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 των μικρών μορίων RNA από δεδομένα αλληλούχησης επόμενης γενιάς με τη χρήση του spipeRNA

Joanna A. Handzlik

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

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση των μικρών RNA Ανάλυση των μικρών μορίων RNA από δεδομένα αλληλούχησης επόμενης γενιάς με τη χρήση του spipeRNA

Joanna A. Handzlik
A.M.: ΠΙΒ0140

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

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Άρτεμις Χατζηγεωργίου**, Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Υ/Η, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας
Γεώργιος Μ. Σπύρου, Επικεφαλής της ομάδας Βιοπληροφορικής, Κυπριακό Ινστιτούτου Νευρολογίας και Γενετικής
Ιωάννης Βλάχος, Ερευνητής Ιατρικής Σχολής του Πανεπιστημίου Χάρβαρντ

Δεκέμβριος 2016

ΠΕΡΙΛΗΨΗ

Μία από τις σημαντικότερες τεχνολογικές εξελίξεις στο χώρο της βιοτεχνολογίας τα τελευταία χρόνια, αφορά την τεχνολογία της αλληλούχησης επόμενης γενιάς (Next Generation Sequencing ή NGS). Ενώ η αποκρυπτογράφηση του πρώτου ανθρωπίνου γονιδιώματος (3 δις βάσεις ανά απλοειδές γονιδίωμα) χρειάστηκε περίπου 15 χρόνια με υλικό κόστος και μόνο, ανερχόμενο στα 10 δις δολάρια Αμερικής, σήμερα η αλληλούχηση ολόκληρου του ανθρωπίνου γονιδιώματος μπορεί να παραχθεί από μία μονάχα συσκευή μέσα σε λίγες μέρες, με κόστος που δεν ξεπερνά τα 1.000 δολάρια Αμερικής. Η τεχνολογία αυτή άνοιξε το δρόμο για πολλές συναρπαστικές εφαρμογές, όπως είναι η *de novo* αλληλούχηση, η ανάλυση του μεταγραφώματος (RNA-Seq) και του μεθυλώματος (methyl-seq), ο προσδιορισμός των θέσεων πρόσδεσης των μεταγραφικών παραγόντων (ChIP-seq), η ανίχνευση των γονιδιακών μεταλλάξεων υπεύθυνων για ασθένειες και πολλές άλλες ακόμη εφαρμογές.

Ο σκοπός της παρούσας διπλωματικής ήταν ο σχεδιασμός και η υλοποίηση ενός υπολογιστικού εργαλείου αφιερωμένου στην ανάλυση των small RNA-Seq δεδομένων, τα οποία αποτελούν ένα κομμάτι της γενικής ανάλυσης του μεταγραφώματος (RNA-Seq). Η ανάλυση αυτή αποσκοπεί στην ποσοτικοποίηση των εκφραζόμενων μικρών μορίων RNA και την εύρεση καινούργιων, μη σχολιασμένων περιοχών έκφρασης, σε βιολογικά δείγματα ποικίλλης προέλευσης.

Ο αλγόριθμος που σχεδιάστηκε, ονομάστηκε “spipeRNA» και απαντά σε πολλές ανοιχτές προκλήσεις: ποσοτικοποιεί όλα τα μικρά RNAs και όχι μόνο τα miRNAs, επιλύει το πρόβλημα των εγγραφών με πολλαπλές θέσεις ευθυγράμμισης πάνω στο γονιδίωμα και χειρίζεται κατάλληλα τις εγγραφές χωρίς υπάρχοντα σχολιασμό.

Στην εργασία παρουσιάζονται και τα αποτελέσματα εκτέλεσης του εργαλείου για την ανάλυση προσομοιωμένων δεδομένων και 8 small RNA-Seq συνόλων δεδομένων, τα οποία περιλαμβάνουν καρκινικά και υγιή δείγματα πνεύμονα και παγκρέατος. Το spipeRNA συγκρίθηκε με ένα δημοφιλές εργαλείο ανάλυσης των miRNAs επιδεικνύοντας υψηλότερη ακρίβεια σε προσομοιωμένα και πραγματικά δεδομένα.

Το εργαλείο spipeRNA, βασίζεται σε μία αξιόπιστη, ευέλικτη και πλήρως αυτοματοποιημένη ροή εργασιών, χρήσιμη για τη γρήγορη και υψηλής απόδοσης ανάλυση των small RNA-Seq δεδομένων από αλληλουχητές επόμενης γενιάς.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μικρά μη-κωδικά RNAs, αλληλούχηση επόμενης γενιάς, NGS, ευθυγράμμιση πάνω στο γονιδίωμα, σχολιασμός γονιδιωματικών περιοχών, microRNA, snoRNA, snRNA, tRNA, rRNA, siRNA

ABSTRACT

Some of the most important technological developments in biotechnology in recent years, are summarized under the term “Next Generation Sequencing (NGS)”. While the sequencing of the first human genome (3 gigabases per haploid genome) took about 15 years and roughly 100 million of US dollars in material costs only, today the raw sequencing data for a complete human genome (100 gigabases at 30x coverage) can be produced by a single machine within a few days and for just 1.000 US dollars. This technological quantum leap has paved the way for numerous exciting applications such as de novo sequencing, transcriptome (RNA-Seq) and methylome (methyl-Seq) analysis, the determination of transcription factor binding sites (ChIP-Seq), the detection of disease-causing mutations, and many others.

The purpose of this study was the design and implementation of the computational tool, dedicated to the analysis of small RNA-Seq data, which form a part of the overall analysis of trascriptome (RNA-Seq). This analysis aims to quantify the expressed small RNA molecules and to detect new non-annotated expression regions in various biological samples.

The implemented algorithm was called "spipeRNA» and tries to overcome many open challenges: it quantifies all types of small RNAs, not only the miRNAs, solves the problem of multi-mapped reads and appropriately handles the reads without existing annotation.

This study presents the results obtained by applying this tool to analyze simulated data and 8 small RNA-Seq datasets, which include tumor/healthy lung and pancreas samples. The comparison between the spipeRNA and very popular tool for miRNAs analysis, showed, that in some cases, the spipeRNA may produce more precise and accurate output.

The spipeRNA is an integrated data analysis pipeline, based on a reliable, flexible and fully automated workflow, useful for fast and efficient analysis of small RNA-Seq data produced by next-generation sequencers.

.

SUBJECT AREA: Bioinformatics

KEYWORDS: small non-coding RNAs, Next Generation Sequencing, NGS, reads alignment, annotation of genomic regions, microRNA, snoRNA, snRNA, tRNA, rRNA, siRNA

Στη μητέρα μου

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα της διπλωματικής εργασίας μου, καθηγήτρια του Πανεπιστημίου Θεσσαλίας κυρία Χατζηγεωργίου για την ευκαιρία που μου έδωσε να ασχοληθώ με το παρόν θέμα. Ευχαριστώ ιδιαίτερα τον ερευνητή της Ιατρικής Σχολής του Χάρβαρντ, κύριο Ιωάννη Βλάχο, για τις πολύτιμες συμβουλές και την καθοδήγησή του, χάρη στα οποία έφερα εις πέρας τη συγγραφή της εργασίας. Επίσης, είμαι ευγνώμων στον Επικεφαλής της ομάδας Βιοπληροφορικής του Κυπριακού Ινστιτούτου Νευρολογίας και Γενετικής, τον κύριο Γεώργιο Σπύρου, για την προσεκτική ανάγνωση της εργασίας μου και τις πολύτιμες υποδείξεις του. Ευχαριστώ τη μεταδιδακτορικό του Πανεπιστημίου Θεσσαλίας, Μαρία Παρασκευοπούλου, για τη συνδρομή της στην επιλογή του λογισμικού για τον καθορισμό των εμπλουτισμένων περιοχών έκφρασης και το συνάδελφό μου Γεώργιο Σκούφο για τη βοήθειά του στη δημιουργία των προσομοιωμένων δεδομένων του Κεφαλαίου 6.2. Ευχαριστώ τους φίλους μου Δημήτρη και Σίλβια για την ηθική τους υποστήριξη. Ευχαριστώ θερμά τη γιαγιά μου, που με στήριζε όλα αυτά τα χρόνια και με συμβούλευε με τον καλύτερο τρόπο. Πάνω από όλα, ευχαριστώ τη μητέρα μου για την αστείρευτη και ολόψυχη αγάπη και υποστήριξη που μου πρόσφερε όλα τα χρόνια που ήταν κοντά μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στην Αθήνα κατά τη διάρκεια του ακαδημαϊκού έτους 2015-2016 στα πλαίσια της φοίτησής μου στο διατμηματικό πρόγραμμα σπουδών "Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία". Το 2015 η καθηγήτρια του Πανεπιστημίου Θεσσαλίας κυρία Άρτεμις Χατζηγεωργίου και ο νυν ερευνητής της Ιατρικής Σχολής του Χάρβαρντ, κύριος Ιωάννης Βλάχος, μου έδωσαν την ευκαιρία να εργαστώ σε μία μελέτη συλλογής ανθρωπίνων small RNA-Seq δεδομένων και ανάλυσης των microRNA εκφράσεων στα δείγματα αυτά. Η ιδέα της διπλωματικής εργασίας αναδύθηκε αυτόματα, λόγω του μεγάλου όγκου δεδομένων που διαθέταμε. Αρχικά, ο στόχος μας ήταν η μελέτη της έκφρασης όλων των μικρών μη κωδικών RNAs, αλλά στην πορεία καταλάβαμε ότι έλειπαν τα βασικά στάδια της ανάλυσης των δεδομένων αυτών, μέχρι να οδηγηθούμε τελικά στην έκφραση, και τα οποία περιγράφονται στην παρούσα εργασία.

1. ΕΙΣΑΓΩΓΗ

Τα μικρά μη κωδικά μόρια RNA διαδραματίζουν σπουδαίο ρόλο στη ρύθμιση της γονιδιακής έκφρασης. Κάποια από τα μικρά RNAs, όπως για παράδειγμα τα microRNAs (miRNAs) και τα small interfering RNAs (siRNAs), μπορούν να επιφέρουν γονιδιακή σίγηση σε μετα-μεταγραφικό επίπεδο, στοχεύοντας ειδικά mRNA μόρια. Η ανάλυση των μικρών μορίων RNA με τη χρήση της αλληλούχησης επόμενης γενιάς (Next Generation Sequencing) αποτελεί ολοένα και πιο δημοφιλή μέθοδο για μελέτη του βιολογικού ρόλου των miRNAs και άλλων μικρών ρυθμιστικών μεταγράφων καθώς και της αποτύπωσης βιολογικού δικτύου γονιδιακής έκφρασης. [1]

Η αλληλούχηση μικρών RNAs (small RNA-Seq), η οποία παρέχει τη δυνατότητα για ανακάλυψη, σχολιασμό και ποσοτικοποίηση των μικρών RNA μορίων, είχε αρχικά σχεδιαστεί για τη μέτρηση της έκφρασης των miRNAs. Μία πιο κοντινή ματιά όμως στις προκύπτουσες αλληλουχίες, αποκάλυψε πολλούς διαφορετικούς τύπους μη κωδικών RNAs, με μήκη παρόμοια με αυτά των miRNAs. Κάποιοι χαρακτηριστικοί τύποι τέτοιων RNA αλληλουχιών, συμπεριλαμβάνουν τα tRNAs (ή ακριβέστερα θραύσματα των tRNAs), τα snoRNAs, 21U-RNAs ή snRNAs. Πρόσφατα, η αλληλούχηση small RNA-Seq, βοήθησε στον εντοπισμό καινούργιων RNA ειδών, όπως είναι τα microRNA offset RNAs (moRs), τα οποία προέρχονται από πρόδρομο RNA. [2] Ως εκ τούτου, τα small RNA-Seq δεδομένα, περιέχουν μια πληθώρα, πιθανώς άγνωστων ειδών RNAs. Παρά το γεγονός αυτό, τα περισσότερα εργαλεία ανάλυσης small RNA-Seq δεδομένων, όπως το miRanalyzer, miRDeep ή miRNAkey, εστιάζουν μόνο στα miRNAs, παραμελώντας σε μεγάλο βαθμό άλλους τύπους RNA. Επιπλέον, η εκτίμηση της έκφρασης των ίδιων των microRNAs παρουσιάζει σημαντικές δυσκολίες, όπως είναι για παράδειγμα το μικρό τους μήκος, miRNAs με παρόμοιες νουκλεοτιδιές αλληλουχίες, χημικές τροποποιήσεις που επιδέχονται κα..

2. ΜΙΚΡΑ ΜΗ ΚΩΔΙΚΑ RNAs

Τα μη-κωδικά RNAs (ncRNAs) είναι μόρια RNA, τα οποία μεταγράφονται από το γονιδίωμα και τα οποία δεν κωδικοποιούν πρωτεΐνες. Τα πρώτα γνωστά μη κωδικά RNAs, ήταν τα ριβοσωμικά RNAs (rRNA) και τα μεταφορικά RNAs (tRNAs), τα οποία εμπλέκονται στη μετάφραση του RNA. Ακολούθως, στη λίστα προστέθηκαν μερικά ακόμη RNAs, μεταξύ των οποίων τα μικρά πυρηνικά RNAs (snRNAs) τα οποία συμμετέχουν στη διαδικασία του ματίσματος, τα μικρά RNAs του πυρηνίσκου (snoRNAs), τα οποία εμπλέκονται στην κατεργασία και χημική τροποίηση των rRNAs και οι οδηγοί RNAs (gRNAs), που παίζουν σπουδαίο ρόλο στο editing του RNA. Από τα μη κωδικά αυτά ncRNAs, τα gRNAs (50-70 nt) όπως και τα snRNAs και τα snoRNAs (συνήθως μικρότερα των 200 nt), είναι σχετικά μικρά σε μήκος.

Τα τελευταία χρόνια ωστόσο, η ανακάλυψη των μικρών μη κωδικών RNAs (sncRNAs), αποκάλυψε μία νέα στρατιά μικρών, αλλά ισχυρών ρυθυμιστών της γονιδιακής έκφρασης. Το χαρακτηριστικό γνώρισμα των μικρών μη κωδικών RNAs, είναι το μικρό τους μέγεθος (20-30 nt), η συσχέτισή τους με τα μέλη της Argonaute (Ago) πρωτεϊνικής οικογένειας και τυπικά η επίδρασή τους στη ρύθμιση/σίγηση της γονιδιακής έκφρασης. [3] Παρόλο που στα ευκαρυωτικά κύτταρα, η μεταγραφή πολλών γονιδίων καταστέλλεται ή υπόκειται στη γονιδιακή σίγηση, υπάρχει και το σενάριο της μεταγραφής των γονιδίων, τα οποία δε μεταφράζονται ποτέ. Τα μικρά μη κωδικά RNAs προσθέτουν ένα επιπλέον επίπεδο ελέγχου στο ήδη πολύπλοκο σύστημα ρύθμισης της ευκαρυωτικής γονιδιακής έκφρασης, είτε αναστέλλοντας τη μετάφραση, είτε προωθώντας την αποδόμηση συγκεκριμένων μεταγράφων RNAs (mRNAs). [4]

Τρεις είναι οι κύριες κατηγορίες των μικρών μη κωδικών RNAs οι οποίες έχουν μελετηθεί εκτενώς, τα small interfering RNA (siRNAs), τα microRNAs και τα Piwi-interacting RNAs (piRNAs). [5]

Η βιβλιογραφία των sncRNA ωστόσο μεγαλώνει και νέα μέλη να προστίθενται συνεχώς, όπως για παράδειγμα τα repeat-associated siRNA (rasiRNAs), transacting siRNA (tasiRNA), natural antisense transcript siRNA (natsiRNA), heterochromatic siRNA (hc-siRNA), small scan RNA (scnRNA), 21-mer με 5’ ουριδίνη (21U-RNA) και QDE2-interacting small RNA (qiRNA). [3]

Η τεχνολογία NGS έχει χρησιμοποιηθεί εκτενώς για τη σκιαγράφηση της έκφρασης και την ανακάλυψη των microRNAs και άλλων μικρών μη κωδικών RNAs σε πολλούς οργανισμούς [6], και περιγράφεται αναλυτικά στην επόμενη ενότητα.

3. ΑΛΛΗΛΟΥΧΗΣΗ ΕΠΟΜΕΝΗΣ ΓΕΝΙΑΣ

Με τη βοήθεια κατάλληλων μεθόδων, είναι εφικτή η ανάγνωση κάθε βάσης μίας αλληλουχίας νουκλεοτιδίων ενός τμήματος DNA ή RNA. Η επικρατέστερη μέθοδος αλληλούχησης του DNA ανακαλύφθηκε στα τέλη της δεκαετίας του ’70 και χρησιμοποιείται με μικρές μόνο παραλλαγές μέχρι σήμερα. Πρόκειται για την ενζυμική μέθοδο αλληλούχησης με «τερματισμό επιμήκυνσης της αλυσίδας» (chain termination) με τη βοήθεια τροποποιημένων δεοξυνουκλεοτιδίων, που ανακάλυψε ο βρετανός βιοχημικός Frederick Sanger το 1977.

Οι αυξανόμενες ανάγκες της έρευνας και της διαγνωστικής για αλληλούχηση μεγαλύτερων περιοχών DNA (ή ακόμα και ολόκληρων γονιδιωμάτων) σε ολοένα και μεγαλύτερο αριθμό δειγμάτων, σε σύντομο χρονικό διάστημα και με λογικό κόστος, οδήγησαν στην ανάπτυξη νέων τεχνολογιών αλληλούχησης, που συνδυάζουν τις πρόσφατες εξελίξεις των κλάδων της χημείας, της μηχανολογίας, της μοριακής βιολογίας και της πληροφορικής. Τα πρώτα εμπορικά διαθέσιμα μηχανήματα αλληλούχησης επόμενης γενιάς (Next generation sequencing, NGS) παρουσιάστηκαν το 2005, οδηγώντας σε μεγάλη αύξηση του ρυθμού παραγωγής γονιδιωματικών δεδομένων (αλληλούχηση ενός ανθρωπίνου γονιδιώματος σε λιγότερο από τρεις ημέρες). Σήμερα, είναι εμπορικά διαθέσιμες τουλάχιστον τέσσερις διαφορετικές πλατφόρμες αλληλλούχησης επόμενης γενιάς. Καθεμιά από αυτές χρησιμοποιεί διαφορετική χημεία και διαφορετικό τρόπο ανίχνευσης της αλληλουχίας των βάσεων του DNA. Κοινό χαρακτηριστικό όλων, είναι η μαζικά παράλληλη φύση της αλληλούχησης, δηλαδή η ταυτόχρονη αλληλούχηση πολλών μορίων DNA ή RNA και στη συνέχεια η συναρμολόγηση των επιμέρους αλληλουχιών με εξελιγμένους αλγόριθμους πληροφορικής.

Παρακάτω παρουσιάζεται μία απλουστευμένη περιγραφή των βασικών αρχών της αλληλούχησης επόμενης γενιάς.

1. Η μαζικά παράλληλη αλληλούχηση ξεκινά με την επιλογή του μορίου στόχος, το οποίο μπορεί να είναι για παράδειγμα ολόκληρο το γονιδιωματικό DNA ενός οργανισμού, μόρια του mRNA ενός οργανισμού ή ιστού, το σύνολο των μη κωδικοποιούντων RNA ενός οργανισμού ή ιστού κ.α.

2. Στη συνέχεια, το μόριο στόχος κατακερματίζεται σε κομμάτια συγκεκριμένου μεγέθους. Στα άκρα των θραυσμάτων προστίθενται γνωστές αλληλουχίες πρόσδεσης (adapters). Στο στάδιο αυτό, έχει δημιουργηθεί μια βιβλιοθήκη θραυσμάτων με γνωστά άκρα, τα οποία μπορούν να χρησιμοποιηθούν για ενίσχυση ή άλλο χειρισμό τους. Σε όλες τις πλατφόρμες αλληλούχησης επόμενης γενιάς, η ενίσχυση των βιβλιοθηκών γίνεται σε ένα στερεό υπόστρωμα. Αυτό μπορεί να είναι είτε η επιφάνεια ενός σφαιριδίου σε γαλάκτωμα (emulsion PCR), είτε η επιφάνεια μιας γυάλινης πλάκας (bridge amplification). Κάθε προσδεδεμένο μόριο DNA αποτελεί μια ανεξάρτητη θέση ενίσχυσης, διατηρώντας τη λογική της κλωνικής ενίσχυσης. Η ενίσχυση εξασφαλίζει ότι η ποσότητα του DNA είναι επαρκής για τη μέτρηση του σήματος κατά τη διαδικασία της αλληλούχησης.

3. Μετά την πρόσδεση και την ενίσχυση των θραυσμάτων DNA που απαρτίζουν τη βιβλιοθήκη θραυσμάτων ακολουθεί η διαδικασία της αλληλούχησης. Στο στάδιο αυτό, γίνεται ιδιαίτερα σαφής η μαζικά παράλληλη φύση της διαδικασίας, καθώς όλα τα θραύσματα DNA αλληλουχούνται ταυτόχρονα. Τα στάδια της διαδικασίας είναι:

- Προσθήκη κάθε φορά ενός διαφορετικού νουκλεοτιδίου.

- Ανίχνευση των νουκλεοτιδίων που ενσωματώθηκαν σε κάθε θραύσμα με οπτική ή ηλεκτρονική μέθοδο.

- Έκπλυση των μη ενσωματωμένων νουκλεοτιδίων.

- Κυκλική επανάληψη των παραπάνω βημάτων μέχρι να ολοκληρωθεί η αλληλούχηση των θραυσμάτων.

4. Η περαιτέρω ανάλυση των αλληλουχιών που προκύπτουν γίνεται με ειδικούς αλγόριθμους. Αποσκοπεί αφενός στην αξιολόγηση της ποιότητας της ανάγνωσης, ώστε να επιλεγούν οι αλληλουχίες με το χαμηλότερο ποσοστό σφάλματος, και αφετέρου στην εξαγωγή της αλληλουχίας με μορφή η οποία είναι κατάλληλη για περαιτέρω επεξεργασία. Ανάλογα με την εφαρμογή, οι αλληλουχίες μπορούν:

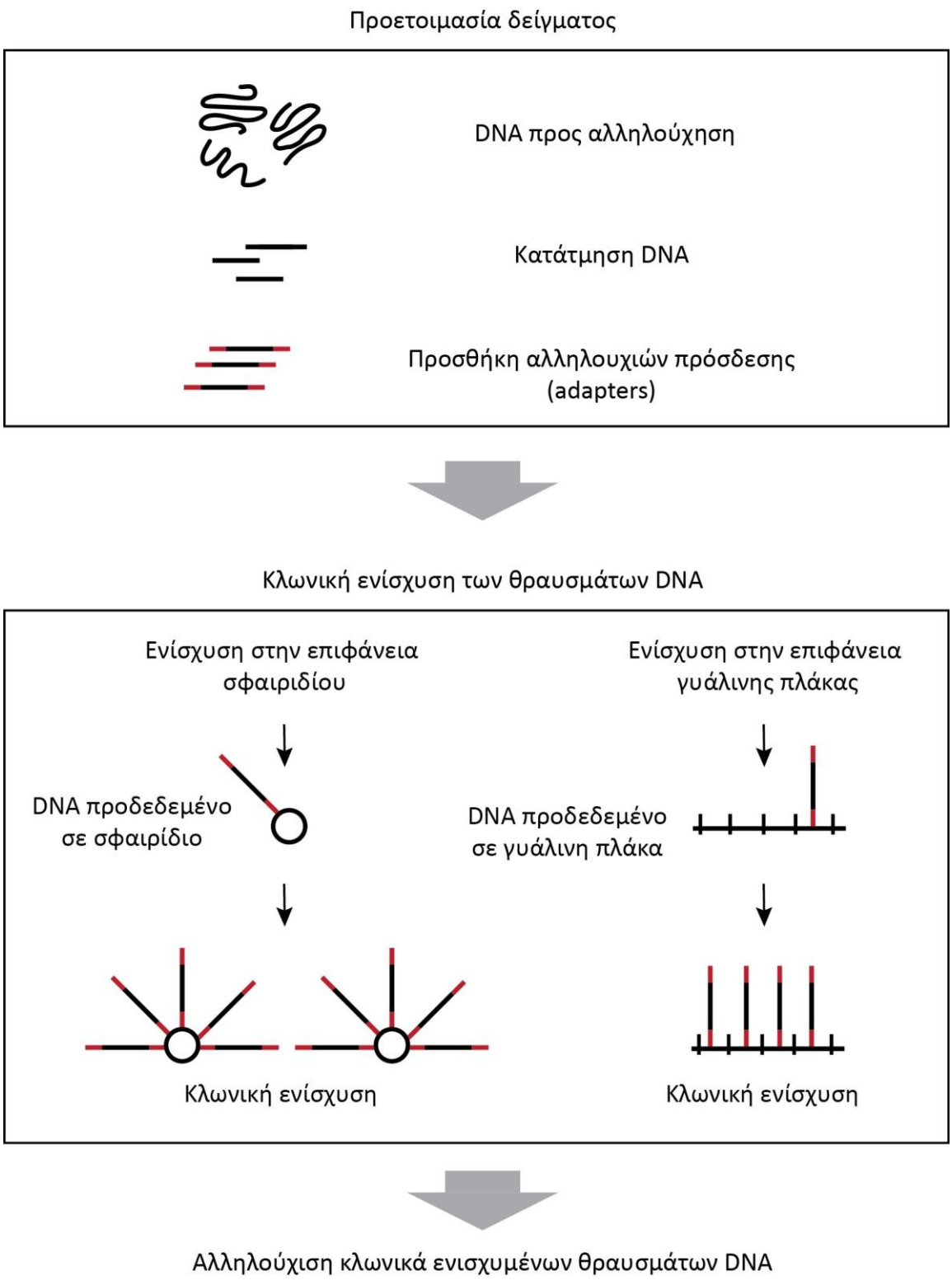
- Να στοιχιστούν σε μια αλληλουχία αναφοράς, ώστε να διαπιστωθούν νουκλεοτιδικές ή δομικές αλλαγές.

- Να συναρμολογηθούν εκ νέου με βάση τις κοινές τους περιοχές, έτσι ώστε να σχηματιστεί μια νέα μεγαλύτερη αλληλουχία (Εικόνα 2).

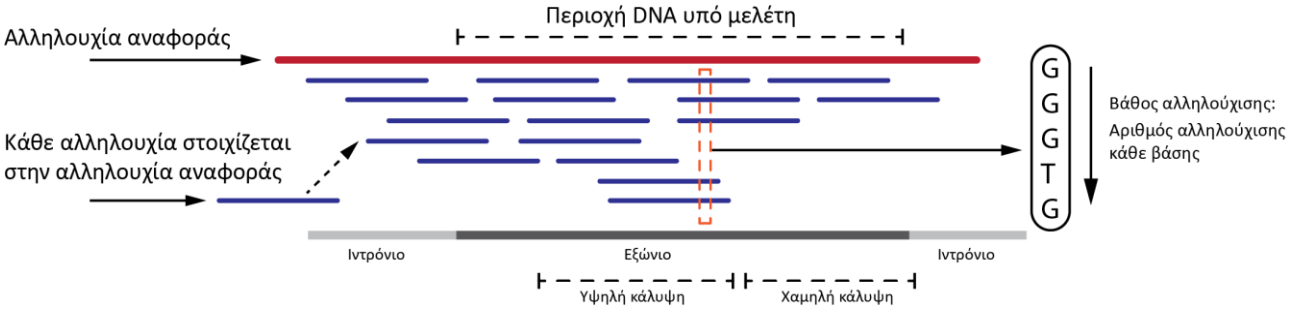
- Να προσδιοριστεί ο σχετικός αριθμός τους, έτσι ώστε να προκύψουν ποσοτικά δεδομένα που αφορούν την έκφραση ενός γονιδίου ή τη συχνότητα αλληλόμορφων σε έναν πληθυσμό.

