ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΕΡΓΑΣΤΗΡΙΟ ΚΥΤΤΑΡΟΓΕΝΕΤΙΚΗΣ ΚΑΙ ΜΟΡΙΑΚΗΣ ΓΕΝΕΤΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΓΕΝΕΤΙΚΗ ΤΟΥ ΑΝΘΡΩΠΟΥ-ΓΕΝΕΤΙΚΗ ΣΥΜΒΟΥΛΕΥΤΙΚΗ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Εργαλεία βιοπληροφορικής στην ανάλυση του μεταγραφώματος (RNA Seq)»

ΤΑΣΣΟΠΟΥΛΟΥ ΡΑΦΑΕΛΑ ΑΝΑΣΤΑΣΙΑ
ΜΕΤΑΠΤΥΧΙΑΚΗ ΦΟΙΤΗΤΡΙΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ
Μπαλής Χαράλαμπος, PH.D. ΕΔΙΠ, Επιβλέπων
Παπαθανασίου Ιωάννα, Επικ. Καθ. Ιατρικής Βιολογίας, Μέλος
Τσέζου Ασπασία, Καθ. Ιατρ. Γενετικής , Μέλος

ΛΑΡΙΣΑ, 2021

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

UNIVERSITY OF THESSALY
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE

POSTGRADUATE MASTER PROGRAM
"HUMAN GENETICS – GENETIC COUNSELING"

MASTER'S THESIS

«Bioinformatics' tools for RNA sequencing analysis»

TASSOPOULOU RAFAELA ANASTASIA

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Ευχαριστίες

Θα ήθελα να ευχαριστήσω του καθηγητές μου Χαράλαμπο Μπαλή, Ιωάννα Παπαθανασίου και Ασπασία Τσέζου για την καθοδήγηση και την πολύτιμη βοήθεια τους.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Περίληψη

Η αλληλούχηση επόμενης γενιάς (NGS) αποτελεί ένα σημαντικό εργαλείο των επιστημόνων. Πολλές έρευνες χρησιμοποιούν το συγκεκριμένο εργαλείο για ανάλυση του γονιδιώματος, του μεταγραφώματος, του επιγενώματος αλλά και πρωτεομική ανάλυση.

Ειδικότερα, για την μελέτη του μεταγραφώματος χρησιμοποιείται η αλληλούχηση του RNA (RNA-Seq), η οποία θα αναλυθεί στη συγκεκριμένη διπλωματική εργασία. Για την RNA-Seq είναι απαραίτητη η απομόνωση του RNA, ο κατακερματισμός του, η δημιουργία βιβλιοθήκης, η προσθήκη adapters, η ενίσχυση των τμημάτων και η επιλογή των κατάλληλων εργαλείων για την αλληλούχηση και την ανάλυση των δεδομένων που προκύπτουν από αυτή.

Η δημοφιλέστερη πλατφόρμα αλληλούχησης έχει δημιουργηθεί από την εταιρεία Illumina και η μέθοδος που χρησιμοποιείται είναι η αλληλούχηση μέσω σύνθεσης. Η ανάλυση αυτή πραγματοποιείται συχνότερα για μελέτες της διαφορετικής έκφρασης των γονιδίων μεταξύ των δειγμάτων και μελέτες για την ανακάλυψη νέων ισομορφών γονιδίων.

Για την ανάλυση των δεδομένων είναι απαραίτητη η συνεισφορά των εργαλείων βιοπληροφορικής. Μερικά από τα εργαλεία βιοπληροφορικής που χρησιμοποιούνται είναι το FASTQC που χρησιμοποιείται για τον ποιοτικό έλεγχο των δεδομένων, το STAR που χρησιμοποιείται για το alignment των αλληλουχιών, το featureCounts που χρησιμοποιείται για τον υπολογισμό των reads και το edgeR που χρησιμοποιείται για την ανάλυση της διαφορετικής έκφρασης των γονιδίων. Αυτά είναι μόνο μερικά από τα εργαλεία που χρησιμοποιούνται και αναλύονται παρακάτω, παρόλα αυτά υπάρχουν διάφορα εργαλεία, καθένα από τα οποία εμφανίζει διαφορετικά μειονεκτήματα και πλεονεκτήματα ανάλογα με το ερευνητικό ερώτημα που τίθεται.

Συνεπώς, η ανάπτυξη της τεχνολογίας δίνει τη δυνατότητα ανάλυσης μεγάλης κλίμακας δεδομένων που προκύπτουν από τις μεθόδους αλληλούχησης επόμενης γενιάς, όμως μέχρι σήμερα δεν υπάρχει «χρυσό» εργαλείο βιοπληροφορικής, για κάθε ερευνητικό ερώτημα.

Λέξεις- Κλειδιά**:** FASTQC, Spliced Transcripts Alignment to Reference (STAR), featureCounts, Empirical Analysis of Digital Gene Expression Data in R (edgeR)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Abstract

Next Generation Sequencing is an important tool for scientists. Many studies use this tool for genomics, epigenomics, transcriptomics and proteomics analysis.

Particularly, RNA-Sequencing is used for transcriptomics analysis, which will be described in this study. RNA-Seq analysis consists of RNA isolation, RNA fragmentation, library preparation, adapters ligation, amplification, the decision about the best bioinformatics' tool and raw data analysis.

The most famous sequencing platform is designed by the company Illumina which uses sequencing-by-synthesis method. This analysis is used to identify differential expressed genes or different isoforms of genes.

Bioinformatics' tools are really important for analyzing data from RNA-Seq. Some of them are FASTQC for data quality control, STAR for alignment, featureCounts for number read counts and edgeR for differential gene expression analysis. These tools are used and analyzed in this study and each one has its benefits and its drawbacks, depending on biological questions.

Therefore, the development of technology enables "big data" analysis from Next Generation Sequencing , but until now there isn't a "golden" bioinformatics' tool for every biological question.

Key words: FASTQC, Spliced Transcripts Alignment to Reference (STAR), featureCounts, Empirical Analysis of Digital Gene Expression Data in R (edgeR)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Πίνακας περιεχομένων

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Εισαγωγή

Η ραγδαία ανάπτυξη της τεχνολογίας και της επιστήμης αποτελούν αλληλένδετες έννοιες που στοχεύουν στη βελτίωση του τρόπου ζωής των ατόμων. Μέσω αυτών ανακαλύπτονται γρηγορότερα και με μικρότερο κόστος σε σχέση με προηγούμενες δεκαετίες νέες θεραπείες, φάρμακα και εμβόλια. Βελτιώνεται ο τρόπος ζωής των ανθρώπων ενώ ταυτόχρονα το προσδόκιμο ζωής συνεχώς αυξάνεται. Νέα τεχνολογικά επιτεύγματα έρχονται στο φως κάνοντας πολλές φορές δύσκολη την κατανόηση τους από τον ανθρώπινο εγκέφαλο. Η πληροφορία και τα δεδομένα είναι πλέον τεράστια για να αποθηκευτούν, οργανωθούν και αναλυθούν από τον άνθρωπο και τότε έρχεται στο προσκήνιο η ανάλυση των βιολογικών δεδομένων μέσω υπολογιστικών εργαλείων. Τα προγράμματα όπως το HapMap, το Human Genome Project, το 100.000 Genome Project, το ENCODE και άλλα αποτελούν μόνο μερικά από όλα τα προγράμματα που για να πραγματοποιηθούν απαιτούν μεγάλο όγκο δεδομένων και τα εργαλεία του παρελθόντος δεν μπορούν να ανταποκριθούν επαρκώς.

Το πρόβλημα αναλαμβάνει να λύσει ο κλάδος της βιοπληροφορικής, μίας ταχέως αναπτυσσόμενης επιστήμης. Ως βιοπληροφορική ορίζεται η επιστήμη που εφαρμόζει υπολογιστικές μεθόδους με σκοπό την οργάνωση, διαχείριση και εν τέλει την κατανόηση της πληροφορίας που συνδέεται με βιομόρια (Κοσσίδα, 2008). Μέσω αυτής ανακαλύπτονται διαρκώς νέα εργαλεία που βοηθούν τους επιστήμονες να κατανοήσουν καλύτερα τα δεδομένα που προκύπτουν. Κάθε ερευνητικό ερώτημα και κάθε πείραμα έχει συγκεκριμένους περιορισμούς που δυσχεραίνουν την επιστημονική διεργασία, συνεπώς κάθε εργαλείο που υπάρχει δίνει λύση σε ένα διαφορετικό ερευνητικό πρόβλημα. Η πληθώρα των εργαλείων αποτελεί όμως και τροχοπέδη κάνοντας περίπλοκη την επιλογή του κατάλληλου εργαλείου βιοπληροφορικής για κάθε μελέτη. Παρόλα αυτά τα εργαλεία αυτά σε συνδυασμό με την ανακάλυψη νέων τεχνολογιών όπως η αλληλούχηση επόμενης γενιάς (NGS) ανοίγουν νέο δρόμο στη μελέτη των βιολογικών συστημάτων. Δημιουργούνται ερευνητικοί κλάδοι που ασχολούνται με τη γενομική (genomics), την επιγενετική (epigenomics), τη μεταγραφωματική (transcriptomics), τη πρωτεομική (proteomics), τη μικροβιοματική (microbiomics) και τη μεταβολομική (metabolomics). Η αλληλούχηση επόμενης γενιάς αποτελεί βασικό κομμάτι των κλάδων αυτών.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Η ανάλυση του DNA και του RNA των οργανισμών δίνει σημαντικά στοιχεία για τον τρόπο που αυτοί λειτουργούν και δίνει απαντήσεις σε πολλά ερευνητικά ερωτήματα. Γίνεται ανάλυση των κωδικών και μη κωδικών περιοχών του γονιδιώματος, των ρυθμιστικών στοιχείων τους, του τρόπου μου αλληλοεπιδρούν μεταξύ τους, των πρωτεϊνών τους και άλλων. Αποκτάται έτσι μεγάλος όγκος πληροφοριών που δίνουν απαντήσεις τόσο για το τρόπο λειτουργίας των οργανισμών που μελετώνται αλλά και για τις εξελικτικές σχέσεις που υπάρχουν. Είναι επίσης δυνατή η μελέτη συσχέτισης γονοτύπου- φαινοτύπου και η μελέτη των επιπτώσεων του περιβάλλοντος στην εμφάνιση διαφόρων ασθενειών.

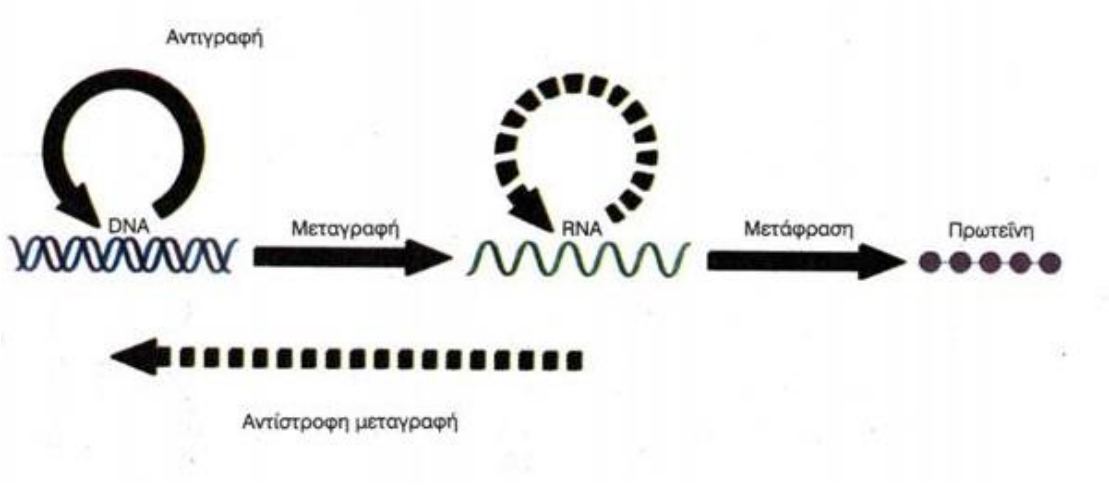
Ο κλάδος της βιοπληροφορικής σε συνδυασμό με την αλληλούχηση επόμενης γενιάς αποτελούν βασικό κομμάτι της σύγχρονης επιστημονικής έρευνας. Τα εργαλεία βιοπληροφορικής για την ανάλυση των δεδομένων που προκύπτουν είναι πολλά και καθένα έχει διαφορετικά μειονεκτήματα και πλεονεκτήματα. Έτσι ο ερευνητής καλείται να επιλέξει τα εργαλεία που είναι κατάλληλα για το ερευνητικό ερώτημα που θέλει να λύσει μέσω μίας πληθώρας εργαλείων. Στη συγκεκριμένη μελέτη γίνεται μία ανασκόπηση μερικών εργαλείων βιοπληροφορικής που χρησιμοποιούνται για την ανάλυση των δεδομένων που προκύπτουν από την αλληλούχηση του RNA των οργανισμών που μελετώνται. Το RNA-Seq όπως θα αναφερθεί και παρακάτω αποτελεί βασικό εργαλείο των επιστημόνων για τη μελέτη του μεταγραφώματος και έχουν δημιουργηθεί πολλά εργαλεία βιοπληροφορικής για την ανάλυση των δεδομένων που δημιουργούνται από την αλληλούχηση. Ο βασικός σχεδιασμός ενός πειράματος RNA-Sequencing και τα βασικά εργαλεία ανάλυσης των δεδομένων περιγράφονται στη παρακάτω εργασία. Αναφέρονται μερικά από τα βασικά πλεονεκτήματα και μειονεκτήματα αυτών, ο τρόπος με τον οποίο λειτουργούν καθώς και η βασική αρχή πάνω στην οποία σχεδιάστηκαν. Στόχος της διπλωματικής είναι η εξοικείωση με τη μέθοδο του RNA-Sequencing, η ανασκόπηση των εργαλείων ανάλογα με το ερευνητικό ερώτημα με στόχο την επιλογή του καλύτερου εργαλείου και η ανάλυση των βασικότερων εργαλείων, τα οποία και χρησιμοποιούνται συχνότερα από την ερευνητική κοινότητα.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Γενικό Μέρος

Η γενετική πληροφορία μεταφέρεται από το DNA στο RNA με τη βοήθεια της μεταγραφής (Εικόνα 1). Το σύνολο των RNA που παράγονται κατά τη μεταγραφή σε μία συγκεκριμένη χρονική στιγμή αποτελούν το μεταγράφωμα (Chatterjee et al., 2018). Το μεταγράφωμα αποτελείται από mRNAs, non-coding RNAs, small RNAs, microRNAs και άλλα. Η ανάλυση του μεταγραφώματος αποτελεί βασικό αντικείμενο μελέτης, με στόχο την κατανόηση των λειτουργικών στοιχείων του γονιδιώματος και των συστατικών που αποτελούν τα κύτταρα. Οι διαφορές που εμφανίζουν τα κύτταρα στο επίπεδο του μεταγραφώματος, δηλαδή η διαφορετική έκφραση των γονιδίων μπορεί να δώσει απαντήσεις σε πολλά ερωτήματα που σχετίζονται με την ανάπτυξη των οργανισμών και τη δημιουργία ασθενειών.

Εικόνα από (*Μοριακή Γενετική*, n.d.)



Εικόνα 1: Το κεντρικό δόγμα της μοριακής βιολογίας

Ανάλυση και η ποσοτικοποίηση του μεταγραφώματος

Η ανάλυση και η ποσοτικοποίηση του μεταγραφώματος γίνονται με διάφορους τρόπους όμως δύο είναι οι κύριες τεχνολογίες στις οποίες βασίζονται όλοι. Η υβριδοποίηση και η αλληλούχιση.

Υβριδοποίηση

Η υβριδοποίηση περιλαμβάνει τη χρήση microarrays και αποτελεί μια οικονομική τεχνική με υψηλή απόδοση, όμως απαιτείται προηγούμενη γνώση για την αλληλουχία ενδιαφέροντος. Άλλοι περιορισμοί της μεθόδου είναι το υψηλό background, το

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

περιορισμένο εύρος της ανίχνευσης και η δυσκολία σύγκρισης των επιπέδων έκφρασης δειγμάτων που προέρχονται από διαφορετικά πειράματα.

Αλληλούχηση

Η ανάλυση του μεταγραφώματος με τη μέθοδο της αλληλούχησης γίνεται με τη βοήθεια της αλληλουχίας του cDNA που δημιουργείται. Αρχικά χρησιμοποιήθηκαν μέθοδοι αλληλούχησης με υψηλό κόστος, χαμηλή ανάλυση και μη ποσοτικές όπως η αλληλούχηση κατά Sanger και οι βιβλιοθήκες EST (Expressed Sequence Tags). Τα EST δημιουργήθηκαν στα τέλη της δεκαετίας του ’80 και βασίζονται στη δημιουργία ακολουθιών 200-500 βάσεων με ένα και μόνο ενιαίο πέρασμα αλληλούχησης και χαρακτηρίζονται από υψηλό ποσοστό λάθους (3%)(Κοσσίδα, 2008). Άλλες μέθοδοι που δημιουργήθηκαν ήταν το Serial Analysis of Gene Expression (SAGE), το Cap Analysis of Gene Expression (CAGE) και το Massively Parallel Signature Sequencing (MPSS) (Wang et al., 2009). Το SAGE και το MPSS βασίζονται σε ετικέτες (tags) που προκύπτουν από mRNA και χαρτογραφούνται με τη βοήθεια γονιδιακών βάσεων δεδομένων. Αυτές οι μέθοδοι έχουν υψηλή απόδοση και το βασικότερο είναι ότι μπορούν με ακρίβεια να υπολογίσουν τα επίπεδα έκφρασης των γονιδίων που μελετώνται. Βασικό πρόβλημα αποτελεί όμως ότι είναι κοστοβόροι, δεν επιτρέπουν την ακριβή χαρτογράφηση με το γονιδίωμα αναφοράς και δεν μπορούν να διακρίνουν τις διαφορετικές ισομορφές των mRNA των ευκαρυωτικών οργανισμών (Wang et al., 2009).

Αλληλούχηση επόμενης γενιάς

Η ανακάλυψη της αλληλούχησης επόμενης γενιάς (NGS) έφερε επανάσταση και έδωσε την ευκαιρία στους ερευνητές να μελετήσουν σε μικρό χρόνο ολόκληρα γονιδιώματα, να αλληλουχήσουν σε βάθος περιοχές με ιδιαίτερο ενδιαφέρον, να αναλύσουν επιγενετικούς παράγοντες και να ταυτοποιήσουν νέα παθογόνα. Ακόμα έδωσε τη δυνατότητα ανακάλυψης νέων παραλλαγών RNA, νέων θέσεων ματίσματος και ποσοτικοποίησης των mRNA μέσω του RNA Sequencing (RNA seq) (*Introduction to NGS*, n.d.). Έτσι η ανακάλυψη του RNA seq έχει ξεκάθαρα πλεονεκτήματα έναντι των άλλων μεθόδων για την ανάλυση του μεταγραφώματος. Τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου παρουσιάζονται στον πίνακα 1.

Πίνακας Table 1 από (Wang et al., 2009)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Πίνακας 1:Θετικά και αρνητικά για κάθε τεχνική (Tiling microarray, cDNA/EST sequencing, RNA-seq)

Technology	**Tiling microarray**	**cDNA or EST sequencing**	**RNA-seq**
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
cost for mapping transcriptomes of large genomes	High	High	Relatively low

RNA-Sequencing

Λόγω της ροής της γενετικής πληροφορίας καλύτερα μελετημένο είναι το mRNA, όμως η μελέτη και των υπόλοιπων ειδών RNA μπορεί να αποτελέσει σημαντικό εργαλείο για τους επιστήμονες (Chatterjee et al., 2018). Η μελέτη των υπόλοιπων ειδών του RNA έγινε πιο έντονη τα τελευταία χρόνια και κυρίως με το πρόγραμμα ENCODE (Encyclopedia of DNA Elements) το οποίο είχε σαν στόχο τη χαρτογράφηση περιοχών που μεταγράφονται και τη μελέτη των σχέσεων των μεταγραφικών παραγόντων, της δομής της χρωματίνης και των τροποποιήσεων των ιστονών. Μέσω του προγράμματος αυτού έγινε δυνατή η σύνδεση μεταξύ λειτουργικών στοιχείων όπως το mRNA και διαφόρων σπάνιων ασθενειών ή και καρκίνου. Έτσι η ανάλυση του συνόλου των RNAs μπορεί να δώσει απαντήσεις σε πολλά βιολογικά ερωτήματα διότι γίνεται ανάλυση όχι μόνο των γονιδίων που εκφράζονται αλλά και των μηχανισμών που ελέγχουν την έκφραση τους (Chatterjee et al., 2018, p. 36; Dunham et al., 2012). Η συνεισφορά του RNA-Seq είναι εμφανής και στο πρόγραμμα Cancer Genome Atlas που ξεκίνησε το 2006 και έχει καταφέρει να βελτιώσει την ικανότητα των επιστημόνων να προλαμβάνουν, διαγιγνώσκουν και να θεραπεύουν τον καρκίνο μέσω της μελέτης του μεταγραφώματος (Chatterjee et al., 2018, p. 36; *The Cancer Genome Atlas Program*, n.d.).

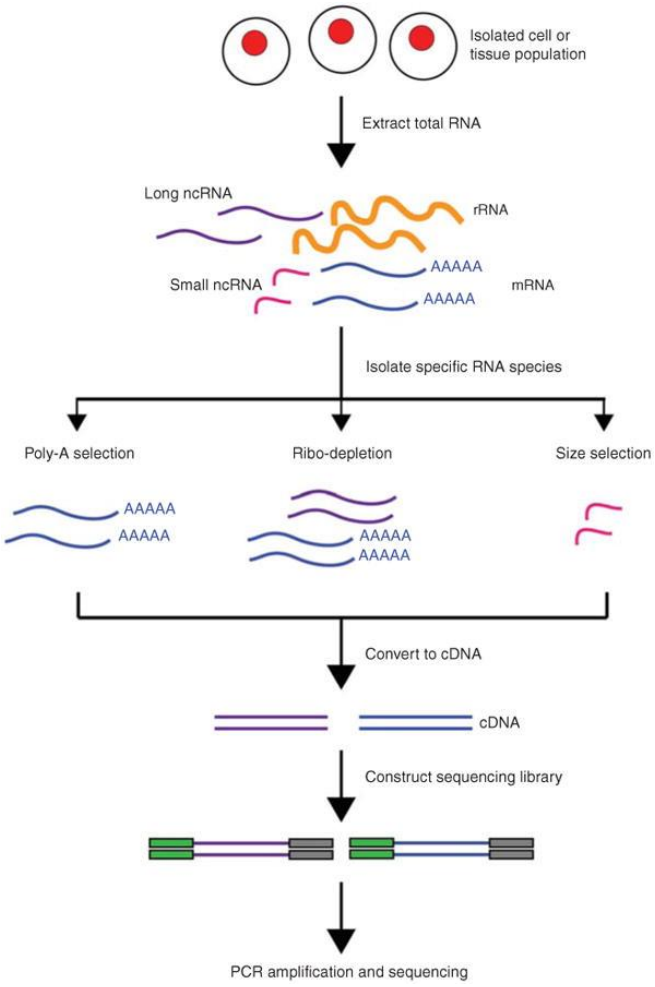
Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Η αλληλούχηση του RNA (RNA-Seq) είναι βασισμένη στη χρήση της αντίστροφης μεταγραφάσης, η οποία μετατρέπει το RNA σε cDNA (Illumina, n.d.). Υπάρχουν όμως πολλά διαφορετικά πρωτόκολλα τα οποία χρησιμοποιούνται ανάλογα με το ερευνητικό ερώτημα που πρόκειται να μελετηθεί. Όλες οι μέθοδοι όμως αποτελούνται από κάποια βασικά βήματα τα οποία είναι η απομόνωση του RNA, η προσθήκη adapters στο ένα ή και στα δυο άκρα του RNA, η δημιουργία της cDNA βιβλιοθήκης, η ενίσχυση του σήματος, η αλληλούχηση και τέλος η ανάλυση των δεδομένων που προκύπτουν (Wang et al., 2009).

Ο σχεδιασμός των βημάτων για την αλληλούχηση του RNA αποτελεί απαραίτητο στοιχείο για την επιτυχία της μεθόδου. Πρέπει να γίνει επιλογή της κατάλληλης βιβλιοθήκης, να εκτιμηθεί το επιθυμητό βάθος αλληλούχησης και να προσδιοριστεί ο αριθμός των αντιγράφων που χρειάζονται (Conesa et al., 2016). Ακόμα βασικά ερωτήματα για το σχεδιασμό είναι το ποιο μόριο RNA θα χρησιμοποιηθεί και πως θα απομονωθεί, πως θα γίνει η μετατροπή του RNA σε cDNA ώστε να έχουν συγκεκριμένο μέγεθος και πως θα τοποθετηθούν οι adapters για να γίνει η ενίσχυση και η αλληλούχηση (Hrdlickova et al., 2017). Αυτά αποτελούν μερικά βασικά βήματα του RNA-Seq και παρουσιάζονται στην εικόνα 2.

Εικόνα figure 1 από (Kukurba & Montgomery, 2015)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 2: Σύνοψη των βασικών βημάτων για το RNA-Seq

Επιλογή του RNA

Η επιλογή του RNA αποτελεί βασικό βήμα για το σχεδιασμό του RNA-Seq. Συνήθως η μελέτη αφορά το mRNA των κυττάρων καθώς τα ερευνητικά ερωτήματα σχετίζονται κυρίως με την έκφραση νέων ισομορφών, την ταυτοποίηση νέων αλληλομορφικών παραλλαγών (allele variant identification) και τη μελέτη των επιπέδων των microRNAs (Conesa et al., 2016). Η επιλογή των mRNAs γίνεται είτε με την επιλογή των poly(A) ουρών είτε με τη μέθοδο της απομάκρυνσης του rRNA (RNA depletion). Η επιλογή αυτή συμβαίνει διότι μόνο το 1-2% του συνολικού RNA αποτελεί αντικείμενο μελέτης και πάνω από το 90% του συνολικού RNA ενός κυττάρου αποτελείται από rRNA που συνήθως δεν παρουσιάζει ερευνητικό ενδιαφέρον (Conesa et al., 2016). Για τους ευκαρυωτικούς οργανισμούς είναι δυνατές και οι δύο προσεγγίσεις, οπού κάθε μία έχει διαφορετικούς περιορισμούς.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Επιλογή RNA με τη χρήση των poly(A) ουρών

Η επιλογή των mRNA με τη χρήση των poly(A) ουρών είναι η πιο συχνή και μπορεί να γίνει και επιλογή lncRNAs που περιέχουν και αυτά poly(A) ουρές. Απαιτεί όμως σχετικά υψηλά επίπεδα mRNA που δεν έχει κοπεί και υπολογίζεται από τον αριθμό RIN (RNA integrity number) (Conesa et al., 2016; Hrdlickova et al., 2017). Για τη σωστή δημιουργία της βιβλιοθήκης ο αριθμός RIN, που υπολογίζεται με τη βοήθεια ενός ειδικού αλγορίθμου, είναι επιθυμητό να είναι μεγαλύτερος από 8 για το ολικό RNA (Chatterjee et al., 2018). Η επιλογή των RNAs με poly(A) γίνεται με τη βοήθεια ειδικών μαγνητών ή με cellulose beads που περιβάλλονται από oligo-dT μόρια (Hrdlickova et al., 2017). Η δεύτερη μέθοδος βασίζεται στη χρήση χρωματογραφίας συγγένειας όπου το RNA περνάει από μία στήλη χρωματογραφίας που περιέχει τα σφαιρίδια με τα oligo(dT). Οι poly(A) ουρές, μέσω υβριδοποίησης, ενώνονται με τα σφαιρίδια ενώ τα υπόλοιπα μόρια εξέρχονται από τη στήλη. Στο τέλος η σύνδεση των σφαιριδίων με τις poly(A) ουρές αποσταθεροποιείται και απομονώνεται τα RNAs που έχουν poly(A) ουρές [poly(A)+ RNAs] (*Oligo(DT) Cellulose Columns- Life Technologies*, n.d.). Μία άλλη μέθοδος απομόνωσης των poly(A)+ mRNAs είναι η χρήση oligo-dT priming για την αντίστροφη μεταγραφή, μία μέθοδος η οποία δημιουργεί αρκετά λάθη και δεν επιλέγεται συχνά (Hrdlickova et al., 2017).

