



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

**Ανάλυση χαμηλής και υψηλής απόδοσης
πειραματικών μεθόδων για συλλογή επιβεβαιωμένων
γονιδίων στόχων των miRNAs**

Σοφία Π. Μανιού

Επιβλέπουσα: Άρτεμις Χατζηγεωργίου, Καθηγήτρια

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση χαμηλής και υψηλής απόδοσης πειραματικών μεθόδων για συλλογή επιβεβαιωμένων γονιδίων στόχων των miRNAs

Σοφία Π. Μανιού

A.M.: ΠΙΒ115

ΕΠΙΒΛΕΠΟΥΣΑ: Άρτεμις Χατζηγεωργίου, Καθηγήτρια

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: Άρτεμις Χατζηγεωργίου, Καθηγήτρια
Γεώργιος Σπύρου, Κάτοχος Έδρας Βιοπληροφορικής &
Δ/ντής Ομάδας Βιοπληροφορικής
Μαρία Παρασκευοπούλου, Ερευνήτρια/Επιστημονικός
Συνεργάτης

Δεκέμβριος 2016

ΠΕΡΙΛΗΨΗ

Τα microRNAs είναι μικρά μη κωδικά μόρια RNA που θεωρούνται οι κύριοι μετα-μεταγραφικοί ρυθμιστές της γονιδιακής έκφρασης. Το μεγάλο ενδιαφέρον που υπάρχει τα τελευταία χρόνια για την κατανόηση της λειτουργίας τους, έχει δημιουργήσει μία έντονη δραστηριότητα έρευνας στο συγκεκριμένο πεδίο.

Μία από τις πιο σημαντικές διαδικασίες στην έρευνα των miRNAs, είναι η αναγνώριση των αλληλεπιδράσεων που έχουν με τα γονίδια στόχους τους. Τέτοιες αλληλεπιδράσεις μπορούν να επιβεβαιωθούν από πειραματικές μεθόδους υψηλής και/ή χαμηλής απόδοσης. Η υψηλής απόδοσης μέθοδος των μικροσυστοιχιών χρησιμοποιείται κατά κόρον στο πεδίο έρευνας των miRNAs, με σκοπό να μελετηθεί η επίδραση που έχει η υπερέκφραση ή καταστολή ενός miRNA στην γονιδιακή έκφραση. Από μία λεπτομερειακή επεξεργασία δεδομένων των μικροσυστοιχιών, μπορούν να προκύψουν αλληλεπιδράσεις των miRNAs με τα γονίδια στόχους τους.

Στα πλαίσια αυτής της διπλωματικής εργασίας εφαρμόστηκαν διαφορετικές προσεγγίσεις για την επεξεργασία τέτοιων δεδομένων, και πραγματοποιήθηκε μία σύγκριση μεταξύ τους για την αναζήτηση της πιο αξιόπιστης μεθόδου. Έναν ακόμη στόχο της εργασίας αποτέλεσε η συλλογή miRNA:gene αλληλεπιδράσεων από την χαμηλής απόδοσης μέθοδο των γονιδίων αναφοράς σε διάφορους κυτταρικούς τύπους και διαφορετικές πειραματικές συνθήκες. Δόθηκε ιδιαίτερη έμφαση στα δεδομένα της συγκεκριμένης μεθόδου, η οποία μπορεί να επιβεβαιώσει το ακριβές σημείο πρόσδεσης του miRNA πάνω στο μετάγραφο του γονιδίου στόχου του.

Η συλλογή τέτοιων δεδομένων μπορεί να χρησιμοποιηθεί σε μεγάλο εύρος εφαρμογών, όπως την αναβάθμιση μίας σχετικής βάσης δεδομένων ή την εκπαίδευση αλγορίθμων πρόβλεψης στόχων των miRNAs.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: microRNAs, μικροσυστοιχίες, μέθοδος γονιδίων αναφοράς, πειραματικά επιβεβαιωμένοι στόχοι, επεξεργασία δεδομένων μικροσυστοιχιών

ABSTRACT

MicroRNAs are small non-coding RNA molecules considered to be the main post-transcriptional regulators of gene expression. In the past few years, the great interest of understanding their functional role has lead to vigorous research efforts concerning the biogenesis and function of miRNAs.

One of the most important processes in miRNA research is the identification of the interactions they have with their target genes. These interactions can be verified by high and/or low throughput experimental techniques. Microarrays are a high throughput technique vastly applied in the miRNA research, as a means to examine the transcriptome-wide effects of the overexpression or inhibition of a miRNA under study. The wealth of information that lies in microarray datasets can be meticulously processed in order to infer interactions of miRNAs and their target genes.

In this thesis, different approaches of microarray data processing were applied, and a comparison of them was made to examine the most reliable data processing method. Apart from the miRNA:gene interactions that emerged from the Microarrays, the scope of this thesis was also to collect data regarding interactions verified by the low throughput technique of the Reporter genes in various cell types and under different experimental conditions. Considering that this method can provide the exact location of the binding site, great emphasis was given on its derived data.

The collection of data related to these interactions can be used in a wide range of applications, including the update of a relevant database, or the training of miRNA target prediction algorithms.

SUBJECT AREA: Bioinformatics

KEYWORDS: microRNAs, Microarrays, Reporter genes, experimentally verified targets, microarray data processing

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ	9
1. ΕΙΣΑΓΩΓΗ	10
1.1 microRNAs	10
1.1.1 Βιογένεση	10
1.1.2 Λειτουργία	12
1.1.3 Ρόλος στις ασθένειες	13
1.2 Αλληλεπιδράσεις miRNAs με γονίδια	15
1.2.1 Πειραματικές μέθοδοι	15
1.2.2 Βάσεις δεδομένων	17
1.3 Microarrays	18
1.3.1 Περιγραφή τεχνολογίας	19
1.3.2 Κανονικοποίηση δεδομένων	20
1.3.2.1 MAS5 αλγόριθμος	21
1.3.2.2 RMA αλγόριθμος	22
1.3.2.3 gcRMA αλγόριθμος	22
1.3.3 Υπολογισμός τιμής έκφρασης	23
2. ΑΠΟΤΕΛΕΣΜΑΤΑ	24
2.1 Επεξεργασία δεδομένων από μικροσυστοιχίες	24
2.1.1 Εφαρμογή μεθόδων	24
2.1.2 Σύγκριση μεθόδων	29
2.2 Συλλογή αλληλεπιδράσεων από Reporter gene assays	34
2.2.1 Διαδικασία καταγραφής	34
2.2.2 Στατιστικά δεδομένων	36
3. ΣΥΜΠΕΡΑΣΜΑΤΑ	38
ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ	40
ΣΥΝΤΗΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ	41
ΑΝΑΦΟΡΕΣ	42

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Σχήμα 1: Αριθμός γονιδίων για κάθε αντιστοιχία γονιδίου:τιμές έκφρασης	24
Σχήμα 2: Κατανομή τιμών έκφρασης από μέση τιμή για καταστολή.....	27
Σχήμα 3: Κατανομή τιμών έκφρασης από μέση τιμή για υπερέκφραση.....	27
Σχήμα 4: Κατανομή τιμών έκφρασης από διάμεσο για καταστολή	28
Σχήμα 5: Κατανομή τιμών έκφρασης από διάμεσο για υπερέκφραση	28
Σχήμα 6: Venn διάγραμμα τιμών έκφρασης από μέση τιμή.....	29
Σχήμα 7: Κατανομή επιπλέον γονιδίων από μέση τιμή για κάθε μέθοδο	31
Σχήμα 8: Κατανομή επιπλέον γονιδίων από διάμεσο για κάθε μέθοδο	31
Σχήμα 9: Μέση τιμή επιπλέον γονιδίων από μέση τιμή για κάθε αντιστοιχία γονιδίου:τιμές έκφρασης	33
Σχήμα 10: Μέση τιμή επιπλέον γονιδίων από διάμεσο για κάθε αντιστοιχία γονιδίου:τιμές έκφρασης	33

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Βιογένεση των miRNAs [8].....	11
Εικόνα 2: Μηχανισμοί δράσης των miRNAs [16]	13
Εικόνα 3: Στιγμιότυπο της διεπαφής του DIANA-TarBase. Ο χρήστης μπορεί να καθορίσει τους όρους της αναζήτησής του (1), ώστε να εμφανιστούν όλες οι σχετικές αλληλεπιδράσεις (2). Επεκτείνοντας την λίστα της αλληλεπίδρασης, ή πατώντας το κουμπί (i) δίπλα από κάθε miRNA ή γονίδιο παρουσιάζονται περισσότερες πληροφορίες (4). Ο χρήστης έχει την δυνατότητα να δει όλα τα πειράματα που χρησιμοποιήθηκαν για την επιβεβαίωση της αλληλεπίδρασης (5), καθώς και να περιορίσει τα αποτελέσματα της αναζήτησής του χρησιμοποιώντας τα διαθέσιμα φίλτρα (3). [23]..	17
Εικόνα 4: Γενωμικό chip Affymetrix	19
Εικόνα 5: Διαδικασία πειράματος μικροσυστοιχιών.....	20
Εικόνα 6: Πεδία κάθε αλληλεπίδρασης.....	35

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1: Ο ρόλος των miRNA σε διάφορους τύπους καρκίνου [21].....	14
Πίνακας 2: Πειραματικές τεχνικές αναγνώρισης miRNA:gene αλληλεπιδράσεων [23]	16
Πίνακας 3: Datasets σύγκρισης.....	25
Πίνακας 4: Αποτελέσματα στατιστικού ελέγχου.....	32
Πίνακας 5: Στατιστικά δεδομένων.....	36
Πίνακας 6: Εγγραφές ανά οργανισμό.....	36

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών “Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία”. Το θέμα της εργασίας βασίστηκε στην ύψιστη σημασία που έχουν τα δεδομένα που σχετίζονται με τις πειραματικά υποστηριζόμενες αλληλεπιδράσεις μεταξύ των miRNAs και των γονιδίων που αποτελούν στόχους τους.

Τέτοιου είδους δεδομένα μπορούν να χρησιμοποιηθούν σε ένα μεγάλο εύρος εφαρμογών, που μπορεί να περιλαμβάνει την ενημέρωση μίας σχετικής βάσης δεδομένων ή την εκπαίδευση των αλγορίθμων που προβλέπουν στόχους των miRNAs. Η συλλογή όμως τέτοιων δεδομένων είναι μία δύσκολη και χρονοβόρα διαδικασία, καθώς οι αλληλεπιδράσεις συνήθως είναι κρυμμένες σε επιμέρους ή συμπληρωματικό υλικό των επιστημονικών δημοσιεύσεων.

Οι miRNA:gene αλληλεπιδράσεις μπορούν να αναγνωριστούν με τη χρήση πειραματικών μεθόδων. Η μέθοδος των μικροσυστοιχιών χρησιμοποιείται κατά κόρον στην έρευνα των miRNAs για να μελετηθεί η επίδραση που έχει η υπερέκφραση ή καταστολή ενός υπό μελέτη miRNA στη γονιδιακή έκφραση. Μέσω της επεξεργασίας τέτοιων δεδομένων μπορούν να αναγνωριστούν αλληλεπιδράσεις μεταξύ miRNAs και γονιδίων στόχων τους.

Στα δεδομένα των μικροσυστοιχιών, ένα γονίδιο μπορεί να αντιστοιχίζεται με περισσότερες από μία τιμές έκφρασης, και για την ανάλυση της γονιδιακής έκφρασης κάθε γονίδιο πρέπει να συσχετιστεί με μία τελική τιμή έκφρασης. Συνήθως επιλέγεται η μέση τιμή ή διάμεσος όλων των τιμών έκφρασης του γονιδίου. Ωστόσο σε αυτή την εργασία, διερευνήθηκε μία διαφορετική προσέγγιση για τον υπολογισμό της τελικής τιμής έκφρασης.

Βασικό στόχο της εργασίας αποτέλεσε η συλλογή και ανάλυση δεδομένων που αφορούν τις πειραματικά επιβεβαιωμένες αλληλεπιδράσεις των miRNAs με τα γονίδια στόχους τους. Τα σχετικά δεδομένα συλλέχθηκαν από την λεπτομερειακή επεξεργασία δεδομένων που προέρχονται από τις miRNA μικροσυστοιχίες, καθώς και με την αναγνώρισή ειδικών αλληλεπιδράσεων από ένα μεγάλο πλήθος επιστημονικών δημοσιεύσεων.

Σε αυτό το σημείο, θα ήθελα να ευχαριστήσω την Καθηγήτρια κα. Άρτεμις Χατζηγεωργίου για την ανάθεση και επίβλεψη αυτής της διπλωματικής εργασίας, καθώς και όλα τα μέλη του DIANA lab για την πολύτιμη βοήθεια τους. Ιδιαίτερα θα ήθελα να ευχαριστήσω τη μεταδιδακτορική ερευνήτρια Μαρία Παρασκευοπούλου για τις ιδέες και την καθοδήγηση που μου προσέφερε κατά τη διάρκεια της εργασίας μου. Ακόμη ευχαριστώ τον Σπύρο Ταστσόγλου για τη συνεισφορά του στη διπλωματική με τη συλλογή και προ-επεξεργασία των πειραματικών δεδομένων.

1. ΕΙΣΑΓΩΓΗ

1.1 microRNAs

Ο όρος micro-RNA (miRNA) χρησιμοποιήθηκε για να περιγράψει μικρά μόρια RNA που δεν παράγουν πρωτεΐνες, αλλά προσδένονται στα μεταγράφα (mRNAs) των γονιδίων στόχων τους και αναστέλλουν την έκφραση τους μειώνοντας έτσι την παραγωγή πρωτεΐνης. Εντοπίζονται συνήθως σε συντηρημένες περιοχές του γονιδιώματος των φυτών και ζώων.

Τα πρώτα miRNAs ανακαλύφθηκαν το 1993 στον οργανισμό *Caenorhabditis elegans*, και ήταν το *lin-4* και το *let-7* [1]. Ωστόσο η έρευνα σχετικά με αυτά ξεκίνησε το 2001, όταν ανακαλύφθηκε πως αποτελούν μία μεγάλη κλάση μορίων που υπάρχουν σε μεγάλες ποσότητες στα κύτταρα πολλών οργανισμών, όπως του ανθρώπου, καθώς και των *C. elegans* και *D. melanogaster* [2-4].

Τα RNA μόρια αυτής της νέας κλάσης παρουσίαζαν πολλές ομοιότητες με τα μόρια *lin-4* και *let-7* RNAs. Παρ' όλα αυτά, τα μοτίβα έκφρασης αυτών των μορίων σε διάφορους τύπους κυττάρων διαφοροποιούνταν, κάτι που υποδείκνυε πως πιθανότατα συμμετέχουν σε διαφορετικού είδους ρυθμιστικά μονοπάτια [2-4].

Τα miRNAs πλέον θεωρούνται οι βασικοί μετα-μεταγραφικοί ρυθμιστές της γονιδιακής έκφρασης, είτε μέσω της αποικοδόμησης ή τομής του mRNA, είτε μέσω της καταστολής της μετάφρασης του mRNA σε πρωτεΐνη. Σήμερα, έχουν καταγραφεί στη βάση δεδομένων miRBase (<http://www.mirbase.org/>) περισσότερα από 2,500 ανθρώπινα miRNAs.

