ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Σύγκριση αλγορίθμων στοίχισης για την ανάλυση δεδομένων μικρών RNAs

Νιζάμης Ευάγγελος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνη
Χατζηγεωργίου Άρτεμις
Καθηγήτρια

Λαμία, 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Σύγκριση αλγορίθμων στοίχισης για την ανάλυση δεδομένων μικρών RNAs

Νιζάμης Ευάγγελος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπουσα
Χατζηγεωργίου Άρτεμις
Καθηγήτρια

Λαμία, 2022

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Σύγκριση αλγορίθμων στοίχισης για την ανάλυση δεδομένων μικρών RNAs

Νιζάμης Ευάγγελος

Τριμελής Επιτροπή:

Χατζηγεωργίου Άρτεμις, Καθηγήτρια

Μπάγκος Παντελής, Καθηγητής

Μπράλιου Γεωργία, Επίκουρος Καθηγήτρια

ΠΡΟΛΟΓΟΣ

Σε αυτό το σημείο οφείλω ένα μεγάλο ευχαριστώ στην επιβλέπουσα καθηγήτριά μου κ. Χατζηγεωργίου Άρτεμις, Καθηγήτρια του τμήματος Πληροφορικής με εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, για την εμπιστοσύνη, τη συμπαράσταση και την καθοδήγησή της καθ' όλη τη διάρκεια της πτυχιακής μου εργασίας.

Θερμές ευχαριστίες θα ήθελα να εκφράσω και στα υπόλοιπα μέλη της τριμελούς επιτροπής, τον κ. Μπάγκο Παντελή και την κα Μπράλιου Γεωργία για τις επισημάνσεις και τις υποδείξεις τους.

Ιδιαίτερα ευχαριστώ τον υποψήφιο διδάκτορα Θάνο Αλεξίου για την πολύτιμη βοήθεια και καθοδήγηση του κατά την διάρκεια της εκπόνησης της πτυχιακής μου εργασίας.

Τέλος, θα ήταν παράλειψή μου να μην ευχαριστήσω από τα βάθη της καρδιάς μου τον οικογενειακό και φιλικό μου περίγυρο που με βοήθησε να ολοκληρώσω αυτό το ταξίδι.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια έχουν χαρακτηριστεί τα μικρά μη-κωδικά μόρια RNA (microRNAs-miRNAs) ως εξαιρετικοί βιοδείκτες, καθώς έχει δειχθεί πως συμμετέχουν στη ανάπτυξη και εξέλιξη πληθώρα παθήσεων συμπεριλαμβανομένου και διάφορων καρκινικών τύπων, ενώ μπορούν να χρησιμοποιηθούν ως δείκτες ανταπόκρισης στη θεραπεία. Για το λόγο αυτό τα miRNAs, αποτελούν έναν πολύ ελκυστικό στόχο στην έρευνα και ανάπτυξη νέων θεραπειών. Η ανάπτυξη της βιοπληροφορικής έδωσε στους ερευνητές μια πληθώρα εργαλείων τα οποία μπορούν να χρησιμοποιήσουν για την ανάλυση ήδη γνωστών miRNAs, καθώς επίσης και για την πρόβλεψη νέων τέτοιων ακολουθιών με σημαντική λειτουργία. Επίσης, δίνεται η δυνατότητα να ερευνηθούν οι αλληλεπιδράσεις των miRNAs και των διαφόρων στόχων τους, με αποτέλεσμα να βρεθεί ή/και προβλεφθεί η λειτουργικότητά τους.

Σημαντικό βήμα στην βιοπληροφορική ανάλυση των miRNAs και των στόχων τους, αποτελεί η στοίχιση των ακολουθιών που προκύπτουν από δεδομένα αλληλούχισης επόμενης γενιάς στο γονιδίωμα αναφοράς. Δεδομένης της πληθώρας αλγορίθμων στοίχισης, στόχος της παρούσας πτυχιακής εργασίας είναι η σύγκριση τέτοιων αλγορίθμων στοίχισης, ως προς την αποτελεσματικότητα τους στην διαδικασία ποσοτικοποίησης των miRNAs.

Για το σκοπό αυτό, χρησιμοποιήθηκαν τέσσερις αλγόριθμοι στοίχισης, οι Bowtie, Bowtie 2, HISAT2 και STAR, καθώς και συνθετικά simulated δεδομένα, για τα οποία γνωρίζουμε από πριν την πραγματική ποσοτικοποίηση των δεδομένων που περιέχουν, καθώς ακόμη και από ποια περιοχή του γονιδιώματος προέρχεται κάθε ακολουθία (ground truth).

Βρέθηκε πως το STAR έχει την καλύτερη απόδοση στην στοίχιση, τόσο των mature, όσο και των hairpin ακολουθιών. Παρόλ΄αυτά ο χρόνος που απαιτείται για την ολοκλήρωση της στοίχισης, είναι σημαντικά περισσότερος από τον χρόνο που χρειάζονται οι υπόλοιποι aligners. Επιπλέον, παρατηρήθηκε πως το HISAT2 επιτυγχάνει καλύτερο ποσοστό στοίχισης (μετά το STAR), συγκριτικά με το Bowtie και το Bowtie 2, όσον αφορά τα mature miRNAs, που είναι και μικρότερα σε μήκος αλληλουχίας. Ωστόσο, ο χρόνος που χρειάζεται για την ολοκλήρωση της στοίχισης είναι σχεδόν τριπλάσιος. Τέλος, στην περίπτωση των hairpin miRNAs, παρατηρήθηκε

πως το καλύτερο πρόγραμμα για την στοίχισή τους, μετά το STAR, είναι το Bowtie 2, καθώς επιτυγχάνει το καλύτερο ποσοστό στοίχισης.

Τελικά, το βέλτιστο πρόγραμμα στοίχισης, τόσο για τα hairpin, όσο και για τα mature miRNAs είναι το STAR, το οποίο ως μειονέκτημα εμφανίζει τον σημαντικά υψηλότερο χρόνο που χρειάζεται για την ολοκλήρωση της διαδικασίας.

ABSTRACT

Over the last years, small non-coding RNAs(microRNAs-miRNAs) have been characterized as excellent biomarkers, as it is shown to participate in the development and progression of many diseases, including various types of cancer, while they can be used as indicators of treatment response. That's why miRNAs are a very attractive target in the research and development of new therapies. Nowadays, researchers have a handful of tools that they can use to analyze already known miRNAs, as well as to predict new miRNA sequences with their function. Also, it is possible to investigate the interactions of miRNAs and their various targets, resulting in finding and/or predicting their gene functionality and regulation.

An important step in the bioinformatic analysis of miRNAs is the alignment of the reads, produced from next-generation sequencing data to the reference genome. Given the multitude of alignment algorithms, the aim of this thesis is to compare alignment algorithms, in terms of their effectiveness in the process of quantifying the expression of miRNAs.

For this purpose, four alignment algorithms, Bowtie, Bowtie 2, HiSat 2 and STAR, were used, as well as simulated data. For these data we already know the actual quantification of the data they contain, as well as the exact position of each read on the genome (ground truth).

STAR was found to be the best aligner for both mature and hairpin reads, in terms of alignment rates. However, it is highly time-consuming, as it takes much longer than other aligners to complete the alignment. In addition, it was observed that HISAT2 achieves a better alignment rate (after STAR), compared to Bowtie and Bowtie 2, for mature miRNAs, which are shorter in sequence length. However, the time it takes to complete the alignment is almost 3-times more than the other two aligners. Finally, in the case of hairpin miRNAs, it was observed that the best aligner, after STAR, is Bowtie 2, as it achieves the best alignment rate.

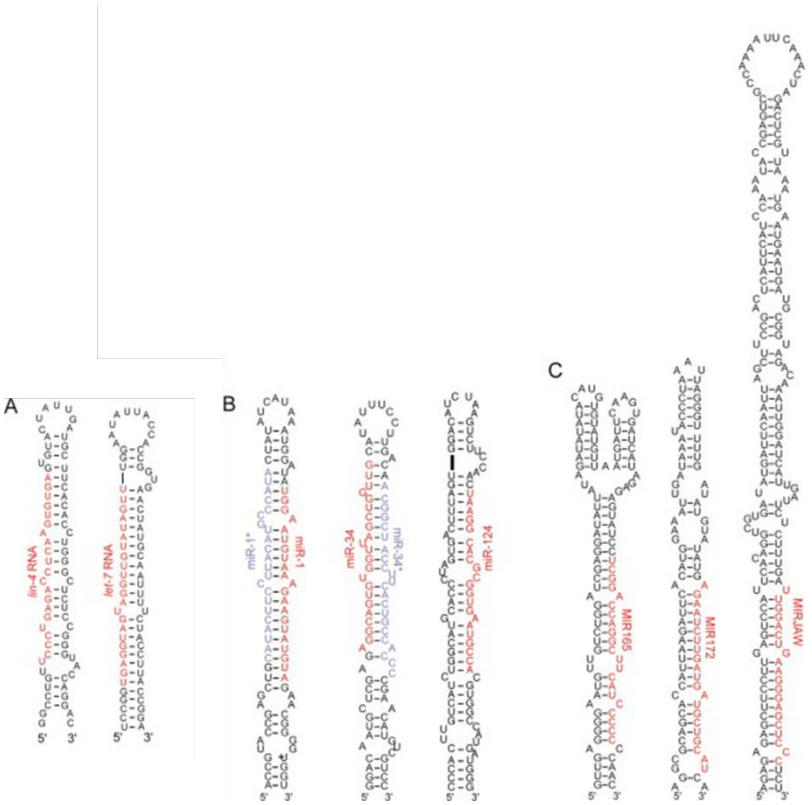
Finally, the optimal alignment program, both for hairpins and for mature miRNAs, is STAR, although it has the disadvantage of significantly higher time needed to complete the process.

ΕΙΣΑΓΩΓΗ

1. Μικρά RNA – microRNA (miRNA)

Τα μικρά RNA μόρια (microRNA – miRNAs) είναι μια κατηγορία μη-κωδικών RNA (17-24 nt) που παίζουν σημαντικό ρόλο στη ρύθμιση της γονιδιακής έκφρασης. Η πλειοψηφία των miRNAs μεταγράφεται από DNA αλληλουχίες σε πρώιμα miRNAs (pri-miRNAs), τα οποία επεξεργάζονται σε πρόδρομα miRNAs (pre-miRNAs) και τελικά σε ώριμα miRNAs. Στις περισσότερες περιπτώσεις, τα ώριμα miRNA αλληλεπιδρούν με την 3' αμετάφραστη περιοχή (3' UTR) των mRNA-στόχων μέσω της συμπληρωματικότητας των βάσεων, με αποτέλεσμα να προκαλούν αποικοδόμηση του mRNA και μεταφραστική αποσιώπηση[1]. Αυτό μπορεί να επιτευχθεί με μία ή περισσότερες από τις ακόλουθες διαδικασίες[2,3]:

- θραύση του mRNA κλώνου σε δύο κομμάτια
- αποσταθεροποίηση του mRNA μέσω βράχυνσης της ουράς poly(A) ουράς του
- μειωμένη αποτελεσματικότητα στη μετάφραση του mRNA σε πρωτεΐνες από τα ριβοσώματα.



Εικόνα 1. Σχηματική αναπαράσταση των προβλεπόμενων βρόγχων (loops) που περιλαμβάνουν τα ώριμα miRNA (κόκκινο) και την πλευρική αλληλουχία. Με μπλε εμφανίζονται τα miRNA που έχουν εντοπιστεί πειραματικά. (Α) Προβλεπόμενοι βρόγχοι των lin-4 και let-7 miRNAs στο C.elegans, με κοντινή ομολογία με τα miRNAs μυγών και θηλαστικών. (Β) Παραδείγματα miRNAs από γονίδια μεταζώων, με κοντινή ομολογία με τα miRNAs μυγών και θηλαστικών. (C) Παραδείγματα miRNAs από γονίδια φυτών[3].*

Το ανθρώπινο γονιδίωμα κωδικοποιεί περίπου 2.600 ώριμα miRNAs (miRBase v.22) και, σύμφωνα με τα δεδομένα της βάσης δεδομένων GENCODE (v.29), κωδικοποιεί περισσότερα από 200.000 μετάγραφα, συμπεριλαμβανομένων ισομορφών με μικρές παραλλαγές[4]. Ένα συγκεκριμένο miRNA μπορεί να στοχεύει πολλά διαφορετικά mRNAs[5], ενώ ένα συγκεκριμένο mRNA μπορεί να στοχευθεί από πολλά miRNA, είτε μεμονωμένα, είτε ταυτόχρονα[6].

Συνολικά, έχουν βρεθεί περισσότερες από 45.000 θέσεις-στόχος των miRNA εντός των ανθρώπινων 3'UTRs, οι οποίες είναι πολύ καλά διατηρημένες, ενώ >60% των γονιδίων που κωδικοποιούν ανθρώπινες πρωτεΐνες έχουν υποστεί επιλεκτική πίεση για να διατηρήσουν τη σύνδεση με τα miRNA[7].

Λόγω του ότι τα miRNA κυκλοφορούν και εξωκυτταρικά, μπορούν να χρησιμοποιηθούν ως βιοδείκτες σε πολλές ασθένειες (π.χ. Alzheimer), καθώς απελευθερώνονται σε σωματικά υγρά[8].

1.1. Ιστορική αναδρομή miRNAs

Το πρώτο miRNA ανακαλύφθηκε το 1993, από την ερευνητική ομάδα του Victor Ambros, οι οποίοι μελετούσαν το γονίδιο *lin-4* του οργανισμού *C. Elegans*, το οποίο ελέγχει το ρυθμό ανάπτυξης των προνυμφών του, μέσω καταστολής του *lin-14* γονιδίου. Κατά την απομόνωση του μετάγραφου του *lin-4* γονιδίου, διαπίστωσαν την ύπαρξη μικρών μη-κωδικών RNA, μήκους ~22 νουκλεοτιδίων, που περιείχαν αλληλουχίες μερικώς συμπληρωματικές με την 3΄ UTR περιοχή του *lin-14* mRNA, αντί να απομονώσουν το mRNA που κωδικοποιεί την *lin-4* πρωτεΐνη. Εκείνη τη χρονική στιγμή προτάθηκε πως τα μικρά αυτά μη-κωδικά μόρια RNA, αναστέλλουν την μετάφραση του *lin-14* mRNA στην αντίστοιχη πρωτεΐνη[9].

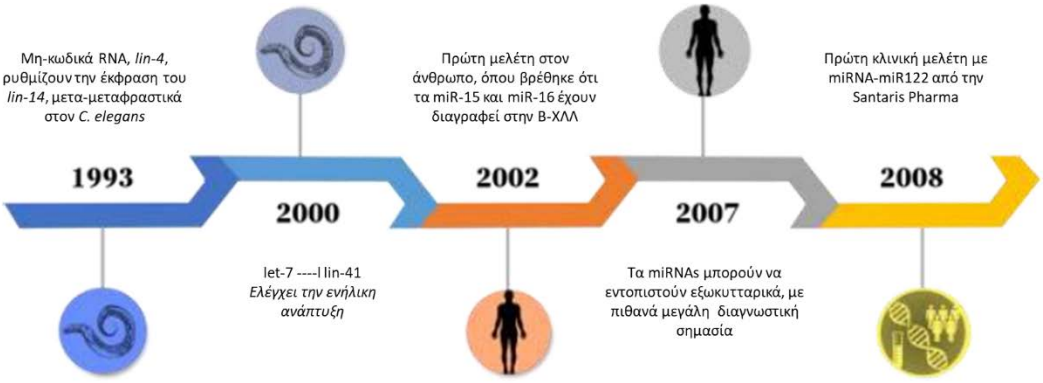
Στις αρχές του 2000, η ερευνητική ομάδα του Gary Ruvkun απομόνωσε ένα δεύτερο μη-κωδικό μόριο RNA, το *let-7* μήκους 21 νουκλεοτιδίων, το οποίο καταστέλλει το *lin-41* γονίδιο, προκειμένου να προωθήσει μια μεταγενέστερη αναπτυξιακή μετάβαση στο *C. elegans*. Το συγκεκριμένο μικρό μόριο RNA βρέθηκε πως είναι συμπληρωματικό με την 3΄ αμετάφραστη περιοχή των γονιδίων *lin-14, lin-28, lin-41, lin-42* και *daf-12*, υποδεικνύοντας ότι η έκφραση αυτών των γονιδίων μπορεί να ελέγχεται άμεσα από το *let-7*[10].

Η απάντηση δόθηκε αργότερα, το 2021, όπου βρέθηκε πως, τα lin-4 και let-7 RNAs ανήκουν σε μεγαλύτερες ομάδες μικρών ρυθμιστικών μη-κωδικών RNA, τα οποία εντοπίστηκαν, τόσο στον C. elegans και στη δροσόφιλα, όσο και σε ανθρώπινα κύτταρα. Σε αυτό το σημείο προτάθηκε ο όρος microRNA ή miRNA[11].

Η πρώτη μελέτη με miRNA σε ανθρώπινη ασθένεια ήταν το 2002, όπου οι Calin et al. διαπίστωσαν ότι οι περιοχές miR-15 και miR-16 είχαν διαγραφεί στη χρόνια Β-λεμφοκυτταρική λευχαιμία (Β-ΧΛΛ), σε δείγματα περιφερικού αίματος ασθενών που μελετήθηκαν[12].

Το 2008 η Santaris Pharma ανέπτυξε μια θεραπεία miRNA (Miravirsen) για την ηπατίτιδα C, όπου ένα ολιγονουκλεοτίδιο στοχεύει το miR-122 και χρησιμοποιεί μια έξυπνη στρατηγική για να εμποδίσει την αναπαραγωγή του ιού[13,14].

Τέλος, μέχρι τον Σεπτέμβριο του 2022, υπάρχουν καταχωρημένες 1180 κλινικές μελέτες σχετιζόμενες με microRNA (ανεξάρτητα της φάσης που βρίσκονται) στην Παγκόσμια βάση δεδομένων κλινικών μελετών (https://clinicaltrials.gov/).



Εικόνα 2. Χρονοδιάγραμμα με τα κομβικά σημεία από την ανακάλυψης των miRNA στο C. elegans έως τις κλινικές δοκιμές σε ανθρώπους (https://www.journaloflifesciences.org/archives/1537/snapshot-of-mirna-biology-progress-from-1993-to-2020.htm).