Επιλογή RNA με τη μέθοδο του rRNA depletion

Το rRNA depletion (απομάκρυνση των rRNA) πραγματοποιείται όταν η ακεραιότητα του mRNA δεν είναι ικανοποιητική ή η ποσότητα του είναι μικρή ή όταν μελετώνται προκαρυωτικοί οργανισμοί που δεν περιέχουν poly(A)+ mRNAs (Conesa et al., 2016). Γενικά ο αριθμός των rRNAs είναι μεγάλος σε ένα κύτταρο όμως τα μόρια αυτά έχουν μικρή ερευνητική αξία και επομένως είναι αναγκαίο να απομακρυνθούν από το δείγμα. Για την απομάκρυνση αυτή έχουν δημιουργεί πολλές μέθοδοι με διαφορετικά μειονεκτήματα και πλεονεκτήματα.

Η χρήση ειδικών ως προς μια αλληλουχία ανιχνευτών που υβριδίζονται στο rRNA και έπειτα η απομάκρυνση των rRNAs είναι μία βασική μέθοδος. Μια εναλλακτική μέθοδος είναι η χρήση αντινοηματικών ολιγονουκλεοτιδικών DNA που ακολουθείται από την πέψη με RNase H (probe-directed degradation PDD). Μία άλλη μέθοδος, που απαιτεί υψηλό αριθμό ολικού RNA, περιλαμβάνει τη χρήση ολικού cDNA το οποίο γίνεται κυκλικό, υβριδοποίηση με ανιχνευτές rRNA και πέψη με ειδική νουκλεάση (duplex-specific nuclease, DSN), κάνοντας έτσι όλα τα rRNAs αδύνατο να ενισχυθούν

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

κατά το amplification. Εναλλακτικά, χρησιμοποιούνται not-so-random primers (NSR) που συνδέονται με επιθυμητά μόρια RNA στο βήμα της αντίστροφης μεταγραφάσης. Στη μέθοδο αυτή χρησιμοποιούνται εξαμερή ή εφταμερή που δεν υπάρχουν στο rRNA και έτσι γίνεται απομάκρυνση τους από τη μελέτη. Επίσης πολλές φορές ανάλογα με το πείραμα μπορεί να γίνει και απομάκρυνση των μετάγραφων που εμφανίζονται υψηλή έκφραση. Η μέθοδος αυτή παρότι έχει ως μειονέκτημα το off-target priming και εξαρτάται από το είδος του οργανισμού που μελετάται, εμφανίζει σημαντικό πλεονέκτημα έναντι άλλων μεθόδων διότι μπορεί να δουλέψει ικανοποιητικά με degraded RNAs και με δείγματα με μικρό input (low input samples). Μία άλλη μέθοδος βασίζεται στην αποδιάταξη μέσω θερμότητας, στο re-annealing και στην επιλεκτική αποδόμηση μέσω νουκλεάσης (duplex-specific nuclease, DSN). Η μέθοδος αυτή ονομάζεται C_0T-hybridization. Η αποικοδόμηση συμβαίνει πιο συχνά σε cDNAs που υπάρχουν συχνότητα στο δείγμα και σε αυτό το φαινόμενο βασίζεται η επιτυχία της. Η αποικοδόμηση των rRNAs και των tRNAs γίνεται επίσης και με ειδικό ένζυμο το οποίο αναγνωρίζει μόρια RNA που έχουν 5'- μονοφωσφορικά άκρα. Το ένζυμο αυτό ονομάζεται TEX και είναι μία 5' φωσφόρο-εξαρτώμενη εξωνουκλεάση (terminator 5′-phosphate-dependent exonuclease) (Hrdlickova et al., 2017).

Όλες οι παραπάνω μέθοδοι χρησιμοποιούνται ανάλογα με τον τύπο του ερευνητικού ερωτήματος και το είδος του δείγματος όμως οι τρεις μέθοδοι που παρουσιάζονται και στον πίνακα 2 χρησιμοποιούνται συχνότερα (Hrdlickova et al., 2017).

Πίνακας 2: Μέθοδοι για την επιλογή του επιθυμητού RNA

Ευκαρυωτικοί οργανισμοί	Προκαρυωτικοί οργανισμοί/ degraded RNA/noncoding RNA	
poly(A)+ RNAs selection	Depletion of rRNA	
Oligo-dT bead purification	PDD (probe-directed degradation)	NSR-priming (not-so-random priming)
Εύκολη χρήση	-	Degraded RNAs
Σχετικά χαμηλό κόστος	-	Αρκετά off-target

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

-	Απαραίτητος ο σχεδιασμός διαφορετικών ανιχνευτών για κάθε είδος	Εξαρτάται από το είδος του οργανισμού
Low-input samples	High-input samples/ ακατάλληλο για κλινικά δείγματα	Low-input samples

Κατακερματισμός RNA

Μετά την επιλογή των RNAs που πρόκειται να μελετηθούν είναι απαραίτητο να γίνει κατακερματισμός (fragmentation) του RNA σε μικρότερα κομμάτια για να είναι συμβατά με τα μηχανήματα αλληλούχησης που έχουν δημιουργηθεί. Ο κατακερματισμός είναι απαραίτητος μόνο για τα μεγάλα μόρια RNA και όχι για τα μικρότερα μόρια όπως τα miRNAs, τα piRNAs και τα siRNAs. Έτσι κατά το fragmentation τα μεγάλα μόρια RNA κατακερματίζονται σε τμήματα συνήθως 200-500 βάσεων ανάλογα με την πλατφόρμα αλληλούχησης που πρόκειται να χρησιμοποιηθεί (Wang et al., 2009). Παραδείγματος χάριν, μία από τις γνωστότερες πλατφόρμες αλληλούχησης, η Illumina, απαιτεί τμήματα RNA μικρότερα από 600 βάσεις (Hrdlickova et al., 2017). Ο κατακερματισμός μπορεί να γίνει πριν το στάδιο της αντίστροφης μεταγραφής, όπου κατακερματίζεται το RNA και ονομάζεται RNA fragmentation ή μετά το στάδιο της αντίστροφής μεταγραφής, όπου έχει δημιουργηθεί το cDNA και είναι αυτό που κατακερματίζεται και ονομάζεται DNA fragmentation (Wang et al., 2009).

RNA fragmentation (Κατακερματισμός του RNA)

Ο κατακερματισμός του RNA γίνεται με τη χρήση αλκαλικών διαλυμάτων, διαλυμάτων με δισθενή κατιόντα ή με τη χρήση ενζύμων. Τα δισθενή κατιόντα είναι συνήθως μαγνησίου ή ψευδαργύρου και χρησιμοποιούνται κυρίως σε συνδυασμό με υψηλές θερμοκρασίες (συνήθως 70^{o}C). Υψηλές θερμοκρασίες χρησιμοποιούνται και κατά το fragmentation με αλκαλικά διαλύματα διότι η υψηλή θερμοκρασία μειώνει το λάθος κατακερματισμό που μπορεί να προκληθεί λόγω της δομής του RNA, παρόλα αυτά ο κατακερματισμός δεν είναι ποτέ εντελώς τυχαίος. Ο κατακερματισμός με τη χρήση ενζύμου περιλαμβάνει το ένζυμο RNase III που και αυτή η μέθοδος προκαλεί λάθη. Παρόλα αυτά, ο άνισος κατακερματισμός του RNA αποτελεί πρόβλημα, προκαλώντας συγκεκριμένες περιοχές του RNA να αναπαρίστανται με διαφορετικό

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

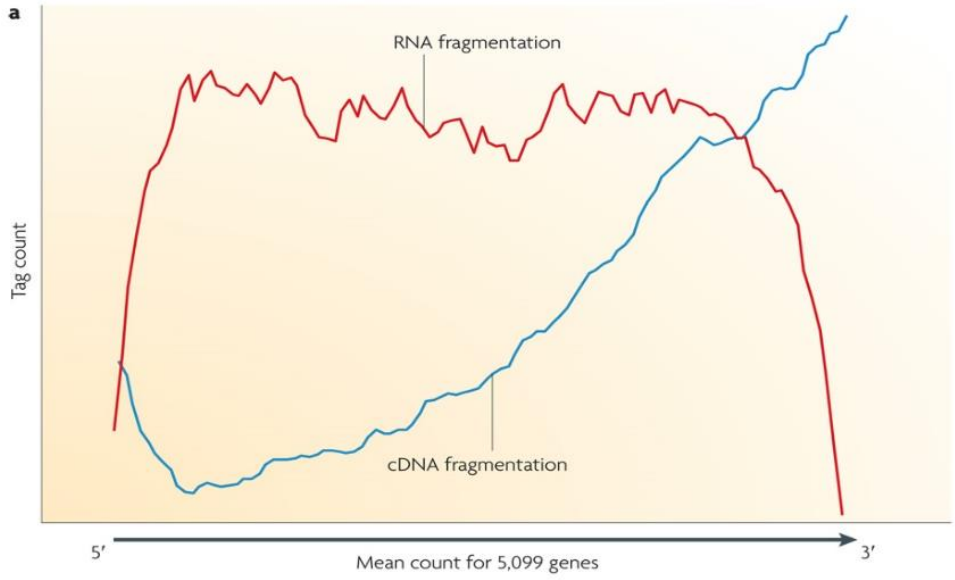
τρόπο (Hrdlickova et al., 2017). Γενικά φαίνεται από την εικόνα 3 ότι η μέθοδος αυτή κάνει λίγα συστηματικά λάθη (bias) στο transcript body και πολύ λίγα συστηματικά λάθη στα δύο άκρα του μετάγραφου (transcript) (Wang et al., 2009).

DNA fragmentation (Κατακερματισμός του DNA)

Το DNA fragmentation γίνεται με τη χρήση acoustic shearing, DNases ή τρανσποζονίων (Hrdlickova et al., 2017). Η μέθοδος του acoustic shearing αποτελείται από κύματα μικρού μήκους και υψηλής συχνότητας ενέργεια που προκαλούν θραύσματα DNA μήκους χιλιάδων ή εκατοντάδων βάσεων. Δεν χρειάζεται μεταβολή της θερμοκρασίας και είναι μία αυτοματοποιημένη μέθοδος που χρησιμοποιείται ευρύτατα (*DNA Fragmentation | NEB*, n.d.; Hrdlickova et al., 2017). Ο κατακερματισμός του cDNA με τη χρήση των τρανσποζονίων ονομάζεται tagmentation και χρησιμοποιεί την τρανσποζάση Tn5, που κάνει τον κατακερματισμό και ταυτόχρονα προσθέτει ολιγονουκλεοτιδικούς adapters στα δύο άκρα των θραύσεων (Hrdlickova et al., 2017). Σε αντίθεση με το RNA fragmentation, το cDNA fragmentation εμφανίζει συστηματικά λάθη στον προσδιορισμό των 3’ άκρων των μετάγραφων, παρέχοντας χρήσιμες πληροφορίες για την ταυτοποίηση αυτών των άκρων (Wang et al., 2009). Στο cDNA fragmentation η αναλογία του ενζύμου και του DNA που χρησιμοποιούνται πρέπει να είναι συγκεκριμένη και για αυτό το λόγο το RNA fragmentation είναι αυτό που προτιμάται (Hrdlickova et al., 2017). Όπως φαίνεται και στη εικόνα 3, η μπλε καμπύλη που αναπαριστά το cDNA fragmentation κάνει συστηματικά λάθη στο 3’ άκρο ενώ το RNA fragmentation παρέχει καλή κάλυψη στο μετάγραφο άλλα χωλαίνει στα 5’ και 3’ άκρα (Wang et al., 2009).

Εικόνα figure 3 από (Wang et al., 2009)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 3: cDNA fragmentation (μπλε καμπύλη) και RNA fragmentation (κόκκινη καμπύλη)

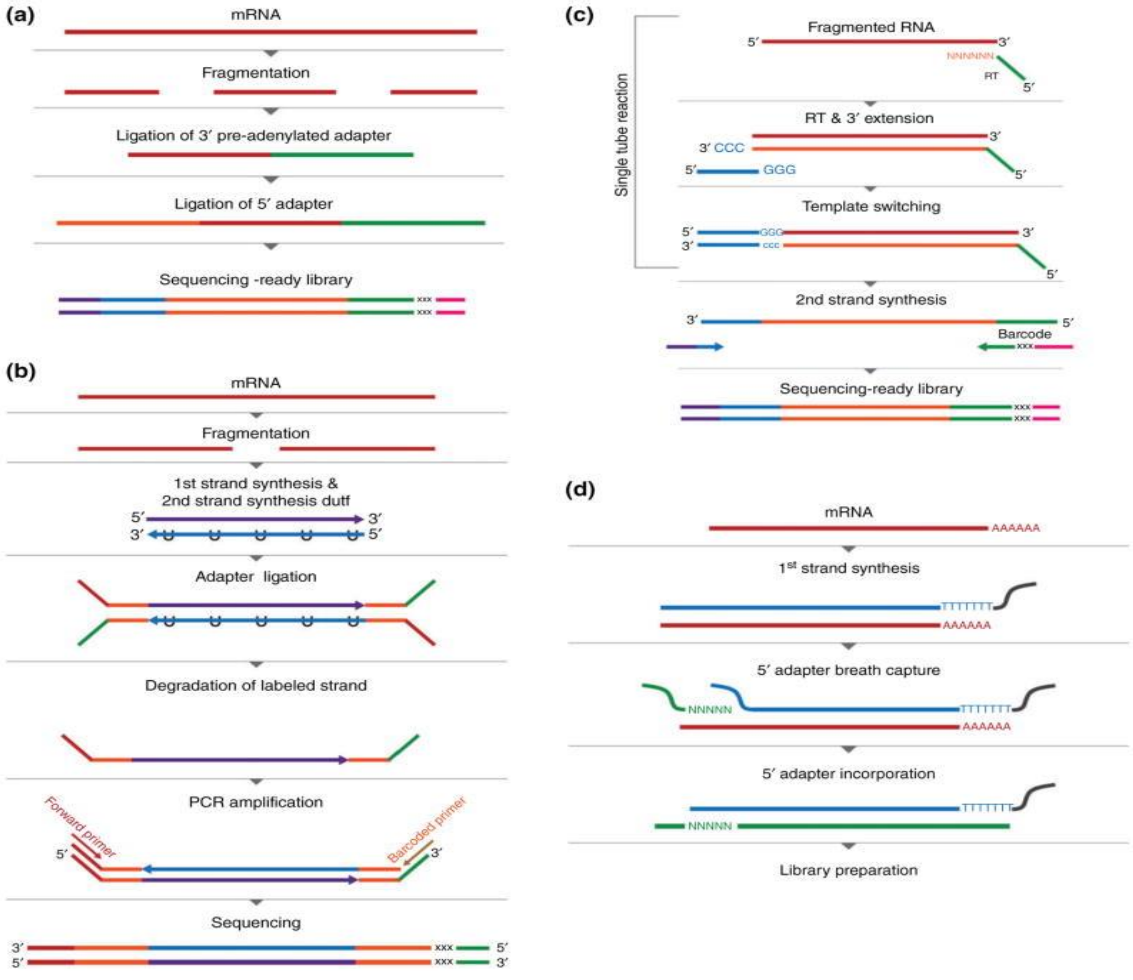
Προσθήκη adapters

Στα θραύσματα που έχουν δημιουργηθεί χρειάζεται να προστεθούν DNA adapters ώστε να γίνει η ενίσχυση και η αλληλούχηση τους. Η μέθοδος αυτή έχει σαν μειονέκτημα ότι χάνεται πληροφορία σχετικά με το ποια από τις δύο αλυσίδες του DNA αντιστοιχεί στην κωδική αλυσίδα του RNA, δηλαδή για το ποια αλυσίδα του DNA εκφράζεται πραγματικά. Το μειονέκτημα αυτό περιπλέκει αρκετά την ανάλυση των μετάγραφων. Πιο συγκεκριμένα, δεν αναγνωρίζονται νέα είδη RNA, antisense RNAs και RNAs που επικαλύπτονται, κάνοντας δύσκολη την ανάλυση και ποσοτικοποίηση των αποτελεσμάτων. Λύση σε αυτό το πρόβλημα δίνει η δημιουργία strand-preserving βιβλιοθήκης, όπου αποτυπώνεται το directionality του RNA (Conesa et al., 2016; Hrdlickova et al., 2017). Ένα από τα πρωτόκολλα που χρησιμοποιούνται περιλαμβάνει τη προσθήκη διαφορετικών adapters στα 5’ και 3’ άκρα του RNA και είναι σχεδιασμένο κυρίως για small RNA-Seq. Αρχικά γίνεται απομάκρυνση της 3’ φωσφορικής ομάδας και προσθήκη μίας 5’ φωσφορικής ομάδας. Στη συνέχεια μια κολοβωμένη (truncated) RNase II προσθέτει έναν 5′ adenylated 3′ adapter και RNA λιγάση I προσθέτει έναν 5’ adapter, όπως φαίνεται και στην εικόνα 4a (Hrdlickova et al., 2017). Μία άλλη μέθοδος περιλαμβάνει την προσθήκη ουρακίλης (dUTP) στη δεύτερη αλυσίδα του cDNA (εικόνα 4b). Με τον τρόπο αυτό, η δεύτερη αλυσίδα του cDNA δεν χρησιμοποιείται για ενίσχυση και αλληλούχηση καθώς δεν αποτελεί καλό υπόστρωμα για την πολυμεράση στο στάδιο της ενίσχυσης. Επιπρόσθετα,

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

χρησιμοποιείται το ένζυμο γλυκοσιλάση της ουρακίλης (uracil DNA glycosylase, UDG) που αναγνωρίζει και κόβει την ουρακίλη από το DNA, οπότε η δεύτερη αλυσίδα του DNA που περιέχει την ουρακίλη αποικοδομείται. Υπάρχουν όμως και μέθοδοι που η προσθήκη των adapters γίνεται στα θραύσματα RNA και όχι στα θραύσματα cDNA, όπως η μέθοδος Pedegrine (εικόνα 4c) και η BrAD-Seq (εικόνα 4d). Παρόλα αυτά η μέθοδος που περιλαμβάνει την προσθήκη ουρακίλης (dUTP) στη δεύτερη αλυσίδα του cDNA είναι αυτή χρησιμοποιείται στα περισσότερα πρωτόκολλα, όμως δεν είναι κατάλληλη για low-input δείγματα καθώς η χρήση του ενζύμου οδηγεί σε μείωση του υλικού που υπάρχει (Conesa et al., 2016; Hrdlickova et al., 2017).

Εικόνα figure 1 από (Hrdlickova et al., 2017)



Εικόνα 4: Μέθοδοι για strand-specific RNA-Seq

Η δημιουργία strand-specific βιβλιοθηκών εμφανίζει σαφές πλεονέκτημα στο annotation του μεταγραφώματος παρόλα αυτά δεν χρησιμοποιείται ιδιαίτερα διότι

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

χρειάζονται πολλά βήματα ή απευθείας σύνδεση RNA-RNA που συνηθώς δεν είναι αποτελεσματική (Wang et al., 2009).

Ενίσχυση των αλληλουχιών της βιβλιοθήκης με τη μέθοδο PCR και προσθήκη των labels

Η αλληλούχηση του cDNA πολλές φορές απαιτεί την ενίσχυση (amplification) των αλληλουχιών της βιβλιοθήκης με PCR (Polymerase Chain Reaction). Η PCR προκαλεί συστημικά λάθη λόγω του διαφορετικού μεγέθους των cDNAs και οδηγεί στη δημιουργία τμημάτων που ενισχύονται περισσότερο από άλλα. Για το λόγο αυτό έχουν αναπτυχθεί πολλές τεχνικές που διορθώνουν τα λάθη που προκαλούνται από την PCR. Μία από τις μεθόδους αυτές, λαμβάνοντας υπόψιν ότι έγινε τυχαίο RNA fragmentation, θεωρεί ότι τα προϊόντα της αλληλούχησης (final sequencing reads) που έχουν ίδια αρχή και τέλος προέρχονται από λάθη στη διαδικασία της PCR και τα συγχωνεύει. Άλλες μέθοδοι περιλαμβάνουν τη χρήση labels ώστε να ξεχωρίζουν τα διαφορετικά προϊόντα της PCR. Οι προσθήκες αυτών των περιοχών πραγματοποιούνται συνήθως κατά την προσθήκη των adapters ή του fragmentation ή της αντίστροφης μεταγραφής. Γενικά χρησιμοποιούνται labels που η αλληλουχία τους είναι γνωστή ή αλληλουχίες τυχαίων νουκλεοτιδίων. Αυτοί οι δύο διαφορετικοί τύποι διαφέρουν ως προς το μέγεθος και την πολυπλοκότητα τους. Ειδικότερα τα γνωστής αλληλουχίας labels κατανέμονται ομοιόμορφα στη βιβλιοθήκη αλλά κατασκευάζονται δυσκολότερα σε σύγκριση με τις αλληλουχίες τυχαίων νουκλεοτιδίων που εμφανίζουν υψηλή αστάθεια (Hrdlickova et al., 2017).

Επιλογή Single-end ή Paired-end sequencing

Το RNA-Seq γίνεται είτε με τη μέθοδο Single-end sequencing (SE) είτε με τη μέθοδο Paired-end sequencing (PE) και η επιλογή της μεθόδου που θα χρησιμοποιηθεί αποτελεί ένα από τα σημαντικότερα βήματα του σχεδιασμού του πειράματος. Το SE παρέχει τα reads από την αλληλούχηση του cDNA από το ένα άκρο κάθε θραύσματος cDNA (cDNA fragment). Το PE παρέχει τα reads της αλληλούχησης και από τα δύο άκρα κάθε θραύσματος cDNA, δημιουργώντας έτσι διπλάσιο αριθμό reads από το SE. Η επιλογή ανάμεσα στις δύο μεθόδους έγκειται στο ερευνητικό ερώτημα που τίθεται και στο κόστος της κάθε μεθόδου. Ειδικότερα, το SE αποτελεί μία οικονομικότερη μέθοδο και είναι κατάλληλη για μελέτες έκφρασης γονιδίων καλά μελετημένων οργανισμών (well-annotated organisms). Αντίθετα το PE είναι πιο ακριβό αλλά αυξάνει το mappability (την ικανότητα χαρτογράφησης στις επαναλαμβανόμενες

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

περιοχές του γονιδιώματος) και την ικανότητα ταυτοποίησης δομικών ανακατατάξεων, ενθέσεων, διπλασιασμών και αναστροφών. Ακόμα το PE χρησιμοποιείται για την ανάλυση της έκφρασης διαφορετικών ισομορφών μιας πρωτεΐνης, την ανακάλυψη νέων μετάγραφων (de novo transcripts) και γενικά τη μελέτη φτωχά χαρακτηρισμένων μεταγραφωμάτων (poorly characterize annotated transcriptomes) (Chatterjee et al., 2018; Conesa et al., 2016).

Προσδιορισμός sequencing depth (βάθους αλληλούχησης) και library size (μεγέθους της βιβλιοθήκης)

Ο σχεδιασμός του πειράματος περιλαμβάνει τον προσδιορισμό του βάθους της αλληλούχησης (sequencing depth) και του μεγέθους της βιβλιοθήκης (library size). Το sequencing depth ορίζεται ως ο αριθμός των reads ανά δείγμα συνεπώς το ερώτημα που προκύπτει είναι το πόσα reads χρειάζονται ώστε να προσδιοριστεί επαρκώς το μεταγράφωμα που μελετάται. Όσο το βάθος της αλληλούχησης μεγαλώνει τόσα περισσότερα μετάγραφα θα ανιχνευθούν και θα προσδιοριστούν ακριβέστερα σε ένα δείγμα. Γενικά ο αριθμός των reads που χρειάζονται διαφέρει ανάλογα με το πείραμα. Ειδικότερα για τη μελέτη γονιδίων ευκαρυωτικών οργανισμών που έχουν μέτρια έως υψηλή έκφραση χρειάζονται πέντε εκατομμύρια reads ενώ για χαμηλής έκφρασης γονίδια χρειάζονται έως και δέκα εκατομμύρια reads. Για την αλληλούχηση ενός single cell το εύρος των reads κυμαίνεται από είκοσι χιλιάδες έως ένα εκατομμύριο reads. Όσον αφορά το library size, αυτό εξαρτάται από την πολυπλοκότητα του μεταγραφώματος του οργανισμού που μελετάται. Τέλος είναι απαραίτητος ο προσδιορισμός των αντιγράφων που πρέπει να αλληλουχιθούν. Για το RNA-Seq δεν υπάρχει κάποιος κανόνας για το πόσο replicates πρέπει να χρησιμοποιηθούν, παρόλα αυτά τα περισσότερα πειράματα έχουν τουλάχιστον τρία replicates. Γενικά για να ανακαλυφθεί η διαφορετική έκφραση των γονιδίων έχει φανεί ότι η αύξηση του sequencing depth πάνω από δέκα εκατομμύρια reads δεν βελτιώνει τα αποτελέσματα. Αντίθετα όμως η προθήκη παραπάνω replicates φαίνεται να τα βελτιώνει. Έχει φανεί έτσι ότι η προσθήκη παραπάνω replicates είναι η καλύτερη λύση για την ανακάλυψη της διαφορετικής έκφρασης των γονιδίων σε σχέση με την αύξηση του sequencing depth όταν υπάρχουν παραπάνω από είκοσι εκατομμύρια reads ανά δείγμα (Chatterjee et al., 2018; Conesa et al., 2016).