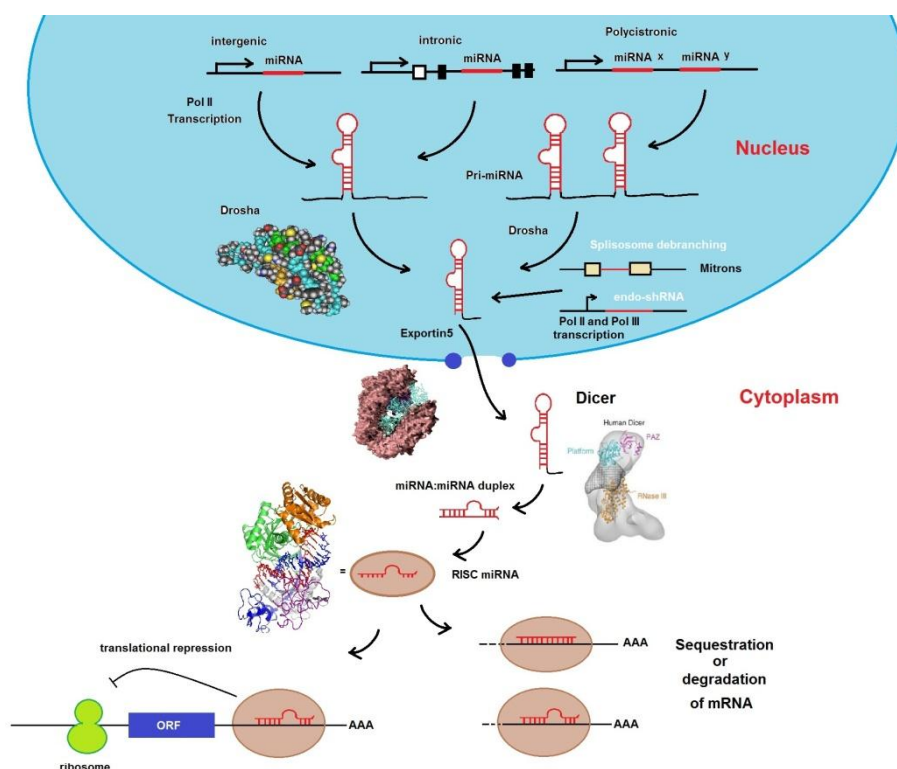
Έχει αποδειχθεί πως τα miRNAs ρυθμίζουν πάνω από το 50% των ανθρώπινων γονιδίων, και η απορρύθμιση τους έχει συσχετιστεί με πολλές ανθρώπινες ασθένειες. Η πρώτη ασθένεια με την οποία συσχετίστηκε ήταν η Χρόνια Λεμφοκυτταρική Λευχαιμία, όταν βρέθηκε πως εμφανίζεται συχνά μία διαγραφή στα γονίδια που παράγουν τα miR-15 και miR-16 στο 65% των ασθενών [5].

1.1.1 Βιογένεση

Υπάρχουν διαφορετικοί τρόποι με τους οποίους μπορεί να δημιουργηθεί ένα miRNA. Ο πιο συχνός τρόπος γένεσης των miRNAs είναι μέσω ειδικών γονιδίων, που σκοπό τους έχουν την παραγωγή του συγκεκριμένου miRNA. Περίπου το 40% των miRNAs όμως περιλαμβάνονται σε ιντρόνια γονιδίων που προκύπτουν μετά από το μάτισμα (splicing) των μορίων pre-mRNA κατά τη διαδικασία της μεταγραφής των γονιδίων. Τα ιντρόνια αυτά μπορεί να προέρχονται από γονίδια που είτε μεταφράζονται σε πρωτεΐνη είτε όχι. Επίσης, τα miRNAs μπορεί να περιλαμβάνονται ακόμα και σε εξώνια μεγάλων μεταγράφων που δεν μεταφράζονται σε πρωτεΐνη. Τα miRNAs που προέρχονται από ιντρόνια ή εξώνια άλλων μεταγράφων συνήθως ρυθμίζονται μαζί με τα γονίδια αυτά [6].

Τα γονίδια που παράγουν miRNAs, μεταγράφονται με τη βοήθεια της RNA πολυμεράσης II, η οποία προσδένεται στον υποκινητή (promoter). Ο

υποκινητής αποτελεί μία περιοχή που βρίσκεται ακριβώς πριν από τη DNA (Deoxyribonucleic Acid) ακολουθία των γονιδίων και περιέχει την πληροφορία για τη δημιουργία του μορίου precursor miRNA (pre-miRNA) που αποτελεί μία φουρκέτα (hairpin loop). Το μετάγραφο που προκύπτει περιέχει μία σειρά από χαρακτηριστικές περιοχές, όπως το 5' cap άκρο και το poly-A tail που είναι μία σειρά από αδενίνες. Τα ιντρόνια αφαιρούνται από το μετάγραφο με τη διαδικασία του ματίσματος. Τα μετάγραφα των miRNAs των ζωικών οργανισμών, αποτελούν αρχικά έναν βρόγχο με έναν μόνο βραχίονα (single-arm stem loop) μήκους περίπου 80 βάσεων. Στη συνέχεια αυτός ο βρόγχος σχηματίζει ένα πρώιμο μόριο miRNA που ονομάζεται primary miRNA (pri-miRNA). [7]



Εικόνα 1: Βιογένεση των miRNAs [8]

Τα μόρια pri-miRNA αποτελούνται από μία διπλή έλικα και μία φουρκέτα. Η πρωτεΐνη Pasha αναγνωρίζει το μόριο pri-miRNA στον πυρήνα του κυττάρου, και μαζί με το ένζυμο Drosha το οποίο έχει την ιδιότητα να κόβει το RNA, σχηματίζουν ένα σύμπλοκο που πραγματοποιεί την επεξεργασία του miRNA ώστε να δημιουργηθεί το μόριο pre-miRNA [9, 10]. Ωστόσο, η παραπάνω διαδικασία πραγματοποιείται μόνο στα miRNAs που προέρχονται από τα δικά τους εξειδικευμένα γονίδια, ενώ τα miRNAs που προέρχονται από ιντρόνια από το μάτισμα άλλων mRNA μορίων προσπερνούν αυτή τη διαδικασία και ονομάζονται “Mirtrons”.

Στη συνέχεια τα μόρια pre-miRNA εξέρχονται από τον πυρήνα του κυττάρου και εισέρχονται στο κυτταρόπλασμα. Εκεί αναγνωρίζονται από το ένζυμο Dicer το οποίο πραγματοποιεί τη διάσπασή τους. Το ένζυμο Dicer

αλληλεπιδρά με τα 3' και 5' άκρα της φουρκέτας του pre-miRNA και αποκόπτει τον βρόγχο, με αποτέλεσμα να δημιουργείται μία διπλή έλικα (miRNA-miRNA*) από ζεύγη νουκλεοτιδίων με μη τέλεια συμπληρωματικότητα μεταξύ τους. [11] Οι δύο κλώνοι της διπλής έλικας, miRNA και miRNA*, μπορούν να λειτουργήσουν ξεχωριστά ως ένα miRNA. Ωστόσο συνήθως μόνο ο ένας από τους δύο κλώνους ενσωματώνεται στο σύμπλοκο RISC (RNA-induced silencing complex), μέσω του οποίου τα miRNAs μπορούν να αλληλεπιδράσουν με τους mRNA στόχους τους.

1.1.2 Λειτουργία

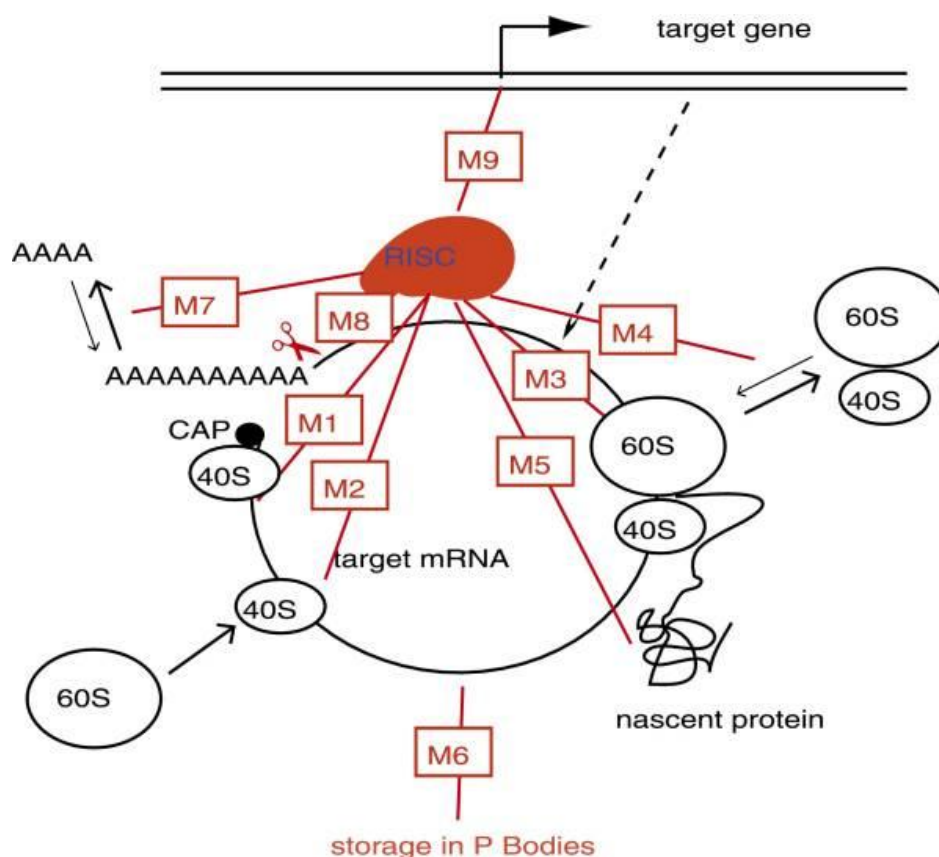
Τα miRNAs είναι μονόκλωνα μόρια RNA μήκους 18 με 25 νουκλεοτίδια, και παρουσιάζουν μία συμπληρωματικότητα με κάποιο σημείο ενός ή και περισσότερων μορίων mRNA. Η σημαντικότερη περιοχή τους είναι το seed region που περιλαμβάνει την περιοχή από το 2^ο μέχρι το 7^ο νουκλεοτίδιο του 5' άκρου τους. Η συμπληρωματικότητα μεταξύ miRNA και mRNA μπορεί να είναι τέλεια ή ατελής. Για τη περίπτωση της ατελής συμπληρωματικότητας υπάρχουν δύο είδη μη τέλειου ταιριάσματος. Το 5' dominant binding προκύπτει από ένα τέλειο ταιρίασμα περίπου 7 βάσεων από το seed region του miRNA, με λίγο ή και καθόλου ταιρίασμα βάσεων από το 3' άκρο του miRNA. Αντίστοιχα, το 3' compensatory binding προκύπτει από ένα ανεπαρκές ταιρίασμα βάσεων από το seed region του miRNA, με ισχυρό ταιρίασμα βάσεων του 3' άκρου. [12]

Η βασική τους λειτουργία σχετίζεται με την ρύθμιση της έκφρασης του γονιδίου μέσω της πρόσδεσής τους στους mRNA στόχους. Στους ζωικούς οργανισμούς, τα σημεία πρόσδεσης των miRNAs βρίσκονται στην περιοχή 3'UTR (3' Untranslated Region) των mRNAs. Ωστόσο έχουν αναφερθεί πολλές περιπτώσεις όπου έχουν βρεθεί στόχοι των miRNAs στην κωδική περιοχή (coding sequence - CDS) των mRNAs [13]. Η συμπληρωματικότητα μεταξύ miRNAs και mRNAs είναι ατελής, με αποτέλεσμα τα μόρια mRNA να μην αποικοδομούνται, αλλά να καταστέλλεται η μετάφρασή τους σε πρωτεΐνη. Αντίθετα, στα φυτά τα miRNAs έχουν στόχους κυρίως στην κωδική περιοχή των mRNAs, και έχουμε τέλεια (ή σχεδόν τέλεια) συμπληρωματικότητα, κάτι που έχει ως αποτέλεσμα την ολοκληρωτική αποικοδόμηση των μορίων mRNA. [14, 15]

Η δράση των miRNAs συσχετίζεται με τους εξής 9 διαφορετικούς μηχανισμούς: [16]

1. Καταστολή έναρξης της μετάφρασης προσδένοντας στην μικρή ριβοσωμική υπομονάδα (40S)
2. Παρεμπόδιση σύνδεσης της μεγάλης ριβοσωμικής υπομονάδας (60S)
3. Καταστολή της επιμήκυνσης (elongation) της μετάφρασης
4. Πρώιμος τερματισμός της μετάφρασης με αποκόλληση του ριβοσώματος (Ribosome drop-off)
5. Διάσπαση του πεπτιδίου κατά τη διάρκεια της μετάφρασης
6. Μεταφορά στα P-bodies

7. Αποικοδόμηση του mRNA
8. Τομή του mRNA
9. Καταστολή της μεταγραφής μέσω αναδιοργάνωσης της χρωματίνης που ακολουθείται από σίγαση των γονιδίων



Εικόνα 2: Μηχανισμοί δράσης των miRNAs [16]

1.1.3 Ρόλος στις ασθένειες

Έχει αποδειχθεί ότι τα miRNAs παίζουν σημαντικό ρόλο σε ένα ευρύ φάσμα αναπτυξιακών διαδικασιών, όπως είναι ο μεταβολισμός, ο πολλαπλασιασμός και η απόπτωση των κυττάρων, το αναπτυξιακό πρόγραμμα, και η εξέλιξη των νευρωνικών κυττάρων. Συσχετίζονται επίσης με την έκφραση νευρωνικών γονιδίων, τη μορφογένεση του εγκεφάλου, τη διαφοροποίηση των μυών, και τη διαίρεση των βλαστοκυττάρων. [17]

Η απορρύθμιση (έλλειψη ή περίσσεια) των miRNAs στα κύτταρα έχει συνδεθεί με πολλές ανθρώπινες ασθένειες. Μεταλλάξεις σε ένα μόλις νουκλεοτίδιο στην ακολουθία του miRNA ή του γονιδίου στόχου του, ή σίγαση της μεταγραφής του miRNA μέσω εξωτερικών παραγόντων, έχει αποδειχθεί ότι μπορούν να επηρεάσουν τη λειτουργία των miRNAs μέσα στα κύτταρα [18].

Τα miRNAs φαίνεται να έχουν έναν καθοριστικό ρόλο στην εξέλιξη του καρκίνου, και υπάρχουν υποψίες πως θα μπορούσαν να συμβάλλουν στη θεραπεία του λόγω της συσχέτισής τους με την ανάπτυξη και τη

διαφοροποίηση των κυττάρων [13, 19]. Έχει αποδειχθεί πως μερικές αλλαγές που ωθούν τα καρκινικά κύτταρα να γίνουν κακοήγη, οφείλονται σε διαφορετικά μοτίβα έκφρασης των miRNAs [20].

Στον Πίνακα 1 παρουσιάζεται μία λίστα από διάφορους τύπους καρκίνων και τα miRNAs που έχουν συσχετιστεί με αυτούς, λόγω των ιδιαίτερων προφίλ έκφρασης που παρουσιάζουν [21]. Η υπερέκφραση των miRNAs στον καρκίνο έχει ως αποτέλεσμα να λειτουργούν ως ογκογονίδια μέσω της καταστολής ογκοκατασταλτικών γονιδίων, ή ως ρυθμιστές διάφορων κυτταρικών διαδικασιών, όπως η διαφοροποίηση και η απόπτωση των κυττάρων [22].

Πίνακας 1: Ο ρόλος των miRNA σε διάφορους τύπους καρκίνου [21]

Cancer Type	miRNAs	Regulation
Breast	miR-21, miR-155, miR-23, miR-191	Up
	miR-205, miR-145, miR-10b, miR-125b	Down
Ovary	miR-200a, miR-200c, miR-141	Up
	miR-199a, miR-140, miR-145, miR-125bl	Down
Endometrioid adenocarcinoma	miR-205, miR-449, miR-429	Up
	miR-193a, miR-99b, miR-204	Down
Colon	miR-135, miR-21, miR-15b, miR-181b, miR-191, miR-200c, miR143, miR145, miR-133b, miR-126	Up
	miR-143, miR-145, miR-133b, miR-126	Down
AML	miR-191, miR-199a, miR-155	Up
CLL	miR-21, miR-155	Up
	miR-15a, miR-16, miR-29, miR-143, miR-45, miR-30d, let-7a, miR-181a	Down
Esophagus	miR-194, miR-192, miR-200c	Up
Gastrointestinal	miR-106b-25	Up
	miR-15b, miR-16	Down
Lung	miR-17-92	Up
	let-7	Down
Bladder	miR-29c, miR-26a, miR-30c, miR-30e-5p, miR-145, miR-30a-3p, miR-133a, miR-133b, miR-195, miR-125b, miR-199a	Down

Πέρα από τον καρκίνο, τα miRNAs έχουν συσχετιστεί με ένα ευρύ φάσμα ασθενειών που περιλαμβάνει τα αυτοάνοσα νοσήματα, την καρδιακή ανεπάρκεια, καρδιαγγειακές παθήσεις, τη φλεγμονή, ασθένειες του

μυοσκελετικού συστήματος, καθώς και του ήπατος. Επιπλέον, υπάρχουν έρευνες που συνδέουν τη δράση τους με νευροαναπτυξιακές και νευροεκφυλιστικές ασθένειες, όπως το σύνδρομο Fragile X (SFX), το σύνδρομο Rett, το σύνδρομο Down, η νόσος του Alzheimer, η νόσος του Huntington, και η σχιζοφρένεια. [17]

1.2 Αλληλεπιδράσεις miRNAs με γονίδια

Μία από τις πιο σημαντικές διαδικασίες στην έρευνα των miRNAs, είναι η αναγνώριση των αλληλεπιδράσεων που έχουν με τα γονίδια που αποτελούν στόχους τους. Η μελέτη τέτοιου είδους αλληλεπιδράσεων παίζει καθοριστικό ρόλο για την βαθύτερη κατανόηση της λειτουργίας τους. Η αναγνώριση τέτοιων αλληλεπιδράσεων μπορεί να γίνει με τη χρήση υπολογιστικών και/ή πειραματικών μεθόδων.

