ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

του Τμήματος Βιοχημείας και Βιοτεχνολογίας

«ΕΦΑΡΜΟΓΕΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ - ΜΟΡΙΑΚΗ ΓΕΝΕΤΙΚΗ, ΔΙΑΓΝΩΣΤΙΚΟΙ ΔΕΙΚΤΕΣ»

ΠΑΥΛΙΝΑ ΧΑΛΥΒΟΠΟΥΛΟΥ

Βιοπληροφορική ανάλυση μικροβιακών γονιδιωμάτων / μεταγονιδιωμάτων από δεδομένα νέων τεχνολογιών αλληλούχισης (Next Generation Sequencing)

Λάρισα 2014

Βιοπληροφορική ανάλυση μικροβιακών γονιδιωμάτων / μεταγονιδιωμάτων από δεδομένα νέων τεχνολογιών αλληλούχισης (Next Generation Sequencing)

Bioinformatics analysis of microbial genome / metagenome data from Next Generation Sequencing Technologies

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

κ. Γρηγόριος Αμούτζιας (Επιβλέπων)

Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

κ. Παναγιώτης Μαρκουλάτος

Καθηγητής Εφαρμοσμένης Μικροβιολογίας με έμφαση στη Βιοτεχνολογία, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

κ. Δημήτριος Μόσιαλος

Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, τον κ. Γρηγόριο Αμούτζια για την καθοδήγηση, την κατανόηση και την πολύτιμη βοήθειά του, κατά την διάρκεια της διεκπεραίωσης της διπλωματικής αυτής εργασίας. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Παναγιώτη Μαρκουλάτο και τον κ. Δημήτριο Μόσιαλο για τις πολύτιμες γνώσεις που μου πρόσφεραν κατά τη διάρκεια του μεταπτυχιακού προγράμματος. Γενικά, ένα μεγάλο ευχαριστώ σε όλους τους καθηγητές μου που συνέβαλαν με τη μεθοδικότητα και τη διδασκαλία τους στην επιτυχή ολοκλήρωση των μεταπτυχιακών μου σπουδών.

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος «Εφαρμογές Μοριακής Βιολογίας - Μοριακή Γενετική - Διαγνωστικοί Δείκτες», με τίτλο "Βιοπληροφορική ανάλυση μικροβιακών γονιδιωμάτων / μεταγονιδιωμάτων από δεδομένα νέων τεχνολογιών αλληλούχισης (Next Generation Sequencing)". Ο σκοπός της εργασίας αυτής είναι η δημιουργία ενός υπολογιστικού πρωτόκολλου για την ανάλυση μεταγονιδιωματικών αλληλουχιών χρησιμοποιώντας κατάλληλα και ελεύθερα προσβάσιμα προγράμματα βιοπληροφορικής. Το συγκεκριμένο πρωτόκολο διαχειρίζεται i) την απόκτηση μεταγονιδιωματικών δεδομένων αλληλούχισης από την βάση δεδομένων SRA χρησιμοποιώντας ως παράδειγμα δημοσιευμένα δεδομένα από ανθρώπινα κόπρανα και ενεργό λυματολάσπη, ii) τον ποιοτικό έλεγχο και το φιλτράρισμα των δεδομένων, iii) τη de novo συναρμολόγισή τους προς τον σχηματισμό contigs με δύο διαφορετικά προγράμματα, iv) την πρόβλεψη των γονιδίων και v) την ταξινομική τους ανάλυση. Τέλος, τα αποτελέσματα της ανάλυσης της εργασίας αυτής συγκρίνονται με εκείνα των δημοσιευμένων εργασιών από τις οποίες πάρθηκαν τα πρωτόλεια δεδομένα ενώ γίνεται και μια σύγκριση της αποτελεσματικότητας των προγραμμάτων που χρησιμοποιήθηκαν στο στάδιο της δημιουργίας contigs.

ABSTRACT

This thesis, entitled "Bioinformatics analysis of microbial genome / metagenome data from Next Generation Sequencing Technologies", was part of the master program "Molecular Biology and Genetics Applications - Diagnostic Markers". The aim of this study is the creation of a computational protocol for analyzing metagenomics sequences using appropriate and freely available bioinformatics programs. This protocol manages i) obtaining metagenomics sequencing data from the SRA database using as an example two published data, one from human faeces and one from activated sludge, ii) the quality control and filtering of data, iii) their de novo assembly to form the contigs via two different programs, iv) their gene prediction and v) their taxonomic analysis. Finally, the analysis results of this study are compared with those of the two published studies from which the data were taken, while a comparison of the effectiveness of the programs used in the process of creating contigs is made.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

Ένα από τα πιο αξιοσημείωτα γεγονότα στον τομέα μικροβιακής οικολογίας την τελευταία δεκαετία είναι η εμφάνιση και η ανάπτυξη της μεταγονιδιωματικής. Η μεταγονιδιωματική ορίζεται ως η μελέτη του μεταγονιδιώματος, δηλαδή του ολικού γονιδιωματικού DNA από περιβαλλοντικά δείγματα.

Πληροφορίες της μεταγονιδιωματικής από την άμεση ανάλυση του DNA ενός δεδομένου οικοσυστήματος χωρίς προηγούμενη καλλιέργεια, επιτρέπουν μια βαθύτερη κατανόηση του οικολογικού ρόλου, του μεταβολικού προφίλ, και της εξελικτικής ιστορίας των μικροβίων (Kim et al., 2013). Με τη μεταγονιδιωματική παρέχονται πληροφορίες για το λειτουργικό ρόλο των γονιδίων των μικροβιακών κοινοτήτων και έτσι δίνεται μια ευρύτερη περιγραφή τους σε σχέση με τα αντίστοιχα αποτελέσματα των φυλογενετικών ερευνών. Από μόνη της, η μεταγονιδιωματική δίνει γενετικές πληροφορίες σχετικά με δυνητικά νέα ένζυμα, τη λειτουργία και φυλογένεση των οργανισμών χωρίς να προηγείται η καλλιέργειά τους και τα εξελικτικά προφίλ της λειτουργίας και δομής της κοινότητας (Thomas et al., 2012). Μπορεί επίσης να συμπληρωθεί με προσεγγίσεις μετατρανσκριπτομικής ή μετα πρωτεομικής για να περιγράψει δραστηριότητες έκφρασης (Wilmes and Bond, 2006) (Gilbert et al., 2010a).

1.1 Τεχνολογίες Αλληλούχισης Νέας Γενιάς (Next Generation Sequencing technologies / NGS technologies)

Η ραγδαία ανάπτυξη των τεχνολογιών αλληλούχισης επιτρέπει την ενδελεχή αξιολόγηση των μικροβιακών κοινοτήτων μέσω της ανάγνωσης νουκλεοτιδικών αλληλουχιών με σχετικά χαμηλό κόστος. Οι τεχνολογίες Next Generation Sequencing (NGS) έχουν φέρει επανάσταση στον τομέα της μικροβιακής οικολογίας, δεδομένου ότι επιτρέπουν στους ερευνητές να προσεγγίσουν το πραγματικό επίπεδο βιοποικιλότητας μιας μικροβιακής κοινότητας.

1.1.1 Χαρακτηριστικά των Τεχνολογιών Αλληλούχισης και κόστος Αλληλούχισης

Εδώ και επτά χρόνια η μέθοδος αυτοματοποιημένης αλληλούχισης κατά Sanger είναι πλέον παρωχημένη για την ανάλυση γονιδιωμάτων και μεταγονιδιωμάτων. Η μέθοδος Sanger είχε κυριαρχήσει στην βιομηχανία για σχεδόν δύο δεκαετίες, και οδήγησε σε

μια σειρά από μνημειώδη επιτεύγματα, συμπεριλαμβανομένης της ολοκλήρωσης της αλληλούχισης του ανθρώπινου γονιδιώματος (International Human Genome Sequencing Consortium, 2004). Παρά τις πολλές τεχνικές βελτιώσεις κατά τη διάρκεια αυτής της περιόδου, οι περιορισμοί της τεχνολογίας αυτής κατέδειξαν την ανάγκη για νέες και βελτιωμένες τεχνολογίες για την αλληλούχιση μεγάλου αριθμού ανθρώπινων γονιδιωμάτων.

Generation	**Platform**	**Pros**	**Cons**	**Read length (bp)**	**Gb per run**
First	Sanger	Long read lengths; high single-pass accuracy; good ability to call repeats and homopolymer regions	Requires relatively large amounts of DNA	1000-2000	
Second	454 pyrosequencing	Longer reads improve mapping in repetitive regions, fast run times	High reagent cost, high error rates in the number of bases in homopolymer repeats	~330bp	0.45
	Solid	Two-base encoding provides inherent error correction	Many reads do not fit anywhere in the genome, long run times	50bp	30-50
	Illumina	Currently the most widely used	Lower accuracy in	~75-100	18-35

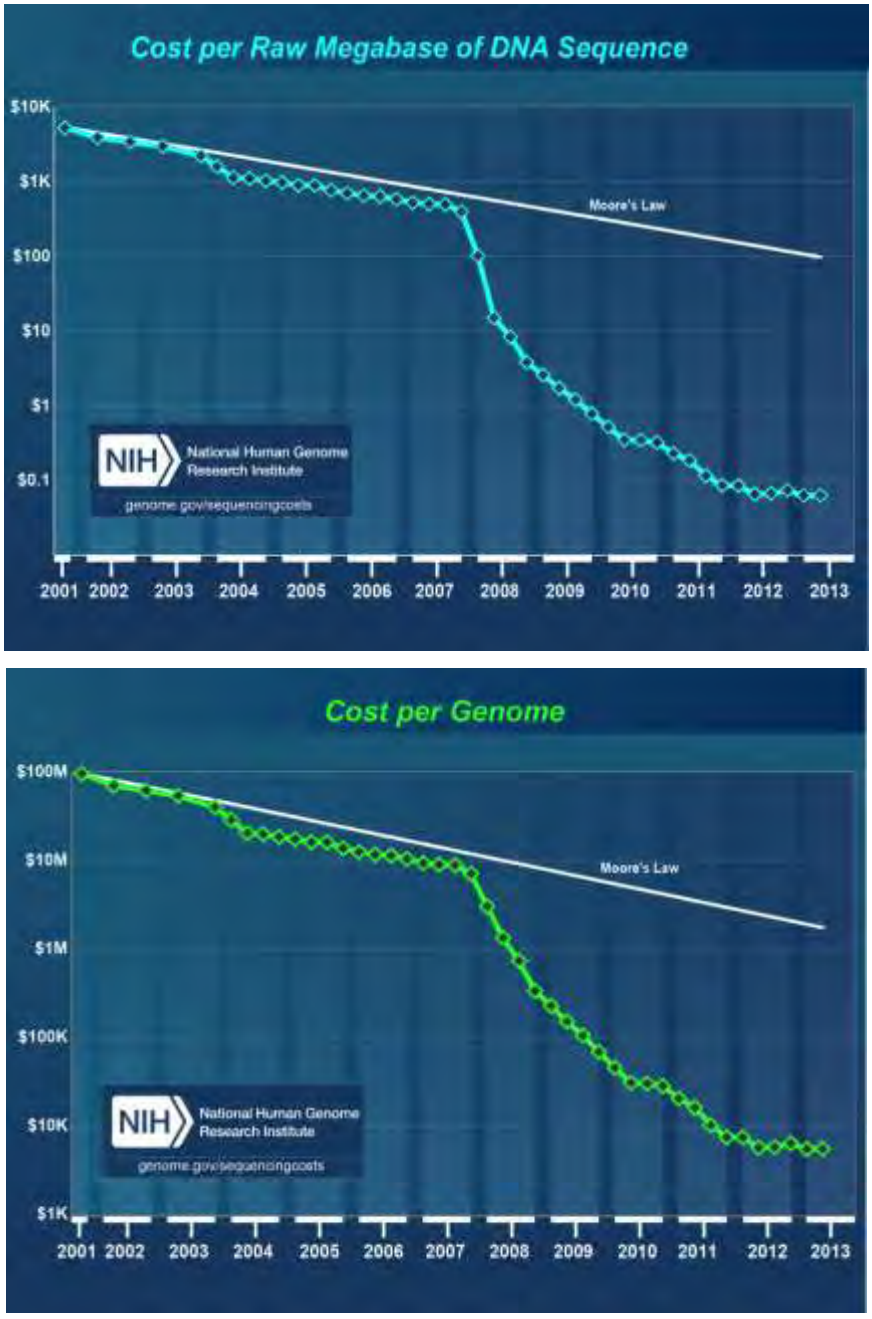
		platform in the field	base recognition		
	Pacific Biosciences	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	964	Not Available
Third	Ioan torrent/ Ion proton	Direct measurement of nucleobase incorporation events; DNA synthesis reaction operates under natural	Sequential washing steps can lead to accumulation of errors; potential difficulties in reading through highly	100–200	
	Oxford nanopore	Potential for long read lengths; low cost of **α**HL nanopore production; no fluorescent labeling or optics necessary	Cleaved nucleotides may be read in the wrong order; difficult to fabricate a device with multiple parallel pores	Not yet quantified	

Πίνακας 1. Συγκριτικός πίνακας των τεχνολογιών αλληλούχισης πρώτης, δεύτερης και τρίτης γενιάς.

Πλέον υπάρχει ένας σημαντικός αριθμός νέων τεχνολογιών αλληλούχισης, ο

καθένας με τα ιδιαίτερα χαρακτηριστικά του. Η επιλογή της πλατφόρμας αλληλούχισης γίνεται σύμφωνα με το είδος της ανάλυσης που θέλουμε να γίνει.

Το κύριο πλεονέκτημα που προσφέρεται από τις NGS τεχνολογίες είναι η ικανότητα να παράγουν ένα τεράστιο όγκο των δεδομένων φτηνά και μάλιστα σε ορισμένες περιπτώσεις πάνω από ένα δισεκατομμύριο short reads ανά κύκλο! (Metzker, 2010).



Εικόνα 1. (http://www.genome.gov/sequencingcosts/) Για κάθε ένα γράφημα φαίνονται τα στοιχεία του NHGRI (National Human Genome Research Institute) από

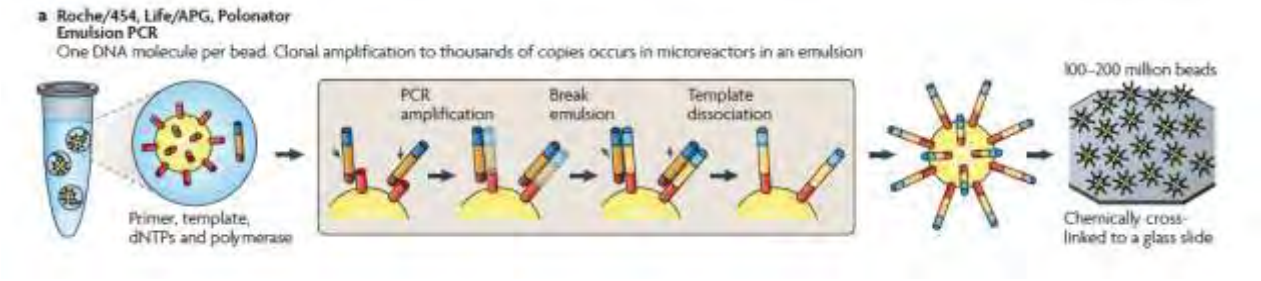
το 2001 έως το 2013 (πάνω) του κόστους ανά megaβάση της αλληλουχίας του DNA και (κάτω) του κόστους ανά ανθρώπινο γονιδίωμα.

Τα δεδομένα κοστολόγησης που παρουσιάζονται στην εικόνα 1 συνοψίζονται σε σχέση με δύο παραμέτρους: (1) «το κόστος ανά megaβάση της αλληλούχισης DNA», (1Mb = ένα εκατομμύριο βάσεις) (2) «το κόστος ανά ανθρώπινο γονιδίωμα». Για να φανεί η φύση των μειώσεων του κόστους της αλληλούχισης DNA, κάθε γράφημα δείχνει επίσης υποθετικά πώς θα έπρεπε να μειώνεται το κόστος αν ίσχυε ο νόμος του Moore. Ο συγκεκριμένος νόμος βασίζεται σε εμπειρική παρατήρηση και περιγράφει μια μακροπρόθεσμη τάση στη βιομηχανία των ηλεκτρονικών υπολογιστών όπου η υπολογιστική ισχύς διπλασιάζεται κάθε δύο χρόνια. Οι τεχνολογικές εξελίξεις που «τηρούν» το Νόμο του Moore θεωρούνται ευρέως ότι πηγαίνουν εξαιρετικά καλά, καθιστώντας το χρήσιμο για σύγκριση.

Ένα ιδιαίτερο πρόβλημα είναι ότι η αύξηση της απόδοσης των NGS πλατφορμών δημιουργεί προκλήσεις στη συναρμολόγηση των αλληλουχιών, ιδίως όσον αφορά στις απαιτήσεις σε μνήμη υπολογιστή RAM. Για παράδειγμα, ένας αλληλουχητής Illumina HiSeq2000 μπορεί να παράγει 600 Gb σε ένα ενιαίο τρέξιμο. Ως αποτέλεσμα, η μεταγονιδιωματική σήμερα είναι μοιρασμένη μεταξύ των μικρότερων, πιο στοχευμένων project με συναρμολογήσεις και project μεγάλης κλίμακας χωρίς συναρμολογήσεις (Teeling and Glöckner, 2012). Η τάση της μεταγονιδιωματικής για τις τεράστιες κλίμακες δεδομένων είχε ήδη προβλεφθεί πριν ακόμα καταστούν διαθέσιμες οι δεύτερης και τρίτης γενιάς NGS πλατφόρμες και έχει ονομαστεί «μεγαγονιδιωματική» (Handelsman, 2005). Αυτά τα projects των μεγαγονιδιωμάτων περιλαμβάνουν το Ανθρώπινο Μικροβίωμα (Turnbaugh et al., 2007) και το Μικροβίωμα της Γης ανάμεσα σε άλλα (Gilbert et al., 2010a) *(Gilbert et al.,* 2010b).

1.1.2 Roche 454 – Pyrosequencing

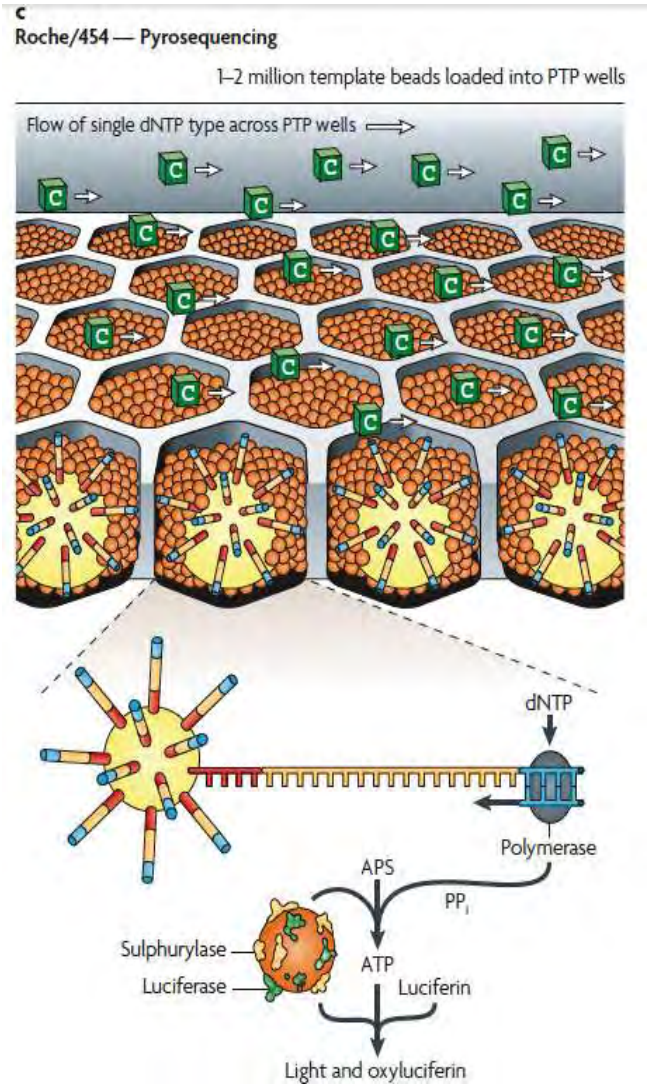
Το 2005, η εταιρία 454 Life Sciences παρουσίασε την πρώτη νέας γενιάς (NGS) πλατφόρμα αλληλούχισης. Η τεχνολογία ονομάζεται pyrosequencing και πραγματοποιεί αλληλούχιση με σύνθεση, σε πραγματικό χρόνο. Ήδη όμως θεωρείται παρωχημένη.