1.2. Βιοσύνθεση και Ωρίμανση των miRNAs

Η βιοσύνθεση και ωρίμανση των miRNAs αποτελεί μια αρκετά πολύπλοκη διαδικασία, η οποία περιλαμβάνει αρκετά στάδια, με κυριότερα την μεταγραφή του πρώιμου pri-miRNA από ένα δίκλωνο DNA, την μετατροπή του σε πρόδρομο pre-miRNA και τέλος την ωρίμανσή του σε miRNA.

Οι δύο υποψήφιες RNA πολυμεράσες για τη μεταγραφή σε pri-miRNA είναι οι πολυμεράση II (polymerase II -pol II) και η πολυμεράση III (polymerase III- pol III). Η pol II είναι υπεύθυνη για τη μεταγραφή τμημάτων του DNA που μεταφράζονται σε πρωτεΐνες (mRNA), καθώς και ορισμένα μη-κωδικά RNA, συμπεριλαμβανομένων των μικρών πυρηνικών RNA (snoRNAs) και τεσσάρων από τα μικρά πυρηνικά RNA (snRNAs) που εμπλέκονται στη διαδικασία του ματίσματος. Η pol III μεταγράφει μερικά από τα βραχύτερα σε μήκος μη-κωδικά RNA, όπως τα tRNAs, του ριβοσωμικού RNA 5S και το U6 snRNA[3].

Τα miRNAs που προέρχονται από τα ιντρόνια των γονιδίων που κωδικοποιούν πρωτεΐνες μεταγράφονται από τη πολυμεράση II. Ωστόσο, και τα υπόλοιπα miRNA είναι επίσης προϊόντα κατά κύριο λόγο της πολυμεράσης II και πιο σπάνια της πολυμεράσης III. Η βιοσύνθεσή τους ταξινομείται σε 2 οδούς: την κανονική και τη μη-κανονική.

1.2.1. Κανονική οδός βιοσύνθεσης των miRNAs

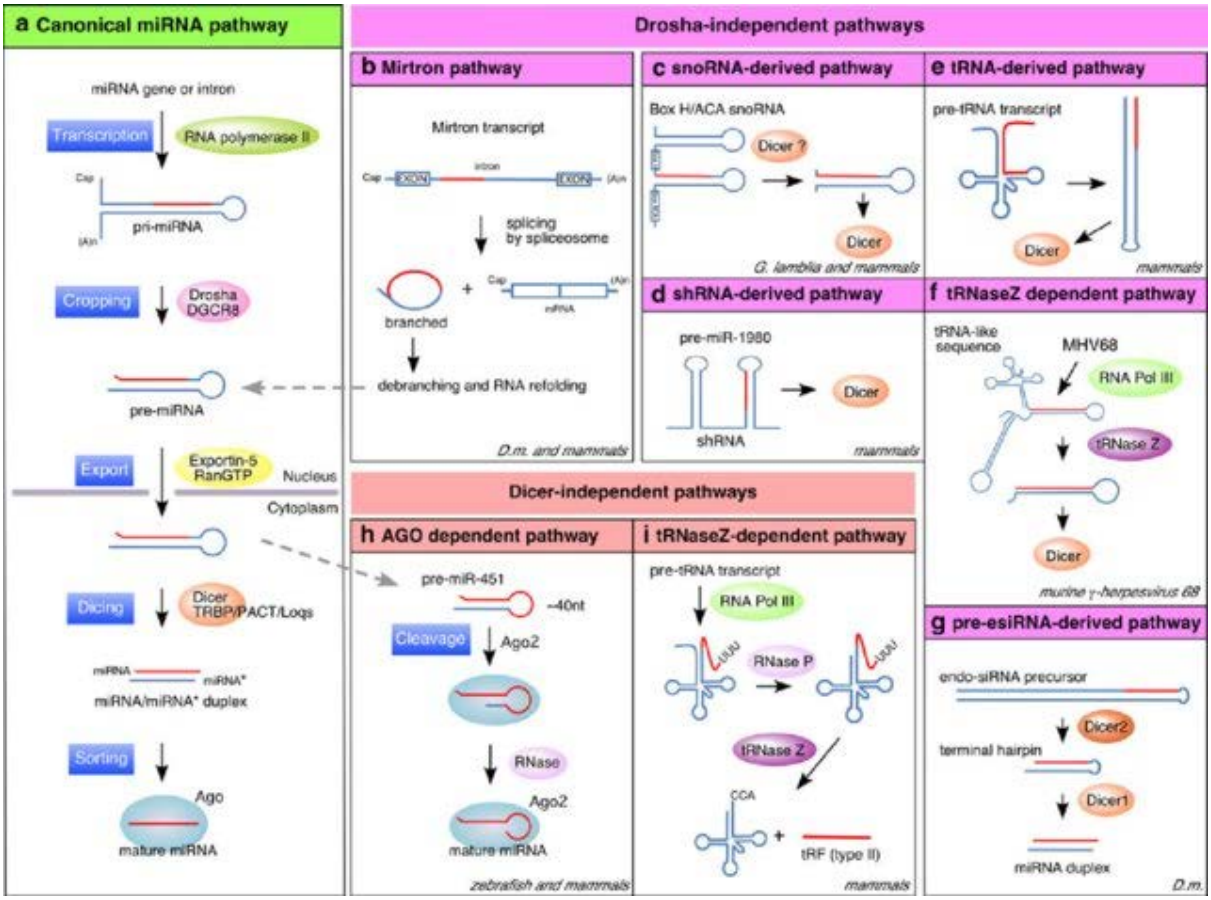
Η κανονική οδός βιοσύνθεσης είναι η κυρίαρχη οδός με την οποία επεξεργάζονται τα miRNA.

- Σε αυτό το μονοπάτι, τα pri-miRNA μεταγράφονται από τα γονίδιά τους και στη συνέχεια μετατρέπονται σε pre-miRNA από το σύμπλεγμα Drosha-DGCR8 [πρωτεΐνη δέσμευσης RNA - DiGeorge Syndrome Critical Region 8 (DGCR8) και ένα ένζυμο της ριβονουκλεάσης III – Drosha]. Το σύμπλεγμα DGCR8 αναγνωρίζει ένα N6-μεθυλαδενυλιωμένο GGAC και άλλα μοτίβα εντός του pri-miRNA για την ακριβή θέση πέψης του pri-miRNA, ενώ το ένζυμο Drosha διασπά το δίκλωνο pri-miRNA στη βάση της χαρακτηριστικής δομής φουρκέτας (hairpin) του pri-miRNA, οδηγώντας στο σχηματισμό του pre-miRNA με μία προεξοχή 2 νουκλεοτιδίων στο 3’ άκρο που αναγνωρίζεται από την Εξπορτίνη 5.

- Έτσι, τα pre-miRNA, εξάγονται από τον πυρήνα στο κυτταρόπλασμα από ένα σύμπλεγμα εξπορτίνης 5 (XPO5)/RanGTP.
- Στη συνέχεια τα pre-miRNA που βρίσκονται στο κυτταρόπλασμα υδρολύονται από το ένζυμο Dicer της ομάδας της ενδονουκλεάσης RNase III. Αυτή η επεξεργασία περιλαμβάνει την αφαίρεση του τερματικού βρόχου, με αποτέλεσμα ένα ώριμο διπλό miRNA. Η κατεύθυνση του κλώνου miRNA καθορίζει το όνομα της ώριμης μορφής miRNA. Ο κλώνος 5p προκύπτει από το 5' άκρο της φουρκέτας pre-miRNA, ενώ ο κλώνος 3p από το 3'άκρο.
- Και οι δύο κλώνοι που προέρχονται από το ώριμο διπλό miRNA μπορούν να προσδεθούν στην οικογένεια πρωτεϊνών Argonaute (AGO). Η επιλογή του κλώνου 5p ή 3p βασίζεται εν μέρει στη θερμοδυναμική σταθερότητα στα 5' άκρα του διπλού miRNA ή ενός 5' U στη θέση νουκλεοτιδίου 1. Γενικά, ο κλώνος με χαμηλότερη σταθερότητα 5' ή ουρακίλη 5' προσδένεται κατά προτίμηση στο AGO και θεωρείται ο οδηγός κλώνος. Ο κλώνος χωρίς φορτίο ονομάζεται κλώνος επιβατών και αποικοδομείται από την AGO2. Η πρόσδεση του κλώνου-οδηγό miRNA στις πρωτεΐνες AGO, οδηγεί στο σχηματισμό ενός miRNA συμπλόκου που οδηγεί στη γονιδιακή σίγαση (miRNA-Induced Silencing Complex).
- Το σύμπλοκο miRNA της γονιδιακής σίγασης που δημιουργήθηκε (miRNA-Induced Silencing Complex) προσδένεται με την πρωτεΐνη TRBP και δημιουργεί το σύμπλοκο RISC (RNAinduced silencing complex), το οποίο βοηθάει στην πρόσδεση του miRNA πάνω στη συμπληρωματική αλληλουχία του mRNA στόχου και τελικά τη ρύθμιση της γονιδιακής του έκφρασης[1].

1.2.2. Μη κανονική οδός βιοσύνθεσης των miRNAs

Μέχρι σήμερα, έχουν διευκρινιστεί πολλαπλά μη-κανονικά μονοπάτια βιοσύνθεσης των miRNAs. Αυτές οι οδοί χρησιμοποιούν διαφορετικούς συνδυασμούς των πρωτεϊνών που εμπλέκονται στην κανονική οδό, κυρίως των Drosha, Dicer, εξπορτίνη 5 και AGO2. Γενικά, η μη κανονική οδός βιοσύνθεσης του miRNA μπορεί να ομαδοποιηθεί σε μονοπάτια ανεξάρτητα του συμπλόκου Drosha/DGCR8 και σε μονοπάτια ανεξάρτητα του ενζύμου Dicer[1,15].



Εικόνα 3. Σχηματική αναπαράσταση της κανονικής οδού σύνθεσης των ώριμων miRNAs και των μη-κανονικών μονοπατιών που βιοσύνθεσης που έχουν αναγνωριστεί[15].

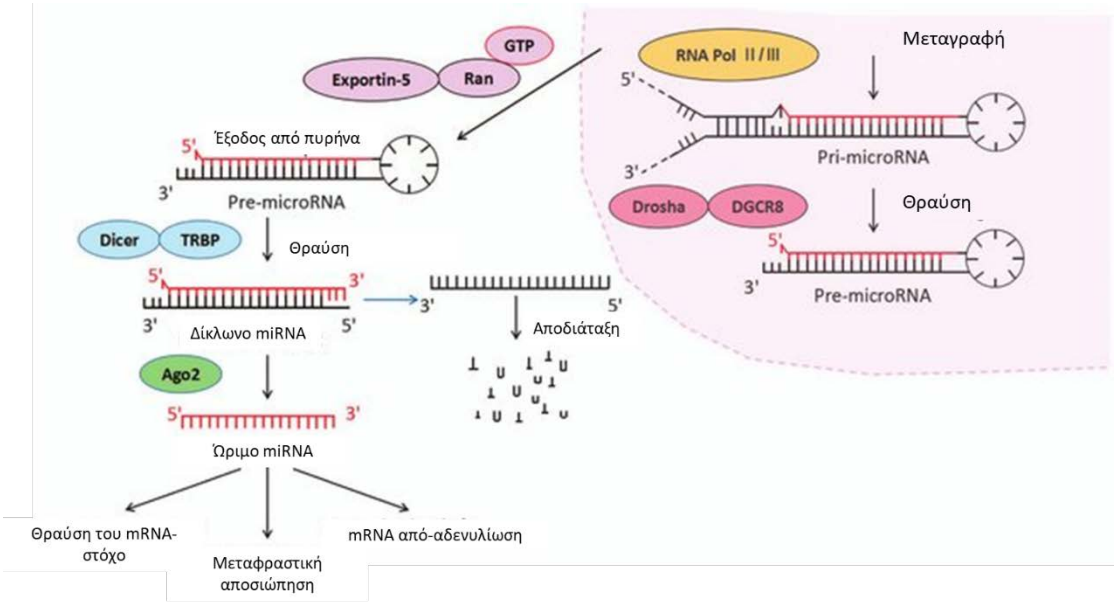
1.3. Μηχανισμοί ρύθμισης της γονιδιακής έκφρασης μέσω miRNA

Οι περισσότερες μελέτες μέχρι σήμερα έχουν δείξει ότι τα miRNA συνδέονται σε μια συγκεκριμένη αλληλουχία στο 3' UTR των mRNA-στόχων τους για να προκαλέσουν καταστολή της μετάφρασης και απο-αδενυλίωση του mRNA (**Εικόνα 4**). Ωστόσο, έχει βρεθεί πως τα miRNAs, έχουν την δυνατότητα να προσδένονται και σε άλλες περιοχές του mRNA, συμπεριλαμβανομένων της 5' UTR, κωδικών περιοχών, καθώς και εντός των περιοχών του υποκινητή.

Η δέσμευση των miRNAs στη 5' UTR και στις κωδικές περιοχές έχουν ως αποτέλεσμα την αποσιώπηση της γονιδιακής έκφρασης, ενώ η πρόσδεση των miRNAs στον υποκινητή των γονιδίων οδηγεί στην επαγωγή της μεταγραφής[1,16].

1.3.1. miRNA-επαγόμενη γονιδιακή αποσιώπηση μέσω του miRISC

Όπως αναφέρθηκε παραπάνω, το miRISC είναι το σύμπλοκο αποσιώπησης που δημιουργείται από τον οδηγό κλώνο και τις πρωτεΐνες AGO. Η εξειδίκευση στόχευσης του συμπλόκου miRISC οφείλεται στην αλληλεπίδρασή του με συμπληρωματικές αλληλουχίες στο mRNA στόχο, που ονομάζονται MREs (miRNA response elements). Ο βαθμός συμπληρωματικότητας των MRE καθορίζει εάν θα υπάρξει θραύση του mRNA στόχου από την AGO2 πρωτεΐνη ή miRISC-επαγόμενη αναστολή της μετάφρασης και διάσπαση του mRNA στόχου[1].



Εικόνα 4. Το miRNA μονοπάτι γονιδιακής ρύθμισης της έκφρασης. Ο λειτουργικός κλώνος του ώριμου miRNA πακετάρετε με τις πρωτεΐνες Argonaute (Ago2) και δημιουργείται το σύμπλεγμα αποσιώπησης - miRISC, όπου οδηγεί σε αποσιώπηση του mRNA-στόχο, είτε μέσω θραύσης του mRNA, είτε μέσω καταστολής της μετάφρασης ή αποαδενυλίωσης του mRNA[17].

1.3.2. miRNA-επαγόμενη ενεργοποίηση της μετάφρασης

Παρόλο που οι περισσότερες μελέτες των microRNAs, επικεντρώνονται στη γονιδιακή αποσιώπηση, υπάρχουν αναφορές για ενεργοποίηση της γονιδιακής έκφρασης. Ωστόσο, η miRNA-επαγόμενη υπερέκφραση των γονιδίων συμβαίνει κάτω από συγκεκριμένες περιπτώσεις και περιλαμβάνει τη συμμετοχή των AGO2 και FXR1 πρωτεϊνών, αντί των GW182 που αναφέρθηκαν προηγουμένως[1].

1.3.3. miRNA-επαγόμενη μεταγραφική και μετα-μεταγραφική γονιδιακή ρύθμιση εντός του πυρήνα

Μέσω της Ιμπορτίνης-8 ή της εξπορτίνης-1, η ανθρώπινη πρωτεΐνη AGO2 μετακινείται μεταξύ του πυρήνα και του κυτταροπλάσματος μέσω της αλληλεπίδρασής του με την TNRC6A (μια πρωτεΐνη της οικογένειας GW182) που περιέχει ένα σήμα πυρηνικού εντοπισμού και εξαγωγής. Το πυρηνικό εντοπισμένο miRISC βρέθηκε ότι ρυθμίζει την γονιδιακή έκφραση, τόσο σε μεταγραφικό, όσο και σε μετα-μεταγραφικό επίπεδο του mRNA. Ωστόσο, η κατανόησή και η γνώση μας για το πότε και πώς τα miRNA ασκούν τις λειτουργίες τους στον πυρήνα είναι ακόμα περιορισμένη[1].

1.4. Ο ρόλος των miRNAs στις ανθρώπινες ασθένειες

Τα microRNAs έχουν αποδειχθεί ότι παίζουν σημαντικό ρόλο σε ένα ευρύ φάσμα αναπτυξιακών διεργασιών, συμπεριλαμβανομένου του μεταβολισμού, του κυτταρικού πολλαπλασιασμού, της απόπτωσης, του ρυθμού ανάπτυξης και της μοίρας των νευρωνικών κυττάρων. Επίσης, άλλοι ρυθμιστικοί ρόλοι περιλαμβάνουν τη νευρωνική γονιδιακή έκφραση, τη μορφογένεση του εγκεφάλου, τη διαφοροποίηση των μυών και τη διαίρεση βλαστοκυττάρων[17]. Υπάρχουν τρεις κατηγορίες γενετικών μεταλλαγών που επηρεάζουν τη λειτουργία των miRNAs:

1. Αλλαγές στον αριθμό των αντιγράφων (Copynumber variations - CNV), που προσομοιάζουν μεταλλάξεις μεγάλης κλίμακας
2. Μονο-νουκλεοτιδικοί πολυμορφισμοί (Single nucleotide polymorphisms-SNP)
3. Επιγενετικές αλλαγές

Αυτές οι μεταλλαγές μπορούν να εντοπιστούν, τόσο στα γονίδια που κωδικοποιούν τα miRNA, όσο και στα γονίδια-στόχους, καθώς και στα γονίδια που είναι υπεύθυνα για την επεξεργασία των miRNA[18].

1.4.1. miRNAs και καρκίνος

Είναι πλέον γνωστό ότι η υπερέκφραση ή η υποέκφραση των miRNAs συμβαίνει σε διάφορους ανθρώπινους καρκίνους. Τα υπερεκφρασμένα miRNAs μπορεί να λειτουργήσουν και ως ογκογονίδια, καθώς οδηγούν στην υποέκφραση των ογκοκατασταλτικών γονιδίων ή/και ως ρυθμιστές κυτταρικών διεργασιών, όπως τη

διαφοροποίηση των κυττάρων ή την απόπτωση[17]. Τα προφίλ έκφρασης των miRNAs μπορεί να είναι χρήσιμα για διάγνωση καρκίνου, την πρόβλεψη της κλινικής έκβασης, καθώς και την επιλογή κατάλληλης φαρμακευτικής αγωγής. Επομένως, η επιδιόρθωση της έκφρασης των κατάλληλων miRNA, μπορεί να αποτελέσει μια νέα θεραπευτική προσέγγιση στον καρκίνο[18].