Οι υπολογιστικές μέθοδοι αποτελούν το πρώτο βήμα στην ανάλυση των miRNAs και είναι πιο αποτελεσματικές όσον αφορά την ανακάλυψη άγνωστων miRNA:gene αλληλεπιδράσεων. Ωστόσο η εξέλιξη και βελτίωση των υπολογιστικών προγραμμάτων, αλλά και η δημιουργία καινούργιων βασίζεται σε πειραματικά δεδομένα που τις περισσότερες φορές εμφανίζονται κατακερματισμένα στη σχετική βιβλιογραφία.

Καθώς όμως η σχετική βιβλιογραφία και ο αριθμός των πειραμάτων αυξάνονται συνεχώς με πολύ μεγάλο ρυθμό, η συλλογή και καταγραφή των πειραματικά επιβεβαιωμένων στόχων των miRNAs είναι υψίστης σημασίας. Σήμερα, υπάρχουν διάφορες βάσεις δεδομένων που περιλαμβάνουν τέτοιου είδους αλληλεπιδράσεις και όλες αποτελούν πολύτιμα εργαλεία στο πεδίο έρευνας των miRNAs.

1.2.1 Πειραματικές μέθοδοι

Οι πειραματικές τεχνικές που χρησιμοποιούνται για να επιβεβαιωθούν τα γονίδια-στόχοι των miRNAs μπορούν να διαχωριστούν σε δύο κατηγορίες, τις χαμηλής απόδοσης (low-yield) και τις υψηλής απόδοσης (high-throughput) μεθόδους. Αυτή η κατηγοριοποίηση βασίζεται στον σκοπό εφαρμογής τους και στον αριθμό των δεδομένων που παράγονται σε κάθε πείραμα.

Οι πιο διαδεδομένες χαμηλής απόδοσης τεχνικές περιλαμβάνουν τις Reporter gene assays, quantitative Polymerase Chain Reaction (qPCR), και Western Blot. Η μέθοδος των Reporter gene assays χρησιμοποιείται κυρίως για την επιβεβαίωση ενός σημείου πρόσδεσης του miRNA στο στόχο του, ενώ οι υπόλοιπες χρησιμοποιούνται συνήθως σε συνδυασμό για την να αναγνωρίσουν αλληλεπιδράσεις που προκαλούν αποσύνθεση του mRNA ή/και καταστολή της μετάφρασης.

Η μέθοδος των μικροσυστοιχιών (Microarrays) που χρησιμοποιείται για την αναγνώριση αλληλεπιδράσεων μεταξύ των miRNAs και των γονιδίων στόχων τους, αποτελεί μία υψηλότερης απόδοσης έκδοση των qPCR και Northern Blot μεθόδων. Αντίστοιχα, τεχνικές ποσοτικοποίησης του συνόλου των

κωδικοποιημένων πρωτεϊνών όπως η pSILAC είναι υψηλότερης απόδοσης εκδόσεις των μεθόδων ELISA και Western Blot.

Τα Next-Generation Sequencing (NGS) πειράματα προσφέρουν ένα συνδυασμό υψηλότερης απόδοσης και ακρίβειας, φέρνοντας την επανάσταση στον τρόπο που αναγνωρίζονται οι αλληλεπιδράσεις των miRNAs με τα γονίδια στόχους τους. Οι τεχνικές αυτές βασίζονται σε σημεία στόχους του mRNA που προσδένεται η πρωτεΐνη Argonaute (AGO). Μερικές από τις NGS τεχνικές περιλαμβάνουν τις HITS-CLIP, PAR-CLIP και CLASH.

Πίνακας 2: Πειραματικές τεχνικές αναγνώρισης miRNA:gene αλληλεπιδράσεων [23]

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

Στον Πίνακα 2 παρουσιάζεται μία σύνοψη των διαθέσιμων πειραματικών μεθόδων για την αναγνώριση των αλληλεπιδράσεων μεταξύ των miRNAs και των γονιδίων στόχων τους. Για κάθε μέθοδο γίνεται η κατηγοριοποίηση της σε χαμηλής και/ή υψηλής απόδοσης, και δίνονται κάποια συνοπτικά σχόλια που αφορούν τον σκοπό εφαρμογής της και κατ' επέκταση τα αποτελέσματα που προκύπτουν. [23]

1.2.2 Βάσεις δεδομένων

Η βάση δεδομένων του DIANA-TarBase δημιουργήθηκε το 2006 και είναι μία συλλογή από πειραματικά επιβεβαιωμένες αλληλεπιδράσεις μεταξύ miRNAs και γονιδίων στόχων τους, οι οποίες έχουν αναγνωριστεί χειροκίνητα από ένα πολύ μεγάλο όγκο επιστημονικών δημοσιεύσεων. Κάθε αλληλεπίδραση συνοδεύεται από λεπτομερή μεταδεδομένα, τα οποία παρέχουν πληροφορίες σχετικά με το αποτέλεσμα του πειράματος (θετικό ή αρνητικό αποτέλεσμα), την πειραματική μέθοδο που εφαρμόστηκε, τις πειραματικές συνθήκες συμπεριλαμβανομένου του τύπου κυττάρων και ιστού, το σημείο που προσδένεται το miRNA πάνω στο γονίδιο, καθώς και τις ακολουθίες των primers που χρησιμοποιήθηκαν σε πειράματα κλωνοποίησης.

The screenshot displays the DIANA-TarBase database interface. At the top, a search bar contains the query 'hsa-miR-655'. Below the search bar, a table lists search results. The first row shows 'PTTG1 (hsa)' as the gene name, 'hsa-miR-655' as the miRNA name, and a predicted score of 0.947. The table includes columns for 'Gene name', 'miRNA name', 'Methods', 'Pred.Score', 'Publication', 'Methods', 'Tissue', 'Cell line', 'Tested cell line', and 'Exp. condition'. Below the table, there are several filters and interaction information sections. The 'Filters' section includes options for Species, Method Type, Regulation type, Validation type, and Validated as. The 'Interaction info' section provides details about the location, method, result, regulation, valid type, and source of the interaction. The 'Methods' section lists various methods used in the experiments, such as 'Luciferase Reporter Assay', 'IP', 'WB', and 'RS'. The interface also includes a 'Click (i) for further info' button and a 'Publication year' filter.

Εικόνα 3: Στιγμιότυπο της διεπαφής του DIANA-TarBase. Ο χρήστης μπορεί να καθορίσει τους όρους της αναζήτησής του (1), ώστε να εμφανιστούν όλες οι σχετικές αλληλεπιδράσεις (2). Επεκτείνοντας την λίστα της αλληλεπίδρασης, ή πατώντας το κουμπί (i) δίπλα από κάθε miRNA ή γονίδιο παρουσιάζονται περισσότερες πληροφορίες (4). Ο χρήστης έχει την δυνατότητα να δει όλα τα πειράματα που χρησιμοποιήθηκαν για την επιβεβαίωση της αλληλεπίδρασης (5), καθώς και να περιορίσει τα αποτελέσματα της αναζήτησής του χρησιμοποιώντας τα διαθέσιμα φίλτρα (3). [23]

Η σημερινή έκδοση της βάσης περιλαμβάνει περισσότερες από μισό εκατομμύριο αλληλεπιδράσεις σε 24 διαφορετικούς οργανισμούς. Αυτό το πλήθος των αλληλεπιδράσεων ανταποκρίνεται σε 9 με 250 φορές περισσότερες εγγραφές σε σχέση με κάθε άλλη σχετική βάση δεδομένων. Οι αλληλεπιδράσεις αυτές έχουν επιβεβαιωθεί από 28 διαφορετικές υψηλής ή χαμηλής απόδοσης πειραματικές μεθόδους, σε 356 διαφορετικούς τύπους κυττάρων και 59 διαφορετικούς ιστούς. Πέρα από τις αλληλεπιδράσεις που συλλέχθηκαν χειροκίνητα από επιστημονικές δημοσιεύσεις, περιέχονται και αλληλεπιδράσεις που προέκυψαν από την ανάλυση δεδομένων που προέρχονται από τα NGS πειράματα. [23]

Το miRTarBase [24] αποτελεί τη μοναδική βάση δεδομένων παρόμοια με το DIANA-TarBase, που δίνει εξίσου έμφαση στη συλλογή δεδομένων από μεθόδους υψηλής απόδοσης. Η τελευταία έκδοση της βάσης, η οποία ενημερώθηκε το 2015, περιέχει 366,181 επιβεβαιωμένες αλληλεπιδράσεις των miRNAs και των στόχων τους από πειράματα υψηλής ή χαμηλής απόδοσης που έχουν συλλεχθεί χειροκίνητα από επιστημονικές δημοσιεύσεις. Οι αλληλεπιδράσεις αυτές σχετίζονται με περίπου 22,500 γονίδια καθώς και παραπάνω από 3,500 miRNAs σε 18 διαφορετικούς οργανισμούς.

Πολλές από τις υπόλοιπες σχετικές βάσεις δεδομένων, δεν ενημερώνονται συχνά ή ο αριθμός των αλληλεπιδράσεων που έχουν καταγεγραμμένες είναι πολύ μικρότερος σε σύγκριση με του DIANA-TarBase και του miRTarBase. Η βάση δεδομένων miRecords [25] κυκλοφόρησε το 2009 με στόχο την συλλογή αλληλεπιδράσεων μεταξύ miRNAs και στόχων τους, που έχουν επιβεβαιωθεί κυρίως από χαμηλής απόδοσης πειραματικές τεχνικές. Η τελευταία έκδοση της βάσης που ενημερώθηκε το 2013, περιέχει μόλις 2,705 αλληλεπιδράσεις εκ των οποίων οι 2,028 προέρχονται από χαμηλής απόδοσης μεθόδους.

Υπάρχουν διάφορες άλλες βάσεις δεδομένων στις οποίες έχουν καταχωρηθεί αποτελέσματα από τη CLIP-Seq sequencing μέθοδο, και οι οποίες δίνουν την δυνατότητα της διαδικτυακής ανάλυσης τέτοιων δεδομένων. Μερικές από αυτές είναι οι βάσεις δεδομένων starBase [26] και CLIPZ [27]. Ωστόσο, τέτοιου είδους βάσεις διαφέρουν κατά πολύ από αυτές που αναφέραμε παραπάνω, καθώς ο σκοπός τους είναι η συλλογή των δεδομένων που προέρχονται από την CLIP-Seq sequencing μέθοδο, και των περιοχών πρόσδεσης κάθε RNA binding protein (RBP).

1.3 Microarrays

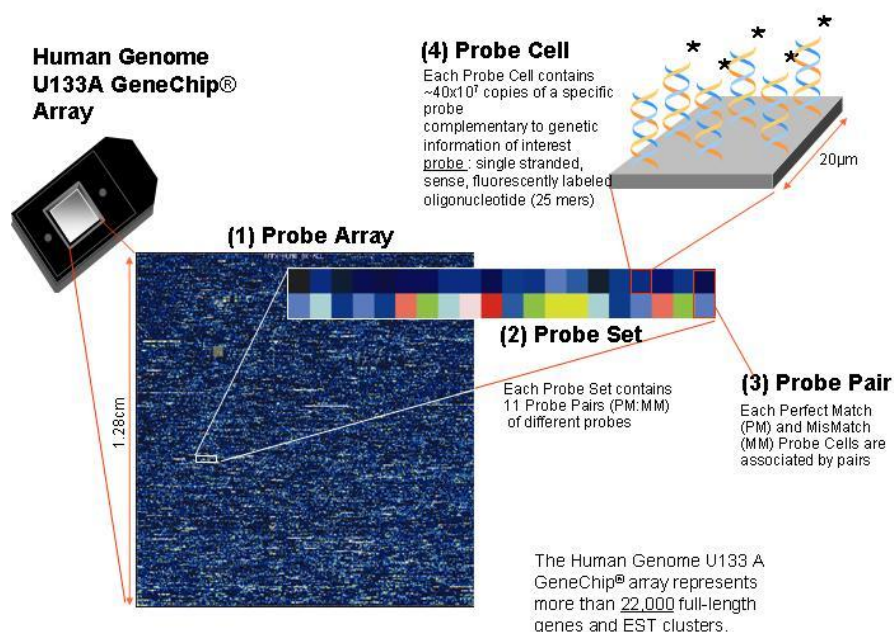
Η αρχή της συμπληρωματικότητας αποτελεί τη βασική ιδέα πάνω στην οποία στηρίζονται τα πιο χρήσιμα εργαλεία της μοριακής βιολογίας, όπως είναι οι μέθοδοι των Northern και Western Blots, της qPCR και της *in situ* υβριδοποίησης. Η τεχνολογία των μικροσυστοιχιών βασίζεται εξίσου στην αρχή της συμπληρωματικότητας με τη διαφορά ότι στόχος της δεν είναι μόνο η ανίχνευση αλλά και η μέτρηση της γονιδιακής έκφρασης. [28] Θεωρείται ως μία από τις πιο αξιόπιστες μεθόδους για τις υψηλής απόδοσης και μεγάλου όγκου γονιδιωματικές αναλύσεις.

Οι μικροσυστοιχίες είναι μία στερεή επιφάνεια στην οποία βρίσκονται σε συγκεκριμένη διάταξη χιλιάδες ανιχνευτές, που είναι συμπληρωματικοί προς

διαφορετικά τμήματα DNA ή RNA. Ο αρχικός σχεδιασμός της τεχνολογίας είχε ως στόχο την εξέταση της έκφρασης πολλών γονιδίων ταυτόχρονα. Ωστόσο υπάρχει ένα μεγάλο εύρος εφαρμογών, που περιλαμβάνει τη χρήση τους στην έρευνα των miRNAs. Αποτελεί μία από τις πιο διαδεδομένες τεχνολογίες για την ανάλυση της επίδρασης που έχει η υπερέκφραση ή καταστολή του miRNA στην έκφραση του γονιδίου.