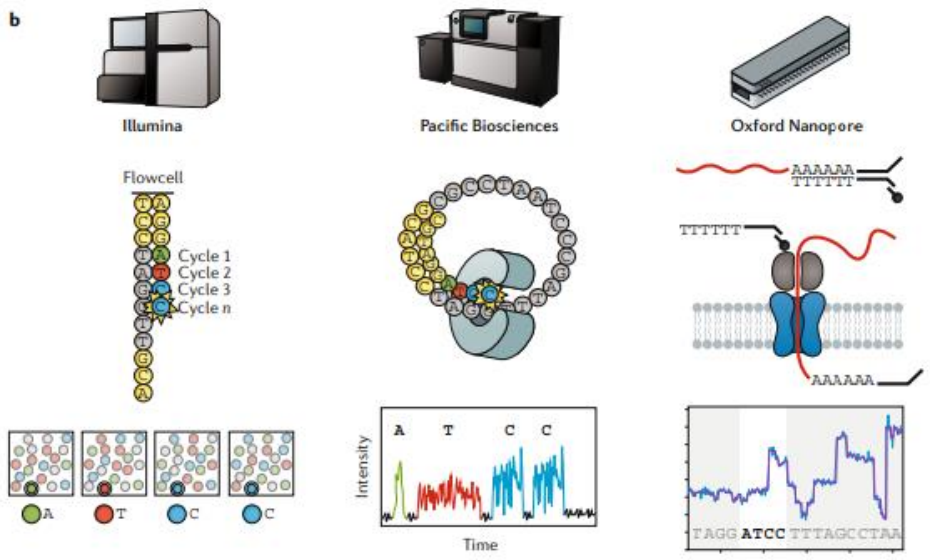
Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Επιλογή κατάλληλης πλατφόρμας αλληλούχησης

Η επιλογή της πλατφόρμας που γίνεται η αλληλούχηση αποτελεί βασικό κομμάτι του πειράματος. Στην αγορά υπάρχουν διαθέσιμες πολλές πλατφόρμες οι περισσότερες από τις οποίες χρησιμοποιούν τη μέθοδο της αλληλούχησης με σύνθεση (sequencing-by-synthesis, SBS). Η εταιρία που έχει κυριαρχήσει το τομέα αυτό είναι η Illumina με τις πλατφόρμες HiSeq και MiSeq. Βασιζόμενο στο sequencing-by-synthesis χρησιμοποιεί reversible-terminator νουκλεοτίδια που είναι σημασμένα με φθορίζουσα χρωστική ενώ το DNA είναι ακινητοποιημένο σε γυάλινη επιφάνεια για να γίνει η αλληλούχηση (Kukurba & Montgomery, 2015). Το συμπληρωματικό προς την αλληλουχία reversible-terminator νουκλεοτίδιο προσδένεται στο DNA και απελευθερώνεται η φθορίζουσα χρωστική, η οποία είναι διαφορετική ανάλογα με τη βάση που προστίθεται. Ο ρόλος του terminator είναι να εμποδίζει άλλα νουκλεοτίδια να προσδεθούν στην επόμενη βάση του DNA και απομακρύνεται όταν έχει απελευθερωθεί η φθορίζουσα χρωστική ώστε να «διαβαστεί» η επόμενη βάση (*NGS Workflow Steps | Illumina Sequencing Workflow*, n.d.). Το HiSeq της Illumina είναι η συχνότερα χρησιμοποιούμενη πλατφόρμα για την αλληλούχηση του RNA καθώς εμφανίζει μικρά ποσοστά λάθους (Kukurba & Montgomery, 2015). Εναλλακτικά χρησιμοποιούνται πλατφόρμες όμως η PacBio που βασίζεται στο single-molecule real-time sequencing και το Oxford Nanopore, οι οποίες όμως δεν χρησιμοποιούνται τόσο συχνά όσο το HiSeq της Illumina και παρουσιάζονται στην εικόνα 5 (Kukurba & Montgomery, 2015; Stark et al., 2019).

Εικόνα Figure1b (Stark et al., 2019)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 5: Οι τρεις βασικές τεχνολογίες αλληλούχησης που χρησιμοποιούνται για το RNA-Seq

Κατά την αλληλούχηση, η βιβλιοθήκη που έχει δημιουργηθεί φορτώνεται στο flow cell και τοποθετείται στον sequencer. Το HiSeq της Illumina αποτελείται από δύο flow cells με οχτώ ξεχωριστά lanes. Σε αυτά τα cDNA fragments περνούν μια διαδικασία ενίσχυσης που ονομάζεται cluster generation και έτσι δημιουργούνται εκατομμύρια αντίγραφα DNA με μία αλυσίδα τα οποία στη συνέχεια αλληλουχούνται με τη μέθοδο SBS που αναφέρθηκε παραπάνω (Kukurba & Montgomery, 2015; *NGS Workflow Steps | Illumina Sequencing Workflow*, n.d.).

Single Cell RNA-Sequencing

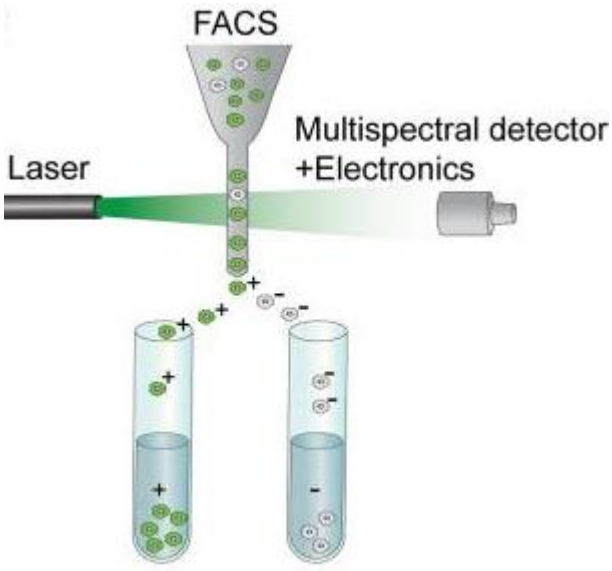
Ένα είδος του RNA-Seq αποτελεί το single-cell RNA sequencing (scRNA-Seq) το οποίο δίνει στους επιστήμονες τη δυνατότητα ανακάλυψης νέων ευρημάτων σχετικών με σπάνιους πληθυσμούς κυττάρων και ρυθμιστικών σχέσεων μεταξύ των γονιδίων (Hwang et al., 2018). Με το κλασσικό RNA-Seq λαμβάνονται πληροφορίες για την έκφραση των γονιδίων κατά προσέγγιση και όχι για την έκφραση των γονιδίων ενός συγκεκριμένου κυττάρου (Hwang et al., 2018). Αυτό αποτελεί πρόβλημα καθώς λαμβάνουμε σαν δεδομένο ότι όλα τα κύτταρα του ίδιου τύπου εκφράζουν τα ίδια γονίδια κάτι που δεν ισχύει από τα δεδομένα που υπάρχουν (Hwang et al., 2018). Συνεπώς οι περισσότερες μελέτες που παίρνουν σαν δεδομένο τη συγκεκριμένη παραδοχή στερούνται τις πληροφορίες σχετικά με την ποικιλομορφία που υπάρχει μεταξύ των κυττάρων του ίδιου τύπου (Hwang et al., 2018). Η πρώτη περιγραφή της

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

μεθόδου που χρησιμοποιήθηκε και κάνει ανάλυση του μεταγραφώματος ενός κυττάρου με τη βοήθεια του next generation sequencing πραγματοποιήθηκε το 2009, οπού άνοιξε το δρόμο για το single cell RNA-seq (Hwang et al., 2018). Μέσω των μελετών αυτών είναι δυνατόν να δοθούν απαντήσεις σχετικά με την αντοχή στα καρκινικά φάρμακα ή την επανεμφάνιση του καρκίνου (Hwang et al., 2018). Αντικείμενο έρευνας με τη μέθοδο αυτή αποτελούν επίσης οι σχέσεις των κυττάρων στα πρώτα στάδια της ανάπτυξης και ο καθορισμός της τύχης των λεμφοκυττάρων (Hwang et al., 2018).

Απομόνωση κυττάρου για scRNA-Seq

Το πρώτο βήμα για το single cell RNA-Seq είναι η απομόνωση του κυττάρου το οποίο πρόκειται να αλληλουχιθεί (Hwang et al., 2018). Για να γίνει αυτό υπάρχουν πολλές μέθοδοι με πιο διαδεδομένη αυτή των flow-activated cell sorting (FACS) (Hwang et al., 2018). Στη μέθοδο αυτή αρχικά χρησιμοποιούνται φθορίζοντα μονοκλωνικά αντισώματα τα οποία αναγνωρίζουν συγκεκριμένους επιφανειακούς δείκτες (markers) των κυττάρων που πρόκειται να απομονωθούν, όπως φαίνεται και στην εικόνα 6 (Hwang et al., 2018).



Εικόνα 6: Τα FACS απομονώνουν single cells με τη χρήση φθοριζόντων δεικτών

Στη συνέχεια ακολουθούν τα βασικά στάδια για τη δημιουργία της βιβλιοθήκης (Hwang et al., 2018). Πιο συγκεκριμένα, πραγματοποιείται λύση των κυττάρων, δημιουργία του cDNA με αντίστροφη μεταγραφή και ενίσχυση των τμημάτων που απομονώθηκαν (Hwang et al., 2018). Η αλληλούχηση γίνεται και εδώ με διάφορες

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

πλατφόρμες με πιο διαδεδομένες αυτές της Illumina (HiSeq, MiSeq) που αναφέρθηκαν παραπάνω (Hwang et al., 2018).

Ανάλυση των δεδομένων από το RNA-Seq

FASTQ files

Μετά την αλληλούχηση δημιουργείται ένα αρχείο που περιέχει όλες τις πληροφορίες από την πλατφόρμα του NGS που χρησιμοποιήθηκε. Η Illumina, χρησιμοποιώντας τη μέθοδο sequencing-by-synthesis και το λογισμικό Real Time Analysis (RTA), δημιουργεί και αποθηκεύει τα base calls για κάθε κύκλο αλληλούχησης σε αρχεία της μορφής individual base call, bcl. Τα αρχεία αυτά μετατρέπονται σε δεδομένα που περιέχουν τα δεδομένα της αλληλούχησης, αφού αυτή ολοκληρωθεί. Η διαδικασία ονομάζεται BCL to FASTQ conversion και μετατρέπει τα αρχεία BCL σε FASTQ. Τα FASTQ αρχεία περιέχουν τα δεδομένα από την αλληλούχηση των clusters που περνούν μέσα από ένα flow cell. Ανάλογα με το αν η αλληλούχηση ήταν single-end ή paired-end δημιουργούνται διαφορετικά αρχεία. Πιο συγκεκριμένα, για ένα single-end πείραμα δημιουργείται ένα αρχείο Read 1 (R1) FASTQ αρχείο για κάθε δείγμα που υπάρχει σε κάθε flow cell lane. Για paired-end αλληλούχηση δημιουργούνται δυο αρχεία για κάθε δείγμα σε κάθε lane, ένα R1 FASTQ αρχείο και ένα R2 FASTQ αρχείο (*FASTQ Files Explained*, n.d.). Τα αρχεία αυτά έχουν επέκταση *.fastq.gz. και περιέχουν πληροφορίες σε μορφή κειμένου (text form) για την αλληλουχία που δημιουργήθηκε και πληροφορίες για nucleotide base calls και τα quality-scores ανά βάση (*FASTQ Files Explained*, n.d.; *File Format Guide*, n.d.). Κάθε FASTQ αρχείο αποτελείται από τέσσερις σειρές, όπως φαίνεται στην εικόνα 7 (*File Format Guide*, n.d.).

```
@<identifier and expected information>
<sequence>
+<identifier and other information OR empty string>
<quality>
```

Εικόνα 7: Μορφή αρχείου FASTQ

Η πρώτη σειρά αποτελείται από το identifier της αλληλουχίας (sequence identifier) και εξαρτάται από το λογισμικό που κάνει τη μετατροπή των αρχείων από BCL σε FASTQ (*FASTQ Files Explained*, n.d.). Η δεύτερη σειρά περιέχει τις βάσεις, δηλαδή την

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

ακολουθία που προκύπτει από την αλληλούχηση και αποτελείται από τις κλασσικές βάσεις A,T,C,G,a,t,c,g ή N,n για βάσεις που δεν αναγνωρίζονται (unknown bases). Η σειρά «Quality score identifier line» είναι τρίτη και αποτελείται από το σύμβολο + που ακολουθείται συνήθως από κενό. Τέλος η τέταρτη σειρά περιέχει τα quality scores για κάθε base call (*FASTQ Files*, n.d.-b; *FASTQ Files Explained*, n.d.; *File Format Guide*, n.d.). Τα quality scores αναπαρίστανται στο αρχείο με διάφορους τρόπους όπως φαίνεται στην εικόνα 8 (*File Format Guide*, n.d.).

Εικόνα από

https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/#fastq-files .

Decimal-encoding, space-delimited	`[0-9]+	<quality>\s[0-9]+`
Phred-33 ASCII	`[\!\"\#\$\%\&\'\(\)\*\+,\-\.\/0-9:;<=>\?\@A-I]+`
Phred-64 ASCII	`[\@A-Z\[\\\]\^_`a-h]+`

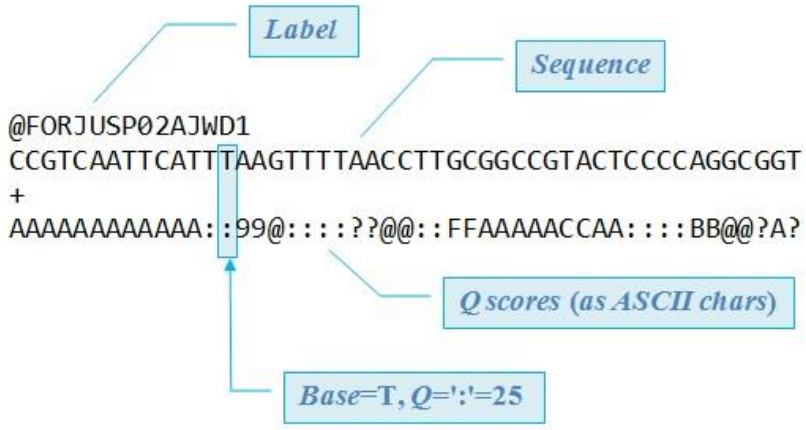
Εικόνα 8: Οι διαφορετικοί τρόποι που αναπαρίστανται τα quality scores στα αρχεία FASTQ

Q score

Το quality score ή Q score ονομάζεται αλλιώς και Phred και η τιμή του αναπαριστά την πιθανότητα η βάση που του αντιστοιχεί να οφείλεται σε λάθος. Στα FASTQ αρχεία συνήθως τα scores αναπαρίστανται με ASCII χαρακτήρες και συνηθέστερα με τους χαρακτήρες Phred-33 ASCII (*Quality (Phred) Scores*, n.d.).

Δηλαδή τα αρχεία FASTQ έχουν τη μορφή που φαίνεται στην εικόνα 9 (*FASTQ Files*, n.d.-a).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 9: Μορφή αρχείου FASTQ

Γενικά πρέπει το μήκος της αλληλουχίας που υπάρχει στη δεύτερη σειρά να συμβαδίζει με το μήκος της σειράς με τα quality scores (*File Format Guide*, n.d.). Παρόλα αυτά ένα αρχείο FASTQ έχει μεγάλο μέγεθος και μπορεί να ανοιχτεί είτε με text editors που χειρίζονται μεγάλου μεγέθους αρχεία είτε μέσω του command line των Unix και Linux (*FASTQ Files Explained*, n.d.).

Ποιοτικός έλεγχος των raw data

Τα αρχεία που δημιουργούνται και περιέχουν τα raw reads προτείνεται να περάσουν μία διαδικασία ελέγχου προτού γίνει η ανάλυση τους. Αυτό γίνεται για να προσδιοριστεί η ποιότητα τους και να γίνουν αλλαγές όπου χρειάζονται, μειώνοντας με τον τρόπο αυτό τα συστημικά λάθη του τυχόν προκύπτουν (*FastQC*, n.d.-a). Πιο συγκεκριμένα, γίνεται έλεγχος της αλληλούχησης, της παρουσίας adapters, του ποσοστού των βάσεων CG, της ύπαρξης επιμόλυνσης και άλλων παραγόντων (Conesa et al., 2016). Υπάρχουν δύο βασικά εργαλεία για τον έλεγχο αυτών το δεδομένων. Το πρώτο εργαλείο ονομάζεται NGSQC και χρησιμοποιείται σε όλες τις πλατφόρμες αλληλούχησης που υπάρχουν, ενώ το εργαλείο FASTQC χρησιμοποιείται για την ανάλυση δεδομένων από τις πλατφόρμες της Illumina (Conesa et al., 2016). Το NGSQC δίνει τη δυνατότητα επιβεβαίωσης των βιολογικών ανακαλύψεων που προκύπτουν, στηρίζοντας τα δεδομένα αυτά, σε πραγματικά βιολογικά γεγονότα και όχι σε λάθη κατά την αλληλούχηση (Dai et al., 2010).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Ποιοτικός έλεγχος των αρχείων FASTQ με το πρόγραμμα FASTQC

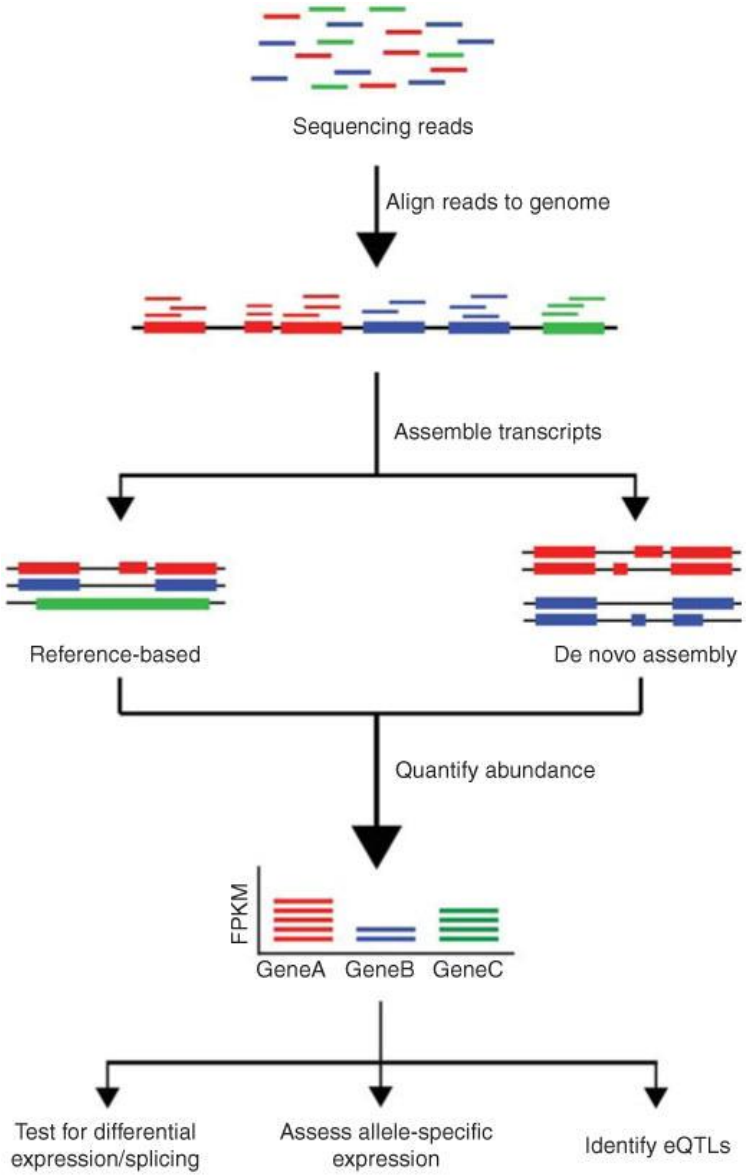
Το FastQC είναι μία εφαρμογή JAVA που λαμβάνει αρχεία της μορφής BAM, SAM και FASTQ ως input και το output είναι ένας φάκελος HTML-based permanent report folder (Chatterjee et al., 2018). Αποτελεί χρήσιμο εργαλείο για την ανάλυση των raw sequencing data καθώς χρησιμοποιείται είτε μέσω του command line είτε μέσω του Graphical User Interface και προσφέρει μία εκτίμηση για τυχόν προβλήματα που υπάρχουν στα δεδομένα (Chatterjee et al., 2018; *FastQC-A Quality Control Tool for High Throughput Sequence Data*, n.d.; *FastQC*, n.d.-b). Το εργαλείο αυτό παρέχει πληροφορίες σχετικά με το αν δημιουργήθηκαν duplicates κατά την PCR, το ποσοστό των GC βάσεων, την ποιότητα των reads και την ύπαρξη rRNA ή tRNA (Chatterjee et al., 2018). Ακόμα δημιουργεί γραφήματα και πίνακες, διευκολύνοντας το χρήστη στον έλεγχο των δεδομένων (*FastQC-A Quality Control Tool for High Throughput Sequence Data*, n.d.).

Επιλογή reads καλής ποιότητας

Αν τα δεδομένα που μελετώνται δεν είναι τα επιθυμητά υπάρχουν εργαλεία βιοπληροφορικής που απομακρύνουν τα reads που έχουν χαμηλή ποιότητα, μειώνουν τις βάσεις που δεν εμφανίζουν καλή ποιότητα και βελτιώνουν τους adapters (Conesa et al., 2016). Πιο συγκεκριμένα, το πρόγραμμα Cleanadaptors βασίζεται στη γλώσσα προγραμματισμού C και αφαιρεί τις αλληλουχίες των adapters ενώ το πρόγραμμα Trimmomatic είναι μία εφαρμογή JAVA που κόβει διάφορες αλληλουχίες από τα δεδομένα αλληλούχησης (Chatterjee et al., 2018). Το FASTX-toolkit αποτελεί ένα ακόμα εργαλείο επεξεργασίας των raw data που προκύπτουν σε αρχεία της μορφής FASTQ ή FASTA και βασίζεται στο κέλυφος του UNIX (Chatterjee et al., 2018). Αφού γίνει η βελτιστοποίηση των δεδομένων, πραγματοποιείται ξανά η ανάλυση FASTQC και μετά τη λήψη των επιθυμητών αποτελεσμάτων ακολουθεί η στοίχιση (alignment) των reads (Chatterjee et al., 2018). Μία αρχική σύνοψη των βημάτων που ακολουθούνται για την ανάλυση με τη μέθοδο του RNA-Seq παρουσιάζεται παρακάτω, στην εικόνα 10.

Εικόνα figure 2 από (Kukurba & Montgomery, 2015)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 10: Σύνοψη της ανάλυσης των δεδομένων από το RNA-Seq

Alignment (Στοίχιση αλληλουχιών)

Το alignment των reads γίνεται με τέσσερις διαφορετικούς τρόπους ανάλογα με το σκοπό του πειράματος που πραγματοποιείται. Οι τρόποι αυτοί είναι η στοίχιση που βασίζεται στο γένωμα (Genome based alignment), η στοίχιση που βασίζεται στο μεταγράφωμα (Transcriptome based alignment), η στοίχιση που δεν βασίζεται πουθενά (Reference free, de novo alignment) και η ποσοτικοποίηση των διάφορων ισομορφών που δεν βασίζονται σε στοίχιση (Alignment free isoform quantification) (Chatterjee et al., 2018).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Genome based alignment

Το Genome based alignment βασίζεται στην αλληλουχία ολόκληρου του γονιδιώματος αναφοράς (whole reference genome sequence) και λαμβάνει υπόψιν τη διαδικασία του ματίσματος, για το alignment (Chatterjee et al., 2018). Αυτό αποτελεί βασικό πλεονέκτημα των εργαλείων που το πραγματοποιούν καθώς στους ευκαρυωτικούς οργανισμούς πραγματοποιείται η διαδικασία του ματίσματος όπου δύο εξώνια ενώνονται και συνεπώς κατά το alignment αυτά θα φαίνονταν σαν δύο διαφορετικά θραύσματα. Όταν όμως το εργαλείο το αναγνωρίζει, του παρέχει ένα σημαντικό πλεονέκτημα έναντι άλλων, κάνοντας τη μελέτη πιο αξιόπιστη και τα εργαλεία αυτά πιο δημοφιλή (Kukurba & Montgomery, 2015). Μερικά από τα εργαλεία αυτά είναι το TopHat2, το STAR και το HISAT2 με τα δύο πρώτα να χρησιμοποιούνται συχνότερα (Chatterjee et al., 2018). Υπάρχουν ακόμα τα εργαλεία MapSplice και RUM, όμως καθένα από αυτά έχει δικά του μειονεκτήματα και πλεονεκτήματα σχετικά με την απόδοση και την ταχύτητα και η επιλογή του κατάλληλου εργαλείου εξαρτάται από τη φύση της μελέτης που γίνεται (Kukurba & Montgomery, 2015). Κατά το alignment ένα read ταιριάζει είτε με μία μοναδική περιοχή του γονιδιώματος αναφοράς είτε με ταιριάζει με πολλές διαφορετικές περιοχές του γονιδιώματος (multireads). Τα multireads εμφανίζονται λόγω επαναλαμβανόμενων αλληλουχιών του γονιδιώματος είτε λόγω κοινών περιοχών μεταξύ των παράλογων γονιδίων (paralogous genes) (Conesa et al., 2016). Δημιουργούνται έτσι αρχεία της μορφής GTF(General Transfer Format) ή GFF (Generic Feature Format) για να γίνει στη συνέχεια η χαρτογράφηση (mapping) και η συναρμολόγηση (assembly) των μετάγραφων (Chatterjee et al., 2018). Γενικά η μέθοδος αυτή χρησιμοποιείται όταν υπάρχει το γονιδίωμα αναφοράς του οργανισμού που μελετάται και ανακαλύπτει νέες ισομορφές μετάγραφων που προκύπτουν από το εναλλακτικό μάτισμα (Chatterjee et al., 2018).

Transcriptome based alignment

Το transcriptome based alignment χρησιμοποιεί τις αλληλουχίες των μετάγραφων σαν αλληλουχία αναφοράς αντί για ολόκληρο το γονιδίωμα (Chatterjee et al., 2018). Η μέθοδος αυτή δεν αναγνωρίζει νέες ισομορφές μετάγραφων που έχουν προκύψει από εναλλακτικό μάτισμα, όμως ο χρόνος που χρειάζεται η μέθοδος αυτή είναι μικρότερος σε σχέση με το genome based alignment, χωρίς όμως να αποτελεί σημαντικό πλεονέκτημα της μεθόδου αυτής (Chatterjee et al., 2018). Ακόμα με τη μέθοδο αυτή τα multireads είναι περισσότερα λόγω των διαφόρων ισομορφών που χαρτογραφούνται

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

στην ίδια περιοχή του μεταγραφώματος όταν μοιράζονται τα ίδια εξώνια (Conesa et al., 2016). Υπάρχουν αρκετά εργαλεία που χρησιμοποιούν αυτή τη μέθοδο όπως το Bowtie που βασίζεται στη γλώσσα προγραμματισμού C και είναι ένα γρήγορο και ευρέος χρησιμοποιούμενο εργαλείο, το Burrow-Wheeler Aligner (BWA) που χρησιμοποιείται κυρίως για μικρές αλληλουχίες, Mapping and Assembly with Quality (MAQ) και το GSNAP (Genomic Short-read Nucleotide Alignment Program) που βασίζεται στις γλώσσες προγραμματισμού C και Perl χρησιμοποιείται για τη γρήγορη ανίχνευση σύνθετων αλλαγών και εναλλακτικού ματίσματος σε μικρού μήκους reads (Chatterjee et al., 2018).

Reference free ή de novo alignment

Η μέθοδος Reference free ή de novo alignment χρησιμοποιείται όταν η αλληλουχία του γονιδιώματος ή του μεταγραφώματος του οργανισμού που μελετάται δεν έχει ανακαλυφθεί (Chatterjee et al., 2018). Χρησιμοποιείται για τον ποιοτικό έλεγχο της έκφρασης των γονιδίων που μελετώνται και πολλές φορές βοηθάει στη μελέτη της έκφρασης των γονιδίων σε καρκινικά κύτταρα (Chatterjee et al., 2018; Stark et al., 2019). Για τη μέθοδο αυτή προτιμώνται PE strand-specific αλληλουχίες μεγάλου μήκους καθώς οδηγούν σε πιο σωστή στοίχιση των de novo αλληλουχιών (Chatterjee et al., 2018; Conesa et al., 2016). Αυτό συμβαίνει διότι το πρώτο βήμα της ανάλυσης αυτής είναι η ένωση των επικαλυπτόμενων περιοχών των reads, δημιουργώντας τα contigs (Chatterjee et al., 2018). Αφού οι επικαλυπτόμενες αλληλουχίες (contigs) ανακαλυφθούν, τα reads χαρτογραφούνται στο νέο μεταγράφωμα που δημιουργήθηκε λόγω των contigs και γίνεται ποσοτικοποίηση της έκφρασης τους (Chatterjee et al., 2018). Το πιο διαδεδομένα εργαλείο για την ανάλυση αυτή είναι το Trinity που βασίζεται στη γλώσσα προγραμματισμού C, όμως υπάρχουν και άλλα εργαλεία όπως το Trans-ABySS, το Oases και το SOAPdenovo-Trans (Chatterjee et al., 2018).

