 2008, doi: 10.1038/nature07242.
[53] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, “Most mammalian mRNAs are conserved targets of microRNAs,” *Genome Res.*, vol. 19, no. 1, pp. 92–105, Jan. 2009, doi: 10.1101/gr.082701.108.
[54] K. Srivastava and A. Srivastava, “Comprehensive review of genetic association studies and meta-analyses on miRNA polymorphisms and cancer risk,” *PloS One*, vol. 7, no. 11, pp. e50966–e50966, 2012, doi: 10.1371/journal.pone.0050966.
[55] L. Han, P. D. W. Witmer, E. Casey, D. Valle, and S. Sukumar, “DNA methylation regulates microRNA expression,” *Cancer Biol. Ther.*, vol. 6, no. 8, pp. 1290–1294, Aug. 2007, doi: 10.4161/cbt.6.8.4486.
[56] M. A. Diroma, L. Ciaccia, G. Pesole, and E. Picardi, “Elucidating the editome: bioinformatics approaches for RNA editing detection,” *Brief. Bioinform.*, vol. 20, no. 2, pp. 436–447, Mar. 2019, doi: 10.1093/bib/bbx129.
[57] Z. Wang *et al.*, “RES-Scanner: a software package for genome-wide identification of RNA-editing sites,” *GigaScience*, vol. 5, no. 1, p. 37, Aug. 2016, doi: 10.1186/s13742-016-0143-4.
[58] D. John, T. Weirick, S. Dimmeler, and S. Uchida, “RNAEditor: easy detection of RNA editing events and the introduction of editing islands,” *Brief. Bioinform.*, vol. 18, no. 6, pp. 993–1001, Nov. 2017, doi: 10.1093/bib/bbw087.
[59] K. Stellos *et al.*, “Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation,” *Nat. Med.*, vol. 22, no. 10, pp. 1140–1150, Oct. 2016, doi: 10.1038/nm.4172.
[60] Q. Zhang, “Analysis of RNA Editing Sites from RNA-Seq Data Using GIREMI,” in *Transcriptome Data Analysis: Methods and Protocols*, Y. Wang and M. Sun, Eds. New York, NY: Springer New York, 2018, pp. 101–108.
[61] F. Zhang, Y. Lu, S. Yan, Q. Xing, and W. Tian, “SPRINT: an SNP-free toolkit for identifying RNA editing sites,” *Bioinforma. Oxf. Engl.*, vol. 33, no. 22, pp. 3538–3548, Nov. 2017, doi: 10.1093/bioinformatics/btx473.
[62] H. Xiong *et al.*, “RED-ML: a novel, effective RNA editing detection method based on machine learning,” *GigaScience*, vol. 6, no. 5, pp. 1–8, May 2017, doi: 10.1093/gigascience/gix012.
[63] E. Picardi and G. Pesole, “REDItools: high-throughput RNA editing detection made easy,” *Bioinforma. Oxf. Engl.*, vol. 29, no. 14, pp. 1813–1814, Jul. 2013, doi: 10.1093/bioinformatics/btt287.
[64] M. Piechotta, E. Wyler, U. Ohler, M. Landthaler, and C. Dieterich, “JACUSA: site-specific identification of RNA editing events from replicate sequencing data,” *BMC Bioinformatics*, vol. 18, no. 1, pp. 7–7, Jan. 2017, doi: 10.1186/s12859-016-1432-8.
[65] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, “Predicting effective microRNA target sites in mammalian mRNAs,” *eLife*, vol. 4, p. e05005, Aug. 2015, doi: 10.7554/eLife.05005.
[66] R. Gumienny and M. Zavolan, “Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G,” *Nucleic Acids Res.*, vol. 43, no. 3, pp. 1380–1391, Feb. 2015, doi: 10.1093/nar/gkv050.
[67] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, “Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites,” *Genome Biol.*, vol. 11, no. 8, pp. R90–R90, 2010, doi: 10.1186/gb-2010-11-8-r90.
[68] M. D. Paraskevopoulou *et al.*, “DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows,” *Nucleic Acids Res.*, vol. 41, no. Web Server issue, pp. W169–W173, Jul. 2013, doi: 10.1093/nar/gkt393.
[69] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, “The role of site accessibility in microRNA target recognition,” *Nat. Genet.*, vol. 39, no. 10, pp. 1278–1284, Oct. 2007, doi: 10.1038/ng2135.
[70] N. Wong and X. Wang, “miRDB: an online resource for microRNA target prediction and functional annotations,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D146–D152, Jan. 2015, doi: 10.1093/nar/gku1104.