1.3.1 Περιγραφή τεχνολογίας

Οι μικροσυστοιχίες αποτελούνται από τους ανιχνευτές (probes), οι οποίοι περιέχουν μία συγκεκριμένη γονιδιακή ακολουθία και χρησιμοποιούνται για να δεσμεύσουν με ομοιοπολικούς δεσμούς τα συμπληρωματικά δείγματα DNA ή RNA. Με βάση την επιμέρους δομή των ανιχνευτών τους, οι μικροσυστοιχίες διαχωρίζονται σε δύο κατηγορίες. Η μία κατηγορία περιλαμβάνει ανιχνευτές με μονόκλωνα ολιγονουκλεοτίδια μήκους περίπου 22 βάσεων τα οποία έχουν δημιουργηθεί *in situ*, ενώ η άλλη κατηγορία περιλαμβάνει ανιχνευτές με συμπληρωματικά δείγματα του γονιδιώματος που έχουν προκύψει από την αντίστροφη μεταγραφή των mRNAs των γονιδίων. Οι μικροσυστοιχίες με ανιχνευτές ολιγονουκλεοτιδίων, αποτελούν την προτεινόμενη τεχνολογία χρήσης και βασικός προμηθευτής της είναι η Affymetrix. [28]

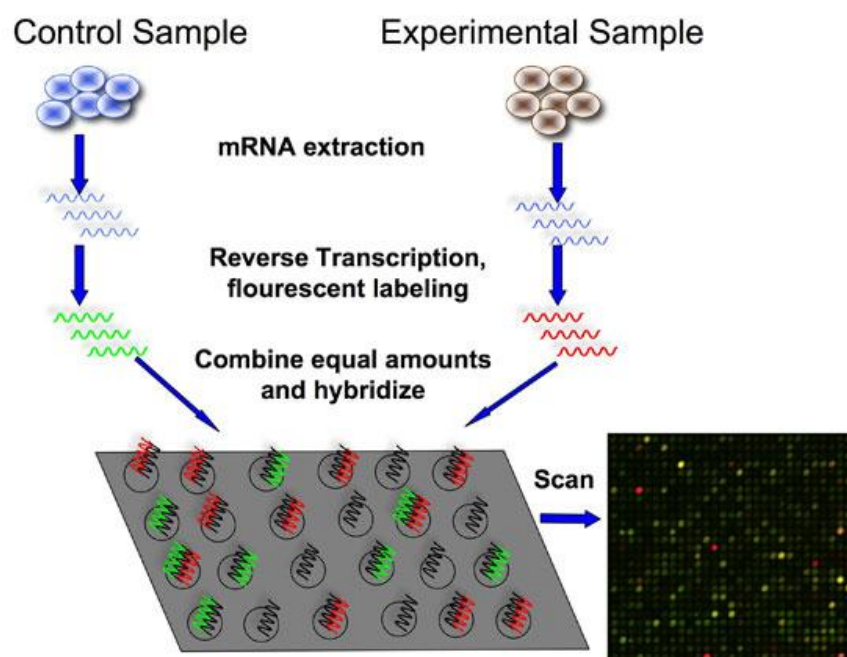


Εικόνα 4: Γενωμικό chip Affymetrix

Το γενωμικό chip της Affymetrix αποτελείται από σύνολα ανιχνευτών (probe sets) που το καθένα περιλαμβάνει συνήθως 11 με 20 ανιχνευτές, καθένας από τους οποίους περιέχει εκατοντάδες αντίγραφα ενός ολιγονουκλεοτιδίου μήκους περίπου 25 βάσεων. Οι ανιχνευτές συνήθως σχεδιάζονται έτσι ώστε να αντιστοιχίζονται σε ζεύγη, με κάθε ζεύγος να αποτελείται από έναν

ανιχνευτή ολιγονουκλεοτιδίων με τέλεια αντιστοιχία (Perfect Match, PM) και έναν με μερική αντιστοιχία (MisMatch, MM). Τα ολιγονουκλεοτίδια του PM ανιχνευτή είναι από άκρη σε άκρη συμπληρωματικά προς μια γονιδιακή περιοχή, ενώ τα ολιγονουκλεοτίδια του MM έχουν υποστεί μία μετάλλαξη στη βάση που βρίσκεται στο κέντρο τους. Η χρήση των MM ανιχνευτών έχει σκοπό να συμβάλλει στην διαδικασία της εκτίμησης του θορύβου. [29]

Η διαδικασία που ακολουθείται στα πειράματα των μικροσυστοιχιών φαίνεται στην παρακάτω εικόνα. Αρχικά, το συμπληρωματικό DNA ή mRNA που απομονώθηκε από τον ιστό ή τα κύτταρα του δείγματος προς ανάλυση, επισημαίνεται με τη χρήση φθορίζουσών ουσιών. Τα επισημασμένα αυτά δείγματα σχηματίζουν δεσμούς υδρογόνου πάνω στις μικροσυστοιχίες. Οι φθορίζουσες ουσίες διεγείρονται με τη χρήση ενός λέιζερ, ώστε ένας σαρωτής να εντοπίσει το φως που παράγεται και να απεικονίσει τις μικροσυστοιχίες σε ψηφιακή εικόνα. Η εικόνα υφίσταται περαιτέρω επεξεργασία από εξειδικευμένα προγράμματα, με στόχο να ποσοτικοποιηθεί το σήμα της έντασης φθορισμού που εκπέμπει κάθε ανιχνευτής. Η ποσοτικοποίηση του σήματος έχει ως στόχο να συσχετιστούν οι ανιχνευτές με αριθμητικά δεδομένα. [28]



Εικόνα 5: Διαδικασία πειράματος μικροσυστοιχιών

1.3.2 Κανονικοποίηση δεδομένων

Σε πολλές από τις διαδικασίες του πειράματος των μικροσυστοιχιών, όπως είναι η υβριδοποίηση, η επισημάνση και η επεξεργασία εικόνας, συμβαίνουν συστηματικά λάθη που επηρεάζουν τον εντοπισμό και την μέτρηση της γονιδιακής έκφρασης. Αυτό καθιστά απαραίτητη την κανονικοποίηση των δεδομένων, ώστε να εξασφαλιστεί η εγκυρότητα της βιολογικής πληροφορίας

και να δημιουργηθούν τιμές που μπορούν να συγκριθούν μεταξύ τους, παρά το γεγονός ότι προέρχονται από διαφορετικά πειράματα. [28]

Υπάρχουν διάφορες μέθοδοι κανονικοποίησης, και η επιλογή της κατάλληλης μεθόδου εξαρτάται συνήθως από τα δεδομένα [28]. Παρόλο που οι μικροσυστοιχίες γονιδίων είναι μία καθιερωμένη μεθοδολογία για την εξέταση της γονιδιακής έκφρασης, η κανονικοποίηση των δεδομένων εξακολουθεί να αποτελεί μία ανοικτή πρόκληση, καθοριστική για τα συμπεράσματα του πειράματος και το σχεδιασμό των επόμενων βημάτων της πειραματικής διαδικασίας. Σε αυτή την ενότητα θα παρουσιαστούν οι τρεις πιο διαδεδομένες μέθοδους κανονικοποίησης που περιλαμβάνουν την RMA, τη gcRMA καθώς και τη μέθοδο MAS5 της Affymetrix.

1.3.2.1 MAS5 αλγόριθμος

Η Affymetrix για να συσχετίσει τα σύνολα ανιχνευτών με αριθμητικά δεδομένα, ανέπτυξε αρχικά τον MAS4 (MicroArray Suite 4) αλγόριθμο. Για κάθε σύνολο ανιχνευτών υπολογίζεται η μέση τιμή της διαφοράς που προκύπτει αν από την τιμή έντασης φθορισμού που αντιστοιχεί στον PM ανιχνευτή αφαιρεθεί η τιμή που αντιστοιχεί στον MM ανιχνευτή. Ακραίες τιμές που απέχουν απόσταση μεγαλύτερη από τρεις φορές την τυπική απόκλιση της μέσης τιμής απορρίπτονται. Ένα πρόβλημα της συγκεκριμένης μεθόδου ήταν ότι πολλές φορές οι MM ανιχνευτές είχαν μεγαλύτερες τιμές για την ένταση φθορισμού από τους PM ανιχνευτές. Αυτό είχε ως αποτέλεσμα περίπου για το 30% των γονιδίων να υπολογίζεται μία αρνητική τιμή για την ένταση του φθορισμού. Κατά συνέπεια, αυτή η μέθοδος εγκαταλείφθηκε σύντομα και αντικαταστήθηκε από τη MAS5. Η μέθοδος MAS5 είναι μία βελτιωμένη έκδοση της MAS4 στην οποία έχει προστεθεί μία συνθήκη για την πρόληψη των αρνητικών τιμών έντασης φθορισμού.

Η μέθοδος MAS5 υπολογίζει για κάθε ζεύγος ανιχνευτών (PM και MM) ένα βάρος που σχετίζεται με την απόσταση που έχει το ζεύγος από την διάμεσο του συνόλου ανιχνευτών στο οποίο ανήκει, χρησιμοποιώντας τον One-Step Tukey's Biweight αλγόριθμο. Για κάθε σύνολο ανιχνευτών υπολογίζεται μία τιμή που αντιστοιχεί στη μέση τιμή των βαρών που έχει κάθε ζεύγος που περιλαμβάνεται στο συγκεκριμένο σύνολο. Αν όλες οι τιμές της έντασης φθορισμού των PM ανιχνευτών είναι μεγαλύτερες από τις τιμές των MM ανιχνευτών, τότε η διαδικασία που ακολουθεί ο MAS5 δεν διαφοροποιείται καθόλου από αυτή του MAS4. Αν όμως υπάρχουν μερικές ή πολλές τιμές έντασης φθορισμού των MM ανιχνευτών που είναι μεγαλύτερες από τις τιμές των PM ανιχνευτών τότε διακρίνονται δύο περιπτώσεις. Στη περίπτωση των μερικών τιμών, αντί για τη διαφορά των τιμών έντασης φθορισμού του ζεύγους, χρησιμοποιούνται τα βάρη που έχουν υπολογιστεί για το συγκεκριμένο ζεύγος. Στη περίπτωση των πολλών τιμών, η τιμή της έντασης φθορισμού του MM ανιχνευτή αντικαθίσταται με μία νέα τιμή λίγο μικρότερη από την αντίστοιχη τιμή του PM ανιχνευτή. Η συνοπτική τιμή της έντασης φθορισμού μετατρέπεται σε λογάριθμο έτσι ώστε να μειωθεί η διακύμανση. [30]

1.3.2.2 RMA αλγόριθμος

Το πρόβλημα που προκύπτει με τις τιμές έντασης φθορισμού των MM ανιχνευτών, δεν καθιστά τη χρήση τους ως έναν αξιόπιστο εκτιμητή του θορύβου. Αυτός ήταν ένας από τους σημαντικούς λόγους που οδήγησαν την ερευνητική κοινότητα προς αναζήτηση νέων αλγορίθμων για την ανάλυση των δεδομένων που παράγονται από τις μικροσυστοιχίες. Ένας ακόμη λόγος για την ανάγκη της ανάπτυξης νέων αλγορίθμων, ήταν το γεγονός ότι η εφαρμογή της κανονικοποίησης είναι προτιμότερο να γίνεται σε επίπεδο ενός ανιχνευτή και όχι ενός συνόλου ανιχνευτών. Σε αυτή τη λογική στηρίζεται ο RMA (Robust Multi-Array Analysis) αλγόριθμος, ο οποίος χρησιμοποιεί μόνο τα δεδομένα των PM ανιχνευτών. Για την εκτίμηση του θορύβου του υποβάθρου θεωρείται ότι οι τιμές των PM ανιχνευτών είναι μία μίξη του πραγματικού σήματος, του οπτικού θορύβου και του σήματος της μη ειδικής υβριδοποίησης. Ο αλγόριθμος για να συσχετίσει τους ανιχνευτές με αριθμητικά δεδομένα ακολουθεί τρία βασικά βήματα.

Το πρώτο βήμα του αλγορίθμου πραγματοποιεί για κάθε chip ξεχωριστά μία μη γραμμική διόρθωση του υποβάθρου, με βάση την κατανομή που ακολουθούν οι τιμές των PM ανιχνευτών. Η εκτίμηση του θορύβου του υποβάθρου που περιλαμβάνει τον οπτικό θόρυβο και την μη ειδική υβριδοποίηση, γίνεται χρησιμοποιώντας τη μέθοδο της εκτίμησης πυκνότητας με χρήση πυρήνων (Kernel Density Estimation). Ο εκτιμώμενος θόρυβος υποβάθρου που προκύπτει αφαιρείται από τις τιμές των PM ανιχνευτών. Στο επόμενο βήμα, εφαρμόζεται μία κανονικοποίηση ποσοστημορίων σε επίπεδο ανιχνευτή, που μετασχηματίζει τις κατανομές των PM ανιχνευτών ώστε να αποκτήσουν παρόμοιες στατιστικές ιδιότητες σε όλα τα chips. Στο τελευταίο βήμα, γίνεται η παραδοχή ότι οι μετασχηματισμένες σε λογάριθμο τιμές των PM ανιχνευτών ακολουθούν ένα γραμμικό προσθετικό μοντέλο, το οποίο δίνει ως έξοδο μία τιμή έκφρασης για κάθε σύνολο ανιχνευτών σε κάθε chip. Για την εκτίμηση των παραμέτρων του μοντέλου ακολουθείται μία διαδικασία κανονικοποίησης σε όλα τα chips, η οποία βασίζεται στη διάμεσο (median polish) και εξασφαλίζει την αποφυγή των ακραίων τιμών. [29]

1.3.2.3 gcRMA αλγόριθμος

Ο gcRMA αλγόριθμος βασίζεται στη γενική αρχή του RMA αλγορίθμου με τη διαφοροποίηση ότι χρησιμοποιεί τις ακολουθίες του κάθε ανιχνευτή για να πραγματοποιήσει μία διόρθωση υποβάθρου με βάση την περιεκτικότητά τους σε GC ζεύγη βάσεων. Για κάθε ανιχνευτή μοντελοποιείται η δυνατότητα που έχει για να σχηματίσει δεσμούς με την συμπληρωματική του ακολουθία. Η μοντελοποίηση γίνεται με βάση την επίδραση που έχει κάθε μεμονωμένη βάση στην αντίστοιχη θέση της μέσα στον ανιχνευτή. Αυτό επιτρέπει την διόρθωση του υποβάθρου χρησιμοποιώντας τις τιμές των PM καθώς και των MM ανιχνευτών. Για την εκτίμηση του θορύβου του υποβάθρου θεωρείται ότι τόσο οι τιμές των PM ανιχνευτών όσο και οι τιμές των MM ανιχνευτών, αποτελούν μίξη του πραγματικού σήματος, του οπτικού θορύβου και της μη ειδικής υβριδοποίησης.

Η διαδικασία της διόρθωσης του υποβάθρου μπορεί να διαχωριστεί σε τρία βήματα. Στο πρώτο βήμα, γίνεται η διόρθωση του οπτικού θορύβου που προκαλείται κατά τη διάρκεια της απεικόνισης των μικροσυστοιχιών σε ψηφιακή εικόνα. Στο δεύτερο βήμα, οι τιμές έντασης φθορισμού των ανιχνευτών προσαρμόζονται με βάση την πληροφορία για την δυνατότητα τους να σχηματίσουν δεσμό με τις συμπληρωματικές τους ακολουθίες, καθώς και τις διορθωμένες από οπτικό θόρυβο τιμές των MM ανιχνευτών. Κατά αυτό τον τρόπο, πραγματοποιείται η διόρθωση της μη ειδικής υβριδοποίησης. Στο τελευταίο βήμα, οι τιμές των PM ανιχνευτών διορθώνονται με βάση τη δυνατότητά τους για σχηματισμό δεσμού με τις συμπληρωματικές τους ακολουθίες, καθώς και αφαιρούνται από αυτές οι διορθωμένες από οπτικό θόρυβο τιμές των MM ανιχνευτών. Στη συνέχεια ακολουθούν οι διαδικασίες της κανονικοποίησης ποσοστιμορίων και του μετασχηματισμού των ανιχνευτών σε τιμές έκφρασης, οι οποίες δεν διαφέρουν καθόλου από τις αντίστοιχες του RMA αλγορίθμου. [31]

1.3.3 Υπολογισμός τιμής έκφρασης

Μετά τον μετασχηματισμό των τιμών έντασης φθορισμού, κάθε γονίδιο αντιστοιχίζεται με έναν αριθμό από τιμές έκφρασης που προκύπτουν από τα διαφορετικά σύνολα ανιχνευτών που αναπαριστούν το ίδιο γονίδιο. Ο αριθμός αυτός είναι κυμαινόμενος και εξαρτάται από τον σχεδιασμό του chip που χρησιμοποιήθηκε στο πείραμα των μικροσυστοιχιών. Ωστόσο για την περαιτέρω ανάλυση της γονιδιακής έκφρασης, κάθε γονίδιο πρέπει να συσχετιστεί με μία τελική τιμή έκφρασης που μπορεί να υπολογιστεί από τις επιμέρους τιμές που αντιστοιχούν στο γονίδιο.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις για τον υπολογισμό της τελικής τιμής έκφρασης του γονιδίου. Η μέθοδος που χρησιμοποιείται συνήθως είναι η μέση τιμή ή διάμεσος των αντίστοιχων τιμών έκφρασης. Ωστόσο υπάρχουν και περιπτώσεις που κάποιοι ερευνητές επιλέγουν να μην συμπεριλάβουν καθόλου στην ανάλυσή τους τα γονίδια που αντιστοιχίζονται με παραπάνω από μία τιμή. Στα πλαίσια αυτής της εργασίας, για τον συσχετισμό του γονιδίου με μία τελική τιμή έκφρασης εφαρμόστηκε μία προσέγγιση που βασίζεται στον κανόνα της πλειοψηφίας (Majority Rule).