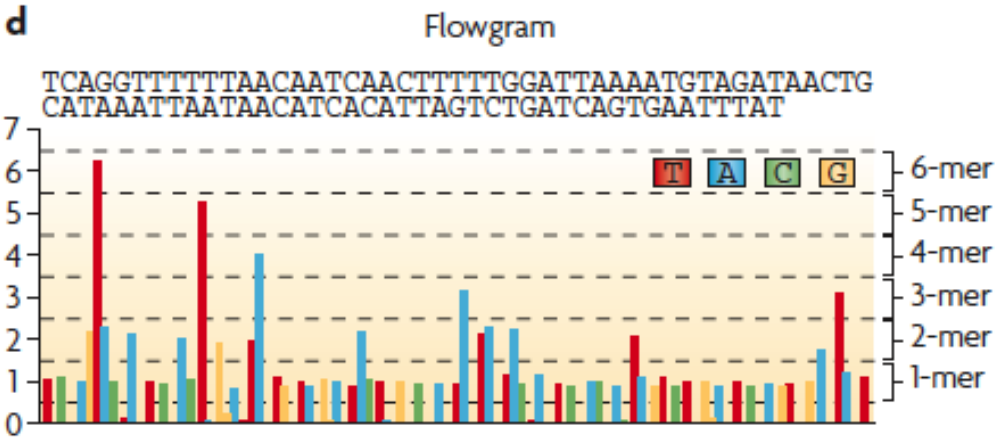
Εικόνα 2. Ένα μόριο DNA ανά σφαιρίδιο. Κλωνική ενίσχυση σε χιλιάδες αντίγραφα σε μικροαντιδραστήρες σε ένα γαλάκτωμα (Metzker, 2010)

Η DNA πολυμεράση αρχίζει την επιμήκυνση χρησιμοποιώντας τα dNTPs που έχουμε βάλει ανά φορά (ένα είδος από τα τέσσερα κάθε φορά). Αν το dNTP έχει ενσωματωθεί, παράγεται πυροφωσφορικό (PPi). Το PPi μετατρέπεται σε ATP από τη σουλφορυλάση. Η λουσιφεράση χρησιμοποιεί το ATP για να οξειδώσει τη λουσιφερίνη και έτσι παράγεται μια λάμψη λόγω χημειοφωταύγειας (εικόνα 3).

Το φώς που παράγεται εμφανίζεται σαν μια κορυφή για κάθε τύπο νουκλεοτιδίου που ενσωματώνεται. Ανάλογα με το ύψος της κορυφής βλέπουμε αν ενσωματώθηκαν παραπάνω από ένα νουκλεοτίδια τη φορά. Έτσι διαβάζουμε την αλληλουχία. Αν κάποιο dNTP δεν ενσωματώνεται, αποικοδομείται, δεν παράγεται ATP ούτε και φως.



Εικόνα 3. 1-2 εκατομμύρια σφαιρίδια προτύπου φορτώνονται σε PTP πηγαδάκια (Metzker, 2010)



Εικόνα 4. Παράδειγμα flowgram. (Metzker, 2010)

Το φως που παράγεται από τον ενζυματικό καταρράκτη καταγράφεται ως μια σειρά κορυφών, ονομάζεται flowgram (εικόνα 4).

Το βίντεο στον παρακάτω σύνδεσμο απεικονίζει τις βασικές αρχές αυτής της τεχνολογίας αλληλούχισης.
http://www.youtube.com/watch?v=nFfgWGFe0aA

1.1.3 Pacific Biosciences

Το 2010, η εταιρία Pacific Biosciences παρουσίασε μια πλατφόρμα DNA – αλληλούχισης μονόκλωνου μορίου σε πραγματικό χρόνο (SMRT – Single Molecule Real Time). Η συγκεκριμένη τεχνολογία αλληλούχισης στηρίζεται σε πραγματικό χρόνο με βάση τον φθορισμό. Για την προετοιμασία του δείγματος δεν απαιτείται στάδιο ενίσχυσης καθώς πραγματοποιείται αλληλούχιση με σύνθεση μονόκλωνου μορίου (Shokralla et al., 2012). Η πρωτοπορία της πλατφόρμας στηρίζεται σε δύο τεχνολογίες:

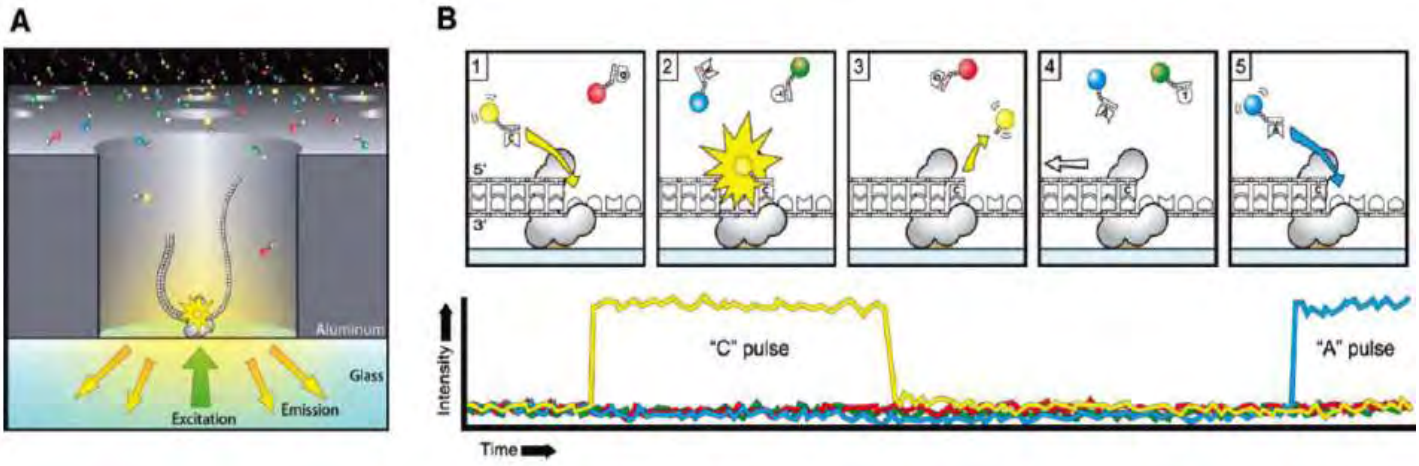
- α) στην σύνδεση της φθορίζουσας χρωστικής με την 5’ φωσφορική ομάδα του κάθε δεοξυριβονουκλεοτιδίου και όχι με την βάση (παλιά μέθοδος) και
- β) στην χρησιμοποίηση μιας νανο-δομής, ονομαζόμενη ως Zero Mode Waveguide (ZMW) στην οποία παίρνει μέρος ο πολυμερισμός του DNA, σε πραγματικό χρόνο (Niedringhaus et al., 2011).

Η νανο-συσκευή Zero Mode Waveguide (ZMW) αποτελείται από δεκάδες χιλιάδες οπές, με διάμετρο νανομέτρων και είναι ο μικρότερος όγκος ανίχνευσης στον κόσμο (10^{-21} liters). Κατασκευάζεται με διάτρηση μιας λεπτής μεταλλικής μεμβράνης και υποστηρίζεται από ένα διαφανές υπόστρωμα (Shokralla et al., 2012).

Η DNA πολυμεράση , ο εκκινητής και το DNA εκμαγείο (δημιουργία συμπλόκου) σταθεροποιούνται στον πυθμένα της νανο-συσκευής (ZMWs). Στη συνέχεια στο θάλαμο του ZMW προστίθενται φθορίζοντα νουκλεοτίδια. Η οπτική κάμερα και το σύστημα λέιζερ καταγράφει σε πραγματικό χρόνο την δραστικότητα της DNA πολυμεράσης με τα φθορίζοντα μόρια. Κατά τη διαδικασία της ενσωμάτωσης αυτών, η φθορίζουσα ομάδα διαχέεται μακριά, και ακολουθεί μετατόπιση του κλώνου, κάνοντας έτσι χώρο για την ανάγνωση του επόμενου νουκλεοτιδίου της αλυσίδας (εικόνα 5).

Το είδος της τεχνολογίας που προώθησε η Pacific Biosciences, χρησιμοποιεί

τη φυσική ικανότητα της DNA πολυμεράσης να ενσωματώνει δέκα ή περισσότερα νουκλεοτίδια ανά δευτερόλεπτο σε αρκετές χιλιάδες παράλληλες οπές ZMWs (Shokralla et al., 2012). Η αλληλούχιση γίνεται με γρήγορο ρυθμό και δημιουργεί μεγάλου μήκους reads (της τάξεως των 1000 βάσεων) κάτι που αυξάνει την ποιότητα της συναρμολόγησης και τον εντοπισμό SNP (Single- Nucleotide Polymorphisms) (Niedringhaus et al., 2011).



Εικόνα 5. (Niedringhaus et al., 2011). Σχηματική απεικόνιση της PacBio μοριακής αλληλούχισης πραγματικού χρόνου. (A) Η πλάγια όψη μιας ενιαίας νανοδομής ZMW που περιέχει μία μόνο DNA πολυμεράση (Φ 29) δεσμευμένη στο κάτω μέρος της γυάλινης επιφάνειας. Το ZMW και το σύστημα συνεστιακής απεικόνισης επιτρέπουν την ανίχνευση του φθορισμού μόνο από την επιφάνεια του πυθμένα κάθε ZMW. (B) Αναπαράσταση της ενσωμάτωσης ενός νουκλεοτιδίου υπόστρωμα επισημασμένου με φθορίζον υπόστρωμα σε ένα εκμαγείο που αλληλουχείται. Η αντίστοιχη χρονική ανίχνευση φθορισμού σε σχέση με καθένα από τα πέντε βήματα ενσωμάτωσης φαίνεται ακριβώς από κάτω (Copyright 2009 American Association for the Advancement of Science).

Τα βίντεο στον παρακάτω σύνδεσμο απεικονίζουν τις βασικές αρχές αυτής της τεχνολογίας αλληλούχισης.

http://www.youtube.com/watch?v=NHCJ8PtYCFc

http://www.youtube.com/ watch?v=GX6RSKh4J7E

1.1.4 Illumina/Solexa

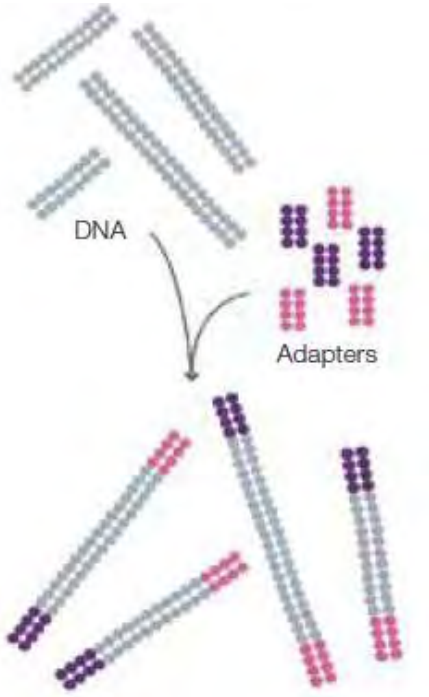
Το 2007, η εταιρεία Illumina απέκτησε τη Solexa, η οποία ανέπτυξε μια πολύ επιτυχημένη τεχνολογία αλληλούχισης των γονιδιωμάτων. Η συγχώνευση αποτέλεσε καταλύτη για την ανάπτυξη εργαλείων και προηγμένων μηχανημάτων αλληλούχισης. Κάποια από τα μηχανήματα της Illumina είναι το HiSeq System, HiScan SQ, Genome Analyzer MiSeq. Η επιλογή χρήσης του κάθε μηχανήματος εξαρτάται από το μέγεθος της αλληλουχίας που είναι προς αλληλούχιση. (http://www.illumina.com)

Η τεχνολογία της Illumina προσφέρει τη δυνατότητα δημιουργίας εκατομμυρίων αντιδράσεων διαφορετικών δειγμάτων με μαζικό παράλληλο τρόπο. Επομένως, μπορούν να αναλυθούν πολλά δείγματα μαζί σε ένα μόνο τρέξιμο. Αυτό επίσης συνεπάγει και μείωση του χρόνου της αλληλούχισης. (http://www.illumina.com)

Ο τρόπος κατά τον οποίο γίνεται αλληλούχιση γονιδιωμάτων με Illumina είναι ο εξής:

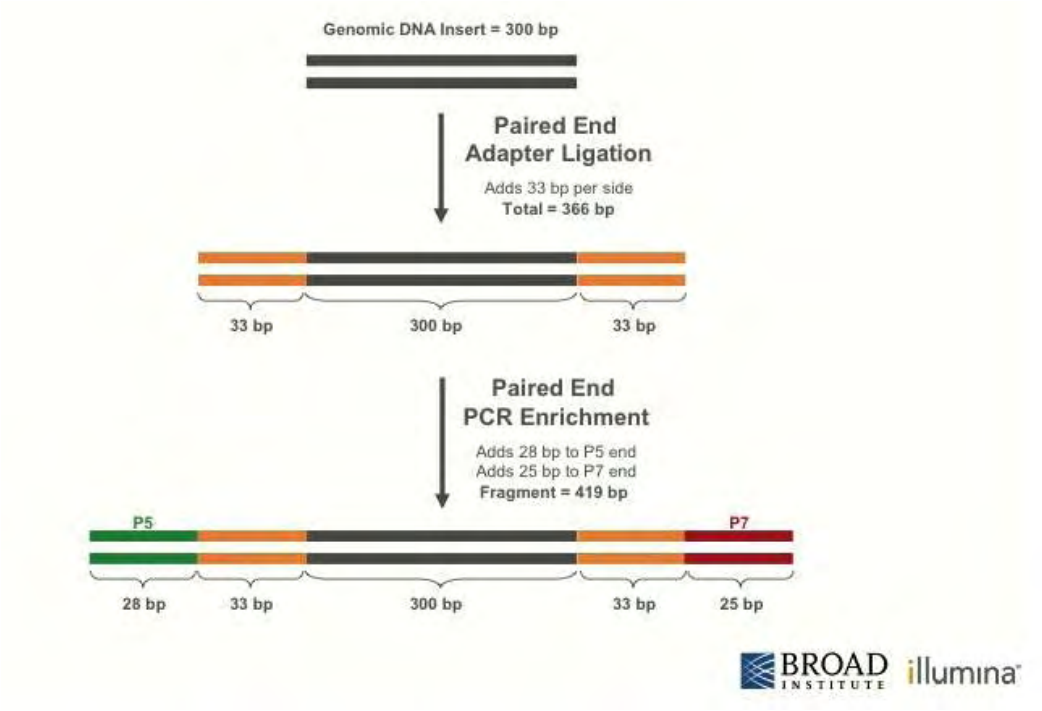
1 Προετοιμασία βιβλιοθήκης (Library Preparation)

Τα δίκλωνα μόρια του DNA των δειγμάτων τεμαχίζονται σε τυχαία κομμάτια και στα άκρα τους ενσωματώνονται οι λεγόμενοι adapters. Οι adapters έχουν συγκεκριμένα αλλά διαφορετικά barcodes για το κάθε δείγμα. Τα barcodes είναι μεμονωμένες αλληλουχίες οι οποίες προστίθενται στα δείγματα ώστε να μπορεί να γίνει αντιστοίχιση του θραύσματος , κατά την ανάλυση των δεδομένων, με το δείγμα στο οποίο ανήκει.



Εικόνα 6. Τυχαία κατακερματισμένο γονιδιωματικό DNA και adapters στα δύο άκρα των θραυσμάτων http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Μετά το DNA με τους adaptors ηλεκτροφορείται και επιλέγονται κομμάτια συγκεκριμένου μεγέθους (συνήθως 200 – 300 bp). Σε αυτά τα κομμάτια προστίθενται μέσω PCR P7 και P5 άκρα, τα οποία χρησιμεύουν για την πρόσδεση στην επιφάνεια εργασίας. Τα δίκλωνα μόρια γίνονται μονόκλωνα.



Εικόνα 7. Σύνδεση των adapters και των P5 και P7 άκρων http://openwetware.org/wiki/Image:Libprep3.jpg

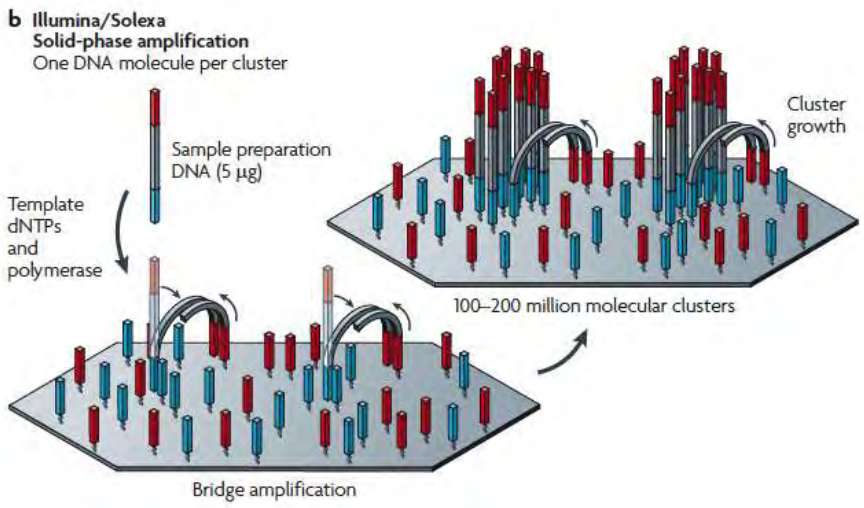
2 Δημιουργία συμπλέγματος-cluster (Cluster Generation)

Τα μονόκλωνα μόρια τοποθετούνται πάνω σε μία επιφάνεια-πλάκα εργασίας, (workflow-glass flow cell). Η κάθε πλάκα αποτελείται εσωτερικά από ολιγονουκλεοτίδια τα οποία είναι συμπληρωματικά ως προς τους adapters, και χωρίζεται σε οχτώ ξεχωριστές λωρίδες.

Εικόνα 8. Πολλά δείγματα μπορούν να φορτώνονται στις οκτώ λωρίδες για ταυτόχρονη ανάλυση σε ένα σύστημα Illumina Sequencing http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.p

df

Πραγματοποιείται τυχαίος υβριδισμός (μέσω εναλλαγής υψηλής με χαμηλή θερμοκρασία) μεταξύ των ολιγονουκλεοτιδίων της πλάκας με τους adapters του ενός άκρου των μονόκλωνων θραυσμάτων DNA. Όπως φαίνεται στην εικόνα 9, οι ελεύθεροι adapters των μονόκλωνων μορίων υβριδίζονται με τα ολιγονουκλεοτίδια της πλάκας δημιουργώντας γέφυρες– bridge amplification (Mardis, 2011). Μία ισοθερμική πολυμεράση ενισχύει για την δημιουργία κλώνων. Επίσης, και οι adapters της πλάκας δρουν ως εκκινητές για την ενίσχυση (Zhou et al., 2010). Η κάθε βιβλιοθήκη θραυσμάτων αποτελείται πλέον από εκατοντάδες εκατομμύρια μοναδικά συμπλέγματα (clusters). Τα συμπληρωματικά συμπλέγματα αποκόπτονται και απομακρύνονται ξεπλένοντας. Οι αντίστροφοι κλώνοι διασπώνται και ξεπλένονται. Τα άκρα μπλοκάρονται και ο εκκινητής αλληλούχισης υβριδοποιείται με τα DNA. Μετά τη δημιουργία συμπλέγματος, οι βιβλιοθήκες είναι έτοιμες για αλληλούχιση.