Τα δεδομένα αλληλούχησης από τις πλατφόρμες επόμενης γενιάς είναι εκ φύσεως ψηφιακά. Η φύση των δεδομένων επιτρέπει την εκτίμηση ποσοτικών παραμέτρων της αλληλουχίας, που είναι ιδιαίτερα χρήσιμες κατά τη μελέτη βιολογικών συστημάτων. Για παράδειγμα, μπορούν να μελετηθούν χρωμοσωμικές ενισχύσεις, που είναι συχνές σε νεοπλασίες. Ακόμα, ο αριθμός αναγνώσεων σε μια βιβλιοθήκη RNA μπορεί να προσφέρει πληροφορίες για το επίπεδο έκφρασης των μεταγράφων. Σε μελέτες πληθυσμών και μεταγενομικής, είναι δυνατή η αξιολόγηση της εκπροσώπησης ενός γονιδιώματος επί του συνόλου του γενετικού υλικού που εξετάστηκε.

Η τεχνολογία αλληλούχησης επόμενης γενιάς έχει επιταχύνει την ανάπτυξη της βιολογικής έρευνας και έχει αρχίσει να καταλαμβάνει κεντρική θέση σε διαγνωστικές εφαρμογές. [7]



Εικόνα 1: Αλληλούχηση επόμενης γενιάς. Διαδικασία προετοιμασίας των δειγμάτων προκειμένου να υποβληθούν σε αλληλούχηση επόμενης γενιάς. Το DNA κατατμείται σε θραύσματα επιθυμητού μεγέθους. Στα άκρα των θραυσμάτων προστίθενται γνωστές αλληλουχίες πρόσδεσης. Τα θραύσματα ακινητοποιούνται μέσω των αλληλουχιών πρόσδεσης σε στερεό υπόστρωμα (επιφάνεια σφαιριδίου ή γυάλινη πλάκα). Τα ακινητοποιημένα θραύσματα υφίστανται πολλαπλασιασμό (κλωνική ενίσχυση) με τη μέθοδο της PCR. Τα ενισχυμένα θραύσματα είναι έτοιμα προς αλληλούχηση, (Γ. Παπανικολάου, 2015)



Εικόνα 2: Βάθος αλληλούχησης. Οι αλληλουχίες που προκύπτουν από τα κλωνικά ενισχυμένα θραύσματα DNA μπορούν να στοιχηθούν σε μια αλληλουχία αναφοράς. Τα θραύσματα του DNA παράγονται με τέτοιο τρόπο ώστε να έχουν αλληλοεπικαλυπτόμενες περιοχές. Όταν σε μια περιοχή της αλληλουχίας αναφοράς έχουν αναγνωστεί (και στοιχιστεί) πολλές αλληλουχίες θραυσμάτων, τότε θεωρούμε ότι η περιοχή έχει υψηλή κάλυψη. Σε περιοχές με υψηλή κάλυψη μπορύν να ανιχνευτούν αλλαγές της αλληλουχίας που απαντώνται σε πολύ μικρό ποσοστό, ακόμη και της τάξης του 1%. Αντίθετα, σε περιοχές με χαμηλή κάλυψη η ευαισθησία της αλληλούχησης μειώνεται δραματικά (Γ. Παπανικολάου, 2015)

4. ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ SMALL RNA-SEQ ΠΕΙΡΑΜΑΤΑ

Η ποιοτική και ποσοτική ανάλυση των μικρών μη-κωδικών RNAs από αλληλουχήσεις επόμενης γενιάς (small RNA-Seq) είναι μια νέα τεχνολογία, που χρησιμοποιείται, όλο και περισσότερο, για να διερευνήσει με υψηλή ευαισθησία και ειδικότητα, τους RNA πληθυσμούς, που περιλαμβάνουν τα microRNAs, καθώς και άλλα ρυθμιστικά μικρού μήκους μετάγραφα. Η ανάλυση των small RNA-Seq δεδομένων έχει ως στόχο την άντληση των βιολογικών πληροφοριών, όπως για παράδειγμα την ανίχνευση, την ποσοτικοποίηση και την ανάλυση των γνωστών και νέων μη-κωδικών RNAs. Για κάθε απαιτούμενο βήμα της ανάλυσης, απαιτείται η εφαρμογή ειδικών στατιστικών και βιοπληροφορικών εργαλείων. [1]

Ένα κοινό προκύπτον πρόβλημα ανάλυσης των small RNA-Seq δεδομένων, αποτελούν οι εγγραφές με πολλαπλές έγκυρες θέσεις ευθυγράμμισης πάνω στο γονιδίωμα. Η πλειοψηφία των μεταγράφων από small RNA-Seq πειράματα, ανήκει στην κλάση των miRNAs, τα οποία εκτός του μικρού τους μήκους, τείνουν να εμφανίζονται σε οικογένειες με πολύ παρόμοιες αλληλουχίες. Μια περαιτέρω περιπλοκή, αποτελεί η προσθήκη μονού νουκλεοτιδίου 3’ αδενοσίνης ή ουρακίλης, ανεξάρτητα από το αντίστοιχο νουκλεοτίδιο του DNA κλώνου, σε αρκετά από τα ώριμα miRNAs. Ένα χαρακτηριστικό παράδεγμα αποτελεούν τα δύο miRNAs let-7b και το let-7c, των οποίων τα ώριμα miRNAs διαφέρουν σε ένα μονάχα νουκλετοδίο. Εάν λοιπόν, στο let-7b, υπάρξει μία προσθήκη 3’ αδενοσίνης, τότε η προκύπτουσα αλληλουχία ευθυγραμμίζεται με το ίδιο σκορ είτε στη γονιδιωματική περιοχή του let-7b είτε στη γονιδιωματική περιοχή του let-7c. Συνήθως ο τρόπος με τον οποίον ποσοτικοποιείται μία τέτοια εγγραφή, είναι να κατανεμηθεί μεταξύ των υποψήφιων μεταγράφων, να κατανεμηθεί τυχαία σε ένα μετάγραφο ή να απορριφθεί εντελώς. [14]

Παρακάτω παρουσιάζονται επιλεγμένα υπάρχοντα εργαλεία για την ανάλυση των small RNA-Seq δεδομένων, καθένα από τα οποία προσεγγίζει διαφορετικά ή με κάποιες παραλλαγές, το πρόβλημα της ανάλυσης των μικρών RNAs, με χρήση διαφορετικών βιοπληροφορικών εργαλείων και στατιστικών προσεγγίσεων.

4.1 miRanalyzer

Το miRanalyzer (Michael Hackenberg et al., 2009) είναι μία web υπηρεσία, η οποία έχει ως στόχο την ανίχνευση των υπαρχόντων miRNAs και την πρόβλεψη των καινούργιων miRNA αλληλουχιών, από δείγματα βαθιάς αλληλούχησης. Παρακάτω, παρουσιάζονται τα βασικά βήματα του αλγορίθμου:

- Ανίχνευση των γνωστών miRNA αλληλουχιών
- Ευθυγράμμιση εγγραφών πάνω σε μεταγράφωμα
- Πρόβλεψη καινούργιων miRNA αλληλουχιών

Κατά την ανίχνευση των γνωστών miRNAs, οι εγγραφές ευθυγραμμίζονται πάνω στις ώριμες και πρόδρομες αλληλουχίες miRNAs, προερχόμενες από τη βάση δεδομένων miRbase. Εγγραφές οι οποίες δεν αντιστοιχίστηκαν σε γνωστές miRNA αλληλουχίες του

προγηγούμενου βήματος, ευθυγραμμίζονται σε γνωστά mRNA, μη κωδικά RNA (RFam) και (ρετρο)-τρανσποζόνια μετάγραφα. Τέλος, η πρόβλεψη των καινούργιων miRNA αλληλουχιών, συντελείται με μία μέθοδο μηχανικής μάθησης βασισμένη στη μέθοδο του random forest με ένα μεγάλο εύρος χαρακτηριστικών. Για την υλοποίηση του τελικού μοντέλου πρόβλεψης, χρησιμοποιήθηκαν τρία διαφορετικά σετ δεδομένων από ανθρώπους (hsa), μύες (rat) και νηματοειδή σκουλήκια (cel). [13] Το miRanalyzer είναι διαθέσιμο στον παρακάτω σύνδεσμο:

http://bioinfo5.ugr.es/miRanalyzer/miRanalyzer.php.

4.2 miRDeep2

Το miRDeep2 (Marc R. Friedländer et al., 2011) είναι ένα δημοφιλές εργαλείο ποσοτικοποίησης γνωστών και άγνωστων miRNA αλληλουχιών από δεδομένα βαθιάς αλληλούχησης. Τα κύρια βήματα του αλγορίθμου περιλαμβάνουν τα παρακάτω:

- Προεργασία δεδομένων και ευθυγράμμιση των εγγραφών πάνω στο γονιδίωμα
- Ποσοτικοποίηση των γνωστών miRNAs ή/και πρόβλεψη καινούργιων miRNA αλληλουχιών

Το βασικό κομμάτι της ποσοτικοποίησης και του προφίλ έκφρασης των γνωστών miRNAs, βασίζεται στην ευθυγράμμιση των εγγραφών και των ώριμων γνωστών miRNAs πάνω σε αλληλουχίες των miRNA προδρόμων. Η δημιουργία του αρχείου index, καθώς και η ευθυγράμμιση των εγγραφών, πραγματοποιούνται με το εργαλείο Bowtie. [11]

Για την πρόβλεψη των καινούργιων miRNA αλληλουχιών, το miRDeep2 χρησιμοποιεί ένα πιθανοκρατικό μοντέλο βασισμένο στη βιογένεση του μορίου miRNA, για την αξιολόγηση του ταιριάσματος της θέσης και συχνότητας του αλληλουχημένου RNA με τη δευτεροταγή δομή του miRNA precursor.

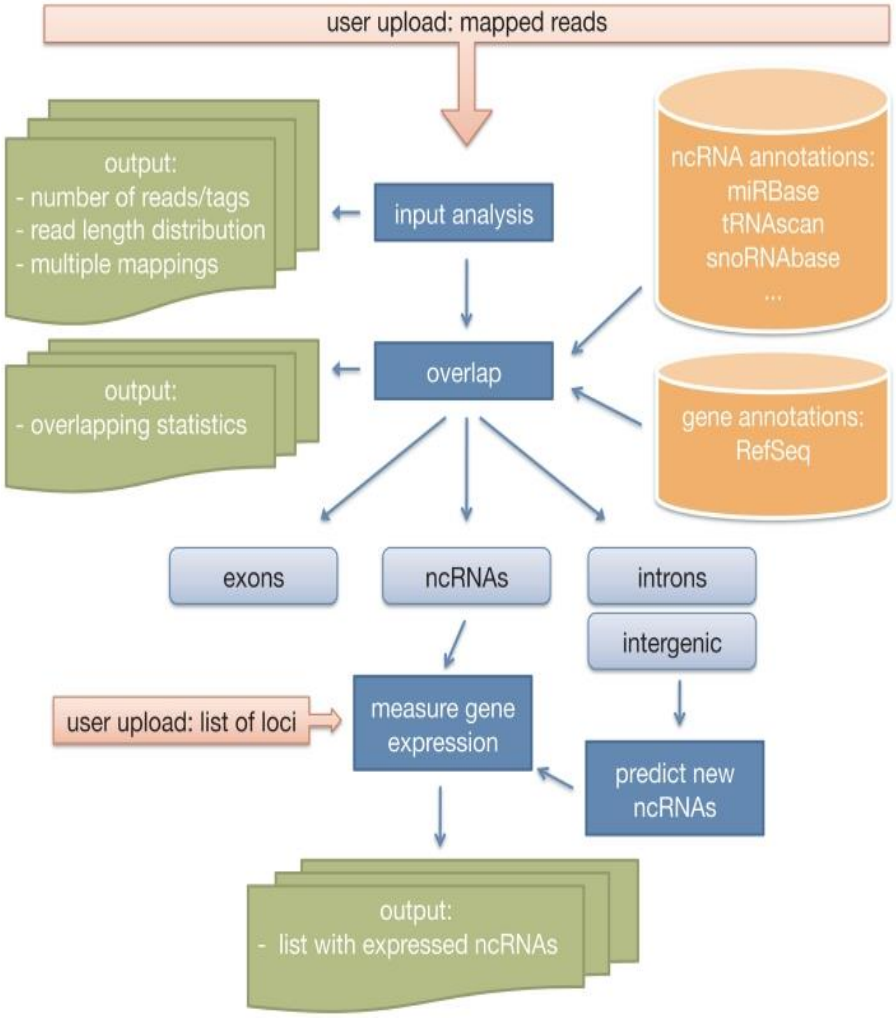
Μετά την ευθυγράμμιση των εγγραφών αλληλούχησης πάνω στο γονιδίωμα, ο αλγόριθμος πραγματοποιεί το «DNA bracketing» γύρω από τις περιοχές ευθυγράμμισης, υπολογίζοντας τη δευτεροταγή δομή του RNA. Αναγνωρίζονται, σύμφωνα με πιθανότητα, οι αξιόπιστες αλληλουχίες πρόδρομων miRNAs. Αναγνωρίζονται οι αξιόπιστες miRNA precursor αλληλουχίες με βάση την πιθανότητά τους να είναι πραγματικά miRNA precursors. Η έξοδος του αλγορίθμου είναι μία λίστα με σκορ γνωστών και αγνώστων miRNA προδρόμων και ωρίμων miRNA κλώνων του δείγματος βαθιάς αλληλούχησης, καθώς και οι εκτιμήσεις των ψευδώς θετικών αποτελεσμάτων. [12] Το miRDeep2 είναι διαθέσιμο στον παρακάτω σύνδεσμο: https://www.mdc-berlin.de/8551903/en/

4.3 DARIO

Το DARIO (Mario Fasold et al., 2011) είναι μία web υπηρεσία η οποία δέχεται σαν είσοδο τις ήδη ευθυγραμμισμένες σε γονιδίωμα αναφοράς εγγραφές, αποθηκευμένες σε αρχεία με μορφή BAM ή BED. Τα αρχικά βήματα εκτέλεσης της ανάλυσης του DARIO

περιλαμβάνουν τον ποιοτικό έλεγχο των δεδομένων εισόδου. Στη συνέχεια, οι εγγραφές επικαλύπτονται με γονιδιακά μοντέλα του επιλεγμένου προς μελέτη είδους. Από την ανάλυση εξαιρούνται εκείνες οι εγγραφές, οι οποίες επικαλύπτονται με περιοχές εξονίων. Στην ανάλυση της έκφρασης, οι εγγραφές επικαλύπτονται με σχολιασμένα μη κωδικά RNAs τα οποία έχουν συλλεχθεί από διάφορες πηγές-βάσεις δεδομένων. Ο τρόπος με τον οποίο χειρίζονται οι εγγραφές με πολλαπλές έγκυρες θέσεις ευθυγράμμισης πάνω στο γονιδίωμα, βασίζεται στη διαίρεση του πλήθους των εγγραφών για μία αλληλουχία δια του πλήθους των περιοχών ευθυγράμμισης. Έτσι, παράγεται μία λίστα με τα εκφρασμένα ncRNAs και το αντίστοιχο πλήθος τους.

Για την πρόβλεψη των καινούργιων μη κωδικών RNA, το DARIO χρησιμοποιεί μέθοδο μηχανικής μάθησης βασισμένη σε ταξινομητή random forest. Η μέθοδος αυτή, βασίζεται στα χαρακτηριστικά μοτίβα που σχηματίζουν οι εγγραφές, οι οποίες αντιπροσωπεύουν διαφορετικές κλάσεις των ncRNAs. [2] Το DARIO είναι διαθέσιμο στον παρακάτω σύνδεσμο: http://dario.bioinf.uni-leipzig.de/help.py.



Σχήμα 1: Διάγραμμα ροής του εργαλείου DARIO (Mario Fasold, 2011)

Σχόλια

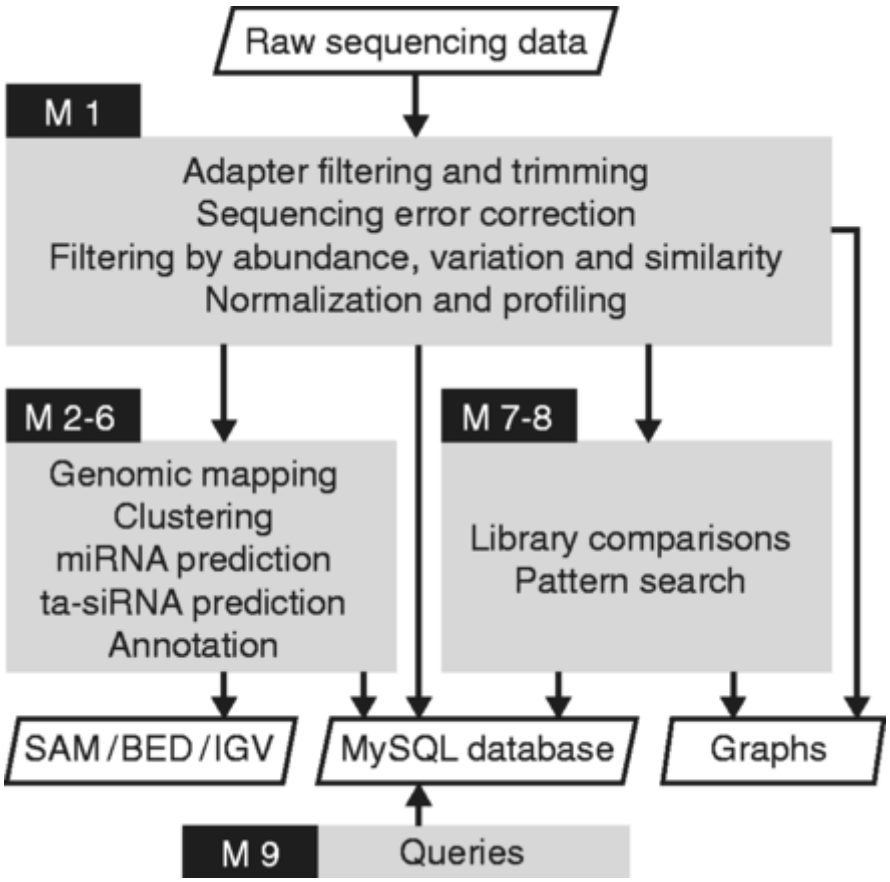
Ενώ το DARIO παράγει ευανάγνωστα αρχεία εξόδου με απλή αναπαράσταση, έχει ένα βασικό μειονέκτημα, το οποίο είναι ο περιορισμός στο μέγεθος του αρχείου εισόδου (60MB). Παρόλο που τα αρχεία εισόδου πρέπει να έχουν BED φορμάτ, ώστε να συντελείται η μέγιστη συμπίεση, αυτό δεν είναι αρκετό όταν πρόκειται για μεγάλα αρχεία τα οποία περιέχουν multi-mapped εγγραφές. Όπως έχει δειχθεί προηγουμένως, πολλά από τα μικρά RNAs, παρουσιάζουν περισσότερες από μία έγκυρες θέσεις ευθυγράμμισης πάνω στο γονιδίωμα. Κατά συνέπεια, το τελικό αρχείο ευθυγράμμισης, συμπεριλαμβάνει όλες αυτές τις θέσεις με αποτέλεσμα, το μέγεθός του να αυξάνεται εκθετικά.

4.4 shortran

Το shortran (Gupta V et al., 2012) περιλαμβάνει τα παρακάτω εννέα βήματα για την ανάλυση των μικρών ζωικών και φυτικών RNA δεδομένων από πειράματα NGS.

- Προεργασία των δεδεομένων εισόδου
- Ευθυγράμμιση των εγγραφών πάνω στο γονιδίωμα
- Ομαδοποίηση των εγγραφών
- Πρόβλεψη καινούργιων miRNA αλληλουχιών
- Πρόβλεψη των καινούργιων trans acting siRNA (ta-si-RNA) αλληλουχιών
- Σχολιασμός των εγγραφών-μεταγράφων
- Σύγκριση των βιβλιοθηκών
- Διαγράμματα έκφρασης
- Εύρεση των προτύπων – Μαρκοβιανή πιθανότητα των δύο και τριών τελευταίων 5’ νουκλεοτιδίων

Τα κυριότερα σημεία του αλγορίθμου, περιλαμβάνουν τον καθαρισμό των NGS δειγμάτων με τη βοήθεια του εργαλείου FASTX, την ευθυγράμμιση των εγγραφών πάνω στο γονιδίωμα με τη χρήση του εργαλείου Bowtie και την πρόβλεψη των μικρών RNAs με τους αλγορίθμους miRDeep ή miRDeep-P, όταν πρόκειται για φυτικά δεδομένα εισόδου. Επιπλέον, ο αλγόριθμος πραγματοποιεί είτε μία απλή ανάλυση κατά συστάδες, είτε μπορεί να χρησιμοποιήσει τον αλγόριθμο NiBLs. Τέλος, πραγματοποιείται ο σχολιασμός των περιοχών έκφρασης οι οποίες προκύπτουν από την ανάλυση κατά συστάδες και παράγονται γραφήματα έκφρασης καθώς και παράγοντες συσχέτισης μεταξύ των δειγμάτων με σκοπό την εύρεση των συνεκφραζόμενων βιβλιοθηκών. [8] Το shortran είναι διαθέσιμο στον παρακάτω σύνδεσμο https://omictools.com/shortran-tool.



Σχήμα 2: Διάγραμμα ροής του εργαλείου shortran (Vikas Gupta, 2012)

Σχόλια

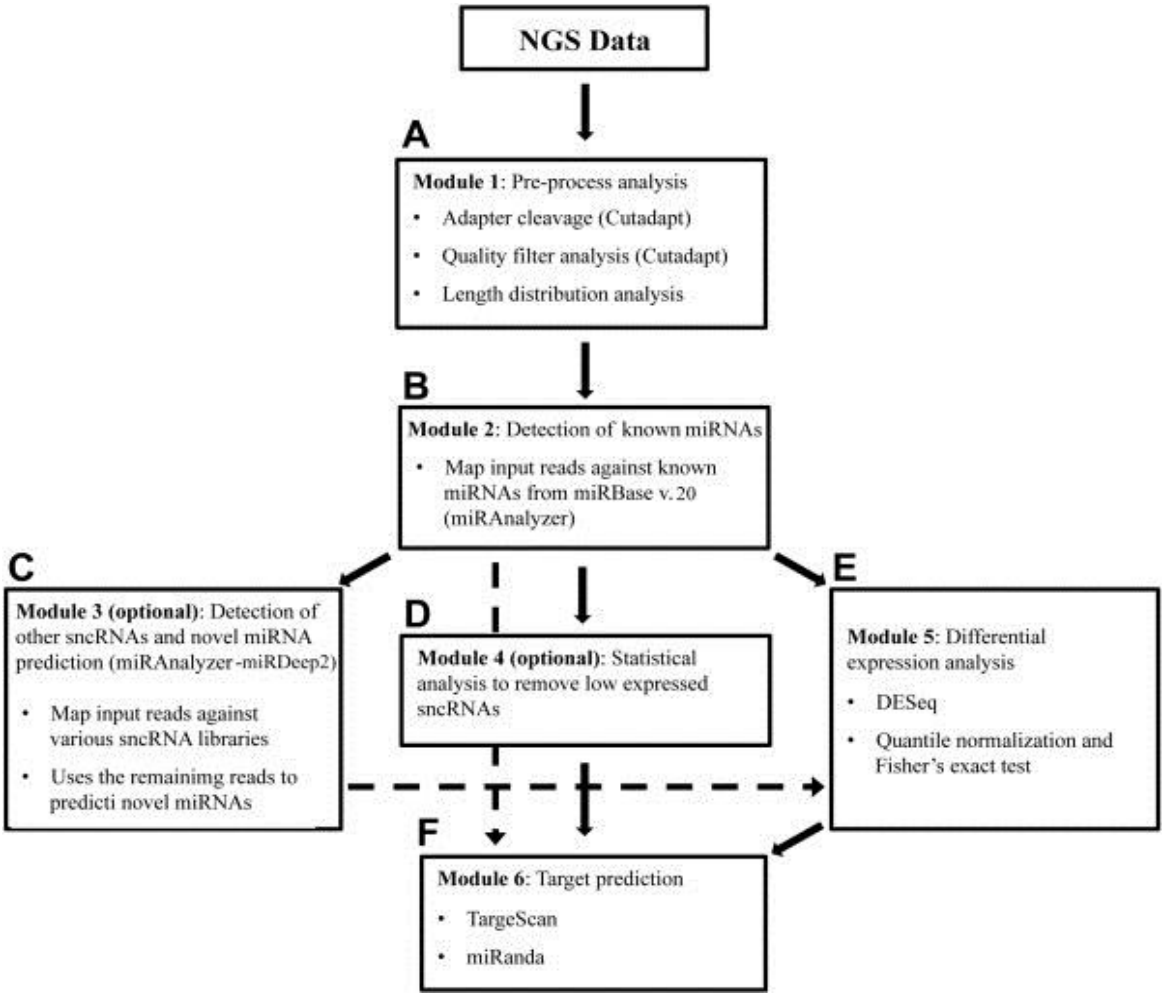
Ένα σημαντικό μειονέκτημα του shortran είναι το γεγονός ότι κάθε φορά που εκτελείται, χτίζει εκ νέου το Index του εκάστοτε γονιδιώματος. Αυτό έχει σαν συνέπεια μία σημαντική καθυστέρηση στην εκτέλεση, ιδίως όταν πρόκειται για μεγάλα γονιδιώματα, όπως είναι το ανθρώπινο. Επίσης, τα αρχεία εξόδου που παράγονται κατά την εκτέλεση το εργαλείου shortran, είναι πολλά και δυσανάγνωστα, ενώ το βήμα Clustering που είναι υπεύθυνο για την ομαδοποίηση των εγγραφών σε συστάδες, δεν προβαίνει σε εύκολη κατανόηση του σχολιασμού τους, παρά την επιλογή που υπάρχει.

4.5 iMir

Το εργαλείο iMir (Giurato G et al., 2013) δημιουργήθηκε με σκοπό την ανάλυση των smallRNA-Seq δεδομένων, αποτελούμενο από τα παρακάτω έξι ξεχωριστά βήματα εκτέλεσης.

- Προεργασία των δεδομένων εισόδου
- Ανίχνευση των γνωστών miRNAs
- Ανίχνευση άλλων μη κωδικών RNAs και πρόβλεψη των καινούργιων miRNAs αλληλουχιών
- Αφαίρεση των μικρών μη κωδικών RNAs με χαμηλή έκφραση
- Διαφορική έκφραση
- Πρόβλεψη στόχων των miRNAs

Αρχικά, από τα NGS αρχεία εισόδου αφαιρούνται οι αλληλουχίες πρόσδεσης, με τη χρήση του εργαλείου cutadapt. Στη συνέχεια, με τη χρήση του miRanalyzer, ανιχνεύονται οι γνωστές miRNA ακολουθίες. Στο βήμα αυτό, υπάρχει και η δυνατότητα εκτέλεσης ανάλυσης κατά συστάδες (cluster analysis), της ανάλυσης κύριων συνιστωσών και/ή η εφαρμογή των διαφορετικών ιεραρχικών αλγορίθμων ομαδοποίησης. Η δυνατότητα εκτέλεσης των παραπάνω είναι πολύ χρήσιμη, όταν υπάρχει μεγάλος αριθμός δειγμάτων για την αξιολόγηση των ομοιοτήτων και των διαφορών μεταξύ τους, όπως για παράδειγμα κατά την ανάλυση αποτελεσμάτων ομάδων από βιοψίες όγκων. Καινούργιες miRNA αλληλουχίες προβλέπονται με τη χρήση των miRanalyzer και miRDeep. Πριν την εκτέλεση της διαφορικής έκφρασης, το iMir παρέχει τη δυνατότητα εκτέλεσης ενός ενδιάμεσου σταδίου, που αφορά την αφαίρεση του θορύβου, ο οποίος μπορεί να προκληθεί από μετάγραφα με πολύ χαμηλό αριθμό εγγραφών. [1] Το εργαλείο iMir είναι διαθέσιμο στον παρακάτω σύνδεσμο http://www.labmedmolge.unisa.it/inglese/research/imir.