1.4.2. miRNAs και λοιπές ασθένειες

Έχει αποδειχθεί ότι τα miRNAs, εκτός από τον καρκίνο, παίζουν καθοριστικό ρόλο στα αυτοάνοσα νοσήματα, στις νευρολογικές ασθένειες και στις καρδιαγγειακές παθήσεις.

Τα miRNA συμμετέχουν στην ανάπτυξη των κυττάρων και τη διατήρηση των λειτουργιών του ανοσοποιητικού συστήματος. Τροποποιημένη έκφραση των microRNAs έχει συσχετιστεί με διάφορες αυτοάνοσες διαταραχές, όπως η ρευματοειδής αρθρίτιδα, ο Συστημικός ερυθηματώδης Λύκος και η σκλήρυνση κατά πλάκας.

Επιπλέον, αρκετές ερευνητικές ομάδες ανέφεραν ότι ορισμένα miRNA εκφράζονται μόνο στο κεντρικό νευρικό σύστημα, συμμετέχοντας ουσιαστικά στην ανάπτυξη του εγκεφάλου και στις φυσιολογικές λειτουργίες. Η χρησιμότητα των miRNAs στη λειτουργία του ΚΝΣ επιβεβαιώθηκε, με μελέτες knockdown των μονοπατιών βιοσύνθεσης των miRNAs, όπου η αφαίρεση του Dicer οδήγησε σε ανώριμη νευρογένεση, ενώ η διαγραφή του DGCR προκάλεσε αλλοιωμένο σχηματισμό της σπονδυλικής στήλης. Τέλος, πολλές μελέτες υποδηλώνουν ότι τα miRNA διαδραματίζουν συγκεκριμένο ρόλο στην καρδιαγγειακή ανάπτυξη και διαταραχές[18].

1.5. Χρήση των miRNA ως βιοδείκτες για διαγνωστικές και θεραπευτικές προσεγγίσεις

Τα χαρακτηριστικά ενός ιδανικού βιοδείκτη είναι τα εξής:

- Πρέπει να είναι εύκολα προσβάσιμος,
- Να έχει υψηλή εξειδίκευση για τον τύπο ιστού ή κυττάρου προέλευσης και να είναι ευαίσθητο στον τρόπο που ποικίλλει ανάλογα με την εξέλιξη της νόσου.
- Να είναι εύκολα μεταφράσιμος από την έρευνα στην κλινική πράξη

Τα τελευταία χρόνια έχει βρεθεί πως τα miRNAs αποτελούν έναν ιδανικό βιοδείκτη, καθώς μπορούν να εντοπιστούν και να απομονωθούν από εύκολα προσβάσιμους ιστούς, όπως μέσω υγρών βιοψιών από αίμα, ούρα και άλλα σωματικά υγρά. Επίσης, έχει δειχθεί σε αρκετές μελέτες, πως συμμετέχουν στη διαφοροποίηση των καρκινικών σταδίων και άλλων ασθενειών, ενώ ακόμη χρησιμοποιούνται και για τη μέτρηση της ανταπόκρισης στη θεραπεία. Επιπλέον, οι τεχνολογίες για την ανίχνευση νουκλεϊκών οξέων υπάρχουν ήδη, με αποτέλεσμα η ανάπτυξη νέων αναλύσεων να συνεπάγεται λιγότερο χρόνο και χαμηλότερο κόστος σε σύγκριση με την παραγωγή νέων αντισωμάτων για πρωτεϊνικούς βιοδείκτες. Τέλος, ακόμη ένα πλεονέκτημα των miRNA έγκειται στη δυνατότητά τους να χρησιμοποιηθούν ως μοντέλα πολλαπλών δεικτών για ακριβή διάγνωση, καθοδηγούμενη και εξατομικευμένη θεραπεία, καθώς και αξιολόγηση της ανταπόκρισης στη θεραπεία[19].

2. Προγράμματα στοίχισης

Τα τελευταία χρόνια η βελτίωση της αποτελεσματικότητας στην αλληλούχιση του DNA και του RNA έχουν διευρύνει τις εφαρμογές αλληλούχισης και έχουν αυξήσει δραματικά το μέγεθος των συνόλων δεδομένων (sequencing datasets). Εταιρείες τεχνολογίας και ανάπτυξης, όπως η Illumina (San Diego, CA, USA) και Applied Biosystems (Foster City, CA, USA), έχουν αναπτύξει μεθοδολογίες για την εύρεση προφίλ προτύπων μεθυλίωσης (MeDIP-Seq), για τη χαρτογράφηση των αλληλεπιδράσεων DNA-πρωτεΐνης (ChIP-Seq), καθώς και για τον εντοπισμό διαφορικά εκφραζόμενων γονιδίων (RNA-Seq), τόσο στο ανθρώπινο γονιδίωμα, όσο και σε άλλα είδη.

Κάθε μία από αυτές τις μελέτες απαιτούσε την στοίχιση μεγάλου αριθμού μικρών ακολουθιών DNA («short reads») στο ανθρώπινο γονιδίωμα. Με τις υπάρχουσες μεθόδους, το υπολογιστικό κόστος της στοίχισης πολλών small reads σε ένα γονιδίωμα θηλαστικού είναι πολύ μεγάλο. Για παράδειγμα, προκειμένου να γίνει στοίχιση των 140 δισεκατομμυρίων βάσεων, θα απαιτούνταν περισσότεροι από 5 CPU-μήνες χρησιμοποιώντας τον αλγόριθμο Maq, ενώ περισσότερα από 3 CPU-έτη στην περίπτωση του SOAP[20].

Αν και η χρήση αυτών των εργαλείων έχει αποδειχθεί αποδοτική, είναι σαφές πως υπάρχει ανάγκη εύρεσης νέων εργαλείων που καταναλώνουν λιγότερο χρόνο και υπολογιστικούς πόρους. Κάποια από τα εργαλεία που πληρούν αυτές τις απαιτήσεις

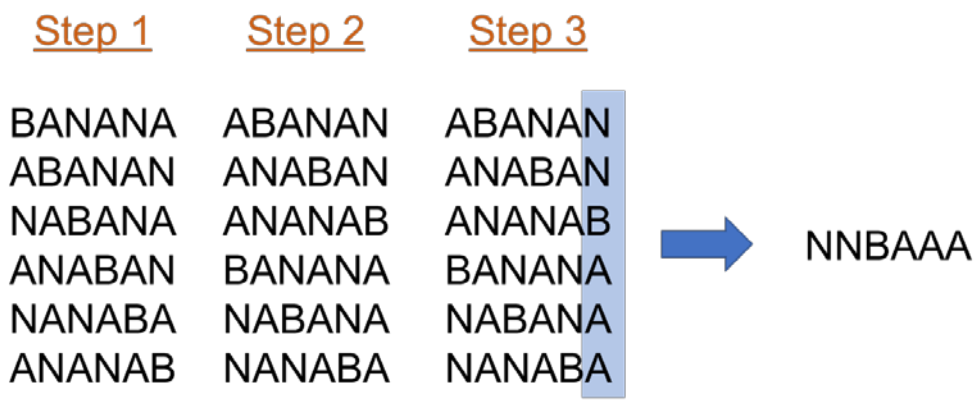
και που θα μελετηθούν στην παρούσα εργασία είναι το Bowtie, το Bowtie 2, το HiSat2 και το STAR.

2.1. Bowtie aligner

Το Bowtie είναι ένα εξαιρετικά γρήγορο πρόγραμμα στοίχισης ανοιχτού κώδικα (http://bowtie.cbcb.umd.edu), με αποδοτική μνήμη, για την στοίχιση μικρών τμημάτων DNA αλληλουχιών (έως 50 bp) σε μεγάλα γονιδιώματα. Είναι γραμμένο σε C++, χρησιμοποιεί τη SeqAn βιβλιοθήκη και υποστηρίζει FASTQ και FASTA μορφές.

Το Bowtie δημιουργεί indexes του γονιδιώματος αναφοράς χρησιμοποιώντας ένα πρότυπο που βασίζεται στον μετασχηματισμό Burrows-Wheeler (BWT) και στο FM index. Με το Burrows-Wheeler Index το αποτύπωμα μνήμης χαμηλό˙ για το ανθρώπινο γονιδίωμα το τυπικό index για unpaired alignment είναι 2,2 GB, ενώ για paired-end alignment είναι 2,9 GB. Η μέθοδος αναζήτησης στο FM index είναι ο αλγόριθμος ακριβούς αντιστοίχισης των Ferragina και Manzini.

Ο Burrows-Wheeler μετασχηματισμός είναι ένας τρόπος ταξινόμησης και ομαδοποίησης δεδομένων, με τέτοιο τρόπο ώστε παρόμοια γράμματα να ομαδοποιούνται για να είναι ευκολότερη η συμπίεση των δεδομένων. Για παράδειγμα, η λέξη «BANANA» μετά τον Burrows-Wheeler μετασχηματισμό μετατρέπεται στη λέξη «NNBAAA. Επειδή τα "A" είναι ομαδοποιημένα, είναι πλέον πολύ πιο εύκολο να συμπιεστούν. Ο τρόπος λειτουργίας του συγκεκριμένου αλγορίθμου αποτελείται από 3 βήματα. Αρχικά δημιουργείται ένας πίνακας ξεκινώντας από τη λέξη «BANANA», στον οποίο το τελευταίο γράμμα της λέξης έρχεται στην αρχή δημιουργώντας όλους τους πιθανούς συνδυασμούς. Στη συνέχεια, ακολουθεί η ταξινόμηση των δεδομένων του πίνακα αλφαβητικά. Αν το πρώτο γράμμα ξεκινώντας από την αριστερή στήλη είναι ίδιο, χρησιμοποιείται αυτό της δεύτερης στήλης, κ.ο.κ. Στο τέλος, χρησιμοποιείται ως output η σειρά των γραμμάτων της τελευταίας στήλης, όπου στην περίπτωσή μας είναι «NNBAAA» (**Εικόνα 5**). Ο Burrows-Wheeler μετασχηματισμός είναι ένας αλγόριθμος συμπίεσης χωρίς απώλειες, που προσφέρει τόσο συμπίεση, όσο και ανάκτηση της συμβολοσειράς, όποτε χρειάζεται.



Εικόνα 5. Βήματα που ακολουθεί ο Burrows-Wheeler αλγόριθμος για τον σχηματισμό δεδομένων.

Το FM-index είναι ένας suffix-tree αλγόριθμος που βασίζεται στον Burrow-Wheeler μετασχηματισμό. Χρησιμοποιεί έναν μηχανισμό αναζήτησης προς τα πίσω, στο αποτέλεσμα του BWT, ο οποίος επιτρέπει την εύρεση ακριβών αντιστοιχίσεων μοτίβων κατά τη διάρκεια βημάτων, που είναι αναλογικά με το μήκος του μοτίβου που αναζητείται και επίσης είναι ανεξάρτητα από το μέγεθος της ακολουθίας αναφοράς. Εκτός από την καλή υπολογιστική του πολυπλοκότητα, επιτυγχάνει υψηλές αναλογίες συμπίεσης, επιτρέποντας το indexing του πλήρους ανθρώπινου γονιδιώματος σε λιγότερο από 1,5 GB χώρου μνήμης.

Ωστόσο, το Bowtie δεν υιοθετεί ακριβώς αυτόν τον αλγόριθμο, λόγω του ότι η ακριβής αντιστοίχιση δεν επιτρέπει σφάλματα αλληλουχίας ή γενετικές παραλλαγές. Για το λόγο αυτό εισήγαγαν δύο νέες επεκτάσεις που κάνουν την τεχνική εφαρμόσιμη σε στοίχιση small reads: α) έναν αλγόριθμο οπισθοδρόμησης με επίγνωση της ποιότητας που επιτρέπει αναντιστοιχίες και ευνοεί τις στοιχίσεις υψηλής ποιότητας και β) «double indexing», μια στρατηγική για την αποφυγή υπερβολικής οπισθοδρόμησης[20].

Το Bowtie μετά την στοίχιση παράγει αρχεία τύπου SAM, επιτρέποντας με αυτόν τον τρόπο στο Bowtie να λειτουργεί συνδυαστικά με άλλα εργαλεία που υποστηρίζουν τα αρχεία SAM, συμπεριλαμβανομένων των SAMtools. Τέλος, αποτελεί την βάση άλλων εργαλείων, όπως το TopHat, το Cufflinks, το Crossbow και το Myrna.

Συνοπτικά λοιπόν το Bowtie έχει σχεδιαστεί για να είναι εξαιρετικά γρήγορο για σύνολα μικρών reads όπου, (α) πολλά από τα reads έχουν τουλάχιστον μία καλή, έγκυρη στοίχιση, (β) πολλά από τα reads είναι σχετικά υψηλής ποιότητας και (γ) ο αριθμός των στοιχίσεων που αναφέρεται ανά read είναι μικρός (ιδανικά στο 1). Αυτά τα κριτήρια ικανοποιούνται στο πλαίσιο αναλύσεων όπως RNA-seq, ChIP-seq, κ.α.

2.1.1. Τρόπος λειτουργίας του Bowtie

Το bowtie παίρνει ένα index και ένα σύνολο από reads ως input και εξάγει ως output μια λίστα στοιχίσεων. Οι στοιχίσεις επιλέγονται σύμφωνα με

- έναν συνδυασμό των επιλογών *-v*/*-n*/*-e*/*-l* (συν τις επιλογές *-I*/*-X*/*--fr*/*--rf*/ *--ff* για paired-end στοίχιση), οι οποίες ορίζουν ποιες στοιχίσεις είναι νόμιμες
- τις επιλογές *-k*/*-a*/*-m*/*-M*/*--best*/*--strata* που ορίζουν ποιες και πόσες νόμιμες στοιχίσεις θα πρέπει να αναφέρονται.
- Οι στοίχιση που περιλαμβάνουν έναν ή περισσότερους διφορούμενους χαρακτήρες αναφοράς (N, -, R, Y, κ.λπ.) θεωρούνται άκυρες από το Bowtie. Αυτό ισχύει μόνο για διφορούμενους χαρακτήρες στην αναφορά.

Από προεπιλογή, το Bowtie επιβάλλει μια πολιτική στοίχισης παρόμοια με την προεπιλεγμένη πολιτική Maq (*-n* 2 *-l* 28 *-e* 70). Επίσης, επιβάλει μια απλούστερη πολιτική end-to-end k-διαφορών (*-v* 2). Οι δύο λειτουργίες στοίχισης *-n* και *-v* αποκλείονται αμοιβαία. Η διαδικασία με την οποία το Bowtie επιλέγει να αναφέρει μια στοίχιση, είναι τυχαία, προκειμένου να αποφευχθεί η «προκατάληψη της χαρτογράφησης (mapping bias)».

Ωστόσο, στην προεπιλεγμένη λειτουργία, το Bowtie μπορεί να εμφανίζει μεροληψία, όσον αφορά στον κλώνο. Η μεροληψία κλώνου συμβαίνει όταν το γονιδίωμα αναφοράς και τα reads είναι τέτοια ώστε

a. Ορισμένα reads στοιχίζονται εξίσου καλά και με τους δυο κλώνους της αναφοράς και
b. ο αριθμός τέτοιων θέσεων μεταξύ των δύο κλώνων είναι διαφορετικός.

Όταν αυτό συμβαίνει για ένα read, το Bowtie επιλέγει αποτελεσματικά τον ένα κλώνο ή τον άλλο με πιθανότητα 50% και, στη συνέχεια, αναφέρει μια τυχαία επιλεγμένη στοίχιση του read. Ωστόσο, η λειτουργία *--best* του Bowtie, οδηγεί στην εξάλειψη της προκατάληψης του κλώνου αναγκάζοντας το Bowtie έναν από τους δύο κλώνους με πιθανότητα ανάλογη με τον αριθμό των καλύτερων θέσεων στον κλώνο.

Οι στοιχίσεις με κενά δεν υποστηρίζονται αυτήν τη στιγμή στο Bowtie, αλλά υποστηρίζονται στο Bowtie 2.

Η *-n* λειτουργία στοίχισης

Η επιλογή του bowtie *-n* (η οποία είναι η προεπιλογή), καθορίζει ποιες στοιχίσεις είναι έγκυρες σύμφωνα με την ακόλουθη πολιτική, η οποία είναι παρόμοια με την προεπιλεγμένη πολιτική του Maq.

1. Οι στοιχίσεις δεν πρέπει να έχουν περισσότερες από *N* αναντιστοιχίες (όπου *N* είναι αριθμός 0-3, ορίζεται με *-n*) στις πρώτες βάσεις *L* (όπου L είναι αριθμός 5 ή μεγαλύτερος, ορίζεται με *-l*) στο υψηλής ποιότητας άκρο (αριστερό) του read. Οι πρώτες βάσεις *L* ονομάζονται «seeds».
2. Το άθροισμα των τιμών ποιότητας Phred σε όλες τις αναντιστοιχίες (όχι μόνο στο seed) δεν μπορεί να υπερβαίνει το *E* (που ορίζεται με *-e*). Όπου οι τιμές-ποιότητας δεν είναι διαθέσιμες (π.χ. εάν τα reads προέρχονται από ένα αρχείο FASTA), η προεπιλεγμένη ποιότητα Phred είναι 40.

Η *-v* λειτουργία στοίχισης

Στη λειτουργία *-v*, οι στοιχίσεις δεν μπορούν να έχουν περισσότερες από V αναντιστοιχίες (όπου το V μπορεί να είναι ένας αριθμός από το 0 έως το 3 που ορίζεται χρησιμοποιώντας την επιλογή *-v*). Οι τιμές ποιότητας σε αυτήν την επιλογή αγνοούνται.

Γραμμή εντολής

bowtie [options]* -x <ebwt> {-1 <m1> -2 <m2> | --12 <r> | --interleaved <i> | <s>} [<hit>]

όπου,

-x <ebwt> : Το όνομα του index όπου θα ψάξει το Bowtie

<m1> : Λίστα αρχείων διαχωρισμένων με κόμμα που περιέχουν τα #1 ζεύγη (το όνομα του αρχείου συνήθως περιλαμβάνει _1) ή, εάν έχει καθοριστεί *-c*, τις ίδιες τις αλληλουχίες του ζεύγους.