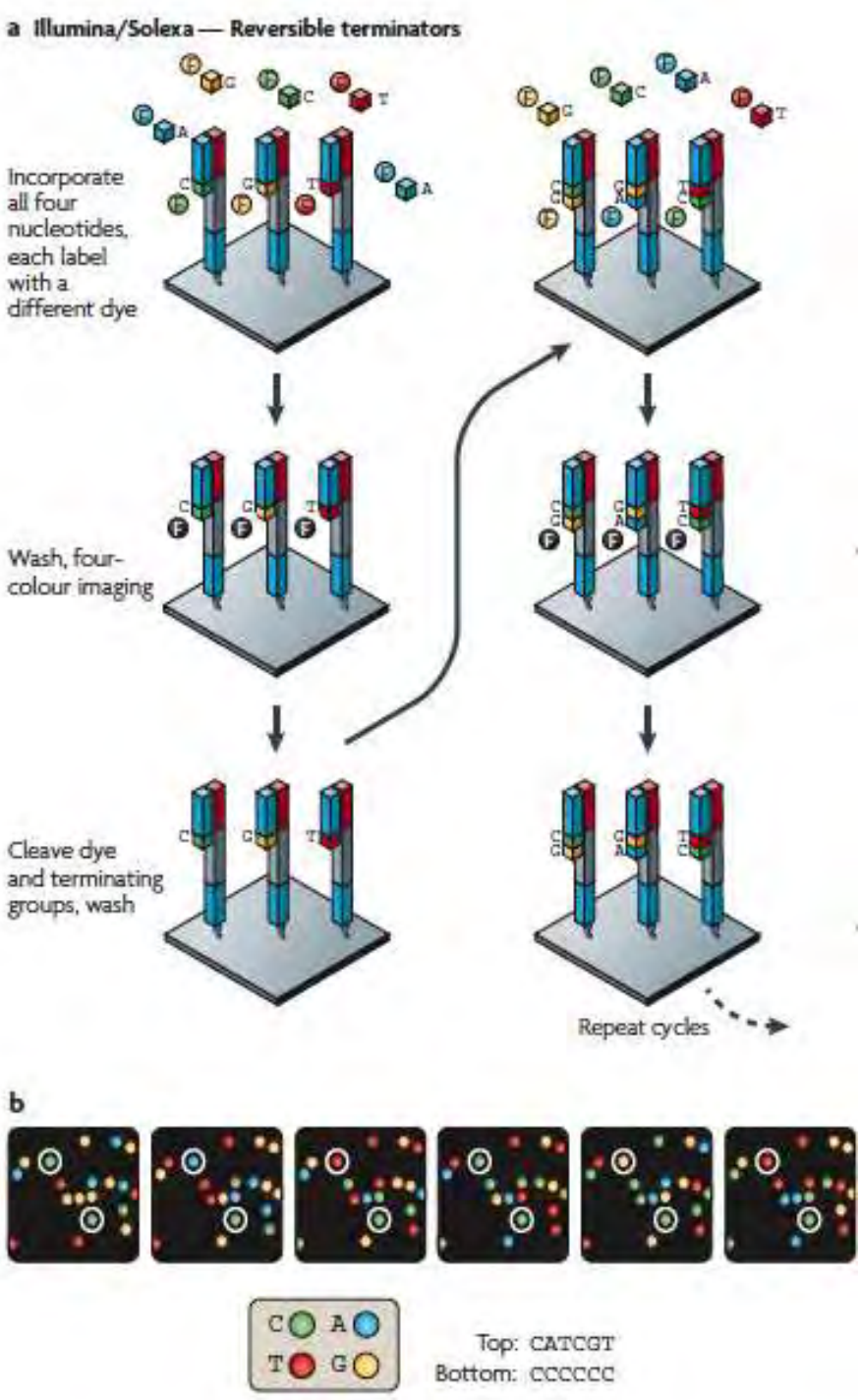


Εικόνα 9. Bridge amplification και σχηματισμός των clusters (Metzker, 2010)

3 Αλληλούχιση (Sequencing)

Η αλληλούχιση όλων των clusters γίνεται ταυτόχρονα βάση προς βάση με παράλληλο τρόπο χρησιμοποιώντας τέσσερις διαφορετικές φθορίζουσες χρωστικές συνδεδεμένες με τέσσερα διαφορετικά ολιγονουκλεοτίδια (A, T, G και C) (Zhou et al., 2010). Οι τέσσερις φθορίζουσες με τις βάσεις πλησιάζουν την βάση του cluster αλλά μόνο μία θα ενωθεί. Και οι τέσσερις βάσεις ανταγωνίζονται μεταξύ τους για να

συνδεθούν με το εκμαγείο. Αυτός ο ανταγωνισμός εξασφαλίζει την υψηλότερη δυνατή ακρίβεια. Μόλις το λέιζερ – CCD camera ανιχνεύσει ότι όντως η συμπληρωματική βάση είναι σωστή (από το χρώμα που εκπέμπει) τότε καταγράφεται το χρώμα της βάσης, η φθορίζουσα χρωστική αφαιρείται και μένει η βάση μετά από ξέπλυμα. Το ίδιο γίνεται και για την επόμενη βάση της αλυσίδας του cluster μέχρι να τερματιστεί. Έτσι δημιουργούνται συμπληρωματικές αλυσίδες των clusters (εικόνα 10).



Εικόνα 10. α | Ενσωμάτωση των βασεων, έκπλυση και δημιουργία αλυσίδας β | Οι εικόνες των τεσσάρων χρωμάτων επισημαίνουν τα δεδομένα αλληλούχισης από δύο κλωνικά ενισχυμένα πρότυπα (Metzker, 2010).

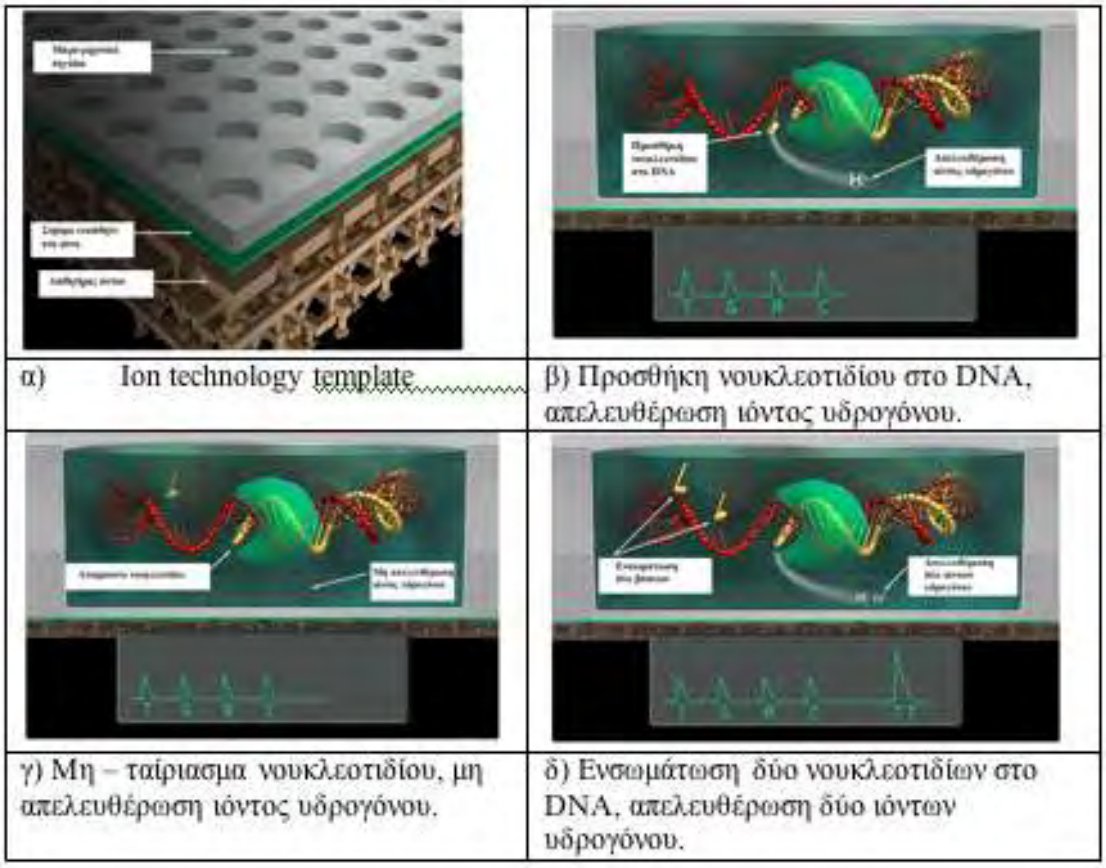
Τα βίντεο στον παρακάτω σύνδεσμο απεικονίζουν τις βασικές αρχές αυτής της τεχνολογίας αλληλούχισης.

http://www.youtube.com/watch?v=77r5p8IBwJk&feature=related

http://www.youtube.com/watch?v=l99aKKHcxC4

1.1.5 Ion Torrent

Το 2010, η Life Technologies παρουσίασε μία καινούργια τεχνολογία αλληλούχισης, την Ion Torrent (Metzker, 2010). Η συγκεκριμένη τεχνολογία δημιουργεί μια άμεση σύνδεση μεταξύ χημικών και ψηφιακών πληροφοριών, επιτρέποντας έτσι τη γρήγορη, απλή και μαζικά κλιμακούμενη αλληλούχιση. Χρησιμοποιεί την απλή χημεία του νουκλεϊκού οξέος, σύμφωνα με τον Watson, σε μια απίστευτα ισχυρή, τεχνολογία ημιαγωγών. Η αρχή της τεχνολογίας Ion Torrent βασίζεται σε μια αναπτυγμένη βιοχημική διαδικασία, στην οποία ένα νουκλεοτίδιο ενσωματώνεται σε μια αλυσίδα του DNA από μια πολυμεράση, με αποτέλεσμα την απελευθέρωση ιόντων υδρογόνου, ως παραπροϊόν (Pareek et al., 2011).



Εικόνα 11. Μέθοδος αλληλούχισης Ion Torrent (http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/Semiconductor-Sequencing-Technology/Ion-Torrent-Technology-How-Does-It-Work.html)

Η εκτέλεση της βιοχημικής διαδικασίας γίνεται με μαζικό και παράλληλο τρόπο, σε chip υψηλής πυκνότητας μικρο-μηχανικών πηγαδιών. Κάθε πηγάδι δέχεται ένα μόνο πρότυπο DNA από βιβλιοθήκη. Κάτω από το πηγάδι είναι ένα ιοντικά-ευαίσθητο στρώμα και ένας αισθητήρας ιόντων για την ανίχνευση της αλλαγής της συγκέντρωσης των ιόντων υδρογόνου λόγω της ενσωμάτωσης των νουκλεοτιδίων. (Shokralla et al., 2012). Η διαδικασία φαίνεται στην εικόνα 11.

Το 2012, η Life Technologies παρουσίασε μια νέα γενιά αλληλούχισης, την Ion Proton. Η Ion Proton είναι ένα chip το οποίο έχει τη δυνατότητα αλληλούχισης του ανθρώπινου γονιδίωματος και του ανθρώπινου εξονιώματος μέσα σε λίγες ώρες (Pareek et al., 2011). Το chip χρησιμοποιεί την τεχνολογία ημιαγωγών CMOS, παρόμοια με εκείνη των ψηφιακών φωτογραφικών μηχανών, με τη διαφορά ότι αντί για τον εντοπισμό φωτός το chip αναγνωρίζει τη χημεία των μορίων και μεταφράζει-αποκωδικοποιεί κατευθείαν σε ψηφιακά δεδομένα.
(http://www.vincentabry.com/en/ion-proton-sequencer-decodes-human-genome-in-1-day-for-1000-dollars-1543)

Εικόνα 12. Ion Proton chip Ουσιαστικά είναι ένα πολύ μικρό pH-meter. Δεν βασίζεται σε ανίχνευση φωτός [http://www.vincentabry.com/en/ion-proton-sequencer- decodes-human-genome-in-1-day-for-1000-dollars-1543]

Τα βίντεο στον παρακάτω σύνδεσμο απεικονίζουν τις βασικές αρχές αυτής της τεχνολογίας αλληλούχισης.
http://www.youtube.com/watch?v=yVf2295JqUg

https://www.youtube.com/watch?v=MxkYa9XCvBQ

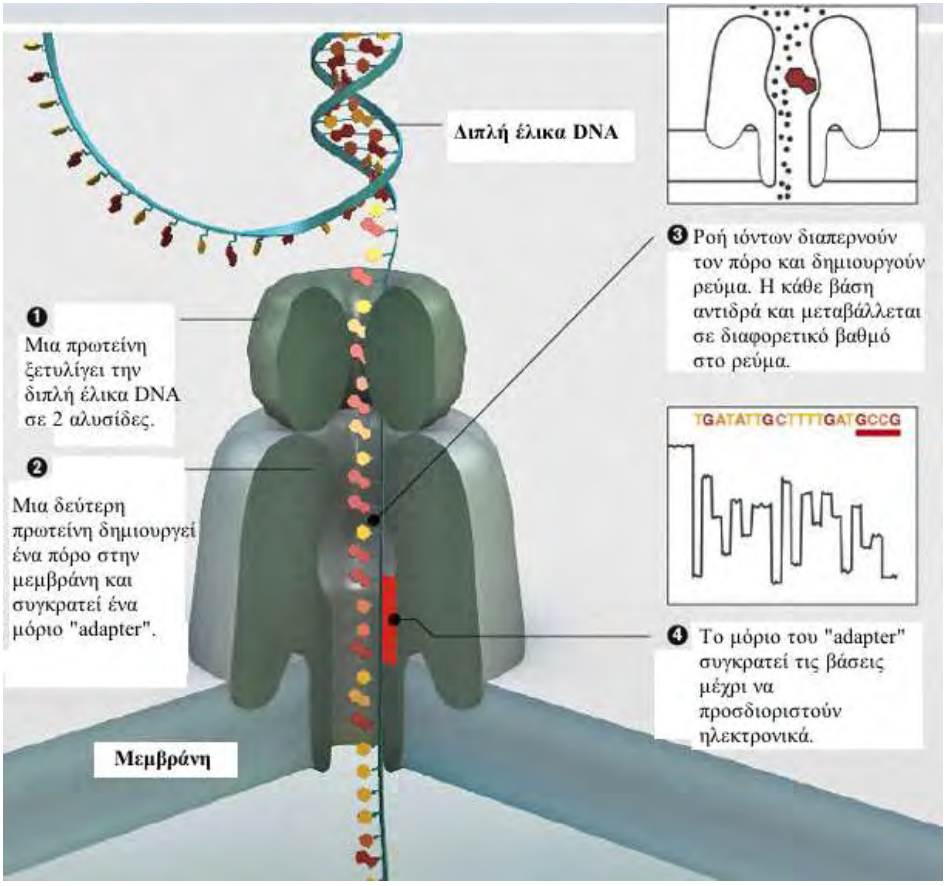
1.1.6 Nanopore Oxford Technologies

Στο μέλλον, πιστεύεται ότι η τρίτη γενιά αλληλούχισης θα πραγματοποιείται από array που θα αποτελούνται από μικρούς σε σχήμα πρωτεϊνικούς πόρους (nanopores). Η ιδέα της αλληλούχισης, βασισμένη σε nanopores, προτάθηκε πριν από περίπου 20 χρόνια. Τα πρώτα πειράματα διεξήχθησαν το 1996, χρησιμοποιώντας ως πόρους την πρωτεΐνη α-αιμολυσίνη. Τα αποτελέσματα των πειραμάτων δεν ήταν τα επιθυμητά καθώς η μέθοδος ήταν ατελής και η τεχνολογική γνώση δεν επαρκούσε (Schneider and Dekker, 2012).

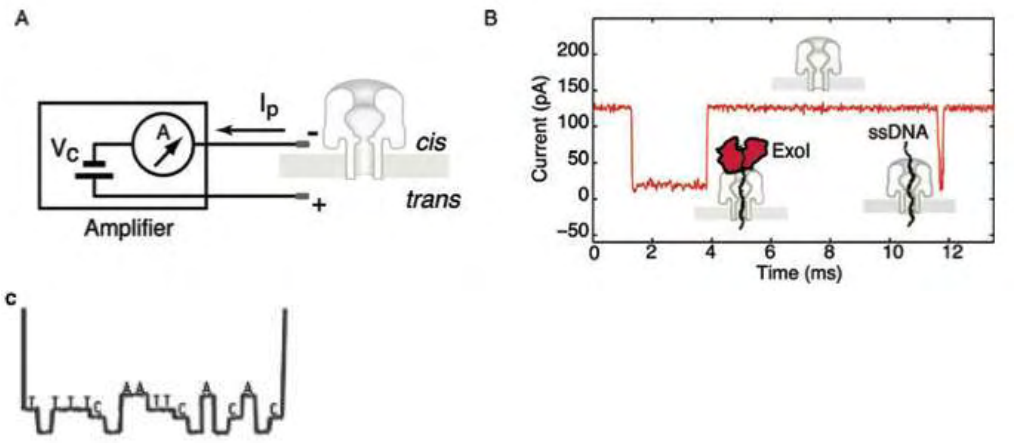
Σήμερα, η μέθοδος αυτή είναι πολλά υποσχόμενη, εξαιτίας της ανάπτυξης της τεχνολογίας, καθώς υπολογίζεται ότι θα είναι πιο φτηνή σε κόστος και πιο γρήγορη με ελάχιστη ποσότητα δείγματος. Ένα ακόμη πλεονέκτημα αυτής της μεθόδου, είναι η μη σήμανση του μορίου, χωρίς ενίσχυση του δείγματος, για την ανίχνευση των αζωτούχων βάσεων (Anselmetti, 2012). Επίσης, αναμένεται να προσφέρει λύσεις στον περιορισμό των τεχνολογιών αλληλούχισης με short reads και να καταστήσει δυνατή την αλληλούχιση μεγάλων μορίων σε μερικά λεπτά, χωρίς την τροποποίηση ή την προετοιμασία δειγμάτων (2012).

Η τεχνολογία των nanopores για την ανάλυση των νουκλεϊκών οξέων βασίζεται σε δύο διαφορετικές ιδέες: i) τα βιολογικά nanopores (Biological nanopores) όπως είναι η αιμολυσίνη και ii) τα συνθετικά nanopores στερεής κατάστασης (Solid-state- Graphene) nanopore. Η κεντρική ιδέα κατά τη διαδικασία της αλληλούχισης με nanopore (εικόνα 13) και στις δύο προσεγγίσεις είναι ίδια, με τη διαφορά ότι τα βιολογικά nanopores στηρίζονται σε μια λιπιδική διπλοστιβάδα ενώ τα συνθετικά nanopores στηρίζονται σε μία συνθετική μεμβράνη. Συγκεκριμένα, το μόριο του δίκλωνου DNA εισέρχεται σε ένα πρωτεϊνικό πόρο. Στον πρωτεϊνικό πόρο στηρίζεται μία άλλη πρωτεΐνη (πολυμεράση) η οποία κατα την διέλευση του δίκλωνου DNA, το ξετυλίγει και έτσι δημιουργούνται δύο αλυσίδες (Venkatesan and Bashir, 2011). Εσωτερικά του πόρου, καθώς ξετυλίγεται η μονόκλωνη αλυσίδα του DNA, περνούν ιόντα και εφαρμόζεται τάση ρεύματος στα άκρα της μεμβράνης (2012). Το δυναμικό που εφαρμόζεται διαμέσου της μεμβράνης, δημιουργεί ιονικό ρεύμα (Maitra et al., 2012). Στο δυναμικό που δημιουργείται , σε σχέση με την πολικότητα του μορίου, στηρίζεται ο εντοπισμός των βάσεων (A,T,G ή C) την κάθε

χρονική στιγμή και ύστερα γίνεται η καταγραφή αυτών, όπως φαίνεται στην εικόνα 14 (Venkatesan and Bashir, 2011).



Εικόνα 13. Διαδικασία αλληλούχισης με τεχνολογία Nanopore (http://www2.technologyreview.com/article/427677/nanopore-sequencing/)



Εικόνα 14. Α) Δημιουργία διηλεκτρικής έντασης στην μεμβράνη για τον εντοπισμό των βάσεων. Β) Γράφημα βάσεων αλληλουχίας από αλληλούχιση με Nanopore. C)

Γράφημα υποθετικής ανάγνωσης αλληλουχίας από αλληλούχιση με Nanopore. (Maitra et al., 2012). http://www.nature.com/scientificamerican/journal/v294/n1/box/scientificamerican0106-46_BX4.html]

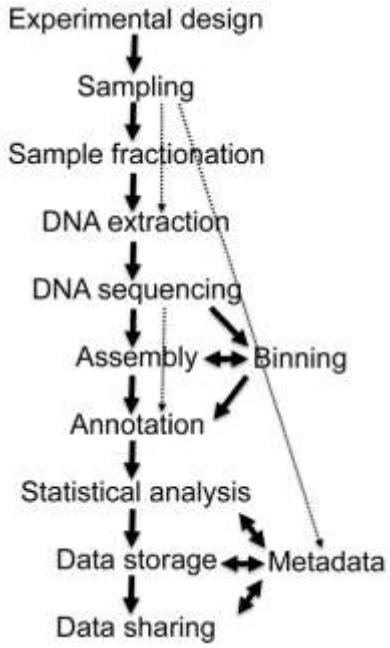
Τέλος, υπάρχουν πολλά αναπάντητα ερωτήματα για την τεχνική αλληλούχισης Graphene nanopore, όσον αφορά τη χημεία που χρησιμοποιεί και για το πόσο ακριβής (πιστή) είναι η μέθοδος αυτή.