Σχήμα 3: Διάγραμμα ροής του εργαλείου iMir (Giurato G, 2013)

5. ΑΛΓΟΡΙΘΜΜΟΣ spipeRNA

5.1 Επισκόπηση

Το spipeRNA δημιουργήθηκε λόγω της υπάρχουσας ανάγκης, για μία ολοκληρωμένη λύση στη διαχείριση όλων των μικρών μορίων RNA από NGS πειράματα. Παρατηρήθηκε έλλειψη εύκολου στη χρήση εργαλείου, με απλή και κατανοητή έξοδο, το οποίο να μην περιορίζει το χρήστη με το μέγεθος των αρχείων εισόδου. Επιπλέον, υπάρχει ανάγκη να αντιμετωπιστούν τα βασικά προβλήματα στη μελέτη της έκφρασης των μικρών μη κωδικών RNAs, τα οποία περιλαμβάνουν τον τρόπο ευθυγράμμισης, τη διαχείριση των multi-mapped εγγραφών, τον τρόπο σχολιασμού των εγγραφών και το χειρισμό των εγγραφών χωρίς υπάρχοντα σχολιασμό.

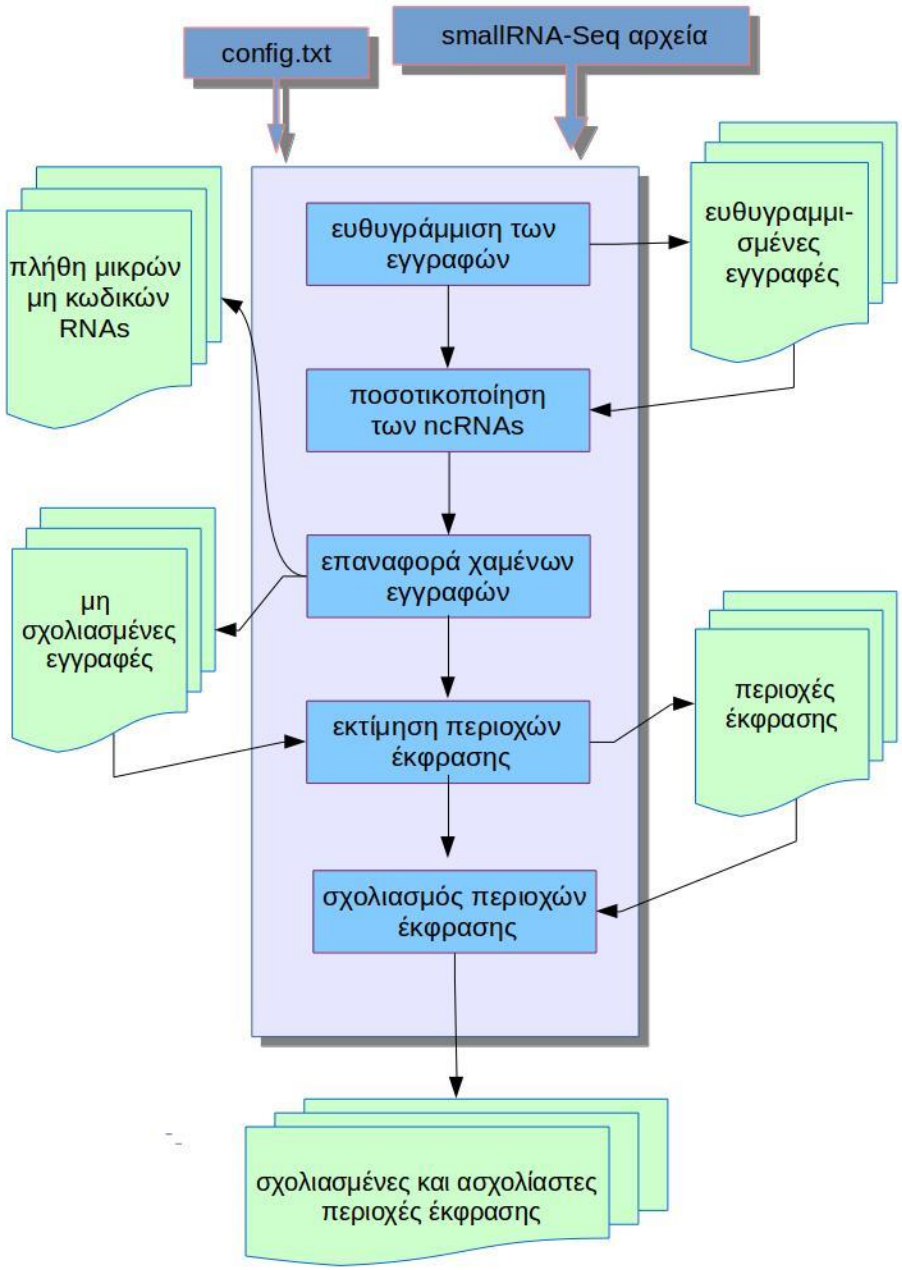
Το spipeRNA επιτρέπει την ανάλυση των small RNA-Seq δεδομένων, η οποία περιλαμβάνει την ποσοτικοποίηση των διάφορων ειδών sncRNAs καθώς και την εύρεση των περιοχών έκφρασης χωρίς υπάρχοντα σχολιασμό.

Εκτελείται με τη βοήθεια της γραμμής εντολών και δύο αρχείων ρύθμισης. Το spipeRNA είναι μία συλλογή scripts γραμμένων στη γλώσσα Perl, τα οποία είτε καλούν άλλα έτοιμα πακέτα και εργαλεία, είτε πραγματοποιούν κάποιες απαραίτητες διενέργειες στην ανάλυση των μικρών μη κωδικών RNAs.

5.2 Βήματα εκτέλεσης

Στο σχήμα 4 παρουσιάζεται το διάγραμμα ροής των βημάτων του εργαλείου spipeRNA. Σαν είσοδο, το εργαλείο δέχεται ψηφιακά αρχεία αλληλούχησης επόμενης γενιάς σε μορφή fastq, fasta ή csfasta, από τα οποία έχουν αφαιρεθεί οι αλληλουχίες πρόσδεσης (adapters, barcodes κτλ.) και στα οποία έχει εφαρμοστεί ο έλεγχος ποιότητας.

Το spipeRNA είναι μία σειρά από αρχεία γραμμένα στη γλώσσα Perl και R, τα οποία εκτελούνται σειριακά. Το εργαλείο έχει τη δυνατότητα παράλληλης εκτέλεσης, σε περίπτωση που ο χρήστης επιθυμεί να τρέξει πολλά δείγματα μαζί. Το spipeRNA δέχεται δύο αρχεία εισόδου, ένα αρχείο με τις διαδρομές προς τα διάφορα υπολογιστικά εργαλεία, τα οποία χρησιμοποιούνται κατά την εκτέλεσή του, και ένα αρχείο με τις διαδρομές προς τα αρχεία με τα δείγματα εισόδου. Σημειώνεται, πως το spipeRNA απαιτεί την προεγκατάσταση του εργαλείου samtools στο σύστημα στο οποίο εκτελείται.



Σχήμα 4: Διάγραμμα ροής του εργαλείου spipeRNA

Η πρώτη φάση του αλγορίθμου περιλαμβάνει την ευθυγράμμιση των εγγραφών. Σημαντική διαφορά του spipeRNA σε σχέση με άλλους αλγορίθμους ανάλυσης των μικρών RNAs ή των miRNAs, έγκειται στο γεγονός ότι η ευθυγράμμιση πραγματοποιείται πάνω στο γονιδίωμα, και όχι στο μεταγράφωμα. Ευθυγράμμιση των εγγραφών πάνω στο μεταγράφωμα μπορεί να ελλοχεύει κινδύνους, οι οποίοι περιγράφονται στην ενότητα 6.2.1. Λεπτό σημείο της ευθυγράμμισης των μικρών σε μήκος εγγραφών, αποτελούν τα multi-mapped reads, δηλαδή οι εγγραφές από τα δεδομένα των NGS πειραμάτων, που αντιστοιχούν σε πάνω από μία περιοχές του γονιδιώματος, με την ίδια πιθανότητα. Τα multi-mapped reads είναι ένα γενικό πρόβλημα στην ευθυγράμμιση όλων των τύπων των RNA-Seq δεδομένων, ειδικά όμως, αποτελεί ιδιαίτερο πρόβλημα στα small RNA-Seq δεδομένα. Το μικρό μήκος των εγγραφών-μεταγράφων, εγγραφές με παρόμοιες αλληλουχίες και μία νουκλεοτιδική αναντιστοιχία μπορούν να οδηγήσουν σε λανθασμένη τοποθέτηση του εγγράφου πάνω στο γονιδίωμα.

Μία απλή αντιμετώπιση των παραπάνω, είναι να αγνοηθούν στο σύνολό τους όλα τα multi-mapped reads, κάτι, όμως, που είναι ισοδύναμο με την απώλεια τεραστίου όγκου πληροφορίας. Μία άλλη, απλή προσέγγιση είναι η αντιστοίχιση των multi-mapped reads, σε μία από τις πιθανές θέσεις πάνω στο γονιδίωμα, με τρόπο τυχαίο, με συνέπεια όμως τη λάθος αντιστοίχιση πολλών εγγραφών. Με τον τρόπο αυτόν, μία εγγραφή, για παράδειγμα, που εμφανίζει δύο πιθανές θέσεις πάνω στο γονιδίωμα, θα έχει 50% πιθανότητα να αντιστοιχιστεί σε λάθος θέση, μία εγγραφή με τρεις πιθανές θέσεις θα έχει 67% πιθανότητα να αντιστοιχιστεί λάθος κτλ. Άλλη λύση, για τα multi-mapped reads, είναι η καταγραφή όλων των πιθανών θέσεων για ένα multi-mapped read. Η μέθοδος αυτή είναι, όμως, επιρρεπής σε λανθασμένη ερμηνεία της γενικής ανάλυσης των αποτελεσμάτων του πειράματος, επειδή οι ευθυγραμμίσεις δεν είναι πλέον εκπρόσωποι της ποσότητας του αρχικού γενετικού υλικού και, κατά συνέπεια, θα υπερεκτιμούνται ποσοτικά κάποιες επαναλαμβανόμενες περιοχές του γονιδιώματος. [9] Μία παραλλαγή της τελευταίας μεθόδου είναι η καταγραφή του κλάσματος της εγγραφής δια του πλήθους των περιοχών, ώστε σε κάθε τέτοια περιοχή να αντιστοιχιστεί μόνο ένα μέρος της εγγραφής. Για την ευθυγράμμιση και το χειρισμό των multi-mapped εγγραφών στο spipeRNA, επιλέχθηκε ο state-of-the-art αλγόριθμος butter.

Στον αλγόριθμο spipeRNA, η κάθε ευθυγραμμισμένη εγγραφή, σχολιάζεται χωριστά, ενώ οι εγγραφές, για τις οποίες υπάρχουν πάνω από ένα μετάγραφο, σχολιάζονται βάσει σχέσης, η οποία επιλέγει το κατάλληλότερο μετάγραφο σύμφωνα με το μήκος της εγγραφής και του μετάγραφου, καθώς και το μήκος της επικάλυψης μεταξύ τους. Οι εγγραφές που ευθυγραμμίζονται στο γονιδίωμα αλλά δεν αντιστοιχούνται σε κάποιο γνωστό μετάγραφο, ελέγχονται περαιτέρω για σχηματισμούς εγγραφών οι οποίοι υποδηλώνουν έκφραση. Ο καθορισμός των περιοχών έκφρασης συντελείται με τη βοήθεια ενός κρυφού μοντέλου Markov στο οποίο οι πιθανότητες εκπομπής και μετάβασης βασίζονται σε κατανομή Poisson.

5.2.1 Ευθυγράμμιση των εγγραφών με τη βοήθεια του butter

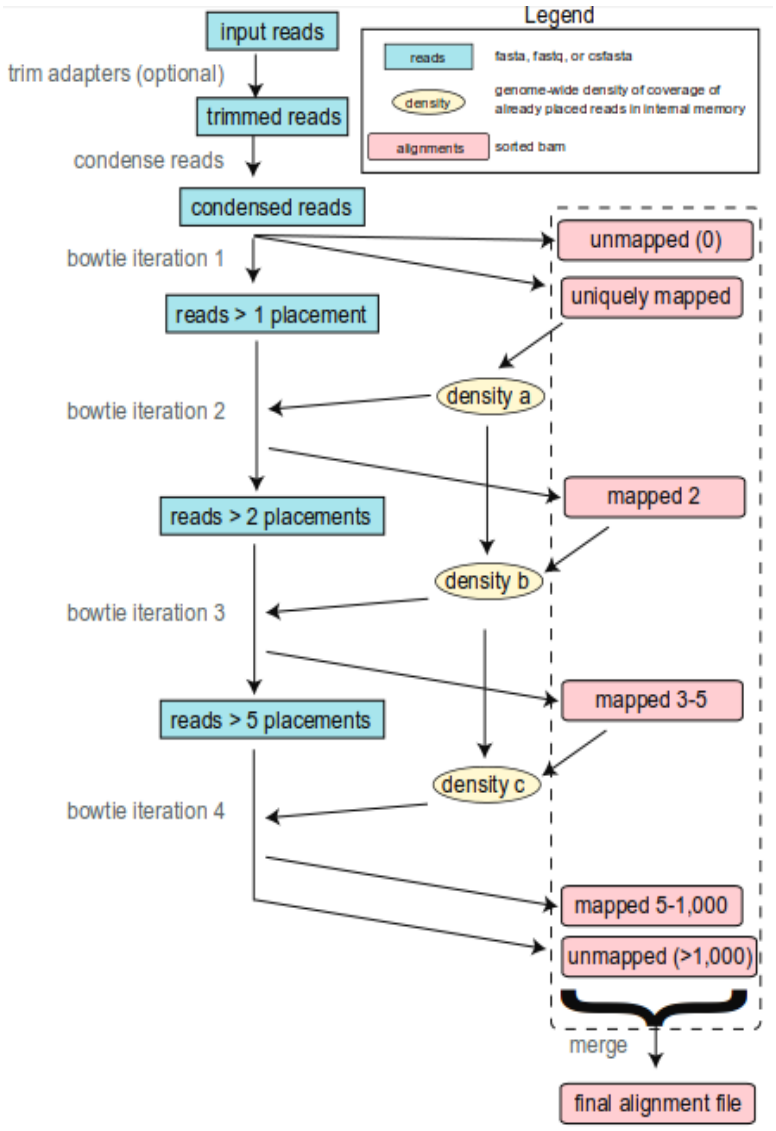
Το butter είναι ένα script γραμμένο στη γλώσσα Perl, το οποίο βασίζεται σε δύο δωρεάν διαθέσιμα εργαλεία, το samtools και το bowtie. Το σχήμα 5 περιγράφει τη γενική μεθοδολογία της ευθυγράμμισης που χρησιμοποιείται στο butter. Ο χρήστης εισάγει το γονιδίωμα αναφοράς και αρχείο προερχόμενο από small RNA-Seq πείραμα σε fasta, fastq ή csfasta μορφή. Ίδιες εγγραφές που εμφανίζονται πολλές φορές στα αρχεία εισόδου, ομαδοποιούνται σε αντιπροσώπους, με ταυτόχρονη αποθήκευση του αριθμού της εμφάνισής τους.

Όπως αναφέρθηκε πιο πάνω, οι multi-mapped εγγραφές είναι ένα γενικό πρόβλημα στα NGS δεδομένα και αποτελεί ένα ιδιαίτερο πρόβλημα στις μικρού μήκους εγγραφές. Παρακάτω, παρουσιάζεται ο αλγόριθμος με τον οποίο το εργαλείο butter χειρίζεται τις εγγραφές που αντιστοιχούν σε παραπάνω από μία περιοχές του γονιδιώματος.

Αλγόριθμος butter

Η πρώτη επανάληψη του bowtie περιορίζει την έξοδο της ευθυγράμμισης σε εγγραφές που έχουν καμία ή μία έγκυρη θέση ευθυγράμμισης πάνω στο γονιδίωμα. Εγγραφές που αντιστοιχούν σε περισσότερες από μία θέσεις πάνω στο γονιδίωμα αποθηκεύονται σε ένα προσωρινό fasta ή csfasta αρχείο, για χρήση στην επόμενη εκτέλεση. Στη συνέχεια, υπολογίζονται και αποθηκεύονται οι πυκνότητες των εγγραφών με μοναδική θέση πάνω στο γονιδίωμα, με τη χρήση της ανάλυσης του συρόμενου παραθύρου (μέγεθος=250, βήμα=50). Η πυκνότητα του κάθε παραθύρου αντιστοιχεί στο άθροισμα όλων των εγγραφών της κάθε θέσης (συνδυασμός και των δύο κλώνων). Στη δεύτερη επανάληψη του bowtie, χρησιμοποιούνται οι εγγραφές με δύο πιθανές θέσεις πάνω στο γονιδίωμα, οι οποίες αποθηκεύτηκαν στο προηγούμενο βήμα. Κάθε πιθανή περιοχή πέφτει σε πέντε διαφορετικά παράθυρα. Από τα πέντε αυτά παράθυρα, ως παράθυρο αναφοράς επιλέγεται εκείνο το οποίο εμφανίζει τη μεγαλύτερη πυκνότητα εγγραφών, η οποία υπολογίστηκε στο προηγούμενο βήμα. Αν, για παράδειγμα, η πρώτη περιοχή παρουσιάζει σκορ πυκνότητας 70 εγγραφών και η δεύτερη περιοχή έχει σκορ πυκνότητας 30 εγγραφών, η πιθανότητα να διατηρηθεί η υπό εξέταση εγγραφή για την πρώτη περίπτωση είναι 70 / (70 + 30) [πχ. 70%] και για τη δεύτερη είναι 30 / (70 + 30) [πχ. 30%]. Σε περίπτωση που όλες οι πιθανές περιοχές έχουν μηδενική πυκνότητα εγγραφών, η τελική τοποθέτηση της εγγραφής γίνεται τυχαία. Στη συνέχεια η διαδικασία επαναλαμβάνεται για δύο ακόμη επαναλήψεις. Στην τρίτη επανάληψη ελέγχονται οι εγγραφές με τρεις, τέσσερις ή πέντε πιθανές θέσεις πάνω στο γονιδίωμα, ενώ η τέταρτη επανάληψη ελέγχει εγγραφές, οι οποίες έχουν από πέντε μέχρι χίλιες πιθανές θέσεις πάνω στο γονιδίωμα.

Το τελικό αρχείο που παράγεται κατά τη φάση εκτέλεσης του butter είναι ένα ταξινομημένο αρχείο ευθυγραμμίσεων σε BAM μορφή. Κάθε εγγραφή του αρχείο εισόδου εμφανίζεται στο αρχείο εξόδου μία φορά, συμπεριλαμβανομένων και των εγγραφών χωρίς έγκυρη θέση πάνω στο γονιδίωμα.



Σχήμα 5: Διάγραμμα ροής του εργαλείου butter (Michael J. Axtell, 2014)

Εκτός από τα πεδία ευθυγράμμισης που προστίθενται από το εργαλείο bowtie, το butter προσθέτει επιπλέον τρεις προσαρμοσμένες στην εκτέλεσή του ετικέτες για κάθε ευθυγραμμισμένη εγγραφή. Η ετικέτα XX:i:[ακέραιος] δηλώνει τον αριθμό των έγκυρων θέσεων ευθυγράμμισης οι οποίες υπολογίζονται σύμφωνα με το εργαλείο bowtie. Η ετικέτα XY:Z:[συμβολοσειρά] δηλώνει τον τρόπο με τον οποίο επιλέχτηκε η τελική θέση ευθυγράμμισης της εγγραφής πάνω στο γονιδίωμα και ο οποίος φέρει μία από τις παρακάτω τιμές:

U: εγγραφή η οποία ταιριάζει σε μία μοναδική έγκυρη θέση πάνω στο γονιδίωμα αναφοράς,

P: εγγραφή με πολλαπλές έγκυρες θέσεις ευθυγράμμισης και τελική επιλεγμένη θέση λόγω ομαδοποίησης,

R: εγγραφή με πολλαπλές έγκυρες θέσεις ευθυγράμμισης και τελική επιλεγμένη θέση με τυχαίο τρόπο,

N: μη ευθυγραμμισμένη εγγραφή λόγω μηδενικής ύπαρξης έγκυρης θέσης ευθυγράμμισης πάνω στο γονιδίωμα,

M: εγγραφή με πολλαπλές έγκυρες θέσεις ευθυγράμμισης των οποίων ο αριθμός υπερβαίνει τη μεταβλητή εισόδου --max_rep, επομένως δεν εκτελείται καμία τοποθέτηση της εγγραφής

O: εγγραφή με πολλαπλές έγκυρες θέσεις ευθυγράμμισης για τις οποίες δεν υπάρχει κατανομή πυκνότητας των εγγραφών καθώς επίσης και ο αριθμός των θέσεων υπερβαίνει τη μεταβλητή εισόδου --ranmax, επομένως δεν εκτελείται καμία τοποθέτηση της εγγραφής πάνω στο γονιδίωμα.

Η ετικέτα XZ:Z:[αριθμός κινητής υποδιαστολής] δηλώνει την πιθανότητα με την οποία μία εγγραφή τοποθετείται σε μία θέση του γονιδώματος. Η πιθανότητα αυτή ισούται με τη μονάδα για τιμές M, U, N και O της ετικέτας XY:Z.

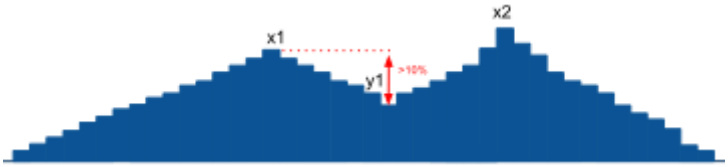
Σημειώνεται, πως το spipeRNA εκτελείται για τη μεταβλητή ranmax, ορισμένη στο 3 και για τη μεταβλητή max_rep ορισμένη στο 1000.

5.2.2 Επαναφορά των χαμένων εγγραφών

Από την ενότητα 5.2.1, λόγω της μεταβλητής ranmax ορισμένης στο 3, οι εγγραφές με πάνω από τρεις έγκυρες περιοχές ευθυγράμμισης πάνω στο γονιδίωμα, απορρίπτονται. Έτσι στο σημείο αυτό, προστίθεται ένας επιπλέον έλεγχος, όπου οι εν λόγω εγγραφές, ελέγχονται επιπρόσθετα και εάν ικανοποιείται μία από τις συνθήκες που παρουσιάζονται στη συνέχεια, η αρχικώς απορριφθείσα εγγραφή υπολογίζεται τελικά στις ncRNA εκφράσεις. Εάν υπάρχει μόνο ένα σχολιασμένο μετάγραφο που μπορεί να αντιστοιχιστεί σε μία από τις θέσεις ευθυγράμμισης, τότε η εγγραφή αντιστοιχίζεται στο μετάγραφο αυτό. Εάν υπάρχουν παραπάνω από ένα μετάγραφο που ταιριάζουν με τις πιθανές θέσεις ευθυγράμμισης της εγγραφής, τότε επιλέγεται το μετάγραφο το οποίο παρουσιάζει τη μεγαλύτερη συγκέντρωση εγγραφών από το προηγούμενο βήμα. Εάν όλα τα μετάγραφα παρουσιάζουν μηδενική πυκνότητα εγγραφών, επιλέγεται το μετάγραφο σύμφωνα με τη σχέση της ενότητας 5.2.2. Επισημαίνεται πως η multi-mapped εγγραφή απορρίπτεται, εάν δεν υπάρχει ούτε ένα μετάγραφο το οποίο να μπορεί να αντιστοιχιστεί σε κάποια από τις θέσεις ευθυγράμμισής του.

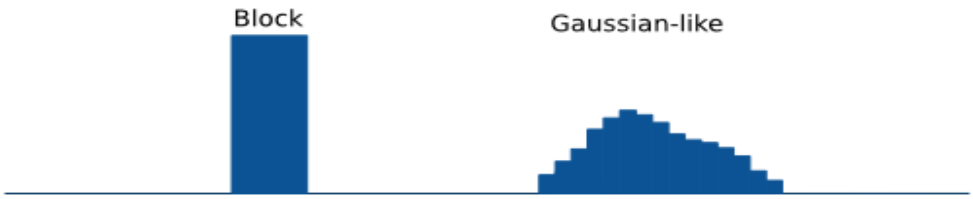
5.2.3 Εκτίμηση περιοχών έκφρασης

Εγγραφές οι οποίες δεν έχουν αντιστοιχιστεί σε σχολιασμένα γονίδια, υπάγονται στην περαιτέρω ανάλυση με σκοπό την εύρεση των περιοχών εμπλουτισμένης έκφρασης. Η εικόνα 5 παρουσιάζει περίπτωση κατανομής των NGS εγγραφών, στην οποία δεν είναι ευδιάκριτο εάν θα πρέπει οι δύο σχηματισμένες κορυφές να συνενωθούν ή να διαχωριστούν και πού πρέπει να οριστεί η αρχή και το τέλος των περιοχών έκφρασης.



Εικόνα 3: Κατανομή NGS εγγραφών πάνω στο γονιδίωμα (pyicos, 2016, διαθέσιμη στο: http://regulatorygenomics.upf.edu/Software/Pyicoteo/_images/Split.png)

Στην Εικόνα 6 ο σχηματισμός εγγραφών στα αριστερά (μπλοκ εγγραφών) μοιάζει με σφάλμα αλληλούχησης, ενώ η γκαουσιανή κατανομή στα δεξιά, αποτελεί μία πιθανή περιοχή έκφρασης.



Εικόνα 4: Κατανομή NGS εγγραφών πάνω στο γονιδίωμα (pyicos, 2016, διαθέσιμη στο: http://regulatorygenomics.upf.edu/Software/Pyicoteo/_images/Artifact.png)

Τα παραπάνω αποτελούν κάποιους από τους χαρακτηριστικούς προβληματισμούς στον ορισμό και τη διαχείριση των σχηματισμένων από εγγραφές κορυφών. Η λύση πρέπει να αναζητηθεί σε μαθηματικά και στατιστικά μοντέλα, που προσπαθούν να ορίσουν τα κατάλληλα για το κάθε δείγμα ή μία ευρύτερη γονιδιακή περιοχή φράγματα, τα οποία να επιτρέπουν έναν ορθότερο χειρισμό τέτοιων σχηματισμών.

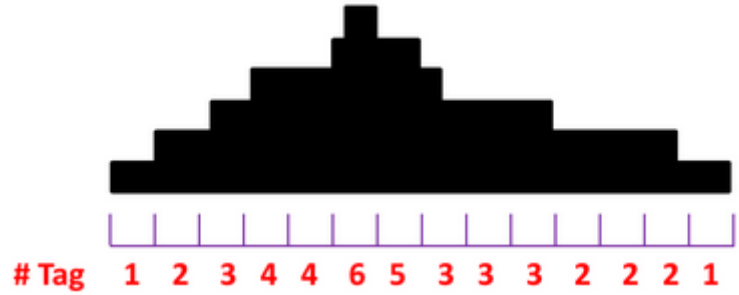
Για τον εντοπισμό των πιθανών εμπλουτισμένων περιοχών από εγγραφές, οι οποίες δεν εμπίπτουν στην κατηγορία των γνωστών μεταγράφων, το spipeRNA χρησιμοποιεί το πακέτο MiClip της γλώσσας R. Αν και το πακέτο αυτό προορίζεται κυρίως για τα CLIP-Seq δεδομένα, μπορεί να χρησιμοποιηθεί και για άλλους τύπους δεδομένων, όπως είναι τα small RNA-Seq ή RNA-Seq δεδομένα, λόγω ενός γενικευμένου μαθηματικού μοντέλου, στο οποίο βασίζεται.