<m2> : Λίστα αρχείων διαχωρισμένων με κόμμα που περιέχουν τα #2 ζεύγη (το όνομα του αρχείου συνήθως περιλαμβάνει _2) ή, εάν έχει καθοριστεί *-c*, τις ίδιες τις αλληλουχίες του ζεύγους.

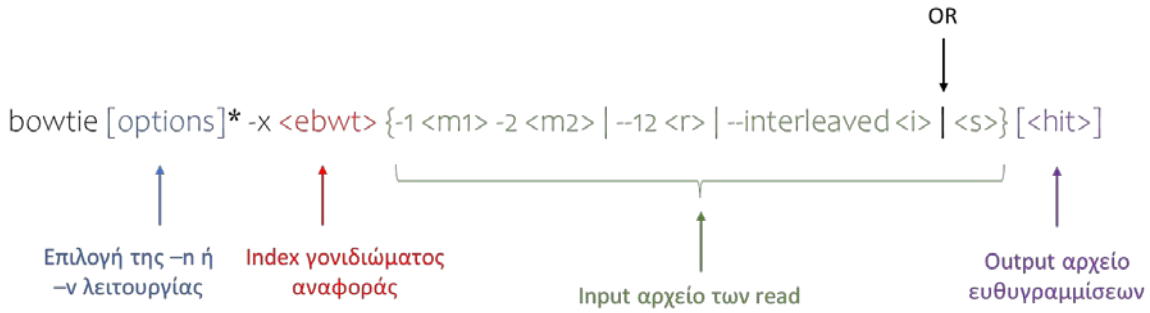
<r> : Λίστα αρχείων διαχωρισμένων με κόμμα που περιέχει έναν συνδυασμό unpaired και paired-end reads σε Tab-οριοθετημένη μορφή. Η Tab-

οριοθετημένη μορφή είναι μια μορφή 1 read ανά γραμμή, όπου τα unpaired reads αποτελούνται από ένα όνομα, μια ακολουθία και μια συμβολοσειρά ποιότητας που διαχωρίζονται με καρτέλες. Η αντίστοιχη μορφή για τα paired-end reads αποτελείται από ένα όνομα, ακολουθία του #1 ζεύγους, τιμές ποιότητας του #1 ζεύγους, ακολουθία του #2 ζεύγους και τιμές ποιότητας του #2 ζεύγους διαχωρισμένες με καρτέλες. Οι τιμές ποιότητας μπορούν να εκφραστούν χρησιμοποιώντας οποιαδήποτε από τις κλίμακες που υποστηρίζονται στα αρχεία FASTQ. Τα reads μπορεί να είναι ένας συνδυασμός διαφορετικών μηκών και unpaired και paired-end reads μπορούν να αναμιγνύονται στο ίδιο αρχείο.

<i> : Μια λίστα διαχωρισμένη με κόμματα με παρεμβαλλόμενα paired-end αρχεία FASTQ, όπου οι εγγραφές για το ζεύγος #1 παρεμβάλλονται με τις εγγραφές για το ζεύγος #2. Τα reads μπορεί να είναι ένας συνδυασμός διαφορετικών μηκών.

<s> : Μια λίστα αρχείων διαχωρισμένων με κόμματα που περιέχουν unpaired reads για στοίχιση ή, εάν έχει καθοριστεί -c, οι ίδιες οι unpaired ακολουθίες των reads.

<hit> : Αρχείο για την εγγραφή των στοιχίσεων. Από προεπιλογή, οι στοιχίσεις εγγράφονται σε ένα "standard out" αρχείο.



Εικόνα 6. Επεξήγηση της γραμμής εντολών του Bowtie αλγορίθμου.

Επιλογές του Bowtie για να ελέγξει τα αποτελέσματα της στοίχισης

-k<int> : Αναφορά έως και <int> στοιχίσεις για κάθε read (default *-k* 1)

-a : Αναφορά όλων των έγκυρων στοιχίσεων για κάθε read

-m<int> : Αποτρέπει την αναφορά των reads με >int στοιχίσεις (χρήση του *-m* 1 για την αναφορά μοναδικών χαρτογραφημένων reads)

-S: για το SAM output

2.1.2. Χρήση του Bowtie

Ενδεικτικά ο τρόπος χρήσης του Bowtie είναι ο εξής:

1. Δημιουργία index του γονιδιώματος αναφοράς χρησιμοποιώντας το Bowtie-build
2. Γραμμή εντολής

   Προεπιλογή εντολής του Bowtie (default setting):

   ```
   $ bowtie -S index reads.fq bowtie-out.sam 2> bowtie-out.stderr
   ```

   Στην περίπτωση των paired-end reads:

   ```
   $ bowtie -S index -1 fwd_reads.fq -2 rev_reads.fq bowtie-out.sam
   ```

2.2. Bowtie 2 aligner

Από την άλλη, το Bowtie 2, είναι γρηγορότερο, εμφανίζει μεγαλύτερη ακρίβεια και είναι ιδιαίτερα καλό στην στοίχιση μεγαλύτερων τμημάτων 50 έως 100 ή 1.000 χαρακτήρων με σχετικά μεγάλα (π.χ. θηλαστικά) γονιδιώματα. Το Bowtie 2 χρησιμοποιεί ως index γονιδίωμα σε συνδυασμό με το FM index, προκειμένου να διατηρεί το αποτύπωμα μνήμης μικρό. Συγκεκριμένα, για το ανθρώπινο γονιδίωμα, το αποτύπωμα μνήμης του είναι συνήθως περίπου 3,2 GB.

Αναλυτικά οι διαφορές μεταξύ του Bowtie και του Bowtie 2 συνοψίζονται ως εξής:

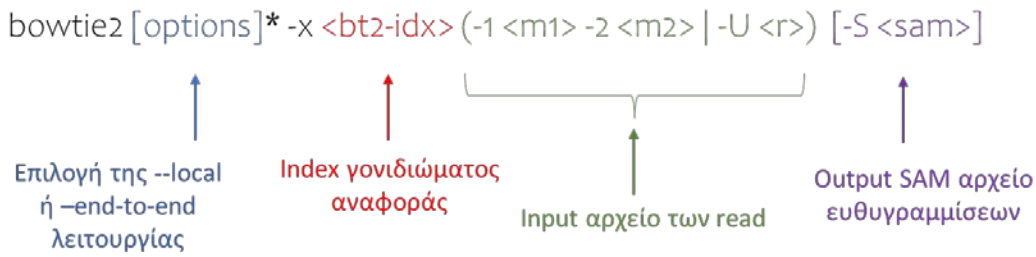
1. Το Bowtie 2 υποστηρίζει πλήρως την στοίχιση με κενό με γραμμικές ποινές για τα κενά, ενώ το Bowtie βρίσκει μόνο στοιχίσεις χωρίς κενό.
2. Για τμήματα μεγαλύτερα από ~50 bp, το Bowtie 2 είναι ταχύτερο, πιο ευαίσθητο και χρησιμοποιεί λιγότερη μνήμη από το Bowtie. Για σχετικά

σύντομες ακολουθίες (π.χ. λιγότερο από 50 bp) το Bowtie είναι μερικές φορές πιο γρήγορο ή/και πιο ευαίσθητο.

3. Το Bowtie 2 υποστηρίζει την τοπική στοίχιση, η οποία δεν απαιτεί στοίχιση από άκρο σε άκρο «end-to-end». Ωστόσο, μπορεί να υποστηρίξει και την «end-to-end» στοίχιση, η οποία, όπως και στην περίπτωση του Bowtie, απαιτεί την πλήρη στοίχιση του τμήματος.
4. Δεν υπάρχει ανώτατο όριο στο μήκος των ακολουθιών στο Bowtie 2, ενώ το Bowtie έχει ένα ανώτερο όριο περίπου στα 1000 bp.
5. Στο Bowtie 2 όλες οι στοιχίσεις βρίσκονται σε ένα συνεχές φάσμα βαθμολογιών στοίχισης.
6. Υπάρχει μόνο ένα σχήμα βαθμολόγησης, παρόμοιο με το Needleman-Wunsch και το Smith-Waterman.
7. Η λειτουργία paired-end στοίχισης του Bowtie 2 είναι πιο ευέλικτη από αυτή του Bowtie. Για παράδειγμα, για ζεύγη που δεν στοιχίζονται με αυτόν τον τρόπο, θα προσπαθήσει να βρει unpaired στοιχίσεις.
8. Το Bowtie 2 δεν στοιχίζει τα χρωματισμένα reads[21,22].

Γραμμή εντολής

bowtie2 [options]* -x <bt2-idx> (-1 <m1> -2 <m2> | -U <r>) [-S <sam>]



Εικόνα 7. Επεξήγηση της εντολής του bowtie 2.

2.2.1. Bowtie 2 end-to-end στοίχιση

Η συγκεκριμένη λειτουργία αποτελεί προεπιλογή του αλγορίθμου και επιτρέπει την στοίχιση ολόκληρου το read στο γονιδίωμα αναφοράς (*--end-to-end* επιλογή). Επίσης, ελέγχει και βαθμολογεί παραμέτρους όπως ο αριθμός των κενών και το μήκος.
Οι επιλογές χρήσης του αλγορίθμου είναι οι εξής:

- Πολύ γρήγορος
- Γρήγορος
- Πολύ ευαίσθητος
- Ευαίσθητος (προεπιλογή αλγορίθμου) *-D* 15 *-R* 2 *-L* 22 *-i* S,1,1.15

Όπου:

-D: καθορίζει το πόσες επεκτάσεις θα δοκιμάσει για ένα δεδομένο seed που ταιριάζει

-L: υπαγορεύει το μήκος των seed σε ένα μια multiseed στοίχιση

2.2.2. Bowtie 2 τοπική (local) στοίχιση

Η τοπική επιλογή (--local επιλογή) επιτρέπει ένα soft trimming των 3΄ και 5΄ άκρων και περιλαμβάνει κενά. Όπως και η end-to-end στοίχιση, η τοπική στοίχιση έχει αρκετές προ-εγκατεστημένες επιλογές λειτουργίας με την ευαίσθητη λειτουργία να αποτελεί την προεπιλογή (*-D* 15 *-R* 2 *-N* 0 *-L* 20 *-i* S,1,0.75).
Όπου:

-N: υπαγορεύει τις αναντιστοιχίες που επιτρέπονται στην περιοχή του seed

Επιλογές του Bowtie για να ελέγξει τα αποτελέσματα της στοίχισης

Ως προεπιλογή αναφέρει μια μοναδική στοίχιση ανά read

-k : Ρυθμίζει το ποσό των στοιχίσεων ανά read

-a : Αναφορά όλων των έγκυρων στοιχίσεων

2.2.3. Χρήση του Bowtie 2

Ενδεικτικά ο τρόπος χρήσης του Bowtie 2 είναι ο εξής:

1. Δημιουργία index του γονιδιώματος αναφοράς χρησιμοποιώντας το bowtie 2-build. Τα indexes του bowtie δεν είναι συμβατά με τα indexes που δημιουργούνται από το bowtie 2 και αντιστρόφως.
2. Γραμμή εντολής

   Προεπιλογή εντολής του Bowtie (default setting):

   ```
   $ bowtie2 –local -x index reads.fq -S bowtie-out.sam 2> bowtie-out.stderr
   ```

   Στην περίπτωση των paired-end reads:

   ```
   $ bowtie2 –fr --local -x index -1 fwd_reads.fq -2 rev_reads2.fq -S bowtie-out.sam
   ```

2.3. HISAT2 aligner

Το HISAT2 είναι ένα γρήγορο και ευαίσθητο πρόγραμμα στοίχισης για τη χαρτογράφηση τμημάτων DNA και RNA (αποτελέσματα αλληλούχισης επόμενης γενιάς), τόσο σε ένα σύνολο ανθρώπινων γονιδιωμάτων, όσο και σε άλλα μοναδικά γονιδιώματα αναφοράς. Για πρώτη φορά, με βάση μια επέκταση του BWT για γραφήματα, σχεδιάστηκε ένα graph FM index (GFM)[23]. Εκτός από τη χρήση ενός παγκόσμιου GFM index που αντιπροσωπεύει έναν πληθυσμό ανθρώπινων γονιδιωμάτων, το HISAT2 χρησιμοποιεί και ένα μεγάλο σύνολο μικρών GFM indexes που καλύπτουν συλλογικά ολόκληρο το γονιδίωμα.

Αυτά τα μικρά indexes (local indexes), σε συνδυασμό με διάφορες στρατηγικές στοίχισης, επιτρέπουν γρήγορη και ακριβή στοίχιση των τμημάτων[24,25].

2.3.1. HISAT2 flags

Υπάρχουν συγκεκριμένες εντολές για τον συγκεκριμένο aligner.

- Δημιουργία indexes

  ```
  hisat2-build <commands>
  ```

- Χρήση HISAT2

  ```
  hisat2 <commands>
  ```

- Λειτουργία σε πολλαπλούς επεξεργαστές (processors/cores)

  ```
  -p/ --threads
  ```

- Απενεργοποίηση του splice alignment

--no-spliced-alignment

- Αναφορά των στοιχίσεων για τη σύνδεση των μεταγράφων με StringTie

--dta/--downstream-transcriptome-assembly

Από προεπιλογή, το HISAT2 μπορεί να κάνει soft-clip των read κοντά στα 3' και 5' άκρα. Οι χρήστες μπορούν να ελέγξουν αυτήν την επιλογή ορίζοντας διαφορετικές ποινές για soft-clipping (*--sp*) ή απαγορεύοντας το soft-clipping (*[--no-softclip]*).

Γραμμή εντολής

hisat2 [options]* -x <hisat2-idx> {-1 <m1> -2 <m2> | -U <r> | --sra-acc <SRA accession number>} [-S <hit>]

όπου,

-x <hisat2-idx> : Το όνομα του index για το γονιδίωμα αναφοράς

-1 <m1> : Λίστα αρχείων διαχωρισμένων με κόμμα που περιέχουν τα ζεύγη1s (το όνομα του αρχείου συνήθως περιλαμβάνει _1)

-2 <m2> : Λίστα αρχείων διαχωρισμένων με κόμμα που περιέχουν τα ζεύγη 2s (το όνομα του αρχείου συνήθως περιλαμβάνει _2)

-U <r> : Λίστα αρχείων διαχωρισμένων με κόμμα που περιέχει unpaired reads για στοίχιση. Τα reads μπορεί να είναι ένας συνδυασμός διαφορετικών μηκών

--sra-acc <SRA accession number> :

Μια λίστα διαχωρισμένη με κόμματα των SRA accession number

-S <hit> : SAM αρχείο για την εγγραφή των στοιχίσεων. Από προεπιλογή, οι στοιχίσεις εγγράφονται σε ένα "standard out" αρχείο.

2.3.2. Χρήση του HISAT2

Ενδεικτικά ο τρόπος χρήσης του HISAT2 είναι ο εξής:

1. Δημιουργία indexes του γονιδιώματος αναφοράς. Υπάρχουν έτοιμα (prebuilt) indexes στην ιστοσελίδα του HISAT2.
   - Αρχικά γίνεται εξαγωγή των θέσεων ματίσματος και των εξονίων, π.χ.

     ```
     $ hisat2_extract_splice_sites.py chr1_data/chr1.gtf >chr1.ss
     $ hisat2_extract_exons.py chr1_data/chr1.gtf >chr1.ss
     ```

   - Δημιουργία του index

     ```
     $ hisat2-build -p 8 –ss chr1.ss –exon chr1.exon chr1.fa chr1_idx
     ```

2. Στοίχιση των single-end reads

   ```
   $ hisat2 -p 12 –dta -x chr1_data/indexes/chr1_idx -U SRR4098426.fastq -S SRR4098426_chr1.sam
   ```

3. Στοίχιση των paired-end reads

   ```
   $ hisat2 -p 12 –dta -x chr1_data/indexes/chr1_idx –SRR4098026_1.fastq -2 SRR4098026_2.fastq -S SRR4098026_chr1.sam
   ```

4. Το output της στοίχισης είναι ένας πίνακας με στοιχεία, όπως το μέγεθος της βιβλιοθήκης, ο αριθμός των reads που χαρτογραφήθηκαν μία ή περισσότερες φορές και το ποσοστό της συνολικής στοίχισης. Επίσης, προκύπτει ένα SAM αρχείο της στοίχισης.

2.4. STAR

Ο αλγόριθμος STAR (**S**pliced **T**ranscripts **A**lignment to a **R**eference), σχεδιάστηκε ειδικά για να αντιμετωπίσει πολλές από τις προκλήσεις που υπάρχουν κατά την διαδικασία χαρτογράφησης δεδομένων RNA-seq. Σε αντίθεση με τις προηγούμενες προσεγγίσεις, το STAR σχεδιάστηκε για να στοιχίζει τις μη-συνεχόμενες αλληλουχίες απευθείας με το γονιδίωμα αναφοράς. Ο αλγόριθμος STAR αποτελείται από δύο βασικά βήματα: την αναζήτηση των seeds και την ομαδοποίησης/συρραφής/ βαθμολόγησης.

Η κεντρική ιδέα του πρώτου βήματος του STAR είναι η διαδοχική αναζήτηση ενός μέγιστου προθέματος που στοιχίζεται (Maximum Mappable Prefix - MMP). Η συγκεκριμένη ιδέα είναι παρόμοια με αυτή του Maximal Exact Match που χρησιμοποιείται από τα μεγάλης κλίμακας εργαλεία στοίχισης γονιδιώματος Mummer και MAUVE. Η διαδοχική εφαρμογή της αναζήτησης MMP μόνο στα μη

αντιστοιχισμένα τμήματα της ακολουθίας κάνει τον αλγόριθμο STAR εξαιρετικά γρήγορο. Η αναζήτηση MMP στον αλγόριθμο STAR υλοποιείται μέσω μη συμπιεσμένων πίνακα καταλήξεων (Suffix Arrays - SAs). Το πλεονέκτημα είναι πως για κάθε MMP, η αναζήτηση SA μπορεί να βρει όλες τις ευδιάκριτες ακριβείς γονιδιωματικές αντιστοιχίσεις με μικρή υπολογιστική επιβάρυνση, γεγονός που διευκολύνει την ακριβή στοίχιση των ακολουθιών που απεικονίζονται σε πολλαπλούς γονιδιωματικούς τόπους (multimapping reads). Επίσης, η συγκεκριμένη αναζήτηση, εμφανίζει πλεονέκτημα, όσον αφορά στη ταχύτητα, έναντι των συμπιεσμένων SAs που εφαρμόζονται σε πολλά δημοφιλή προγράμματα στοίχισης μικρών ακολουθιών. Ωστόσο, αυτό το πλεονέκτημα ταχύτητας αντισταθμίζεται από την αυξημένη χρήση μνήμης.