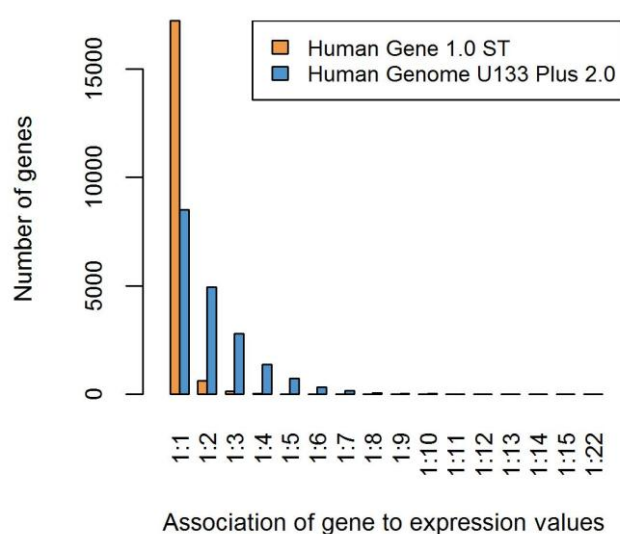
Για κάθε γονίδιο γίνεται μία ψηφοφορία μεταξύ των αντίστοιχων τιμών έκφρασής του, κατά την οποία αποφασίζεται με βάση τον κανόνα της πλειοψηφίας αν υπάρχει υπερέκφραση ή καταστολή του γονιδίου. Η τελική τιμή έκφρασης είναι η μέση τιμή ή διάμεσος που προκύπτει μόνο από το σύνολο των τιμών που υποστήριξαν την απόφαση του κανόνα. Για παράδειγμα, αν η πλειοψηφία των τιμών αποφασίζει την καταστολή του γονιδίου τότε η επιλογή της μέσης τιμής ή διαμέσου θα γίνει μόνο από τις αντίστοιχες τιμές που υποδεικνύουν καταστολή.

2. ΑΠΟΤΕΛΕΣΜΑΤΑ

2.1 Επεξεργασία δεδομένων από μικροσυστοιχίες

Βασικός στόχος της παρούσας διπλωματικής εργασίας αποτέλεσε η σύγκριση της κλασικής προσέγγισης της μέσης τιμής και διαμέσου (Average), με την προσέγγιση που βασίζεται στον κανόνα της πλειοψηφίας (Majority). Για χάρη αυτής της σύγκρισης, χρησιμοποιήθηκε ένας αριθμός από υψηλής ποιότητας δεδομένα των miRNA μικροσυστοιχιών που επιλέχθηκαν από ευρέως αναγνωρισμένες επιστημονικές δημοσιεύσεις. Η συλλογή των δεδομένων πραγματοποιήθηκε από τα γενωμικά chips της Affymetrix Human Gene 1.0 ST (HG-1.0-ST) και Human Genome U133 Plus 2.0 (HG-U133-Plus-2.0).

Ο σχεδιασμός των δύο chips διαφέρει και κατ' επέκταση, ο αριθμός των τιμών έκφρασης που αντιστοιχούν σε κάθε γονίδιο. Στο σχήμα που ακολουθεί απεικονίζεται ο αριθμός των γονιδίων για κάθε αντιστοιχία γονιδίου προς τιμές έκφρασης (γονίδιο:τιμές). Η μέγιστη αντιστοιχία γονιδίου προς τιμές έκφρασης είναι 1:22, η οποία παρουσιάζεται στο HG-U133-Plus-2.0. Αντίστοιχα στο HG-1.0-ST η μέγιστη τιμή είναι 1:10. Μία σημαντική παρατήρηση που πρέπει να επισημανθεί είναι ότι στο HG-U133-Plus-2.0 περίπου το 45% των γονιδίων αντιστοιχίζονται με 1 τιμή έκφρασης, ενώ στο HG-1.0-ST το αντίστοιχο ποσοστό ισούται περίπου με 96%.



Σχήμα 1: Αριθμός γονιδίων για κάθε αντιστοιχία γονιδίου:τιμές έκφρασης

2.1.1 Εφαρμογή μεθόδων

Για την εφαρμογή των μεθόδων χρησιμοποιήθηκαν συνολικά 29 σύνολα δεδομένων. Για τη μοναδική ονομασία κάθε συνόλου δημιουργήθηκε ένα αναγνωριστικό, με βάση τον συγγραφέα της δημοσίευσης, τα κύτταρα προέλευσης του δείγματος, το υπό μελέτη miRNA και τη replication μέθοδο.

Στον Πίνακα 3 για κάθε σύνολο παρουσιάζονται κάποιες πληροφορίες σχετικά με το miRNA που μελετήθηκε στα πλαίσια του εκάστοτε πειράματος, τη συνθήκη υπό την οποία εξετάστηκε η γονιδιακή έκφραση, καθώς και το chip από το οποίο προέρχεται.

Πίνακας 3: Datasets σύγκρισης

Dataset	miRNA	miRNA Treatment		Chip	
		<i>Silencing</i>	<i>Overexpression</i>	<i>HG-1.0-ST</i>	<i>HG-U133-Plus-2.0</i>
bossel-miR20a-0.5h [32]	hsa-miR-20a-5p	✓		✓	
bossel-miR20a-0h [32]	hsa-miR-20a-5p	✓		✓	
bossel-miR20a-1h [32]	hsa-miR-20a-5p	✓		✓	
bossel-miR20a-2h [32]	hsa-miR-20a-5p	✓		✓	
bossel-miR671-0.5h [32]	hsa-miR-671-5p	✓		✓	
bossel-miR671-0h [32]	hsa-miR-671-5p	✓		✓	
bossel-miR671-1h [32]	hsa-miR-671-5p	✓		✓	
bossel-miR671-2h [32]	hsa-miR-671-5p	✓		✓	
rao-si-mir222 [33]	hsa-miR-222-3p	✓			✓
rao-si-mir221 [33]	hsa-miR-221-3p	✓			✓
selbach_8h_let [34]	hsa-let-7b-5p		✓		✓
selbach_8h_mir1 [34]	hsa-miR-1-3p		✓		✓
selbach_8h_mir16 [34]	hsa-miR-16-5p		✓		✓
selbach_8h_mir30a [34]	hsa-miR-30a-5p		✓		✓
selbach_8h_mir155 [34]	hsa-miR-155-5p		✓		✓
selbach_32h_let [34]	hsa-let-7b-5p		✓		✓
selbach_32h_mir1 [34]	hsa-miR-1-3p		✓		✓
selbach_32h_mir16 [34]	hsa-miR-16-5p		✓		✓
selbach_32h_mir30a [34]	hsa-miR-30a-5p		✓		✓
selbach_32h_mir155 [34]	hsa-miR-155-5p		✓		✓
hausser-mir7 [35]	hsa-miR-7-5p		✓		✓
hausser-mir124 [35]	hsa-miR-124-3p		✓		✓
frankel-mir101 [36]	hsa-miR-101-3p		✓		✓
frankel-2014-miR95 [37]	hsa-miR-95-3p		✓		✓
hollander-mir212 [38]	hsa-miR-212-3p		✓		✓
hu-miR941-HEK293 [39]	hsa-miR-941		✓		✓
hu-miR941-HEK293T [39]	hsa-miR-941		✓		✓
schneider-12h-miR28 [40]	hsa-miR-28-5p		✓		✓
schneider-24h-miR28 [40]	hsa-miR-28-5p		✓		✓

Από τους δύο αλγόριθμους που εφαρμόστηκαν, μόνο ο αλγόριθμος της μεθόδου Average υλοποιήθηκε στα πλαίσια της παρούσας εργασίας, καθώς ο αλγόριθμος της μεθόδου Majority ήταν ήδη υλοποιημένος. Ο αλγόριθμος αυτός παίρνει σαν είσοδο καθένα από τα σύνολα δεδομένων και για κάθε γονίδιο του υπολογίζει δύο τελικές τιμές έκφρασης, τη μέση τιμή και τη διάμεσο όλων των τιμών έκφρασής του. Επομένως στην έξοδο του αλγόριθμου, κάθε γονίδιο αντιστοιχίζεται με δύο τιμές. Οι τιμές αυτές είναι οι τελικές τιμές έκφρασης που προέκυψαν από τη μέση τιμή και τη διάμεσο.

Δεδομένου ότι η εφαρμογή των δύο μεθόδων δεν πρόκειται να επιφέρει κάποια αλλαγή στον υπολογισμό της τελικής τιμής έκφρασης των γονιδίων που προέρχονται από τις 1:1 αντιστοιχίες, έπρεπε να αφαιρεθούν τα συγκεκριμένα γονίδια από τα δεδομένα. Για το σκοπό αυτό, υλοποιήθηκε αλγόριθμος που τα επεξεργάζεται πριν την εφαρμογή οποιασδήποτε μεθόδου, ώστε να δημιουργήσει για καθένα από αυτά μία λίστα με τα γονίδια που προέρχονται από τέτοιες αντιστοιχίες. Η λίστα αυτή χρησιμοποιείται ώστε να αφαιρεθούν από το κάθε σύνολο δεδομένων τα αντίστοιχα γονίδια.

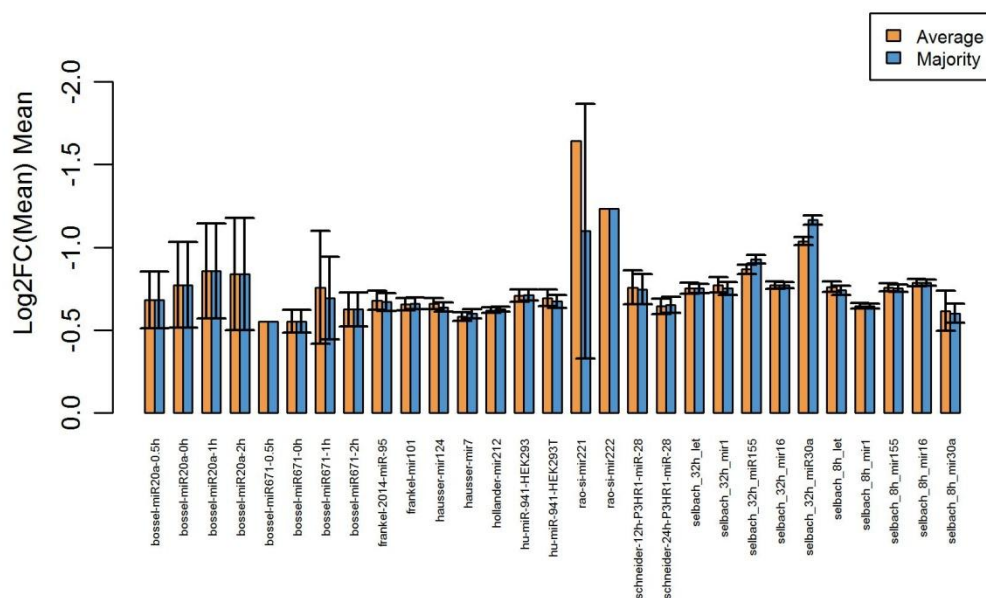
Μετά την απαλοιφή των γονιδίων με τις 1:1 αντιστοιχίες, υλοποιήθηκε ένας αλγόριθμος για να αναπαρασταθεί γραφικά η κατανομή των τελικών τιμών έκφρασης των γονιδίων του κάθε συνόλου δεδομένων μετά την εφαρμογή των υπό μελέτη μεθόδων. Για την αναπαράσταση της κατανομής, ο αλγόριθμος υπολογίζει τη μέση τιμή και την τυπική απόκλιση των τελικών τιμών έκφρασης με τη χρήση δύο κατώφλιων. Τα κατώφλια χρησιμοποιούνται για να ξεχωρίσουν τα γονίδια με στατιστικά σημαντική μεταβολή στην έκφρασή τους από εκείνα των οποίων η έκφραση παρέμεινε αμετάβλητη.

Με βάση αυτά τα κατώφλια, ο αλγόριθμος διαχωρίζει το εκάστοτε σύνολο δεδομένων σε δύο υποσύνολα. Τα υποσύνολα αυτά αποτελούνται από τελικές τιμές έκφρασης που υποδεικνύουν υπερέκφραση και καταστολή αντίστοιχα. Οι τιμές που επιλέχθηκαν ως κατώφλια είναι η 0.5 για την υπερέκφραση και η -0.5 για την καταστολή. Η μέση τιμή και η τυπική απόκλιση των τελικών τιμών έκφρασης υπολογίζεται ξεχωριστά για το κάθε υποσύνολο. Στα σχήματα που ακολουθούν φαίνονται οι γραφικές παραστάσεις της κατανομής των τελικών τιμών έκφρασης που δίνονται ως έξοδος από τον αλγόριθμο.

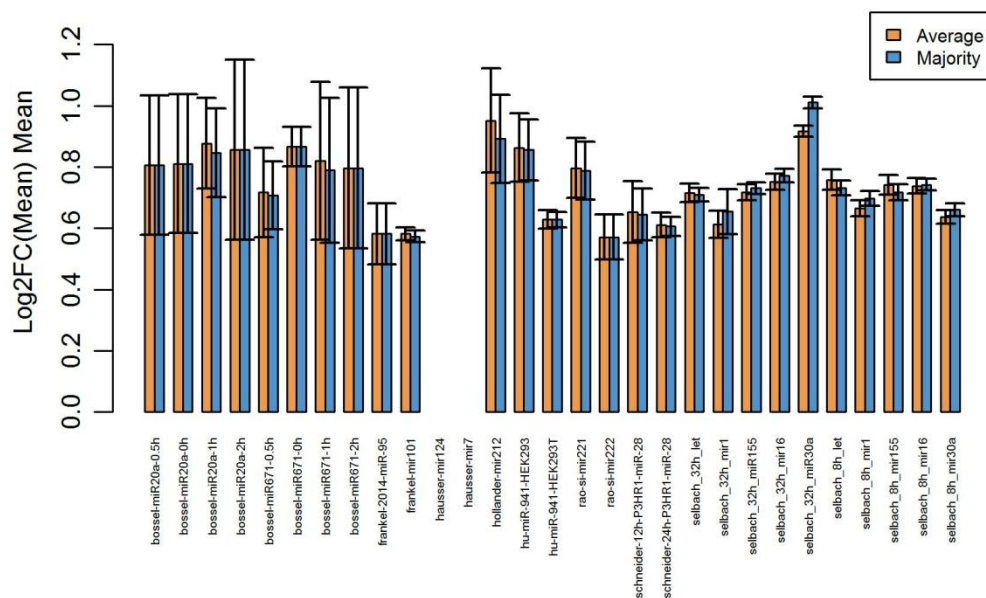
Το ύψος της μπάρας αναπαριστά τη μέση τιμή έκφρασης του συνόλου των τελικών τιμών έκφρασης, ενώ η κατακόρυφη γραμμή κάθε μπάρας συμβολίζει τη διασπορά του συνόλου από τη μέση τιμή έκφρασης. Στα Σχήματα 2 και 3, απεικονίζεται η κατανομή των τελικών τιμών έκφρασης που προκύπτουν από τη μέση τιμή για τα υποσύνολα των γονιδίων που υποδεικνύουν καταστολή και υπερέκφραση αντίστοιχα. Τα Σχήματα 4 και 5, απεικονίζουν την κατανομή των τελικών τιμών έκφρασης που προκύπτουν από τη διάμεσο για καθένα από τα δύο υποσύνολα της καταστολής και της υπερέκφρασης.

Γενικά παρατηρείται ότι στην πλειοψηφία των περιπτώσεων η μέση τιμή έκφρασης του κάθε συνόλου δεδομένων της Majority, δεν διαφέρει πολύ από την μέση τιμή έκφρασης του αντίστοιχου συνόλου της Average. Η ίδια παρατήρηση ισχύει και για τη διασπορά των τιμών έκφρασης ανά σύνολο των δύο μεθόδων. Ωστόσο, διακρίνονται κάποιες περιπτώσεις συνόλων δεδομένων στις οποίες δεν απεικονίζεται καμία μπάρα. Αυτό οφείλεται στο γεγονός ότι το υποσύνολο της καταστολής/υπερέκφρασης αυτών των

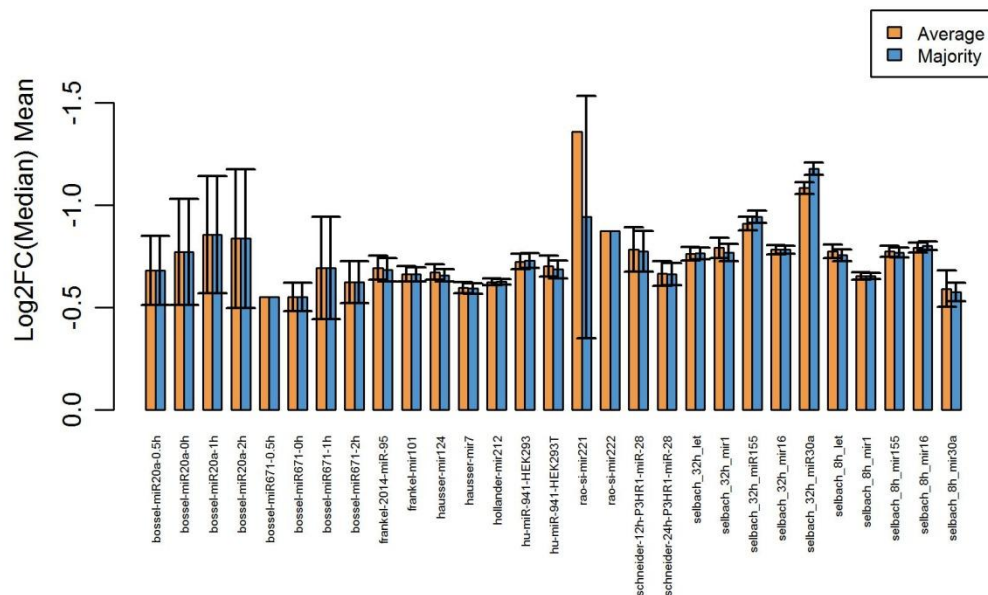
συνόλων δεν περιλαμβάνει κάποιο γονίδιο, καθώς δεν υπήρχε σημαντική μεταβολή στη γονιδιακή τους έκφραση.