Με τη χρήση του, MiClip, εκτελούνται δύο γύροι Hidden Markov Model (HMM) για να εντοπιστούν οι εμπλουτισμένες περιοχές. Σε αυτήν τη διαδικασία, οι διπλότυπες εγγραφές (duplicate reads), δηλαδή οι εγγραφές με ακριβώς ίδιες συντεταγμένες ευθυγράμμισης, ομαδοποιούνται σε μοναδική ετικέτα (tag). Ετικέτες οι οποίες επικαλύπτονται με ένα τουλάχιστον νουκλεοτίδιο, ομαδοποιούνται σε συστάδες (clusters), ενώ αυτές που δεν επικαλύπτονται με άλλες ετικέτες, απορρίπτονται.

Εντοπισμός των εμπλουτισμένων περιοχών - πρώτος γύρος HMM

Για την αναγνώριση των εμπλουτισμένων περιοχών, οι συστάδες διαιρούνται σε κάδους (bins) των 5 βάσεων. Εάν $x_1^{(k)}$ είναι ο συνολικός αριθμός των ετικετών (ουσιαστικά των εγγραφών σε μία ορισμένη θέση) στον t-κάδο της k συστάδας, η συστάδα k μπορεί να αναπαρασταθεί ως μία σειρά από πλήθη ετικετών

$$\vec{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_{T_k}^{(k)})$$



Εικόνα 5: Το «# Tag» αναφέρεται στο στρογγυλοποιημένο αριθμό των ετικετών που καλύπτουν τον κάθε κάδο (Tao Wang, 2014)

Στην Εικόνα 7 παρατηρούμε μία συστάδα χωρισμένη σε κάδους των πέντε βάσεων, όπου στον πρώτο κάδο, υπάρχει μία ετικέτα, δηλαδή μία εγγραφή ή ένα read, στο δεύτερο κάδο υπάρχουν δύο ετικέτες, στον τρίτο τρεις κτλ.

Για να προσδιοριστούν οι εμπλουτισμένες περιοχές από το πλήθος των παρατηρηθέντων ετικετών, χρησιμοποιείται HMM με τις δύο εξής καταστάσεις:

$$\begin{cases} I_t^{(k)} = 0, \text{εάν ο κάδος δεν είναι εμπλουτισμένος} \\ I_t^{(k)} = 1, \text{εάν ο κάδος είναι εμπλουτισμένος} \end{cases}$$

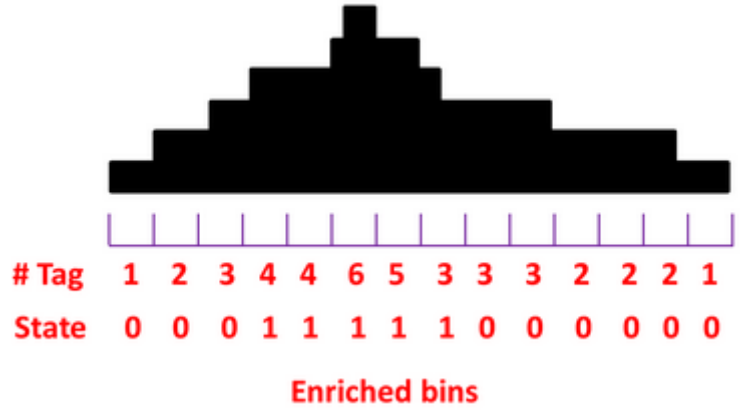
Μία δημοφιλής μέθοδος μοντελοποίησης της κατανομής των ετικετών κατά μήκος του γονιδιώματος πραγματοποιείται με τη βοήθεια της κατανομής Poisson. Δεδομένης της κατάστασης $I_t^{(k)}$, το παρατηρηθέν πλήθος των ετικετών μοντελοποιήθηκε με μεικτό Poisson μοντέλο δύο συνιστωσών:

$$\begin{cases} X_t^{(k)} \sim Poisson(\lambda_0) | I_t^{(k)} = 0 \\ X_t^{(k)} \sim Poisson(\lambda_1) | I_t^{(k)} = 1 \end{cases}$$

Η πιθανότητα εκπομπής μπορεί να γραφτεί ως

$$Pr(X_t^{(k)} = x \mid \lambda_0, \lambda_1, \omega) = (1-\omega)\frac{\lambda_0^x \cdot e^{-\lambda_0}}{x!} + \omega\frac{\lambda_1^x \cdot e^{-\lambda_1}}{x!}, (\lambda_0 < \lambda_1)$$

όπου το ω είναι το ποσοστό των εμπλουτισμένων κάδων στις συστάδες που δημιουργούνται. Η μήτρα μετάβασης Π είναι ένας $2x2$ πίνακας, όπου $\Pr(I_t^{(k)} = s \mid I_{t-1}^{(k)} = r)$ είναι η πιθανότητα εκπομπής. Οι παράμετροι λ_0, λ_1 και ω εκτιμώνται από τα δεδομένα εισόδου με τη μέθοδο των στιγμών, ενώ οι κρυφές καταστάσεις $I_t^{(k)}$, δηλαδή οι εμπλουτισμένοι και μη κάδοι, υπολογίζονται με τον αλγόριθμο Viterbi. Ο αλγόριθμος Viterbi υπολογίζει την κρυφή κατάσταση του κάθε κάδου, σύμφωνα με το κριτήριο ότι η posterior πιθανότητα του κάθε κάδου σε μία από τις δύο καταστάσεις των κάδων, δηλαδή του εμπλουτισμένου ή όχι κάδου, θα πρέπει να είναι μεγαλύτερη από την posterior πιθανότητα του κάδου αυτού να είναι σε άλλη κατάσταση, δεδομένου του μοντέλου και της παρατήρησης. Τελικά, οι γειτονικοί εμπλουτισμένοι κάδοι συνενώνονται σε εμπλουτισμένες περιοχές.



Εικόνα 6: Πρώτος γύρος HMM. Το «# Tag» αναφέρεται στο στρογγυλοποιημένο αριθμό των ετικετών που καλύπτουν τον κάθε κάδο και το «State» γνωστοποεί εάν ο κάδος είναι εμπλουτισμένος. (Tao Wang, 2014)

Η Εικόνα 8 αναπαριστά το αποτέλεσμα του πρώτου γύρου HMM. [10]

Από την εκτέλεση του πρώτου HMM του πακέτου MiClip, επιλέγονται μόνο εκείνες οι περιοχές οι οποίες είναι επισημασμένες ως «enriched», δηλαδή οι εμπλουτισμένες κορυφές. Το αποτέλεσμα της εκτέλεσης του πρώτου HMM αποθηκεύεται σε ένα αρχείο με την παρακάτω μορφή.

region_id	chr	strand	start	end	enriched	sites	count	rpm
181	2	-	47280755	47280794	TRUE	FALSE	27	2.965223203
254	2	-	64875965	64876014	TRUE	FALSE	29	3.184869367
277	2	-	71046390	71046424	TRUE	FALSE	46	5.051861754
405	2	-	114937535	114937584	TRUE	FALSE	115	12.62965438
460	2	-	132254380	132254444	TRUE	FALSE	94	10.32336967
461	2	-	132254465	132254504	TRUE	FALSE	66	7.248323386
463	2	-	132254645	132254699	TRUE	FALSE	137	15.04576218
464	2	-	132254700	132254789	TRUE	FALSE	103	11.31177741
465	2	-	132254800	132254859	TRUE	FALSE	51	5.600977162
466	2	-	132254865	132254944	TRUE	FALSE	143	15.70470067

Εικόνα 7: Δείγμα αρχείου εμπλουτισμένων περιοχών

Τα χαρακτηριστικά των κορυφών συμπεριλαμβάνουν, μεταξύ άλλων, το χρωμόσωμα, τον κλώνο, την αρχή και το τέλος τη κορυφής, καθώς και το εκτιμώμενο πλήθος των εγγραφών από το NGS αρχείο.

Αλλαγή σε αρχεία του πακέτου MiClip

Μία μικρή αλλαγή, που καλό είναι να πραγματοποιηθεί στο πακέτο MiClip, πριν τη χρήση του στο εργαλείο spipeRNA, αφορά τα αρχεία «cluster.pl» και «cluster_p.pl». Τα αρχεία αυτά επεξεργάζονται τα διάφορα χαρακτηριστικά των γραμμών του αρχείου εισόδου SAM, και τα ομαδοποιούν κατά τη φάση της εκτέλεσης με το διαχωριστικό «_», με το οποίο στη συνέχεια της εκτέλεσης τα διαχωρίζουν. Ο συγκεκριμένος χαρακτήρας που χρησιμοποιείται μπορεί να προκαλέσει πρόβλημα όταν συμπεριλαμβάνεται σε κάποιο στοιχείο της γραμμής. Ένα χαρακτηριστικό παράδειγμα είναι όνομα χρωμοσώματος όπως το CHR_HSCHR17_5_CTG4 της γονιδιωματικής έκδοσης GRCh38.85.

Συγκεκριμένα, λοιπόν, για την έκδοση MiClip 1.3, στο αρχείο «cluster.pl», στις γραμμές 47, 65 και 76, ο χαρακτήρας «_» θα αλλαχτεί σε «|», ενώ στο αρχείο «cluster_p.pl», την ίδια αλλαγή θα έχουν οι γραμμές 33, 41, 52.

5.2.4 Σχολιασμός Περιοχών Έκφρασης

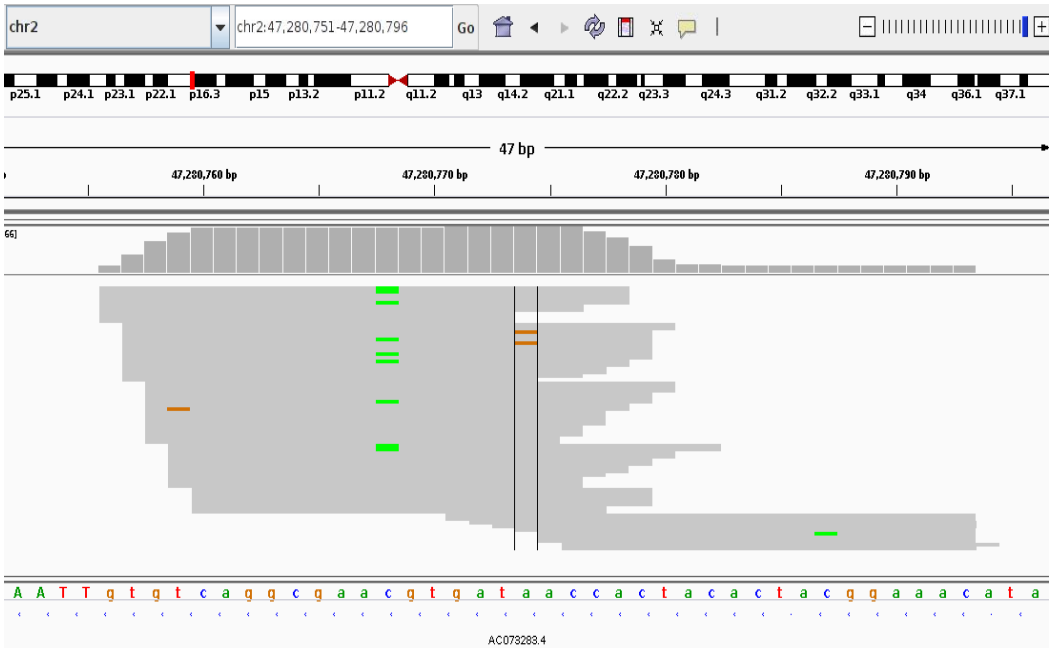
Το τελευταίο στάδιο του pipeline περιλαμβάνει την ονοματολογία-σχολιασμό των κορυφών ή εμπλουτισμένων περιοχών που βρέθηκαν στο προηγούμενο βήμα. Για το σκοπό αυτό, χρησιμοποιείται αρχείο σχολιασμένων γονιδιακών συντεταγμένων, που περιλαμβάνουν τα κωδικά και μη γονίδια. Ο λόγος για τον οποίον ελέγχονται τα κωδικά γονίδια είναι, επειδή υπάρχει περίπτωση οι NGS εγγραφές των small RNA-Seq πειραμάτων, να προέρχονται, για παράδειγμα, από θραύσματα μεγαλύτερων μεταγράφων mRNAs, τα οποία διασπώνται μέσα στο κύτταρο από κάποια μετα-μεταγραφική διαδικασία.

Η αντιστοίχιση της εκτιμώμενης περιοχής και του αντίστοιχου γονιδίου, πραγματοποιείται με τρόπο παρόμοιο με αυτόν που παρουσιάστηκε στην ενότητα 5.2.2 με την εξαίρεση ότι, το ποσοστό κάλυψης της εκτιμώμενης περιοχής και του σχολιασμένου γονιδίου πρέπει να είναι ίσο ή μεγαλύτερο από 80%. Παράδειγμα αρχείου εμπλουτισμένων περιοχών έκφρασης για τις οποίες έχει αναζητηθεί ο κατάλληλος σχολιασμός, παρουσιάζεται στην Εικόνα 10.

Chromosome	Strand	Start Loci	Stop Loci	Read Counts	Transcript ID	Transcript Name	Region Length	Sequence
2	-	47280755	47280794	25	ENSG00000234690	AC073283.4	39	TGTGTCAGGCGAACGTGATAACCACTACACTACGGAAACA
2	-	64875960	64876014	27	no_annotation	no_annotation	54	AGTTGGCCACTGTCCTGCTGTCCATATCAACCAACACCTTTTCTGGGATCTGAT
2	-	71046390	71046424	45	no_annotation	no_annotation	34	TGGATGAGAGCGCCGAATCCTAACCACTAGACCGC
2	+	78701590	78701704	146	no_annotation	no_annotation	114	TATCCTCGGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGAGCTTGACTC
2	+	78701750	78701784	42	no_annotation	no_annotation	34	GCCGGTGTAATACCACTACTCTGATCGTTTTTTTT
2	-	114937535	114937584	112	no_annotation	no_annotation	49	TAAGGTTTCACGCCCTCTTGAACTCTCTCTTCAAAGTTCTTTCCAACTTT
2	-	132254380	132254439	103	no_annotation	no_annotation	59	TTTGTTACGACTTTTACTTCCTCTAGATAGTCAAGTTCGACTGTCTTCTCAGCG
2	-	132254465	132254499	65	no_annotation	no_annotation	34	CTGATCCGAGGGCCTCACTAAACCATCCAATCAGG
2	-	132254645	132254789	137	no_annotation	no_annotation	144	CATAGGGTAGGCACACGCTGAGCCAGTCAGTGTAGCGCGCGTGCAGCCCTG
2	-	132254800	132254859	46	no_annotation	no_annotation	59	AACTAGTTGGCATGCCAGAGTCTCGTTCGTTATTGGAATTAGCCAGACAAATC

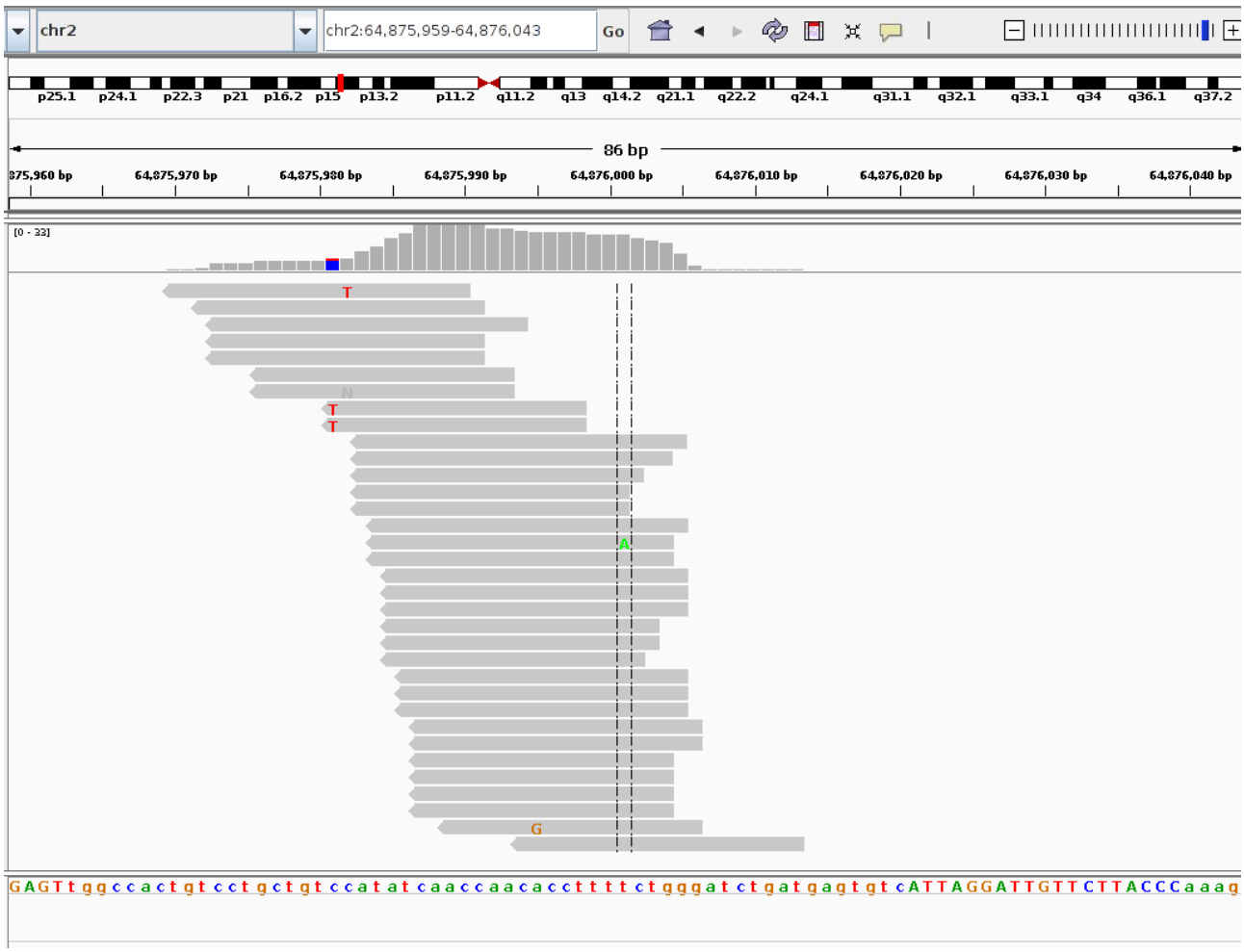
Εικόνα 8: Μέρος αρχείου με τις σχολιασμένες περιοχές εμπλουτισμένης έκφρασης

Παρατηρεί κανείς, για παράδειγμα, ότι η πρώτη περιοχή έκφρασης του παραπάνω αρχείου, αντιστοιχεί στο γονίδιο δύο του μεταγραφόμενου κλώνου «-», στη νουκλεοτιδική περιοχή «47.280.755-47.280.794». Αυτή η περιοχή έκφρασης έχει αντιστοιχιστεί με το γονίδιο AC073283.4-005 με το αναγνωριστικό ENSG00000234690. Η περιοχή αυτή αναπαραστάθηκε στον περιηγητή IGV όπως φαίνεται στην εικόνα 11.



Εικόνα 9: Μεγέθυνση της γονιδιακής περιοχής chr2:47.280.755-47.280.794, όπως παρίσταται στον περιηγητή IGV. Στην περιοχή αυτή διακρίνονται οι ευθυγραμμισμένες στο γονιδίωμα εγγραφές οι οποίες σχηματίζουν στατιστικά σημαντική κορυφή.

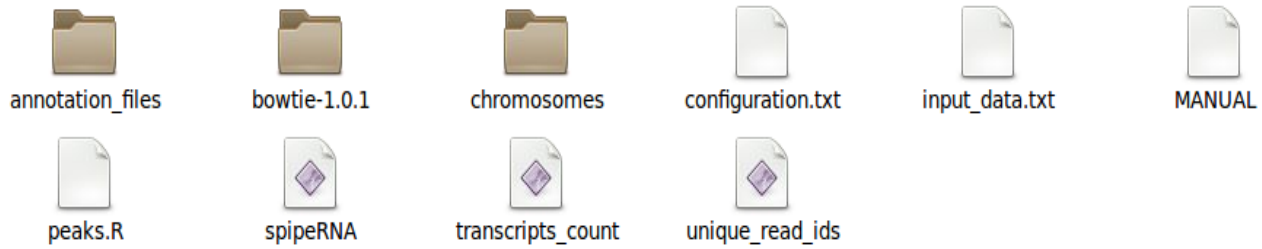
Για τη τη συγκεκριμένη περιοχή έκφρασης, υπάρχει πράγματι ένα σχολιασμένο μετάγραφο AC073283.4, σύμφωνα με τη βάση refGene. Σημειώνεται επίσης, ότι στο παραχθέν αρχείο εξόδου, οι εκτιμώμενες περιοχές έκφρασης χωρίς υπάρχοντα σχολιασμό, σημειώνονται με το χαρακτηριστικό no_annotation. Στην εικόνα 12 παρουσιάζεται μία περιοχή του γονιδιώματος για την οποία υπάρχει στατιστικά σημαντική, σχηματισμένη από εγγραφές κορυφή, για την οποία είτε δεν υπήρξε ο κατάλληλος σχολιασμός στο αρχείο με τα σχολιασμένα γονίδια, είτε δεν υπάρχει κανένας σχολιασμός της περιοχής αυτής.



Εικόνα 10: Μεγέθυνση της γονιδιακής περιοχής chr2:64,875,965-64,876,014, όπως παρίσταται στον περιηγητή IGV. Στην περιοχή αυτή διακρίνονται οι ευθυγραμμισμένες στο γονιδίωμα εγγραφές οι οποίες σχηματίζουν στατιστικά σημαντική κορυφή.

5.3 Περιεχόμενα

Η εικόνα 13 παρουσιάζει το περιεχόμενο του πακέτου spipeRNA.



Εικόνα 11: Το περιεχόμενο του εργαλείου spipeRNA

Ο φάκελος annotation_files περιέχει δύο αρχεία με σχολιασμένα μετάγραφα. Το πρώτο αρχείο αφορά τα μη-κωδικά μετάγραφα και το δεύτερο αρχείο όλα τα διαθέσιμα μετάγραφα για τον οργανισμό που μελετάται. Ο φάκελος bowtie-1.0.1 περιέχει το

πρόγραμμα bowtie και butter, ενώ ο άδειος φάκελος chromosomes, θα πρέπει να περιέχει τα fasta και index αρχεία των χρωμοσωμάτων από το γενικό fasta αρχείο του γονιδιώματος αναφοράς, όταν εκτελεστεί το script chromosomes_creation.

Στο αρχείο ρυθμίσεων configuration.txt, ο χρήστης οφείλει να ορίσει τα παρακάτω:

1. Θέση του εργαλείο butter

2. Θέση του bowtie index

3. Θέση του αρχείο με τις γνωστές σχολιασμένες μη κωδικές αλληλουχίες

4. Θέση του αρχείο με τις γνωστές σχολιασμένες γονιδιακές αλληλουχίες

5. Θέση του φακέλου με τα αρχεία των χρωμοσωμάτων

6. Αριθμός των διεργασιών

7. Αριθμός των πυρήνων που χρησιμοποιούνται στην ευθυγράμμιση των εγγραφών

Παρακάτω παρουσιάζεται το αρχείο config.txt όπως δίνεται προς χρήση.

Path_to_butter=./bowtie-1.0.1

Path_to_bowtie_index=/path_to_bowtie_index

Path_to_merged_non_coding_annotation_file=./annotation_files/merged_ncRNA_annotation.gtf

Path_to_general_annotation_file=./annotation_files/Homo_sapiens.GRCh38.85.gtf

Path_to_chromosomes_directory=./chromosomes

Number_of_processes=1

Number_of_alignment_cores=1

Τα αρχεία με τα σχολιασμένα μη κωδικά γονίδια και το γενικευμένο αρχείο με τις γνωστές σχολιασμένες γονιδιακές αλληλουχίες, πρέπει να έχουν τα εξής γονιδιακά χαρακτηριστικά, χωρισμένα με το διαχωριστικό 'tab': χρωμόσωμα, κλώνος, αρχή και τέλος μεταγραφής, τύπος μικρού μορίου RNA, χαρακτηριστικό μεταγραφής και όνομα μεταγραφής. Παράδειγμα αρχείου μη κωδικών μεταγράφων παρουσιάζεται παρακάτω.

Chromosome	Strand	Start Loci	End Loci	Small RNA Type	Transcript ID	Transcript Name
19	+	53701267	53701287	miRNA	MIMAT0002843	hsa-miR-520b
18	+	42270589	42270715	rRNA	ENSG00000253040	RNA5SP454-201
21	-	16035413	16035509	snRNA	ENSG00000252273	RNU6-426P-201
21	-	17454789	17454859	tRNA	UCSC	tRNA-Gly-GCC-1-5
13	-	107320895	107320963	snoRNA	ENSG00000201847	SNORD31.1-201

Εικόνα 12: Υποχρεωτική μορφή των αρχείων με τα σχολιασμένα μετάγραφα.

Ο λόγος, που χρησιμοποιείται η συγκεκριμένη μορφή, είναι επειδή η συλλογή των σχολιασμένων γονιδίων προέρχεται συνήθως από διαφορετικές βάσεις δεδομένων, οι οποίες δεν ακολουθούν απαραίτητα ένα κοινό πρότυπο αναπαράστασης των δεδομένων τους. Το spipeRNA παρέχει έτοιμα αρχεία με σχολιασμένα κωδικά και μη RNAs, τα οποία έχουν αντληθεί από βάσεις Ensembl, miRBase και UCSC, όλα σε έκδοση GRCh38.

Στο αρχείο input_data.txt, ο χρήστης εισάγει τα μονοπάτια προς τα αρχεία εισόδου που επιθυμεί να αναλύσει.

5.4 Εγκατάσταση και χρήση

Το spipeRNA, απαιτεί την προεγκατάσταση του εργαλείου samtools, ενώ η προεγκατάσταση του πακέτου MiClip της R, είναι προαιρετική. Η εντολή εκτέλεσης του εργαλείου spipeRNA, παρουσιάζεται παρακάτω.

```
>./spipeRNA configuration.txt input_file.txt
```

Πριν την εκτέλεση της παραπάνω εντολής, εάν ο χρήστης επιθυμεί να προσθέσει τις νουκλεοτιδικές αλληλουχίες στις καινούργιες περιοχές έκφρασης που υπολογίζονται κατά τη φάση εκτέλεσης του spipeRNA (με εγκατεστημένο το πακέτο MiClip), θα πρέπει να εκτελέσει το script chromosome_creation με παράμετρο εισόδου, το μονοπάτι προς το fasta αρχείο του γονιδιώματος αναφοράς. Παράδειγμα εκτέλεσης του chromosome_creation παρουσιάζεται παρακάτω.