1.2 Μεταγονιδιωματική Ανάλυση

Αφού γίνει η απομόνωση του γενετικού υλικού, θα ακολουθήσει η αλληλούχιση των θραυσμάτων DNA και στη συνέχεια τα εκατομύρια διαβασμένα κομμάτια θα πρέπει να αναλυθούν με μεθόδους βιοπληροφορικής.

Πιο συγκεκριμένα, τα βασικά βήματα που ακολουθώνται σε μία ανάλυση μεταγονιδιώματος είναι :

1. επεξεργασία του βιολογικού δείγματος
2. αλληλούχιση
3. ποιοτικός έλεγχος και φιλτράρισμα δεδομένων αλληλούχισης
4. συναρμολόγηση (assembly)
5. binning
6. σχολιασμός (annotation)
7. στατιστική ανάλυση



Εικόνα 15. Διάγραμμα ενός τυπικού project μεταγονιδιώματος. Τα διακεκομμένα βέλη δείχνουν τα βήματα που μπορούν να παραλειφθούν *(Thomas et al., 2012)*

1.2.1 Επεξεργασία του δείγματος

Το ενδιαίτημα από το οποίο παίρνουμε το δείγμα έχει σημαντική επίδραση στην μετέπειτα ανάλυση. Ενδιαιτήματα με λίγα μικροβιακά είδη ή με άνισο πληθυσμό από λίγα κυρίαρχα είδη είναι πιο υποσχόμενοι στόχοι σε σχέση με τα ενδιαιτήματα που έχουν πολλά είδη ίσης αφθονίας. Ωστόσο, πιο σημαντικό από τον απόλυτο αριθμό των ειδών είναι το επίπεδο της γονιδιακής συνοχής. Ακόμη και φαινομενικά ιδανικά ενδιαιτήματα με σταθερή σύνθεση λίγων κυρίαρχων ειδών μπορεί να είναι δύσκολο να συναρμολογηθούν όταν οι εξελικτικές προσαρμογές έχουν οδηγήσει σε μεγάλα παν-γονιδιώματα και επομένως, σε ένα χαμηλό επίπεδο κλωνικότητας του πληθυσμού. Σε αντίθεση, φαινομενικά ακατάλληλα ενδιαιτήματα που φιλοξενούν μια πληθώρα ειδών με δυναμικά μεταβαλλόμενες συνθέσεις μπορεί να δώσουν καλή συναρμολόγιση, όταν τα είδη που ευδοκιμούν και κυριαρχούν είναι σε μεγάλο βαθμό κλωνικά (Teeling and Glöckner, 2012).

Το DNA που εξάγεται πρέπει να είναι αντιπροσωπευτικό όλων των κυττάρων που υπάρχουν στο δείγμα και πρέπει να ληφθούν επαρκείς ποσότητες υψηλής ποιότητας νουκλεϊκών οξέων για την παραγωγή βιβλιοθήκης και μετέπειτα για την αλληλούχιση. Η επεξεργασία απαιτεί ειδικά πρωτόκολλα για κάθε τύπο δείγματος και

είναι διαθέσιμες διάφορες αξιόπιστες μέθοδοι για την εξαγωγή του DNA. Αν η κοινότητα στόχος συνδέεται με έναν ξενιστή (π.χ. ένα ασπόνδυλο ή φυτό), πρέπει να εξασφαλιστεί ότι λαμβάνεται έστω και ελάχιστο DNA του ξενιστή. Ο φυσικός διαχωρισμός και η απομόνωση των κυττάρων από τα δείγματα θα μπορούσαν επίσης να είναι σημαντικά για τη μεγιστοποίηση της απόδοσης του DNA ή την αποφυγή της συν-εξαγωγής των ενζυματικών αναστολέων που θα μπορούσαν να παρεμβαίνουν με μετέπειτα επεξεργασία (Thomas et al., 2012).

Η παραγωγή βιβλιοθήκης για τις περισσότερες τεχνολογίες αλληλούχισης απαιτεί υψηλές ποσότητες νανογραμμαρίων ή μικρογραμμαρίων DNA, και ως εκ τούτου μπορεί να απαιτηθεί η ενίσχυση του αρχικού υλικού (π.χ. Multiple displacement amplification (MDA)) (Lasken, 2009).

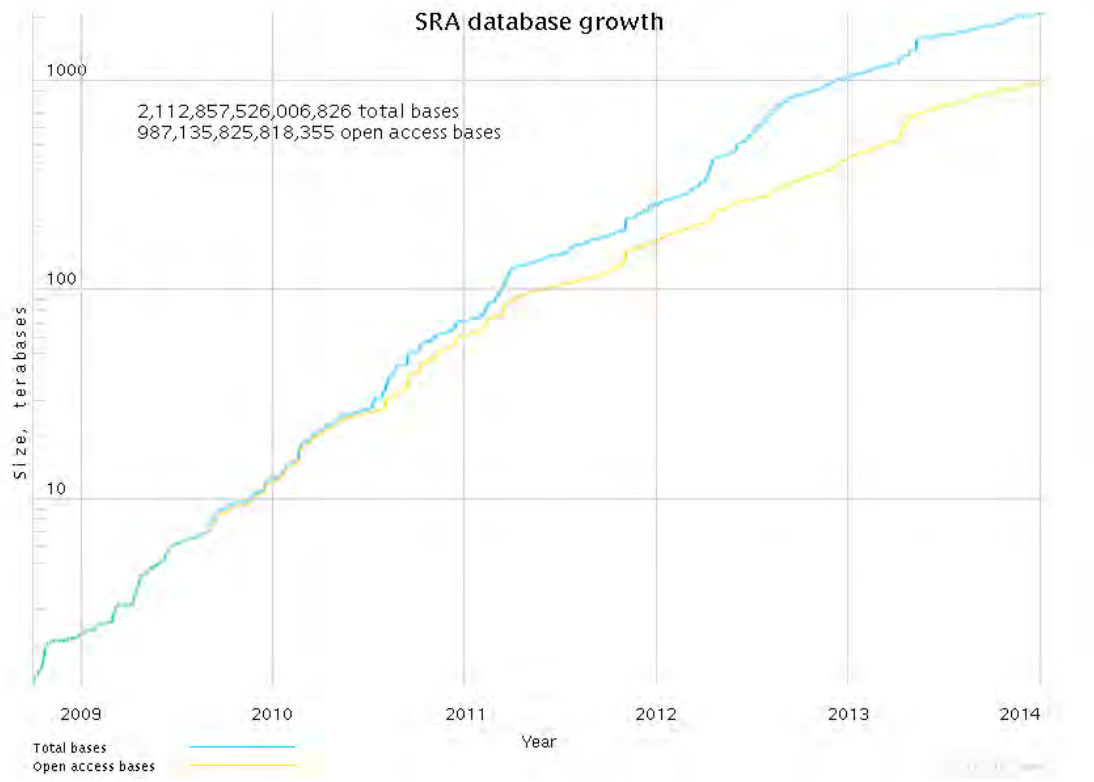
1.2.2 Αλληλούχιση

Αρχικά, η 454/Roche Pyrosequencing ήταν πιο διαδεδομένη, διότι δημιουργούσε σημαντικά μεγαλύτερα reads από ό, τι οι ανταγωνιστικές πλατφόρμες. Εν τω μεταξύ, σε εργασίες μεταγονιδιώματος μεγάλης κλίμακας γινόταν και συνεχίζει να γίνεται αυξημένη χρήση της Illumina και σε μικρότερο βαθμό, της πλατφόρμας SOLID. Το χαμηλότερο κόστος της τεχνολογίας Illumina (~ 50 USD ανά Gbp) και η πρόσφατη επιτυχία της εφαρμογής της στη μεταγονιδιωματική και ακόμη η παραγωγή των προσχεδίων των γονιδιωμάτων από ένα σύνθετο σύνολο δεδομένων (Hess et al., 2011) (Qin et al., 2010), την καθιστούν σήμερα μια όλο και πιο δημοφιλή επιλογή. Από αυτή λοιπόν την τεχνολογία αλληλούχισης έχουμε λάβει τα αλληλουχημένα δεδομένα μας για την παρούσα εργασία. Μένει να δούμε τι αντίκτυπο θα έχουν στο μεταγονιδιωματικό τομέα οι νεότερες πλατφόρμες αλληλούχισης, όπως η Ion Torrent Proton και η Oxford Nanopore Technologies.

1.2.3 Αποθήκευση δεδομένων αλληλούχισης σε ελεύθερα διαθέσιμες βάσεις δεδομένων

Τα δεδομένα αλληλούχισης από μια πλατφόρμα συνήθως αποθηκεύονται στη βάση δεδομένων Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement) του National Center for Biotechnology Information (NCBI). Η SRA αποθηκεύει δεδομένα από τεχνολογίες NGS συμπεριλαμβανομένων των 454, IonTorrent, Illumina, SOLiD,

Helicos and ολόκληρων γονιδιωμάτων. Η SRA αποθηκεύει τώρα πια και πληροφορίες στοίχισης με τη μορφή των reads τοποθετημένων σε μία αλληλουχία αναφοράς. Η SRA αποτελεί πρωτεύον αρχείο του National Institutes of Health (NIH) με δεδομένα αλληλούχισης υψηλής απόδοσης και είναι μέρος της International Partnership of archives (INSDC) στο NCBI, του European Bioinformatics Institute (EBI)και της DNA Database of Japan (DDJ). Τα τελευταία χρόνια παρατηρείται ανάπτυξη της SRA καθώς όλο και περισσότερα δεδομένα αποθηκεύονται.



Εικόνα 16. http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement Ανάπτυξη της βάσης δεδομένων SRA.

Η αποθήκευση των δεδομένων στην SRA γίνεται με τέτοιο τρόπο ούτως ώστε να αναφέρονται πληροφορίες σχετικά με τη μελέτη, την προέλευση του εξεταζόμενου δείγματος, το πείραμα, την πλατφόρμα που χρησιμοποιήθηκε, την ανάλυση που έγινε, καθώς και τη χρονική περίοδο που υποβλήθηκαν τα δεδομένα (Kodama et al., 2012).

1.2.4 Ποιοτικός έλεγχος και φιλτράρισμα δεδομένων αλληλούχισης

Το πρώτο βήμα της επεξεργασίας των δεδομένων, είναι ο ποιοτικός έλεγχος

αλληλουχιών από αρχεία σε format FastQ:

Τα αρχεία FastQ έχουν χαρακτηριστική μορφή. Αποτελούνται από τέσσερις γραμμές για το κάθε read.

#	FASTQ Data
1	@SRR032209.2000 length=36
2	GTTGTGGCTGAGATGGGATGTAAACTTGANGANANN
3	+SRR032209.2000 length=36
4	B=A@?@BBB<285:<?8%3;#########!##!#!!

Εικόνα 17. Παράδειγμα ενός read σε FastQ μορφή

Η πρώτη γραμμή αρχίζει πάντοτε με το σύμβολο **@** και προσδιορίζει το όνομα του read (εικόνα 17). Πολλές φορές (όπως στην περίπτωση αρχείων από Illumina) μπορεί να αναφέρονται πληροφορίες σχετικά με την θέση του read στο flow cell. Στην δεύτερη γραμμή εμφανίζεται η αλληλουχία του read. Δηλαδή A,T,G και C. Η εμφάνιση του γράμματος N δηλώνει ότι η βάση δεν μπόρεσε να διαβαστεί. Η τρίτη γραμμή περιέχει μόνο το σύμβολο ¨+¨ ή άλλοτε μπορεί και να συνοδεύεται από το όνομα του read. Τέλος, η ποιότητα της κάθε βάσης του read εμφανίζεται με κωδικοποιημένη μορφή (ASCII) στην τελευταία γραμμή. Ουσιαστικά στην τελευταία γραμμη, τα σύμβολα αντιστοιχούν σε τιμές Q, για την κάθε μια βάση που αλληλουχίθηκε.

Το ASCII (American Standard Code for Information Interchange) είναι μία μορφή κωδικοποίησης κειμένου με την μορφή χαρακτήρων της αγγλικής αλφαβήτου (εικόνα 18).

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Εικόνα 18. Σύστημα κωδικοποίησης ASCII (http://en.wikipedia.org/wiki/ASCII)

Πρότυπο Sanger (Standard Sanger) ή αλλιώς βαθμολογία ποιότητας PHRED (quality score) όπως φαίνεται:

$$Q = -10 \log_{10} P$$

Ο βαθμός - τιμή ποιότητας Q (Quality ή Q - score) είναι μια ακέραια τιμή που προκύπτει από την πιθανότητα να έχει γίνει λάθος στην αλληλούχιση μιας συγκεκριμένης βάσης. Αν p = πιθανότητα να έχει γίνει λάθος στην αλληλούχιση της συγκεκριμένης βάσης, τότε:

Q=-10log10(p)

Q=30 -> p=0.001 (πολύ καλής ποιότητας αλληλούχιση)
Q=13 -> p=0.05

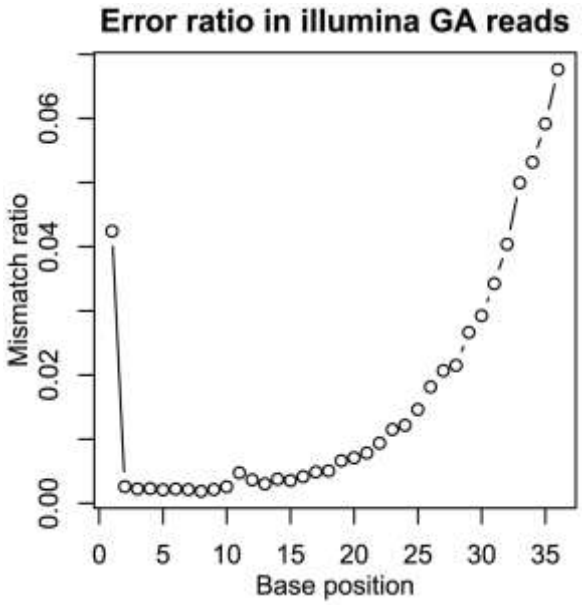
Το PHRED ήταν το πρώτο πρόγραμμα το οποίο ανέπτυξε ακριβή και ισχυρή ποιότητα βαθμολόγισης για την κάθε βάση. Έχει τη δυνατότητα υπολογισμού εξαιρετικά υψηλής ακρίβειας αποτελεσμάτων, που συνδέονται λογαριθμικά με τις πιθανότητες λάθους. Η πιο σημαντική χρήση του PHRED score είναι ο αυτόματος προσδιορισμός ακριβείας και ποιότητας των αλληλουχιών. Μπορεί, επίσης να χρησιμοποιηθεί για να εκτιμηθεί εάν οι διαφορές μεταξύ των δύο επικαλυπτόμενων ακολουθιών είναι πιο πιθανό να προκύψουν από τυχαία σφάλματα ή από διάφορα αντίγραφα μιας επαναλαμβανόμενης αλληλουχίας. Το Q20 υποδηλώνει ότι η πιθανότητα η βάση να είναι λανθασμένη είναι 0.01 ενώ το Q30 είναι 0.001 (πίνακς 2). Όπως είναι φανερό, όσο μεγαλύτερο είναι το Quality score (π.χ Q30 πολύ καλής ποιότητας αλληλούχιση) τόσο μεγαλύτερη ακρίβεια έχει η βάση και άρα τόσο μικρότερη είναι η πιθανότητα λάθους.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Πίνακας 2. Phred quality scores (http://en.wikipedia.org/wiki/Phred_quality_score)

Η αξιολόγιση της ποιότητας του κάθε read είναι πολύ σημαντική διεργασία καθώς υπάρχει το ενδεχόμενο λάθους ανάγνωσης μίας ή περισσότερων βάσεων εξαιτίας συστηματικού λάθους, που μπορεί να έχει είτε η τεχνολογία της αλληλούχισης που χρησιμοποιήθηκε είτε η ποιότητα της ίδιας της αλληλουχίας. Στην εικόνα 19 φαίνεται η κατανομή λάθους ανάγνωσης βάσεων σε Illumina reads.



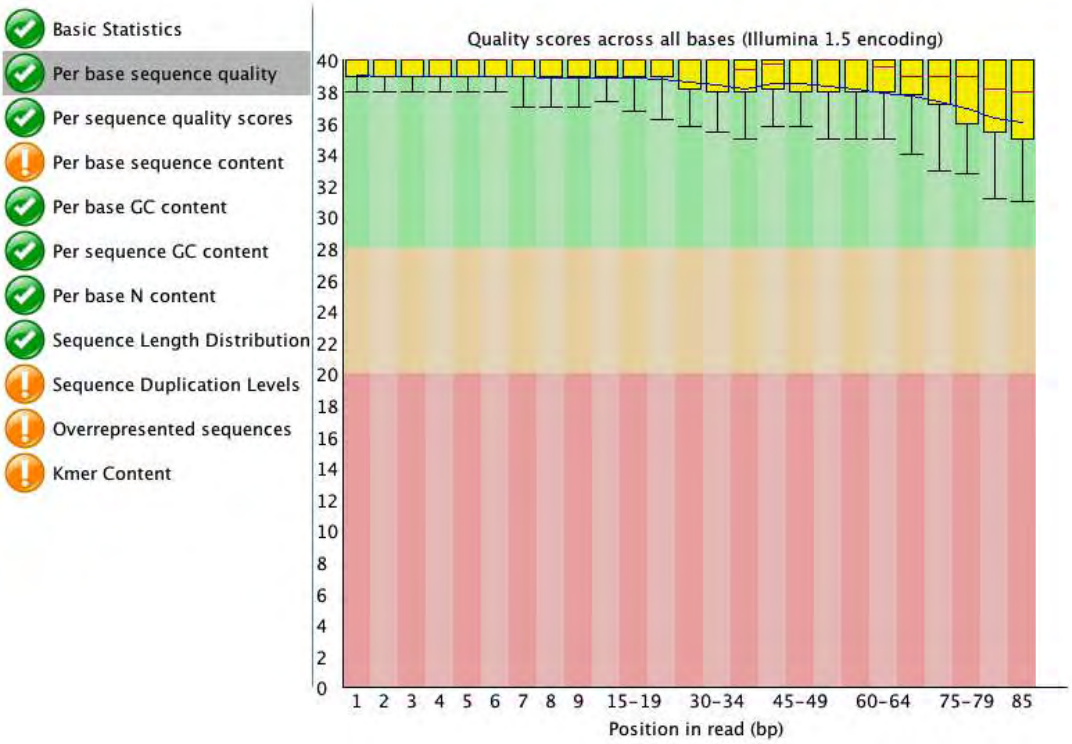
Εικόνα 19. Κατανομή λάθους ανάγνωσης βάσεων σε Illumina reads. Το πρόβλημα εντοπίζεται στη συσσώρευση λαθών κατά την ενσωμάτωση φθοριζόντων dNTPs. Ο βαθμός λάθους στα GA reads εξαρτάται από τη θέση κάθε βάσης στο read. Ο βαθμός του mismatch ανάμεσα στα mapped reads και στην αλληλουχία αναφοράς σε σχέση με τον ολικό αριθμό των mapped reads απεικονίζεται σε διάγραμμα έναντι της θέσης της κάθε βάσης στα reads. Ο βαθμός του mismatch αυξάνεται μαζί με τη θέση της βάσης υποδεικνύοντας τη μείωση της ακρίβειας των base calls.