[71] T. Lappalainen *et al.*, “Transcriptome and genome sequencing uncovers functional variation in humans,” *Nature*, vol. 501, no. 7468, pp. 506–511, Sep. 2013, doi: 10.1038/nature12531.
[72] H. Parkinson *et al.*, “ArrayExpress--a public database of microarray experiments and gene expression profiles,” *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D747–D750, Jan. 2007, doi: 10.1093/nar/gkl995.
[73] B. Zhou *et al.*, “Comprehensive, Integrated, and Phased Whole-Genome Analysis of the Primary ENCODE Cell Line K562,” *bioRxiv*, p. 192344, Dec. 2017, doi: 10.1101/192344.

[74] "Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562." [Online]. Available: https://genome.cshlp.org/content/29/3/472. [Accessed: 27-Nov-2019].
[75] I. Dunham *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.
[76] C. A. Davis *et al.*, "The Encyclopedia of DNA elements (ENCODE): data portal update," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D794–D801, Jan. 2018, doi: 10.1093/nar/gkx1081.
[77] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, no. 3, p. R46, Mar. 2014, doi: 10.1186/gb-2014-15-3-r46.
[78] S. E. Hunt *et al.*, "Ensembl variation resources," *Database*, vol. 2018, Jan. 2018, doi: 10.1093/database/bay119.
[79] W. J. Kent *et al.*, "The Human Genome Browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002, doi: 10.1101/gr.229102.
[80] G. Ramaswami and J. B. Li, "RADAR: a rigorously annotated database of A-to-I RNA editing," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D109–D113, Jan. 2014, doi: 10.1093/nar/gkt996.
[81] W. J. Kent, "BLAT--the BLAST-like alignment tool," *Genome Res.*, vol. 12, no. 4, pp. 656–664, Apr. 2002, doi: 10.1101/gr.229202.
[82] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.
[83] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinforma. Oxf. Engl.*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[84] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics*, vol. 26, no. 7, pp. 873–881, Apr. 2010, doi: 10.1093/bioinformatics/btq057.
[85] *Broad Institute. (Accessed: 2018/02/21; version 2.17.8). "Picard Tools." Broad Institute, GitHub repository. http://broadinstitute.github.io/picard/*. .
[86] M. A. DePristo *et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat. Genet.*, vol. 43, no. 5, pp. 491–498, May 2011, doi: 10.1038/ng.806.
[87] S. H. Bernhart, I. L. Hofacker, and P. F. Stadler, "Local RNA base pairing probabilities in large sequences," *Bioinformatics*, vol. 22, no. 5, pp. 614–615, Dec. 2005, doi: 10.1093/bioinformatics/btk014.
[88] C. B. Do, D. A. Woods, and S. Batzoglou, "CONTRAfold: RNA secondary structure prediction without physics-based models," *Bioinformatics*, vol. 22, no. 14, pp. e90–e98, Jul. 2006, doi: 10.1093/bioinformatics/btl246.
[89] M. Khorshid, J. Hausser, M. Zavolan, and E. van Nimwegen, "A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets," *Nat. Methods*, vol. 10, no. 3, pp. 253–255, Mar. 2013, doi: 10.1038/nmeth.2341.
[90] J. Harrow *et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res.*, vol. 22, no. 9, pp. 1760–1774, Sep. 2012, doi: 10.1101/gr.135350.111.
[91] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, Mar. 2009, doi: 10.1186/gb-2009-10-3-r25.
[92] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades," *Nucleic Acids Res.*, vol. 40, no. 1, pp. 37–52, Jan. 2012, doi: 10.1093/nar/gkr688.
[93] A.-S. Espadinha *et al.*, "A tyrosine kinase-STAT5-miR21-PDCD4 regulatory axis in chronic and acute myeloid leukemia cells," *Oncotarget*, vol. 8, no. 44, pp. 76174–76188, Jul. 2017, doi: 10.18632/oncotarget.19192.
[94] T. Barrett *et al.*, "NCBI GEO: archive for functional genomics data sets--update," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D991–D995, Jan. 2013, doi: 10.1093/nar/gks1193.
[95] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, N.J: L. Erlbaum Associates, 1988.
[96] R. Rosenthal, "Parametric measures of effect size," in *The handbook of research synthesis*, New York, NY, US: Russell Sage Foundation, 1994, pp. 231–244.
[97] C. O. Fritz, P. E. Morris, and J. J. Richler, "Effect size estimates: Current use, calculations, and interpretation," *J. Exp. Psychol. Gen.*, vol. 141, no. 1, pp. 2–18, 2012, doi: 10.1037/a0024338.
[98] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nat. Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.