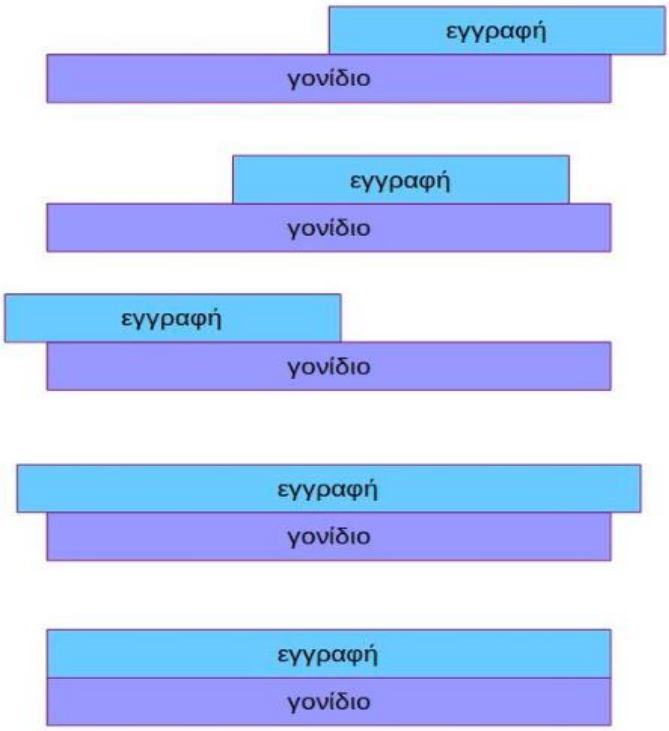
```
>./chromosome_creation genome.fa
```

Το spipeRNA εκτελείται σε συστήματα Linux/Unix.

5.4.1 Ποσοτικοποίηση των γνωστών μικρών RNAs

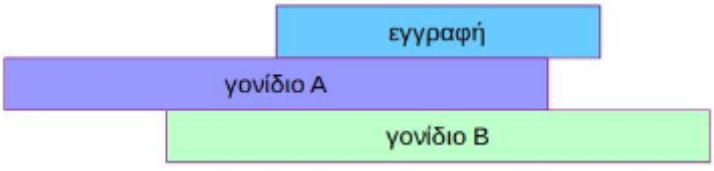
Με τη βοήθεια του εργαλείου samtools, το παραχθέν αρχείο BAM του προηγούμενου βήματος, μετατρέπεται σε μία πιο ευανάγνωστη μορφή SAM. Το SAM αρχείο, μεταξύ

άλλων, περιέχει τη γονιδιακή ακολουθία που έχει ευθυγραμμιστεί (ή που δεν έχει ευθυγραμμιστεί) πάνω στο γονιδίωμα, το χρωμόσωμα και τον κλώνο του DNA στον οποίο αυτή αντιστοιχεί και την περιοχή αρχής πάνω στο χρωμόσωμα στο οποίο ανήκει. Ο σκοπός του βήματος αυτού είναι να ευρεθούν οι περιοχές που αντιστοιχούν σε μη κωδικές επισημασμένες περιοχές του γονιδιώματος, ή πιο απλά να βρεθούν ποια και πόσα μικρά RNAs εκφράζονται στο εκάστοτε δείγμα. Για το σκοπό αυτό συντελείται ο έλεγχος της κάθε ευθυγραμμισμένης εγγραφής του SAM αρχείου, για πιθανή αντιστοιχία σε σχολιασμένη μη κωδική περιοχή. Στην εικόνα 3 παρουσιάζονται οι πέντε περιπτώσεις επικάλυψης μίας εγγραφής ευθυγραμμισμένης πάνω στο γονιδίωμα και ενός σχολιασμένου γονιδίου.



Εικόνα 13: Τρόποι επικάλυψης μίας εγγραφής και ενός γονιδίου

Εάν ικανοποιείται μία από τις συνθήκες της εικόνας 3 και επιπλέον, η επικάλυψη της εγγραφής και του γονιδίου είναι μεγαλύτερη από κατώφλι το οποίο έχει οριστεί στο 20%, τότε η συγκεκριμένη εγγραφή αντιστοιχίζεται στο συγκεκριμένο γονίδιο. Υπάρχει και περίπτωση, όπου μία εγγραφή μπορεί να αντιστοιχιστεί σε παραπάνω από ένα γονίδιο. Παρακάτω, παρουσιάζεται ένα παράδειγμα εγγραφής για την οποία υπάρχουν δύο σχολιασμένα γονίδια.



Εικόνα 14: Εγγραφή με δύο έγκυρους γονιδιακούς σχολιασμούς

Η επιλογή του γονιδίου στο οποίο θα αντιστοιχιστεί μία τέτοια εγγραφή, βασίζεται στην παρακάτω σχέση

$$\alpha\nu\alpha\lambda o\gamma\acute{\iota}\alpha = \left| 1 - \frac{\mu\acute{\eta}\kappa o\varsigma\_\varepsilon\gamma\gamma\rho\alpha\phi\acute{\eta}\varsigma + \mu\acute{\eta}\kappa o\varsigma\_\mu\varepsilon\tau\alpha\gamma\rho\acute{\alpha}\phi o\upsilon}{\frac{2}{\varepsilon\pi\iota\kappa\acute{\alpha}\lambda\upsilon\psi\eta}} \right| .$$

Η επικάλυψη που εμφανίζεται στον παρανομαστή, αντιστοιχεί στον αριθμό των κοινών νουκλεοτιδίων μεταξύ της εγγραφής και της σχολιασμένης περιοχής. Για κάθε πιθανή σχολιασμένη περιοχή, υπολογίζεται η παραπάνω σχέση και επιλέγεται εκείνο το γονίδιο που παρουσιάζει τη μικρότερη αναλογία. Μία σπάνια περίπτωση εμφανίζεται όταν υπάρχουν δύο ή περισσότερα γονίδια με την ακριβώς ίδια αναλογία. Τότε, επιλέγεται γονίδιο με το μεγαλύτερο πλήθος αντιστοιχισμένων εγγραφών, ενώ, στην περίπτωση ισάριθμων πληθών, το γονίδιο επιλέγεται τυχαία.

Μία δεύτερη σχέση, η οποία ελέγχθηκε και η οποία δίνει ακριβώς τα ίδια αποτελέσματα με την πρώτη, είναι η παρακάτω.

$$. \ \alpha\nu\alpha\lambda o\gamma\acute{\iota}\alpha = \left| 1 - \frac{\mu\acute{\eta}\kappa o\varsigma\_\varepsilon\gamma\gamma\rho\alpha\phi\acute{\eta}\varsigma}{\varepsilon\pi\iota\kappa\acute{\alpha}\lambda\upsilon\psi\eta} \right| + \left| 1 - \frac{\mu\acute{\eta}\kappa o\varsigma\_\mu\varepsilon\tau\alpha\gamma\rho\acute{\alpha}\phi o\upsilon}{\varepsilon\pi\iota\kappa\acute{\alpha}\lambda\upsilon\psi\eta} \right|$$

Στο βήμα αυτό, πέρα από την ποσοτικοποίηση των μεταγράφων, όσες εγγραφές δεν αντιστοιχούν σε σχολιασμένα μη κωδικά γονίδια, αποθηκεύονται σε αρχείο SAM για μελλοντική χρήση.

6. ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην παρούσα ενότητα θα γίνει ανασκόπηση της μορφής των αρχείων εξόδου του εργαλείου spipeRNA καθώς και μερικές μετρικές και αποτελέσματα από την εκτέλεσή του.

Το πρώτο κατά σειρά αρχείο, που δημιουργείται κατά την εκτέλεση του εργαλείου spipeRNA, είναι το αρχείο με τις ευθυγραμμισμένες και μη εγγραφές του εκάστοτε δείγματος. Οι εγγραφές αυτές προέρχονται από την εκτέλεση του εργαλείου butter και στη συνέχεια από αυτές επιλέγονται μόνο εκείνες, οι οποίες παρουσιάζουν έγκυρη ευθυγράμμιση πάνω στο γονιδίωμα αναφοράς. Έτσι, στη φάση αυτή, παράγονται δύο αρχεία εξόδου, ένα αρχείο το οποίο φέρει το παρακάτω χαρακτηριστικό όνομα: «sampleID_smallRNA_counts.tsv», και το οποίο περιέχει τα μετρημένα μη κωδικά RNAs από το δείγμα με το χαρακτηριστικό sampleID, και ένα δεύτερο αρχείο με το χαρακτηριστικό όνομα: «sampleID_no_annotation_coverage.sam», το οποίο περιέχει τις έγκυρες ευθυγραμμισμένες εγγραφές, χωρίς όμως τον αντίστοιχο σχολιασμό και οι οποίες υποβάλλονται στην περαιτέρω ανάλυση. Στην εικόνα 15 παρουσιάζεται ένα μέρος αρχείου, με τα μετρημένα μη κωδικά μετάγραφα, το οποίο προέρχεται από ένα δείγμα υγιούς νεφρού.

Transcript ID	Biotype	Transcript Name	Count	RPM
ENSG00000207261	snRNA	RNU6-738P-201	1	0.109
UCSC	tRNA	tRNA-Val-TAC-1-2	57	6.225
ENSG00000212040	miRNA	MIR543-201	1	0.109
ENSG00000234492	lincRNA	RPL34-AS1-003	1	0.109
MIMAT0004556	miRNA	hsa-miR-10b-3p	240	26.21
ENSG00000278048	snRNA	U2.16-201	304	33.2
ENSG00000272688	lincRNA	RP13-270P17.3-001	1	0.109
ENSG00000221500	snoRNA	SNORD100-201	110	12.01
ENSG00000263934	snoRNA	SNORD3A-202	3657	399.4
ENSG00000274309	snoRNA	SNORA71E-201	5	0.546
ENSG00000251632	lincRNA	RP11-714L20.1-001	1	0.109
MIMAT0025848	miRNA	hsa-miR-6511b-3p	3	0.328
ENSG00000224023	lincRNA	FLJ37035-004	1	0.109
MIMAT0004507	miRNA	hsa-miR-92a-1-5p	35	3.822
MIMAT0027578	miRNA	hsa-miR-6838-5p	2	0.218
MIMAT0019867	miRNA	hsa-miR-4738-3p	1	0.109
ENSG00000259438	lincRNA	CTD-2650P22.1-001	1	0.109
ENSG00000233864	lincRNA	TTTY15-001	2	0.218
MIMAT0000267	miRNA	hsa-miR-210-3p	3127	341.5

Εικόνα 15: Μέρος αρχείου με τα εκφρασμένα μη κωδικά γονίδια το οποίο παράγεται κατά τη φάση της εκτέλεσης του εργαλείου spipeRNA

Το αρχείο αυτό αποτελείται από τις εξής πέντε στήλες: το αναγνωριστικό του μεταγράφου, το βιότυπο του μεταγράφου, το όνομα του μεταγράφου, το πλήθος των εγγραφών που αντιστοιχούν στο συγκεκριμένο μετάγραφο, και τέλος το RPM (Reads Per Milion) του πλήθους αυτού. Το RPM υπολογίζεται σύμφωνα με την παρακάτω σχέση:

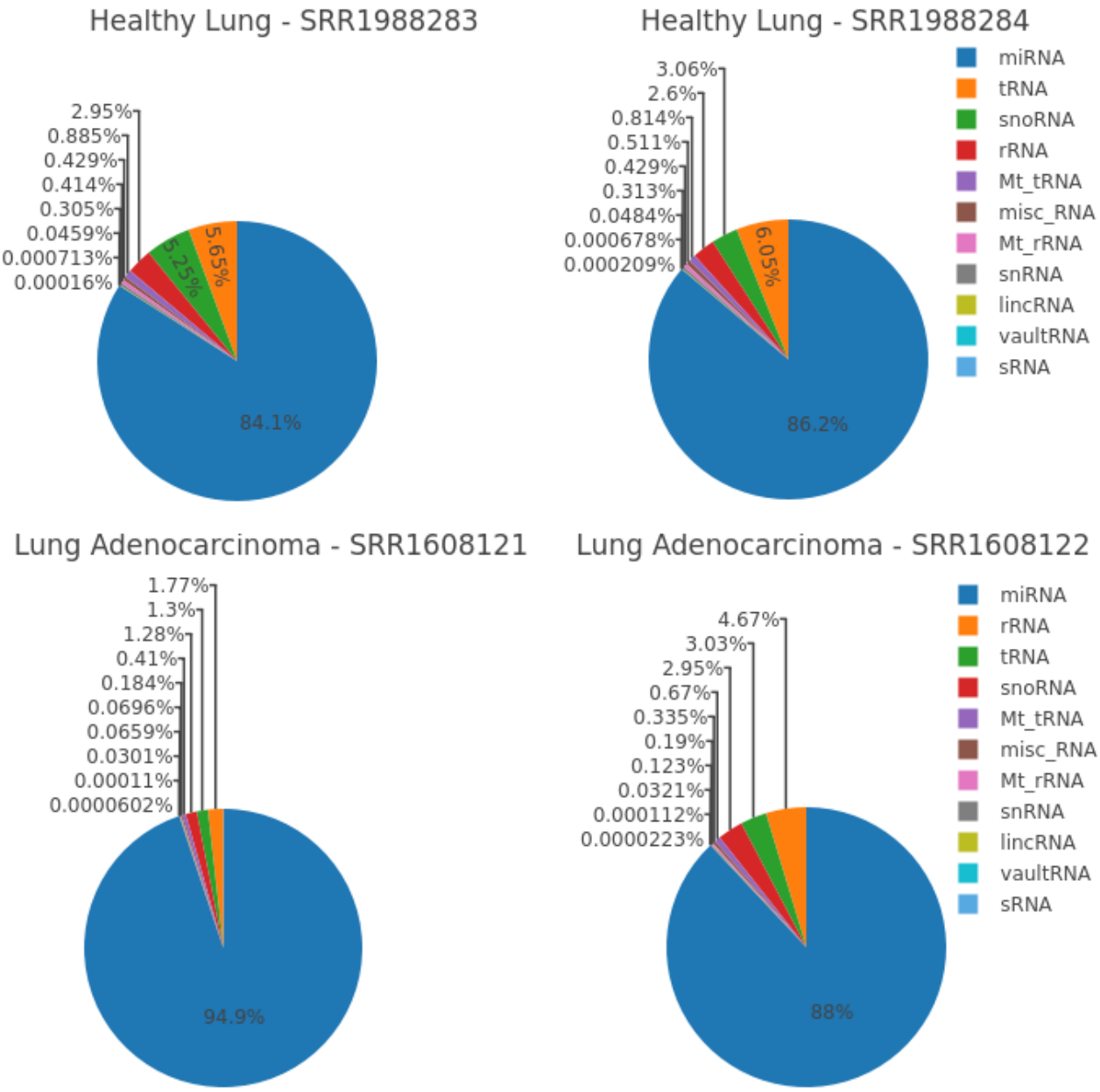
$$RPM = \frac{\textit{πλήθος εγγραφών που αντιστοιχούν σε συγκεκριμένο μετάγραφο}}{\textit{πλήθος όλων των εγγραφών}} \cdot 1000000$$

Το πλήθος όλων των εγγραφών του παρανομαστή, αντιστοιχεί σε όλες τις εγγραφές που ευθυγραμμίζονται πάνω στο γονιδίωμα. Η μετρική του RPM επιτρέπει τη σύγκριση των πληθών των μεταγράφων μεταξύ διαφορετικών δειγμάτων, με διαφορετικό βάθος αλληλούχησης.

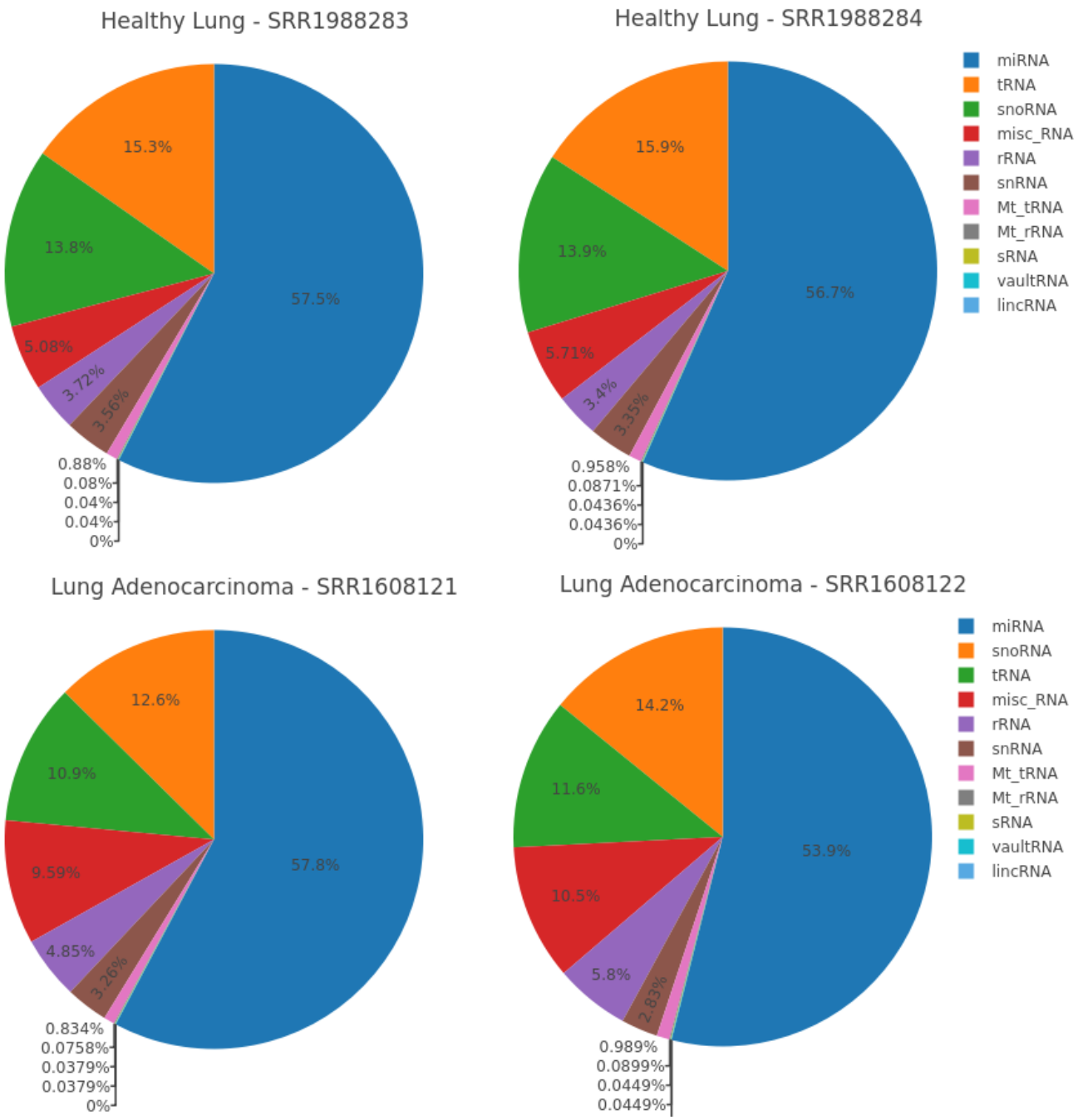
Το δεύτερο κατά σειρά αρχείο που παράγεται κατά τη φάση της εκτέλεσης του spipeRNA, είναι το αρχείο με τις εμπλουτισμένες περιοχές έκφρασης, για τις οποίες δε βρέθηκε ο κατάλληλος σχολιασμός των μη κωδικών γονιδίων. Παράδειγμα τέτοιου αρχείου παρουσιάστηκε στην ενότητα 5.2.3. Τέλος, οι περιοχές εμπλουτισμένης έκφρασης ελέγχονται για ενδεχόμενο κωδικό σχολιασμό, και ένα χαρακτηρστικό αρχείο εξόδου του βήματος αυτού παρουσιάστηκε στην ενότητα 5.2.4.

6.1 Δεδομένα από πειράματα NGS

Το spipeRNA εκτελέστηκε για 8 υγιή και καρκινικά small RNA-Seq δεδομένα νεφρού και πνεύμονα. Στο σχήμα 6 διακρίνονται τα ποσοστά των εγγραφών που αντιστοιχούν σε διάφορους τύπους των μη κωδικών RNA. Παρατηρεί κανείς πως τα miRNAs έχουν τη μεγαλύτερη συγκέντρωση των εγγραφών. Το σχήμα 7 παρουσιάζει τα ποσοστά των διαφόρων τύπων snRNA.

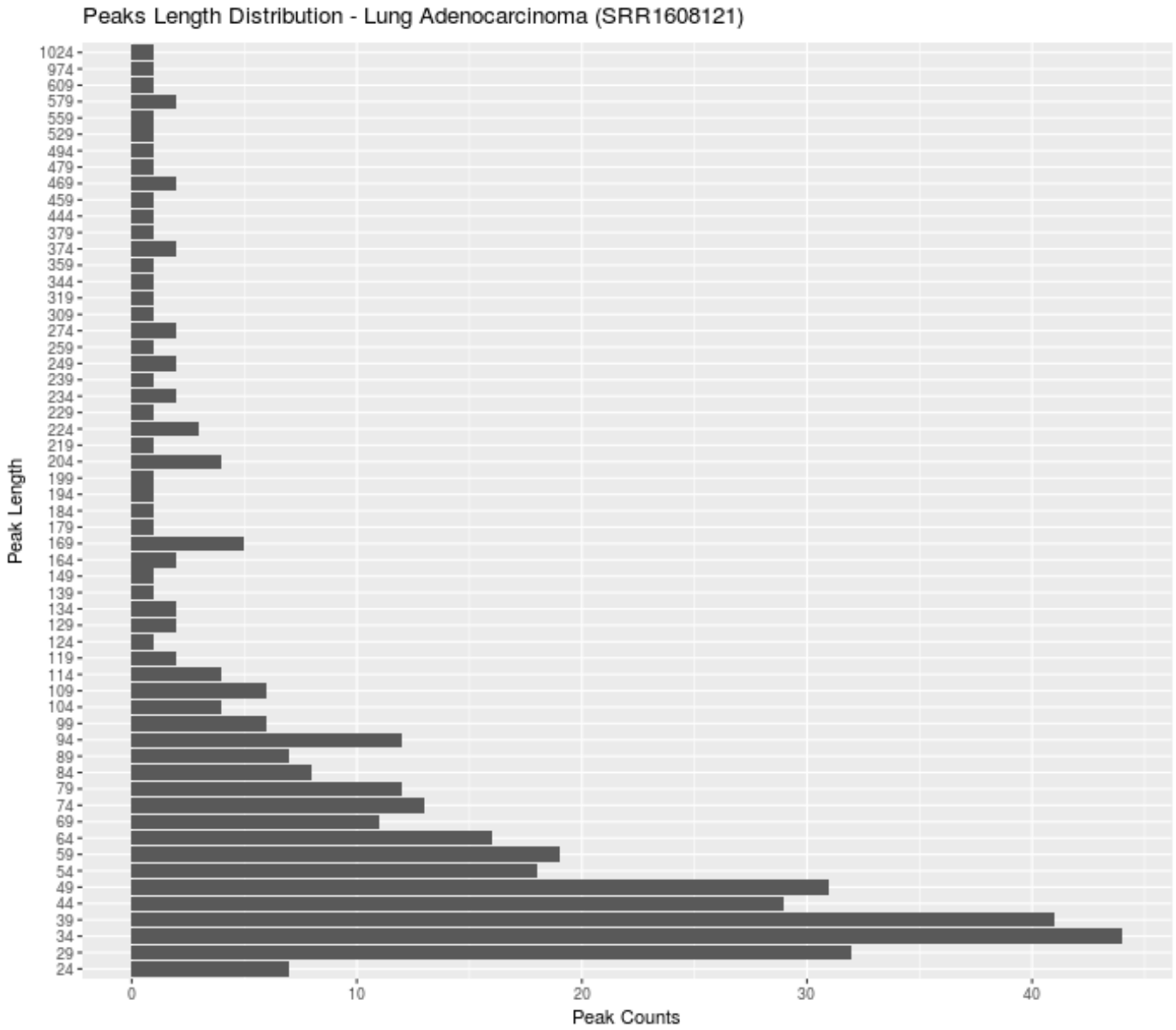


Σχήμα 6: Ποσοστά των εγγραφών που αντιστοιχούν σε διάφορους τύπους snRNA ανά δείγμα.



Σχήμα 7: Ποσοστά των διάφορων τύπων ncRNA ανά δείγμα.

Το σχήμα 8 αναπαριστά τα μήκη των προβλεπόμενων περιοχών έκφρασης σε συνάρτηση με το πλήθος τους, σύμφωνα με το spipeRNA. Παρατηρεί κανείς ότι οι περισσότερες καινούργιες προβλεπόμενες περιοχές έχουν μήκος ~40nt, κάτι που συνάδει με το προφίλ μήκους των μικρών μη κωδικών RNAs.



Σχήμα 8: Κατανομή μηκών των των προβλεπόμενων περιοχών έκφρασης

6.2 Έλεγχος ακρίβειας του spipeRNA με χρήση των προσομοιωμένων δεδομένων

Το εργαλείο spipeRNA εκτελέστηκε και για προσομοιωμένα small RNA-Seq δεδομένα εισόδου, τα οποία δημιουργήθηκαν με το εργαλείο ART. Οι προς ανάλυση fasta εγγραφές, κατασκευάστηκαν με βάση το αρχείο των σχολιασμένων μικρών μη κωδικών μεταγράφων. Το μήκος των σχηματισμένων εγγραφών, ορίστηκε στα 22 νουκλεοτίδια. Το παραχθέν fasta αρχείο με τις προς ανάλυση εγγραφές, εκτελέστηκε και με το εργαλείο miRDeep, και υπολογίστηκαν σε κάθε περίπτωση, οι λόγοι αλλαγής, μερικές μετρικές ομοιότητας και συντελεστές συσχέτισης μεταξύ των εκτιμώμενων και αναμενόμενων πληθών των μεταγράφων.