(http://openi.nlm.nih.gov/detailedresult.php?img=3096631_pone.0019534.g001&req=4)

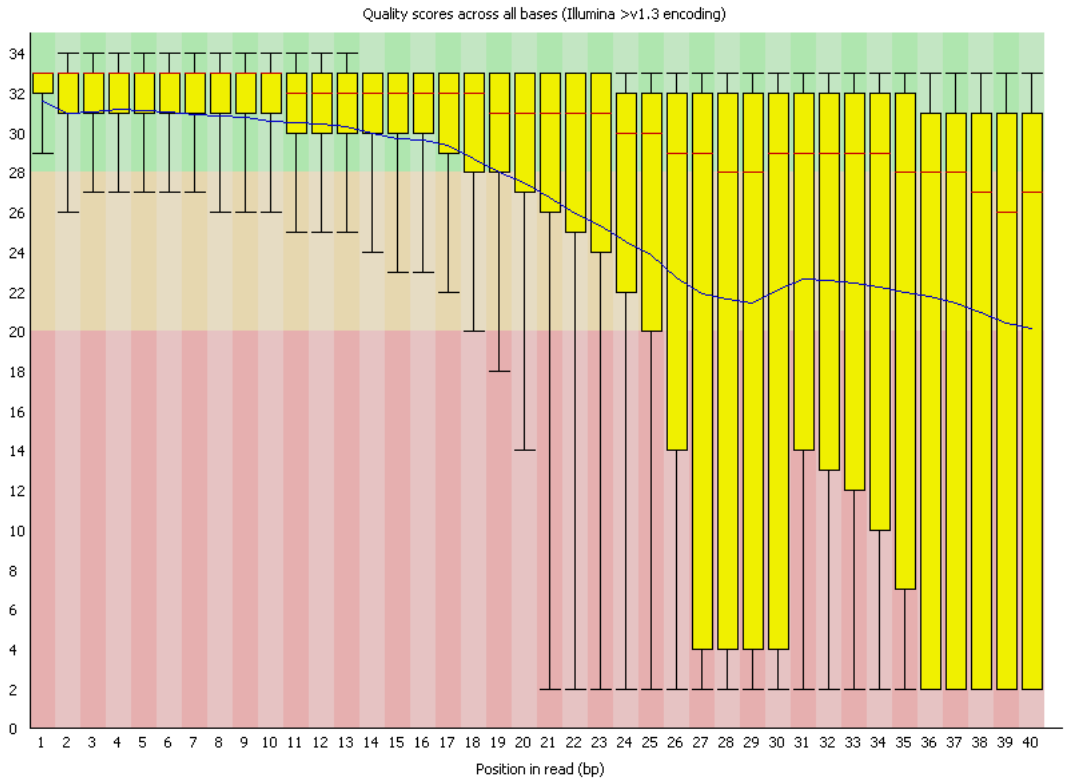
1.2.5 Ποιοτικός έλεγχος αρχείων FastQ

Κάποια από τα προγράμματα ελέγχου ποιότητας των αλληλουχημένων reads είναι τα: Reaper, FastQC, Bcbio-nextgen., Chipster, GeneProf και Biopieces. Θα πρέπει να σημειωθεί ότι πολλά προγράμματα για την λειτουργία τους στηρίζονται στη γνώση και χρήση δυναμικού προγραμματισμού όπως της Perl, Biopython, Java ή C++. Πριν την επιλογή του προγράμματος θα πρέπει να εξακριβωθούν δύο παράμετροι. Η πρώτη είναι το πρόγραμμα να υποστηρίζει τα δεδομένα της συγκεκριμένης τεχνολογίας (π.χ Illumina) και η δεύτερη είναι να δέχεται αρχεία FastQ. Σε περίπτωση που το πρόγραμμα δεν δέχεται τη μορφή FastQ αλλά κάποια άλλη, υπάρχει η επιλογή μετατροπής του στην επιθυμητή μορφή με την χρήση όμως κάποιων προγραμμάτων που έχουν συγκεκριμένα αυτό το σκοπό.

Παράδειγμα υψηλής και χαμηλής ποιότητας δεδομένων με την χρήση του προγράμματος FastQC φαίνονται αντίστοιχα στις παρακάτω δύο εικόνες. Το πρόγραμμα FastQC παρέχει πληροφορίες για την ποιότητα του κάθε read, της κάθε βάσης του read, του ποσοστού GC του read, του ποσοστού των αδιάβαστων βάσεων (N), του ποσοστού των διπλασιασμένων reads και της κατανομής μήκους της αλληλουχίας.



Εικόνα 20. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Διάγραμμα απεικόνισης υψηλού βαθμού ποιότητας των βάσεων των reads



Εικόνα 21. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Διάγραμμα απεικόνισης χαμηλού βαθμού ποιότητας των βάσεων των reads

1.2.6 Φιλτράρισμα των χαμηλής ποιότητας βάσεων/reads

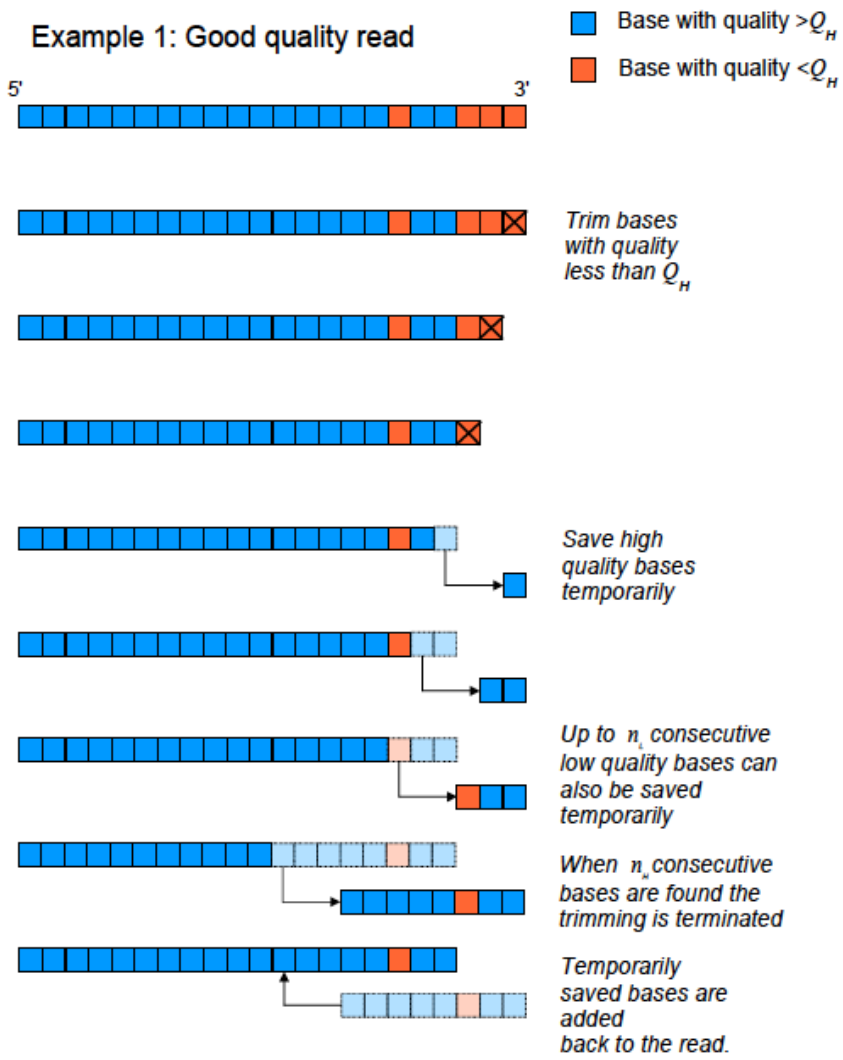
Συνεπώς, ερχόμαστε στο επόμενο βήμα της επεξεργασίας των δεδομένων αλληλούχισης, που είναι το φιλτράρισμα (trimming) των βάσεων. Όπως αναφέρθηκε και παραπάνω, έχει πραγματοποιηθεί ο ποιοτικός έλεγχος τόσο της κάθε βάσης του read αλλά και του συνόλου του κάθε read. Έτσι, με βάση αυτόν τον έλεγχο μπορεί να γίνει το φιλτράρισμα. Κατά προτίμηση τα νουκλεοτίδια που έχουν ποιότητα κάτω από Q20 διότι η πιθανότητα να είναι λάθος θα είναι μεγαλύτερη από 0.01.

Με το φιλτράρισμα μπορούν να αποκοπούν οι adapters των sequence reads καθώς επίσης και ολόκληρα κομμάτια των sequence reads που έχουν πολύ μικρό μήκος, μετά το αρχικό φιλτράρισμα.

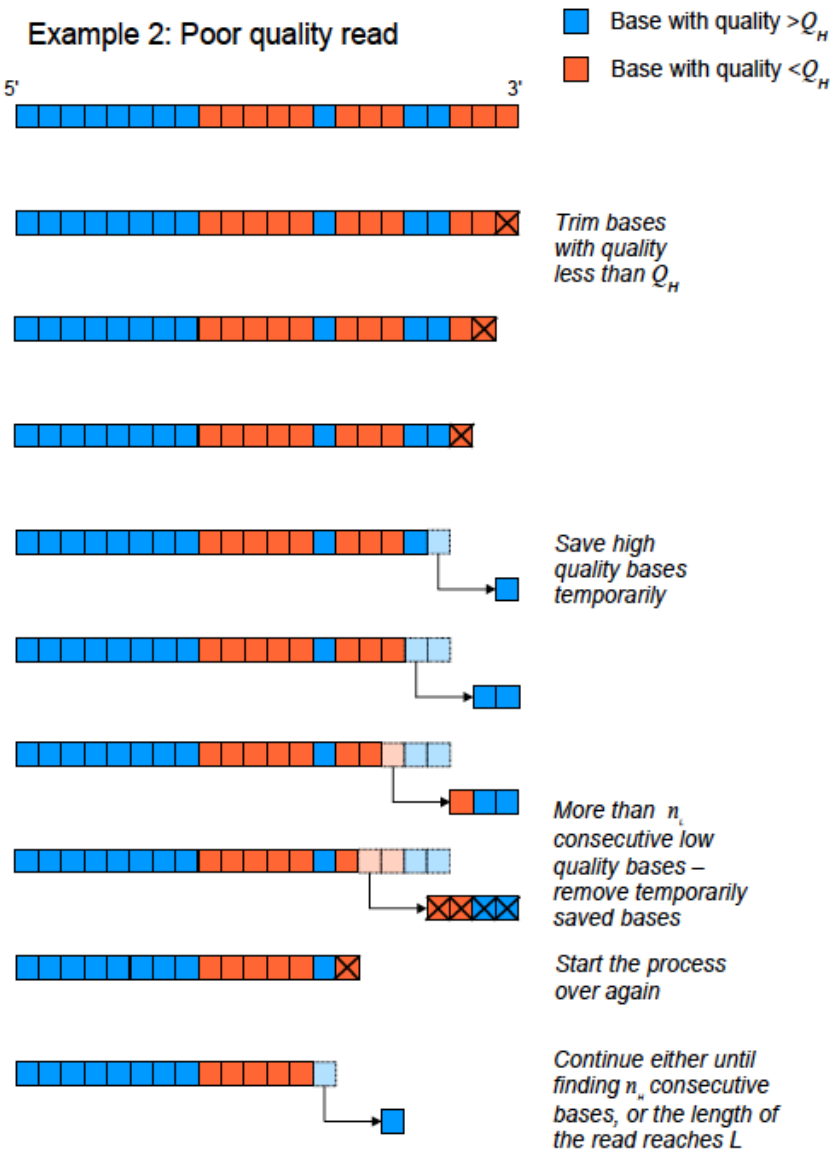
Προτεινόμενα προγράμματα για trimming των sequence reads είναι τα ακόλουθα: Condetri, NGS Toolkit, Chipster, Reaper, SeqTrim και SolexaQA. Σε αυτό το σημείο, προαιρετικά, μπορεί να γίνει αφαίρεση των PCR διπλασιασμένων reads για την καλύτερη στοίχιση των reads στην αλληλουχία αναφοράς. Το Condetri εκτός από το trimming παρέχει και αυτή την δυνατότητα δυνατότητα με το πρόγραμμα filterPCRduplicates.pl. Το trimming πραγματοποιείται σε δύο βήματα (Smeds and Künstner, 2011):

(1) Trimming των βάσεων (εικόνα 22) χαμηλής ποιότητας από το 3'-άκρο (εικόνα 23)

(2) Συνολικός έλεγχος της ποιότητας των read/pair



Εικόνα 22. Πράδειγμα trimming read υψηλής ποιότητας (Smeds and Künstner, 2011)



Εικόνα 23. Πράδειγμα trimming read χαμηλής ποιότητας (Smeds and Künstner, 2011)

Μετά το βήμα του trimming, τα quality scores των reads που παραμένουν ξαναελέγχονται. Ένα read εγκρίνεται εάν ένα συγκεκριμένο κλάσμα (frac) των βάσεων έχει quality score υψηλότερο από hq, και δεν υπάρχει καμία βάση στο read που να έχει quality score κάτω από ένα κατώτερο όριο (lq = low quality threshold). Αν τα inputs έιναι reads paired – end και τα δύο reads του ζευγαριού πρέπει να εγκριθούν για τη διατήρηση του ζεύγους. Εάν εγκριθεί μόνο το ένα από τα reads, αποθηκεύεται σε ένα πρόσθετο, unpaired αρχείο το οποίο μπορεί να χρησιμοποιηθεί ως δεδομένα single-end.

Τέλος με το condetri γίνεται το φιλτράρισμα των pcr duplicates. Κατά την αλληλούχιση με τις τεχνικές NGS, η PCR είναι συχνά απαραίτητη για την προσθήκη των adapters ή / και για να υπάρξει μια επαρκής ποσότητα θραυσμάτων για αλληλούχιση. Ωστόσο, υπάρχει κίνδυνος ενίσχυσης των ίδιων θραυσμάτων ξανά και ξανά, και στην χειρότερη των περιπτώσεων αλληλούχισης μόνο ενός μικρού μέρους του επιθυμητού υλικού. Εντοπίζοντας και αφαιρώντας αυτά τα λεγόμενα PCR-duplicates (αντίγραφα) μπορούμε να αποτρέψουμε τα λανθασμένα αποτελέσματα κατά την ανάλυση. Το πρόγραμμα παίρνει ένα ζευγάρι των αρχείων FASTQ και αφαιρεί τα περιττά αντίγραφα που μπορεί να έχουν προκύψει στο στάδιο της PCR συγκρίνοντας όλα τα ζεύγη των reads μεταξύ τους. Εάν υπάρχουν πολλά αντίγραφα ενός ζεύγους read, μόνο το ένα ζεύγος διατηρείται (αυτό με την υψηλότερη ποιότητα).

1.2.7 Συναρμολόγηση (Assembly)

Συναρμολόγηση είναι η διαδικασία της συγχώνευσης των επικαλυπτόμενων κοντών reads σε μεγαλύτερες συνεχόμενες αλληλουχίες (contigs), με βάση την αλληλοεπικάλυψη μεταξύ των reads (Kim et al., 2013). Σήμερα υπάρχουν λίγοι μεταγονιδιωματικοί συναρμολογητές, καθώς τα μεταγονιδιωματικά reads είναι πιο περίπλοκα, λόγω των μη κλωνικών ετερογενών reads που προκύπτουν από τα πολλαπλά στελέχη που διαφέρουν μόνο σε μερικές περιοχές ή σε περιοχές που έχουν γίνει ανακατατάξεις, χαμηλότερης ή άνισης κάλυψης σε ολόκληρα τα γονιδιώματα, ανταλλαγές επαναλαμβανόμενων αλληλουχιών μεταξύ συγγενικών ειδών, και γονιδιακές μετατοπίσεις καθώς μοιράζονται παρόμοιες οντότητες μεταξύ μακρινών συγγενικών οργανισμών.

Η συναρμολόγηση, όπως αναφέρθηκε, αποδίδει μεγαλύτερα γονιδιωματικά θραύσματα. Αυτά επιτρέπουν τη μελέτη της διάταξης των γονιδίων. Πολύτιμες γνώσεις για τις λειτουργίες τους μπορούν να συναχθούν από γειτονικά γονίδια, π.χ. όταν ένα γονίδιο άγνωστης λειτουργίας εμφανίζεται πάντα μαζί με ένα γονίδιο του οποίου η λειτουργία είναι γνωστή (Overbeek et al., 1999) (Osterman and Overbeek, 2003). Η συναρμολόγηση αλληλουχιών από μεταγονιδιωματικές βιβλιοθήκες μπορεί να οδηγήσει σε ένα καλό προσχέδιο ή ακόμα και σε πλήρη γονιδιώματα όταν τα στοχευόμενα είδη εμφανίζουν μικρή ενδοειδική παραλλαγή, αλλά αυτό συνήθως απαιτεί να έχει γίνει αλληλούχιση σε μεγάλο βάθος.

Παρά το γεγονός ότι η συναρμολόγηση δίνει μεγαλύτερες ακολουθίες, φέρει επίσης τον κίνδυνο της δημιουργίας χιμαιρικών contigs, ιδίως όταν τα δείγματα έχουν απομονωθεί από βιότοπους με στενά συγγενικά είδη ή με εξαιρετικά συντηρημένες αλληλουχίες οι οποίες εμφανίζονται μεταξύ των ειδών. Επιπλέον, η συναρμολόγηση διαστρεβλώνει πληροφορίες για τις ποσότητες, καθώς οι επικαλυπτόμενες αλληλουχίες από ένα είδος σε αφθονία αναγνωρίζονται οτι ανήκουν στο ίδιο γονιδίωμα και, κατά συνέπεια, ενώνονται με αυτό. Αυτό οδηγεί σε μια σχετική υποεκπροσώπηση των ακολουθιών των ειδών σε αφθονία. Μια εναλλακτική λύση για όλα τα reads που αποτελούν ένα δεδομένο contig (ή γονίδιο) είναι να γίνει μια άμεση χαρτογράφηση/στοίχιση των reads πάνω στα συναρμολογημένα contigs (Teeling and Glöckner, 2012).

Δύο στρατηγικές μπορούν να χρησιμοποιηθούν για τα μεταγονιδιωματικά δείγματα:

- συναρμολόγηση με βάση κάποια αλληλουχία αναφοράς ή αλλιώς reference-based assembly (co-assembly) και
- de novo συναρμολόγηση.

Η reference-based συναρμολόγηση μπορεί να γίνει με πακέτα λογισμικού όπως τα AMOS (http://sourceforge.net/projects/amos/), ή MIRA (http://sourceforge.net/projects/mira-assembler/). Αυτά τα πακέτα λογισμικού περιλαμβάνουν αλγόριθμους που είναι γρήγοροι και αποδοτικοί στη μνήμη και ως εκ τούτου μπορούν συχνά να πραγματοποιηθούν σε μηχανές μεγέθους φορητού υπολογιστή (laptop) σε μια-δυο ώρες. Η reference based συναρμολόγηση λειτουργεί καλά, εάν το μεταγονιδιωματικό σύνολο δεδομένων περιέχει ακολουθίες στενά συγγενικές με τα γονιδιώματα αναφοράς. Ωστόσο, διαφορές μεταξύ γονιδιωμάτων δείγματος και αναφοράς, όπως μια μεγάλη προσθήκη, απαλοιφή ή πολυμορφισμοί, μπορεί να σημαίνουν ότι η συναρμολόγηση είναι κατακερματισμένη ή ότι οι αποκλίνουσες περιοχές δεν έχουν καλυφθεί.

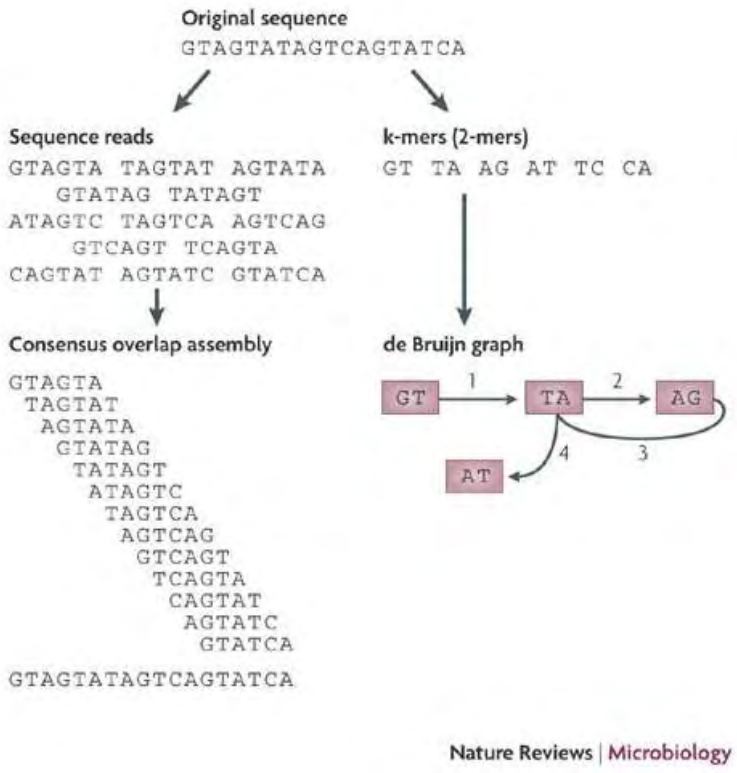
Σε πρώιμα στάδια των μεταγονιδιωματικών μελετών, εφαρμόστηκαν στη διαδικασία της συναρμολόγησης μεταγονιδιώματος προσεγγίσεις που έκαναν χαρτογράφηση με βάση αλληλουχίες αναφοράς (reference mapping approaches) (π.χ., AMOS, MIRA). Ωστόσο, λόγω της χαμηλής κάλυψης των προγραμμάτων αυτών και της πολυπλοκότητας των μεταγονιδιωματικών reads, προτιμήθηκαν οι de novo συναρμολογητές διότι είχαν ως σκοπιμότητα να αντιμετωπίσουν όλα αυτα τα

προβλήματα.