Alignment free isoform qualification

Η μέθοδος Alignment free isoform qualification είναι μία νέα τεχνολογία για τη μελέτη μεγάλου μήκους reads η οποία συσχετίζει τα reads με τα μετάγραφα παραλείποντας το βήματα της ποσοτικοποίησης (Chatterjee et al., 2018; Stark et al., 2019). Η μέθοδος αυτή πραγματοποιείται με τα εργαλεία Sailfish που βασίζεται στη γλώσσα προγραμματισμού C, Kallisto και Salmon (Chatterjee et al., 2018; Stark et al., 2019).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Ποσοτικοποίηση των reads

Στη συνέχεια συνήθως ακολουθεί η εκτίμηση των επιπέδων της έκφρασης των γονιδίων. Γίνεται έτσι ποσοτικοποίηση της παρουσίας των διαφόρων reads με τη χρήση διαφόρων προγραμμάτων μερικά από τα οποία είναι τα RSEM, CuffLinks, MMSeq, HTSeq, featureCounts, MISO και FluxCapacitor (Conesa et al., 2016; Kukurba & Montgomery, 2015; Stark et al., 2019). Τα CuffLinks, MISO και FluxCapacitor υπολογίζουν τον αριθμό των reads που χαρτογραφούνται σε όλο το μήκος των μετάγραφων, ποσοτικοποιώντας έτσι την έκφραση τους (Conesa et al., 2016). Αντίθετα, το HTSeq μετράει των αριθμό των reads που χαρτογραφούνται σε ένα εξώνιο (Conesa et al., 2016). Το HTSeq και το featureCounts δεν συμπεριλαμβάνουν στις μετρήσεις πολλά reads όπως αυτά που είναι multireads και αυτά που εμφανίζουν overlap multiple expression features, εξαλείφοντας έτσι μετάγραφα με επικαλυπτόμενες ή ομόλογες περιοχές (Stark et al., 2019). Δημιουργούνται λοιπόν αρχεία της μορφής expression matrix όπου οι γραμμές αναπαριστούν τα μετάγραφα που ανιχνεύονται και οι στήλες κάθε δείγμα, όπως φαίνεται στον πίνακα 3 (Stark et al., 2019). Οι τιμές των αρχείων αυτών είναι είτε ο πραγματικός αριθμός των reads είτε η αναμενόμενη παρουσία τους (estimated abundances) (Stark et al., 2019).

Πίνακας 3: Expression matrix format

	Sample 1	Sample 2	Sample 3	Sample ….
Gene 1	10	20	15	…
Gene 2	20	40	29	…
Gene 3	5	35	60	…
Gene 4	15	30	15	…
Gene….	…	…	…	…

Κανονικοποίηση

Ακολουθεί η κανονικοποίηση του αριθμού των reads (Kukurba & Montgomery, 2015). Αυτό συμβαίνει διότι υπάρχουν αλληλουχίες με διαφορετικό μήκος, λάθη κατά την αλληλούχηση και διαφορετικό συνολικό αριθμό reads σε κάθε πείραμα κάνοντας αδύνατη τη σύγκριση των αποτελεσμάτων ειδικά όταν μελετώνται οι διαφορές στην έκφραση γονιδίων μεταξύ διαφορετικών δειγμάτων (Conesa et al., 2016). Η μέτρηση PRKM (reads per kilobase per million reads) είναι μία μέθοδος κανονικοποίησης του

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

αριθμού των reads λαμβάνοντας υπόψιν το μήκος του γονιδίου και το συνολικό αριθμό των reads (Kukurba & Montgomery, 2015). Ο αριθμός αυτός υπολογίζεται από τον τύπο RPKM= (Number of mapped reads for a transcript (or a feature) * 1000 bases * 10^6)/(length of the transcript * total mapped reads for the sample) (Chatterjee et al., 2018). Η μέθοδος αυτή οδηγεί στη σύγκριση των μεταγραφικών επιπέδων εντός και μεταξύ των δειγμάτων (Chatterjee et al., 2018). Ο αριθμός FPKM (fragment per kilobase per million reads) είναι όπως ο αριθμός RPKM με τη διαφορά ότι αντί για reads μετριέται ο αριθμός των cDNA fragments. Η κανονικοποίηση με αυτή τη μέθοδο προτιμάται όταν γίνεται Pair-End sequencing (PE) ενώ για το Single-End sequencing (SE) χρησιμοποιούνται και οι δύο μετρήσεις χωρίς διαφορά (Chatterjee et al., 2018; Conesa et al., 2016). Ο αριθμός RPKM και FPKM μετατρέπεται εύκολα σε έναν άλλο αριθμό, τον αριθμό TPM (transcripts per million) που υπολογίζεται από τον τύπο TPM= (read count of transcript * average read length * 10^6)/ (length of the transcript * total transcript count) (Chatterjee et al., 2018). Πλήθος βιοπληροφορικών εργαλείων χρησιμοποιούνται για την κανονικοποίηση των δεδομένων όπως τα RSeQC, ERANGE, πακέτα του R (edgeR) και το Cufflink (Chatterjee et al., 2018). Μετά από το βήμα αυτό τα δεδομένα είναι κατάλληλα για την περαιτέρω ανάλυση τους καθώς είναι ομοιόμορφα και μπορούν να συγκριθούν μεταξύ τους.

Ανάλυση δεδομένων από scRNA-Seq

Όσον αφορά την ανάλυση των δεδομένων που προκύπτουν από την αλληλούχηση με τη μέθοδο scRNA-Seq μέχρι σήμερα δεν έχει βρεθεί κάποιο εργαλείο που να χρησιμοποιείται σε όλες τις έρευνες και να αποτελεί gold standard (Hwang et al., 2018). Αντίθετα υπάρχει πληθώρα διαθέσιμων βιοπληροφορικών εργαλείων που χρησιμοποιούνται ευρύτατα και τα περισσότερα από αυτά έχουν αναφερθεί παραπάνω (Hwang et al., 2018). Ειδικότερα γίνεται ποιοτικός έλεγχος των raw data με το εργαλείο FASTQC και read alignment με τα εργαλεία BWA ή STAR (Hwang et al., 2018). Ακολουθεί η κανονικοποίηση με τις μεθόδους που αναφέρθηκαν παραπάνω και τέλος υπολογίζονται οι παράγοντες που επηρεάζουν το πείραμα (confounding factors) όπως είναι οι βιολογικές διαφορές και ο τεχνικός θόρυβος (Hwang et al., 2018). Από την ανάλυση αυτή δεν προκύπτουν περισσότερα από ένα εκατομμύριο reads λόγω του μικρού αρχικού υλικού και του ορίου της ενίσχυσης από την PCR (Conesa et al., 2016). Αν υπήρχε όμως δυνατότητα για μεγαλύτερο βάθος αλληλούχησης θα βοηθούσε ιδιαίτερα στη μελέτη της έκφρασης των allele- specific γονιδίων. Άλλο ένα πρόβλημα

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

αποτελεί το μικρό μέγεθος του μεταγραφώματος σε ένα scRNA-Seq πείραμα (Conesa et al., 2016). Τα γονίδια που εκφράζονται σε κάθε κύτταρο είναι 3000-8000 συνεπώς είναι δύσκολο να υπολογιστεί ο τεχνικός θόρυβος που προκύπτει λόγω της ευαισθησίας της μεθόδου, από το βιολογικό θόρυβο (true biological noise) που οφείλεται στο γεγονός ότι ένα γονίδιο δεν εκφράζεται σε μία συγκεκριμένη χρονική στιγμή στο κύτταρο αλλά η πρωτεΐνη που δημιουργείται από αυτό είναι παρούσα (Conesa et al., 2016).

Επιλογή των κατάλληλων εργαλείων για κάθε ερευνητικό ερώτημα

Το συχνότερο ερώτημα σε μία RNA-Sequencing ανάλυση είναι η ανακάλυψη γονιδίων που παρουσιάζουν διαφορετική έκφραση μεταξύ των δειγμάτων που μελετώνται (Chatterjee et al., 2018). Για το λόγο αυτό έχουν δημιουργηθεί πολλά εργαλεία για την ανάλυση της διαφορετικής έκφρασης των γονιδίων που προκύπτει από το RNA-Sequencing (Chatterjee et al., 2018). Μερικά από τα εργαλεία αυτά είναι τα Cuffdiff2, DESeq(2), edgeR, Limma Voom, baySeq και το DEGseq (Chatterjee et al., 2018; Kukurba & Montgomery, 2015).

Καθένα από αυτά έχει τους δικούς του περιορισμούς και επομένως κάθε εργαλείο έχει τα δικά του μειονεκτήματα και πλεονεκτήματα. Η επιλογή λοιπόν του κατάλληλου εργαλείου είναι διαφορετική για κάθε ανάλυση και πρέπει να γίνεται με γνώμονα το πείραμα που πραγματοποιείται. Αυτό πολλές φορές δεν είναι εύκολο διότι δεν έχει γίνει σύγκριση των εργαλείων για τα διαφορετικά είδη δεδομένων που υπάρχουν και συνεπώς δεν έχει βρεθεί το καλύτερο πρόγραμμα για κάθε περίπτωση. Συνεπώς η επιλογή του κάθε εργαλείου εξαρτάται από την εμπειρία του ερευνητή, προτείνεται όμως η επιλογή δύο εργαλείων για μεγάλης σημασίας αναλύσεις (Chatterjee et al., 2018; Conesa et al., 2016; Kukurba & Montgomery, 2015).

Συνοπτικά κάποια από τα εργαλεία βιοπληροφορικής που παρουσιάστηκαν παραπάνω, η χρησιμότητά τους και που βασίζονται παρουσιάζονται στις εικόνες 11, 12 και 13 από (Chatterjee et al., 2018).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Tool	Based on	Notes/comments/suitable for
Tools for design and power-calculation in RNA-Seq		
SCOTTY	Web user interface	Determine appropriate sample size and read depth using publicly available or pilot data
RNASeqPowerCalculator	R	Calculates the power and sample size for differential expression analysis
Quality assessment of RNA-Seq data		
FASTQC	Java	Quality control toolkit for high throughput sequencing data
NGS QC	R	Quality control toolkit for high throughput sequencing data
Preprocessing RNA-Seq sequences		
Cleanadaptors	C	Removes adaptor sequences (component of DMAP software)
Trimmomatic	Java	Variety of trimming tasks for sequencing data
FASTX-toolkit	UNIX—shell scripts	Suite of tools for preprocessing of FASTA/FASTQ files
RNA-Seq specific parameter assessment		
RNA-SeQC	Java	Mutiple RNA-Seq specific quality analysis
RSeQC	Python and C	RNA-Seq specific quality and bias analysis
Qualimap 2	Java and R	Detects biases in the sequencing and/or mapping of the data from SAM/BAM file, could be used for ChIP-seq and whole-genome sequencing data

Εικόνα 11: Συνοπτικά εργαλεία βιοπληροφορικής που χρησιμοποιούνται στο RNA-Seq

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Tool	Based on	Notes/comments/suitable for
ComBat	R	Adjusts batch effects in expression data
Tools for RNA-Seq alignment (for Genome-based alignment)		
Tophat2	C	A spliced read mapper for RNA-Seq
STAR	C	Aligns RNA-Seq reads to reference genome using uncompressed suffix arrays
HISAT2	C	Spliced alignment program for mapping RNA-Seq reads
Tools for RNA-Seq alignment (for Transcriptome-based alignment)		
Bowtie	C	Fast, memory-efficient short read aligner
BWA		Short read sequence aligner
GSNAP	C, Perl	Fast detection of complex variants and splicing in short reads
Tools for RNA-Seq alignment (for De novo assembly alignment)		
Trinity	C	De novo reconstruction of transcriptomes from RNA-seq data
Oases	C	De novo transcriptome assembly from short reads
Tools for RNA-Seq alignment (for reference free alignment)		
Sailfish	C	Rapid alignment-free quantification of isoform abundance

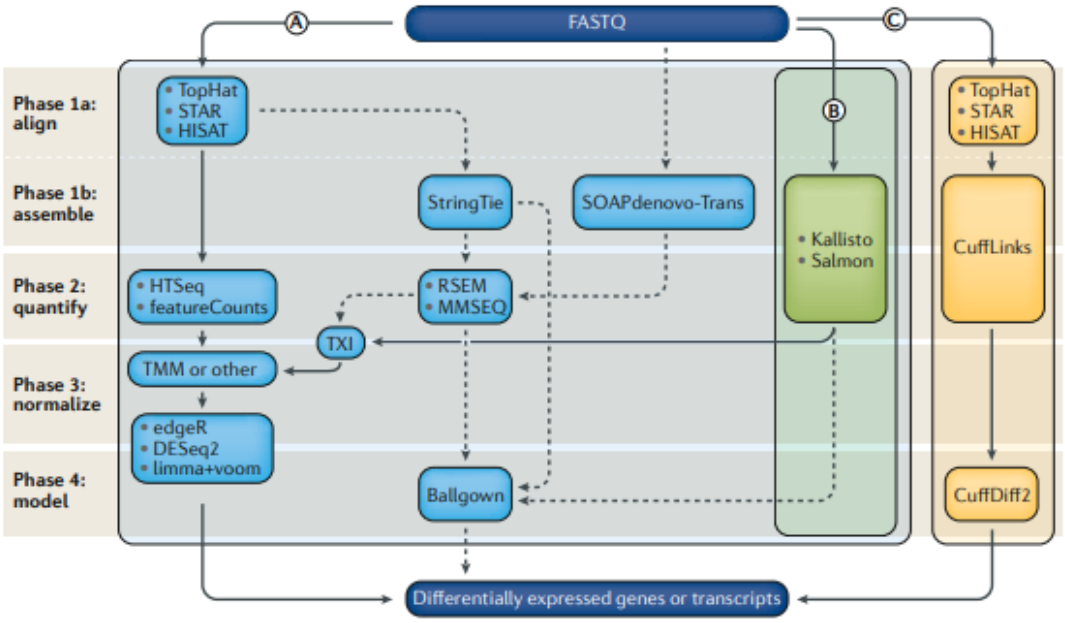
Εικόνα 12: Συνοπτικά εργαλεία βιοπληροφορικής που χρησιμοποιούνται στο RNA-Seq

Tools for differential expression analysis		
Cuffdiff2	C	RNA-Seq differentially expressed transcripts and genes
edgeR	R	RNA-Seq differential expression analyses, with reduced effect of outliers
DESeq2	R	RNA-Seq differential analysis of count data
Voom	R	Linear model analysis for RNA-Seq read counts (component of limma software)

Εικόνα 13: Συνοπτικά εργαλεία βιοπληροφορικής που χρησιμοποιούνται στο RNA-Seq

Ειδικότερα για την ανάλυση της διαφορετικής έκφρασης των γονιδίων οι φάσεις της ανάλυσης των δεδομένων και μερικά από τα εργαλεία βιοπληροφορικής που χρησιμοποιούνται περιγράφονται στην εικόνα 14 (Stark et al., 2019).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 14: Η ροή για την ανάλυση της διαφορετικής έκφρασης των γονιδίων από τα RNA-Seq δεδομένων

Εκτός από την ανάλυση της διαφορετικής έκφρασης των γονιδίων, το RNA-Sequencing χρησιμοποιείται και για τη μελέτη της διαφορετικής έκφρασης των ισομορφών ενός μετάγραφου (Conesa et al., 2016). Η ανάλυση αυτή ονομάζεται Alternative splicing analysis και τα εργαλεία που χρησιμοποιούνται για αυτή είναι το CuffDiff2 και το rSeqDiff (Conesa et al., 2016). Το πρώτο υπολογίζει αρχικά την έκφραση της κάθε ισομορφής και στη συνέχεια συγκρίνει τα αποτελέσματα μεταξύ τους, ενώ το δεύτερο χρησιμοποιεί ένα ιεραρχικό likelihood ratio test για να αναγνωρίσει ταυτόχρονα τη διαφορετική έκφραση των ισομορφών και τη διαφορετική έκφραση των γονιδίων (Conesa et al., 2016). Υπάρχουν και άλλα εργαλεία που βασίζονται στην ανίχνευση των διαφορετικών ισομορφών συγκρίνοντας την κατανομή των reads στα εξώνια και στα junctions των γονιδίων μεταξύ των δειγμάτων που μελετώνται (Conesa et al., 2016). Η μέθοδος αυτή ονομάζεται exon based approach και εργαλεία που χρησιμοποιούν τέτοιες μεθόδους για alternative splicing analysis είναι τα DEXseq, DSGSeq, rMATS και το DiffSplice που είναι κατάλληλα για την ταυτοποίηση μεμονωμένων περιστατικών εναλλακτικού ματίσματος (Conesa et al., 2016).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

RNA-Sequencing χρησιμοποιείται για την ανίχνευση allele-specific expression καθώς λόγω του resolution που παρουσιάζει αυτή η μέθοδος δίνει τη δυνατότητα υπολογισμού της έκφρασης των μητρικών και πατρικών αλλήλιων (Kukurba & Montgomery, 2015). Με τις μελέτες αυτές αναμένεται να δοθούν απαντήσεις σχετικά με τα επιγενετικά φαινόμενα, τη γενετική ποικιλομορφία και τις διαφορές μεταξύ υγιών και παθολογικών ιστών (Kukurba & Montgomery, 2015).

Ακόμα τα δεδομένα από το RNA-Sequencing συνδυάζονται με δεδομένα που σχετίζονται με τη γενετική ποικιλότητα, όπως γονοτυπικά δεδομένα (genotyping data) (Kukurba & Montgomery, 2015). Οι μελέτες που συνδυάζουν τα δεδομένα αυτά έχουν οδηγήσει στην ανακάλυψη γενετικών τόπων που σχετίζονται με τη διαφορετική έκφραση γονιδίων (Kukurba & Montgomery, 2015). Οι γενετικοί αυτοί τόποι ονομάζονται expression quantitative trait loci (eQTLs) και οι διαφορές στην έκφραση φαίνεται να σχετίζονται με variants που σχετίζονται με τα διαφορετικά φαινοτυπικά χαρακτηριστικά και τις ασθένειες που εμφανίζουν τα άτομα (Kukurba & Montgomery, 2015). Επομένως η ανάλυση των eQTLs δίνει τη δυνατότητα ανακάλυψης βιολογικών διεργασιών, γενετικών αλλαγών και μοριακών μονοπατιών που προκαλούν ασθένειες (Kukurba & Montgomery, 2015). Με το RNA-Sequencing γίνεται επίσης ταυτοποίηση των variants που επηρεάζουν την έκφραση των διαφόρων ισομορφών μίας πρωτεΐνης που προκύπτουν από εναλλακτικό μάτισμα (splicing-QTLs, sQTLs) (Kukurba & Montgomery, 2015). Για τον έλεγχο των συσχετίσεων αυτών χρησιμοποιούνται αρκετά εργαλεία όπως το ANOVA ενώ πρόσφατα έχει δημιουργηθεί και το λογισμικό Matrix eQTL (Kukurba & Montgomery, 2015).

Παρακάτω αναλύονται μερικά από τα βασικότερα εργαλεία βιοπληροφορικής. Τα εργαλεία αυτά είναι απαραίτητα για τις περισσότερες αναλύσεις των δεδομένων από RNA-Seq και χρησιμοποιούνται ευρύτατα. Φυσικά, υπάρχουν και άλλα εργαλεία βιοπληροφορικής όπως αυτά που αναφέρθηκαν παραπάνω τα οποία είναι εξίσου εύχρηστα και αξιόπιστα και μεταξύ αυτών ο ερευνητής επιλέγει αυτό που ταιριάζει στο ερευνητικό ερώτημα που θέλει να απαντήσει.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Ειδικό Μέρος

Για την ανάλυση της διαφορετικής έκφρασης των γονιδίων το πρώτο εργαλείο που χρησιμοποιείται είναι το FASTQC για τον ποιοτικό έλεγχο των δεδομένων που προκύπτουν. Ακολουθεί το alignment των reads με το εργαλείο STAR κάνοντας το alignment βάσει του reference genome και με το εργαλείο featureCounts γίνεται η ποσοτικοποίηση των reads. Τέλος η κανονικοποίηση και η ανάλυση της διαφορετικής έκφρασης των γονιδίων γίνεται με εργαλεία του R, όπως το edgeR.

FASTQC

Το FastQC είναι το εργαλείο που χρησιμοποιείται συχνότερα για τον έλεγχο των raw sequencing data και δημιουργεί πίνακες και γραφήματα με στόχο την εύκολη αξιολόγηση των δεδομένων. Σαν δεδομένα για την ανάλυση με το εργαλείο αυτό χρησιμοποιούνται αρχεία της μορφής FASTQ, SAM και BAM και όταν σαν λειτουργικό σύστημα χρησιμοποιείται το Linux, το FASTQC «τρέχει» με την εντολή:

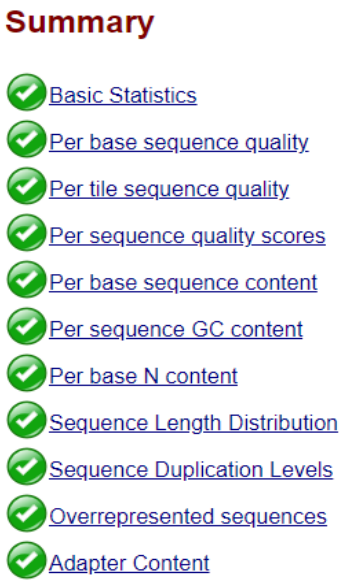
fastqc mysample1_R1.fastq

όπου mysample1_R1.fastq το αρχείο ενδιαφέροντος (Pevsner, 2015). Όταν το FastQC χρησιμοποιείται μέσω του Galaxy επιλέγονται τα Tools>NGS: QC>FastQC και επιλέγεται το αρχείο FASTQ για το οποίο δημιουργούνται τα σχεδιαγράμματα σε μορφή HTML (Pevsner, 2015). Τα αποτελέσματα του FastQC αναλύονται παρακάτω και εξαρτώνται από τα δεδομένα που λαμβάνει το πρόγραμμα.

Όταν τα δεδομένα είναι σωστά στο δεξί μέρος του παραθύρου που έχει ανοίξει εμφανίζονται κάποιοι πράσινοι κύκλοι που υποδηλώνουν την εγκυρότητα των δεδομένων, δηλαδή δηλώνουν ότι είναι φυσιολογικά όπως φαίνεται και στην εικόνα 15, (*FastQC Report Good_sequence_short.Txt* , n.d.)) (*FastQC*, n.d.-a; *FastQC Report Good_sequence_short.Txt* , n.d.).

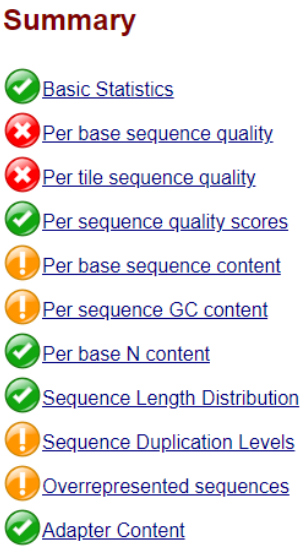
Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Summary



Εικόνα 15:Summary normal

Όπως φαίνεται λοιπόν το FastQC κάνει έλεγχο για όλα τα παραπάνω και με πράσινο παρουσιάζει ότι μετά τους ελέγχους που έγιναν τα δεδομένα είναι σωστά. Ο χρήστης έχει τη δυνατότητα να ελέγξει τα γραφήματα και τους πίνακες που δημιουργήθηκαν στο δεξί μέρος που παραθύρου, όπως θα αναφερθεί και παρακάτω. Όταν τα δεδομένα δεν είναι φυσιολογικά εμφανίζεται στα αριστερά η εικόνα 16 (*FastQC Report Bad_sequence.Txt* , n.d.).



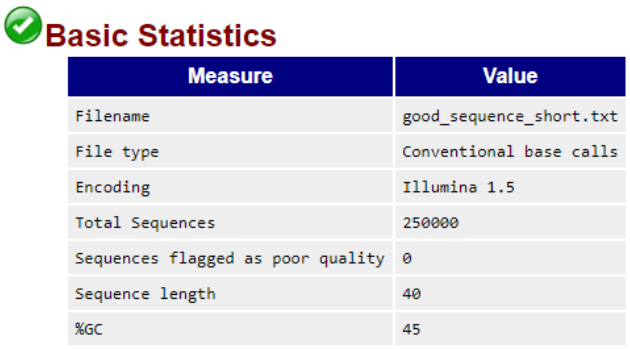
Εικόνα 16:Summary failed

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Σύμφωνα με αυτή την εικόνα (Εικόνα 16), το πορτοκαλί θαυμαστικό υποδηλώνει ότι τα δεδομένα είναι slightly abnormal, ενώ το κόκκινο x ότι τα δεδομένα είναι εκτός των φυσιολογικών πλαισίων. Τα αποτελέσματα αυτά πολλές φορές ξεπερνούνται από το χρήστη ανάλογα με το σκοπό του πειράματος και ανάλογα των γραφημάτων που έχουν δημιουργηθεί στο αριστερό μέρος του παραθύρου, το οποίο θα αναφερθεί παρακάτω (*FastQC*, n.d.-a). Πατώντας σε μία από τις υποκατηγορίες του summary εμφανίζονται οι πίνακες και τα γραφήματα που έχουν δημιουργηθεί.

Basic Statistics

Πατώντας στο Basic Statistics εμφανίζεται η εικόνα 17 (*FastQC Report Good_sequence_short.Txt* , n.d.) που περιέχει πληροφορίες για το όνομα του αρχείο που αναλύθηκε, τον τύπο του αρχείου αυτού (conventional base calls ή colorspace data), το encoding, το συνολικό αριθμό των αλληλουχιών, τις φιλτραρισμένες αλληλουχίες, το μήκος της μικρότερης και μεγαλύτερης αλληλουχίας που μελετήθηκε και το ποσοστό GC των αλληλουχιών. Το Basic Statistics δεν εμφανίζει ποτέ μήνυμα προειδοποίησης ή λάθους (*FastQC*, n.d.-a).



Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Εικόνα 17: Basic statistics

Per Base Sequence Quality

Η επιλογή Per Base Sequence Quality αποτελεί μία οπτική αναπαράσταση των quality values για όλες τις βάσεις που υπάρχουν στο αρχείο που μελετάται. Πιο συγκεκριμένα, δημιουργείται ένα γράφημα τύπου BoxWhisker όπου η κόκκινη γραμμή αναπαριστά τη διάμεσο, το κίτρινο πλαίσιο το ενδοτεταρτημοριακό εύρος και η μπλε γραμμή τη μέση ποιότητα της κάθε βάσης. Τα quality scores εμφανίζονται στον άξονα των y, ο

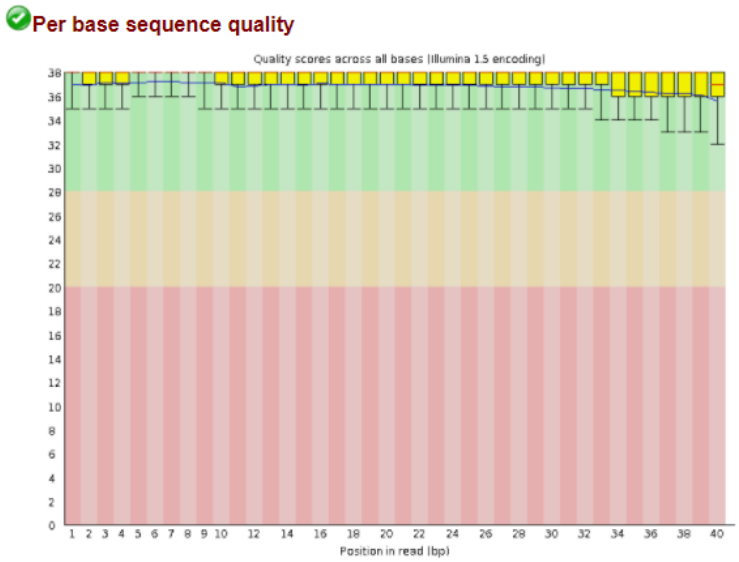
Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

οποίος όπως φαίνεται και στην εικόνα παρακάτω χωρίζεται σε τρεις κατηγορίες. Το πράσινο χαρακτηρίζει τα πολύ καλά quality scores, το πορτοκαλί τα μέτρια quality scores και το κόκκινο τα κακά quality scores. Γενικά όσο πιο υψηλά είναι τα score τόσο πιο αξιόπιστη είναι η ανάλυση (*FastQC*, n.d.-a). Συνήθως προτείνεται να απομακρύνονται από τη μελέτη βάσεις που τα outliers τους έχουν απόκλιση μεγαλύτερη του 30%. Τα 3΄ άκρα των reads έχουν συνήθως χαμηλά quality scores, όπως φαίνεται και στην εικόνα 18 (*FastQC Report Bad_sequence.Txt* , n.d.),οπότε είναι καλό να απομακρύνονται για να βελτιώνεται η ανάλυση των δεδομένων στη συνέχεια (Conesa et al., 2016).

Εικόνα 18: Per base sequence quality failed

Όπως λοιπόν φαίνεται και στην εικόνα 18 η ποιότητα των δεδομένων ειδικά προς το 3΄ άκρο είναι κακή. Αντίθετα στην εικόνα 19 (*FastQC Report Good_sequence_short.Txt* , n.d.) τα quality scores είναι υψηλά και είναι η εικόνα που αναμένεται αν τα δεδομένα μας είναι σωστά.

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

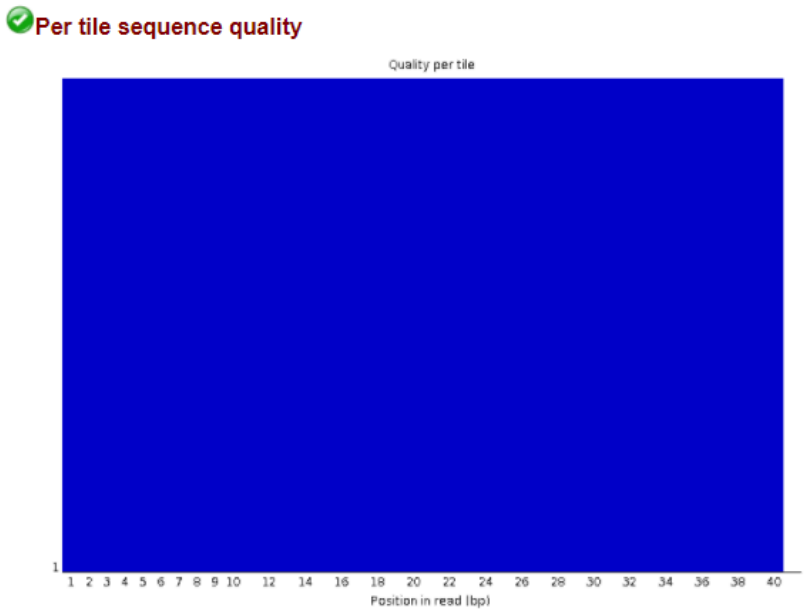


Εικόνα 19: Per base sequence quality normal

Per Tile Sequence Quality

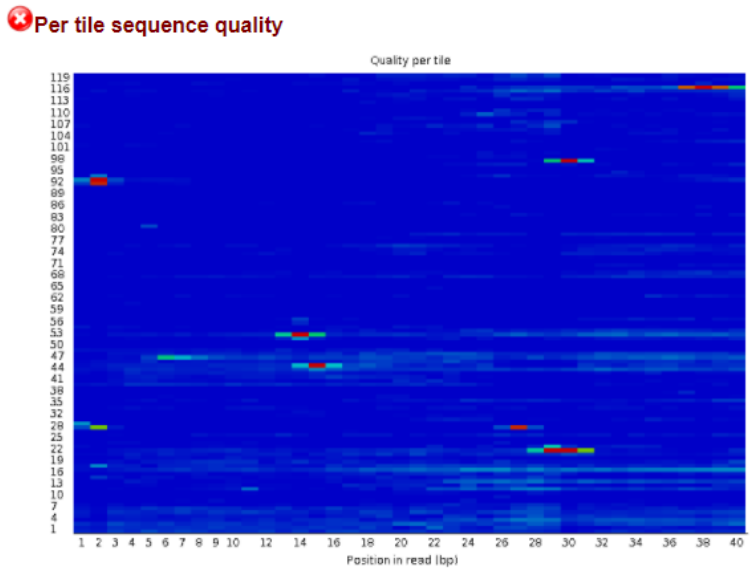
Το Per Tile Sequence Quality εμφανίζεται όταν χρησιμοποιούνται βιβλιοθήκες της Illumina που περιέχουν τους sequence identifiers και σχετίζονται με το flowcell tile από το οποίο προήλθε κάθε read. Το γράφημα που δημιουργείται παρουσιάζει την απόκλιση που έχει κάθε tile από το μέσο quality (*Per Tile Sequence Quality*, n.d.). Το επιθυμητό γράφημα που δημιουργείται είναι το γράφημα της εικόνας 20 (*FastQC Report Good_sequence_short.Txt* , n.d.).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 20:Per tile sequence quality normal

Αντίθετα, η εικόνα 21 (*FastQC Report Bad_sequence.Txt* , n.d.) είναι αυτή που δείχνει ότι υπάρχει πρόβλημα στα δεδομένα. Πιο συγκεκριμένα, τα χρώματα των γραφημάτων αυτών παρουσιάζουν με θερμά χρώματα τις περιοχές όπου το tile έχει χειρότερα qualities scores σε σχέση με άλλα tiles για την ίδια βάση (*Per Tile Sequence Quality*, n.d.).



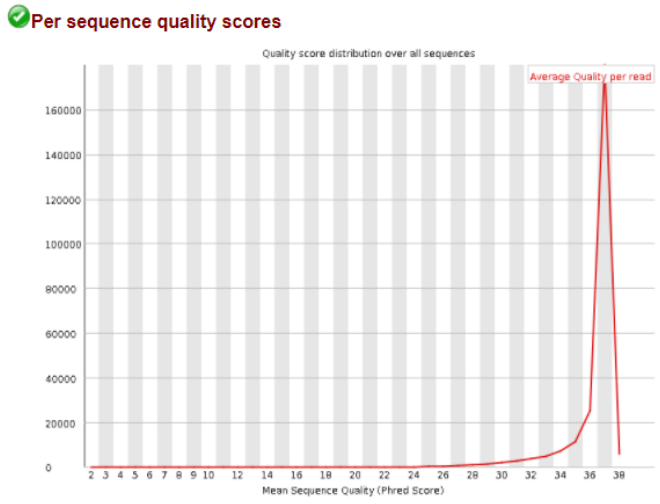
Εικόνα 21: Per tile sequence quality failed

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Συνεπώς ένα καλό γράφημα είναι παντού μπλε ενώ ένα γράφημα με κόκκινο έχει αρκετά προβλήματα (*Per Tile Sequence Quality*, n.d.). Τα προβλήματα αυτά πιθανώς δημιουργούνται από φυσαλίδες ή κηλίδες ή θραύσματα που υπάρχουν στο flowcell (*Per Tile Sequence Quality*, n.d.).

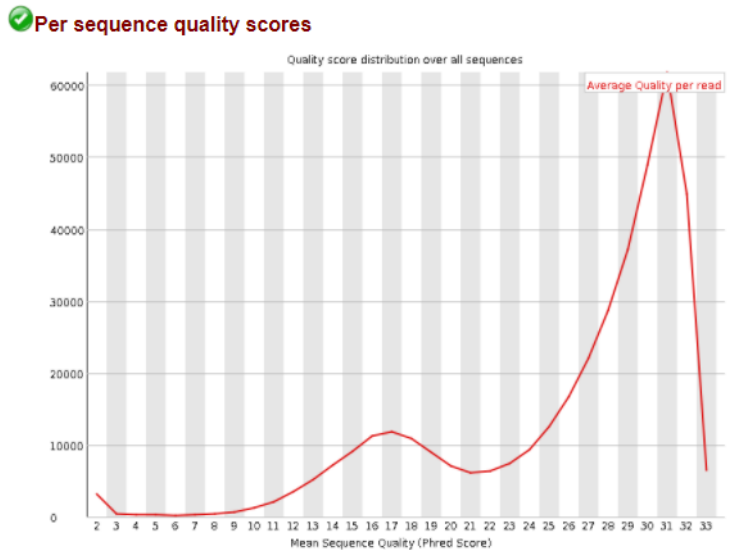
Per Sequence Quality Scores

Το Per Sequence Quality Scores παρέχει πληροφορίες σχετικά με το αν κάποια από τις αλληλουχίες που μελετήθηκαν έχει συνολικά χαμηλό quality value. Αν συμβαίνει αυτό σημαίνει συνήθως ότι υπάρχει κάποιο συστημικό λάθος. Όπως φαίνεται και στις εικόνες 22 και 23 (*FastQC Report Bad_sequence.Txt* , n.d.; *FastQC Report Good_sequence_short.Txt* , n.d.), η καμπύλη πρέπει να τείνει στο μηδέν στην αρχή του γραφήματος και στη συνέχεια να αυξάνεται χωρίς να υπάρχουν διακυμάνσεις (*FastQC*, n.d.-a).



Εικόνα 22: Per sequence quality scores normal

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

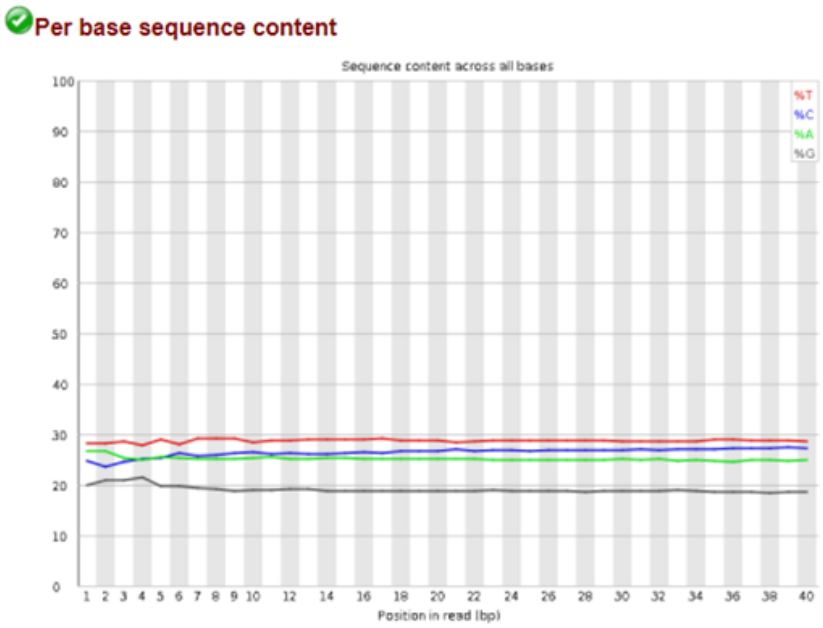


Εικόνα 23: Per sequence quality scores normal

Per Base Sequence Content

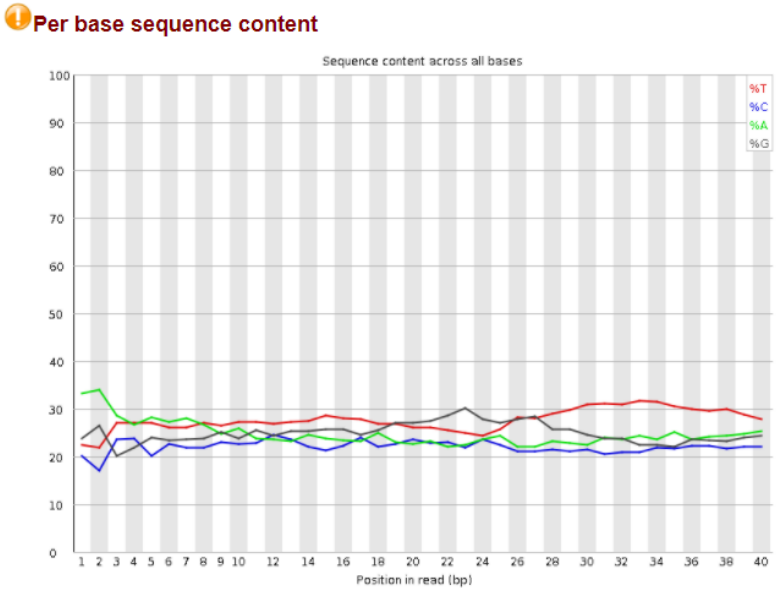
Στη συνέχεια του Summary είναι το Per Base Sequence Content. Σε αυτό δημιουργείται ένα γράφημα στο οποίο εμφανίζεται το ποσοστό που κάθε βάση εμφανίζεται σε κάθε θέση (*FastQC*, n.d.-a). Στην εικόνα 24 (*FastQC Report Good_sequence_short.Txt* , n.d.) παρουσιάζεται το αναμενόμενο γράφημα από τη συγκεκριμένη ανάλυση. Σε αυτό φαίνεται ότι για μία τυχαία βιβλιοθήκη το ποσοστό της κάθε βάσης δεν πρέπει να εμφανίζει μεγάλες διαφορές με το ποσοστό των υπόλοιπων βάσεων και γενικά το ποσοστό κάθε βάσης θα πρέπει να αντικατοπτρίζει το ποσοστό της κάθε βάσης στο γένωμα του οργανισμού που μελετάται (*Per Base Sequence Content*, n.d.). Δηλαδή πρέπει οι γραμμές που εμφανίζονται να είναι όσο το δυνατόν παράλληλες (*FastQC*, n.d.-a).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 24: Per base sequence content normal

Αντίθετα, η εικόνα 25 (*FastQC Report Bad_sequence.Txt* , n.d.) δεν είναι ιδιαίτερα ικανοποιητική και προκαλεί προβλήματα κατά τον έλεγχο των δεδομένων.



Εικόνα 25: Per base sequence content concerning

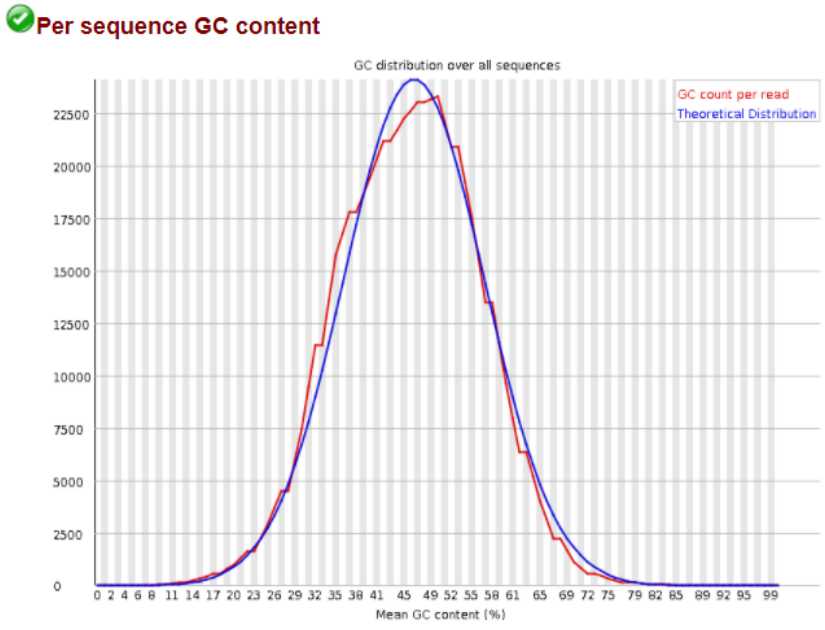
Στην εικόνα αυτή (Εικόνα 25) παρουσιάζεται Warning (πορτοκαλί θαυμαστικό) διότι οι βάσεις διαφέρουν για κάθε θέση σε ποσοστό μεγαλύτερο του 10% (*Per Base*

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Sequence Content, n.d.). Οι διαφορές στο 5΄ άκρο των αλληλουχιών είναι γενικά αποδεκτές καθώς οι περισσότερες βιβλιοθήκες δημιουργούν αυτό το πρόβλημα. Οι βιβλιοθήκες που δημιουργούν αυτό το πρόβλημα είναι αυτές στις οποίες χρησιμοποιήθηκαν τυχαία εξαμερή κατά το priming και οι βιβλιοθήκες στις οποίες έγινε fragmentation με τρανσποζάση (tagmentation) (*Per Base Sequence Content*, n.d.). Τα λάθη αυτά είναι τεχνικά σφάλματα όμως δεν μπορούν να διορθωθούν. Παρόλα αυτά στις περισσότερες περιπτώσεις δεν φαίνεται να επηρεάζουν τα επόμενα βήματα της ανάλυσης (*Per Base Sequence Content*, n.d.).

Per Sequence GC Content

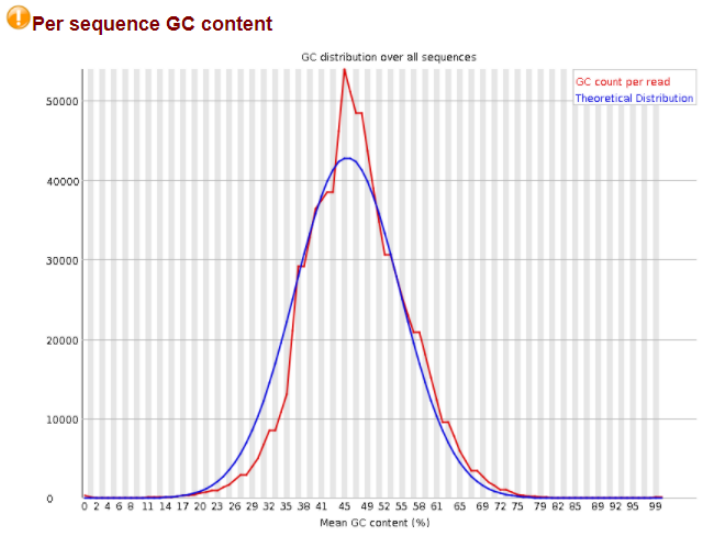
Ακολουθεί το Per Sequence GC Content όπου συγκρίνεται το περιεχόμενο σε GC σε όλο το μήκος κάθε αλληλουχίας με ένα μοντέλο που παρουσιάζει τη θεωρητική κατανομή των GC. Όπως φαίνεται και στην εικόνα 26 (*FastQC Report Good_sequence_short.Txt* , n.d.),οι δύο καμπύλες σχεδόν συμπίπτουν και αυτή η εικόνα αποτελεί ένδειξη ότι τα δεδομένα είναι σωστά. Το peak της καμπύλης πρέπει να ταυτίζεται με το συνολικό περιεχόμενο σε GC του γενώματος του οργανισμού που μελετάται. Πολλές φορές όμως ο αριθμός αυτός δεν είναι γνωστός και για το λόγο αυτό λαμβάνεται υπόψιν μία υποθετική κατανομή (*Per Sequence GC Content*, n.d.).



Εικόνα 26: Per sequence GC content normal

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Πολλές φορές εμφανίζεται μία ελαφρά ανώμαλη κατανομή όπως αυτή της εικόνας 27 (*FastQC Report Bad_sequence.Txt* , n.d.) ή και ακόμα πιο διαφοροποιημένες από την κανονική κατανομή.



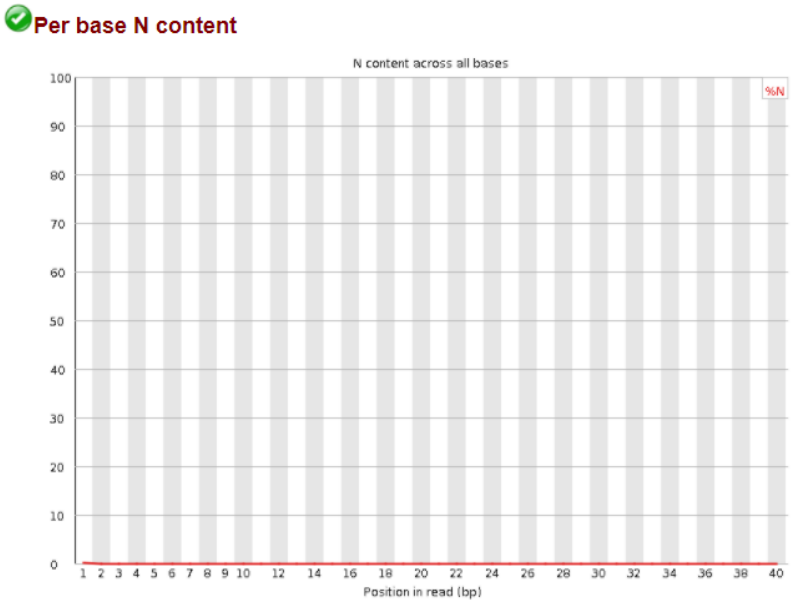
Εικόνα 27:Per sequence GC content concerning

Όταν το άθροισμα των αποκλίσεων από την κανονική κατανομή είναι μεγαλύτερη από το 15% των reads εμφανίζεται μήνυμα προειδοποίησης, ενώ αν είναι μεγαλύτερη από 20% μήνυμα λάθους. Γενικά η ανώμαλη κατανομή αποτελεί ένδειξη συστημικού λάθους ή βιβλιοθήκης με επιμόλυνση (*Per Sequence GC Content*, n.d.).

Per Base N Content

Το Per Base N Content δημιουργεί ένα γράφημα με το ποσοστό των N για κάθε θέση της αλληλουχίας. Ειδικότερα, το N είναι μία βάση που καλείται κατά την αλληλούχηση όταν το μηχάνημα δεν αναγνωρίζει με ακρίβεια μία από τις τέσσερις βάσεις A,T,G,C. Το γράφημα που δημιουργείται πρέπει να είναι το παρακάτω (Εικόνα 28) (*FastQC Report Good_sequence_short.Txt* , n.d.), δηλαδή η γραμμή να τείνει στο μηδέν (*Per Base N Content*, n.d.).

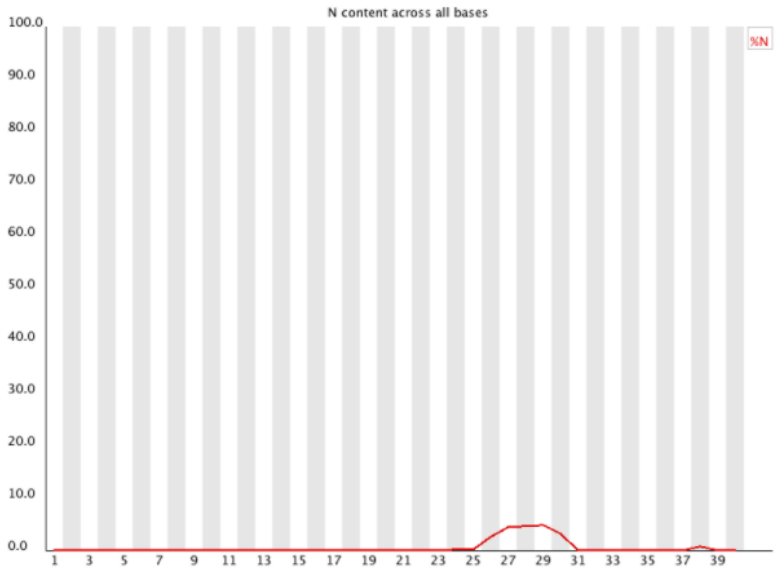
Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 28: Per Base N content normal

Αντίθετα αν η εικόνα είναι η παρακάτω (Εικόνα 29) (*Per Base N Content*, n.d.), υποδηλώνεται ότι η ανάλυση δεν έγινε σωστά και πιο συγκεκριμένη δείχνει ότι υπάρχει μία γενική χαμηλή ποιότητα στη μελέτη. Εμφανίζεται έτσι μήνυμα προειδοποίησης όταν το ποσοστό των N είναι μεγαλύτερο από 5% και μήνυμα λάθους όταν είναι μεγαλύτερο από 20% τα οποία όμως πρέπει να αναλυθούν και με τη βοήθεια των υπολοίπων γραφημάτων και πινάκων από την ανάλυση FastQC (*Per Base N Content*, n.d.).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

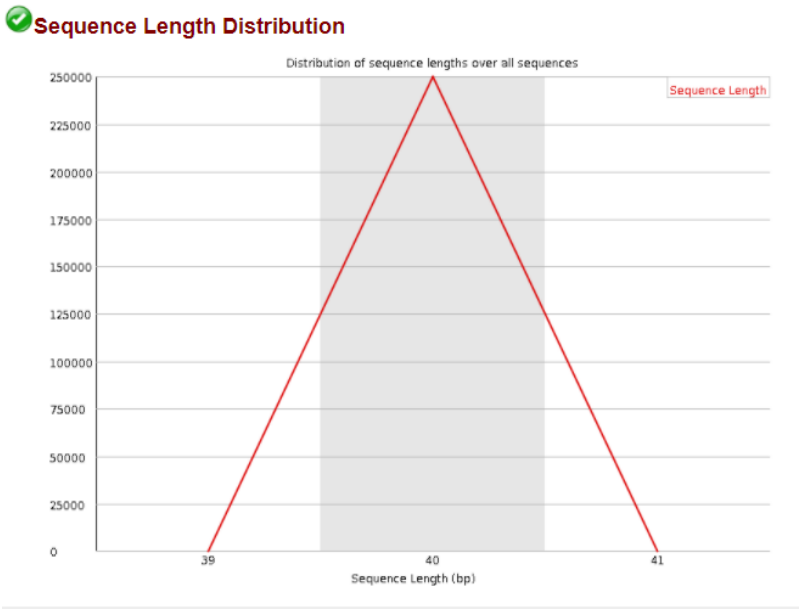


Εικόνα 29: N content across all bases concerning

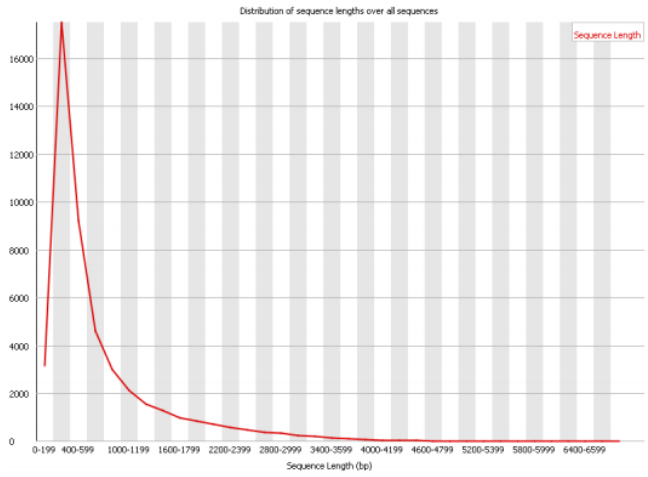
Sequence Length Distribution

Το Sequence Length Distribution αποτελείται από ένα γράφημα που παρουσιάζει το μήκος των θραυσμάτων (fragments) στο αρχείο που μελετήθηκε. Το μήκος των τμημάτων αυτών εξαρτάται από το είδος του πειράματος καθώς πολλές φορές οι βιβλιοθήκες που δημιουργούνται έχουν αλληλουχίες με διαφορετικό μήκος. Για το λόγο αυτό από αυτή την ανάλυση θα προκύψει μήνυμα λάθους μόνο όταν κάποια από αλληλουχία έχει μήκος μηδέν και μήνυμα προειδοποίησης όταν οι αλληλουχίες δεν έχουν το ίδιο μήκος. Επομένως και τα δύο γραφήματα 30 και 31 που παρουσιάζονται παρακάτω είναι αποδεκτά (*FastQC Report Good_sequence_short.Txt* , n.d.) (*FastQC*, n.d.-a) .