Πίνακας 1: Μετρικές ομοιότητας και συσχετίσεων μεταξύ των εκτιμώμενων miRNA μετρήσεων και των προσομοιωμένων miRNA πληθών

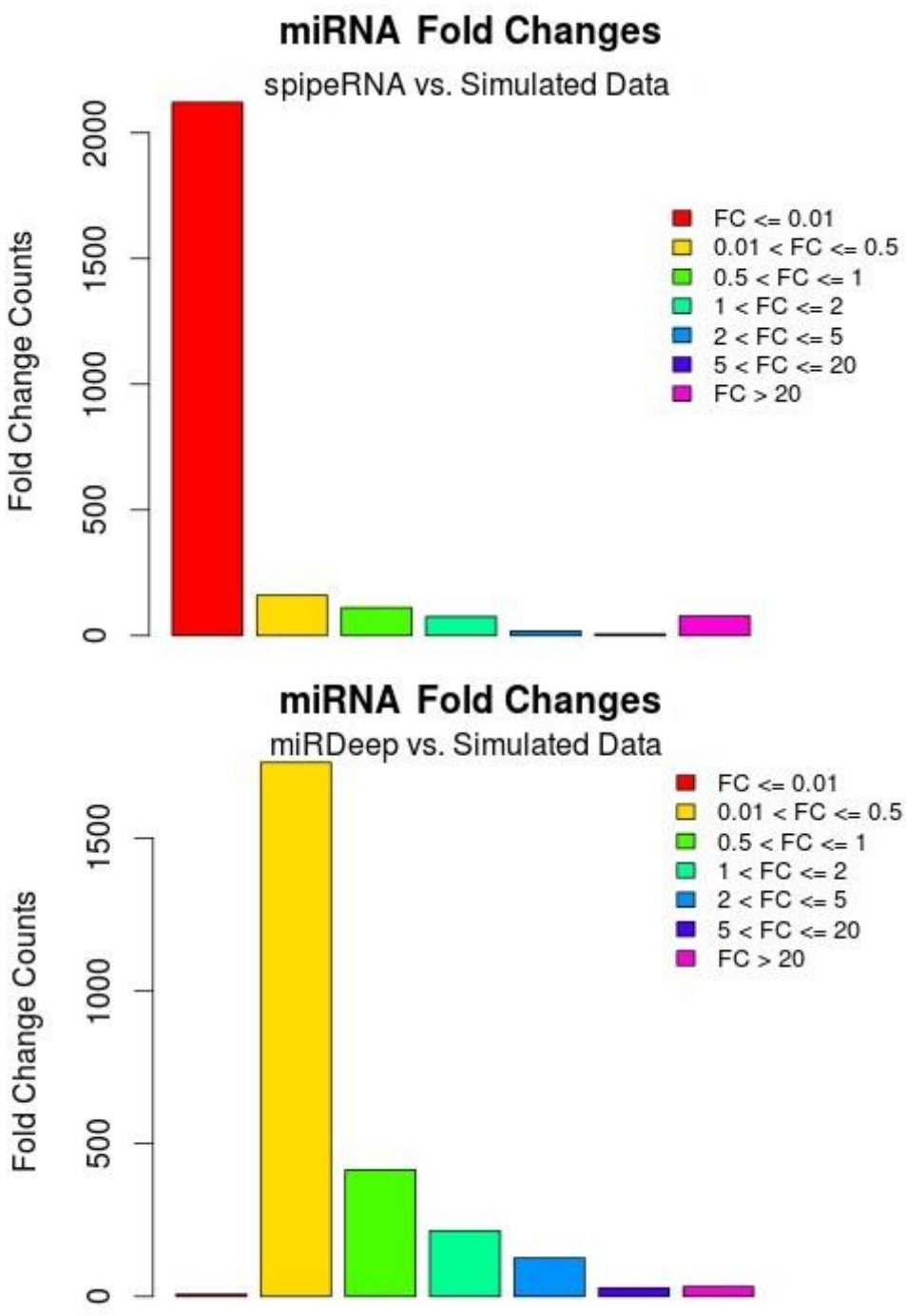
	Εκτέλεση miRDeep	**Εκτέλεση spipeRNA**
Ρίζα μέσης τετραγωνικής απόκλισης (RMSD)	45.50658	23.44123
Συσχέτιση Pearson	0.402474	0.428306
Συσχέτιση Spearman	0.5006901	0.7363603
Συσχέτιση Kendall	0.380359	0.7277342
Ευκλείδεια απόσταση	2306.063	1189.055

Η ρίζα της μέσης τετραγωνικής απόκλισης, υπολογίστηκε σύμφωνα με την παρακάτω σχέση,

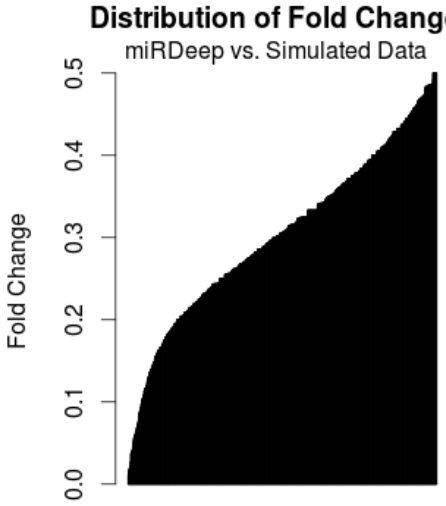
$$RMSD = \sqrt{\frac{1}{\mathrm{N}} \sum_{i=1}^{N} \delta_i^2}$$

όπου δ_i είναι η διαφορά του αναμενόμενου και εκτιμώμενου πλήθους εγγραφών, για ένα συγκεκριμένο μετάγραφο. Παρατηρεί κανείς μεγαλύτερη ακρίβεια στην εκτίμηση της έκφρασης των miRNAs, στην περίπτωση του spipeRNA. Η συσχέτιση Pearson περιγράφει μία γραμμική σχέση μεταξύ των miRNA πληθών, και λαμβάνει την τιμή 1 για μία τέλεια γραμμική συσχέτιση. Οι συντελεστές Spearman και Kendall, υποδηλώνουν μία μονότονη συσχέτιση, η οποία φαίνεται να υπάρχει στην περίπτωση της εκτέλεσης του spipeRNA. Τέλος, όσο μικρότερη είναι η Ευκλείδεια απόσταση, τόσο οι τιμές των εκτιμώμενων και προσομοιωμένων miRNA πληθών, μοιάζουν μεταξύ τους. Και πάλι, το spipeRNA δείχνει καλύτερη επίδοση στη συγκεκριμένη μετρική, σε σχέση με αυτήν του miRDeep. Οι παραπάνω μετρικές, μπορούν να επιβεβαιωθούν και οπτικά από τα σχήματα 11 και 12, όπου παρατηρεί κανείς μία μεγαλύτερη συσχέτιση μεταξύ των αναμενόμενων και των προβλεπόμενων, σύμφωνα με το spipeRNA, miRNA εκφράσεων.

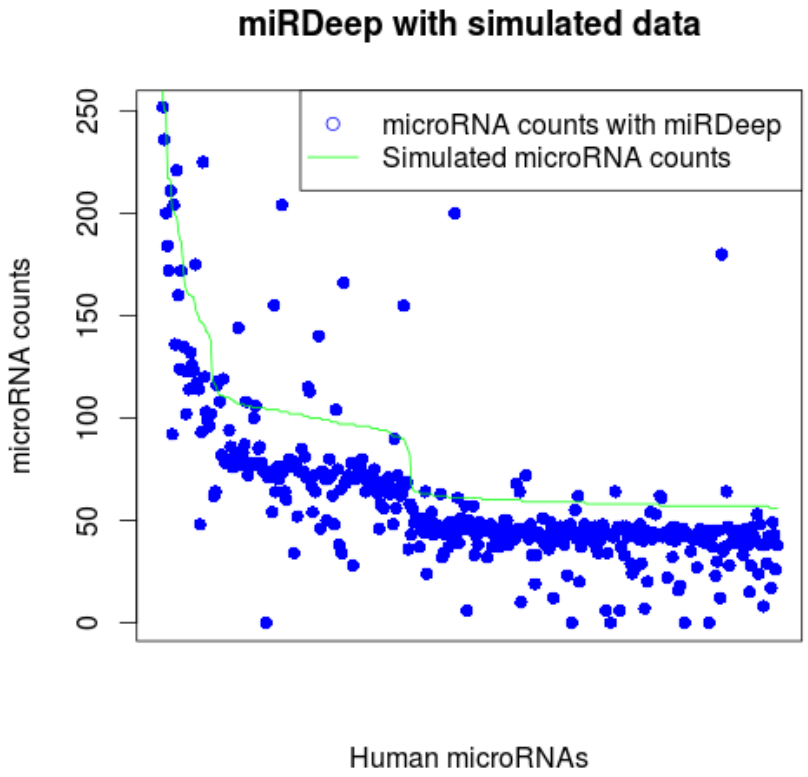
Στα διαγράμματα του σχήματος 9, παρουσιάζονται διάφορα φράγματα των λόγων αλλαγής και στις δύο εκτελέσεις, ενώ από το σχήμα 10 παρατηρεί κανείς την κατανομή των λόγων αλλαγής στην περίπτωση του miRDeep, στο διάστημα $0.01 \leq Fold\ \ Change \leq 0.5$, όπου οι περισσότεροι λόγοι αλλαγής είναι μεγαλύτεροι από 0.2 (20%).



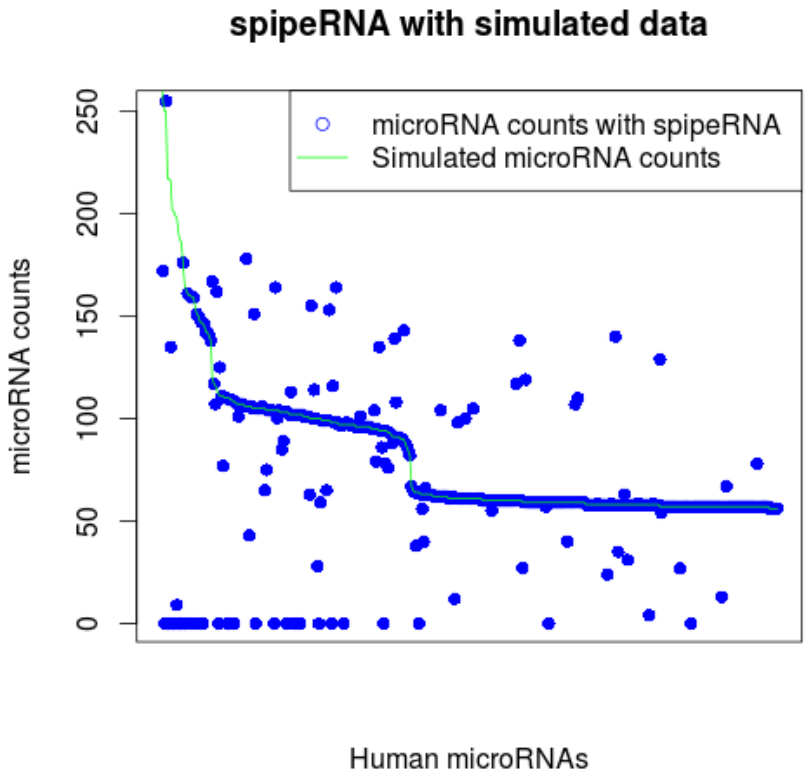
Σχήμα 9: Φάσματα λόγων αλλαγής των εκτιμώμενων και υπολογισμένων με το SpipeRNA και miRDeep, ncRNA ποσοτήτων



Σχήμα 10: Κατανομή των λόγων αλλαγής στο διάστημα από 0.01 μέχρι 0.5, μεταξύ των miRNA ποσοτήτων υπολογισμένων με το εργαλείο miRDeep και των προσομοιωμένων δεδομένων

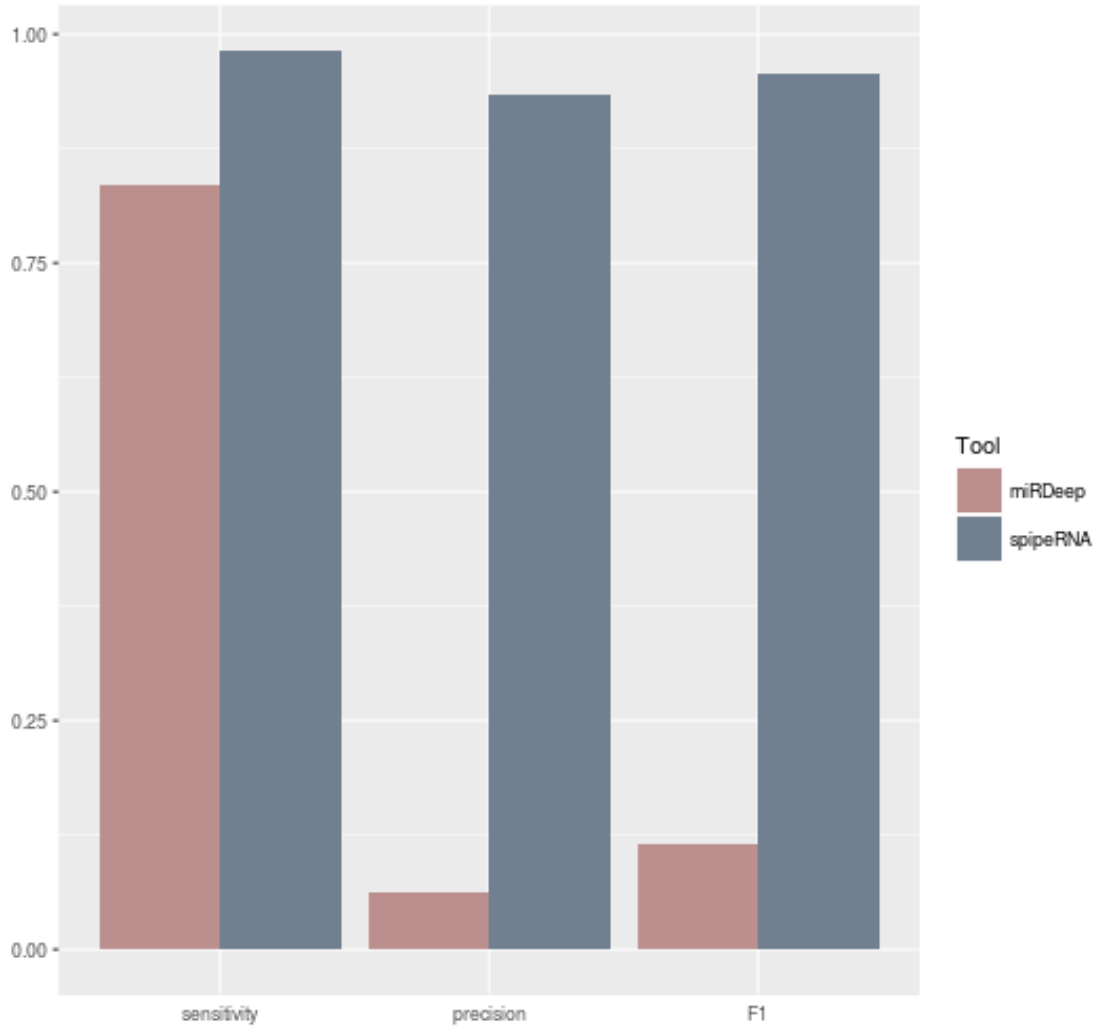


Σχήμα 11: Διάγραμμα σκέδασης των miRNA εκφράσεων υπολογισμένων με το miRDeep, για προσομοιωμένα δεδομένα εισόδου



Σχήμα 12: Διάγραμμα σκέδασης των miRNA εκφράσεων υπολογισμένων με το spipeRNA, για προσομοιωμένα δεδομένα εισόδου

Τέλος, πραγματοποιήθηκε ένας τελευταίος έλεγχος της ακρίβειας των δύο εργαλείων, με τη βοήθεια των σκορ της ευαισθησίας και της ειδικότητας. Τα θετικά αποτελέσματα (True Positive - TP), αντιστοιχίστηκαν στα miRNAs που παρουσίασαν λόγο αλλαγής $\leq$ 0.1 από το αντίστοιχο πλήθος των προσομοιωμένων δεδομένων. Τα ψευδώς θετικά (False Positive - FP) αποτελέσματα ορίστηκαν να είναι εκείνα τα πλήθη των miRNAs τα οποία παρουσίασαν λόγο αλλαγής > 0.1, ενώ τα ψευδώς αρνητικά (False Negative – FN) ορίστηκαν εκείνα τα miRNAs, τα οποία δεν εκφράστηκαν καθόλου, σύμφωνα με το εκάστοτε εργαλείο. Η ευαισθησία υπολογίστηκε βάσει της σχέσης TP/(TP+FN), η ειδικότητα σύμφωνα με τη σχέση TP/(TP+FP) ενώ το σκορ F1 σύμφωνα με τη σχέση F1=2TP/(2TP+FP+FN).



Σχήμα 13: Συνολικά σκορ της ευαισθησίας, της ειδικότητας και του F1, για τα εργαλεία miRDeep και spipeRNA και για είσοδο, τα προσομοιωμένα small RNA-Seq δεδομένα

Το σχήμα 13 αποκαλύπτει συνολικά μεγαλύτερη ακρίβεια του εργαλείου spipeRNA στην ανάλυση των προσομοιωμένων small RNA-Seq δεδομένων.

6.3 Επίδοση

Το spipeRNA δοκιμάστηκε σε ανθρώπινο γονιδίωμα αναφοράς μεγέθους 45.3GB και εκτελέστηκε σε μηχάνημα το οποίο διαθέτει 32 πυρήνες με τα παρακάτω χαρακτηριστικά ο έκαστος: Intel(R) Xeon(R) CPU E5-2630 v3 @2.40GHz, με συνολική μνήμη 264049668 kB, σε λειτουργικό Linux version 3.13.0-39-generic.

Πίνακας 2: Χρόνοι εκτέλεσης του spipeRNA

Μέγεθος αρχείου	**Αριθμός εγγραφών**	**Χρόνος εκτέλεσης**
1.6G	11,252,414	707s
116.9MB	1,001,097	75s

Το εργαλείο spipeRNA παρέχει τη δυνατότητα παράλληλης εκτέλεσης πολλών δειγμάτων. Παρακάτω, παρουσιάζεται ο χρόνος εκτέλεσης του spipeRNA για τέσσερα αρχεία εισόδου, με τη χρήση της παράλληλης εκτέλεσης.

Πίνακας 3: Χρόνοι εκτέλεσης του spipeRNA

Μέγεθος αρχείου	**Αριθμός εγγραφών**
1.6G	11,252,414
1.7G	11,708,733
1.9G	12,895,133
1.8G	12,763,204
Χρόνος εκτέλεσης	713s

Σημειώνεται πως για τις παραπάνω εκτελέσεις, στο αρχείο configuration.txt, ο αριθμός των παράλληλων διεργασιών ορίστηκε να είναι 4 και ο αριθμός των πυρήνων για την ευθυγράμμιση ορίστηκε να είναι 1.

7. ΣΥΓΚΡΙΣΗ ΤΩΝ miRDeep KAI spipeRNA

Το spipeRNA είναι μία ολοκληρωμένη πρόταση για την ποσοτικοποίηση κάθε είδους μικρών σε μήκος RNA μορίων, που εκφράζονται σε βιολογικό δείγμα και αναπαρίστανται με δεδομένα της βαθιάς αλληλούχησης. Όλες οι εγγραφές-μεταγραφές, χρησιμοποιούνται με τον ίδιο τρόπο, ανεξαρτήτως της βιολογικής τους προέλευσης, των πιθανών βιολογικών ιδιαιτεροτήτων του κάθε μορίου κτλ.

Υπάρχουν πολλά διαθέσιμα εργαλεία, των οποίων οι αλγόριθμοι είναι αφιερωμένοι στην ποσοτικοποίηση συγκεκριμένων μικρών μορίων RNA, όπως είναι το miRDeep, το οποίο αναπτύχθηκε για να ανακαλύπτει γνωστά και άγνωστα miRNAs από δεδομένα βαθιάς αλληλούχησης, το tDRmapper το οποίο καθορίζει, ονοματολογεί και ποσοτικοποιεί τα tRNAs που εκφράζονται σε ένα δείγμα κα.

Η παρούσα ενότητα έχει στόχο τη σύγκριση των αποτελεσμάτων μεταξύ του πιο δημοφιλούς εργαλείου διαχείρισης των microRNAs, του miRDeep2, και του spipeRNA, και την περαιτέρω ανάλυση των τυχόντων διαφοροποιήσεων μεταξύ τους.

7.1 Σύγκριση αποτελεσμάτων

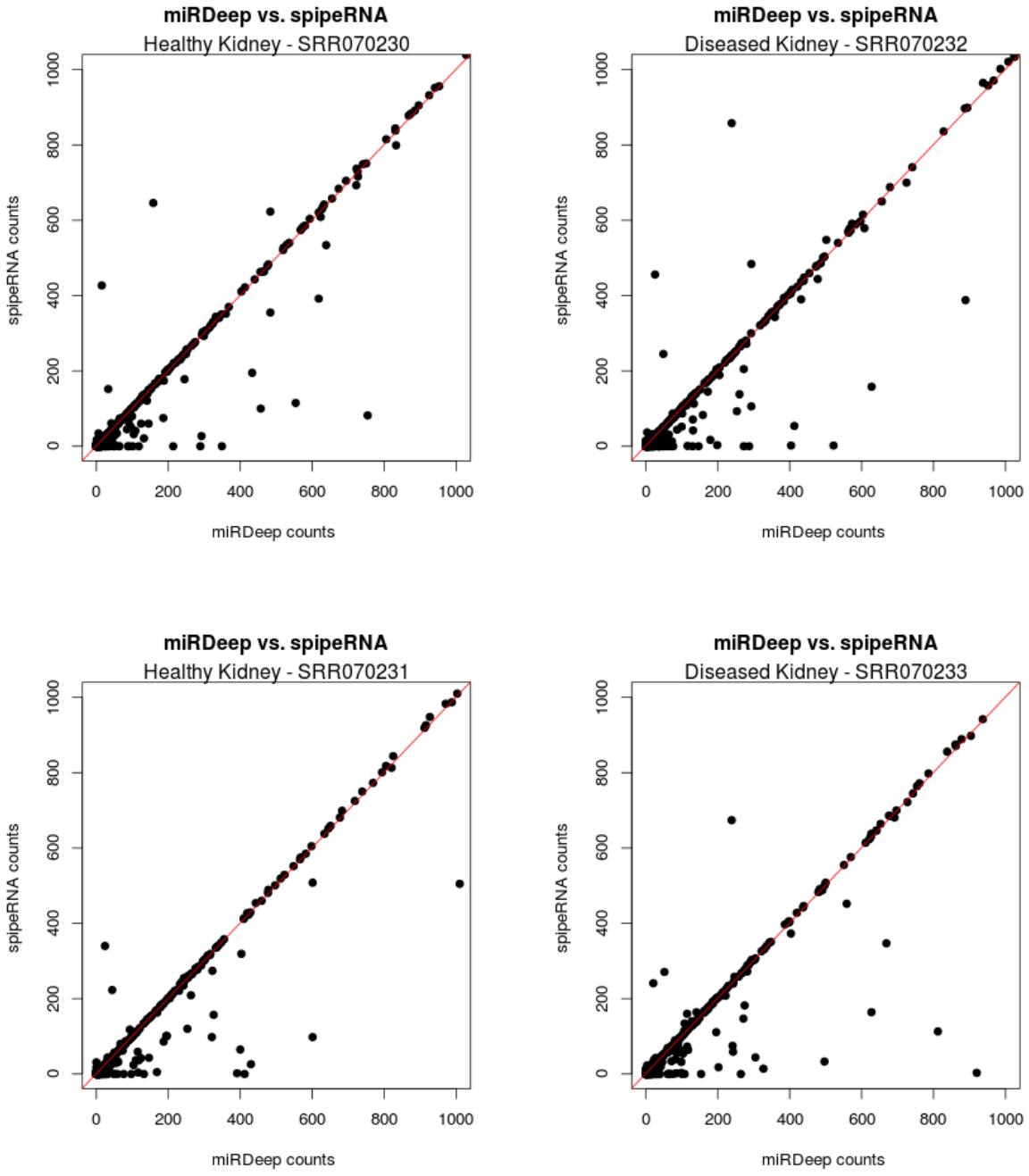
Για τη σύγκριση των αποτελεσμάτων μεταξύ των δύο εργαλείων, επιλέχθηκαν τα παρακάτω δείγματα NGS πειραμάτων.

Πίνακας 4: Βιολογικά δείγματα από small RNA-Seq πειράματα

Αναγνωριστικό Μελέτης (Study ID)	**Αναγνωριστικό Δείγματος (Sample ID)**	**Τύπος Δείγματος**
SRP003902	SRR070230	Healthy Kidney (Υγιής Νεφρός)
SRP003902	SRR070231	Healthy Kidney (Υγιής Νεφρός)
SRP003902	SRR070232	Kidney Cancer (Καρκινικός Νεφρός)
SRP003902	SRR070233	Kidney Cancer (Καρκινικός Νεφρός)
SRP048750	SRR1608121	Lung Adenocarcinoma (Αδενοκαρκίνωμα Πνεύμονα)
SRP048750	SRR1608122	Lung Adenocarcinoma (Αδενοκαρκίνωμα Πνεύμονα)
SRP057590	SRR1988283	Healthy Lung (Υγιής Πνεύμονας)
SRP057590	SRR1988284	Healthy Lung (Υγιής Πνεύμονας)

Για το κάθε δείγμα του πίνακα 4, υπολογίστηκαν οι λόγοι αλλαγής των miRNA εκφράσεων μεταξύ των εκτελέσεων miRDeep και spipeRNA. Στις μετρήσεις

συμπεριλήφθηκαν μόνο εκείνα τα miRNAs, για τα οποία σε μία, τουλάχιστον, από τις δύο εκτελέσεις, ο αριθμός εμφάνισής τους ξεπερνά τις δέκα εγγραφές. Στις παρακάτω εικόνες, ο άξονας των x αναπαριστά την έκφραση των miRNAs που υπολογίστηκαν με τον αλγόριθμο miRDeep, ενώ ο άξονας των y αναπαριστά την έκφραση των miRNAs που υπολογίστηκαν με το spipeRNA.



Σχήμα 14: Διαγράμματα σκέδασης των miRNA εκφράσεων μεταξύ των miRDeep και spipeRNA εκτελέσεων

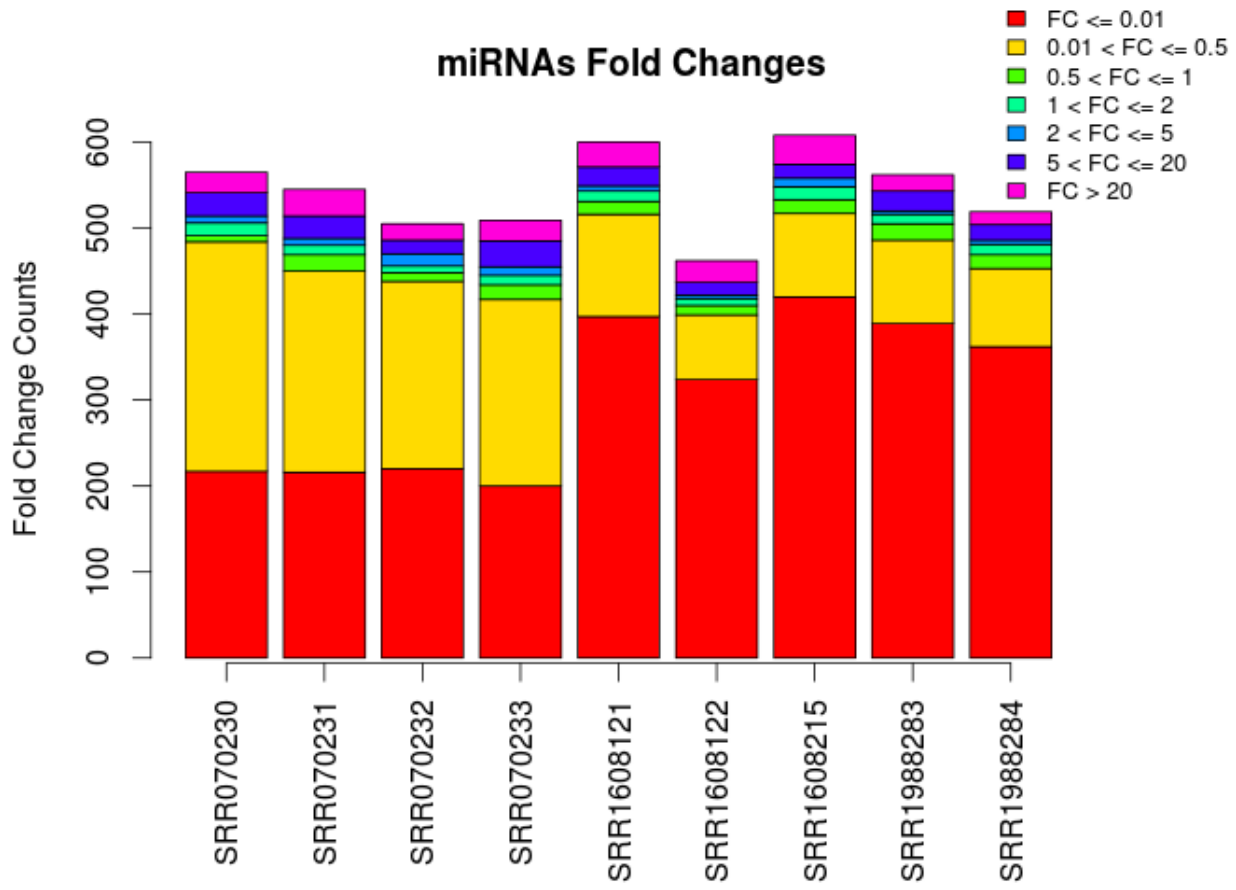
Από τα παραπάνω διαγράμματα σκέδασης, μπορεί κανείς να παρατηρήσει μία γραμμική σχέση των περισσοτέρων miRNA ποσοτήτων μεταξύ των δύο εκτελέσεων,

καθώς και αρκετά miRNA για τα οποία οι δύο αλγόριθμοι δε συμφωνούν στις ποσοτικές τους εκτιμήσεις.