Η de novo συναρμολόγηση, την οποία και θα χρησιμοποιήσουμε, συνήθως απαιτεί μεγάλους υπολογιστικούς πόρους. Έτσι, μια ολόκληρη κατηγορία των εργαλείων συναρμολόγησης που βασίζεται στα διαγράμματα de Bruijn δημιουργήθηκε ειδικά για να χειριστεί πολύ μεγάλες ποσότητες δεδομένων (Miller et al., 2010). Οι απαιτήσεις των μηχανημάτων για τους de Bruijn συναρμολογητές Velvet (Zerbino and Birney, 2008) *και SOAP (Li et al., 2008)* εξακολουθούν να είναι σημαντικά υψηλότερες από ό, τι για τη reference- based συναρμολόγηση, και συχνά απαιτούν εκατοντάδες gigabytes μνήμης σε ένα μοναδικό μηχάνημα και ολόκληρες ημέρες για να τρέξουν.

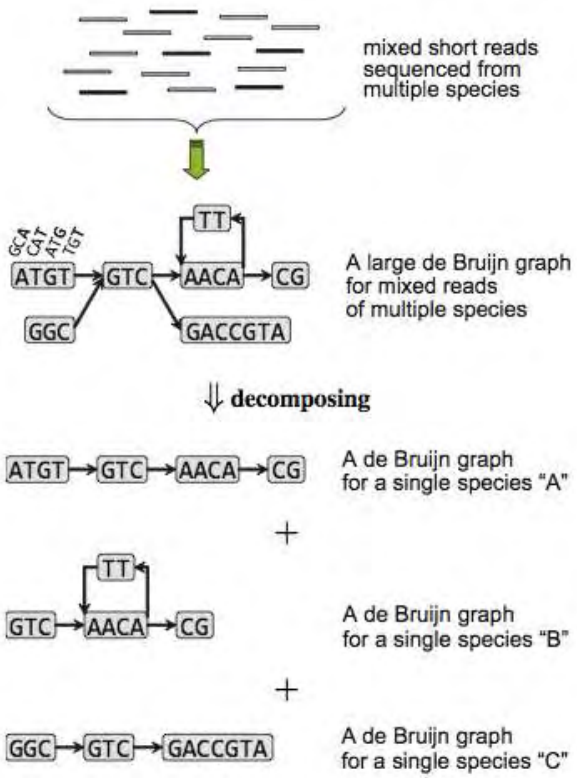
Το γεγονός ότι οι περισσότερες (αν όχι όλες) μικροβιακές κοινότητες περιλαμβάνουν σημαντικές διακυμάνσεις σε ένα στέλεχος και στο επίπεδο των ειδών καθιστά τη χρήση των αλγορίθμων συναρμολόγησης κλωνικών γονιδιωμάτων λιγότερο κατάλληλους για μεταγονιδιωματική ανάλυση. Η "κλωνική" παραδοχή στην οποία στηρίζονται πολλοί συναρμολογητές μπορεί να οδηγήσει σε καταστολή του σχηματισμού ενός contig για ορισμένα ετερογενή taxa με κάποιες παραμέτρους. Βελτιωμένοι διαθέσιμοι συναρμολογητές είναι οι : Genovo (Laserson et al., 2011), Meta-IDBA (Peng et al., 2011), MetaVelvet (Namiki et al., 2012), MAP (Lai et al., 2012), και Ray Meta (Boisvert et al., 2012). Πρόσφατα δύο συναρμολογητές τύπου de Bruijn οι MetaVelvet and Meta-IDBA (Peng et al., 2011) έχουν κυκλοφορήσει και έχουν να κάνουν με τη μη κλωνικότητα των φυσικών πληθυσμών. Και οι δύο συναρμολογητές έχουν στόχο να προσδιορίσουν σε ολόκληρο το de Bruijn γράφημα ένα υπογράφημα που αντιπροσωπεύει τα συναφή γονιδιώματα. Τα Meta-IDBA, MetaVelvet και Ray Meta αναπτύχθηκαν για να αποδίδουν καλά σε μικρού μήκους reads (π.χ. αλληλούχιση Illumina). Από την άλλη πλευρά, άλλα εργαλεία, όπως τα Genovo και MAP, αναπτύχθηκαν για μεγαλύτερα reads (π.χ. αλληλούχιση 454).

Το γράφημα de Bruijn είναι μια δομή δεδομένων για προγράμματα συναρμολόγησης και αναπαριστά συμπυκνωμένα τις αλληλοεπικαλύψεις ανάμεσα στα short reads. Αυτά τα προγράμματα που βασίζονται σε γραφήματα de Bruijn αναγνωρίζουν τις αλληλοεπικαλύψεις ανάμεσα στα reads και συνχωνεύουν τα reads για να φτιάξουν μεγαλύτερες αλληλουχίες. Το γράφημα de Bruijn χρησιμοποιήθηκε πρώτη φορά για αλληλούχιση μέσω υβριδισμού (Idury and Waterman, 1995).



Εικόνα 24. (MacLean et al., 2009) Σε ένα γράφημα de Bruijn όλα τα reads σπάζουν σε k-mers και το μονοπάτι ανάμεσα στα k-mers υπολογίζεται.

Για τη συναρμολόγηση του μεταγονιδιώματος πρέπει να ληφθεί υπόψην ότι το γράφημα de Bruijn κατασκευάζεται από τα διάφορα sequence reads που προέρχονται από τα διάφορα είδη και είναι ισοδύναμο με τα διάφορα πολλαπλά υπογραφήματα de Bruijn, κάθε ένα από τα οποία κατασκευάζεται από τα sequence reads ενός, κάθε φορά, είδους. Επιπλέον, το πολλαπλό γράφημα de Bruijn πρέπει να μπορεί να αποσυναρμολογείται σε ξεχωριστά υπογραφήματα και να μπορεί να δημιουργεί scaffolds που βασίζονται σε κάθε ένα αποσυναρμολογημένο υπογράφημα. Αυτή είναι και η αρχή των προγραμμάτων Velvet – MetaVelvet, όπως φαίνεται και στην εικόνα 25.



Εικόνα 25. (Namiki et al., 2012) Η στρατηγική του προγράματος MetaVelvet για την αποσυναρμολόγηση ενός mixed de Bruijn γραφήματος που προέρχεται από το πρόγραμμα Velvet.

1.2.7.1 Velvet – MetaVelvet

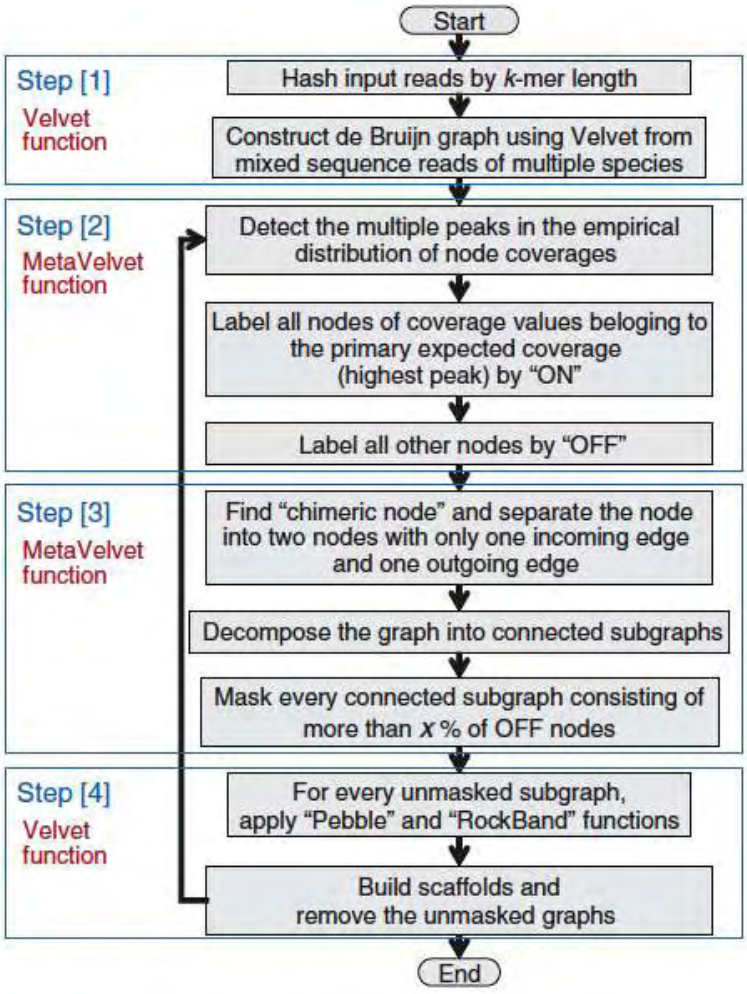
Στην πραγματικότητα πρόκειται για δύο προγράμματα και μάλιστα το MetaVelvet προήλθε έπειτα από βελτιώσεις από το Velvet. Ουσιαστικά το Velvet χρησιμεύει στη de novo συναρμολόγηση με σκοπό τη δημιουργία contigs για γονιδιώματα, ενώ το MetaVelvet για μεταγονιδιώματα. Και τα δύο προγράμματα είναι διαθέσιμα να τα κατεβάσει κανείς από τις ιστοσελίδες τους: http://www.ebi.ac.uk/~zerbino/velvet/ και http://metavelvet.dna.bio.keio.ac.jp/ .

Το Velvet είναι ένας συναρμολογητής γονιδιωμάτων για κοντά reads, γεγονός που το προτιμήσαμε στην εργασία καθώς τα reads μας είναι κοντά αφού αλληλουχίθηκαν με Illumina. Αναπτύχθηκε από τους Daniel Zerbino και Ewan Birney στο European Bioinformatics Institute (EMBL-EBI). Το Velvet παίρνει τα κοντά reads, και παράγει contigs. Οι αλγόριθμοί του χρησιμοποιούν γραφήματα de Bruijn.

Όσον αφορά τον de novo συναρμολογητή μεταγονιδιώματος MetaVelvet,

αποτελεί όπως είπαμε επέκταση του προγράμματος Velvet έχοντας δεχθεί κάποιες βελτιώσεις που το καθιστούν κατάλληλο για ευρείες αναλύσεις μεταγονιδιωμάτων. Επίσης, δίνει μεγαλύτερη τιμή N50 από ότι το Velvet κατά τη συναρμολόγηση μεταγονιδιωματικών κοντών reads. Η τιμή αυτή αντιστοιχεί σε εκείνο το μήκος contigs, ώστε το 50% του γονιδιώματος (μετά από de novo συναρμολόγηση) να εντοπίζεται σε contigs αυτού το μήκους ή μεγαλύτερου. Μεγάλη τιμή του N50 σημαίνει ότι το μεγαλύτερο μέρος του γονιδιώματος βρίσκεται σε λίγα και μεγάλα contigs. Δηλαδή, τόσο καλύτερη η συναρμολόγιση. Μικρή τιμή σημαίνει ότι το γονιδίωμα δεν έχει συναρμολογηθεί καλά. Τέλος, έχει καλύτερη γονιδιακή πρόβλεψη από το Velvet. Το MetaVelvet μπορεί να αποφεύγει αποτελεσματικά τα χιμαιρικά scaffolds, να δίνει μεγαλύτερα scaffolds, να αυξήσει τον αριθμό των προβλεπόμενων γονιδίων και να μειώνει τις κενές θέσεις που δημιουργούνται με το scaffolding.

Το διάγραμμα ροής του MetaVelvet φαίνεται στο επόμενο σχήμα.



Εικόνα 26. (Namiki et al., 2012) Διάγραμμα ροής του MetaVelvet

1.2.7.2 Trinity

Το Trinity είναι ένα άλλο πρόγραμμα για τη συναρμολόγιση των φιλτραρισμένων reads σε μεγαλύτερα κομμάτια, δηλαδή contigs, το οποίο χρησιμοποιήσαμε στην εργασία αυτή. Είναι διαθέσιμο από την ιστοσελίδα http://trinityrnaseq.sourceforge.net/ . Το πρόγραμμα αναπτύχθηκε από το Broad Institute και Hebrew University of Jerusalem, για de novo συναρμολόγηση μεταγραφωμάτων από δεδομένα RNA-seq (γεγονός που δεν μας επηρεάζει εμάς). Για να αντιμετωπίσει το μεγάλο όγκο δεδομένων συνδυάζει τρία ξεχωριστά software:

- Inchworm: συναρμολογεί τα δεδομένα σε μοναδικές αλληλουχίες μεταγράφων, συχνά παράγοντας μετάγραφα πλήρους μήκους για μια κυρίαρχη ισομορφή, αλλά στη συνέχεια δίνει μόνο τα μοναδικά τμήματα των εναλλακτικά ματισμένων μεταγράφων,
- Chrysalis: παίρνει τα contigs του Inchworm και κάνει clustering, ενώ για κάθε cluster φτιάχνει ένα γράφημα de Bruijn. Κάθε cluster αντιπροσωπεύει ολόκληρη τη μεταγραφική πολυπλοκότητα ενός γονιδίου (ή σετ γονιδίων που μοιράζονται τις ίδιες αλληλουχίες). Στη συνέχεια διαχωρίζει τα σετ σε ξεχωριστά γραφήματα και
- Butterfly: επεξεργάζεται τα μεμονωμένα γραφήματα παράλληλα, εντοπίζοντας τα μονοπάτια των reads (ή των ζευγαριών των reads) και τελικά δίνει τα μετάγραφα πλήρους μήκους για τα εναλλακτικά ματισμένα μετάγραφα καθώς επίσης χωρίζει τα μετάγραφα εκείνα που αντιστοιχούν σε παράλογα γονίδια.

1.2.8 Binning

Το binning αναφέρεται στη διαδικασία της ταξινόμησης αλληλουχιών του DNA σε ομάδες που μπορεί να αντιπροσωπεύουν ένα μεμονωμένο γονιδίωμα ή γονιδιώματα από στενά συγγενικούς οργανισμούς (ταξινομική κατάταξη των reads). Το binning είναι το πιο βασικό βήμα για τον χαρακτηρισμό των μικροβιακών κοινοτήτων κατά την ανάλυση του μεταγονιδιώματος. Ωστόσο, το ταξινομικό binning των shotgun αλληλουχιών του μεταγονιδιώματος είναι ένα δύσκολο έργο για τους ερευνητές, ειδικά όταν εργάζονται με κοντά reads που προέρχονται από NGS.

Υπάρχουν πολλοί λόγοι που κάνουν το binning μια προβληματική εργασία.
Ένας από τους λόγους είναι ότι οι τεχνολογίες NGS γενικά παράγουν κοντά reads. Όταν γίνεται shotgun αλληλούχιση σε μικροβιακές κοινότητες, το βάθος της αλληλούχισης δεν φθάνει το επίπεδο που απαιτείται για να γίνει καλή συναρμολόγηση, ακόμη και από τις σημερινές μεθόδους που πραγματοποιούν την αλληλούχιση σε σχετικά μεγάλο βάθος. Η πλειοψηφία των reads παραμένει ασυναρμολόγητη. Τα μικρά κομμάτια δημιουργούν προβλήματα στο binning, όπως έλλειψη εμπιστοσύνης στα αποτελέσματα της στοίχισης ή δυσκολίες στην πρόβλεψη των πρωτεϊνικών αλληλουχιών ορισμένων γονιδίων καθώς και έλλειψη ανάλυσης λόγω ανεπαρκών φυλογενετικών πληροφοριών (Kim et al., 2013).

Μια άλλη πρόκληση είναι το μέγεθος των δεδομένων (τόσο των reads που λαμβάνονται από τα μεταγονιδιώματα όσο και των αλληλουχιών στην βάση δεδομένων αναφοράς) που πρέπει να υποβληθούν σε επεξεργασία κατά τη διαδικασία του binning. Κάποια μικρόβια δεν έχουν αλληλουχημένο το γονιδίωμά τους ακόμη, γεγονός που παρακωλύει επίσης το binning, καθώς δεν αντιπροσωπεύονται από κάποια αλληλουχία αναφοράς στη βάση δεδομένων. Μια τελευταία πρόκληση για τους ερευνητές είναι η ποικιλία των εργαλείων του binning. Υπάρχουν δεκάδες επιλογές, που διαφέρουν και σε λογικές και σε πρακτικές πτυχές (Kim et al., 2013).

Μια άμεση προσέγγιση για το binning των μεταγονιδιωματικών shotgun reads είναι η αναζήτηση για μια παρόμοια ακολουθία, σε μια συλλογή από γνωστές αλληλουχίες, που φέρει την ταξινομική ταυτότητα της κάθε ακολουθίας. Μέθοδοι που βασίζονται στην ομοιότητα (similarity-based methods) ποικίλουν σε τουλάχιστον τρεις διαστάσεις:

1) Στην επιλογή της βάσης δεδομένων αναφοράς: ο χώρος αναζήτησης μπορεί να περιορίζεται σε ένα συγκεκριμένο γονίδιο δείκτη (π.χ., SSU rRNA για προκαρυωτικά) ή ένα μικρό αριθμό επιλεγμένων πρωτεϊνικών οικογενειών ή θα μπορούσε να είναι τόσο γενικός, όπως το σύνολο της βάσης δεδομένων nr του NCBI.

2) Στον αλγόριθμο αναζήτησης: από το BLASTN (στην περίπτωση χρήσης δείκτη γονιδίου rRNA), BLASTX, και BLASTP μέχρι ένα hidden Markov model (HMM) - έρευνα με βάση την ομολογία (αναζήτηση για πρωτεϊνικές οικογένειες).

3) Στην ταξινομική καταχώρηση από τις πληροφορίες του hit: από το αποτέλεσμα του καλύτερου χτυπήματος μεχρι τη χρήση του LCA (lowest common ancestor =

ελάχιστος κοινός πρόγονος), του τροποποιημένου LCA ή πιο περίπλοκων φυλογενετικών συμπερασμάτων. Μεταξύ των ευρέως χρησιμοποιούμενων εργαλείων, το MEGAN χρησιμοποιεί τη βάση δεδομένων nr του NCBI για την αναζήτηση και τον LCA αλγόριθμο για την καταχώρηση (Huson et al., 2011). Τα προγράματα MTR and SOrt-ITEMS τροποποίησαν τον LCA αλγόριθμο του MEGAN χρησιμοποιώντας τις ταξινομικές πληροφορίες που μοιράζονται από τα hits (MTR) ή κάνοντας ένα BLAST hit να μειώσει τα ψευδώς θετικά hits (SOrt-ITEMS) (Gori et al., 2011). Το πρόγραμμα MG-RAST εκμεταλλεύεται και τις αλληλουχίες γονιδιακού rRNA και αυτές που κωδικοποιούν για πρωτεΐνες (Glass et al., 2010). Το πρόγραμμα CARMA χρησιμοποιεί έναν αλγόριθμο όμοιο με τον LCA και δύο τρόπους αναζήτησης στις βάσεις δεδομένων γνωστών πρωτεϊνικών αλληλουχιών: στην nr βάση δεδομένων του NCBI με BLASTX και στη βάση δεδομένων Pfam (Gerlach and Stoye, 2011). Το πρόγραμμα MetaPhyler χρησιμοποιεί 31 γονίδια δείκτες για να περιορίσει το χώρο αναζήτησης *(Wu and Eisen,* 2008).