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 30: Sequence Length Distribution normal



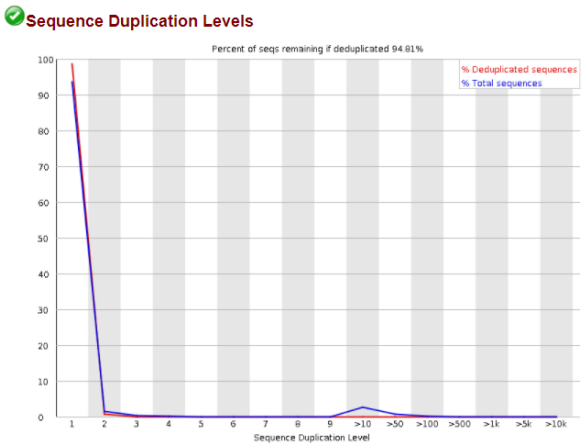
Εικόνα 31: Sequence Length Distribution normal

Sequence Duplication Levels

Ακολουθεί το Sequence Duplication Levels όπου μετριέται ο αριθμός των αντιγράφων για κάθε αλληλουχία. Δημιουργείται έτσι ένα γράφημα όπως φαίνεται παρακάτω. Για την ανάλυση αυτή λαμβάνονται κάποιοι περιορισμοί ώστε για γίνει εξοικονόμηση μνήμης και πληροφοριών που δεν είναι απαραίτητα. Στο γράφημα που δημιουργείται η μπλε γραμμή παρουσιάζει την κατανομή των duplication levels στο σύνολο των αλληλουχιών ενώ η κόκκινη γραμμή το ποσοστό των deduplicated αλληλουχίες. Το

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

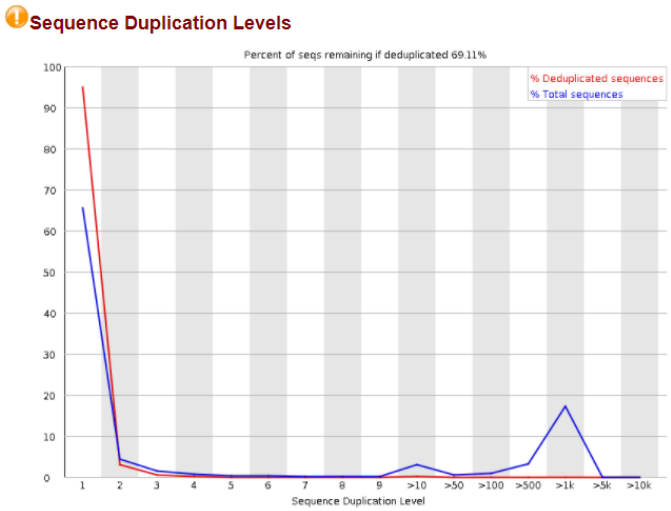
επιθυμητό γράφημα είναι το παρακάτω (εικόνα 32)(*FastQC Report Good_sequence_short.Txt* , n.d.) που όπως φαίνεται οι περισσότερες αλληλουχίες εμφανίζονται μόνο μία φορά και βρίσκονται συνήθως στο αριστερό τμήμα του γραφήματος. Ακόμα στο πάνω μέρος του γραφήματος υπολογίζεται ένα ποσοστό που δίνει μία ένδειξη του συνολικού επιπέδου της αλληλουχίας που πιθανόν έχει χαθεί (*Duplicate Sequences*, n.d.).



Εικόνα 32: Sequence Duplication Levels normal

Αντίθετα το παρακάτω γράφημα (Εικόνα 33) (*FastQC Report Bad_sequence.Txt* , n.d.) είναι ένα γράφημα που οδηγεί σε προειδοποίηση από το σύστημα. Πιο συγκεκριμένα, αν αλληλουχίες που δεν είναι μοναδικές αποτελούν παραπάνω από το 20% του συνόλου εμφανίζεται μήνυμα προειδοποίησης, ενώ αν είναι πάνω από 50% εμφανίζεται μήνυμα λάθους (*Duplicate Sequences*, n.d.).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

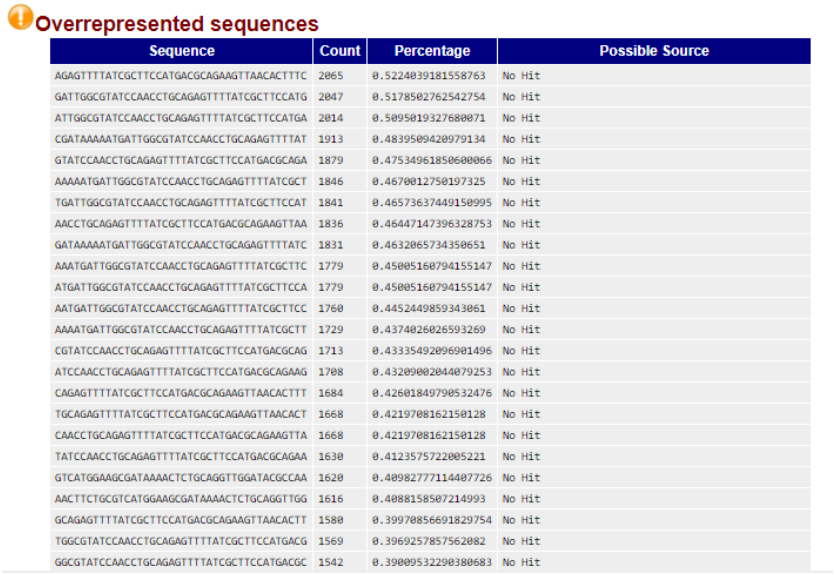


Εικόνα 33: Sequence Duplication Levels concerning

Overrepresented sequences

Όσον αφορά τις overrepresented sequences, όταν τα δεδομένα είναι σωστά εμφανίζεται μήνυμα ενημέρωσης ότι δεν υπάρχουν overrepresented sequences (*FastQC Report Good_sequence_short.Txt* , n.d.). Αντίθετα όταν μία αλληλουχία έχει μεγάλη βιολογική αξία, η βιβλιοθήκη έχει επιμολυνθεί ή οι αλληλουχίες της βιβλιοθήκης δεν είναι διαφοροποιημένες εμφανίζονται overrepresented sequences. Παρουσιάζονται έτσι σε ένα πίνακα, όπως φαίνεται και παρακάτω (Εικόνα 34), όπου παρουσιάζονται οι αλληλουχίες αυτές. Όταν οι αλληλουχίες αυτές αντιπροσωπεύουν περισσότερο από το 0,1% του συνόλου εμφανίζεται μήνυμα προειδοποίησης, ενώ όταν αντιπροσωπεύουν περισσότερο από το 1% που συνόλου εμφανίζεται μήνυμα λάθους. Οι αλληλουχίες που ελέγχονται είναι μόνο οι πρώτες 100.000 συνεπώς όλες οι αλληλουχίες που ακολουθούν δεν υπολογίζονται στην ανάλυση αυτή και είναι κάτι που πρέπει να ληφθεί υπόψιν. Για κάθε αλληλουχία που εντοπίζεται γίνεται έλεγχος στις βάσεις δεδομένων για πιθανές επιμολύνσεις που έγιναν (*Overrepresented Sequences*, n.d.). Πιο συγκεκριμένα, γίνεται αναζήτηση των αλληλουχιών αυτών στα εργαλεία BLAT, BLASTN και BLAST ανάλογα με το αν είναι γνωστό από ποιο είδος προέρχονται και είναι αναγκαία η αναζήτηση όμοιων αλληλουχιών σε όλα τα είδη (Pevsner, 2015). Ο πίνακας που δημιουργείται βρίσκεται παρακάτω (Εικόνα 34) (*FastQC Report Bad_sequence.Txt* , n.d.).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.4753496185060066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.467001275019732S	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.4657363744915099S	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.4644714739632875S	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.463206573435065S	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.4500516079415514S	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.4500516079415514S	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.445244985934306S	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.437402602659326S	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.4333549209690149S	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.4320900204407925S	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.426018497905324S	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.421970816215012S	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.421970816215012S	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.412357572200522S	No Hit
GTCATGGAAGCGATAAAACTCTGCAGGTTGGATACGCCAA	1620	0.4098277711440772S	No Hit
AACTTCTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGG	1616	0.408815850721499S	No Hit
GCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTT	1580	0.3997085669182975S	No Hit
TGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACG	1569	0.396925785756208S	No Hit
GGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGC	1542	0.390095322903806S	No Hit

Εικόνα 34: Overrepresented sequences failed

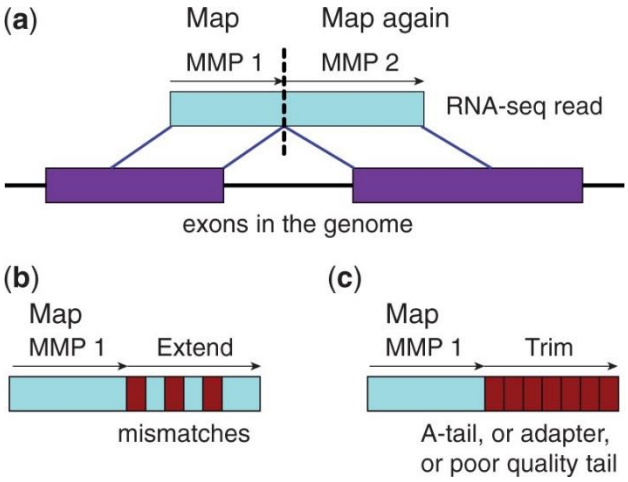
Spliced Transcripts Alignment to Reference (STAR)

Το STAR (Spliced Transcripts Alignment to Reference) είναι ένα λογισμικό που δημιουργήθηκε από τον Alexander Dobin και τους συνεργάτες του το 2012 με στόχο να λύσει τα προβλήματα των άλλων λογισμικών που χρησιμοποιούνταν για το alignment των δεδομένων που προέρχονταν από high-throughput RNA-Sequencing και έχει γραφτεί σε γλώσσα προγραμματισμού C++ (Dobin et al., 2013). Το πρόγραμμα αυτό προσφέρει την καλύτερη ταχύτητα mapping σε σχέση με τις άλλες μεθόδους και βελτιώνει την ακρίβεια και ευαισθησία του alignment (Dobin et al., 2013). Ακόμα κάνει mapping σε ολόκληρες αλληλουχίες RNA και ανακαλύπτει με ικανοποιητική ακρίβεια μετάγραφα που προκύπτουν από εναλλακτικό μάτισμα (Dobin et al., 2013).

Ο αλγόριθμος αυτός βασίζεται στο alignment των ασυνεχών αλληλουχιών από το RNA-Seq κατευθείαν στο γονιδίωμα αναφοράς (reference genome), κάνοντας αρχικά ένα seed searching step που ακολουθείται από ένα clustering-stitching-scoring step (Dobin et al., 2013). Κατά το seed searching step, γίνεται αναζήτηση για το Maximal Mappable Prefix (MMP) όπου το *MMP(R,i,G)* ορίζεται ως το μακρύτερο υπόστρωμα (substring) $(R_i, R_{i+1}, \ldots, R_{i+MML-1})$ που ταυτίζεται με ένα ή περισσότερα υποστρώματα του G. Όπου R είναι η αλληλουχία του read, i η τοποθεσία (location) του read, G η αλληλουχία του reference genome και MML το μέγιστο mappable μήκος (Dobin et al., 2013). Πιο συγκεκριμένα, αρχικά ο αλγόριθμος βρίσκει το MMP αλλά

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

επειδή το read αντιστοιχεί σε μία περιοχή που γίνεται splice junction το read που μελετάται δεν χαρτογραφείται συνεχόμενα στο reference genome (Dobin et al., 2013). Επομένως το μέρος του read που χαρτογραφείται εναποτίθεται σε ένα μέρος του γονιδιώματος αναφοράς (Dobin et al., 2013). Στη συνέχεια, ψάχνεται ξανά MMP για τις περιοχές των reads που δεν έχουν χαρτογραφηθεί στο γονιδίωμα αναφοράς και όταν βρεθεί το MMP εναποτίθεται σε ένα άλλο μέρος του γονιδιώματος (Dobin et al., 2013). Η μέθοδος αυτή εξηγείται στην εικόνα 35a (Dobin et al., 2013).



Εικόνα 35: Αναπαράσταση της μεθόδου Maximum Mappable Prefix (MMP) του STAR. Στην εικόνα γίνεται αναζήτηση των a) θέσεων ματίσματος b) mismatches c) tails

Η αναζήτηση των MMP δίνει τη δυνατότητα ανίχνευσης mismatches, indels και ουρών, όπως poly(A) ουρών, ουρών με χαμηλή ποιότητα αλληλούχησης και αλληλουχιών που περιέχουν adapters, όπως φαίνεται στις εικόνες 35b, 35c (Dobin et al., 2013). Η αναζήτηση των MMPs γίνεται και προς τις δύο κατευθύνσεις (forward, reverse) και η αναζήτηση μόνο των περιοχών των reads που δεν έχουν χαρτογραφηθεί παρέχει ένα σημαντικό πλεονέκτημα χρόνου στον αλγόριθμο STAR. Ακόμα η αναζήτηση αυτή γίνεται με τη χρήση uncompressed suffix arrays (SAs) και ανιχνεύονται έτσι συμβάντα ματίσματος χωρίς καμία προηγούμενη γνώση σχετικά με την περιοχή ή τις ιδιότητες του ματίσματος (Dobin et al., 2013).

Αφού λοιπόν τα τμήματα των reads έχουν εναποτεθεί σε συγκεκριμένες θέσεις του γονιδιώματος αναφοράς ο αλγόριθμος συρράπτει τα τμήματα αυτά (Dobin et al., 2013). Η συρραφή εξαρτάται από ένα score που προκύπτει, όπου ο συνδυασμός των συρραφών που έχουν το μεγαλύτερο score επιλέγεται ως το καταλληλότερο alignment

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

για το read (Dobin et al., 2013). Το score αυτό προκύπτει από τα matches, τα mismatches, τις εισαγωγές (insertions), τις ελλείψεις (deletions) και τα κενά στα splice junctions (Dobin et al., 2013).

Το STAR είναι ένα αξιόπιστο λογισμικό που χρησιμοποιείται από ENCODE και κάνει align στο ανθρώπινο γονιδίωμα σε 550 εκατομμύρια 2* 76 nt reads την ώρα με έναν 12-core server (Dobin et al., 2013).

Για την εγκατάσταση του λογισμικού STAR στα Ubuntu δίνονται οι εντολές:

$ sudo apt-get update

$ sudo apt-get install g++

$ sudo apt-get install make

Ακολουθεί η δημιουργία genome index files όπου ο χρήστης παρέχει στο σύστημα τα αρχεία FASTA που περιέχουν την αλληλουχία αναφοράς και GTF αρχείο με τα annotations. Από τα αρχεία αυτά το λογισμικό δημιουργεί genome indexes τα οποία χρησιμοποιούνται για το mapping. Οι βασικές εντολές για να γίνει αυτό το βήμα είναι:

--runThreadN NumberOfThreads

Όπου καθορίζεται ο αριθμός των threads που θα χρησιμοποιηθούν για τη δημιουργία του genome index και είναι συνήθως ο αριθμός των core που έχει ο υπολογιστής.

--runMode genomeGenerate

Όπου κατευθύνει το STAR να «τρέξει» τη δημιουργία του genome indices.

--genomeDir /path/to/genomeDir

Όπου καθορίζεται το directory στο οποίο είναι αποθηκευμένο το genome indices. Το ίδιο directory path χρησιμοποιείται και στο βήμα του mapping για να ταυτοποιηθεί το γονιδίωμα αναφοράς.

--genomeFastaFiles /path/to/genome/fasta1 /path/to/genome/fasta2 ...

Όπου καθορίζει τα αρχεία FASTA που περιέχουν την αλληλουχία του γονιδιώματος αναφοράς.

--sjdbGTFfile /path/to/annotations.gtf

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Όπου καθορίζεται το path για το αρχείο GTF που περιέχονται τα annotated μετάγραφα, όταν φυσικά αυτά είναι διαθέσιμα.

--sjdbOverhang ReadLength-1

Όπου καθορίζεται το μήκος της αλληλουχίας γύρω από τα annotated junctions που θα χρησιμοποιηθούν για την κατασκευή του splice junction database. Το ReadLength είναι το μήκος των reads, όμως όταν τα μήκη των reads διαφέρουν χρησιμοποιείται by default ο αριθμός 100.

Τα Genome files που δημιουργούνται είναι αρχεία που περιέχουν πληροφορίες για την αλληλουχία, τα suffix arrays, τα χρωμοσώματα, τις συντεταγμένες των splice junctions και πληροφορίες για τα γονίδια.

Αφού δημιουργηθούν τα αρχεία αυτά, που προτείνεται να μην επεξεργάζονται από το χρήστη, ακολουθεί το βήμα του mapping που πραγματοποιείται με τις εντολές

--runThreadN NumberOfThreads

--genomeDir /path/to/genomeDir

Όπου καθορίζεται το path για το directory στο οποίο βρίσκεται το genome indices που δημιουργείται παραπάνω.

--readFilesIn /path/to/read1 [/path/to/read2]

Όπου καθορίζεται το path για τα αρχεία FASTQ που περιέχουν τις αλληλουχίες που θα γίνει το mapping. Για pair-end sequencing χρησιμοποιούνται και τα δύο αρχεία FASTQ (R1, R2) που δημιουργήθηκαν ενώ για single-end sequencing χρησιμοποιείται το μοναδικό αρχείο που έχει δημιουργηθεί.

Μετά το βήμα αυτό το STAR παράγει μία σειρά αρχείων στο τρέχων directory. Το directory που θα βρεθούν όλα τα αρχεία που θα δημιουργηθούν σαν output αλλάζει με την εντολή:

--outFileNamePrefix /path/to/output/dir/prefix

Τα είδη αρχείων που προκύπτουν είναι αρχεία log, SAM, BAM και Splice junctions. Το αρχείο log.out είναι το βασικό αρχείο που περιέχει πληροφορίες για το run, το log.progress.out παρουσιάζει στατιστικά στοιχεία για τις διεργασίες που πραγματοποιούνται και το log.final.out παρουσιάζει τα στατιστικά στοιχεία για το

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

mapping που πραγματοποιήθηκε και χρησιμοποιείται για τον ποιοτικό έλεγχο. Το aligned.out.sam περιέχει τα alignments σε μορφή SAM. Με την εντολή:

--outSAMtype BAM Unsorted

το STAR βγάζει σαν output τα alignments σε μορφή BAM και ειδικότερα unsorted, σε αρχεία τις μορφής aligned.out.bam. Με την εντολή:

--outSAMtype BAM SortedByCoordinate
δημιουργούνται αρχεία aligned.sortedByCoord.out.bam στα οποία τα αρχεία αυτά κατηγοριοποιούνται με βάση τις συντεταγμένες, μία διαδικασίες που είναι πολλές φορές απαραίτητη για περαιτέρω αναλύσεις. Με την εντολή:

--outSAMtype BAM Unsorted SortedByCoordinate
δημιουργούνται και κατηγοριοποιημένα (sorted) και μη κατηγοριοποιημένα (unsorted) αρχεία.

Σαν output υπάρχουν και τα αρχεία SJ.out.tab που περιέχουν πληροφορίες για το high confidence collapsed splice junctions σε μορφή tab-delimited. Το περιεχόμενο των στηλών αυτών των αρχείων περιγράφεται στην εικόνα 36 (Dobin, 2019).

column 1: chromosome

column 2: first base of the intron (1-based)

column 3: last base of the intron (1-based)

column 4: strand (0: undefined, 1: +, 2: -)

column 5: intron motif: 0: non-canonical; 1: GT/AG, 2: CT/AC, 3: GC/AG, 4: CT/GC, 5: AT/AC, 6: GT/AT

column 6: 0: unannotated, 1: annotated (only if splice junctions database is used)

column 7: number of uniquely mapping reads crossing the junction

column 8: number of multi-mapping reads crossing the junction

column 9: maximum spliced alignment overhang

Εικόνα 36: Περιεχόμενο των στηλών των αρχείων SJ.out.tab

Το output του STAR μπορεί να είναι και ένα αρχείο Aligned.toTranscriptome.out.bam όπου τα alignments θα έχουν μεταφραστεί σε συντεταγμένες μετάγραφων (transcript coordinates) με την εντολή:

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

--quantMode TranscriptomeSAM

Η μορφή αυτού του output είναι χρήσιμη όταν χρησιμοποιούνται λογισμικά για ποσοτικοποίηση των reads που απαιτούν το mapping να γίνεται βάσει του μεταγραφώματος, όπως το RSEM.

Με την εντολή:

--quantMode GeneCounts

γίνεται υπολογισμός του αριθμού των reads ανά γονίδιο ενώ γίνεται ταυτόχρονα το mapping ενώ είναι απαραίτητη η χρήση του αρχείου με το annotation με την εντολή –*sjdbGTFfile* που αναφέρθηκε παραπάνω. Δημιουργείται έτσι το αρχείο ReadsPerGene.out.tab που περιέχει τις πληροφορίες που φαίνονται στην παρακάτω εικόνα (Εικόνα 37) (Dobin, 2019).

column 1: gene ID
column 2: counts for unstranded RNA-seq
column 3: counts for the 1st read strand aligned with RNA (htseq-count option -s yes)
column 4: counts for the 2nd read strand aligned with RNA (htseq-count option -s reverse)

Εικόνα 37: Πληροφορίες για τις στήλες του αρχείου ReadPerGene.out.tab

Με την εντολή:

--quantMode TranscriptomeSAM GeneCounts

λαμβάνονται και τα δύο αρχεία Aligned.toTranscriptome.out.bam και ReadsPerGene.out.tab που αναφέρθηκαν παραπάνω.

Τα παραπάνω αποτελούν τις βασικές εντολές του STAR και χρησιμοποιούνται για πληθώρα πειραμάτων. Παρόλα αυτά υπάρχουν και πολλές ακόμα εντολές που διαφοροποιούν την ανάλυση ανάλογα με το σκοπό του πειράματος και τις αναλύσεις που γίνονται στη συνέχεια. Οι παραπάνω εντολές και εντολές που κάνουν την ανάλυση πιο σύνθετη προέρχονται από το STAR manual 2.7.0a που ανανεώθηκε από τον Alexander Dobin στις 23 Ιανουαρίου 2019 (Dobin, 2019).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

featureCounts

Το πρόγραμμα featureCounts δημιουργήθηκε το 2014 από τον Yang Liao και τους συνεργάτες του (Yang Liao et al., 2019; Shi & Liao, 2021). Το πρόγραμμα αυτό είναι διαθέσιμο για χρήση μέσω του command line των Unix και μέσω του πακέτου Rsubread της R και είναι γραμμένο σε γλώσσα προγραμματισμού C (Y. Liao et al., 2014). Χρησιμοποιείται για τον υπολογισμό του αριθμού των reads που προέρχονται από gDNA-Seq και RNA-Seq και εφαρμόζει στρατηγικές όπως το feature blocking και το χρωμοσωμικό hashing για να είναι αποτελεσματικό εργαλείο (Y. Liao et al., 2014; Shi & Liao, 2021).

Το input του προγράμματος αυτού είναι αρχεία SAM ή BAM που περιέχουν πληροφορίες για τα aligned reads και πιο συγκεκριμένα για κάθε read δίνονται πληροφορίες για το όνομα του χρωμοσώματος ή του κωδικονίου αναφοράς στο οποίο έχει χαρτογραφηθεί, τη θέση έναρξής του και για το ακριβές alignment (Y. Liao et al., 2014). Ακόμα σαν input χρειάζεται και μία λίστα από genomic features σε μορφή GFF, GTF ή SAF (Simplified Annotation Format) (Y. Liao et al., 2014; Shi & Liao, 2021). Η μορφή SAF είναι η πιο απλή μορφή αρχείου και δίνει τις απαραίτητες πληροφορίες για τον υπολογισμό του αριθμού των reads καθώς περιέχει πέντε στήλες που περιέχουν το feature identifier, το χρωμοσωμικό όνομα, τη θέση έναρξης, τη θέση τερματισμού και την αλυσίδα (strand) του κάθε feature (Y. Liao et al., 2014). Η μορφή του αρχείου αυτού παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 38) (Shi & Liao, 2021) .

```
GeneID Chr Start End Strand
497097 chr1 3204563 3207049 -
497097 chr1 3411783 3411982 -
497097 chr1 3660633 3661579 -
100503874 chr1 3637390 3640590 -
100503874 chr1 3648928 3648985 -
100038431 chr1 3670236 3671869 -
```

Εικόνα 38: Μορφή αρχείου Simplified Annotation Format (SAF)

Το πρόγραμμα χρησιμοποιείται και για strand-specific υπολογισμό του αριθμού των reads όταν ο σχεδιασμός του πειράματος το υποστηρίζει (Y. Liao et al., 2014). Όταν τα reads προέρχονται από pair-end sequencing, κάθε ζεύγος reads καθορίζει ένα RNA fragment και έτσι το πρόγραμμα μετράει τα fragments, ενώ όταν προέρχονται από single-end sequencing τα reads ταξινομούνται με βάση το όνομα (Y. Liao et al., 2014).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Κάθε feature θεωρείται ως ένα εύρος θέσεων σε μία από τις αλληλουχίες αναφοράς ενώ κάθε meta-feature ορίζεται ως ένα σύνολο από features που αναπαριστούν μία βιολογική δομή ενδιαφέροντος (Y. Liao et al., 2014). Τα features που έχουν το ίδιο identifier ανήκουν στο ίδιο meta-feature και το πρόγραμμα δίνει δεδομένα για τα επίπεδα και των δύο (Y. Liao et al., 2014). Σε ένα πείραμα τα features μπορεί να αντιστοιχούν στα εξώνια ενώ τα meta-features στα γονίδια που ανήκουν τα εξώνια (Y. Liao et al., 2014).

Οι εντολές που χρησιμοποιούνται είναι οι παρακάτω, όταν τα αποτελέσματα από το mapping δίνονται σε μορφή BAM ή SAM, το αρχείο με το annotation είναι σε μορφή GTF και χρησιμοποιείται το featureCounts του Rsubread (Shi & Liao, 2021).

Αρχικά γίνεται φόρτωση της βιβλιοθήκης Rsubread με την εντολή:

library(Rsubread)

Για τον υπολογισμό των single-end reads, όταν χρησιμοποιείται ένα αρχείο GTF με το annotation, από το χρήστη δίνεται η εντολή:

featureCounts(files="mapping_results_SE.bam",annot.ext="annotation.gtf",

isGTFAnnotationFile=TRUE,GTF.featureType="exon",GTF.attrType="gene_id")

Για τον υπολογισμό των single-end reads, από ένα αρχείο BAM, χρησιμοποιείται η εντολή:

featureCounts(files="mapping_results_SE.bam")

Για τον υπολογισμό των fragments αντί των reads σε ένα αρχείο από pair-end sequencing χρησιμοποιείται η εντολή:

featureCounts(files="mapping_results_PE.bam",isPairedEnd=TRUE)

ενώ για τον υπολογισμό των fragments που πληρούν κάποια συγκεκριμένα κριτήρια π.χ. μήκος μεγαλύτερο από 50 βάσεις και μικρότερο από 600 βάσεις χρησιμοποιείται η παρακάτω εντολή.

featureCounts(files="mapping_results_PE.bam",isPairedEnd=TRUE,checkFragLength=TRUE,

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

minFragLength=50,maxFragLength=600)

Περισσότερες πληροφορίες σχετικά με τις διαφορετικές εντολές και χρήσεις του featureCounts δίνονται στην επίσημη ιστοσελίδα του RDocumentation (*FeatureCounts Function*, n.d.).