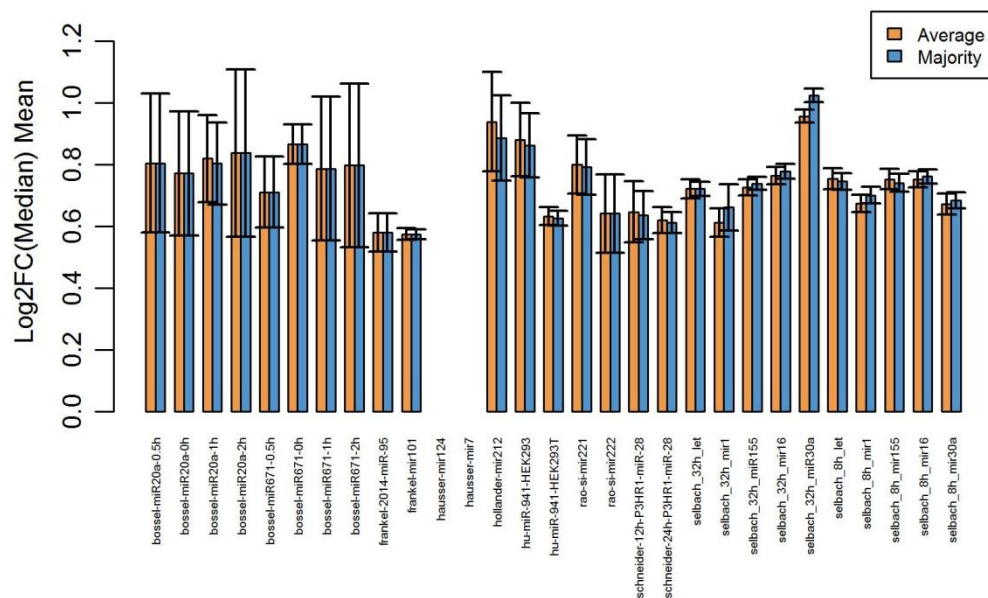
Σχήμα 2: Κατανομή τιμών έκφρασης από μέση τιμή για καταστολή



Σχήμα 3: Κατανομή τιμών έκφρασης από μέση τιμή για υπερέκφραση



Σχήμα 4: Κατανομή τιμών έκφρασης από διάμεσο για καταστολή



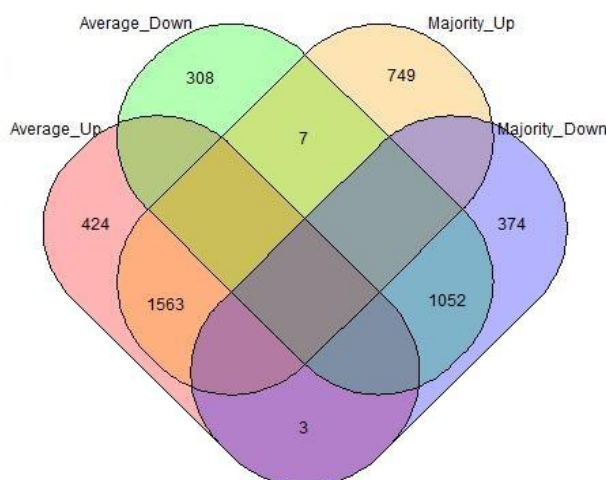
Σχήμα 5: Κατανομή τιμών έκφρασης από διάμεσο για υπερέκφραση

2.1.2 Σύγκριση μεθόδων

Για τη σύγκριση των δύο μεθόδων χρησιμοποιήθηκαν αρχικά τα διαγράμματα Venn, καθώς δύνανται να μελετηθούν όλες οι πιθανές συσχετίσεις μεταξύ των συνόλων που αναπαριστούν. Βασικός στόχος της μελέτης αυτών των συσχετίσεων ήταν να διερευνηθεί το ενδεχόμενο σύγκρουσης μεταξύ των δύο μεθόδων σχετικά με την έκφραση ενός γονιδίου. Πιο συγκεκριμένα, οι περιπτώσεις σύγκρουσης που διερευνήθηκαν είναι η περίπτωση όπου η Majority κρίνει την υπερέκφραση ενός γονιδίου, για το οποίο η Average κρίνει την καταστολή του, και αντίστροφα η περίπτωση όπου η Majority κρίνει την καταστολή ενός γονιδίου, ενώ η Average κρίνει την υπερέκφρασή του.

Για την υλοποίηση των Venn διαγραμμάτων, αναπτύχθηκε αλγόριθμος που υπολογίζει με τη χρήση των κατωφλίων τα 2 υποσύνολα της υπερέκφρασης και καταστολής που αντιστοιχούν στις τελικές τιμές έκφρασης του κάθε συνόλου δεδομένων. Στη συνέχεια, για κάθε σύνολο δημιουργήθηκαν 2 συλλογές, καθεμία από τις οποίες αποτελείται από 4 σύνολα. Η μία συλλογή περιλαμβάνει τα 2 υποσύνολα της Majority και τα 2 υποσύνολα της Average που προέκυψαν από τη μέση τιμή των τελικών τιμών έκφρασης. Η άλλη συλλογή περιλαμβάνει τα 2 υποσύνολα της Majority και τα 2 υποσύνολα της Average που προέκυψαν από τη διάμεσο των τελικών τιμών έκφρασης.

Σε καθεμία από αυτές τις συλλογές εφαρμόζεται ένας ήδη υλοποιημένος αλγόριθμος, ο οποίος υπολογίζει και αναπαριστά σε ένα διάγραμμα όλες τις πιθανές συσχετίσεις μεταξύ των συνόλων κάθε συλλογής. Ο συνολικός αριθμός των διαγραμμάτων που δημιουργήθηκαν είναι 58, εκ των οποίων τα 29 αναπαριστούν τα υποσύνολα που υπολογίζονται από τις τελικές τιμές έκφρασης της μέσης τιμής, ενώ τα άλλα 29 απεικονίζουν τα υποσύνολα που υπολογίζονται από τις τελικές τιμές έκφρασης της διαμέσου. Δεδομένου ότι ο αριθμός των διαγραμμάτων είναι πολύ μεγάλος, θα παρουσιαστεί μόνο 1 ενδεικτικό διάγραμμα το οποίο αξίζει να επισημανθεί.



Σχήμα 6: Venn διάγραμμα τιμών έκφρασης από μέση τιμή

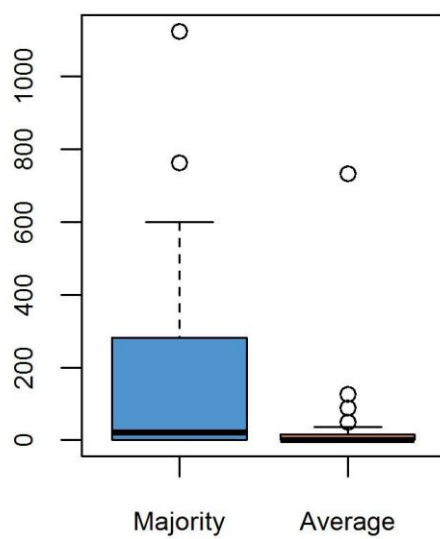
Στο Σχήμα 6 απεικονίζεται το διάγραμμα Venn που προκύπτει από τα υποσύνολα της μέσης τιμής για το σύνολο δεδομένων με αναγνωριστική ονομασία *selbach_32h_mir30a*. Σύμφωνα με τις συσχετίσεις των 4 συνόλων που αναπαριστά, και οι δύο μέθοδοι κρίνουν την υπερέκφραση και την καταστολή σε 1563 και 1052 γονίδια αντίστοιχα. Ο αριθμός των γονιδίων των οποίων η Average κρίνει την καταστολή τους, ενώ η Majority κρίνει την υπερέκφρασή τους, είναι 7. Για την αντίθετη περίπτωση όπου η Majority κρίνει την καταστολή των γονιδίων, ενώ η Average κρίνει την υπερέκφρασή τους, ο αριθμός των γονιδίων είναι 3.

Είναι φανερό ότι για τα παραπάνω 10 γονίδια υπάρχει μία σύγκρουση μεταξύ των δύο μεθόδων για τη γονιδιακή τους έκφραση. Αυτός ήταν και ο λόγος που επιλέχθηκε να παρουσιαστεί το συγκεκριμένο σύνολο δεδομένων, καθώς είναι το μοναδικό στο οποίο φαίνεται μία τέτοια σύγκρουση. Η σύγκρουση αυτή οφείλεται στην ύπαρξη ακραίων τιμών, οι οποίες επηρεάζουν τον υπολογισμό της τελικής τιμής έκφρασης. Ωστόσο, οι ακραίες τιμές επηρεάζουν κυρίως τον υπολογισμό της τελικής τιμής έκφρασης της Average, καθώς η συγκεκριμένη μέθοδος λαμβάνει υπόψη στον υπολογισμό τις συγκεκριμένες τιμές, σε αντίθεση με τη Majority που τις απορρίπτει όταν έρχονται σε διαφωνία με την απόφαση της πλειοψηφίας.

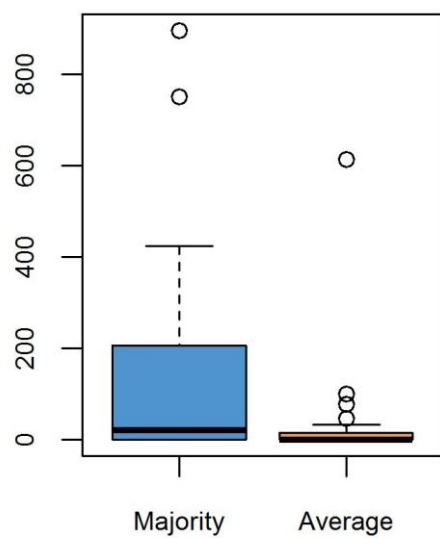
Αυτό που παρατηρείται στο συγκεκριμένο διάγραμμα, καθώς και σε όλα τα υπόλοιπα διαγράμματα των συνόλων δεδομένων, είναι ο μεγαλύτερος αριθμός των επιπλέον γονιδίων των οποίων η Majority κρίνει την υπερέκφραση ή καταστολή τους. Όπως φαίνεται από το διάγραμμα, η Majority κρίνει την υπερέκφραση και καταστολή σε 749 και 374 αντίστοιχα επιπλέον γονίδια. Στην περίπτωση της Average, κρίνεται η υπερέκφραση και η καταστολή σε 424 και 308 αντίστοιχα επιπλέον γονίδια. Ο αριθμός των επιπλέον γονιδίων της κάθε μεθόδου, ήταν ένα από τα βασικά κριτήρια που χρησιμοποιήθηκαν για την περαιτέρω σύγκριση των μεθόδων.

Για τη σύγκριση των μεθόδων με βάση τον αριθμό των επιπλέον γονιδίων που βρίσκουν, υλοποιήθηκε ένας αλγόριθμος ο οποίος για καθεμία από τις συλλογές που έχουν δημιουργηθεί για κάθε σύνολο δεδομένων, υπολογίζει τον συνολικό αριθμό των επιπλέον γονιδίων που προέρχονται από τα υποσύνολα της καταστολής και υπερέκφρασης της κάθε μεθόδου. Στη συνέχεια, με βάση τον συνολικό αριθμό των επιπλέον γονιδίων που βρέθηκαν για κάθε μέθοδο σε όλα τα σύνολα δεδομένων, χρησιμοποιήθηκε ένας ήδη υλοποιημένος αλγόριθμος που αναπαριστά σε δύο θηκογράμματα την κατανομή αυτών των τιμών για κάθε μέθοδο ξεχωριστά. Τα θηκογράμματα που δημιουργήθηκαν φαίνονται στα σχήματα που ακολουθούν.

Τα Σχήματα 7 και 8 απεικονίζουν τα θηκογράμματα της κάθε μεθόδου με βάση τον αριθμό των επιπλέον γονιδίων που προκύπτουν από τη μέση τιμή και τη διάμεσο, αντίστοιχα. Τα θηκογράμματα της Majority δείχνουν ότι η κατανομή των τιμών της παρουσιάζει μία μεγάλη θετική ασυμμετρία, καθώς η διάμεσος βρίσκεται πιο κοντά στην κάτω πλευρά του ορθογωνίου. Τα θηκογράμματα της Average δείχνουν ότι παρουσιάζει μία μικρή εξίσου θετική ασυμμετρία. Το ύψος του ορθογωνίου αντιπροσωπεύει το διάστημα μέσα στο οποίο ανήκει το 50% των τιμών. Στην περίπτωση της μεθόδου Average, το ορθογώνιο αυτό παρουσιάζεται αρκετά “συμπιεσμένο”, ενώ είναι πολύ πιο εκτεταμένο στην περίπτωση της μεθόδου Majority.



Σχήμα 7: Κατανομή επιπλέον γονιδίων από μέση τιμή για κάθε μέθοδο



Σχήμα 8: Κατανομή επιπλέον γονιδίων από διάμεσο για κάθε μέθοδο

Η κατανομή της Average παρουσιάζει και στις δύο περιπτώσεις 3 ακραίες τιμές και μία εξαιρετικά ακραία τιμή. Από την άλλη στην κατανομή της Majority, στη μία περίπτωση παρουσιάζονται μία ακραία τιμή και μία εξαιρετικά ακραία, ενώ στην άλλη περίπτωση παρουσιάζονται 2 εξαιρετικά ακραίες τιμές. Από τη σύγκριση των κατανομών των τιμών της κάθε μεθόδου, καταλήγουμε στο συμπέρασμα ότι η Majority βρίσκει μεγαλύτερο αριθμό επιπλέον γονιδίων σε σύγκριση με την Average. Παρ' όλα αυτά, ήταν απαραίτητο να διερευνηθεί το κατά πόσο η διαφορά των δύο μεθόδων ως προς τον αριθμό των επιπλέον γονιδίων για κάθε σύνολο δεδομένων είναι στατιστικά σημαντική.

Για τον έλεγχο αυτής της σημαντικότητας, υλοποιήθηκε ένας αλγόριθμος που παίρνει ως είσοδο 2 σύνολα, καθένα από τα οποία αποτελείται από το συνολικό αριθμό των επιπλέον γονιδίων που έχουν υπολογιστεί για κάθε σύνολο δεδομένων από τις μεθόδους Majority και Average αντίστοιχα. Ένας ήδη υλοποιημένος αλγόριθμος εφαρμόζει τον μη παραμετρικό έλεγχο της προσημασμένης διάταξης του Wilcoxon με υπόθεση διπλής κατεύθυνσης. Η διαφορά μεταξύ των 2 συνόλων θεωρείται στατιστικά σημαντική εάν η μηδενική υπόθεση μπορεί να απορριφτεί σε διάστημα εμπιστοσύνης μεγαλύτερο του 95% ($p < 0.05$).