Εκτός απο τις similarity-based methods και τις παραμέτρους τους, για τις οποίες μιλήσαμε ακριβώς πριν, για το binning των μεταγονιδιωματικών reads χρησιμοποιείται και η σύνθεση με βάση την κατάταξη των αλληλουχιών του DNA με βάση την σύστασή τους (composition-based classification of DNA sequences). Αυτή η μέθοδος έχει αποδειχθεί ότι είναι πολύ χρήσιμη. Εκμεταλλεύεται τη μοναδικότητα της σύνθεσης των βάσεων που βρίσκονται σε όλα τα γονιδιώματα διαφορετικών ταξινομικών οντοτήτων. Όλες αυτές οι μέθοδοι χρησιμοποιούν τις ταξινομικές πληροφορίες της βάσης δεδομένων αναφοράς για να αντιστοιχίσουν μια ταξινομική ταυτότητα με τα reads. Ωστόσο, μπορούν να χωριστούν σε:

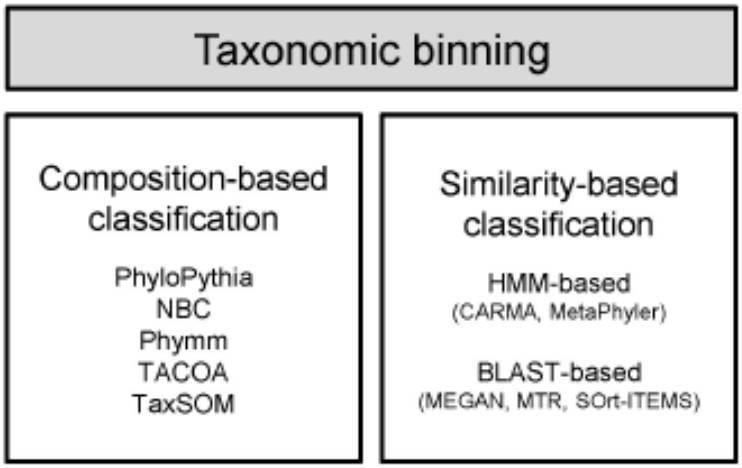
- supervised και
- unsupervised μέθοδοι,

ανάλογα με την εξάρτησή τους από το σύνολο επεξεργασίας του εκάστοτε γονιδιώματος αναφοράς κατά τη διάρκεια της αρχικής διαδικασίας μάθησης (Kim et al., 2013).

Τα πιο δημοφιλή εργαλεία που χρησιμοποιούν τη supervised μέθοδο είναι τα PhyloPythia (McHardy et al., 2007), NBC (Rosen et al., 2011), and Phymm (Brady and Salzberg, 2009). Μεταξύ των παραπάνω, τα NBC και Phymm είναι κατάλληλα

για την ταξινόμηση των κοντών reads που παράγονται από NGS αλληλουχιτές.

Τα δημοφιλή εργαλεία που χρησιμοποιούν την unsupervised μέθοδο είναι τα TACOA (Diaz et al., 2009) και TaxSOM (Weber et al., 2011). Τόσο το TaxSOM όσο και το TACOA δεν είναι κατάλληλα για ασυναρμολόγητα κοντά NGS reads.



Εικόνα 27. (Kim et al., 2013) Ταξινόμηση των προγραμμάτων που κάνουν binning

Τέλος, υπάρχουν εργαλεία που συνδυάζουν τόσο composition-based προσεγγίσεις όσο και similarity-based προσεγγίσεις. Σε αυτά τα εργαλεία περιλαμβάνονται τα PhymmBL(Brady and Salzberg, 2009) και RITA (MacDonald et al., 2012). Αν και οι προσεγγίσεις αυτές έχουν αναφερθεί πως λειτουργούν καλά με short reads, έχουν το μειονέκτημα πως καταναλώνουν πολύ χρόνο. Υπάρχει και ένας τρίτος τύπος μεθόδου binning που έχει εφαρμοστεί στο εργαλείο MetaPhlAn, που χρησιμοποιεί markers που αντιστοιχούν στο κάθε κλαδί (Segata et al., 2012).

1.2.8.1 Phylosift

Το πρόγραμμα PhyloSift εφαρμόζει μια μέθοδο για την ανάλυση της δομής μιας μικροβιακής κοινότητας χρησιμοποιώντας απευθείας τα δεδομένα της μεταγονιδιωματικής ανάλυσης, όπως έχουν προκύψει στην εργασία αυτή από το Prodigal (https://peerj.com/articles/243/). Η ανάλυση μπορεί να διαιρεθεί σε τέσσερα στάδια *(Darling et al.,* 2014):

1. Αρχικά, προσπαθεί να ταυτοποιήσει τις input αλληλουχίες με μια βάση δεδομένων γνωστών οικογενειών αλληλουχιών αναφοράς. Βάση δεδομένων που χρησιμοποιεί το PhyloSift περιλαμβάνει ένα set από 37 “elite” γονιδιακές οικογένειες

(Wu et al., 2013), που έχουν αναγνωριστεί σχεδόν παγκοσμίως και παρουσιάζονται ως single-copy. Αυτές οι οικογένειες αντιπροσωπεύουν περίπου το 1% ενός μέσου βακτηριακού γονιδιώματος. Μαζί με αυτές, η βάση δεδομένων του PhyloSift περιλαμβάνει επίσης 4 πρόσθετα sets από γονιδιακές οικογένειες:

- 16S και18S ριβοσωμικά RNA γονίδια,
- μιτοχονδριακές γονιδιακές οικογένειες,
- ειδικές- ευκαρυωτικές γονιδιακές οικογένειες και
- ϊϊκές γονιδιακές οικογένειες.

2. Σε δεύτερο βήμα, το πρόγραμμα στοιχίζει τις input αλληλουχίες με τα γονίδια αναφοράς. Αυτό το κάνει χρησιμοποιώντας το LAST (Kiełbasa et al., 2011). Το LAST είναι σε θέση να επεξεργάζεται τα δεδομένα γρήγορα και υποστηρίζει την ανίχνευση τόσο της μετατόπισης πλαισίου όσο και των αλληλουχιών οποιουδήποτε μήκους. Το LAST επίσης υποστηρίζει και τους πρωταρχικούς τρεις τύπους αναζήτησης: DNA vs. DNA, DNA vs. AA και AA vs. AA.

Το PhyloSift εφαρμόζει ένα πρόγραμμα στοίχισης hmm για να προσθέσει τις υποψήφιες αλληλουχίες σε πολλαπλές αλληλουχίες αναφοράς. Κατά την κατασκευή της βάσης δεδομένων αναφοράς του PhyloSift ένα προφίλ-HMM παράγεται από μία πολλαπλή στοίχιση των αλληλουχιών αναφοράς των γονιδιακών οικογενειών. Όταν επεξεργάζεται τις υποψήφιες αλληλουχίες το PhyloSift χρησιμοποιεί το χάρτη των HMM προφίλ που έχουν δημιουργηθεί απο τις γονιδιακές οικογένειες αναφοράς για να στοιχίσει τις υποψήφιες. Τέλος, το PhyloSift συνενώνει τις στοιχίσεις των 37 δεικτών σε μια ενιαία πολλαπλή στοίχιση ακολουθιών. Όταν μια ενιαία ακολουθία εισόδου ευθυγραμμίζεται με πολλαπλά γονίδια, η ευθυγραμμισμένη αλληλουχία γίνεται μια ενιαία γραμμή. Όλες οι άλλες αλληλουχίες αντιπροσωπεύονται σε ξεχωριστές γραμμές ευθυγράμμισης.

3. Σε αυτό το στάδιο οι input αλληλουχίες τοποθετούνται στο φυλογενετικό δέντρο των γονιδίων αναφοράς. Το PhyloSift χρησιμοποιεί τον pplacer, ο οποίος όταν τρέχει με maximum likelihood (ML, the default) αναγνωρίζει και αναφέρει ένα σύνολο από τα πιθανότερα σημεία σύνδεσης για κάθε ευθυγραμμισμένη αλληλουχία στη φυλογένεση αναφοράς, καθώς και μια αναλογία που αντιπροσωπεύει τη σχετική πιθανότητα για το επιλεγμένο σημείο προσάρτησης έναντι άλλων πιθανών, ενώ όταν τρέχει με τη Bayesian mode, ο pplacer υπολογίζει την εκ των υστέρων πιθανότητα η αλληλουχία επερώτησης να παρεκλίνει από συγκεκριμένους κλάδους του δέντρου

αναφοράς μέσω της άμεσης ενσωμάτωσης.

4. Στο σημείο αυτό το πρόγραμμα φτιάχνει τις ταξινομικές συνόψεις. Πριν από αυτό όμως, για κάθε γονιδιακή οικογένεια, η βάση δεδομένων του PhyloSift περιλαμβάνει ένα mix με την ταξινόμιση του NCBI (αυτό γίνεται με το "jplace" format των αρχείων).

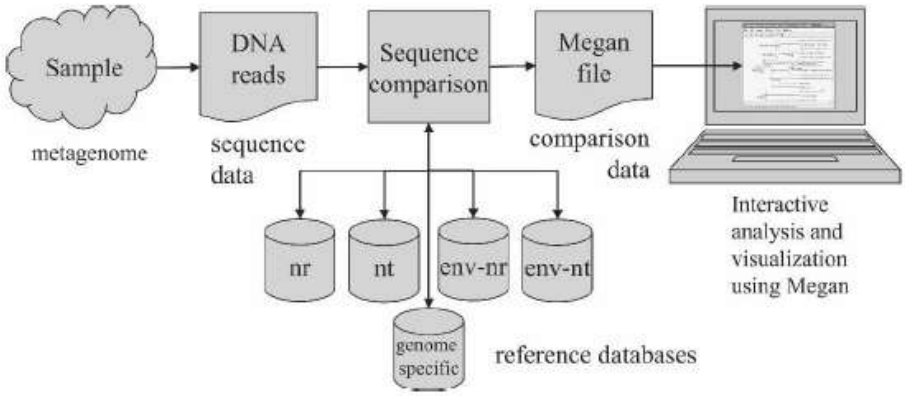
Για την οπτική παρουσίαση της ταξινομικής σύνοψης και τη διερευνητική ανάλυση των δεδομένων παράγονται τα διαγράμματα Krona, δηλαδή zoomable διαγράμματα πίτας.

1.2.8.2 Archaeopteryx

Το Archaeopteryx είναι ένα software για την οπτικοποίηση, ανάλυση και editing των φυλογενετικών δέντρων (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx).
Χρησιμοποιήθηκε γι' αυτό το σκοπό με τα δεδομένα που πήραμε από το Phylosift.

1.2.8.3 MEGAN

Το πρόγραμμα MEGAN (Metagenome Analyzer) (http://ab.inf.uni-tuebingen.de/software/megan) επιτρέπει την ανάλυση μεγάλου όγκου δεδομένων από ένα μόνο επιστήμονα. Σε ένα στάδιο προ-επεξεργασίας, το σύνολο των DNA reads (ή contigs) συγκρίνεται με τις βάσεις δεδομένων γνωστών αλληλουχιών χρησιμοποιώντας ένα εργαλείο σύγκρισης όπως το BLAST. Το MEGAN στη συνέχεια χρησιμοποιείται για να εκτιμήσει το ταξινομικό περιεχόμενο του συνόλου των δεδομένων, χρησιμοποιώντας την ταξινόμηση του NCBI για να συνοψίσει και να βάλει σε σειρά τα αποτελέσματα (Huson et al., 2011), όπως φαίνεται στην εικόνα 28.



Εικόνα 28. (Huson et al., 2011) Για ένα συγκεκριμένο δείγμα οργανισμών, γίνεται αλληλούχιση των κομματιών DNA. Τα reads που προκύπτουν συγκρίνονται με μια ή περισσότερες databases χρησιμοποιώντας το κατάλληλο πρόγραμμα BLAST. Τα δεδομένα που προκύπτουν υποβάλλονται προς επεξεργασία από το MEGAN για να γίνει ταξινομική ανάλυση.

Το πρόγραμμα χρησιμοποιεί έναν απλό αλγόριθμο που αντιστοιχίζει κάθε read στον ελάχιστο κοινό πρόγονο (LCA = lowest common ancestor) του συνόλου των taxa που χτύπησε στη σύγκριση. Ως αποτέλεσμα, οι αλληλουχίες των συγκεκριμένων ειδών εκχωρούνται σε taxa κοντά στα φύλλα του δέντρου του NCBI, ενώ οι ευρέως συντηρημένες αλληλουχίες εκχωρούνται σε taxa ανώτερης τάξης πλησιέστερα προς τη ρίζα.

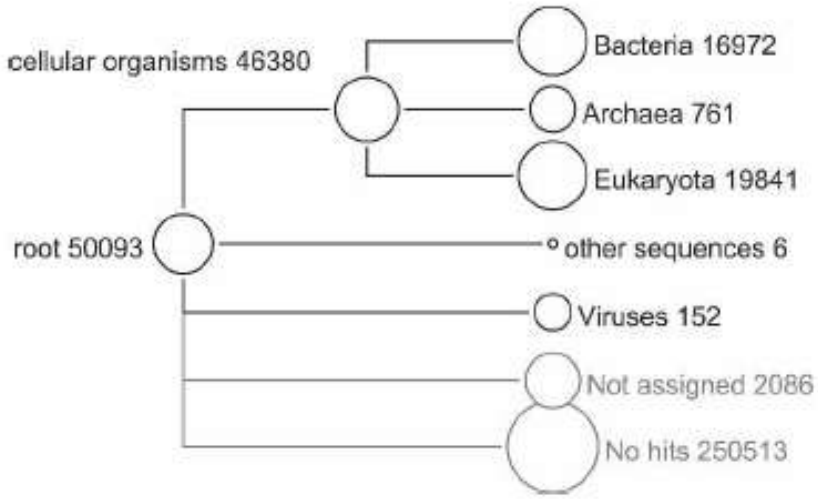
Κατά την εκκίνηση, το MEGAN φορτώνει την πλήρη ταξινόμιση του NCBI (NCBI taxonomy), που περιέχει σήμερα > 280.000 taxa, η οποία μπορεί στη συνέχεια να διερευνηθεί χρησιμοποιώντας προσαρμοσμένες tree-navigation λειτουργίες. Ωστόσο, η κύρια εφαρμογή του Megan είναι να επεξεργάζεται τα αποτελέσματα της σύγκρισης των reads έναντι μιας βάσης δεδομένων γνωστών αλληλουχιών. Το πρόγραμμα αναλύει τα αρχεία που δημιουργούνται από τα BLASTX, BLASTN, ή BLASTZ, και αποθηκεύει τα αποτελέσματα ως μια σειρά από ζευγάρια read-taxon σε ένα ειδικό για το πρόγραμμα μετα-αρχείο (με την κατάληξη .rma).

Το πρόγραμμα αντιστοιχίζει τα reads σε taxa, χρησιμοποιώντας τον αλγόριθμο LCA και στη συνέχεια εμφανίζει την επαγόμενη ταξινόμιση. Οι κόμβοι στην ταξινόμιση μπορεί να χαθούν ή να επεκταθούν. Επιπλέον, το πρόγραμμα παρέχει ένα εργαλείο αναζήτησης για συγκεκριμένα taxa, καθώς και ένα εργαλείο “επιθεώρησης”

για τα αποτελέσματα μεμονωμένων ζευγαριών από BLAST.

Κάθε κόμβος χαρακτηρίζεται από ένα taxon και έναν αριθμό απο reads που καταχωρούνται στο taxon. Το μέγεθος ενός κόμβου κλιμακώνεται λογαριθμικά αντιπροσωπεύοντας τον αριθμό των εκχωρημένων reads στον κόμβο αυτό (εικόνα 30). Το πρόγραμμα δίνει τη δυνατότητα στον ερευνητή να δει τον αριθμό των reads από έναν κόμβο και να εστιάσει στα μεμονωμένα hits του BLAST. Επιπλέον, μπορεί κανείς να επιλέξει ένα σύνολο taxa και στη συνέχεια να χρησιμοποιήσει το MEGAN να παράγει διαφορετικούς τύπους γραφημάτων για αυτά και να εστιάσει σε μεμονωμένα hits του BLAST για το κάθε read καθώς και να επιλέξει ένα σύνολο taxa και στη συνέχεια χρησιμοποιήσει το MEGAN για να παράγει διαφορετικούς τύπους γραφημάτων για αυτά.

Το αποτέλεσμα του αλγόριθμου LCA παρουσιάζεται στο χρήστη ως μερική – ταξινόμιση που επάγεται από το σύνολο των taxa που έχουν εντοπιστεί (εικόνα 29).



Εικόνα 29. (Huson et al., 2011) Ανάλυση MEGAN του mammoth data set, βασισμένη σε BLASTX σύγκριση 302,692 reads έναντι της βάσης δεδομένων NCBI-NR.

Για να εκτελεστεί μια λειτουργική ανάλυση χρησιμοποιώντας την ταξινόμηση του SEED, το MEGAN επιχειρεί να χαρτογραφήσει κάθε read σε ένα λειτουργικό ρόλο του SEED, χρησιμοποιώντας το υψηλότερο σκορ του BLAST για μια αλληλουχία πρωτεΐνης η οποία έχει γνωστό λειτουργικό ρόλο. Η κατάταξη SEED απεικονίζεται ως ένα δέντρο με ρίζα του οποίου οι εσωτερικοί κόμβοι αντιπροσωπεύουν τα διάφορα υποσυστήματα και τα φύλλα του αποτελούν τους λειτουργικούς ρόλους.

Για να εκτελέσει μια ανάλυση KEGG, το MEGAN επιχειρεί να ταιριάζει κάθε read σε κάθε KEGG Orthology (KO) accession number, χρησιμοποιώντας το καλύτερο hit σε μια αλληλουχία αναφοράς για την οποία είναι γνωστός ο KO accession number. Αυτή η πληροφορία χρησιμοποιείται έπειτα για να καταχωρηθούν τα reads σε ένζυμα και μεταβολικά μονοπάτια. Η κατάταξη KEGG αντιπροσωπεύεται από ένα δέντρο με ρίζα (με περίπου 13 000 κόμβους) και φύλλα που αντιπροσωπεύουν διαφορετικά μονοπάτια.

Για να συγκριθεί μια συλλογή από διαφορετικά σύνολα δεδομένων οπτικά, το MEGAN παρέχει μια προβολή σύγκρισης που βασίζεται σε ένα δέντρο στο οποίο κάθε κόμβος δείχνει τον αριθμό των reads που ανατίθεται για κάθε ένα από τα σύνολα δεδομένων. Για την κατασκευή μια τέτοιας άποψης, χρησιμοποιώντας το MEGAN, τα σύνολα δεδομένων πρέπει πρώτα να ανοίξουν όλα μεμονωμένα στο πρόγραμμα.

1.2.9 Σχολιασμός (Annotation)

Ο σχολιασμός των δεδομένων των μεταγενομικών αλληλουχιών πραγματοποιείται γενικά σε δύο βήματα:

- Πρώτον, αναγνωρίζονται τα χαρακτηριστικά που μας ενδιαφέρουν, δηλαδη τα γονίδια (feature prediction) και
- Δεύτερον, γίνεται προσδιορισμός των υποθετικών λειτουργιών των γονιδίων και η καταχώρηση των ταξοινομικών γειτόνων` (functional annotation).

Η εύρεση γονιδίου ή η πρόβλεψη γονιδίου είναι ένα θεμελιώδες βήμα για το σχολιασμό. Οι ανιχνευτές γονιδίων που έχουν αναπτυχθεί για ένα μόνο γονιδίωμα είναι ακατάλληλοι για μεταγονιδιωματική ανάλυση, επειδή τα δεδομένα των μεταγονιδιωμάτων αποτελούνται από ένα μίγμα των αλληλουχιών από διαφορετικούς οργανισμούς και συχνά περιλαμβάνουν κυρίως μικρές συναρμολογήσεις και μη συναρμολογημένα reads. Επιπλέον, το υψηλό ποσοστό σφαλμάτων των NGS μπορεί να οδηγήσει σε μετατοπίσεις του αναγνωστικού πλαισίου και να κάνει την πρόβλεψη γονιδίων πιο δύσκολη. Για το λόγο αυτό, ειδικά προγράμματα πρόβλεψης γονιδίων έχουν αναπτυχθεί, όπως το MetaGene *[Noguchi H, Park J, Takagi T. MetaGene: prokaryotic genefinding from environmental genome shotgun sequences. Nucleic Acids Res 2006;34:5623-5630]*, το MetaGeneAnnotator (Noguchi et al., 2008), το

Orphelia (Hoff et al., 2009), το Glimmer-MG *(Kelley et al.,* 2012)], και το MetaGenemark (Zhu et al., 2010).