Το δεύτερο βήμα του αλγορίθμου περιλαμβάνει τη δημιουργία στοιχίσεων ολόκληρης της ακολουθίας, συρράπτοντας τα seeds που στοιχίστηκαν με το γονιδίωμα στην πρώτη φάση. Σημαντικά, τα seeds από τις paired-end ακολουθίες RNA-seq συγκεντρώνονται και συρράπτονται ταυτόχρονα, με κάθε ακολουθία paired-end να αντιπροσωπεύεται ως μια ενιαία αλληλουχία, επιτρέποντας ένα πιθανά γονιδιωματικό χάσματα ή επικάλυψη μεταξύ των εσωτερικών άκρων. Η συρραφή καθοδηγείται από ένα τοπικό σχήμα βαθμολόγησης της στοίχισης, με καθορισμένες από τον χρήστη, ποινές για τις αντιστοιχίσεις, τις ανα-ντιστοιχίες, τις εισαγωγές, τις διαγραφές και τα κενά, επιτρέποντας μια ποσοτική αξιολόγηση των ιδιοτήτων και των βαθμών στοίχισης[26].

2.4.1. Χρήση του STAR

1. Εγκατάσταση του STAR σε συμβατό λογισμικό

```
$ tar -zxvfSTAR-STAR_version.tar.gz
$ cd STAR-STAR_version/bin/chooseTourOS
$ sudo chmod 755 STAR
$ sudo cp STAR/usr/local/bin
```

2. Δημιουργία indexes του γονιδιώματος αναφοράς. Χρήση γονιδιωμάτων από το NCBI ή το ENSEMBL, χρησιμοποιώντας .gtf ή .gff αρχεία.

```
$ STAR –runThreadN 8 –runMode genomeGEnerate – genomeDir
pathToGenome/ --genomeFastaFiles pathToGenome/hg38.fa –sjdbGTFfile
pathToGTF/gencode.v21.annotation.gtf –sjdbOverhang 100
```

3. Στοίχιση paired-end reads

```
$ STAR –runThreadN 8 –runMode alignReads –genomeDir
pathToGenomeIndex/ --readFilesIn pathToRNAseqFile/myRNAseqReads_
mate1.fastq pathToRNAseqFile/myRNAseqReads_mate2.fastq
```

4. Output αρχείο στοιχίσεων σε bam αρχείο

```
$ STAR –runThreadN 8 –runMode alignReads –genomeDir
pathToGenomeIndex/ --readFilesIn pathToRNAseqFile/myRNAseqReads_
mate1.fastq pathToRNAseqFile/myRNAseqReads_mate2.fastq –
outSAMtype BAM SortedByCoordinate
```

3. Βιοπληροφορική ανάλυση των miRNAs

Είναι γνωστό ότι τα miRNAs είναι σημαντικά στη ρύθμιση του κυτταρικού πολλαπλασιασμού, της απόπτωσης, της φλεγμονής, με αποτέλεσμα να επηρεάζουν πολλές την ανάπτυξη και την πορεία πολλών παθήσεων.

Τα τελευταία χρόνια με την ανάπτυξη της βιοτεχνολογίας και των προγραμμάτων βιοπληροφορικής, τα miRNA με γνωστές αλληλουχίες και τα προφίλ έκφρασής τους μπορούν να δημιουργηθούν από τα δεδομένα αλληλούχισης επόμενης γενιάς. Επίσης, επιτρέπεται η αναγνώριση νέων miRNA και η εξερεύνηση των παραλλαγών της αλληλουχίας υπό διαφορετικές συνθήκες. Επί του παρόντος, υπάρχουν πολλά διαθέσιμα εργαλεία για την ανάλυση των miRNA, όπως προ-επεξεργασία της ακολουθίας, χαρτογράφηση και ανάλυση της διαφορικής τους έκφρασης.

Μια γενική μεθοδολογία βιοπληροφορικής ανάλυσης των αλληλουχιών των miRNAs περιλαμβάνει τα εξής βήματα:

1. Προ-επεξεργασία των δεδομένων, που συμπεριλαμβάνει το φιλτράρισμα των δεδομένων με βάση την ποιότητα των ακολουθιών, καθώς και αφαίρεση του adapter στο 3΄ άκρο (adapter trimming)
2. Χαρτογράφηση και σχολιασμό (mapping and annotation)
3. Ανάλυση των χαρακτηριστικών της αλληλουχίας, συμπεριλαμβανομένης της πρόβλεψης νέων miRNA

4. Ανάλυση διαφορικής έκφρασης για τα γνωστά και για τα νέα miRNAs
5. Ανάλυση της λειτουργικότητας, με βάση την πρόβλεψη του miRNA-στόχου[26].

Βασική προϋπόθεση για την μετέπειτα ανάλυση όλων των βιολογικών δεδομένων, άρα και των miRNAs, είναι η σωστή στοίχιση (alignment) στο γονιδίωμα αναφοράς.

Τέλος, υπάρχουν πολλές βάσεις δεδομένων αλληλουχιών miRNAs, καθώς και των γνωστών στόχων τους που εξυπηρετούν στην ανάλυσή τους, όπως οι miRBase, deepBase, miRTarBase, StarBase, StarScan, Cupid, TargetScan, Diana-microT, κ.α.

ΣΤΟΧΟΣ

Όπως αναφέρθηκε και προηγουμένως, πολλές μελέτες έχουν αναδείξει την σημαντικότητα των miRNAs στη δημιουργία και ανάπτυξη ασθενειών, καθώς επίσης

και στη χρήση τους ως ισχυροί βιοδείκτες πρόγνωσης και διάγνωσης. Γιο το λόγο αυτό η ανάλυσή τους, αλλά και η πρόβλεψη νέων τέτοιων ακολουθιών αποτελεί κομβικό σημείο στην εξέλιξη της επιστήμης. Σημαντικό βήμα στην βιοπληροφορική ανάλυση τέτοιων δεδομένων, αποτελεί η στοίχιση και η στοίχιση των ακολουθιών, που προέκυψε από αλληλούχιση επόμενης γενιάς, στο γονιδίωμα αναφοράς.

Σκοπός λοιπόν της παρούσας εργασίας είναι ο έλεγχος τεσσάρων αλγορίθμων στοίχισης των Bowtie, Bowtie 2, HISAT 2 και STAR, γνωστών miRNAs, με στόχο την ποσοτικοποίησή τους. Οι συγκεκριμένοι αλγόριθμοι θα συγκριθούν και θα χαρακτηριστούν ως προς την αποτελεσματικότητά τους, την ευκολία χρήσης τους και τον χρόνο που απαιτούν προκείμενου να ολοκληρώσουν την διαδικασία.

ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

1. Συνθετικά δεδομένα

Ένα σύνολο δεδομένων short read δημιουργήθηκε με χρήση τυχαίας δειγματοληψίας με μια τεχνική αντιστροφής Monte Carlo. Το annotation έγινε στα Ensembl v85, GtRNAdb 2.0 και miRBase v21. Επίσης, χρησιμοποιήθηκαν τρεις τυχαία επιλεγμένες βιβλιοθήκες sRNA-Seq από το GEO. Τα δείγματα μετά από αφαίρεση του τμήματος στο 3΄ άκρο (adapters), χρησιμοποιώντας το Cutadapt, στοιχίστηκαν έναντι του ανθρώπινου γονιδιώματος αναφοράς GRCh38, χρησιμοποιώντας το Bowtie και επιτρέποντας μόνο 1 αναντιστοιχία (mismatch). Με βάση τα μοναδικά στοιχισμένα reads που παρατηρήθηκαν στα πραγματικά δεδομένα, δημιουργήθηκαν Πιθανότητες Μαζικής Λειτουργίες (Probability Mass Function - PMF) για κάθε βιολογικό τύπο, περιγράφοντας τις θέσεις έναρξης ανάγνωσης. Δημιουργήθηκαν εννέα διαφορετικά PMF για τους ακόλουθους τύπους RNA: miRNA, tRNA, mt-tRNA, rRNA, mt-rRNA, snRNA, snoRNA, lincRNA και επεξεργασμένο μετάγραφο. Ομοίως, εκτιμήθηκαν οι μονο-νουκλεοτιδικοί πολυμορφισμού (SNPs), καθώς και το μέγεθος των reads για κάθε τύπο sRNA με βάση τα PMF των UAR[26]. Τα συγκεκριμένα δεδομένα βρίσκονται στον παρακάτω σύνδεσμο (https://github.com/ jehandzlik/Manatee/tree/simulatedData).

2. Γενικές πληροφορίες ανάλυσης και προετοιμασία λογισμικού

Για την παρούσα πτυχιακή εργασία χρησιμοποιήθηκε το λειτουργικό Ubuntu 22.04 lts. Ακολούθως έγινε εγκατάσταση του ANACONDA (Continuum analytics), το οποίο δημιουργεί ένα περιβάλλον εργασίας, στο οποίο γίνεται δυνατή η απομόνωση και ο διαχωρισμός διαφορετικών project μεταξύ τους που χρησιμοποιούν διαφορετικές εκδόσεις λογισμικών.

Στη συνέχεια μέσω του Conda, που είναι ένα λογισμικό ανοιχτού κώδικα διαχείρισης συστήματος, έγινε εγκατάσταση του channel bioconda χρησιμοποιώντας τον παρακάτω κώδικα:

$ conda config --add channels defaults

$ conda config --add channels bioconda

$ conda config --add channels conda-forge

$ conda install -c conda-forge biopython

$ conda install -c bioconda ucsc-fasomerecords

Τέλος ακολούθησε η εγκατάσταση των προγραμμάτων και των aligners που θα χρησιμοποιηθούν στην παρούσα εργασία, εισάγοντας τον παρακάτω κώδικα:

$ conda activate

$ conda install -c bioconda star

$ conda install -c bioconda bowtie2

$ conda install -c bioconda bowtie

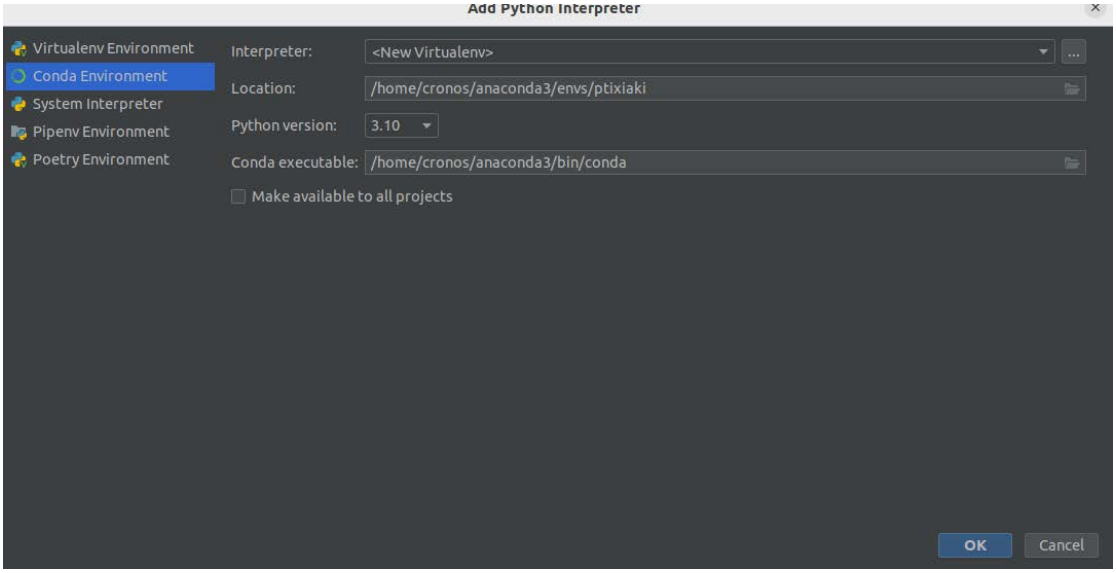
$ conda install -c bioconda hisat2

$ conda install -c bioconda samtools

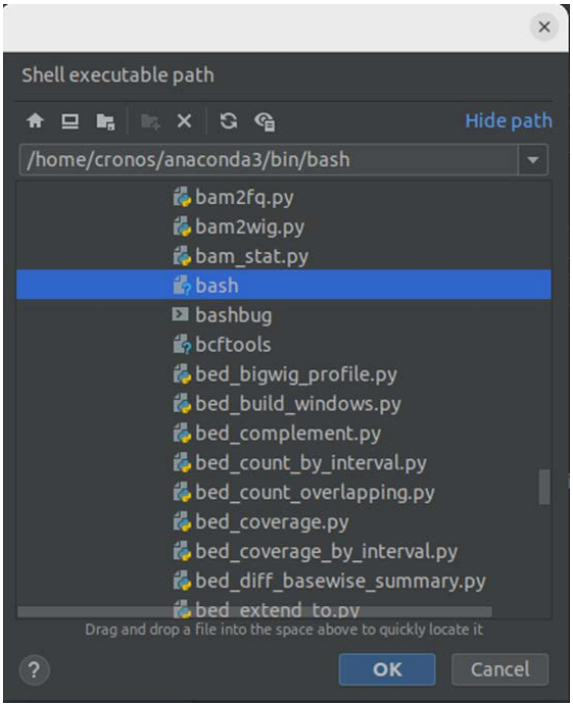
Αναλυτικά, οι εκδόσεις των προγραμμάτων που χρησιμοποιήθηκαν είναι οι εξής:

- Bowtie v1.0.0
- Bowtie 2 v2.2.5
- HISAT2 v2.2.1
- STAR v2.7.10a
- samtools v1.3.1

Για την συγγραφή του κώδικα χρησιμοποιήθηκε το PyCharm (JetBrains), στο οποίο οι παράμετροι του περιβάλλοντος εργασίας ρυθμίστηκαν έτσι ώστε να είναι συμβατό με το ANACONDA, όπως φαίνεται παρακάτω.



Εικόνα 8. Δημιουργία περιβάλλοντος για το ANACONDA και επιλογή python Interpreter.



Εικόνα 9. Επιλογή path για το ANACONDA για τη χρήση συγκεκριμένων εκδόσεων λογισμικών (bowtie, bowtie 2, hisat2 και STAR).

3. Προετοιμασία δεδομένων

Από την βάση δεδομένων miRBase (https://www.mirbase.org/ftp.shtml) έγινε λήψη δυο αρχείων, των *hairpin.fa* (11-Mar-2018) και *mature.fa* (11-Mar-2018). Σε συνδυασμό με αυτά τα δύο αρχεία, χρησιμοποιήθηκαν τα sRNA_MC.fa και sRNA_MC.fa.smallRNA_counts.tsv, τα οποία αποτελούν το reference genome από τα simulated data του manatee και τις πληροφορίες των reads (πόσα και ποια), αντίστοιχα.

Αρχικά λοιπόν, έγινε χρήση της γλώσσα python για την δημιουργία ενός script, το οποίο θα περιλαμβάνει το pipeline της ανάλυσης. Αρχικά δημιουργήθηκε μια συνάρτηση, η make_global_variable(), για τον ορισμό global μεταβλητών, οι οποίες θα χρησιμοποιηθούν μέσα στο script.

```
#make my variables global
make_global_variables()
```

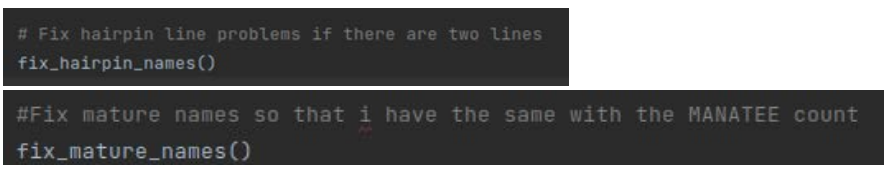
Εικόνα 10. Κλήση της συνάρτησης.

Οι global μεταβλητές που χρησιμοποιήθηκαν, όπως φαίνεται στην **Εικόνα 11** είναι οι εξής:

- index_name: αφορά τα hairpin ή τα mature
- simulated_genome: path από τα simulated δεδομένα
- human_mir_path: path από το αρχείο hairpin και mature
- aligner: τα ονόματα από τα προγράμματα στοίχισης
- aligner_build: οι ρυθμίσεις για την δημιουργία των indexes
- align: οι ρυθμίσεις για την χρήση των aligners

Εικόνα 11. Δημιουργία global μεταβλητών.

Στη συνέχεια, γίνεται κλήση της συνάρτησης fix_hairpin_names() και fix_mature_names(), προκειμένου να ενωθεί η αλληλουχία του read σε μια σειρά και να διορθωθούν τα ονόματα των αλληλουχιών για την ευκολότερη σύγκριση.



```
# Fix hairpin line problems if there are two lines
fix_hairpin_names()
```

```
#Fix mature names so that i have the same with the MANATEE count
fix_mature_names()
```

Εικόνα 12. Κλήση των συναρτήσεων για τα αντίστοιχα αρχεία.

```
def fix_hairpin_names():
    # Fix hairpin line problems if there are two lines
    f = open("new-hairpin.fa", "w")
    handle = open("hairpin.fa", "r")
    for record in SeqIO.parse(handle, "fasta"):
        x = ">" + record.id + "\n" + record.seq
        f.writelines(x + "\n")
    handle.close()
    f.close()
def fix_mature_names():
    subprocess.call(["cut -d '_' -f1 mature.fa > new.mature.fa"], shell=True)
```

Εικόνα 13. Script των συναρτήσεων.

Στο πλαίσιο της παρούσας εργασίας γίνεται χρήση μόνο των ανθρώπινων miRNAs. Για το λόγο αυτό από τα αρχεία *hairpin.fa* και *mature.fa,* δημιουργήθηκαν δυο νέα αρχεία που περιλαμβάνουν, αντίστοιχα, τα miRNAs του ανθρώπου, δηλαδή τις εγγραφές που ξεκινάμε με *>has* (*Homo Sapiens*). Τα βήματα που ακολουθήθηκαν φαίνονται παρακάτω:

1. Δημιουργία αρχείου μόνο με τους τίτλους των miRNAs, χωρίς να περιλαμβάνεται το σύμβολο «>» στην αρχή του κάθε τίτλου.

   $ grep hsa hairpin.fa |tr -d '>' > listFile

2. Χρήση του παρακάτω script για εξαγωγή αρχείων FASTA από ένα multiFASTA αρχείο, με βάση τους τίτλους.