Πίνακας 4: Αποτελέσματα στατιστικού ελέγχου

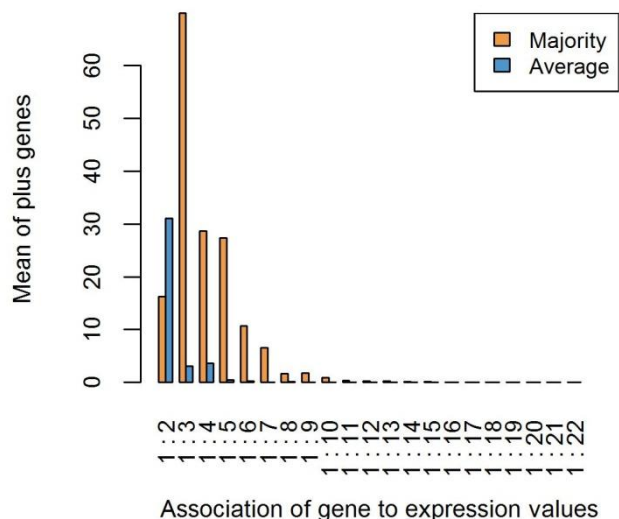
	P-value	Result
Dataset of Mean	2.8857752603724e-05	Reject null hypothesis H0
Dataset of Median	6.40334888718867e-05	Reject null hypothesis H0

Σύμφωνα με τα αποτελέσματα του αλγορίθμου (Πίνακας 4), η μηδενική υπόθεση απορρίφθηκε και για την περίπτωση των επιπλέον γονιδίων που προκύπτουν από τη μέση τιμή, αλλά και για την περίπτωση των επιπλέον γονιδίων που προκύπτουν από τη διάμεσο. Αυτό συνεπάγεται ότι η διαφορά των δύο μεθόδων, όσον αφορά τον αριθμό των επιπλέον γονιδίων, είναι στατιστικά σημαντική. Ωστόσο, είναι απαραίτητο να εφαρμοστεί στα δεδομένα μία περαιτέρω ανάλυση που αφορά τη βιολογική τους λειτουργία, ώστε να μελετηθεί αν τα επιπλέον γονίδια που βρέθηκαν από τη κάθε μέθοδο συσχετίζονται πραγματικά με τα υπό μελέτη miRNAs.

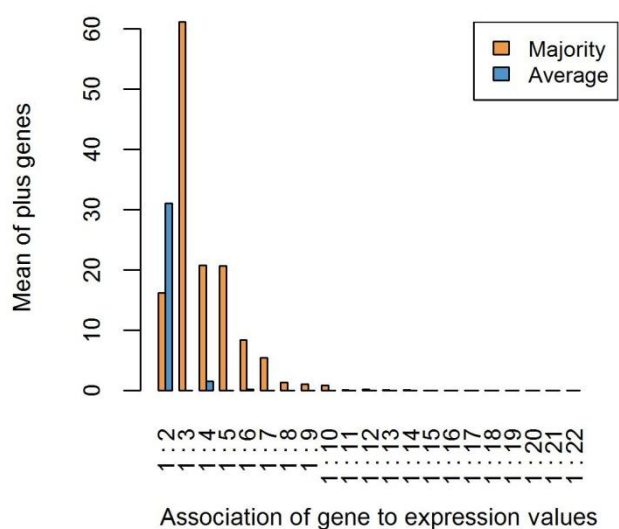
Για τα επιπλέον γονίδια της κάθε μεθόδου, διερευνήθηκαν επίσης οι αντιστοιχίες γονιδίου προς τιμές έκφρασης από τις οποίες προέρχονται. Για το σκοπό αυτό, υλοποιήθηκε ένας αλγόριθμος ο οποίος δημιούργησε για κάθε σύνολο δεδομένων δύο λίστες με τα επιπλέον γονίδια των μεθόδων Majority και Average αντίστοιχα. Πριν την εφαρμογή οποιασδήποτε μεθόδου, για κάθε σύνολο δεδομένων υπολογίστηκε ο αριθμός των τιμών έκφρασης που αντιστοιχεί σε κάθε γονίδιο. Με βάση τις αντιστοιχίες αυτές και τα γονίδια της κάθε λίστας, βρέθηκε σε κάθε σύνολο δεδομένων ο αριθμός των επιπλέον γονιδίων που συσχετίζονται με την κάθε αντιστοιχία.

Για τη γραφική αναπαράσταση αυτών των αντιστοιχιών, υλοποιήθηκε ένας ακόμη αλγόριθμος που υπολογίζει τη μέση τιμή των επιπλέον γονιδίων με τα οποία συσχετίζεται κάθε αντιστοιχία στα σύνολα δεδομένων των μεθόδων.

Στα παρακάτω σχήματα απεικονίζεται η μέση τιμή των επιπλέον γονιδίων που προκύπτουν από τη μέση τιμή και τη διάμεσο για κάθε αντιστοιχία. Η πλειοψηφία των επιπλέον γονιδίων για τη μέθοδο Majority προέρχεται από την 1:3, ενώ για τη μέθοδο Average προέρχεται από την 1:2. Η αντιστοιχία 1:2 είναι η μοναδική περίπτωση όπου η μέθοδος Average συσχετίζει μεγαλύτερο αριθμό επιπλέον γονιδίων με την αντιστοιχία σε σύγκριση με τη Majority.



Σχήμα 9: Μέση τιμή επιπλέον γονιδίων από μέση τιμή για κάθε αντιστοιχία γονιδίου:τιμές έκφρασης



Σχήμα 10: Μέση τιμή επιπλέον γονιδίων από διάμεσο για κάθε αντιστοιχία γονιδίου:τιμές έκφρασης

Το γεγονός αυτό είναι λογικό, καθώς η μέθοδος Average για τον υπολογισμό της τελικής τιμής έκφρασης κάθε γονιδίου λαμβάνει υπόψη και τις δύο τιμές έκφρασης που αντιστοιχούν στο γονίδιο για κάθε περίπτωση της γονιδιακής έκφρασης (όμοια ή διαφορετική). Αντίθετα, η μέθοδος Majority για την αντιστοιχία 1:2, είναι αδύνατο να αποφασίσει με τη χρήση του κανόνα της πλειοψηφίας εάν υπάρχει καταστολή ή υπερέκφραση του γονιδίου στη περίπτωση της διαφορετικής γονιδιακής έκφρασης. Σε αυτή την περίπτωση, η συγκεκριμένη μέθοδος δεν θα συμπεριλάβει καθόλου το γονίδιο αυτό στην περαιτέρω ανάλυση των δεδομένων.

2.2 Συλλογή αλληλεπιδράσεων από Reporter gene assays

Στα πλαίσια αυτής της εργασίας, δόθηκε έμφαση στη συλλογή μεγάλου αριθμού πειραματικά επιβεβαιωμένων miRNA:gene αλληλεπιδράσεων που προέρχονται από την πειραματική μέθοδο των Reporter genes. Αυτή η προτίμηση οφείλεται στο γεγονός ότι η συγκεκριμένη μέθοδος, προσφέρει δύο σημαντικά πλεονεκτήματα έναντι των υπολοίπων. Για μία αλληλεπίδραση που προκύπτει από τη μέθοδο των Reporter genes, επιβεβαιώνεται η *άμεση* αλληλεπίδραση του miRNA και του γονιδίου, καθώς και η ακριβής περιοχή στο μετάγραφο του γονιδίου πάνω στην οποία προσδένεται το miRNA.

Ωστόσο, η αναγνώριση τέτοιων αλληλεπιδράσεων είναι μία δύσκολη και χρονοβόρα διαδικασία. Εκτός από το μεγάλο πλήθος των επιστημονικών δημοσιεύσεων, πολλές από τις αλληλεπιδράσεις είναι κρυμμένες σε εικόνες, πίνακες, ή συμπληρωματικό υλικό των επιστημονικών δημοσιεύσεων. Ακόμα, η μέθοδος των Reporter genes προϋποθέτει τη χρήση επιμέρους υπολογιστικών βημάτων για την ανάκτηση της σχετικής πληροφορίας. Η διαδικασία της καταγραφής των αλληλεπιδράσεων, καθώς και κάποια στατιστικά των δεδομένων που συλλέχθηκαν, περιγράφονται παρακάτω.

2.2.1 Διαδικασία καταγραφής

Για την διευκόλυνση της χειροκίνητης αναγνώρισης των miRNA:gene αλληλεπιδράσεων, χρησιμοποιήθηκαν αλγόριθμοι εξόρυξης κειμένου. Σκοπός αυτών των αλγορίθμων είναι να φιλτράρουν τις διαθέσιμες επιστημονικές δημοσιεύσεις ώστε να δημιουργήσουν ένα σύνολο που αποτελείται μόνο από δημοσιεύσεις που περιέχουν miRNA:gene αλληλεπιδράσεις. Σε αυτό το σύνολο των επιστημονικών δημοσιεύσεων έγινε μετέπειτα η αναζήτηση για την χειροκίνητη αναγνώριση των σχετικών αλληλεπιδράσεων.

Δεδομένου ότι η συλλογή των αλληλεπιδράσεων θα χρησιμοποιηθεί για την αναβάθμιση της βάσης δεδομένων του DIANA-TarBase, ήταν απαραίτητο τα δεδομένα να έχουν μία συγκεκριμένη μορφή που ταυτίζεται με το σχήμα της βάσης. Κάθε αλληλεπίδραση αποτελείται από ένα καθορισμένο αριθμό πεδίων (Εικόνα 6) στα οποία πρέπει να καταχωρηθούν πληροφορίες που αφορούν το γονίδιο, το miRNA, το πείραμα, το σημείο πρόσδεσης, καθώς και την επιστημονική δημοσίευση.

Ανάλυση χαμηλής και υψηλής απόδοσης πειραματικών μεθόδων για συλλογή επιβεβαιωμένων γονιδίων στόχων των miRNAs

[illegible]

Εικόνα 6: Πεδία κάθε αλληλεπίδρασης

Πιο λεπτομερειακά, για το γονίδιο της αλληλεπίδρασης καταγράφονται τα Ensembl Gene και Transcript IDs που αποτελούν τα μοναδικά αναγνωριστικά του όπως αυτά ορίζονται από τη βάση δεδομένων ensembl (<http://www.ensembl.org>), καθώς και το σύμβολο HGNC που συμβολίζει την επίσημη ονομασία του. Αντίστοιχα, για το miRNA καταγράφονται το όνομά του όπως αυτό είναι καταχωρημένο στη βάση δεδομένων miRBase (<http://www.mirbase.org>) και ο οργανισμός στον οποίο αντιστοιχεί.

Για την πειραματική μέθοδο καταγράφονται η ακριβή ονομασία της, το γονίδιο αναφοράς και τα κύτταρα που χρησιμοποιήθηκαν στα πλαίσια του πειράματος, τα κύτταρα και/ή ο ιστός προέλευσης του δείγματος, καθώς και η κατηγοριοποίησή τους σε φυσιολογικά, καρκινικά, κ.ά. Ακόμα, καταγράφονται το αποτέλεσμα του πειράματος (αρνητικό ή θετικό), το είδος της αλληλεπίδρασης (άμεση ή έμμεση), η αναρρύθμιση ή απορύθμιση του γονιδίου, και η περιοχή πρόσδεσης του miRNA στο γονίδιο (π.χ. 3'UTR ή CDS).

Πέρα από την καταγραφή των miRNA:gene αλληλεπιδράσεων που προκύπτουν από τα θετικά αποτελέσματα των πειραμάτων, ήταν εξίσου πολύ σημαντικό να καταγραφούν και οι αλληλεπιδράσεις που προκύπτουν από τα αρνητικά αποτελέσματα. Μια τέτοια πληροφορία είναι πολύ χρήσιμη για τη δημιουργία ενός θετικού και ενός αρνητικού συνόλου δεδομένων αντίστοιχα, με το οποίο θα γίνει η εκπαίδευση και αξιολόγηση των αλγορίθμων που προβλέπουν στόχους των miRNAs.

Όσον αφορά τη καταγραφή του σημείου πρόσδεσης, απαιτούνται κάποια επιπλέον υπολογιστικά βήματα που περιλαμβάνουν αρχικά την εύρεση της ακολουθίας. Η ακολουθία του σημείου πρόσδεσης μπορεί να εντοπιστεί σε επιμέρους υλικό της επιστημονικής δημοσίευσης, είτε να προκύψει από τις ακολουθίες των primers με τη χρήση του εργαλείου In-Silico PCR (<https://genome.ucsc.edu>) στις περιπτώσεις των πειραμάτων κλωνοποίησης, ή μέσω της αντιστοίχισης των γενωμικών συντεταγμένων του μεταγράφου στην περιοχή του μεταγράφου.

Αφού βρεθεί η ακολουθία, εντοπίζεται η θέση της στο γονιδίωμα και αντιστοιχίζεται στην τελευταία έκδοση του Genome Assembly με τη χρήση του εργαλείου BLAT (<https://genome.ucsc.edu>). Για το σημείο πρόσδεσης,

καταγράφονται οι συντεταγμένες του σημείου και η ακολουθία του, καθώς και οι συντεταγμένες και ακολουθίες των primers, εφόσον δίνονται. Όλες οι συντεταγμένες που καταγράφονται αντιστοιχίζονται πάντα στην τελευταία έκδοση του Genome Assembly μέσω του BLAT.

Επιπλέον, καταγράφονται πληροφορίες που αφορούν την επιστημονική δημοσίευση όπως οι συγγραφείς και ο τίτλος της δημοσίευσης, το περιοδικό, η χρονολογία δημοσίευσης, το αναγνωριστικό PMID, και το email επικοινωνίας με τους συγγραφείς. Άλλες σχετικές πληροφορίες που καταγράφονται είναι το σημείο της δημοσίευσης που βρέθηκε η αλληλεπίδραση, τα προβλήματα που μπορεί να προέκυψαν κατά την αναγνώριση της αλληλεπίδρασης, καθώς και γενικά σχόλια σχετικά με τη συγκεκριμένη διαδικασία.

2.2.2 Στατιστικά δεδομένων

Η συλλογή των επιβεβαιωμένων αλληλεπιδράσεων μεταξύ miRNAs και γονιδίων για πειράματα που αφορούν την μέθοδο των Reporter genes πραγματοποιήθηκε χειροκίνητα από ένα σύνολο δεδομένων που περιλάμβανε 204 επιστημονικές δημοσιεύσεις. Κάποιες από τις δημοσιεύσεις δεν ήταν διαθέσιμες στο διαδίκτυο και κατά συνέπεια δεν υπήρχε δυνατότητα να συλλεχθούν τα σχετικά δεδομένα από τις συγκεκριμένες. Στους πίνακες που ακολουθούν φαίνεται μία σύνοψη κάποιων στατιστικών για τα δεδομένα που συλλέχθηκαν, καθώς και τα είδη των οργανισμών που περιλαμβάνουν.

Πίνακας 5: Στατιστικά δεδομένων

	Entries
Total Publications	204
Total Entries	1,586
Positive Interactions	1,335
Negative Interactions	251
Targeted Sequences	1,308
Positive Targeted Sequences	1,124
Negative Targeted Sequences	184

Πίνακας 6: Εγγραφές ανά οργανισμό

Species	Entries
<i>Homo sapiens</i>	1,465
<i>Mus musculus</i>	109
<i>Gallus gallus</i>	10
<i>Sus scrofa</i>	2

Το σύνολο των αλληλεπιδράσεων που συλλέχθηκαν ανέρχεται στις 1,586. Από τις συνολικές εγγραφές, οι 1,335 προέρχονται από θετικά αποτελέσματα του πειράματος των Reporter genes, ενώ οι 251 προέρχονται από αρνητικά αποτελέσματα. Για τις 1,308 εγγραφές βρέθηκε η ακριβής ακολουθία του mRNA στόχου, από τις οποίες οι 1,124 είναι από θετικά αποτελέσματα ενώ οι 184 από αρνητικά. Οι 1,586 εγγραφές αναφέρονται σε 4 διαφορετικούς οργανισμούς, εκ των οποίων οι 1,465 είναι στον *Homo sapiens*, οι 109 στο *Mus musculus*, οι 10 στο *Gallus gallus* και οι 2 στο *Sus scrofa*.

3. ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα miRNAs είναι μία κλάση μικρών μορίων RNA, τα οποία έχουν ως βασική τους λειτουργία τη γονιδιακή ρύθμιση με την πρόσδεσή τους στα μετάγραφα των γονιδίων που αποτελούν στόχους τους. Το μεγάλο ενδιαφέρον για την βαθύτερη κατανόηση της λειτουργίας τους, έχει δημιουργήσει μία έντονη δραστηριότητα στο συγκεκριμένο ερευνητικό πεδίο. Σήμερα, έχει ανακαλυφθεί ένα μεγάλο πλήθος διαφορετικών miRNAs, τα οποία έχουν συσχετιστεί με πολλές σημαντικές ανθρώπινες ασθένειες, όπως ο καρκίνος. Τα miRNAs φαίνεται να παίζουν έναν πολύ σημαντικό ρόλο σε αυτές τις ασθένειες.

Μία από τις πιο σημαντικές διαδικασίες στη μελέτη τους, είναι η αναγνώριση των αλληλεπιδράσεων που έχουν με τα γονίδια στόχους τους. Η αναγνώριση τέτοιων αλληλεπιδράσεων μπορεί να επιβεβαιωθεί με την εφαρμογή πειραματικών μεθόδων. Η χρήση της μεθόδου των μικροσυστοιχιών είναι ευρέως διαδεδομένη για την εξέταση της γονιδιακής έκφρασης μετά την υπερέκφραση ή καταστολή ενός υπό μελέτη miRNA. Με μία λεπτομερειακή επεξεργασία τέτοιων δεδομένων μπορούν να προκύψουν αλληλεπιδράσεις μεταξύ miRNAs και γονιδίων στόχων τους.