Το featureCounts δημιουργεί ένα matrix που είναι κατάλληλο για την περαιτέρω μελέτη των δεδομένων με τη βοήθεια άλλων εργαλείων όπως το limma, edgeR, DESeq2 για την ανάλυση της διαφορετικής έκφρασης των γονιδίων και το DEXSeq για την ανάλυση του εναλλακτικού ματίσματος (Yang Liao et al., 2019). Ακόμα βγάζει σαν output δεδομένα για τη θέση και το μήκος κάθε feature και κάθε read κάνοντας δυνατό τον υπολογισμό μεγεθών όπως το RPKM (Yang Liao et al., 2019). Γενικά όταν το featureCounts χρησιμοποιείται σαν πρόγραμμα του πακέτου Subread βγάζει ένα αρχείο που τον αριθμό των reads και ένα αρχείο που περιέχει το summary των αποτελεσμάτων του counting (Shi & Liao, 2021) . Όταν όμως χρησιμοποιείται σαν πρόγραμμα του Rsubread σαν output δημιουργείται ένα αντικείμενο του R (R 'List' object), ένα matrix που είναι έτοιμο για downstream ανάλυση των δεδομένων (Yang Liao et al., 2019; Shi & Liao, 2021).

Συνεπώς το featureCounts είναι μία αποτελεσματική και γρήγορη μέθοδος λόγω του αλγορίθμου που χρησιμοποιείται έναντι άλλων μεθόδων (Y. Liao et al., 2014). Δίνει επιπλέον στο χρήστη μεγάλη ευελιξία στο χειρισμό των δεδομένων της αλληλούχησης επόμενης γενιάς (Y. Liao et al., 2014). Ειδικότερα η χρήση του featureCounts από το περιβάλλον του R το μετατρέπει σε ένα ιδιαίτερα χρήσιμο εργαλείο όταν πρόκειται το αποτέλεσμα των read count να χρησιμοποιηθεί σε άλλα πακέτα του R όπως το limma και το edgeR (Y. Liao et al., 2014).

Empirical Analysis of Digital Gene Expression Data in R (edgeR)

Το edgeR χρησιμοποιείται για την ανάλυση της διαφορετικής έκφρασης των δεδομένων και αποτελεί ένα πακέτο από το λογισμικό του Bioconductor (Robinson et al., 2010). Χρησιμοποιεί στατιστικές μεθόδους που έχουν δημιουργηθεί από τους Robinson, Smyth, McCarthy, Lund, Chen και Lun και ο βασικός του στόχος είναι η ανάλυση της έκφρασης των δεδομένων που βασίζεται στα replicated counts (Chen et al., 2008; Robinson et al., 2010). Με τη βοήθεια του πακέτου αυτού γίνεται ανάλυση

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

της διαφορετικής έκφρασης τόσο σε επίπεδο γονιδίων όσο και σε επίπεδο εξωνίων, αλλά και μετάγραφων (Chen et al., 2008). Αναλύσεις σε επίπεδο εξωνίων οδηγούν και σε μελέτες της διαφορετικής έκφρασης των ισομορφών που προκύπτουν λόγω του εναλλακτικού ματίσματος που συμβαίνει στους ευκαρυωτικούς οργανισμούς (Chen et al., 2008). Ειδικότερα, βρίσκει διαφορές μεταξύ των groups όταν ένα τουλάχιστον από αυτά έχει replicated measurements (Robinson et al., 2010). Το edgeR χρησιμοποιεί τα raw data, στα οποία δεν έχει γίνει κανονικοποίηση, για να αρχίσει η ανάλυση των δεδομένων (Chatterjee et al., 2018). Για να γίνει κανονικοποίηση για το sequencing depth μετατρέπει τα counts σε CPM και υπολογίζει το composition bias βάσει του TMM (trimmed mean of M-values) (Chatterjee et al., 2018).

Οι μέθοδοι classic edgeR και glm edgeR

Υπάρχουν πολλές στατιστικές μέθοδοι που χρησιμοποιούνται στο πακέτο edgeR όπως η μέθοδος classic edgeR και η μέθοδος glm edgeR (Chen et al., 2008). Οι μέθοδοι αυτοί είναι συμπληρωματικές και πολλές φορές χρησιμοποιούνται συνδυαστικά για την ανάλυση των δεδομένων που προκύπτουν από την αλληλούχηση επόμενης γενιάς (Chen et al., 2008). Βασικότερη για τη στατιστική ανάλυση με τη μέθοδο glm edgeR είναι η μέθοδος quasi-likelihood F-test, που βρίσκεται κάτω από το πρίσμα του glm edgeR και προτιμάται όταν ο αριθμός των δειγμάτων είναι μικρός καθώς έχει higher error rate control (Chatterjee et al., 2018; Chen et al., 2008). Αντίθετα το likelihood ratio test προτιμάται για την ανάλυση δεδομένων από το single cell RNA-Seq και γενικότερα δεδομένων που δεν περιέχουν replicates (Chen et al., 2008).

Απαραίτητα αρχεία

Για την ανάλυση, με τη βοήθεια του quasi-likelihood F-test, είναι απαραίτητος ένας πίνακας που περιέχει τον αριθμό των read counts, ένας vector με το μέγεθος της βιβλιοθήκης και ένας factor που ορίζει το group που ανήκει κάθε δείγμα (Robinson et al., 2010). Στον πίνακα που περιέχει τα read counts οι γραμμές αναπαριστούν τα γονίδια ενώ οι στήλες αντιστοιχούν στις βιβλιοθήκες και τέτοιοι πίνακες δημιουργούνται με προγράμματα όπως το featureCounts που αναφέρθηκε παραπάνω και αποτελεί λειτουργία του πακέτου Rsubread (Chen et al., 2008).

Ανάλυση με τη μέθοδο quasi-likelihood F-test

Για να γίνει η ανάλυση της διαφορετικής έκφρασης των γονιδίων δημιουργείτε ένα αντικείμενο (object) DGEList, όπου περιέχονται τα raw counts (Chatterjee et al., 2018).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Το αντικείμενο αυτό είναι ένα απλό list-based data object στο οποίο αποθηκεύονται τα δεδομένα που χρησιμοποιούνται για το πακέτο edgeR και μπορεί να τροποποιηθεί σαν οποιοδήποτε αντικείμενο της ίδιας μορφής από την R (Chen et al., 2008).

>library(edgeR)

>xE <- DGEList(counts= assay(se), group=dds$Status)

>colnames(xE)<-sample_table$SampleName

>xE

An object of class "DGEList"

$counts

cntrl1 cntrl2 cntrl3 trtmnt1 trtmnt2 trtmnt3

TSPAN6 88 142 93 84 94 51

TNMD 0 0 0 0 0 0

DPM1 249 329 258 327 221 133

22752 more rows ...

$samples

group lib.size norm.factors

cntrl1 control 3299005 1

cntrl2 control 5075669 1

cntrl3 control 4592966 1

trtmnt1 treatment 5094611 1

trtmnt2 treatment 4437821 1

trtmnt3 treatment 3051320 1

Το αντικείμενο αυτό, όπως φαίνεται και παραπάνω αποτελείται από ένα matrix *counts* που περιέχει τα counts και από ένα *samples* που περιέχει πληροφορίες για τα δείγματα ή τη βιβλιοθήκη (Chen et al., 2008). Ειδικότερα στο *samples* εμφανίζεται η στήλη lib.size που δίνει πληροφορίες σχετικά με το μέγεθος της βιβλιοθήκης ή το βάθος της

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

αλληλούχησης (Chen et al., 2008). Προαιρετικά υπάρχει και η δυνατότητα δημιουργίας ενός *genes* που περιέχει πληροφορίες σχετικά με το annotation των γονιδίων (Chen et al., 2008).

Στη συνέχεια απομακρύνονται τα γονίδια που έχουν χαμηλό αριθμό counts (στο συγκεκριμένο παράδειγμα αριθμό counts μικρότερο του 10, CPM<10) και έτσι γίνεται καλύτερη ανίχνευση των γονιδίων που παρουσιάζουν σημαντικά διαφορετική έκφραση μεταξύ των δειγμάτων (Chatterjee et al., 2018). Αυτό συμβαίνει διότι γονίδια με μικρό αριθμό counts (συνήθως μεταξύ 5-10) δεν προσφέρουν επαρκή στοιχεία σχετικά με τη διαφορετική έκφραση (Chen et al., 2008). Η απομάκρυνση αυτώ των γονιδίων γίνεται με τη βοήθεια του CPM (counts-per-million) καθώς λαμβάνονται υπόψιν οι διαφορές του μεγέθους των βιβλιοθηκών μεταξύ των δειγμάτων (Chen et al., 2008).

>cpm(10, mean(xE$samples$lib.size)) #This calculates the CPM

value that corresponds to a count of 10.

[,1]

[1,] 2.35

>keep <- rowSums(cpm(xE) > 2.35) >= 3 #Keeping only the genes

with a CPM value of 2.35 in 3 or more samples.

>table(keep) #Out of 22,757 genes, 11,954 genes are filtered

out.

keep

FALSE TRUE

11954 10803

>xE<- xE[keep, , keep.lib.sizes=FALSE] #The keep.lib.sizes is

set to FALSE so that the library sizes are recalculated after

the filtering.

>dim(xE)

[1] 10803 6

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Ακολουθεί η κανονικοποίηση που συμβαίνει για να απομακρυνθούν τυχόν τεχνικά σφάλματα που προκύπτουν κατά την ανάλυση και δεν σχετίζονται με τη διαφορετική έκφραση των γονιδίων λόγω των διαφορετικών συνθηκών κάτω από τις οποίες πραγματοποιείται το πείραμα (Chen et al., 2008). Η κανονικοποίηση δεν γίνεται βάσει του μεγέθους του κάθε γονιδίου, καθώς το μήκος του κάθε γονιδίου έχει την ίδια επίδραση στον αριθμό των reads για κάθε δείγμα (Chen et al., 2008). Το βάθος της αλληλούχησης αποτελεί ένα παράγοντα που επηρεάζει την ανάλυση όμως δεν γίνεται κανονικοποίηση βάσει αυτού, διότι οι διαφορές που δημιουργούνται λύνονται από το σύστημα χωρίς να χρειάζεται κάποια επεξεργασία των δεδομένων από το χρήστη (Chen et al., 2008). Βασικό παράγοντα για την ανάλυση της έκφρασης αποτελεί το μέγεθος της βιβλιοθήκης (Chen et al., 2008). Η βασική αρχή λειτουργίας του RNA-Seq είναι ο υπολογισμός της σχετικής παρουσίας ενός γονιδίου σε κάθε δείγμα και όχι ο υπολογισμός του συνολικού RNA για κάθε κυττάρου (Chen et al., 2008). Αυτό δημιουργεί πρόβλημα όταν ένας μικρός αριθμός από γονίδια εκφράζεται σε υψηλά επίπεδα σε ένα δείγμα ενώ σε άλλα όχι (Chen et al., 2008). Τα γονίδια αυτά καταλαμβάνουν έτσι ένα μεγάλο ποσοστό της βιβλιοθήκης σε αυτά τα δείγματα και οδηγούν σε ψευδή συμπεράσματα ότι τα υπόλοιπα γονίδια υπο-εκφράζονται στα δείγματα αυτά (Chen et al., 2008). Συνεπώς γίνεται κανονικοποίηση βάσει του μεγέθους της βιβλιοθήκης με τη μέθοδο TMM, που προτιμάται για δεδομένα από RNA-Seq, με τις εντολές που παρουσιάζονται παρακάτω (Chatterjee et al., 2018; Chen et al., 2008). Η μέθοδος αυτή υπολογίζει ένα παράγοντα κανονικοποίησης για κάθε δείγμα και το χρησιμοποιεί για να επανυπολογίσει το μέγεθος της βιβλιοθήκης και να ελέγξει το composition bias που προκύπτει από την ανάλυση.

>xE<- calcNormFactors(xE,method="TMM")

>xE$samples

group lib.size norm.factors

control1 control 3288149 1.142

control2 control 5059628 1.061

control3 control 4580364 0.909

treatment1 treatment 5079636 0.977

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

treatment2 treatment 4424873 0.949

treatment3 treatment 3042193 0.979

Τα δείγματα του RNA μπορούν να κατηγοριοποιηθούν σε δύο διαστάσεις χρησιμοποιώντας τα γραφήματα multi-dimensional scaling (MDS) (Chen et al., 2016). Ένα παράδειγμα εντολών μέσω των οποίων πραγματοποιούνται τα γραφήματα αυτά είναι οι παρακάτω (Chen et al., 2016):

> pch <- c(0,1,2,15,16,17)

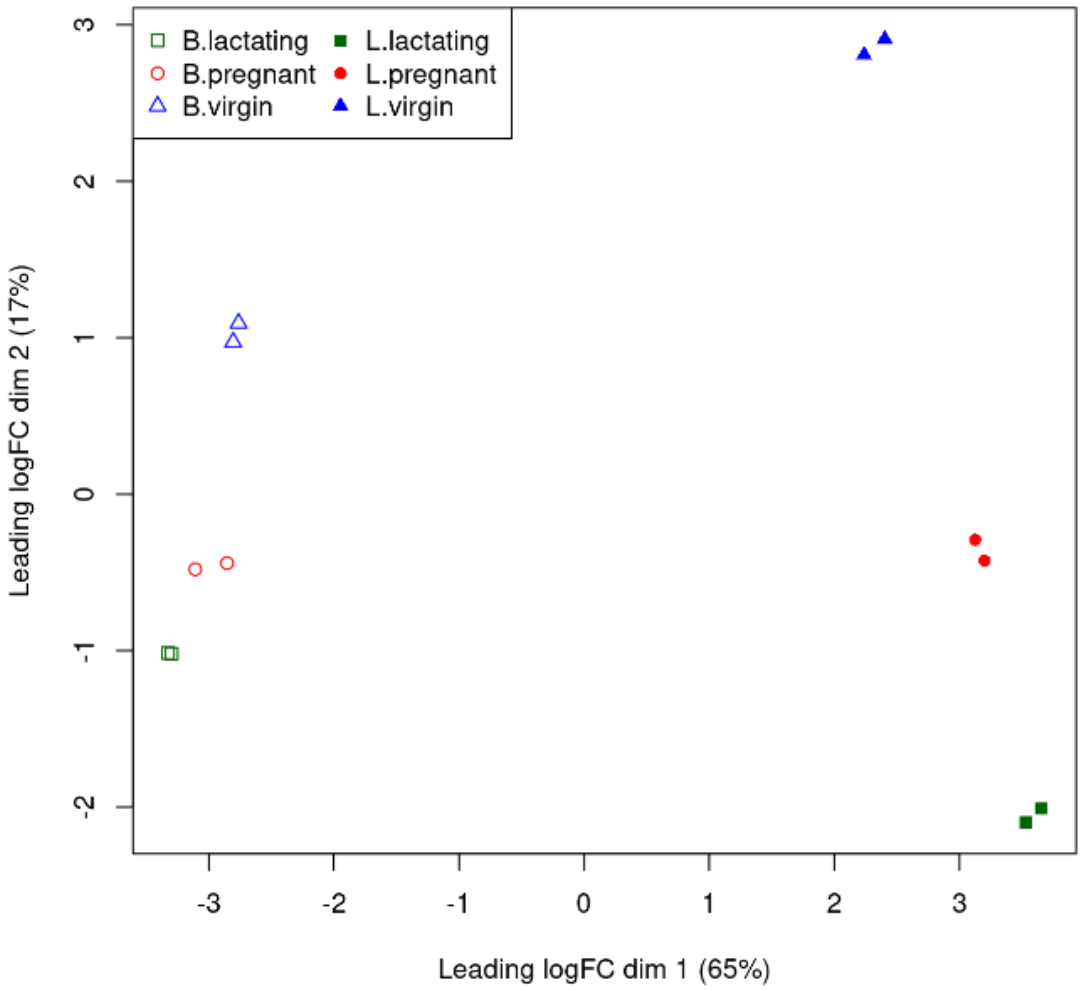
> colors <- rep(c("darkgreen", "red", "blue"), 2)

> plotMDS(y, col=colors[group], pch=pch[group])

> legend("topleft", legend=levels(group), pch=pch, col=colors, ncol=2)

Μέσω αυτών δημιουργούνται γραφήματα της μορφής που φαίνεται στην παρακάτω εικόνα (Εικόνα 39) (Chen et al., 2016).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 39: MDS plot που παρουσιάζει τις αποστάσεις μεταξύ των προφίλ έκφρασης των δειγμάτων

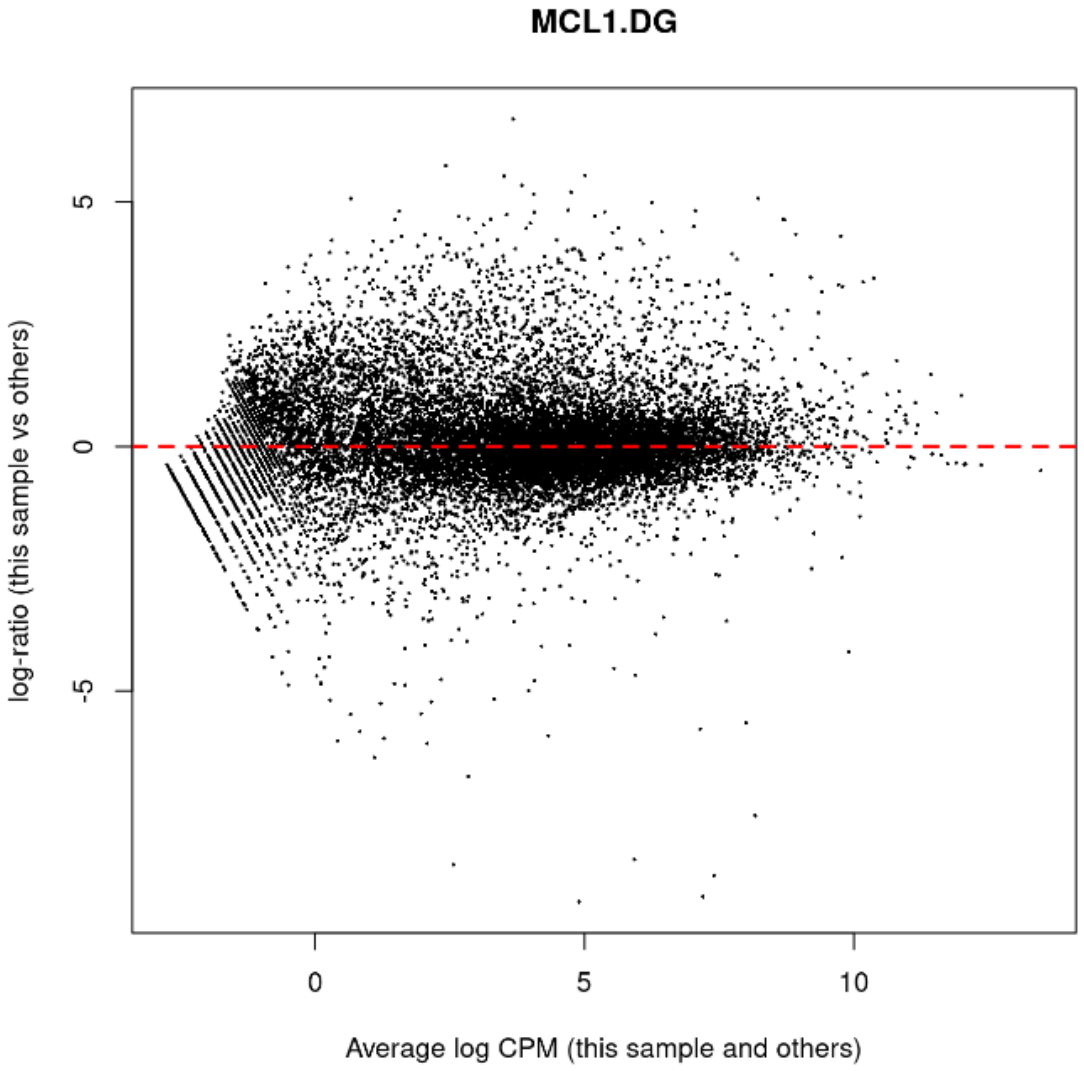
Ο τύπος αυτός του γραφήματος αποτελεί βήμα ανάλυσης και ποιοτικού ελέγχου για τη μελέτη της διαφορετικής έκφρασης μεταξύ των δειγμάτων και όπως φαίνεται και στο παράδειγμα τα δείγματα που προέρχονται από το ίδιο group βρίσκονται κοντά μεταξύ τους σε σχέση με τα δείγματα από τα διαφορετικά groups που είναι απομακρυσμένα (Chen et al., 2016).

Για ένα άλλο σετ δεδομένων, με τις παρακάτω εντολές, δημιουργείται το γράφημα της εικόνας 40 (Chen et al., 2016):

> plotMD(y, column=1)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

> abline(h=0, col="red", lty=2, lwd=2)

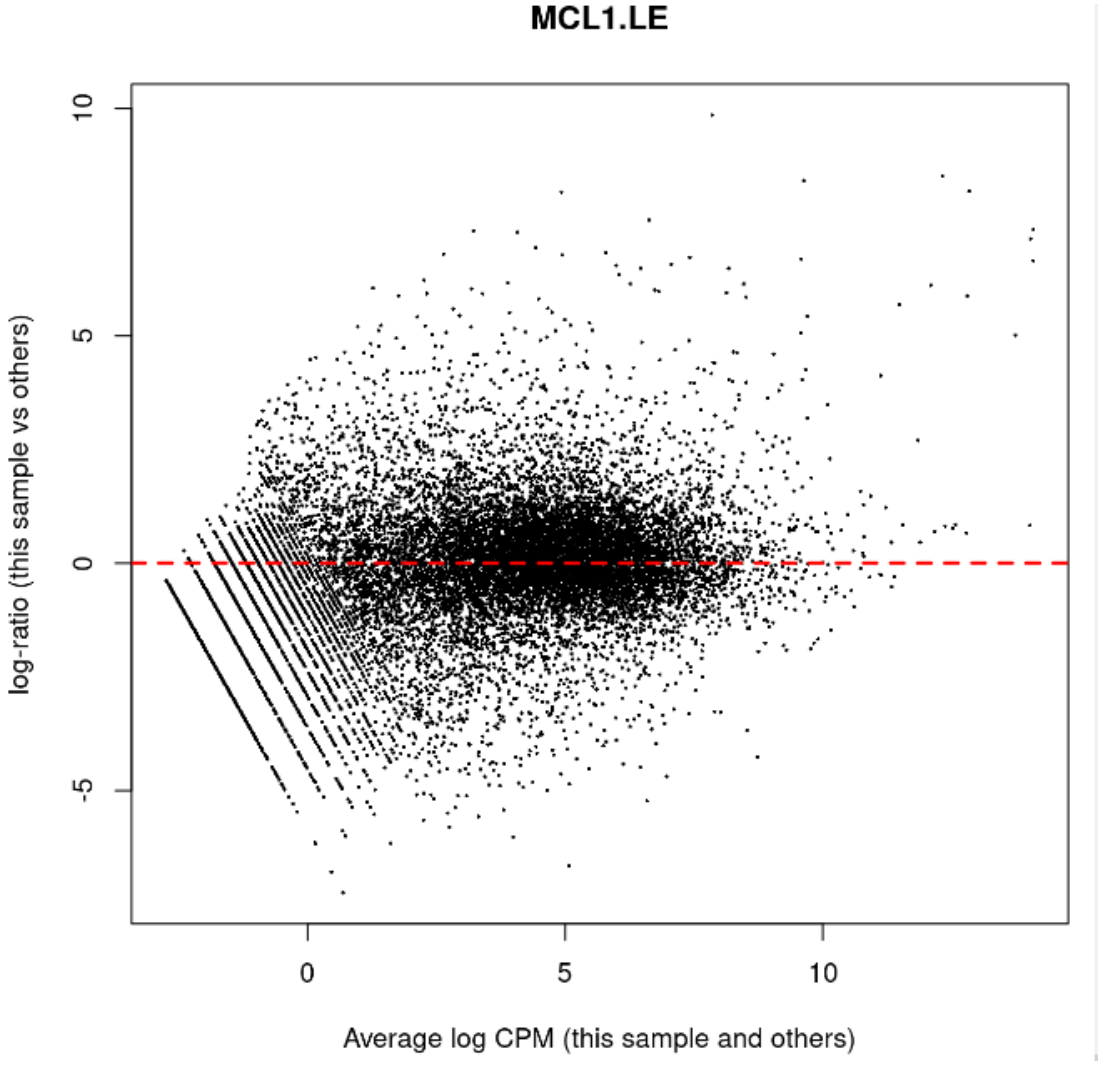


Εικόνα 40: MD plot

Μέσω του γραφήματος αυτού (Εικόνα 40) δίνονται πληροφορίες για το προφίλ έκφρασης μεμονωμένων δειγμάτων. Πιο συγκεκριμένα, παρουσιάζει τις διαφορές στο μέγεθος των βιβλιοθηκών μεταξύ δύο βιβλιοθηκών σε σχέση με το μέσο όρο αυτών των βιβλιοθηκών. Δηλαδή στο παραπάνω γράφημα συγκρίνεται το δείγμα 1 από τα δεδομένα με μία βιβλιοθήκη αναφοράς που έχει δημιουργηθεί από το μέσο όρο όλων των υπόλοιπων δειγμάτων. Στην εικόνα 40 λοιπόν φαίνεται τα περισσότερα γονίδια που αναπαρίστανται με τις μαύρες κουκίδες να βρίσκονται κοντά στην κόκκινη γραμμή (Chen et al., 2016).

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Αντίθετα στην εικόνα 41 υπάρχουν αρκετές μαύρες κουκίδες στο πάνω και δεξιά τμήμα του γραφήματος που αναπαριστούν γονίδια που υπερεκφράζονται στο συγκεκριμένο δείγμα (Chen et al., 2016).