   $ ./faSomeRecords hairpin.fa listFile hsa-hairpin.fa

3. Επανάληψη των βημάτων 1 και 2 για την δημιουργία ενός αρχείου για τα mature miRNAs.

   $ grep hsa mature.fa |tr -d '>' > listFile

   $ ./faSomeRecords mature.fa listFile hsa-mature.fa

4. Δημιουργία indexes για τον κάθε αλγόριθμο ξεχωριστά

```
#Make Human files choose the id that start with >hsa
make_hsa_human_files()
```

Εικόνα 14. Κλήση της συνάρτησης.

```
def make_hsa_human_files():
    subprocess.call([" grep hsa new-hairpin.fa |tr -d '>' > listFile-hairpin"], shell=True)
    subprocess.call([" ./faSomeRecords new-hairpin.fa listFile-hairpin hsa-hairpin.fa"], shell=True)
    subprocess.call([" grep hsa new.mature.fa |tr -d '>' > listFile-mature"], shell=True)
    subprocess.call([" ./faSomeRecords new.mature.fa listFile-mature hsa-mature.fa"], shell=True)
```

Εικόνα 15. Επιλογή των ανθρώπινων εγγραφών.

4. Δημιουργία indexes

Μετά την ολοκλήρωση και προετοιμασία των δεδομένων έρχεται με την σειρά του η δημιουργία των indexes.

```
 # min(14, log2(GenomeLength)/2 - 1)# star aligner
genomeSAindexNbases = 5
counter = 0
# MAKE INDEX WITH BOWTIE, BOWTIE2 AND HISAT2
for i in range(3):
    for x in range(2):
        make_index(x,i)
for y in range(2):
    make_index_with_star(y,genomeSAindexNbases)
```

Εικόνα 16. Διαδοχική κλήση της συνάρτησης και για τους τέσσερις aligners.

Δημιουργήθηκαν οχτώ indexes, τέσσερα για τα hairpin και τέσσερα για τα mature, τα οποία αφορούν τους τέσσερις διαφορετικούς aligners. Το αποτέλεσμα της κάθε εντολής αποθηκεύτηκε ως ένα ξεχωριστό αρχείο για την μετέπειτα ανάλυση και σύγκριση των αποτελεσμάτων.

Bowtie

$ time bowtie-build hsa-hairpin.fa hairpin > index-bowtie-make-hairpin

$ time bowtie-build hsa-mature.fa mature > index-bowtie-make-mature

Bowtie 2

$ time bowtie2-build-s ./hsa-hairpin.fa hairpin > index-bowtie2-make-hairpin

$ time bowtie2-build-s ./hsa-mature.fa mature >index-bowtie2-make-mature

HISAT 2

$ time hisat2-build-s hsa-hairpin.fa hairpin > index-hisat2-make-hairpin

$ time hisat2-build-s hsa-mature.fa mature > index-hisat2-make-mature

STAR

$ time STAR --runMode genomeGenerate --genomeDir ref/ --genomeFastaFiles ./hairpin/hsa-hairpin.fa --runThreadN 7 --genomeSAindexNbases 5 > STAR-hairpin-building-index

$ time STAR --runMode genomeGenerate --genomeDir ./mature/ref/ --genomeFastaFiles ./mature/hsa-mature.fa --runThreadN 7 --genomeSAindexNbases 3 > STAR-mature-building-index

Το genomeSAindexNbases προκύπτει από τον τύπο

min(14, log2(GenomeLength)/2 - 1)# star aligner

Το output κατά την διαδικασία δημιουργίας των indexes, αποθηκεύεται σε ένα αρχείο για την αξιοποίηση της μεταβλητή time και την συνεπακόλουθη σύγκριση των χρόνων.

```
#hisat2 bowtie bowtie2
def make_index(x,i):
    print("\n Building index of " + index_name[x] + " with " + aligner[
        i] + " and as reference genome : " + simulated_genome + "\n")

    subprocess.call(["time  %s-build %s %s > index-%s-make-%s"
                     % (aligner[i], human_mir_path[x], index_name[x], aligner_build[i], index_name[x])]
                    , shell=True)

def make_index_with_star(x,genomeSAindexNbases):
    # MAKE INDEX WITH STAR
    # hairpin has bigger read length than mature that is the reason we need ó for mature.
        print("\n Building index with STAR as reference genome : " + simulated_genome + "\n")
        subprocess.call(
            ["time STAR --runMode genomeGenerate --genomeDir ref/%s --genomeFastaFiles %s --runThreadN 7 "
             "--genomeSAindexNbases %s > index-STAR-%s-building" % (
                 index_name[x], human_mir_path[x], genomeSAindexNbases, index_name[x])]
            , shell=True)
        genomeSAindexNbases -= 2  # smaller genome
```

Εικόνα 17. Script για την δημιουργία των indexes.

5. Στοίχιση με την χρήση των aligners

Το επόμενο βήμα του pipeline, μετά την δημιουργία των indexes, είναι η στοίχιση των δεδομένων στα simulated data.

Οι παράμετροι που χρησιμοποιήθηκαν για τα προγράμματα στοίχισης είναι οι εξής:

Bowtie

$ time bowtie -p 8 -a -n 1 --best --strand -l 8 -x hairpin -f ./sRNA_MC.fa -S hsa-hairpin-bowtie.sam

$ time bowtie -p 8 -a -n 1 --best --strand -l 8 -x mature -f ./sRNA_MC.fa -S hsa-mature-bowtie.sam

Bowtie 2

$ time bowtie2-align-s -p 8 --end-to-end -a -L 8 -f hairpin ./sRNA_MC.fa -S hsa-hairpin-bowtie2.sam

$ time bowtie2-align-s -p 8 --end-to-end -a -L 8 -f mature ./sRNA_MC.fa -S hsa-mature-bowtie2.sam

HISAT 2

$ time hisat2-align-s -a --very-sensitive -x hairpin -f ./sRNA_MC.fa -S hsa-hairpin-hisat2.sam

$ time hisat2-align-s -a --very-sensitive -x mature -f ./sRNA_MC.fa -S hsa-mature-hisat2.sam

STAR

$ time STAR --runMode alignReads --genomeDir ./hairpin/ref/ --outSAMtype SAM --readFilesIn sRNA_MC.fa

$ time STAR --runMode alignReads --genomeDir ./mature/ref/ --outSAMtype SAM --readFilesIn sRNA_MC.fa

```
sam_counter = 0
counter = 0
# MAKE ALIGNMENT !
for i in range(3):
    for x in range(2):
        #align with Hisat2 and bowtie and bowtie2
        align_with_hisat2_and_Bowtie_Bowtie2(sam_counter)
        #align with star
        sam_counter += 1
        if counter < 2:# counter to aligh only 2 times star
            align_with_star(sam_counter)
            counter += 1
            sam_counter +=1
```

Εικόνα 19. Κλήση δύο συναρτήσεων για την στοίχιση των δεδομένων. Η πρώτη αφορά τα προγράμματα bowtie, bowtie 2 kai hisat2,ενώ η δεύτερη αφορά το STAR.

```
def align_with_hisat2_and_Bowtie_Bowtie2(sam_counter):
    print("\n Align with " + aligner[i] + " and using as index : " + index_name[x] + "\n")
    if i == 1:

        subprocess.call(["time %s -f %s  %s  -S hsa-%s-%s-%s.sam "
                         % (align[i], index_name[x], simulated_genome, index_name[x], aligner[i], sam_counter)]
                        , shell=True)
    else:
        subprocess.call(["time %s -x %s  -f %s  -S hsa-%s-%s-%s.sam "
                         % (align[i], index_name[x], simulated_genome, index_name[x], aligner[i], sam_counter)]
                        , shell=True)
```

Εικόνα 18. Συνάρτηση για τα bowtie, bwtie 2 και hisat2.

```
def align_with_star(sam_counter):
    print("\n Align with STAR as reference genome : " + simulated_genome + "\n")
    subprocess.call(["time STAR --runMode alignReads --genomeDir ref/%s --outSAMtype SAM --readFilesIn %s"
                     % (index_name[x], simulated_genome)]
                    , shell=True)
    # star produce some files as output so we change the names
    subprocess.call(["mv Aligned.out.sam hsa-%s-star-%s.sam" % (index_name[x], sam_counter)], shell=True)
    subprocess.call(["mv Log.final.out star-%s-Log.final.out" % (index_name[x])], shell=True)
    subprocess.call(["mv Log.out star-%s-Log.out" % (index_name[x])], shell=True)
    subprocess.call(["mv Log.progress.out star-%s-Log.progress.out" % (index_name[x])], shell=True)
```

Εικόνα 20. Συνάρτηση στοίχισης με το STAR και μετονομασία αρχείων output.

6. Χρήση των Samtools

Με το πέρας της στοίχισης, οι aligners έχουν δημιουργήσει αρχεία της μορφής .sam.

```
# def samtools pipeline
    Samtools_pipeline()
```

Εικόνα 21. Κλήση της συνάρτησης για την χρήση των samtools.

Αρχικά, με τη χρήση των samtools, τα αρχεία αυτά μετατράπηκαν σε binary μορφή .bam με την εντολή

$ samtools view -Sb hsa-hairpin-bowtie2-0.sam > hsa-hairpin-bowtie2-0.bam

Στη συνέχεια γίνεται χρήση της λειτουργίας sort από τα samtools

$ samtools sort hsa-hairpin-bowtie2-0.bam -o hsa-hairpin-bowtie2-0.sorted

Ακολουθεί η κατασκευή ενός index για την δημιουργία ενός αρχείου .bai με την εντολή

$ samtools index hsa-hairpin-bowtie2-0.sorted

Για να γίνει τελικά η χρήση της εντολής idxstats, που παρέχει πληροφορίες σχετικά με τα δεδομένα, όπως την αλληλουχία, το μήκος της και τον αριθμό των εμφανίσεων της.

$ samtools idxstats hsa-hairpin-bowtie2-0.sorted > idxstats-hsa-hairpin-bowtie-4.sorted.txt

Ομοίως, οι παραπάνω εντολές ακολουθήθηκαν τόσο και για τα mature miRNAs, όσο και για τους υπόλοιπους aligners.

```
# samtools pipeline MAKE BAM file , SORT BAM file , MAKE_INDEX , FIND IDXSTATS
def Samtools_pipeline():
    x = os.listdir('./samfiles')  # names of the samfile
    for i in range(len(x)):
        mod_string = x[i][:-4]  # dont want .sam last letters
        subprocess.call(["samtools view -S -b ./samfiles/%s.sam > ./samfiles/%s.bam"
                         % (mod_string, mod_string)],shell=True)
        subprocess.call(["samtools sort  ./samfiles/%s.bam -o ./samfiles/%s.sorted"
                         % (mod_string, mod_string)],shell=True)
        subprocess.call(["samtools index ./samfiles/%s.sorted ./samfiles/%s.sorted.bai"
                         % (mod_string, mod_string)],shell=True)
        subprocess.call(["mv new-idxstats* ./samfiles/"], shell=True)#if there is any from previous run
        subprocess.call(["samtools idxstats ./samfiles/%s.sorted  > ./samfiles/idxstats-%s.txt " %
                         (mod_string, mod_string)],shell=True)
```

Εικόνα 22. Χρήση της σουίτας των samtools.

7. Ποσοτικοποίηση (quantification) των στοιχισμένων miRNAs

Στην παρούσα εργασία μελετώνται δυο μορφές των miRNAs, το πρόδρομο pre-miRNA (hairpin) και το ώριμο miRNA, οι οποίες μεταξύ των άλλων διαφέρουν και στο μέγεθος των αλληλουχιών του. Το πρόδρομο pre-miRNA αποτελείται από 60-120 nt, ενώ το ώριμο miRNA αποτελείται από 17-24 nt.

Για το λόγο αυτό, πριν την ποσοτικοποίηση των reads που στοιχίστηκαν, πρέπει να γίνει επιλογή των ακολουθιών με βάση το μέγεθος τους. Έτσι στην περίπτωση των hairpin miRNAs, η επιλογή έγινε από 60-120 νουκλεοτίδια, ενώ για τα ώριμα miRNA η επιλογή έγινε από 16-25 νουκλεοτίδια, επιτρέποντας με αυτό τον τρόπο το περιθώριο μέχρι και 2 αναντιστοιχίες.

```
# def to choose the right reads
    idx_sam = os.listdir('./samfiles/')
    chose_valid_reads(idx_sam)
```

Εικόνα 23. Κλήση της συνάρτησης.

```
def chose_valid_reads(idx_sam):
    for i in range(len(idx_sam)):
        mod_string = idx_sam[i][0]
        if mod_string == 'i': #chose those that start with i

            file1 = open('./samfiles/%s' % idx_sam[i], 'r')
            lines = file1.readlines()

            # i want to take the length of the reads and if it is the range of hairpin 60-120 and only
            # the reads that did align with count >0
            # or the range or mature 15-27

            if idx_sam[i][13] == 'h':
                length = os.system(
                    "awk '{ if ($2>=60 && $2<=120 && $3>1) {print $1 , $2 ,$3}'} ./samfiles/%s > ./idxstats/new-%s" %
                    (idx_sam[i], idx_sam[i]))
            else:
                length = os.system(
                    "awk '{ if ($2>14 && $2<28 && $3>1) {print $1 , $2 ,$3 }'} ./samfiles/%s > ./idxstats/new-%s" %
                    (idx_sam[i], idx_sam[i]))
```

Εικόνα 24. Script για την επιλογή κατάλληλων reads.

Στη συνέχεια, και αφού επιλέχθηκαν μόνο τα κατάλληλα read ανά κατηγορία, αποθηκεύτηκαν σε ένα ξεχωριστό αρχείο, το οποίο χρησιμοποιήθηκε για την ποσοτικοποίησή τους.

```
#def find_percentage for stats
    idxstats = os.listdir('./idxstats/')
    for i in idxstats:
        print(i)
        find_percentage(i)
```

Εικόνα 26. Κλήση της συνάρτησης για την ποσοτικοποίηση των reads.

```
def find_percentage (file):
    make_lower_all_sRNA_id()#make all lower
    file1 = open("./idxstats/%s" % file, "r")
    lines = file1.readlines()
    file1_data = []
    file1_headers = []
    file1_counts = []
#read every file and make lists 1col = gene_id , 2col, length_fasta , 3rd= counts
    #and from manatee 2 columns 1st = gene_id and 2nd
    for line in lines:
        #make new lists to manipulate the data beter
        file1_data.append(line.strip().split(" ")[1])
        file1_headers.append(line.split(" ")[0])
        file1_counts.append(line.strip().split(" ")[2])
    file2 = open("./sRNA-counts.txt", "r")
    lines2 = file2.readlines()
    file2_counts = []
    file2_headers = []
    for x in lines2:
        file2_headers.append(x.strip().split("\t")[0])
        file2_counts.append(x.strip().split("\t")[1])
    #make counts intiger so we can do mathematical operations
    list_of_file2 = list(map(int, file2_counts))
    list_of_file1 = list(map(int, file1_counts))
    sum = 0
    counter = 0# to find all different samples

    for i in range(len(file1_headers)):
        for j in range(len(file2_headers)):
            if str(file1_headers[i].lower()) == str(file2_headers[j]) :
                    x = list_of_file1[i]
                    y = list_of_file2[j]
                    percentage = x / y
                    temp = sum
                    sum = temp + percentage *100 # percentage of different read that found
                    counter += 1
                    print(str(file1_headers[i]) + "    " + str(file2_headers[j]))
                    break
            else:
                continue

    percentage = sum / counter
    print("the persentage is %.2f" % round(percentage, 2))
    print(counter)
    print("\n")
    file2.close()
    file1.close()
```

Εικόνα 25. Script για την εύρεση των ποσοστών επιτυχούς στοίχισης.

8. Δημιουργία γραφημάτων

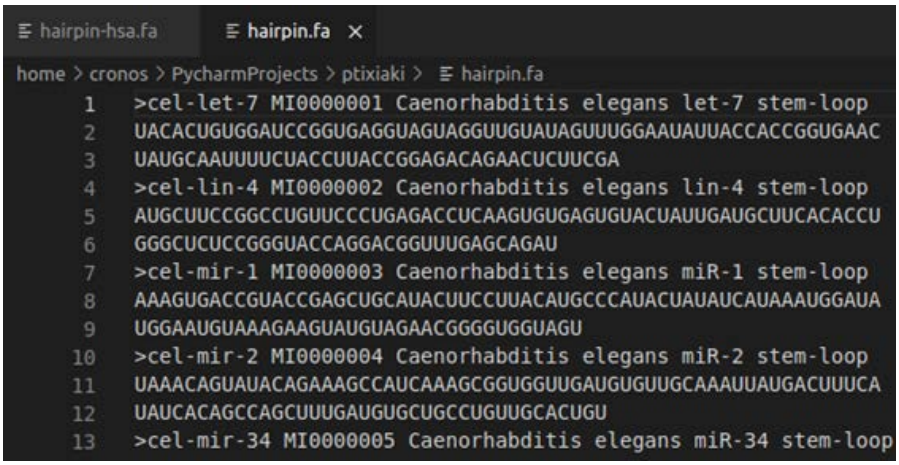
Το αποτέλεσμα της ποσοτικοποίησης είναι ένας πίνακας με το όνομα και τον αριθμό του read που έχει στοιχιστεί, καθώς και το ποσοστό της επιτυχούς στοίχισης. Τα στοιχεία αυτά συλλέχθηκαν και επεξεργάστηκαν με το Graphpad Prism για τον σχεδιασμό τον γραφημάτων.