Στα δεδομένα που προέρχονται από το πείραμα των μικροσυστοιχιών, ένα γονίδιο μπορεί να αντιστοιχίζεται με πολλές τιμές έκφρασης. Για την ανάλυση όμως της γονιδιακής έκφρασης, κάθε γονίδιο πρέπει να συσχετιστεί με μία τελική τιμή έκφρασης. Στη πιο συνηθισμένη μέθοδο που χρησιμοποιείται, επιλέγεται η μέση τιμή ή διάμεσος όλων των τιμών έκφρασης του γονιδίου. Ωστόσο, στα πλαίσια αυτής της εργασίας διερευνήθηκε μία διαφορετική προσέγγιση για τον υπολογισμό της τελικής τιμής έκφρασης, η οποία βασίζεται στον κανόνα της πλειοψηφίας.

Βασικό ζητούμενο της παρούσας εργασίας αποτέλεσε η συλλογή υψηλής ποιότητας δεδομένων που σχετίζονται με miRNA:gene αλληλεπιδράσεις, οι οποίες έχουν επιβεβαιωθεί από υψηλής ή χαμηλής απόδοσης πειραματικές μεθόδους. Λαμβάνοντας υπόψη ότι από την επεξεργασία των δεδομένων που προέρχονται από τις μικροσυστοιχίες μπορούν να αναγνωριστούν τέτοιες αλληλεπιδράσεις, πραγματοποιήθηκε μία σύγκριση των δύο παραπάνω προσεγγίσεων για την ανάδειξη της πιο αξιόπιστης μεθόδου επεξεργασίας τέτοιων δεδομένων.

Για τη σύγκριση αυτή συλλέχθηκαν δεδομένα miRNA μικροσυστοιχιών από ευρέως αναγνωρισμένες επιστημονικές δημοσιεύσεις. Τα δεδομένα αυτά αποτελούνται από 29 σύνολα δεδομένων και προέρχονται από δύο γενωμικά chips της Affymetrix με διαφορετικό σχεδιασμό το καθένα. Σε καθένα από αυτά τα σύνολα εφαρμόστηκε η μέθοδος Average και η μέθοδος Majority για τον υπολογισμό της τελικής τιμής έκφρασης. Μετά την εφαρμογή των δύο μεθόδων ακολούθησε μία σύγκριση των συνόλων που προέκυψαν από τη κάθε μέθοδο.

Κατά τη διαδικασία της σύγκρισης χρησιμοποιήθηκαν τα Venn διαγράμματα, καθώς και γραφικές αναπαραστάσεις σχετικά με την κατανομή των τελικών τιμών έκφρασης που προκύπτουν από την εφαρμογή της κάθε μεθόδου, τον αριθμό των επιπλέον γονιδίων που βρίσκει η κάθε μέθοδος, τον έλεγχο της στατιστικής σημαντικότητας των επιπλέον γονιδίων, και τις αντιστοιχίες γονιδίου προς τιμές έκφρασης από τις οποίες προέρχονται αυτά τα επιπλέον

γονίδια. Από την σύγκριση όλων των παραπάνω, η μέθοδος Majority δείχνει μία καλύτερη απόδοση από την μέθοδο Average.

Με βάση την κατανομή των τελικών τιμών έκφρασης, η πλειοψηφία των περιπτώσεων παρουσιάζει κάποιες μικρές αλλαγές μεταξύ των δύο μεθόδων. Αυτό που παρατηρήθηκε όμως από τα διαγράμματα Venn είναι ότι σε κάθε περίπτωση η μέθοδος Majority βρίσκει πολύ μεγαλύτερο αριθμό επιπλέον γονιδίων σε σύγκριση με τη μέθοδο Average. Από τον στατιστικό έλεγχο αυτών των επιπλέον γονιδίων προέκυψε ότι η διαφορά μεταξύ των δύο μεθόδων όσον αφορά τον αριθμό των επιπλέον γονιδίων που βρίσκουν σε κάθε σύνολο δεδομένων είναι στατιστικά σημαντική.

Πέρα από τις miRNA:gene αλληλεπιδράσεις που αναγνωρίστηκαν από την επεξεργασία δεδομένων των μικροσυστοιχιών, έγινε μία προσπάθεια να συλλεχθούν miRNA:gene αλληλεπιδράσεις οι οποίες έχουν επιβεβαιωθεί από την πειραματική μέθοδο των Reporter genes. Το γεγονός ότι η χαμηλής απόδοσης μέθοδος των Reporter genes έχει το πλεονέκτημα ότι δύναται να αναγνωριστεί η ακριβής περιοχή στο μετάγραφο του γονιδίου πάνω στην οποία προσδένεται το miRNA, αποτέλεσε το λόγο που δόθηκε ιδιαίτερη σημασία στη συλλογή τέτοιων αλληλεπιδράσεων.

Οι αλληλεπιδράσεις αυτές αναγνωρίστηκαν χειροκίνητα από ένα πλήθος 204 επιστημονικών δημοσιεύσεων. Για κάθε αλληλεπίδραση ήταν απαραίτητο να συλλεχθούν κάποιες σημαντικές πληροφορίες σχετικά με την αλληλεπίδραση. Η καταγραφή των αλληλεπιδράσεων που συλλέχθηκαν έχουν συγκεκριμένη μορφή και αποτελούνται από ένα καθορισμένο αριθμό πεδίων με σχετικές πληροφορίες. Αυτό οφείλεται στο γεγονός ότι τα δεδομένα αυτά θα χρησιμοποιηθούν για την αναβάθμιση της βάσης δεδομένων του DIANA-TarBase, οπότε η μορφή τους πρέπει να ταυτίζεται με το σχήμα της βάσης.

Συνολικά συλλέχθηκαν 1,586 αλληλεπιδράσεις, οι οποίες αναφέρονται σε 4 διαφορετικούς οργανισμούς που περιλαμβάνουν τους *Homo sapiens*, *Mus musculus*, *Gallus gallus* και *Sus scrofa*. Από τις συνολικές αλληλεπιδράσεις, οι 1,335 προέρχονται από θετικά αποτελέσματα του πειράματος των Reporter genes, ενώ οι 251 προέρχονται από αρνητικά αποτελέσματα. Η ακριβής ακολουθία του mRNA στόχου βρέθηκε για τις 1,308 αλληλεπιδράσεις, από τις οποίες οι 1,124 προκύπτουν από θετικά αποτελέσματα ενώ οι 184 από αρνητικά.

Πολλές από τις miRNA:gene αλληλεπιδράσεις είναι κρυμμένες σε επιμέρους υλικό των επιστημονικών δημοσιεύσεων, καθιστώντας τη συλλογή τους μία χρονοβόρα και δύσκολη διαδικασία. Παρ' όλα αυτά, η συλλογή τέτοιων αλληλεπιδράσεων είναι υψίστης σημασίας, καθώς πέρα από την ενημέρωση μίας σχετικής βάσης δεδομένων μπορεί να αξιοποιηθεί σε πολλές άλλες εφαρμογές. Η χρήση τέτοιων δεδομένων στην εκπαίδευση αλγορίθμων-μηχανικής μάθησης που προβλέπουν στόχους των miRNAs, μπορεί να συμβάλει στην ανάπτυξη μίας νέας γενιάς αλγορίθμων με πολύ μεγαλύτερη ακρίβεια και ευαισθησία που να αποτελούν αναπόσπαστο κομμάτι της διερευνητικής διαδικασίας και τα οφέλη τους να έχουν άμεση επίδραση στην ιατροβιολογική έρευνά, αλλά και στην κλινική πρακτική.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός Όρος
mRNA	Μετάγγραφο
Splicing	Μάτισμα
Hairpin loop	Φουρκέτα
Promoter	Υποκινητής
Single-arm stem loop	Βραχίονας
Seed region	Περιοχή εκβλάστησης
Coding Sequence	Κωδική περιοχή
Dominant binding	Κυρίαρχο ταίριασμα
Compensatory binding	Ικανοποιητικό ταίριασμα
Elongation	Επιμήκυνση
Ribosome drop-off	Αποκόλληση ριβοσώματος
Fragile	Εύθραυστο
Low-yield method	Μέθοδος χαμηλής απόδοσης
High-throughput method	Μέθοδος υψηλής απόδοσης
Microarray	Μικροσυστοιχία
Reporter genes method	Μέθοδος γονιδίων αναφοράς
Primer sequence	Αναγνωριστική ακολουθία
Gene	Γονίδιο
Probes	Ανιχνευτές
In situ hybridization	Επιτόπια υβριδοποίηση
Kernel Density Estimation	Εκτίμηση πυκνότητας με τη χρήση πυρήνων
Majority Rule	Κανόνας της πλειοψηφίας
Median polish	Κανονικοποίηση της διαμέσου
Dataset	Σύνολο δεδομένων
Replication method	Μέθοδος αντιγραφής
Box plot	Θηκόγραμμα
Wilcoxon signed-rank test	Έλεγχος προσημασμένης διάταξης
Two-tailed	Διπλής κατεύθυνσης
Genome Assembly	Έκδοση χαρτογράφησης γονιδιώματος
Homo Sapiens	Άνθρωπος
Mus Musculus	Ποντίκι
Gallus gallus	Κοτόπουλο
Sus scrofa	Γουρούνι

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

RNA	Ribonucleic acid
miRNA	micro RNA
mRNA	messenger RNA
3'UTR	3' UnTranslated Region
CDS	Coding Sequence
pri-miRNA	primary miRNA
pre-miRNA	precursor miRNA
RISC	RNA-induced silencing complex
Poly-A	Polyadenylation
P-Bodies	Processing bodies
SFX	Syndrome Fragile X
NGS	Next-Generation Sequencing
qPCR	Quantitative Polymerase Chain Reaction
ELISA	Enzyme-Linked Immunosorbent Assay
pSILAC	Pulsed stable isotope labeling by amino acids in cell culture
HITS-CLIP	High-Throughput Sequencing of RNA isolated by Crosslinking Immunoprecipitation
PAR-CLIP	Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation
CLASH	Cross-linking, Ligation and Sequencing of Hybrids
CLIP-Seq	Cross-linking immunoprecipitation sequencing
DNA	Deoxyribonucleic Acid
RBP	RNA-binding protein
AGO	Argonaute
MAS	MicroArray Suite
RMA	Robust Multi-Array Average

ΑΝΑΦΟΡΕΣ

1. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14* Cell, 1993. **75**(5): p. 843-54.
2. Lagos-Quintana, M., et al., *Identification of Novel Genes Coding for Small Expressed RNAs*. Science, 2001. **294**(5543): p. 853-8.
3. Lau, N.C., et al., *An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans*. Science, 2001. **294**(5543): p. 858-62.
4. Lee, R.C. and V. Ambros, *An Extensive Class of Small RNAs in Caenorhabditis elegans*. Science, 2001. **294**(5543): p. 862-4.
5. Calin, G.A., et al., *Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia*. Proc Natl Acad Sci U S A, 2002. **99**(24): p. 15524-9.
6. Rodriguez, A., et al., *Identification of mammalian microRNA host genes and transcription units*. Genome Res, 2004. **14**(10a): p. 1902-10.
7. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II*. EMBO Journal, 2004. **23**(20): p. 4051-60.
8. Rottiers, V. and A.M. Naar, *MicroRNAs in metabolism and metabolic disorders*. Nat Rev Mol Cell Biol, 2012. **13**(4): p. 239-50.
9. Lee, Y., et al., *The nuclear RNase III Drosha initiates microRNA processing*. Nature, 2003. **425**(6956): p. 415-9.
10. Gregory, R.I., T.P. Chendrimada, and R. Shiekhattar, *MicroRNA biogenesis: isolation and characterization of the microprocessor complex*. Methods in Molecular Biology, 2006. **342**: p. 33-47.
11. Lund, E. and J.E. Dahlberg, *Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs*. Cold Spring Harbor Symposia on Quantitative Biology, 2006. **71**: p. 59-66.
12. Brennecke, J., et al., *Principles of microRNA-target recognition*. PLoS Biol, 2005. **3**(3): p. e85.
13. Bartel, D.P., *MicroRNAs: Genomics, Biogenesis, Mechanism, and Function*. Cell, 2004. **116**(2): p. 281-97.
14. Kawasaki, H. and K. Taira, *MicroRNA-196 inhibits HOXB8 expression in myeloid differentiation of HL60 cells*. Nucleic Acids Symposium Series, 2004. **48**: p. 211-2.
15. Moxon, S., et al., *Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening*. Genome Res, 2008. **18**(10): p. 1602-9.
16. Morozova, N., et al., *Kinetic signatures of microRNA modes of action*. RNA, 2012. **18**(9): p. 1635-55.
17. Ardekani, A.M. and M.M. Naeini, *The Role of MicroRNAs in Human Diseases*. Avicenna Journal of Medical Biotechnology, 2010. **2**(4): p. 161-79.
18. Soifer, H.S., J.J. Rossi, and P. Saetrom, *MicroRNAs in disease and potential therapeutic applications*. Mol Ther, 2007. **15**(12): p. 2070-9.
19. Szymanski, M., et al., *A new frontier for molecular medicine: noncoding RNAs*. Biochim Biophys Acta, 2005. **1756**(1): p. 65-75.

20. McManus, M.T., *MicroRNAs and cancer*. Seminars in cancer biology, 2003. **13**(4): p. 253-8.
21. Naeini, M.M. and A.M. Ardekani, *Noncoding RNAs and Cancer*. Avicenna Journal of Medical Biotechnology, 2009. **1**(2): p. 55-70.
22. Zhang, B., et al., *microRNAs as oncogenes and tumor suppressors*. Dev Biol, 2007. **302**(1): p. 1-12.
23. Vlachos, I.S., et al., *DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions*. Nucleic Acids Res, 2015. **43**(Database issue): p. D153-9.
24. Chou, C.H., et al., *miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database*. Nucleic Acids Res, 2016. **44**(D1): p. D239-47.
25. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions*. Nucleic Acids Res, 2009. **37**(Database issue): p. D105-10.
26. Li, J.H., et al., *starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data*. Nucleic Acids Res, 2014. **42**(Database issue): p. D92-7.
27. Khorshid, M., C. Rodak, and M. Zavolan, *CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins*. Nucleic Acids Res, 2011. **39**(Database issue): p. D245-52.
28. Trevino, V., F. Falciani, and H.A. Barrera-Saldaña, *DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research*. Mol Med., 2007. **13**(9-10): p. 527-41.
29. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data*. Nucleic Acids Res., 2003. **31**(4): p. e15.
30. Hubbell, E., W.M. Liu, and R. Mei, *Robust estimators for expression analysis*. Bioinformatics, 2002. **18**(12): p. 1585-92.
31. Wu, Z., et al., *A Model Based Background Adjustment for Oligonucleotide Expression Arrays*. Journal of the American Statistical Association, 2004. **99**(468): p. 909-17.
32. Bossel Ben-Moshe, N., et al., *Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets*. Nucleic Acids Research, 2012. **40**(21): p. 10614-27.
33. Rao, X., et al., *MicroRNA-221/222 confers breast cancer fulvestrant resistance by regulating multiple signaling pathways*. Oncogene, 2011. **30**(9): p. 1082-97.
34. Selbach, M., et al., *Widespread changes in protein synthesis induced by microRNAs*. Nature, 2008. **455**(7209): p. 58-63.
35. Hausser, J., et al., *Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets*. Genome Res, 2009. **19**(11): p. 2009-20.
36. Frankel, L.B., et al., *microRNA-101 is a potent inhibitor of autophagy*. EMBO Journal, 2011. **30**(22): p. 4628-41.
37. Frankel, L.B., et al., *A non-conserved miRNA regulates lysosomal function and impacts on a human lysosomal storage disorder*. Nat Commun, 2014. **5**: p. 5840.
38. Hollander, J.A., et al., *Striatal microRNA controls cocaine intake through CREB signaling*. Nature, 2010. **466**(7303): p. 197-202.

39. Hu, H.Y., et al., *MicroRNA Expression and Regulation in Human, Chimpanzee, and Macaque Brains*. PLoS Genet, 2011. **7**(10): p. e1002327.
40. Schneider, C., et al., *MicroRNA 28 controls cell proliferation and is down-regulated in B-cell lymphomas*. Proc Natl Acad Sci U S A, 2014. **111**(22): p. 8185-90.