Αναλυτικότερα, για κάθε γνωστό miRNA της βάσης miRBase, οι αντίστοιχες μετρήσεις των δύο αλγορίθμων συγκρίθηκαν με τη βοήθεια του λόγου αλλαγής ο οποίος αναπαρίσταται με την παρακάτω σχέση:

$$\lambda\acute{o}\gamma o\varsigma\ \alpha\lambda\lambda\alpha\gamma\acute{\eta}\varsigma = \frac{(B - A)}{A}$$

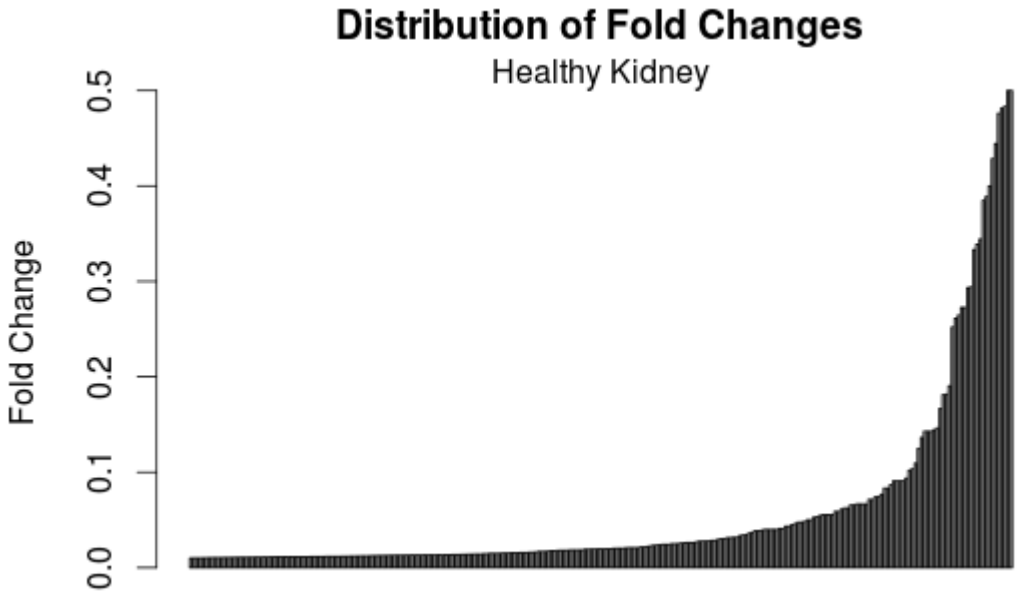
στην οποία το B ισοδυναμεί με την υψηλότερη έκφραση ενός miRNA, σε μία από τις δύο εκτελέσεις. Πριν την εφαρμογή της παραπάνω σχέσης, σε όλες τις miRNA μετρήσεις των δύο εκτελέσεων, προστέθηκε μία εγγραφή. Η προσθήκη αυτή σκοπό έχει να μπορεί να εφαρμοστεί η σχέση fold change χωρίς περιορισμούς. Το σχήμα 15 παρουσιάζει τα φάσματα των λόγων αλλαγής, για τα επιλεγμένα προς σύγκριση δείγματα.



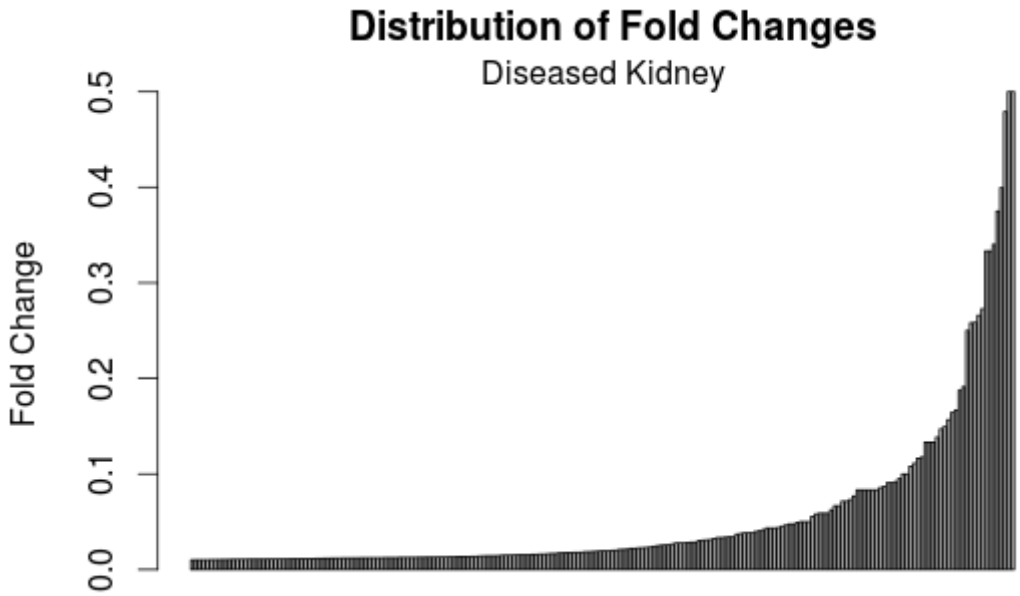
Σχήμα 15: Φάσματα λόγων αλλαγής

Από το παραπάνω σχήμα, μπορεί κανείς να παρατηρήσει πως οι περισσότεροι λόγοι αλλαγής συγκαταλέγονται στα παρακάτω διαστήματα: $0 \le Fold\ \ Change \le 0.01$ και $0.01 \le Fold\ \ Change \le 0.5$. Ο λόγος αλλαγής 0.01 αντιστοιχεί στο ~1% διαφοράς μεταξύ

των δύο εκτιμήσεων. Εάν για παράδειγμα, το miRDeep αντιστοίχισε 100 εγγραφές σε ένα miRNA και το spipeRNA αντιστοίχισε 101 εγγραφές στο ίδιο miRNA, ο λόγος αλλαγής μεταξύ των δύο εκτελέσεων για το συγκεκριμένο miRNA είναι 0.01. Ο λόγος αλλαγής 0.5, αντιστοιχεί σε 1.5 αλλαγή. Στα διαγράμματα των σχημάτων 16 και 17 παρουσιάζονται όλοι οι λόγοι αλλαγής οι οποίοι ανήκουν στο διάστημα $0.01 \leq Fold\ \ Change \leq 0.5$ $0.01 \leqslant FoldChange \leqslant 0.5$. Παρατηρεί κανείς πως στο διάστημα αυτό, οι περισσότεροι λόγοι αλλαγής είναι μικρότεροι από 0.1 και είναι κοντά στο 0.01. Επομένως, στις περισσότερες εκτιμήσεις των miRNA εκφράσεων, οι δύο αλγόριθμοι συμφωνούν. Υπάρχουν όμως αρκετά miRNAs τα οποία εμφανίζουν μεγάλες αποκλίσεις μεταξύ των δύο εκτελέσεων.



Σχήμα 16: Κατανομή των λόγων αλλαγής στο διάστημα από 0.01 μέχρι 0.5 στο δείγμα υγιούς νεφρού



Σχήμα 17: Κατανομή των λόγων αλλαγής στο διάστημα από 0.01 μέχρι 0.5 στο δείγμα καρκινικού νεφρού

7.2 Ανάλυση των μεγάλων αποκλίσεων

Παρόλο που οι περισσότερες miRNA ποσότητες μεταξύ των δύο εκτελέσεων είναι παρόμοιες, όπως έχει δειχτεί στην προηγούμενη ενότητα, υπάρχουν αρκετά miRNAs τα οποία παρουσιάζουν πολύ μεγάλα fold changes μεταξύ των δύο αλγορίθμων. Παρακάτω, αναλύονται μερικά τέτοια miRNAs και διερευνώνται οι πιθανές αιτίες που οδηγούν σε τόσο μεγάλες και σημαντικές αποκλίσεις στις εκτιμήσεις των εκφράσεών τους, μεταξύ των miRDeep2 και spipeRNA αλγορίθμων. Στον παρακάτω πίνακα παρουσιάζονται τέσσερις μεγαλύτεροι λόγοι αλλαγής που προέκυψαν στο δείγμα υγιούς νεφρού (SRR070230).

Πίνακας 5: Τέσσερα microRNAs των οποίων οι εκφράσεις μεταξύ των εκτελέσεων miRDeep και spipeRNA εμφάνισαν τους μεγαλύτερους λόγους αλλαγής

	Αναγνωριστικό μελέτης	**miRNA**	**miRDeep count**	**spipeRNA count**	**fold change**
1.	SRR070230	hsa-miR-1261	349	0	349
2.	SRR070230	hsa-miR-4492	289	0	289
3.	SRR070230	hsa-miR-619-5p	213.5	0	213
4.	SRR070230	hsa-miR-4508	1460	7	204.125

Παρακάτω αναλύονται και τα τέσσερα miRNA ξεχωριστά.

7.2.1 hsa-miR-1261

Από τη βάση δεδομένων miRBase (έκδοση 21), το συγκεκριμένο miRNA φέρει τα παρακάτω χαρακτηριστικά:

Πίνακας 6: Τα βασικά χαρακτηριστικά του ανθρωπίνου miR-1261

Όνομα miRNA αλληλουχίας	hsa-miR-1261
Mature αλληλουχία	**ATGGATAAGGCTTTGGCTT**
Precursor αλληλουχία	TGCT**ATGGATAAGGCTTTGGCTT**ATGGGGA TATTGTGGTTGATCTGTTCTATCCAGATGAC TGAAACTTTCTCCATAGCAGC
Χρωμόσωμα	11
Κλώνος	+
Mature Start Loci	90869180
Mature Stop Loci	90869198
Precursor Start Loci	90869121
Precursor Stop Loci	90869202

Ανάλυση του FASTQ αρχείου

Παρακάτω παρουσιάζονται κάποια αποτελέσματα αναζητήσεων του «καθαρισμένου» από αλληλουχίες πρόσδεσης αρχείου fastq με αναγνωριστικό μελέτης SRR070230, για αλληλουχίες σχετικές με αυτήν του miRNA «hsa-miR-1261»|.

Πίνακας 7: Αναζήτηση της hsa-miR-1261 αλληλουχίας σε αρχείο fastq

	hsa-miR-1261	Αριθμός εμφάνισης	Σχόλιο
miRNA αλληλουχία	ATGGATAAGGCTTTGGCTT	0	
Αντεστραμμένη miRNA αλληλουχία	TTCGGTTTCGGAATAGGTA	0	
Συμπληρωματική miRNA αλληλουχία	TACCTATTCCGAAACCGAA	0	
Αντεστραμμένη συμπληρωματική αλληλουχία miRNA	AAGCCAAAGCCTTATCCAT	0	
miRNA αλληλουχία	ATGGATAAGGCATTGGCTT	1429	Μία αναντιστοιχία
Αντεστραμμένη miRNA αλληλουχία	TTCGGTTACGGAATAGGTA	0	Μία αναντιστοιχία
Συμπληρωματική miRNA αλληλουχία	TACCTATTCCGTAACCGAA	0	Μία αναντιστοιχία
Αντεστραμμένη συμπληρωματική αλληλουχία miRNA	AAGCCTAAGCCTTATCCAT	0	Μία αναντιστοιχία
miRNA αλληλουχία	ATGGATAAGGCATTGGCTT	285	Μία αναντιστοιχία της αλληλουχίας. Εύρεση της ακολουθίας με την παρακάτω κανονική έκφραση ^ATGGATAAGGCATTGGCTT$
Μέρος της miRNA αλληλουχίας	GGATAAGGCATTGGCTT**A**	0	Flanking νουκλεοτίδια της hairpin αλληλουχίας
miRNA αλληλουχίας	**T**ATGGATAAGGCATTGGCTT	0	Flanking νουκλεοτίδια της hairpin αλληλουχίας

Από τον παραπάνω πίνακα γίνεται κατανοητό, πως ο αλγόριθμος του miRDeep έχει αντιστοιχίσει την αλληλουχία ATGGATAAGGCATTGGCTT του fastq αρχείου με την αλληλουχία του ανθρωπίνου miRNA-1261, επιτρέποντας την αναντιστοιχία T→A.

Ενώ στους περισσότερους αλγορίθμους ευθυγράμμισης μικρών RNAs, μία αναντιστοιχία είναι επιτρεπτή, πρέπει να ερευνηθεί εάν στη συγκεκριμένη περίπτωση, η αναντιστοιχία αυτή είναι σωστό να θεωρηθεί ως πολυμορφισμός ενός νουκλεοτιδίου (ή SNP).

Ενδιαφέρον παρουσιάζει μία πιο προσεκτική ματιά στις εγγραφές του fastq αρχείου του συγκεκριμένου δείγματος. Η εντολή

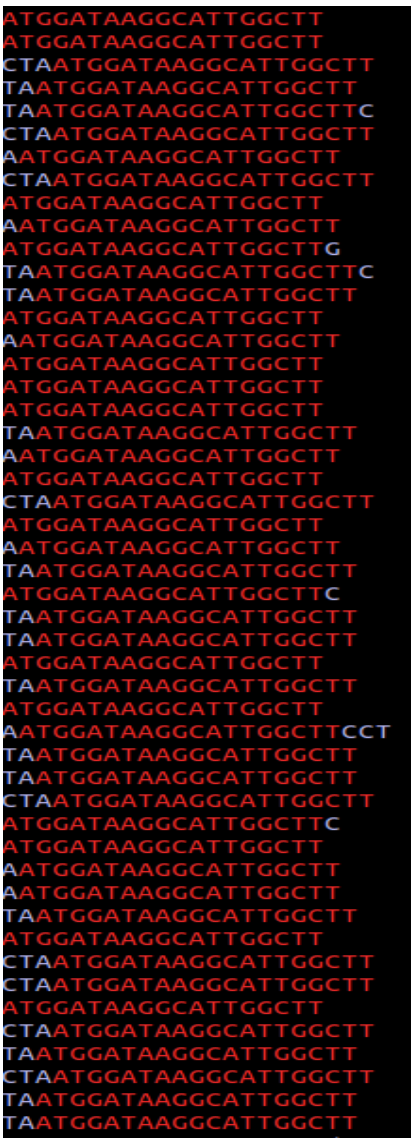
```
>cat SRR070230_trimmed.fastq | grep '^ATGGATAAGGCATTGGCTT$' | wc
-l

285
```

μετρά τις εγγραφές, οι οποίες ταιριάζουν ακριβώς στην αλληλουχία του hsa-mir-1261 με μία αναντιστοιχία T→A.

Η εντολή

```
>cat SRR070230_trimmed.fastq | grep ATGGATAAGGCATTGGCTT
```

επιστρέφει όλες τις εγγραφές, οι οποίες περιέχουν τη συμβολοσειρά ATGGATAAGGCATTGGCTT. Στην εικόνα 16 παρουσιάζεται ένα μέρος της εξόδου της εντολής, δηλαδή κάποιες εγγραφές οι οποίες περιέχουν την αλληλουχία miRNA-1261 και είναι σημειωμένες με κόκκινο χρώμα.



Εικόνα 16: Μέρος των εγγραφών του fastq αρχείου

Από τα σημειωμένα με μπλε χρώμα συνοδευτικά νουκλεοτίδια, μπορεί κανείς να παρατηρήσει ότι οι αντίστοιχες αλληλουχίες δε μοιάζουν να προέρχονται από την αλληλουχία του hsa-mir-1261 προδρόμου. Στο επόμενο βήμα θα διερευνηθεί σε ποιο μετάγραφο ταιριάζουν περισσότερο οι αλληλουχίες αυτές.

Ανάλυση του SAM αρχείου από την εκτέλεση του εργαλείου Butter

Παρακάτω παρουσιάζεται ένα μόνο δείγμα ευθυγραμμισμένης εγγραφής σύμφωνα με το εργαλείο butter. Η εγγραφή φέρει παρόμοια αλληλουχία με αυτήν του hsa-mir-1261 με την αναντιστοιχία που έχει αναφερθεί προηγουμένως.

```
SRR070230_trimmed_264300_14    0    16   3152910  255  23M  *
0    0    CCTAATGGATAAGGCATTGGCTT   IIIIIIIIIIIIIIIIIIIIIII
XA:i:1    MD:Z:21C1     NM:i:1    XX:i:2    XY:Z:P    XZ:f:0.248
```

Η SAM ετικέτα XY:Z:P, σημαίνει πως η συγκεκριμένη εγγραφή έχει ευθυγραμμιστεί πάνω στο γονιδίωμα, με βάση την ομαδοποίηση που περιγράφτηκε στην ενότητα 5.2.1. Η μία αναντιστοιχία έχει επισημανθεί με κόκκινο χρώμα. Από το αρχείο των σχολιασμένων μη κωδικών RNA ευρέθηκε το παρακάτω γονίδιο που κωδικοποιεί το tRNA-Arg (CCT) 3-1 και το οποίο ταιριάζει με τις γονιδιωματικές συντεταγμένες της αλληλουχίας του αρχείου SAM.

16 + 3152900 3152972 tRNA UCSC tRNA-Arg-CCT-3-1

Από τη βάση δεδομένων UCSC, η αλληλουχία του συγκεκριμένου tRNA παρουσιάζεται παρακάτω:

```
>hg38_tRNAs_tRNA-Arg-CCT-3-1 range=chr16:3152900-3152972

5'pad=0 3'pad=0 strand=+ repeatMasking=none

GCCCCGGTGGCCTAATGGATAAGGCATTGGCCTCCTAAGCCAGGGATTGTGGGTTCGAGTCCC

ACCCGGGGTA
```

Σημειώνεται επίσης, πως η αλληλουχία **AGCCAGGGATTGTGGGT**, η οποία επιλέχτηκε τυχαία, εμφανίζεται 161 φορές στο αρχείο fastq και αντιστοιχεί στο άκρο 5’ του tRNA γονιδίου. Στο αρχείο εξόδου του spipeRNA, το οποίο περιέχει τις μετρήσεις των διαφόρων μετάγραφων του συγκεκριμένου δείγματος, παρόλο που δεν υπάρχει ούτε μία μέτρηση για το miR-1261, υπάρχουν 604 μετρήσεις συσχετισμένες με το μετάγραφο tRNA-Arg (CCT) 3-1.

Συμπέρασμα

Χωρίς τη χρήση επιπλέον δεδομένων, δεν μπορεί να επιβεβαιωθεί ποιος αλγόριθμος είναι ορθός ή ορθότερος από τον άλλον. Από τα στοιχεία όμως που μελετήθηκαν πιο πάνω, και ειδικά, από τις αλληλουχίες της εικόνας 16 και τις εντολές αναζήτησης, παρατηρεί κανείς ότι ενώ υπάρχουν πολλές αλληλουχίες οι οποίες ταιριάζουν (με μία αναντιστοιχία) με την αλληλουχία του hsa-miR-1261, αμφίβολο είναι εάν προέρχονται όντως από το πρόδρομο hsa-miR-1261. Οι πολλές εγγραφές με συνοδευτικά στις άκρες τους νουκλεοτίδια, που δεν ταιριάζουν στην αλληλουχία του προδρόμου του hsa-miR-1261, συνηγορούν υπέρ της ποσοτικοποίησης του spipeRNA. Επιπλέον, τα θραύσματα του 5’ άκρου του μεταγράφου tRNA-Arg (CCT) 3-1 που βρέθηκαν στο αρχείο εισόδου, αποδεικνύουν την έκφραση του γονιδίου αυτού.

Το miRDeep2 είναι ένα εργαλείο αφιερωμένο στην ποσοτικοποίηση των miRNAs, πραγματοποιώντας μεταξύ άλλων την ευθυγράμμιση των εγγραφών πάνω στο μεταγράφωμα. Αυτό, ευνοεί περισσότερο τα συγκεκριμένα μόρια σε σχέση με άλλα μικρά RNAs.

7.2.2 hsa-miR-4492

Από τη βάση δεδομένων miRBase (έκδοση 21), το miRNA hsa-miR-4492 φέρει τα παρακάτω χαρακτηριστικά:

Πίνακας 8: Τα βασικά χαρακτηριστικά του ανθρωπίνου miR-4492

Όνομα miRNA αλληλουχίας	hsa-miR-4492
Mature αλληλουχία	**GGGGCTGGGCGCGCGC**
Precursor αλληλουχία	CTGCAGCGTGCTTCTCCAGGCCCCGCGCG CGGACAGACACACGGACAAGTCCCGCCA**G GGGCTGGGCGCGCGC**CAGCCGG
Χρωμόσωμα	11
Κλώνος	+
Mature Start Loci	118910765
Mature Stop Loci	118910781
Precursor Start Loci	118910708
Precursor Stop Loci	118910787

Ανάλυση του FASTQ αρχείου

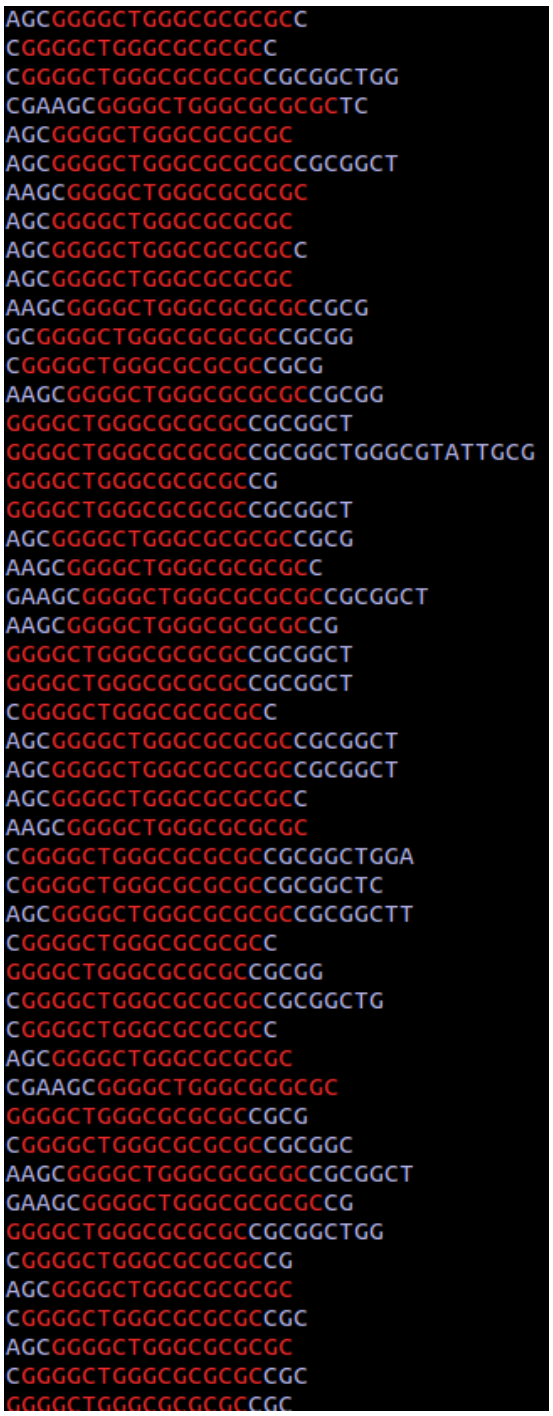
Παρακάτω παρουσιάζονται κάποια αποτελέσματα αναζητήσεων του, «καθαρισμένου» από αλληλουχίες πρόσδεσης αρχείου fastq, με αναγνωριστικό μελέτης SRR070230, για αλληλουχίες σχετικές με αυτήν του miRNA «hsa-miR-4492».

Πίνακας 9: Αναζήτηση της has-miR-4492 αλληλουχίας σε αρχείο fastq

	hsa-miR-4492	Αριθμός εμφάνισης	Σχόλιο
miRNA αλληλουχία	GGGGCTGGGCGCGCGC	6700	
Αντεστραμμένη miRNA αλληλουχία	CGCGCGCGGGTCGGGG	60	
Συμπληρωματική miRNA αλληλουχία	CCCCGACCCGCGCGCG	0	
Αντεστραμμένη συμπληρωματική αλληλουχία miRNA	GCGCGCGCCCAGCCCC	0	
miRNA αλληλουχία	GGGGCTGGGCGCGCGC	0	Αριθμός εμφάνισης αλληλουχιών με την παρακάτω κανονική έκφραση: ^GGGGCTGGGCGCGCGC$

Μία από τις πιθανές αιτίες εμφάνισης των 289 εγγραφών που αντιστοιχούν στο hsa-miR-4492 σύμφωνα με το miRDeep, μπορεί να οφείλεται στην παρουσία των παρακάτω αλληλουχιών του αρχείου fastq.

```
>cat SRR070230_trimmed.fastq | grep '^CGGGGCTGGGCGCGCGCC$' | wc -l
220
```



Εικόνα 17: Μέρος των εγγραφών του fastq αρχείου

Από την εικόνα 17, παρατηρεί κανείς με «γυμνό» μάτι, ότι τα νουκλεοτίδια πέριξ του mature hsa-miR-4492 δεν ταιριάζουν με αυτά της αλληλουχίας του προδρόμου του.

Ανάλυση του SAM αρχείου από την εκτέλεση του εργαλείου Butter

Παρακάτω παρουσιάζεται ένα μόνο δείγμα ευθυγραμμισμένης εγγραφής σύμφωνα με το εργαλείο butter. Η εγγραφή φέρει παρόμοια αλληλουχία με αυτήν του hsa-miR-4492 με την αναντιστοιχία που έχει αναφερθεί προηγουμένως.

Από το αρχείο SAM, αναζητήθηκαν αλληλουχίες, οι οποίες είναι παρόμοιες με αυτήν του hsa-miR-4492. Μεταξύ άλλων, παρακάτω παρουσιάζεται ένα μικρό τμήμα ορισμένων περιοχών, στις οποίες η συγκεκριμένη αλληλουχία ευθυγραμμίζεται χωρίς καμία αναντιστοιχία.

```
SRR070230_trimmed_469297_16    0    KI270733.1    133125    255  21M  *    0
    0    AAGCGGGGCTGGGCGCGCGCC IIIIIIIIIIIIIIIIIIIII XA:i:0   MD:Z:21
    NM:i:0   XX:i:7   XY:Z:P   XZ:f:0.358
SRR070230_trimmed_405646_32    0    KI270733.1    133124    255  21M  *    0
    0    GAAGCGGGGCTGGGCGCGCGC IIIIIIIIIIIIIIIIIIII XA:i:0   MD:Z:21
    NM:i:0   XX:i:7   XY:Z:P   XZ:f:0.358
SRR070230_trimmed_665847_5     0    KI270733.1    133124    255  26M  *    0
    0    GAAGCGGGGCTGGGCGCGCGCCGCGG    IIIIIIIIIIIIIIIIIIIIIIIII
    XA:i:0   MD:Z:26  NM:i:0   XX:i:7   XY:Z:P   XZ:f:0.358
```

Από το αρχείο των σχολιασμένων μη κωδικών RNAs, το μετάγραφο που βρέθηκε ταιριάζει ως προς τις συντεταγμένες του με αυτές των παραπάνω αλληλουχιών, ανήκει στο γονίδιο RNA45S5 (RNA, 45S pre-ribosomal 5). Παρακάτω παρουσιάζεται αναλυτικά το μετάγραφο αυτό.