ΑΠΟΤΕΛΕΣΜΑΤΑ

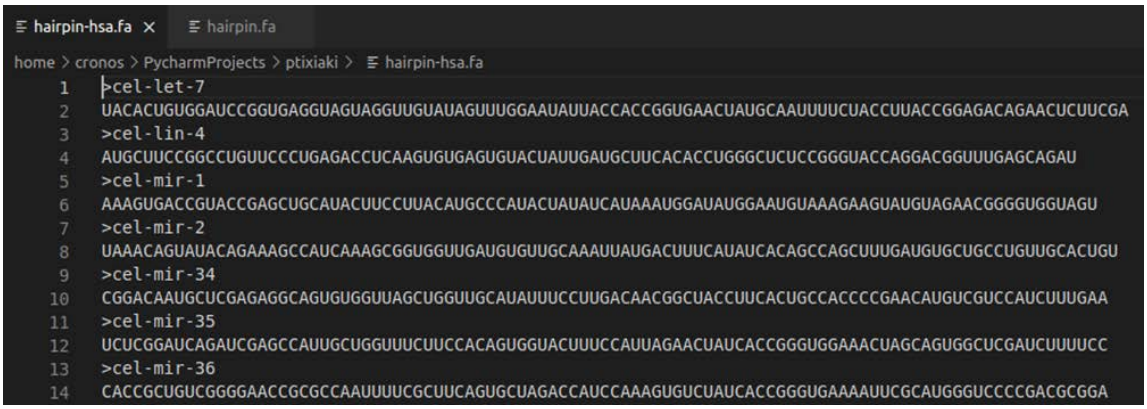
1. Διόρθωση δεδομένων

Βασικό βήμα στην επεξεργασία τον δεδομένων, ήταν η διόρθωση των αρχείων, τα οποία εμφάνιζαν την αλληλουχία κομμένη σε δυο σειρές, γεγονός που επιβάρυνε την αποτελεσματικότητα της στοίχισης με τους τέσσερις aligners.

Για το λόγο αυτό χρησιμοποιήθηκε η εντολή *fix_hairpin_name()* και *fix_mature_name()* για τα πρόδρομα και ώριμα miRNAs αντίστοιχα. Αυτό είχε σαν αποτέλεσμα την επιδιόρθωση όλων τον εγγραφών και στα δυο αρχεία. Ενδεικτικά στις παρακάτω φωτογραφίες παρουσιάζονται το «πριν» και το «μετά» κάποιων ενδεικτικών αλληλουχιών.



Εικόνα 27. Παρουσίαση των δεδομένων των αλληλουχιών, όπου φαίνεται η αλληλουχία να είναι χωρισμένη σε περισσότερες από μια σειρές.



Εικόνα 28. Το αποτέλεσμα της επεξεργασίας των αλληλουχιών, όπου φαίνεται η αλληλουχία να εμφανίζεται σε μια μόνο σειρά, κάτω από το όνομα.

2. Στοίχιση δεδομένων

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα της στοίχισης, τόσο των hairpin, όσο και των mature miRNAs, χρησιμοποιώντας τα Bowtie, Bowtie 2, HISAT2 και STAR.

2.1. Στοίχιση με τη χρήση του Bowtie

Η χρήση του Bowtie στην περίπτωση των hairpin miRNAs, οδήγησε στην στοίχιση περίπου 11.252 ακολουθιών, τουλάχιστον σε μια θέση, αριθμός που αντιστοιχεί περίπου ~1,5%. Ομοίως, χαμηλό παρουσιάστηκε το ποσοστό, όσον αφορά στα mature miRNAs, όπου αυτό άγγιξε το 0,34%, με την πλειοψηφία των ακολουθιών να μην καταφέρνουν να στοιχιστούν πουθενά. Αναλυτικά τα ποσοστά και για τις δύο κατηγορίες φαίνονται στις παρακάτω εικόνες (**Εικ.29 και Εικ.30**).

```
Align with bowtie and using as index : hairpin

# reads processed: 778072
# reads with at least one reported alignment: 11252 (1.45%)
# reads that failed to align: 766820 (98.55%)
Reported 16965 alignments to 1 output stream(s)
30.67user 1.25system 0:04.15elapsed 767%CPU (0avgtext+0avgdata 14972maxresident)k
0inputs+157000outputs (0major+3049minor)pagefaults 0swaps
```

Εικόνα 29. Αποτελέσματα στοίχισης χρησιμοποιώντας το Bowtie στα hairpin miRNAs.

```
Align with bowtie and using as index : mature

# reads processed: 778072
# reads with at least one reported alignment: 2646 (0.34%)
# reads that failed to align: 775426 (99.66%)
Reported 2940 alignments to 1 output stream(s)
25.64user 1.30system 0:03.55elapsed 758%CPU (0avgtext+0avgdata 15112maxresident)k
0inputs+155152outputs (0major+3044minor)pagefaults 0swaps
```

Εικόνα 30. Αποτελέσματα στοίχισης χρησιμοποιώντας το Bowtie στα mature miRNAs.

2.2. Στοίχιση με τη χρήση του Bowtie 2

Χρησιμοποιώντας το Bowtie 2, το 1,74% των hairpin miRNA-ακολουθιών στοιχίστηκε σε μια ακριβώς θέση και το 0,32% των ακολουθιών σε περισσότερες από μια. Όσον αφορά στην περίπτωση των mature miRNAs, το 0,83% στοιχίστηκε ακριβώς μια φορά. Και πάλι η πλειοψηφία των ακολουθιών δεν στοιχίστηκε πουθενά (**Εικ.31 και Εικ.32**).

```
Align with bowtie2 and using as index : hairpin

11.93user 1.55system 0:15.44elapsed 87%CPU (0avgtext+0avgdata 1367056maxresident)k
0inputs+1362696outputs (0major+341340minor)pagefaults 0swaps
778072 reads; of these:
  778072 (100.00%) were unpaired; of these:
    762048 (97.94%) aligned 0 times
    13554 (1.74%) aligned exactly 1 time
    2470 (0.32%) aligned >1 times
2.06% overall alignment rate
56.83user 1.00system 0:07.75elapsed 746%CPU (0avgtext+0avgdata 40456maxresident)k
0inputs+159832outputs (0major+9129minor)pagefaults 0swaps
```

Εικόνα 32. Αποτελέσματα στοίχισης χρησιμοποιώντας το Bowtie 2 στα hairpin miRNAs.

```
Align with bowtie2 and using as index : mature

778072 reads; of these:
  778072 (100.00%) were unpaired; of these:
    771532 (99.16%) aligned 0 times
    6441 (0.83%) aligned exactly 1 time
    99 (0.01%) aligned >1 times
0.84% overall alignment rate
23.37user 0.42system 0:03.26elapsed 729%CPU (0avgtext+0avgdata 37376maxresident)k
0inputs+157360outputs (0major+8355minor)pagefaults 0swaps
```

Εικόνα 31. Αποτελέσματα στοίχισης χρησιμοποιώντας το Bowtie 2 στα mature miRNAs.

2.3. Στοίχιση με τη χρήση του HISAT2

Η στοίχιση με τη χρήση του HISAT2 οδήγησε σε υψηλότερα ποσοστά ακολουθιών που είχαν ακριβώς μια αντιστοίχιση. Στην περίπτωση των hairpin miRNAs, το ποσοστό αυτό άγγιξε το 2,89%, ενώ τα mature miRNAs με ακριβώς μια θέση αντιστοίχισης είναι 2,99% (**Εικ.33 και Εικ.34**).

```
 Align with hisat2 and using as index : hairpin

778072 reads; of these:
  778072 (100.00%) were unpaired; of these:
    752505 (96.71%) aligned 0 times
    22488 (2.89%) aligned exactly 1 time
    3079 (0.40%) aligned >1 times
3.29% overall alignment rate
6.23user 3.03system 0:01.23elapsed 748%CPU (0avgtext+0avgdata 38484maxresident)k
0inputs+162256outputs (0major+8558minor)pagefaults 0swaps
```

Εικόνα 34. Αποτελέσματα στοίχισης χρησιμοποιώντας το HISAT2 στα hairpin miRNAs.

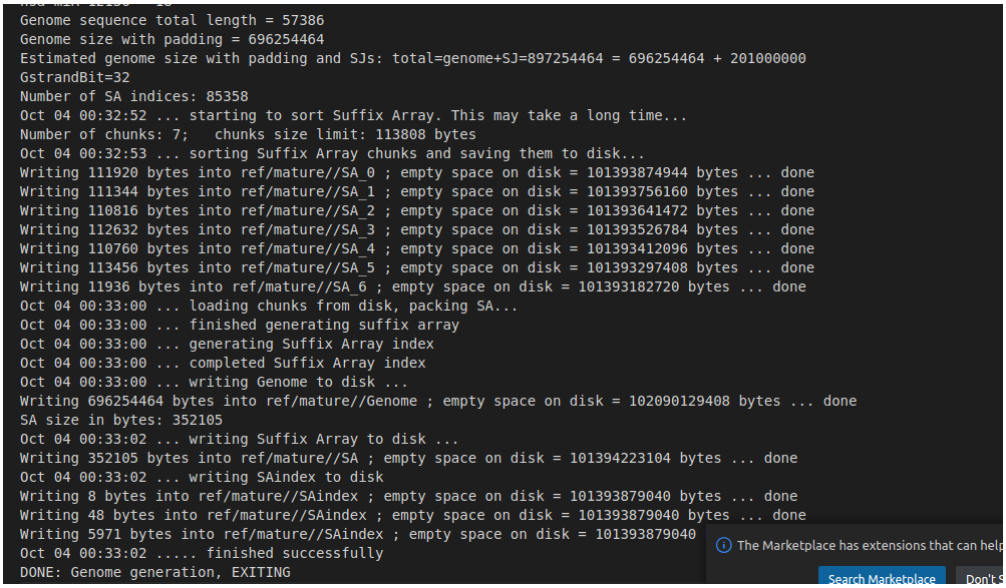
```
 Align with hisat2 and using as index : mature

778072 reads; of these:
  778072 (100.00%) were unpaired; of these:
    753057 (96.79%) aligned 0 times
    23253 (2.99%) aligned exactly 1 time
    1762 (0.23%) aligned >1 times
3.21% overall alignment rate
6.25user 2.78system 0:01.22elapsed 740%CPU (0avgtext+0avgdata 46240maxresident)k
0inputs+161720outputs (1major+10497minor)pagefaults 0swaps
```

Εικόνα 33. Αποτελέσματα στοίχισης χρησιμοποιώντας το Bowtie 2 στα mature miRNAs.

2.4. Στοίχιση με τη χρήση του STAR

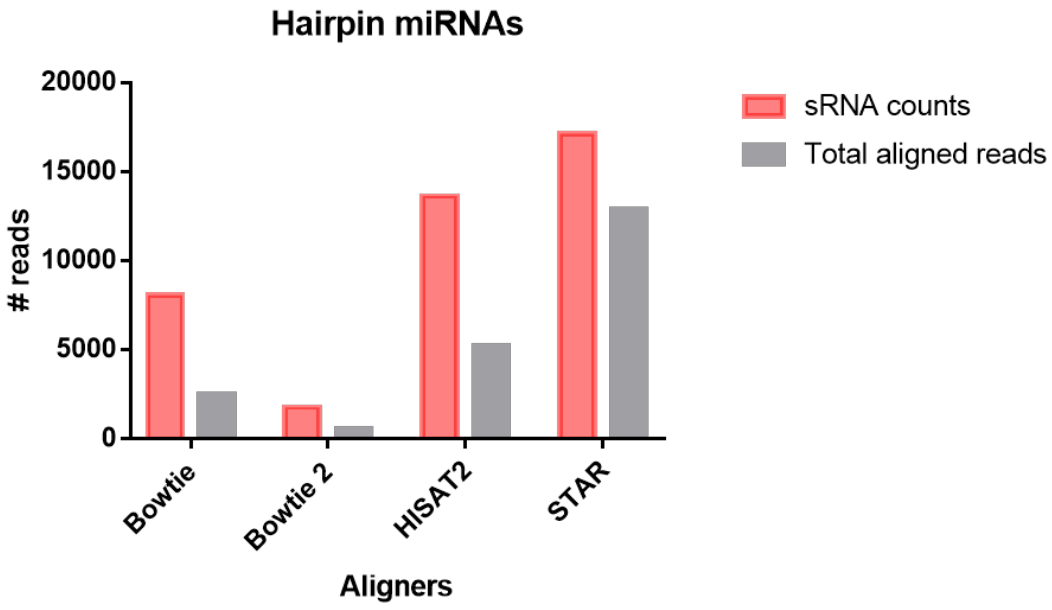
Το STAR παράγει τέσσερα αρχεία μετά την στοίχιση τα οποία περιέχουν τις εξής πληροφορίες: ένα περιλαμβάνει τα ονόματα από τα reads που στοιχίστηκαν, ένα με τα μήκη των ακολουθιών που στοιχίστηκαν, ένα που συνδυάζει τα μήκη και τα ονόματα των reads που στοιχίστηκαν και ένα με πληροφορίες για τις ρυθμίσεις των παραμέτρων και τα συγκεντρωτικά τα δεδομένα. Αντιπροσωπευτικά φαίνεται παρακάτω ένα output αρχείο όπως προέκυψε για τα mature miRNAs (**Εικόνα 34**).



Εικόνα 35. Output αρχείο, όπως προέκυψε μετά την στοίχιση με το STAR για τα mature reads.

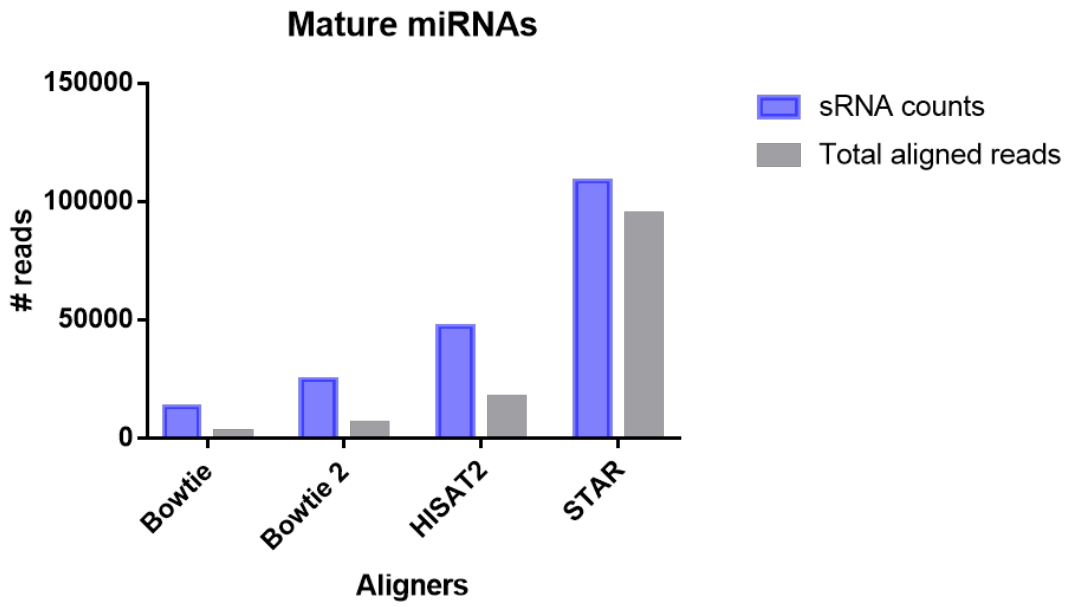
3. Αριθμός των reads που στοιχίστηκαν

Ο αριθμός των ακολουθιών που στοιχίστηκαν επιτυχώς ανάμεσα σε miRNAs με ίδιο ID_name, ήταν χαμηλότερος από τον πραγματικό αριθμό που είναι γνωστό ότι υπάρχουν στα simulated data. Στην περίπτωση των hairpin miRNAs, το Bowtie ανέφερε 2465 αντιστοιχίες έναντι των 8059 που υπάρχουν στην πραγματικότητα.



Εικόνα 36. Σύγκριση του αριθμού των hairpin miRNA reads που στοιχίστηκαν με τους τέσσερις διαφορετικούς αλγόριθμους, αλλά και με τον πραγματικό αριθμό των reads που υπάρχουν.

Ομοίως, το Bowtie 2 ανέφερε 574 αντιστοιχίες έναντι 1766, το HISAT2 5238 αντιστοιχίες έναντι 13619 και το STAR 12851 έναντι 17120 ακολουθίες, αντίστοιχα



Εικόνα 37. Σύγκριση του αριθμού των mature miRNA reads που στοιχίστηκαν με τους τέσσερις διαφορετικούς αλγόριθμους, αλλά και με τον πραγματικό αριθμό των reads που υπάρχουν.

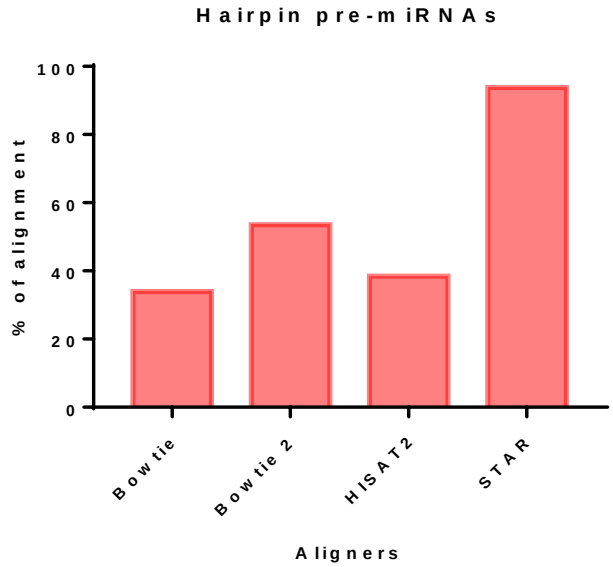
(**Εικόνα 36**).

Ομοίως, στην περίπτωση των mature miRNAs, το Bowtie ανέφερε 2432 αντιστοιχίες έναντι των 13164 που υπάρχουν στην πραγματικότητα. Ομοίως, το Bowtie 2 ανέφερε 6359 αντιστοιχίες έναντι 24872, το HISAT2 17066 αντιστοιχίες έναντι 47156 και το STAR 94742 έναντι 108530 ακολουθίες, αντίστοιχα (**Εικόνα 37**).

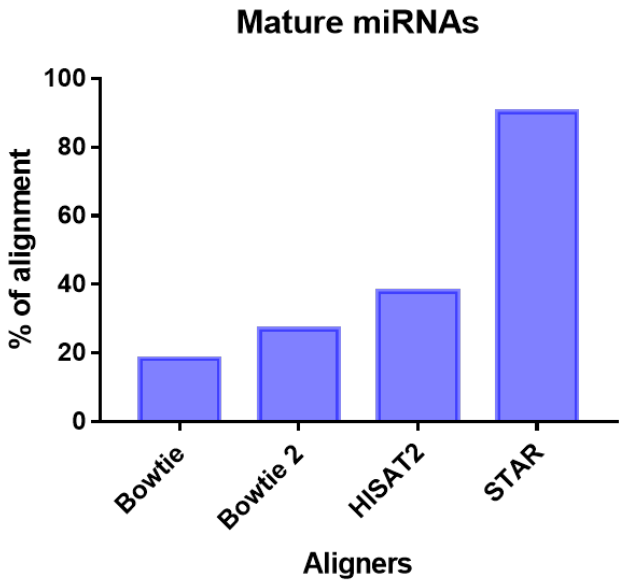
4. Ποσοστό επιτυχούς στοίχισης

Όπως αναφέρθηκε και προηγουμένως, η ποσοτικοποίηση των reads, είχε ως αποτέλεσμα έναν πίνακα, όπου αναφέρεται η ονομασία και ο αριθμός των miRNAs που στοιχίστηκαν, καθώς και το ποσοστό επιτυχίας της στοίχισης με τον εκάστοτε aligner.