Εικόνα 41: MD plot

Το edgeR υπολογίζει για κάθε ξεχωριστό γονίδιο την empirical Bayes moderated διασπορά (Chen et al., 2016). Χρησιμοποιεί ακόμα την negative binomial κατανομή για να διαμορφώσει για κάθε δείγμα τον αριθμό των reads για κάθε γονίδιο (Chen et al., 2016). Για την εκτίμηση της διασποράς χρησιμοποιείται η εντολή estimateDisp() και ειδικότερα (Chatterjee et al., 2018):

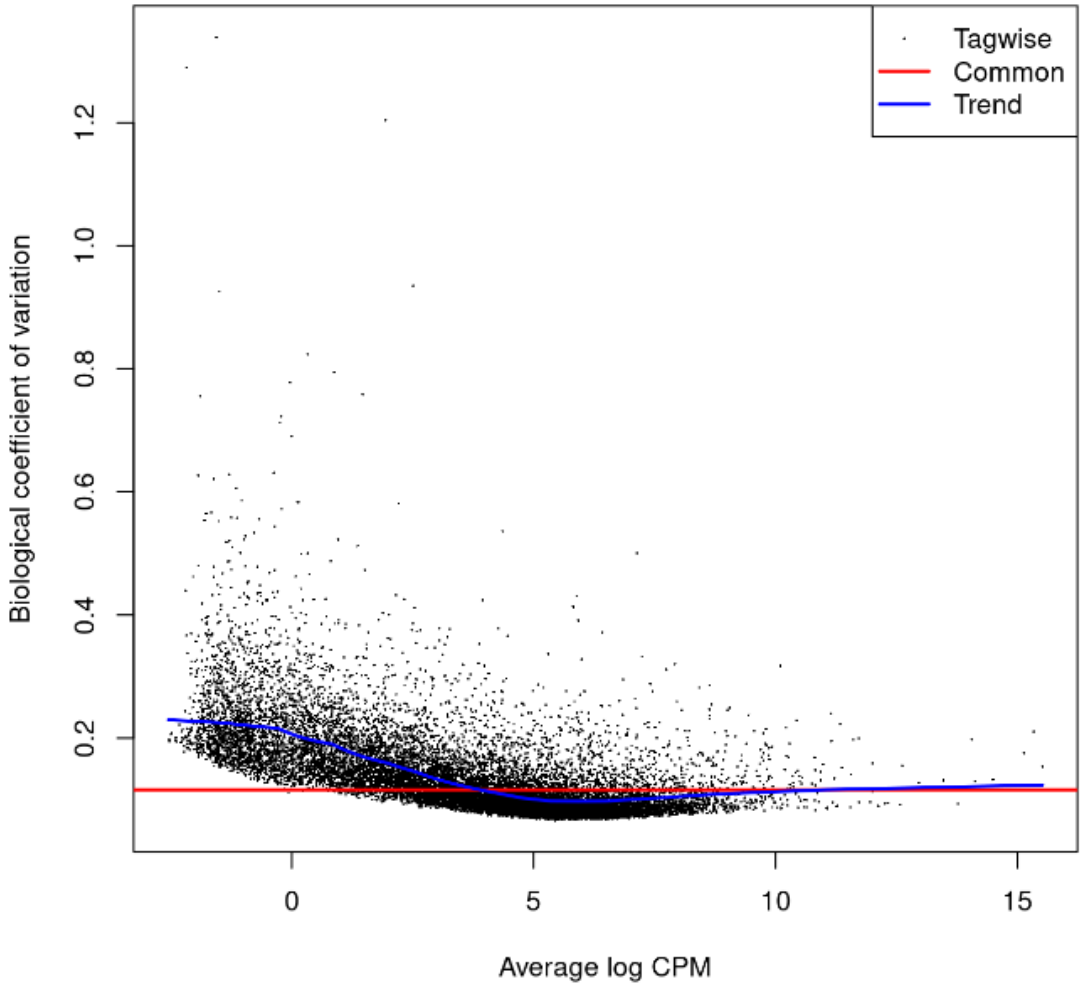
>xE <- estimateDisp(xE, design, robust=TRUE)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Με την εντολή plotBCV παρουσιάζονται σε ένα γράφημα ο συντελεστής διακύμανσης (biological coefficient of variation) σε συνάρτηση με το μέσο $\log_2$CPM (Chatterjee et al., 2018).

>plotBCV(xE)

Η εντολή αυτή οδηγεί στην οπτικοποίηση της διασποράς και το γράφημα που προκύπτει έχει, ανάλογα και τα δεδομένα, την παρακάτω μορφή (Εικόνα 42) (Chen et al., 2016).



Εικόνα 42: BCV plot

Ακολουθεί το quasi-likelihood (QL) test που γίνεται με την παρακάτω εντολή.

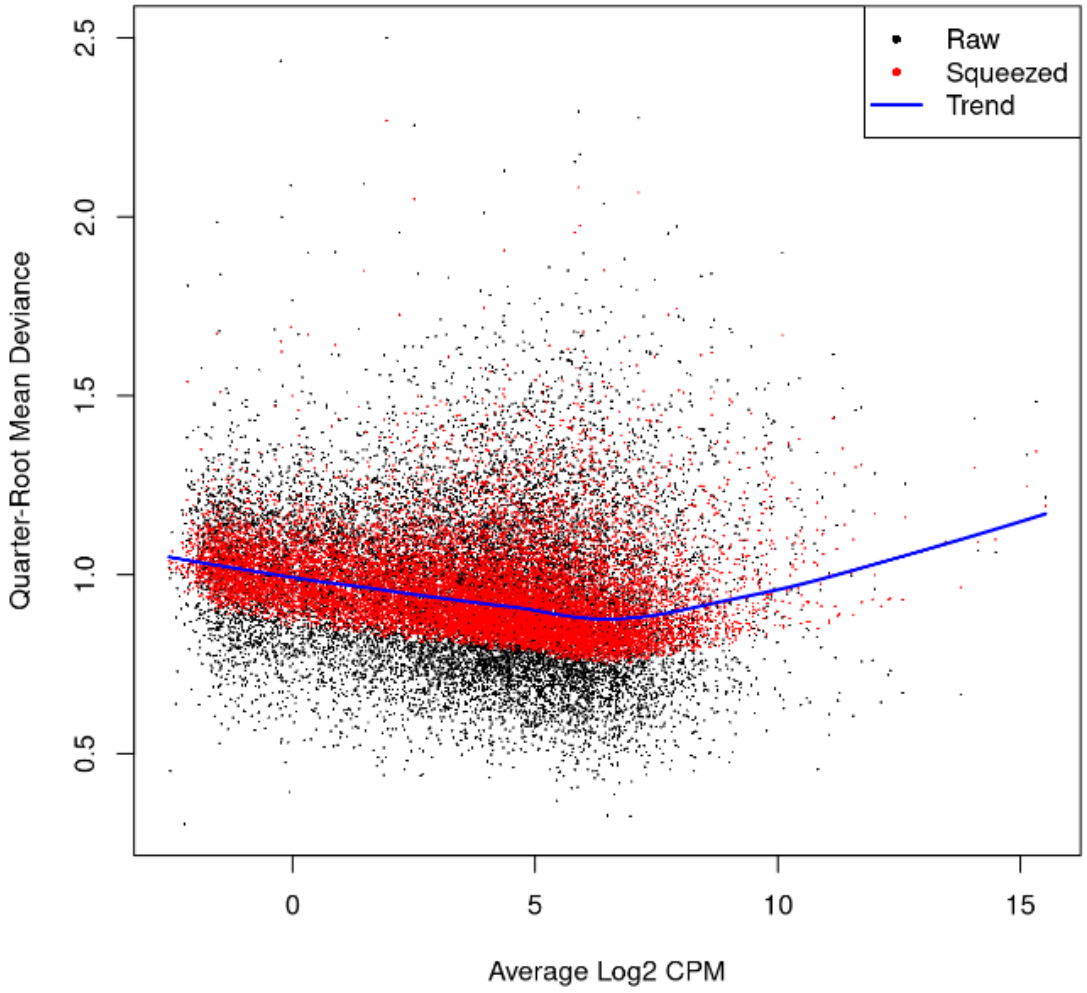
>fit <- glmQLFit(xE, design, robust=TRUE)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Η οπτικοποίηση της διασποράς που προκύπτει από το QL test γίνεται με τη χρήση της εντολής (Chatterjee et al., 2018) :

>plotQLDisp(fit)

Το γράφημα που προκύπτει από την παραπάνω εντολή έχει τη μορφή (Εικόνα 43) (Chen et al., 2016):



Εικόνα 43: QL plot

Στη συνέχεια δημιουργείτε μία contrast matrix για να καθοριστούν τα groups που πρόκειται να συγκριθούν, με τις παρακάτω εντολές (Chatterjee et al., 2018).

>contr.matrix <- makeContrasts(treatment-control, levels=design)

>contr.matrix

Contrasts

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Levels treatment - control

control -1

treatment 1

>resE <- glmQLFTest(fit, contrast=contr.matrix)

>topTags(resE)

*Coefficient: -1*control 1*treatment*

logFC logCPM F PValue FDR

IL1B 7.09 8.88 1336 3.68e-12 3.97e-08

ICAM1 5.28 8.88 1105 9.66e-12 5.22e-08

CXCL8 6.43 10.94 689 1.04e-10 3.76e-07

>is.de <- decideTestsDGE(res, p.value = 0.05,adjust.method="BH")

>summary(is.de)

Με την εντολή *makeContrasts* το θετικό logFC υποδηλώνει ότι ένα γονίδιο υπερεκφράζεται στο group των treatment σε σχέση με το control ενώ ένα αρνητικό logFC υποδηλώνει ότι ένα γονίδιο υπερεκφράζεται στο group των control (Chen et al., 2016). Ακολουθεί το QL F-test όπως είπαμε και παραπάνω καθώς έχει καλύτερο error rate control, με την εντολή *glmQLFTest* και στη συνέχεια με την εντολή *topTags(res)* εμφανίζονται τα γονίδια που εμφανίζουν διαφορετικά επίπεδα έκφρασης (Chen et al., 2016). Στο παραπάνω παράδειγμα δεν υπάρχουν γονίδια με αρνητικό logFC όμως υπάρχουν άλλα σετ δεδομένων που παρουσιάζουν και αρνητικές τιμές όπως φαίνεται παρακάτω (Chen et al., 2016).

*Coefficient: 1*B.lactating -1*B.pregnant*

Length Symbol logFC logCPM F PValue FDR

12992 765 Csn1s2b 6.08 10.19 421 4.43e-11 6.94e-07

211577 2006 Mrgprf 5.14 2.75 342 1.28e-10 7.50e-07

226101 7094 Myof 2.32 6.45 322 1.85e-10 7.50e-07

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

381290 8292 Atp2b4 2.14 6.15 320 1.91e-10 7.50e-07

140474 11281 Muc4 -7.18 6.06 309 2.42e-10 7.57e-07

231830 3346 Micall2 -2.25 5.19 283 4.12e-10 1.08e-06

24117 2242 Wif1 -1.82 6.77 261 6.77e-10 1.52e-06

12740 1812 Cldn4 -5.32 9.88 299 8.83e-10 1.65e-06

21953 667 Tnni2 5.75 3.87 313 9.50e-10 1.65e-06

231991 2873 Creb5 2.57 4.88 241 1.10e-09 1.73e-06

Με την εντολή *decideTestsDGE* επιλέγονται μόνο τα γονίδια με FDR (False Discovery Rate) 5% και με την εντολή *summary* εμφανίζεται ο αριθμός των γονιδίων που υποεκφράζονται και υπερεκφράζονται (Chen et al., 2016). Καταλήγουμε έτσι ότι στο δεύτερο παράδειγμα υπάρχουν 2738 γονίδια που υποεκφράζονται ενώ 2463 υπερεκφράζονται και 10473 δεν έχουν σημαντικές διαφορές στην έκφραση, όπως φαίνεται και παρακάτω (Chen et al., 2016).

*1*B.lactating -1*B.pregnant*

Down 2738

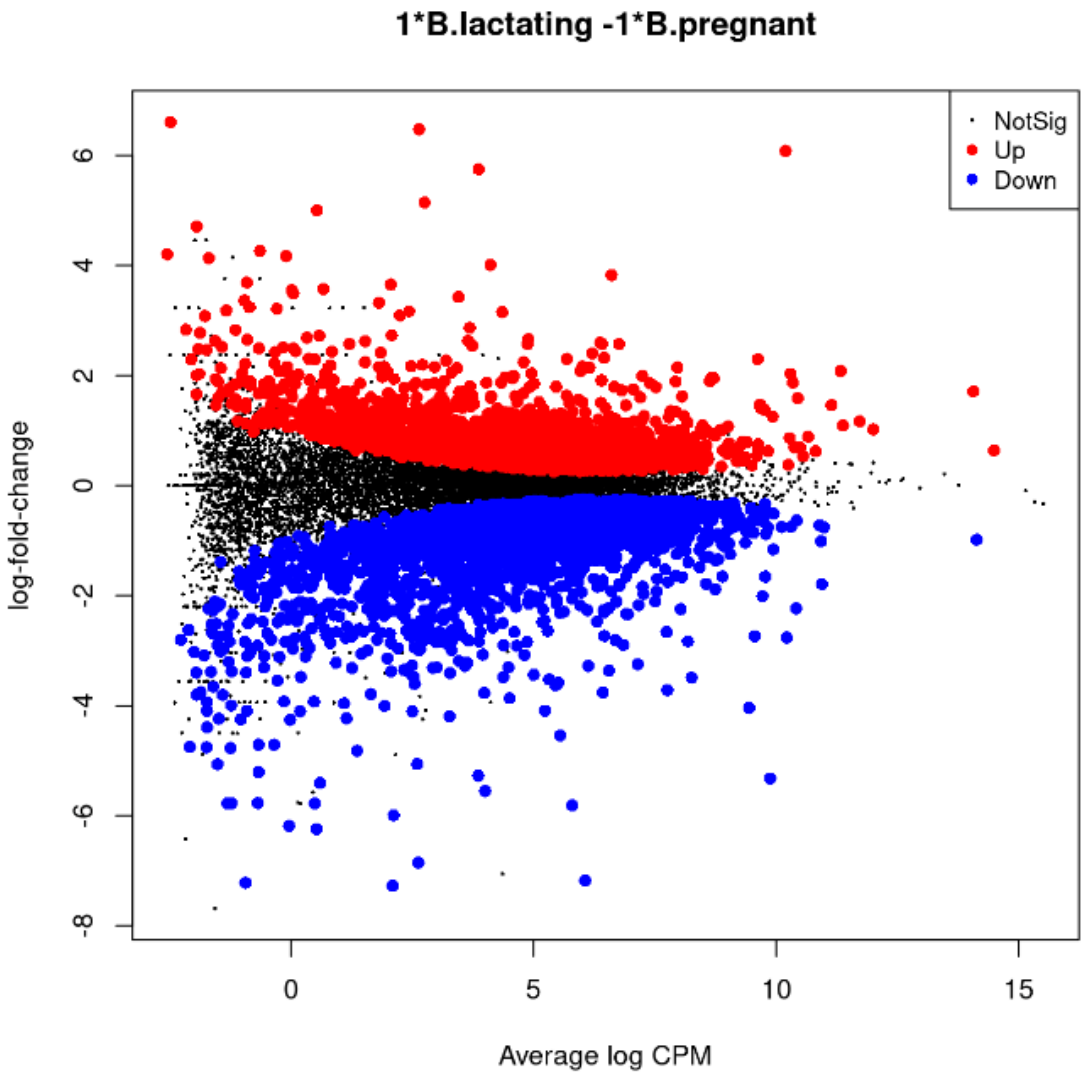
NotSig 10473

Up 2463

Για τη δημιουργία ενός γραφήματος Mean-Difference (MD) όπου παρουσιάζονται τα γονίδια που εμφανίζουν διαφορετική έκφραση χρησιμοποιείται η εντολή (Chen et al., 2016):

>plotMD(resE, status=is.de)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210



Εικόνα 44: MD plot

Στο γράφημα που δημιουργείται από αυτές τι εντολές και παρουσιάζεται παραπάνω (Εικόνα 44), εμφανίζονται τα γονίδια που υπερεκφράζονται με κόκκινο και τα γονίδια που υποεκφράζονται με μπλε (Chen et al., 2016).

Ο πίνακας που περιέχει όλες τις πληροφορίες σχετικά με τα γονίδια που εμφανίζουν διαφορετικά επίπεδα έκφρασης αποθηκεύεται με τη μορφή csv με τις εντολές:

>topEdgeR<- topTags(resE, n=Inf, p.value = 0.05, adjust.

method="BH")

>dim(topEdgeR)

[1] 374 5

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

>write.csv(topEdgeR, file="topEdgeR.csv")

Από τις τιμές του CPM δημιουργούνται ακόμα και γραφήματα όπως το heatmap στο οποίο φαίνονται τα γονίδια που έχουν βρεθεί ότι παρουσιάζουν τις μεγαλύτερες διαφορές στην έκφραση των γονιδίων. Για να γίνει αυτό χρησιμοποιούνται οι εντολές:

>cpm <- cpm(xE)

>lcpm <-cpm(xE, piror.count=2, log=TRUE)

>mat <- lcpm[rownames(tt)[1:374],]

>scaled.mat<-t(scale(t(mat)))

>library(gplots)

>col.pan <- colorpanel(100, "blue","white","red")

>heatmap.2(scaled.mat, col=col.pan, Rowv=TRUE, scale="none",

trace="none") (Chatterjee et al., 2018)

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Αποτελέσματα-Συμπεράσματα

Η διπλωματική εργασία που παρουσιάστηκε παραπάνω αποτελεί μια επισκόπηση των εργαλείων βιοπληροφορικής που χρησιμοποιούνται για την ανάλυση του μεταγραφώματος ή αλλιώς το RNA Sequencing. Αρχικά στο γενικό μέρος αναλύθηκαν τα βασικά βήματα που πραγματοποιούνται για την ανάλυση του μεταγραφώματος και έγινε παρουσίαση των βιοπληροφορικών εργαλείων που χρησιμοποιούνται ανάλογα με το ερευνητικό ερώτημα που μελετάται. Στο γενικό μέρος έγινε ανασκόπηση αυτών των βασικών εργαλείων και μέσω αυτού ο αναγνώστης μπορεί να πάρει τις βασικές γνώσεις που χρειάζεται για τον τρόπο που δουλεύουν τα προγράμματα αυτά. Η παραπάνω εργασία έτσι αποτελεί μια αρχική πηγή πληροφορίων για τα εργαλεία βιοπληροφορικής που μπορούν να χρησιμοποιηθούν για την ανάλυση των δεδομένων που προκύπτουν από την αλληλούχηση επόμενης γενιάς του RNA.

Πιο συγκεκριμένα, όσον αφορά την ανάλυση του μεταγραφώματος που αναλύθηκε σε αυτή την εργασία, το RNA-Seq αποτελεί ένα πολύτιμο εργαλείο στα χέρια των επιστημόνων. Μέσω αυτού γίνεται μελέτη των γονιδίων που εκφράζονται σε ένα κύτταρο ή πληθυσμό κυττάρων ανά πάσα χρονική στιγμή. Ειδικότερα παρέχεται η δυνατότητα ανάλυσης της διαφορετικής έκφρασης ενός γονιδίου ανάλογα με τις συνθήκες που επικρατούν ή του τύπου των κυττάρων που μελετώνται. Γίνονται αναλύσεις της διαφορετικής έκφρασης των ισομορφών ενός μετάγραφου και μελετάται έτσι το εναλλακτικό μάτισμα που συμβαίνει στους ευκαρυωτικούς οργανισμούς. Μέσω του RNA-Seq δίνονται απαντήσεις σχετικά με επιγενετικά φαινόμενα καθώς γίνεται υπολογισμός της έκφρασης των μητρικών και πατρικών αλλήλιων. Γίνεται ακόμα μελέτη της γενετικής ποικιλότητας και των διαφορών που εμφανίζονται μεταξύ υγιών και παθολογικών κυττάρων και ιστών. Ανακαλύπτονται νέοι γενετικοί τόποι που σχετίζονται με τη διαφορετική έκφραση των γονιδίων και variants που σχετίζονται με τα διαφορετικά φαινοτυπικά χαρακτηριστικά των ατόμων και των ασθενειών που αυτοί εμφανίζουν. Επίσης μελετώνται βιολογικές διαδικασίες, γενετικές αλλαγές και μοριακά μονοπάτια τα οποία είναι υπεύθυνα ή σχετίζονται με ασθένειες. Ιδιαίτερο επίσης ενδιαφέρον για το μέλλον παρουσιάζει και το single-cell RNA-Seq που φαίνεται ότι θα δώσει απαντήσεις σε ακόμα περισσότερα ερωτήματα.

Οι ανακαλύψεις αυτές είναι δυνατές μέσω των βιοπληροφορικών εργαλείων καθώς ο όγκος των δεδομένων που προκύπτουν από το RNA-Seq είναι τεράστιος. Έτσι

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

εργαλεία όπως το FASTQC που κάνει τον ποιοτικό έλεγχο των raw data του RNA-Seq, το STAR που πραγματοποιεί το alignment των reads, το featureCounts που πραγματοποιεί την ποσοτικοποίηση και το edgeR που κάνει την ανάλυση της διαφορετικής έκφρασης αποτελούν μόνο μερικά από τα προγράμματα που έχουν αναπτυχθεί. Αυτά σε συνδυασμό με άλλα προγράμματα όπως τα Cufflink, TopHat, Burrow-Wheeler Aligner (BWA), Limma Voom, HTSeq, DESeq(2) και NGSQC αποτελούν μερικά βασικά εργαλεία μέσω των οποίων οι ερευνητές προσπαθούν να δώσουν απαντήσεις στα βιολογικά ερωτήματα που κλίνονται να αντιμετωπίσουν.

Συμπερασματικά, η βιοπληροφορική αποτελεί ένα κλάδο της βιολογίας που εμφανίζει τεράστια ανάπτυξη τα τελευταία χρόνια. Μέσα στην τελευταία εικοσαετία έχουν δημιουργηθεί πληθώρα βιοπληροφορικών εργαλείων και πληθώρα ερευνητικών μεθόδων που αναμένεται να προσφέρουν πολλές απαντήσεις σε ερωτήματα που ταλανίζουν την επιστημονική κοινότητα για δεκαετίες.

Πιο συγκεκριμένα, η ανακάλυψη της αλληλούχησης επόμενης γενιάς άνοιξε το δρόμο για τη μελέτη βιολογικών διεργασιών σε επίπεδο γονιδιώματος, μεταγραφώματος, επιγενώματος, μεταβολώματος και πρωτεινώματος. Οι αναλύσεις μεγάλου όγκου δεδομένων δίνουν τη δυνατότητα ανάλυσης δεδομένων που εμφανίζουν μεγάλη ετερογένεια και ίσως αυτή η ετερογένεια αποτελεί την απάντηση σε πολλά ερευνητικά ερωτήματα.

Μέχρι σήμερα οι επιστήμονες χρησιμοποιούν για κάθε βήμα της μελέτης περισσότερα του ενός εργαλεία έτσι ώστε να γίνεται επιβεβαίωση των αποτελεσμάτων που λαμβάνουν. Παρόλα αυτά για τη βέλτιστη απάντηση των ερευνητικών ερωτημάτων είναι αναγκαία η ανάπτυξη ενός ενιαίου βιοπληροφορικού εργαλείου που θα πραγματοποιεί όλες τις αναλύσεις που χρειάζονται ανεξάρτητα του ερευνητικού ερωτήματος που τίθεται κάθε φορά. Αυτό μέχρι σήμερα φαίνεται δύσκολο να υλοποιηθεί, διότι κάθε ανάλυση έχει τους δικούς της περιορισμούς. Το μέλλον όμως θα δείξει αν η επιστημονική κοινότητα θα καταφέρει να λύσει και αυτό το ζήτημα…

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Βιβλιογραφία

Chatterjee, A., Ahn, A., Rodger, E. J., Stockwell, P. A., & Eccles, M. R. (2018). A guide for designing and analyzing RNA-seq data. In *Methods in Molecular Biology* (Vol. 1783, pp. 35–80). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7834-2_3

Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, *5*. https://doi.org/10.12688/F1000RESEARCH.8987.2

Chen, Y., McCarthy, D., Ritchie, M., Robinson, M., Smyth, G., & Hall, E. (2008, September 17). *edgeR: differential analysis of sequence read count data-User's Guide*. http://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. In *Genome Biology* (Vol. 17, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-016-0881-8

Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., Omenn, G., & Meng, F. (2010). NGSQC: Cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*, *11*(SUPPL. 4), S7. https://doi.org/10.1186/1471-2164-11-S4-S7

DNA Fragmentation | NEB. (n.d.). New England BioLabs. Retrieved March 17, 2021, from https://international.neb.com/applications/ngs-sample-prep-and-target-enrichment/dna-fragmentation

Dobin, A. (2019). *STAR manual 2.7.0a*. https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414),

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

57–74. https://doi.org/10.1038/nature11247

Duplicate Sequences. (n.d.). Babraham BioInformatics. Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/8 Duplicate Sequences.html

FASTQ files. (n.d.-a). Retrieved March 27, 2021, from http://drive5.com/usearch/manual/fastq_files.html

FASTQ Files. (n.d.-b). Illumina BaseSpace. Retrieved March 27, 2021, from https://help.basespace.illumina.com/articles/descriptive/fastq-files/

FASTQ files explained. (n.d.). Illumina. https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

FastQC-A Quality Control tool for High Throughput Sequence Data. (n.d.). Babraham Bioinformatics. Retrieved April 1, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC. (n.d.-a). Retrieved April 1, 2021, from https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

FastQC. (n.d.-b). Illumina. Retrieved April 1, 2021, from https://emea.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastqc.html

FastQC Report bad_sequence.txt . (n.d.). Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

FastQC Report good_sequence_short.txt . (n.d.). Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

featureCounts function. (n.d.). RDocumentation. Retrieved April 17, 2021, from https://www.rdocumentation.org/packages/Rsubread/versions/1.22.2/topics/featureCounts

File Format Guide. (n.d.). National Center for Biotechnology Information Search Database. Retrieved March 27, 2021, from https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/#fastq-files

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, *8*(1). https://doi.org/10.1002/wrna.1364

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, *50*(8), 96. https://doi.org/10.1038/S12276-018-0071-8

Illumina. (n.d.). *RNA SEQUENCING METHODS COLLECTION-An overview of recent RNA-Seq publications featuring Illumina ® technology*. Retrieved March 14, 2021, from https://emea.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/rna-sequencing-methods-review-web.pdf

Introduction to NGS. (n.d.). Illumina. https://emea.illumina.com/science/technology/next-generation-sequencing.html

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liao, Yang, Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, *47*(8), e47. https://doi.org/10.1093/nar/gkz114

NGS Workflow Steps | Illumina sequencing workflow. (n.d.). Illumina. Retrieved March 26, 2021, from https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html

Oligo(dT) Cellulose Columns- Life Technologies. (n.d.). Retrieved March 14, 2021, from https://tools.thermofisher.com/content/sfs/manuals/15939010.pdf

Overrepresented Sequences. (n.d.). Babraham BioInformatics. Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/9 Overrepresented Sequences.html

Per Base N Content. (n.d.). Babraham BioInformatics. Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/6 Per Base N Content.html

Per Base Sequence Content. (n.d.). Babraham BioInformatics. Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/4 Per Base Sequence Content.html

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210

Per Sequence GC Content. (n.d.). Babraham BioInformatics. Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/5 Per Sequence GC Content.html

Per Tile Sequence Quality. (n.d.). Babraham Bioinformatics. Retrieved April 2, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/12 Per Tile Sequence Quality.html

Pevsner, J. (2015). *Bioinformatics and Functional Genomics* (3rd ed.). John Wiley & Sons Inc.

Quality (Phred) scores. (n.d.). Retrieved March 27, 2021, from http://drive5.com/usearch/manual/quality_score.html

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139. https://doi.org/10.1093/BIOINFORMATICS/BTP616

Shi, W., & Liao, Y. (2021). *Rsubread/Subread Users Guide*. https://bioconductor.org/packages/release/bioc/vignettes/Rsubread/inst/doc/SubreadUsersGuide.pdf

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

The Cancer Genome Atlas Program. (n.d.). The Website of the National Cancer Institute (Https://Www.Cancer.Gov). Retrieved March 11, 2021, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Κοσσίδα, Σ. (2008). *Βιοπληροφορική Δυνατότητες και Προοπτικές*. Σοφία Κοσσίδα, Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών. www.bioacademy.gr/bioinformatics

Μοριακή Γενετική. (n.d.). Retrieved April 11, 2021, from http://ebooks.edu.gr/ebooks/v/html/8547/2668/Biologia_B-Lykeiou_html-empl/index4_2.html

Institutional Repository - Library & Information Centre - University of Thessaly
02/06/2024 12:18:27 EEST - 18.223.196.210