KI270733.1 + 130203 135280 rRNA NR_046235 RNA28S5

Μέρος της αλληλουχίας του RNA28S5 σύμφωνα με τη βάση δεδομένων NCBI παρουσιάζεται παρακάτω. Η νουκλεοτιδική αλληλουχία του RNA28S5 φαίνεται να ταιριάζει περισσότερο στο μοτίβο των αλληλουχιών της εικόνας 16.

```
10801 ggataaggattggctctaagggctgggtcggtcgggctggggcgcgaagcggggctgggc
10861 gcgcgccgcggctggacgaggcgccgccgccccccccacgcccggggcacccccctcgcg
```

Σημειώνεται, πως το συγκεκριμένο rRNA μετρήθηκε 104 φορές στο δείγμα SRR070230, σύμφωνα με το spipeRNA.

7.2.3 hsa-miR-619-5p

Από τη βάση δεδομένων miRBase (έκδοση 21), το συγκεκριμένο miRNA φέρει τα παρακάτω χαρακτηριστικά:

Πίνακας 10: Τα βασικά χαρακτηριστικά του ανθρωπίνου miR-619-5p

Όνομα miRNA αλληλουχίας	hsa-miR-619-5p
Mature αλληλουχία	**GCTGGGATTACAGGCATGAGCC**
Precursor αλληλουχία	CGCCCACCTCAGCCTCCCAAAAT**GCTGGGATTACAGGCATGAGCC**ACTGCGGTCGACCATGACCTGGACATGTTTGTGCCCAGTACTGTCAGTTTGCAG

Χρωμόσωμα	12
Κλώνος	-
Mature Start Loci	108836962
Mature Stop Loci	108836983
Precursor Start Loci	108836908
Precursor Stop Loci	108837006

Ανάλυση του FASTQ αρχείου

Παρακάτω παρουσιάζονται κάποια αποτελέσματα αναζητήσεων του «καθαρισμένου» από αλληλουχίες πρόσδεσης αρχείου fastq με αναγνωριστικό μελέτης SRR070230, για αλληλουχίες σχετικές με αυτήν του miRNA «hsa-miR-619-5p».

Πίνακας 11: Αναζήτηση της has-miR-619-5p αλληλουχίας σε αρχείο fastq

	hsa-miR-619-5p	Αριθμός εμφάνισης	Σχόλιο
miRNA αλληλουχία	GCTGGGATTACAGGCATGAGCC	3	
Αντεστραμμένη miRNA αλληλουχία	CCGAGTACGGACATTAGGGTCG	0	
Συμπληρωματική miRNA αλληλουχία	CGACCCTAATGTCCGTACTCGG	0	
Αντεστραμμένη συμπληρωματική αλληλουχία miRNA	GGCTCATGCCTGTAATCCCAGC	0	
miRNA αλληλουχία	GCTGGGATTACAGGCA	85	Μέρος της miRNA αλληλουχίας

Εικόνα 18: Μέρος των εγγραφών του fastq αρχείου

Ανάλυση του SAM αρχείου από την εκτέλεση του εργαλείου Butter

Παρακάτω παρουσιάζεται ένα μικρό μέρος ευθυγραμμισμένων εγγραφών σύμφωνα με το εργαλείο butter. Η εγγραφή φέρει παρόμοια αλληλουχία με την αλληλουχία του «hsa-miR-619-5p».

```
SRR070230_trimmed_62500_1 4    *    0    0    *    *    0    0
    GCTGGGATTACAGGCATG    IIIIIIIIIIIIIIIII    XM:i:1000    XX:i:0
    XY:Z:M   XZ:f:1
SRR070230_trimmed_65571_1 4    *    0    0    *    *    0    0
    TGCTGGGATTACAGGCATG   IIIIIIIIIIIIIIIIII   XM:i:1000    XX:i:0
    XY:Z:M   XZ:f:1
SRR070230_trimmed_155968_9    4    *    0    0    *    *    0    0
    TGCTGGGATTACAGGCATGA  IIIIIIIIIIIIIIIIII  XM:i:1000    XX:i:0
    XY:Z:M   XZ:f:1
SRR070230_trimmed_155968_8    4    *    0    0    *    *    0    0
    TGCTGGGATTACAGGCATGA  IIIIIIIIIIIIIIIIII  XM:i:1000    XX:i:0
    XY:Z:M   XZ:f:1
```

Παρατηρεί κανείς ότι το butter δεν ευθυγραμμίζει καθόλου τις παραπάνω αλληλουχίες στο γονιδίωμα. Η ετικέτα «XM:i:1000» σημαίνει πως για τη συγκεκριμένη αλληλουχία, υπάρχουν πάνω από χίλιες έγκυρες περιοχές ευθυγράμμισης πάνω στο γονιδίωμα αναφοράς. Επειδή η παράμετρος --max_rep έχει οριστεί στο 1000, παρόλο που μπορούν να υπάρχουν σχηματισμένα clusters, η εγγραφή δεν ευθυγραμμίζεται στο γονιδίωμα λόγω του φράγματος αυτού.

7.2.4 hsa-miR-4508

Από τη βάση δεδομένων miRBase (έκδοση 21), το συγκεκριμένο miRNA φέρει τα παρακάτω χαρακτηριστικά:

Πίνακας 12: Τα βασικά χαρακτηριστικά του ανθρωπίνου miR-4508

Όνομα miRNA αλληλουχίας	hsa-miR-4508
Mature αλληλουχία	**GCGGGGCTGGGCGCGCG**
Precursor αλληλουχία	AGGACCCA**GCGGGGCTGGGCGCGCG**GAGCAGCGCTGGGTG CAGCGCCTGCGCCGGCAGCTGCAAGGGCCG
Χρωμόσωμα	15
Κλώνος	-
Mature Start Loci	23562107
Mature Stop Loci	23562123
Precursor Start Loci	23562062
Precursor Stop Loci	23562131

Ανάλυση του FASTQ αρχείου

Παρακάτω παρουσιάζονται κάποια αποτελέσματα αναζητήσεων του «καθαρισμένου» από αλληλουχίες πρόσδεσης αρχείου fastq με αναγνωριστικό μελέτης SRR070230, για αλληλουχίες σχετικές με αυτήν του miRNA «hsa-miR-4508».

Πίνακας 13: Αναζήτηση της hsa-miR-4508 αλληλουχίας σε αρχείο fastq

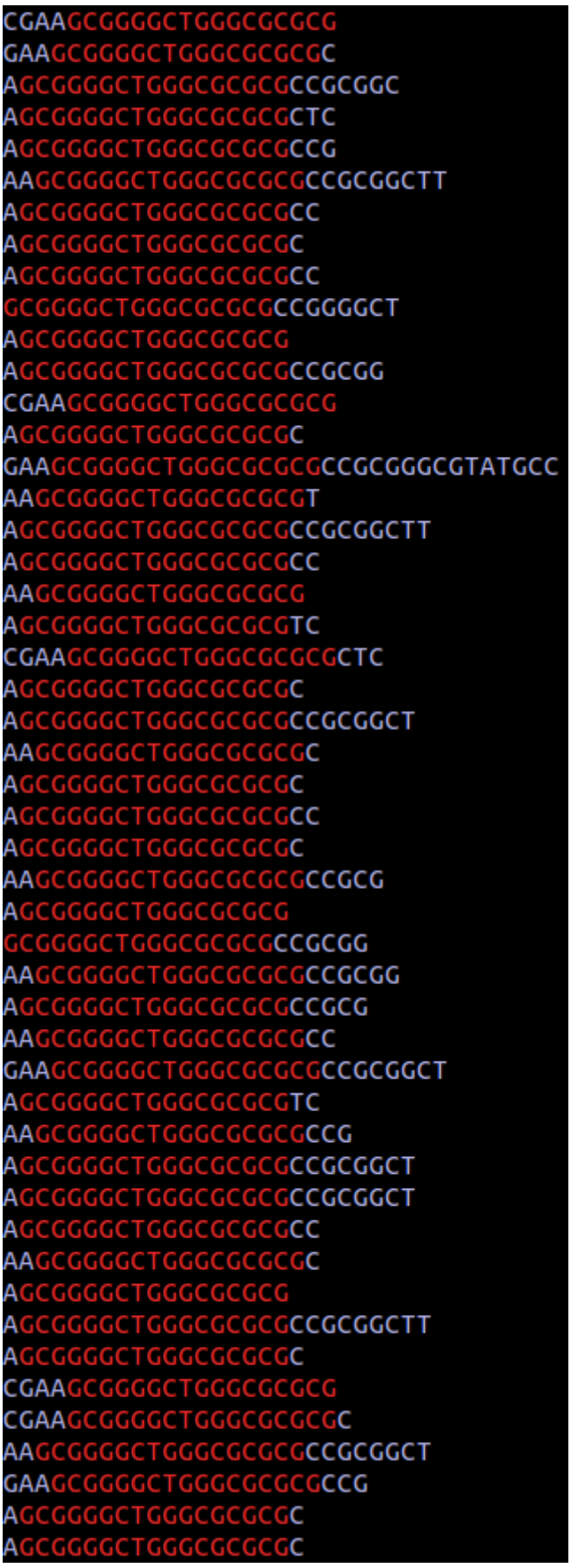
	hsa-miR-4508	Αριθμός εμφάνισης	Σχόλιο
miRNA αλληλουχία	GCGGGGCTGGGCGCGCG	4753	
Αντεστραμμένη συμπληρωματική αλληλουχία miRNA	CGCGCGCCCAGCCCCGC	0	

Η ακριβής αλληλουχία του miR-4508 χωρίς τα flanking νουκλεοτίδια, δεν εμφανίζεται ούτε μία φορά στο αρχείο εισόδου fastq.

```
cat SRR070230_trimmed.fastq | grep '^GCGGGGCTGGGCGCGCG$' | wc -l

0
```

Η αλληλουχία η οποία εμφανίζεται 784 φορές στο αρχείο εισόδου με μία αναντιστοιχία σε σχέση με το ώριμο miR-4508 και με ένα επιπλέον νουκλεοτίδιο που ταιριάζει με αυτό της αλληλουχίας του προδρόμου, παρουσιάζεται παρακάτω.

```
cat SRR070230_trimmed.fastq | grep '^AGCGGGGCTGGGCGCGCGC$' | wc
-l

784
```



Εικόνα 19: Μέρος των εγγραφών του fastq αρχείου

Ανάλυση του SAM αρχείου από την εκτέλεση του εργαλείου Butter

Μέρος του αρχείου SAM από την εκτέλεση του εργαλείου butter παρουσιάζεται παρακάτω.

```
SRR070230_trimmed_52269_1 0 8 140635762 255 21M * 0 0
AGCGGGGCTGGGCGCGCGGTC IIIIIIIIIIIIIIIIIIII XA:i:1 MD:Z:19G1 NM:i:1
XX:i:1 XY:Z:U XZ:f:1
SRR070230_trimmed_13459_1 0 8 140635762 255 19M * 0 0
AGCGGGGCTGGGCGCGCGG IIIIIIIIIIIIIIIIIII XA:i:0 MD:Z:19 NM:i:0 XX:i:2
XY:Z:P XZ:f:0.488
SRR070230_trimmed_136404_326 0 8 140635762 255 18M * 0 0
AGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIII XA:i:0 MD:Z:18 NM:i:0 XX:i:9
XY:Z:P XZ:f:0.026
SRR070230_trimmed_100963_7 0 19 8335366 255 20M * 0 0
GCGGGGCTGGGCGCGCGCTC IIIIIIIIIIIIIIIIIIII XA:i:1 MD:Z:0C19 NM:i:1
XX:i:7 XY:Z:P XZ:f:0.709
SRR070230_trimmed_100963_6 0 19 8335366 255 20M * 0 0
GCGGGGCTGGGCGCGCGCTC IIIIIIIIIIIIIIIIIIII XA:i:1 MD:Z:0C19 NM:i:1
XX:i:7 XY:Z:P XZ:f:0.709
SRR070230_trimmed_505625_1 0 21 8216787 255 31M * 0 0
GGCTGGGGCGGGAAGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:1
MD:Z:10C20 NM:i:1 XX:i:7 XY:Z:P XZ:f:0.343
SRR070230_trimmed_229023_1 0 21 8216791 255 27M * 0 0
GGGGCGCGAAGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:0 MD:Z:27
NM:i:0 XX:i:7 XY:Z:P XZ:f:0.343
SRR070230_trimmed_254617_6 0 21 8216795 255 23M * 0 0
CGCGAAGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIIIIIIII XA:i:0 MD:Z:23 NM:i:0
XX:i:7 XY:Z:P XZ:f:0.274
SRR070230_trimmed_254617_5 0 21 8216795 255 23M * 0 0
CGCGAAGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIIIIIIII XA:i:0 MD:Z:23 NM:i:0
XX:i:7 XY:Z:P XZ:f:0.274
SRR070230_trimmed_771783_1 0 KI270733.1 133111 255 33M * 0 0
CGGGCTGGGGCGCGAAGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
XA:i:0 MD:Z:33 NM:i:0 XX:i:7 XY:Z:P XZ:f:0.438
SRR070230_trimmed_843489_1 0 KI270733.1 133114 255 30M * 0 0
GCTGGGGCGCGAAGCGGGGCTGGGCGCGCG IIIIIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:0
MD:Z:30 NM:i:0 XX:i:7 XY:Z:P XZ:f:0.438
SRR070230_trimmed_778463_1 0 KI270733.1 133115 255 30M * 0 0
CTGGGGCGCGAAGCGGGGCTGGGCGCGCGT IIIIIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:1
MD:Z:29C0 NM:i:1 XX:i:7 XY:Z:P XZ:f:0.438
SRR070230_trimmed_256869_1 0 KI270733.1 133117 255 28M * 0 0
GGGGCGCGAAGCGGGGCTGGGCGCGCGC IIIIIIIIIIIIIIIIIIIIIIIIIIII XA:i:0
MD:Z:28 NM:i:0 XX:i:7 XY:Z:P XZ:f:0.438
```

Παρατηρεί κανείς πως μέρος της αλληλουχίας του miR-4508 εμφανίζεται σε αρκετές θέσεις του γονιδιώματος που ανήκουν σε διαφορετικά χρωμοσώματα. Για τις συγκεκριμένες θέσεις, δεν έχει βρεθεί ο αντίστοιχος σχολιασμός, αλλά για μερικές περιοχές οι οποίες πληρούν τα απαραίτητα κριτήρια, έχουν σημειωθεί καινούργιες περιοχές έκφρασης στο αντίστοιχο αρχείο εξόδου. Παρακάτω, παρουσιάζεται μία γραμμή του αρχείου για την περιοχή έκφρασης του χρωμοσώματος 21 στις θέσεις 8216650-8217039.

```
21 + 8216650 8217039 234 no_annotation no_annotation
CCGGGCCGTACCCATATCCGCAGCAGGTCTCCAAGGTGAACAGCCTCTGGCATGTTGGAACAATGTAGGT
```

```
AAGGGAAGTCGGCAAGCCGGATCCGTAACTTCGGGATAAGGATTGGCTCTAAGGGCTGGGTCGGTCGGGC
TGGGGCGCGAAGCGGGGCTGGGCGCGCGCCGCGGCTGGACGAGGCGCCGCCGCCCCCCCCACGCCCGGGG
CACCCCCCTCGCGGCCCTCCCCCGCCCCACCCCGCGCGCGCCGCTCGCTCCCTCCCCACCCCGCGCCCTC
TCTCTCTCTCTCTCCCCCGCTCCCCGTCCTCCCCCCTCCCCGGGGGAGCGCCGCGTGGGGGCGGCGGCGG
GGGGAGAAGGGTCGGGGCGGCAGGGGCCGGCGGCGGCCGC
```

Σύμφωνα με το spipeRNA, η αλληλουχία miR-4508 εμφανίζεται 7 φορές, όπως προκύπτει από μέρος του αρχείου SAM, στο οποίο η πρώτη κατά σειρά εγγραφή, είναι μία εγγραφή που ευθυγραμμίζεται ακριβώς μία φορά πάνω στο γονιδίωμα, δημιουργώντας μία πυκνότητα εγγραφών στη συγκεκριμένη θέση.

```
SRR070230_trimmed_772025_1 16 15 23562103 255 22M * 0 0
GCTTCGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIIIIII XA:i:1 MD:Z:3C18 NM:i:1
XX:i:1 XY:Z:U XZ:f:1
SRR070230_trimmed_517953_4 16 15 23562106 255 19M * 0 0
ACGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIII XA:i:1 MD:Z:0C18 NM:i:1 XX:i:9
XY:Z:P XZ:f:0.014
SRR070230_trimmed_517953_3 16 15 23562106 255 19M * 0 0
ACGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIII XA:i:1 MD:Z:0C18 NM:i:1 XX:i:9
XY:Z:P XZ:f:0.014
SRR070230_trimmed_136404_357 16 15 23562107 255 18M * 0 0
CGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIII XA:i:0 MD:Z:18 NM:i:0 XX:i:9
XY:Z:P XZ:f:0.014
SRR070230_trimmed_136404_354 16 15 23562107 255 18M * 0 0
CGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIII XA:i:0 MD:Z:18 NM:i:0 XX:i:9
XY:Z:P XZ:f:0.014
SRR070230_trimmed_136404_313 16 15 23562107 255 18M * 0 0
CGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIII XA:i:0 MD:Z:18 NM:i:0 XX:i:9
XY:Z:P XZ:f:0.014
SRR070230_trimmed_136404_255 16 15 23562107 255 18M * 0 0
CGCGCGCCCAGCCCCGCT IIIIIIIIIIIIIIIIII XA:i:0 MD:Z:18 NM:i:0 XX:i:9
XY:Z:P XZ:f:0.014
```

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η τεχνολογία της αλληλούχησης επόμενης γενιάς αξιοποιείται από ερευνητές για την αντιμετώπιση του ολοένα και πιο ποικίλου φάσματος των βιολογικών προβλημάτων. Τα μικρά μη κωδικά RNAs, αποτελούν ένα σημαντικό κομμάτι του μεταγραφώματος και κερδίζουν όλο και μεγαλύτερο επιστημονικό ενδιαφέρον. Μόλις τώρα, αρχίζουμε να συνειδητοποιούμε τη φύση και τη συμμετοχή των μικρών αυτών μορίων, στις φυσιολογικές και παθολογικές διεργασίες.

Πρόκληση αποτελεί η δημιουργία απλών και εύκολων στη χρήση εργαλείων, που να μπορούν να ανταποκρίνονται στη φύση των δεδομένων NGS, λόγω του συνήθως μεγάλου τους μεγέθους και ταυτόχρονα, να μπορούν να παράγουν λογικά και απλά δεδομένα εξόδου, ικανά για περαιτέρω ανάλυση. Τα υπάρχοντα εργαλεία ανάλυσης των small RNA-Seq δεδομένων που παρουσιάστηκαν στη συγκεκριμένη εργασία, χειρίζονται τις NGS εγγραφές με παρόμοιο τρόπο. Τα περισσότερα εργαλεία εστιάζουν ή ευνοούν συγκεκριμένα RNAs έναντι άλλων, ενώ όπως υποδείχθηκε στην ενότητα 7.2, κάτι τέτοιο μπορεί να οδηγήσει σε λανθασμένα προφίλ έκφρασης.

Το spipeRNA, είναι μία ολοκληρωμένη πρόταση για την ανάλυση των small RNA-Seq δεδομένων, το οποίο βασίζεται σε δωρεάν διαθέσιμα εργαλεία και υλοποιεί στατιστικές προσεγγίσεις σε μία αυτοματοποιημένη ροή. Το spipeRNA είναι εύκολο στη χρήση εργαλείο, χωρίς περιττές βάσεις δεδομένων ή πολύπλοκες και δύσκολες στην εγκατάσταση βιβλιοθήκες, και το οποίο δεν περιορίζει το μέγεθος των αρχείων εισόδου, ενώ ταυτόχρονα επιτρέπει παράλληλη εκτέλεση πολλών δειγμάτων. Η λογική και το φορμάτ των δεδομένων εξόδου επιτρέπει στο χρήστη την περαιτέρω ανάλυση των δεδομένων, όπως είναι η διαφορική έκφραση, έλεγχος των εκφρασμένων μη κωδικών RNAs, αναγνώριση των μεταγράφων των εμπλουτισμένων περιοχών κα.

Αν και το spipeRNA είναι ικανό να ποσοτικοποιήσει διαφόρων ειδών μη-κωδικά RNAs παρόντα σε ένα βιολογικό δείγμα, η επέκταση και η βελτίωση ενός εργαλείου είναι μία συνεχής διαδικασία. Μία ενδιαφέρουσα και χρήσιμη προσθήκη θα ήταν η δυνατότητα πρόβλεψης των RNA ειδών που προκύπτουν από τις μη σχολιασμένες περιοχές. Κάτι τέτοιο θα μπορούσε να επιτευχθεί για παράδειγμα, με μεθόδους μηχανικής μάθησης με την ομαδοποίηση των διάφορων κατηγοριών ncRNA με βάση για παράδειγμα τη δευτεροταγή δομή τους, το μήκος τους κτλ.. Επιπλέον, θα μπορούσε να προστεθεί ένας πρόσθετος έλεγχος των εγγραφών οι οποίες δεν είχαν ευθυγραμμιστεί στο γονιδίωμα σε πρώτη φάση. Κάποιο μέρος των εγγραφών αυτών, μπορεί να αποτελούν πραγματικά μετάγραφα που φέρουν διάφορες μεταλλάξεις όπως είναι διαγραφή και εισαγωγή νουκλεοτιδίου/νουκλεοτιδίων. Τέλος, θα μπορούσαν να προστεθούν και άλλοι τρόποι υπολογισμού των καινούργιων περιοχών έκφρασης. Ο αλγόριθμος του NiBLS είναι ένα σχετικό παράδειγμα σχηματισμού συστάδων από εγγραφές με τη βοήθεια των γράφων.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
de novo	Εκ νέου, εδώ πρόκειται για μέθοδο δημιουργίας μεταγραφώματος χωρίς γονιδίωμα αναφοράς
multi-mapped reads	Έγγραφές με πολλαπλές θέσεις ευθυγράμμισης πάνω στο γονιδίωμα
script	Αρχείο γραμμένο με γλώσσα προγραμματισμού σεναρίων
state-of-the-art	Προχωρημένο, τελευταίας τεχνολογίας
random forest	Αλγόριθμος «τυχαίου δάσους»
precursor	Εδώ, πρόδρομος αλληλουχία ενός miRNA
mature	Εδώ, ώριμη αλληλουχία ενός miRNA
pipeline	Ροή εργασιών, εργαλείο
tag	Ετικέτα ή εγγραφή
flanking	Εδώ, τα πέριξ νουκλεοτίδια

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

NGS	Next Generation Sequencing
BED	Browser Extensible Data
SAM	Sequence Alignment/Map
HMM	Hidden Markov Model
RMSD	Root-Mean-Square Deviation

ΠΑΡΑΡΤΗΜΑ Ι

Κώδικας

Ο κώδικας και όλο το συνοδευτικό υλικό βρίσκεται στο CD.

ΠΑΡΑΡΤΗΜΑ ΙΙ

Προεργασία των NGS δεδομένων

Τα small RNA-Seq δεδομένα του Πίνακα 1, αντλήθηκαν από τη βάση δεδομένων GEO. Πριν την εκτέλεση του spipeRNA, τα δεδομένα εισόδου πρέπει να «καθαριστούν» από τις πρόσθετες αλληλουχίες πρόσδεσης και να ελεγχθεί η ποιότητά τους. Οι adapters αφαιρέθηκαν με τη βοήθεια του εργαλείου cutadapt με την παρακάτω εντολή,

```
cutadapt – a adapter_sequence -O 3 -m 18 -q 10 -o input output
```

και στη συνέχεια εφαρμόστηκε ο επιπλέον έλεγχος ποιότητας των εγγραφών, με τη βοήθεια του εργαλείου FastQC.

Προσομοίωση των δεδομένων

Οι εγγραφές προσομοιώθηκαν με τη βοήθεια του εργαλείου art-illumina. Οι αλληλουχίες που χρησιμοποιήθηκαν, αντλήθηκαν από το γονιδίωμα, με βάση τις γονιδιωματικές συντεταγμένες, διαθέσιμες στο σχολιασμένο αρχείο των μη κωδικών RNAs. Εξαιρέθηκαν οι αλληλουχίες των lncRNAs και των miRNA προδρόμων. Σημειώνεται, πως στις συντεταγμένες προστέθηκαν ±2 nt, κάτι που ταιριάζει στο προφίλ των μεταγράφων. Τέλος, το fasta αρχείο δημιουργήθηκε με την παρακάτω εντολή:

```
art-illumina -ss HS20 -I transcripts_sequences.fasta -k 0 -l 22
-f 100 -o transcript_sequence
```

ΑΝΑΦΟΡΕΣ

[1] Giurato G, De Filippo MR, Rinaldi A, Hashim A, Nassa G, Ravo M, Rizzo F, Tarallo R, Weisz A, "iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq", *BMC Bioinformatics* 2013 Dec 13

[2] Mario Fasold, David Langenberger, Hans Binder, Peter F. Stadler and Steve Hoffmann, "DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments", *Nucleic Acids Research* 2011 Jul 1; 39(Web Server issue): W112–W117

[3] Choudhuri S, "Small noncoding RNAs: biogenesis, function and emerging significance in toxicology", *J Biochem Mol Toxicol*. 2010 May-Jun;24(3):195-216. doi: 10.1002/jbt.20325

[4] Theresa Phillips, "Small non-coding RNA and gene expression", *Nature Education 1(1):115*

[5] Yong Huang, Ji Liang Zhang, Xue Li Yu, Ting Sheng Xu, Zhan Bin Wang, and Xiang Chao Cheng, "Molecular Functions of Small Regulatory Noncoding RNA", *Biochemistry,* March 2013, Volume 78, Issue 3, pp 221–230

[6] Po-Jung Huang, Yi-Chung Liu, Chi-Ching Lee, Wei-Chen Lin, Richie Ruei-Chi Gan, Ping-Chiang Lyu and Petrus Tang, "DSAP: deep-sequencing small RNA analysis pipeline", *Nucleic Acids Research* 2010 Jul 1; 38(Web Server issue): W385–W391

[7] Γ. Παπανικολάου, "Εργαστηριακές Ασκήσεις Γενετικής του Ανθρώπου στον Υπολογιστή και στον Πάγκο", 2015

[8] Vikas Gupta, Katharina Markmann, Christian N. S. Pedersen, Jens Stougaard, Stig U. Andersen, "shortran: a pipeline for small RNA-Seq data analysis", *Bioinformatics* 2012 Oct 15; 28(20): 2698–2700

[9] Michael J. Axtell, "Butter: High-precision genomic alignment of small RNA-Seq data", 2014

[10] Tao Wang, Beibei Chen, MinSoo Kim, Yang Xie, Guanghua Xiao, "A Model-Based Approach to Identify Binding Sites in CLIP-Seq Data", *PLOS Journals*, April 8, 2014

[11] Marc R. Friedländer, Sebastian D. Mackowiak, Na Li, Wei Chen, Nikolaus Rajewsky, "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades", *Nucleic Acids Res.* 2012 Jan; 40(1): 37–52

[12] Marc R Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel1 & Nikolaus Rajewsky, "Discovering microRNAs from deep sequencing data using miRDeep*", Nature Biotechnology* 26, 407 - 415 (2008)

[13] Michiel J.L. de Hoon, Ryan J. Taft, Takehiro Hashimoto, Mutsumi Kanamori-Katayama, Hideya Kawaji, Mitsuoki Kawano, Mami Kishima, Timo Lassmann, Geoffrey J. Faulkner, John S. Mattick, Carsten O. Daub, Piero Carninci, Jun Kawai, Harukazu Suzuki, and Yoshihide Hayashizaki, "Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries", *Genome Res*. 2010 Feb; 20(2): 257–264.

[14] Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM, "miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments.", *Nucleic Acids Res.* 2009 Jul;37(Web Server issue):W68-76