Συγκρίνοντας τα ποσοστά των προγραμμάτων στοιχίσεων, παρατηρήθηκαν τα μεγαλύτερα ποσοστά στοίχισης με το STAR, τόσο στα hairpin (**Εικόνα 40**), όσο και στα mature reads (**Εικόνα 41**).



Εικόνα 38. Σύγκριση ποσοστών στοίχισης για κάθε aligner στα hairpin reads.

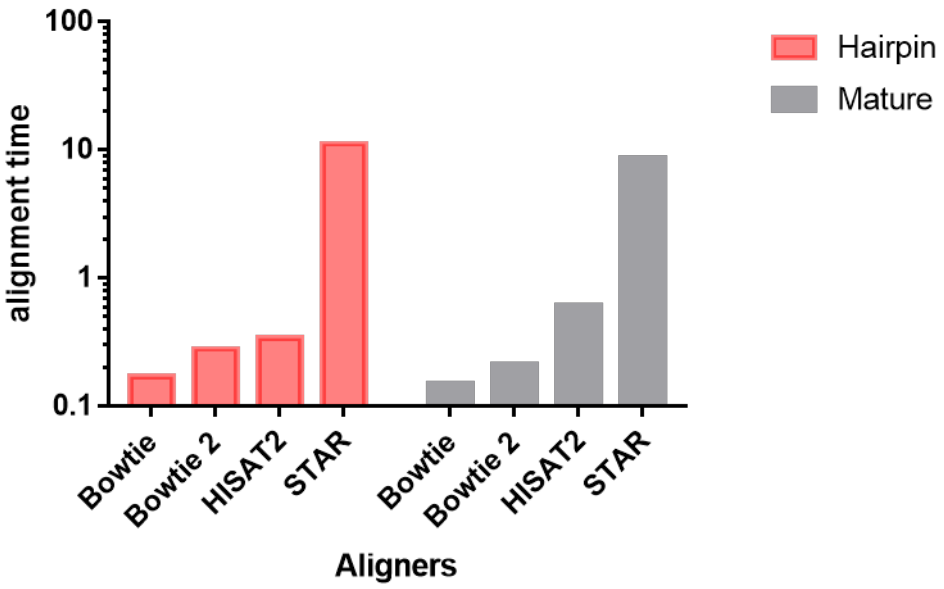


Εικόνα 39. Σύγκριση ποσοστών στοίχισης για κάθε aligner στα mature reads.

5. Χρόνοι για ολοκλήρωση της στοίχισης

Τέλος, τα συγκεκριμένα προγράμματα στοίχισης συγκρίθηκαν και ως προς τον χρόνο που χρειάζονται να καταναλώσουν για να ολοκληρώσουν την δημιουργία των indexes. Παρατηρήθηκε πως γενικά ο χρόνος που απαιτήθηκε στην περίπτωση των hairpin ήταν μεγαλύτερος από τον αντίστοιχο για τα mature miRNAs, με εξαίρεση το HISAT2, όπου οι χρόνοι εμφανίστηκαν αντεστραμμένοι. Παρατηρούμε ότι το Star χρειάζεται πάρα πολύ χρόνο για την δημιουργία του index καθώς και για την στοίχιση.

	Bowtie	Bowtie-2	Star	Hisat-2	
Hairpin	30,67	56,83	290,37	6,23	align
Mature	25,64	23,37	153,59	6,25	
Hairpin	0,16	0,19	10,84	0,42	
Mature	0,12	0,15	7,93	0,44	build



Εικόνα 40. Απαιτούμενοι χρόνοι των aligners για την επιτυχή στοίχιση των hairpin και mature miRNAs.

Εικόνα 41. Output των Bowtie και Bowtie 2 aligners για τα hairpin και mature miRNAs σχετικά με τον χρόνο ολοκλήρωσης της στοίχισης.

Hairpin

```
Total time for call to driver() for forward index: 00:00:00
0.34user 0.19system 0:00.59elapsed 89%CPU (0avgtext+0avgdata 94956maxresident)k
12inputs+39832outputs (0major+44687minor)pagefaults 0swaps
```

Mature

```
Total time for call to driver() for forward index: 00:00:01
0.61user 0.21system 0:00.89elapsed 93%CPU (0avgtext+0avgdata 94952maxresident)k
0inputs+51744outputs (0major+48298minor)pagefaults 0swaps
```

```
 Building index of hairpin with bowtie2 and as reference genome : ./sRNA_MC.fa

Building a SMALL index
0.28user 0.19system 0:00.51elapsed 93%CPU (0avgtext+0avgdata 94996maxresident)k
0inputs+17024outputs (0major+50941minor)pagefaults 0swaps

 Building index of mature with bowtie2 and as reference genome : ./sRNA_MC.fa

Building a SMALL index
0.21user 0.16system 0:00.39elapsed 96%CPU (0avgtext+0avgdata 94964maxresident)k
0inputs+16912outputs (0major+50875minor)pagefaults 0swaps

 Building index of hairpin with bowtie and as reference genome : ./sRNA_MC.fa

0.17user 0.11system 0:00.29elapsed 100%CPU (0avgtext+0avgdata 95176maxresident)k
0inputs+16928outputs (0major+47617minor)pagefaults 0swaps

 Building index of mature with bowtie and as reference genome : ./sRNA_MC.fa

0.15user 0.09system 0:00.24elapsed 99%CPU (0avgtext+0avgdata 94952maxresident)k
0inputs+16896outputs (0major+47546minor)pagefaults 0swaps
```

Εικόνα 42. Output του HISAT2 για τα hairpin και mature miRNAs σχετικά με τον χρόνο ολοκλήρωσης της στοίχισης.

```
 Building index with STAR as reference genome : ./sRNA_MC.fa

8.73user 1.06system 0:10.79elapsed 90%CPU (0avgtext+0avgdata 989320maxresident)k
0inputs+987560outputs (0major+247749minor)pagefaults 0swaps

 Building index with STAR as reference genome : ./sRNA_MC.fa

11.22user 1.60system 0:18.59elapsed 68%CPU (0avgtext+0avgdata 1367200maxresident)k
0inputs+1362664outputs (0major+341391minor)pagefaults 0swaps
```

Εικόνα 43. Output του STAR aligners για τα hairpin και mature miRNAs σχετικά με τον χρόνο ολοκλήρωσης της στοίχισης.

ΣΥΖΗΤΗΣΗ

Τα τελευταία χρόνια έχει βρεθεί πως τα miRNAs αποτελούν έναν ιδανικό βιοδείκτη, καθώς μπορούν να εντοπιστούν και να απομονωθούν από εύκολα προσβάσιμους ιστούς, ενώ παράλληλα έχει δειχθεί σε αρκετές μελέτες, πως συμμετέχουν στη διαφοροποίηση των καρκινικών σταδίων και άλλων ασθενειών, καθώς και χρησιμοποιούνται και για τη μέτρηση της ανταπόκρισης στη θεραπεία. Η ανάπτυξη τόσων πολλών εργαλείων στοίχισης των αλληλουχιών, τα οποία καθορίζουν που στοιχίζονται τα μικρά τμήματα σε μεγαλύτερα γονιδιώματα αναφοράς ή μετάγραφα, αποτελούν ουσιαστικό βήμα καθοριστικής σημασίας για αναλύσεις ολόκληρου το γονιδιώματος και του μεταγραφώματος. Αυτές οι αναλύσεις βρίσκουν εφαρμογές από τη γεωργία και τη κτηνοτροφία, μέχρι την υγεία του ανθρώπου.

Το Bowtie αρχικά πραγματοποιούσε σύντομες αναγνώσεις χωρίς κενά (περίπου 35 ζεύγη βάσεων). Η λογική πίσω από την χωρίς-κενά-στοίχιση ήταν ότι τα μικρά reads θα πρέπει να έχουν μια μοναδική θέση στο γονιδίωμα και η χωρίς-κενά λειτουργία, επιτρέπει στον aligner να λειτουργεί πολύ πιο γρήγορα. Ωστόσο, τα μεγαλύτερα σε μέγεθος reads, οδηγούν σε χαμηλά ποσοστά στοίχισης, ενώ δεν επιτρέπει την στοίχιση των RNA αλληλουχιών σε ένα γονιδίωμα (καθώς οποιοδήποτε εσώνιο στο γονίδιο θα προσθέσει αυτόματα κενά στην στοίχιση). Επιπλέον, το Bowtie εκτελεί στοιχίσεις μόνο στην end-to-end λειτουργία, απαιτώντας με αυτόν τον τρόπο μια ολόκληρη ανάγνωση να στοιχιστεί από το ένα άκρο στο άλλο, προκειμένου η στοίχιση να ολοκληρωθεί και

να αναφερθεί. Αυτό συχνά οδηγεί σε χαμηλότερης ποιότητας στοίχιση, εάν τα άκρα των reads δεν έχουν κοπεί σωστά για έλεγχο ποιότητας.

Για το λόγο αυτό το 2011, κυκλοφόρησε το Bowtie2 το οποίο βασίστηκε στην χωρίς-κενά-στοίχιση του Bowtie, λαμβάνοντας υπόψη την προσθήκη κενών. Το Bowtie2 χρησιμοποιεί ένα FM-Index για να δημιουργήσει ένα index με το γονιδίωμα-αναφοράς και δημιουργεί μια σειρά ερωτημάτων για να βρει πολλαπλές στοιχίσεις χωρίς κενό, οι οποίες στη συνέχεια επεκτείνονται, με τροπο παρόμοιο με το Bowtie. Επιπλέον, το Bowtie2 επιτρέπει τον συνδυασμό πολλών seeds, προσθέτοντας με αυτόν τον τρόπο κενά για τη δημιουργία μεγαλύτερων στοιχίσεων. Το Bowtie 2 έχει επίσης τη δυνατότητα, τόσο end-to-end, όσο και τοπικής λειτουργίας για τη δημιουργία στοιχίσεων. Ανεξάρτητα από τη λειτουργία που εκτελείται, το Bowtie2 δημιουργεί πολλαπλές στοιχίσεις για κάθε read, αλλά αναφέρει μόνο τη μοναδική καλύτερη στοίχιση ανά read. Ένα μειονέκτημα στο Bowtie2 είναι ότι σχεδιάστηκε βασικά για στοιχίσεις DNA ακολουθιών, σε ένα γονιδίωμα αναφοράς και δεν επιτρέπει την προσθήκη ενός αρχείου με στοιχεία της μεταγραφής, προκειμένου να βοηθήσει στην αντιστοίχιση των RNA ακολουθιών.

Το 2014 κυκλοφόρησε το HISAT με σκοπό να στοιχίζει ακολουθίες RNA από RNA-seq δεδομένα. Το HISAT δημιουργεί indexes του γονιδιώματος αναφοράς με παρόμοιο τρόπο με το FM index του Bowtie. Ένα χρόνο αργότερα κυκλοφόρησε το HISAT2, το οποίο για πρώτη φορά χρησιμοποίησε ένα graph-based FM index για να δημιουργήσει το index του γονιδιώματος. Το HISAT2 αντιμετωπίζει επαναλαμβανόμενες αλληλουχίες, συνδυάζοντας τες στο γονιδίωμα αναφοράς σε μια αλληλουχία. Με αυτόν τον τρόπο μειώνεται ο αριθμός των στοιχίσεων που αναφέρονται στο τέλος και αντ΄αυτού εξάγει μόνο μια στοίχιση ανά read για κάθε στοίχιση σε αυτές τις περιοχές, παρά μια στοίχιση για κάθε επαναλαμβανόμενη περιοχή[27].

Τέλος, το STAR σχεδιάστηκε αρχικά για την στοίχιση δεδομένων από RNA-seq με την πρόθεση να διαχειριστεί τα spliced RNA μετάγραφα. Το STAR δημιουργεί indexes του γονιδιώματος αναφοράς μέσω μη συμπιεσμένων suffix arrays, τα οποία χρησιμοποιούν ένα μεγάλο μέρος μνήμης, ενώ είναι και χρονοβόρο[28].

Όσον αφορά στην στοίχιση των μικρών μη-κωδικών μορίων RNA (miRNAs), η αποτελεσματικότητα των aligners ελέγχθηκε, τόσο σε πρόδρομα miRNAs, τόσο και σε ώριμα. Έχει υπολογιστεί πως τα πρόδρομα miRNAs έχουν μήκος από 60-120 νουκλεοτίδια, ενώ τα ώριμα miRNAs έχουν μήκος από 17-22 νουκλεοτίδια. Μετά την

επιλογή των κατάλληλων παραμέτρων και επιλογής μόνο του επιθυμητού μήκους αλληλουχίας, βρέθηκε πως το STAR επιτυγχάνει την καλύτερη στοίχιση, τόσο των mature, όσο και των hairpin ακολουθιών με επιβάρυνση του χρόνο που απαιτείται για την ολοκλήρωση της στοίχισης, καθώς απαιτεί τον περισσότερο χρόνο.

Όσον αφορά τα mature miRNAs, που είναι και μικρότερα σε μήκος αλληλουχίας, το αμέσως επόμενο βέλτιστο πρόγραμμα στοίχισης (μετά το STAR) είναι το HISAT2 που επιτυγχάνει καλύτερο ποσοστό στοίχισης συγκριτικά με το Bowtie και το Bowtie 2. Ωστόσο, ο χρόνος ολοκλήρωσης της στοίχισης τους είναι σχεδόν τριπλάσιος από τον αντίστοιχο που επιτυγχάνει ο Bowtie και το Bowtie 2.

Κατά την στοίχιση των hairpin miRNAs, παρατηρήθηκε πως το καλύτερο πρόγραμμα για την στοίχισή τους, μετά το STAR, είναι το Bowtie 2, καθώς επιτυγχάνει το καλύτερο ποσοστό στοίχισης στα δεδομένα αναφοράς και σχετικά καλό χρόνο ολοκλήρωσης. Είναι γενικά γνωστό πως το Bowtie 2 λειτουργεί καλύτερα για μεγάλου μήκους reads.

Ομοίως, έχει παρατηρηθεί και από άλλες ερευνητικές ομάδες, πως την καλύτερη απόδοση στην στοίχιση RNA-seq δεδομένων έχουν τα HISAT2 και STAR, με το HISAT2 ήταν ~ 3 φορές ταχύτερο από τον επόμενο ταχύτερο aligner στο χρόνο εκτέλεσης, κάτι που θεωρείται δεύτερος σημαντικότερος παράγοντας στις περισσότερες στοιχίσεις[29].

Συνοψίζοντας, στην παρούσα εργασία, το βέλτιστο εργαλείο για την στοίχιση των miRNAs, παρατηρήθηκε πως ήταν το STAR. Ωστόσο, κάθε ερευνητική ομάδα πρέπει με βάση τα ερωτήματά της, πρέπει να κρίνει ποιο εργαλείο είναι καταλληλότερο για το επιστημονικό της ερώτημα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol. (Lausanne).* 2018;9(AUG):402.
2. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*. 2009;136(2):215–233.
3. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*. 2004;116(2):281–297.
4. Plotnikova O, Baranova A, Skoblov M. Comprehensive Analysis of Human microRNA–mRNA Interactome. *Front. Genet.* 2019;10:933.
5. Selbach M, Schwanhäusser B, Thierfelder N, et al. Widespread changes in protein synthesis induced by microRNAs. *Nature*. 2008;455(7209):58–63.
6. Uhlmann S, Mannsperger H, Zhang JD, et al. Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. *Mol. Syst. Biol.* 2012;8:.
7. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 2009;19(1):92–105.
8. Kumar S, Reddy PH. Are circulating microRNAs peripheral biomarkers for Alzheimer's disease? *Biochim. Biophys. Acta*. 2016;1862(9):1617–1627.
9. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*. 1993;75(5):843–854.
10. Reinhart BJ, Slack FJ, Basson M, et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*. 2000;403(6772):901–

906.

11. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. *Science (80-.).* 2001;294(5543):853–858.
12. Calin GA, Dumitru CD, Shimizu M, 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):15524–15529.
13. Mambo E, Szafranska-Schwarzbach AE, Latham G, et al. microRNA Biomarkers as Potential Diagnostic Markers for Cancer. *Genomic Biomarkers Pharm. Dev. Adv. Pers. Heal. Care*. 2013;95–126.
14. Takahashi M, Han S -p., Scherer LJ, Yoon S, Rossi JJ. Current Progress and Future Prospects in Nucleic Acid Based Therapeutics. *Compr. Med. Chem. III*. 2017;280–313.
15. Miyoshi K, Miyoshi T, Siomi H. Many ways to generate microRNA-like small RNAs: Non-canonical pathways for microRNA production. *Mol. Genet. Genomics*. 2010;284(2):95–103.
16. Xu W, Lucas AS, Wang Z, Liu Y. Identifying microRNA targets in different gene regions. *BMC Bioinformatics*. 2014;15 Suppl 7(Suppl 7):
17. Ardekani AM, Naeini MM. The Role of MicroRNAs in Human Diseases. *Avicenna J. Med. Biotechnol.* 2010;2(4):161.
18. Tüfekci KU, Öner MG, Meuwissen RLJ, Genç Ş. The role of microRNAs in human diseases. *Methods Mol. Biol.* 2014;1107:33–50.
19. Condrat CE, Thompson DC, Barbu MG, et al. miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. *Cells*. 2020;9(2):.
20. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):1–10.
21. Bowtie 2: fast and sensitive read alignment.
22. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. 2019;35(3):421–432.
23. Sirén J, Välimäki N, Mäkinen V. Indexing graphs for path queries with

applications in genome research. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 2014;11(2):375–388.

24. HISAT2.
25. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*. 2015;12(4):357–360.
26. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15.
27. Musich R. A Recent (2020) Comparative Analysis of Genome Aligners Shows HISAT2 and BWA are Among the Best Tools. *Theses*. 2020;
28. Chipster.
29. Musich R, Cadle-Davidson L, Osier M V. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Front. Plant Sci.* 2021;12:692.