1.2.9.1 Prodigal

Το πρόγραμμα Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) που χρησιμοποιήσαμε στην εργασία αυτή, είναι ένα πρόγραμμα εύρεσης μικροβίων (βακτηρίων και αρχαίων) που αναπτύχθηκε στο Oak Ridge National Laboratory and the University of Tennessee (http://prodigal.ornl.gov/). Είναι ένα πάρα πολύ γρήγορο πρόγραμμα αναγνώρισης γονιδίων και μπορεί να αναγνωρίσει έναι ολόκληρο μικροβιακό γονιδίωμα το πολύ μέσα σε 30 μόλις sec! Επίσης, είναι πολύ ακριβές σαν εργαλείο (βρίσκει σε γονίδια που δεν περιέχουν ιντρόνια το 3' άκρο). Διαθέτει ένα σύστημα βαθμολόγισης για το ριβοσωμικό binding site, που το διευκολύνει στον εντοπισμό της αρχής του μεταφραστικού site με μεγάλη ακρίβεια. Επιπλέον, το ποσοστό ψευδώς θετικών αποτελεσμάτων είναι πολύ μικρό, γεγονός που αποδίδει στο πρόγραμμα ειδικότητα.

Το πρόγραμμα δουλεύει καλά και με γονιδιώματα με υψηλό ποσοστό GC. Το Prodigal μπορεί να τρέξει και μεταγονιδιώματα ακόμη και με αλληλουχίες που ανήκουν σε άγνωστο οργανισμό. Τέλος, το Prodigal μπορεί να τρέξει σε ένα μόνο βήμα, ενώ δεν χρειάζεται να έχει πάρει πληροφορίες για τον οργανισμό από το χρήστη (το κάνει μόνο του). Ο καθένας μπορεί να χρησιμοποιήσει το πρόγραμμα κατεβάζοντάς το (http://code.google.com/p/prodigal/downloads/detail?name=prodigal.v2_60.linux).

Ο αλγόριθμος του Prodigal για το gene prediction ακολουθεί την αρχή KISS (Keep It Simple, Stupid) έχοντας καλά αποτελέσματα. Αρχικά, επειδή το πρόγραμμα λειτουργεί και για οργανισμούς με high GC content (δηλαδή λιγότερα A, T άρα λιγότερα κωδικόνια λήξης), εξετάζει τα G, C για την 1η, 2η και 3η θέση στο κωδικόνιο. Έτσι γίνεται δυναμικός προγραμματισμός σε όλο το προς ανάλυση γονιδίωμα ή μεταγονιδίωμα χρησιμοποιώντας το frame plot και υπολογίζοντας τα σκορ. Σε δεύτερο βήμα, αφού έχει μαζέψει στατιστικά στοιχεία από το προηγούμενο βήμα, δίνει σκορ για κάθε πιθανό γονίδιο (για κάθε start και stop του καθενός με log-likelihood coding function). Στη συνέχεια, αφού έχει δώσει σκορ για κάθε πιθανό υποψήφιο σε κάθε ORF, δίνει πέναλτι σε start υποψήφιους που βρίσκονται downstream από ένα start με υψηλότερο σκορ. Μετά, προστίθεται και ένας

παράγοντας μήκους που μετατρέπει όλες τις αρνητικές τιμές σε θετικές για να μπορέσουν τα ORFs να αναγνωριστούν ως πραγματικά γονίδια. Σε επόμενο βήμα, για κάθε ORF που έχει γονίδιο με σκορ μεγαλύτερο από κάποιο κατώφλι, θα καταγράψει το translation initiation site με το μεγαλύτερο σκορ. Αυτά που θα βρεθούν θα πάρουν καινούργιο σκορ. Τέλος, ένας τελικός δυναμικός προγραμματισμός πραγματοποιείται για όλα τα ζεύγη start-stop. Το σκορ κάθε πιθανού γονιδίου είναι το σύνολο των προηγούμενων σκορ.

1.2.9.2 Λειτουργικός σχολιασμός (functional annotation)

Ο λειτουργικός σχολιασμός (functional annotation) και η πρόβλεψη των μεταβολικών μονοπατιών είναι τα τελικά βήματα της μεταγονιδιωματικής ανάλυσης που επιτρέπουν το χαρακτηρισμό του λειτουργικού δυναμικού των ακαλλιέργητων μικροβίων ή μικροβιακών κοινοτήτων υπό έρευνα. Η σύνδεση των αλληλουχιών των μεταγονιδιωματικών δεδομένων με συγκεκριμένες λειτουργίες μπορεί να πραγματοποιηθεί με τη χρήση web-based workflows που προσφέρονται δικτυακά χωρίς την απαίτηση υπολογιστών υψηλών επιδόσεων. Αυτές οι ηλεκτρονικές υπηρεσίες σχολιασμού των μεταγονιδιωμάτων όπως η IMG/M (Markowitz et al., 2012), η METAREP (Goll et al., 2010), η CAMERA (Seshadri et al., 2007) παρέχουν πλατφόρμες για την πρόβλεψη των γονιδίων, την καταχώρηση σε λειτουργικές κατηγορίες, οικογένειες πρωτεϊνών και οντολογίες γονιδίων και για την εξαγωγή συμπερασμάτων των αλληλεπιδράσεων των πρωτεϊνών και των μεταβολικών οδών που αντιπροσωπεύονται στα μεταγονιδιωματικά δεδομένα.

Resources	Gene prediction	Functional category	Protein family	Gene ontology	Protein-protein interaction	Pathway and subsystems
MG-RAST	FragGeneScan	COGs, eggNOGs	FIGfams	GO	STRING	KEGG, SEED
IMG/M	FragGeneScan Genemark MetaGene	COGs	Pfam TIGRfam	GO	-	KEGG, SEED
METAREP	MetaGeneAnnotator	COGs	Pfam TIGRfam	GO	-	PRIAM
CAMERA	FragGeneScan Metagene	COGs	Pfam TIGRfam	GO	-	KEGG
MEGAN4	-	COGs	-	-	-	KEGG, SEED

STRING, Search Tool for the Retrieval of Interacting Genes/Proteins; KEGG, Kyoto encyclopedia of genes and genomes.

Εικόνα 30. (Kim et al., 2013) Οι 5 κύριοι πόροι για μεταγονιδιωματικό λειτουργικό σχολιασμό.

Η λειτουργική ανάλυση του μεταγονιδιώματος γενικά χρησιμοποιεί μια

προσέγγιση βασιζόμενη στην ομολογία και περιλαμβάνει μια αναζήτηση BLAST (Altschul et al., 1990) έναντι μιας βάσης δεδομένων, ενσωματώνοντας διάφορες επιμέρους βάσεις δεδομένων για ειδικότερη ανάλυση (Prakash and Taylor, 2012).

1.2.9.3 IMG/M

Το πρόγραμμα IMG/M (integrated microbial genomes and metagenomes) μπορεί να χρησιμοποιηθεί για τη συγκριτική ανάλυση μιας μικροβιακής κοινότητας, η οποία προέρχεται από κάποιο γονιδίωμα ή μεταγονιδίωμα, παρέχοντας ένα σύνολο εργαλείων για την ανάλυση αυτή (https://img.jgi.doe.gov/cgi-bin/m/main.cgi).
Το προς μελέτη μεταγονιδιωματικό σύνολο δεδομένων αφού αλληλουχιθεί (με τις πλατφόρμες 454 ή Illumina) μπορεί να επεξεργαστεί από το pipeline για το σχολιασμό του IMG/M. Αυτό το pipeline μπορεί να εντοπίζει γονιδιακές περιοχές που κωδικοποιούν για πρωτεΐνες (CDSs (Coding Sequence)), επαναλήψεις CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) και μη κωδικοποιούμενα RNAs. Τα RNAs προβλέπονται με εργαλεία όπως tRNAscan-SE (Lowe and Eddy, 1997) και μοντέλα HMM για rRNAs (Lagesen et al., 2007), *(Griffiths-Jones et al.,* 2005), *(Nawrocki et al., 2009).* Τα CDSs αναγνωρίζονται χρησιμοποιώντας τα: Prodigal, Metagene, MetaGenemark και FragGeneScan. Επιπλέον, οι αλληλουχίες με μήκος 100–800 bp συγκρίνονται με την πρωτεϊνική βάση δεδομένων IMG non-redundant χρησιμοποιώντας BlastX για τον εντοπισμό των CDSs που δεν βρέθηκαν με τα προηγούμενα εργαλεία.

Η ανάλυση των συνολικών μεταγονιδιωμάτων των μικροβιακών κοινοτήτων εκτιμά τη φυλογενετική σύνθεση και το λειτουργικό ή μεταβολικό δυναμικό εντός των επιμέρους μικροβιωμάτων, καθώς τα συγκρίνει μεταξύ τους. Το IMG/M υποστηρίζει μια τέτοια ανάλυση, ενσωματώνοντας το συνόλο των δεδομένων με μεμονωμένα κάθε φορά μικροβιακά γονιδιώματα που προέρχονται από το μικροβιακό γονιδίωμα του συστήματος IMG. Το πρόγραμμα IMG χρησιμοποιεί το RefSeq του NCBI ως κύρια πηγή δεδομένων των ακολουθιών για να ενσωματώνει τα μικροβιακά γονιδιώματα με ένα μεγάλο αριθμό πλασμιδίων και ιών. Παρομοίως, το IMG/M καταγράφει τα στοιχεία της κύριας ακολουθίας για τα απομονωμένα γονιδιώματα και μεταγονδιώματα, την οργάνωσή τους σε contigs, καθώς και υπολογιστικά προβλεπόμενες αλληλουχίες που κωδικοποιούν για πρωτεΐνες και γονίδια που κωδικοποιούν για RNA. Τα γονίδια που κωδικοποιούν για πρωτεΐνες σχολιάζονται με

βάση του αν υπάρχουν συντηρημένα μοτίβα και τομείς, αν αποτελούν σηματοδοτούμενα πεπτίδια ή διαμεμβρανικές έλικες, τα μονοπάτια και τις σχέσεις ορθολογίας τους, σχολιασμός που μπορεί να χρησιμεύσει ως ένδειξη των λειτουργιών τους. Αυτά τα σχόλια βασίζονται σε διαφορετικές πηγές δεδομένων, όπως τα Clusters of Orthologous Genes (COG) για τα clusters και τις λειτουργικές κατηγορίες, Pfam, TIGRfam και TIGR για κατηγορίες ανάλογα με το ρόλο που έχουν, InterPro domains και KEGG (Kyoto Encyclopedia of Genes and Genomes) για την ορθολογία και τα μονοπάτια στα οποία συμμετέχουν. Το πρόγραμμα προσφέρει και στατιστικά στοιχεία μαζί με το σχολιασμό.

1.2.10 Στατιστική ανάλυση

Με έναν απλό τρόπο, τα μεταγονιδιωματικά δεδομένα μπορούν να οργανωθούν σε πίνακες όπου οι στήλες αντιστοιχούν στα διάφορα δείγματα και οι σειρές είτε στο ταξινομικό γκρουπ είτε στη λειτουργία των γονιδίων. Τα μεταγονιδιωματικά δεδομένα όμως, συχνά περιέχουν περισσότερα είδη και γονιδιακές λειτουργίες σε σχέση με τον αριθμό των δειγμάτων, οπότε χρειάζεται να γίνουν κάποιες διορθώσεις για multiple hypothesis testing (π.χ. Bonferroni correction for t-test based analyses). Διαθέσιμα εργαλεία για στατιστική ανάλυση υπάρχουν, όπως: Το Primer-E package (http://www.primer-e.com/) που είναι ένα καθιερωμένο εργαλείο, που επιτρέπει στατιστική ανάλυση για μια σειρά πολλών μεταβλητών. Πρόσφατα, η στατιστική ανάλυση με πολυ- μεταβλητές χρησιμοποιήθηκε και από εργαλεία του διαδικτύου όπως το Metastats (White et al., 2009). Επιπλέον, το πακέτο Shotgun-FunctionalizeR παρέχει διάφορες στατιστικές λειτουργίες για την αξιολόγηση των λειτουργικών διαφορών μεταξύ των δειγμάτων, τόσο για μεμονωμένα γονίδια όσο και για ολόκληρα μονοπάτια χρησιμοποιώντας το στατιστικό πακέτο R (Kristiansson et al., 2009).

1.3 Σκοπός της παρούσας εργασίας

Ο σκοπός της εργασίας αυτής ήταν η δημιουργία ενός υπολογιστικού πρωτόκολλου, δηλαδή ενός pipeline, σε έναν Desktop PC για την ανάλυση μεταγονιδιωματικών αλληλουχιών χρησιμοποιώντας κατάλληλα και ελεύθερα προσβάσιμα προγράμματα βιοπληροφορικής.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

Η ανάκτηση, διαχείριση και ανάλυση των δεδομένων της εργασίας αυτής απαιτεί τη χρήση κατάλληλων βάσεων δεδομένων και υπολογιστικών εργαλείων.

2.1 Υπολογιστικό σύστημα

Τα χαρακτηριστικά του υπολογιστικού συστήματος για την διακπεραίωση της εργασίας αυτής είναι:

- Λογισμικό Linux – Ubuntu 12.04
- Intel® Xeon, 2XCPU E5620, Quad Core 2.40 GHz (16 threads)
- Μνήμη 96GB
- Σκληρός δίσκος 3TB

2.2 Προγραμματισμός

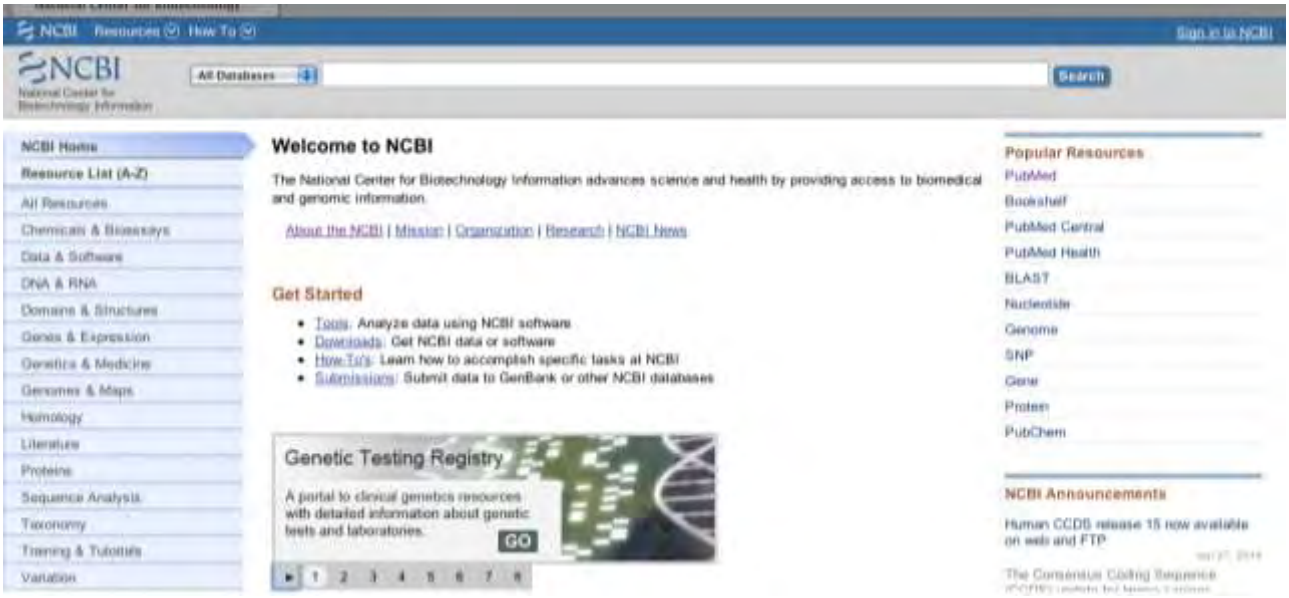
Η επεξεργασία κάποιων δεδομένων έγινε με τη γλώσα προγραμματισμού Perl. Η Perl είναι διαθέσιμη για όλα τα λειτουργικά συστήματα. Το λογισμικό που χρησιμοποιήσαμε στο εργαστήριο ήταν Linux (Ubuntu-12.04) το οποίο είχε ενσωματωμένη την Perl.

2.3 Βάσεις δεδομένων

Για την ανάκτηση και ανάλυση δεδομένων από τεχνολογία NGS χρειάστηκαν αρκετές βάσεις δεδομένων, που αναφέρονται παρακάτω:

2.3.1 National Center for Biotechnology Information (NCBI)

Το NCBI (http://www.ncbi.nlm.nih.gov/) προωθεί την επιστήμη και την υγεία, παρέχοντας πρόσβαση σε βιοϊατρικές και γενετικές πληροφορίες (εικόνα 31).



Εικόνα 31. Ιστοσελίδα εθνικού κέντρου πληροφοριών βιοτεχνολογίας-NCBI (http://www.ncbi.nlm.nih.gov/)

Στεγάζει μια σειρά από βάσεις δεδομένων όπως τη GenBank που περιλαμβάνει πληροφορίες για αλληλουχίες DNA, την PubMed, μια βιβλιογραφική βάση δεδομένων για τη βιοϊατρική βιβλιογραφία κ.α. Όλες αυτές οι βάσεις δεδομένων είναι διαθέσιμες στο διαδίκτυο μέσω της μηχανής αναζήτησης Entrez (http://www.ncbi.nlm.nih.gov/gquery/).

2.3.2 Sequence Read Archive (SRA)

Η βάση δεδομένων SRA (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement) του NCBI, όπως αναφέρθηκε και στην εισαγωγή αποθηκεύει ανεπεξέργαστα δεδομένα (raw reads) από τεχνολογίες NGS συμπεριλαμβανομένων των 454, IonTorrent, Illumina, SOLiD, Helicos and ολόκληρων γονιδιωμάτων. Η SRA αποθηκεύει τώρα πια και πληροφορίες στοίχισης με τη μορφή των reads τοποθετημένων σε μία αλληλουχία αναφοράς (εικόνα 32).

Εικόνα 32. Ιστοσελίδα SRA (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=about)

Η αποθήκευση των δεδομένων στην SRA γίνεται με τέτοιο τρόπο ουτως ώστε να είναι διακριτή η προέλευση των δεδομένων, αναφέροντας πληροφορίες σχετικά με τη μελέτη, την προέλευση του εξεταζόμενου δείγματος, το πείραμα, την πλατφόρμα που χρησιμοποιήθηκε, την ανάλυση που έγινε καθώς και τη χρονική περίοδο που υποβλήθηκαν τα δεδομένα (Kodama et al., 2011).

2.4 Πηγές μεταγονιδιωματικών δεδομένων

Τα raw reads που κατεβάσαμε ήταν τα μεταγονιδιώματα που προέρχονται από:

- Ανθρώπινα κόπρανα, στα πλαίσια της εργασίας για την ταχεία εξέλιξη του ανθρώπινου εντερικού virome (Minot et al., 2013) και
- Ενεργός λυματολάσπη, στα πλαίσια της εργασίας για τη μελέτη του πλασμιδιακού της μεταγονιδιώματος (Zhang et al., 2011).

  Και τα δύο παραπάνω κατεβαίνουν με τη μορφή που είναι αποθηκευμένα στο SRA (δηδ. με την κατάληξη .sra).

2.5 Προγράμματα βιοπληροφορικής

Σήμερα υπάρχει μια πληθώρα προγραμμάτων βιοπληροφορικής τόσο για γονιδιωματική όσο και για μεταγονιδιωματική ανάλυση. Τα τελευταία πολλές φορές

είναι βασισμένα στον τρόπο με τον οποίο λειτουργούν τα πρώτα, αλλά προφανώς πιο εξελιγμένα ώστε να ανταποκρίνονται στις ιδιαίτερες απαιτήσεις τους. Στην παρούσα εργασία, για την ανάπτυξη πρωτοκόλλου, χρησιμοποιήσαμε προγράμματα βιοπληροφορικής με ευρεία χρήση από ερευνητές και που αντιστοιχούν στα πλαίσια μεταγονιδιωματικής de novo συναρμολόγησης από δεδομένα που έχουν αλληλουχηθεί με την τεχνολογία του NGS, Illumina, η οποία παράγει short reads.

2.5.1 Condetri

Το πρόγραμμα Condetri κάνει το λεγόμενο trimming, όπως έχει αναφερθεί στην εισαγωγή, σε FASTQ reads από το 3’ άκρο τους, ούτως ώστε να εξαχθούν reads καλύτερης ποιότητας.

2.5.2 FastQC

Το πρόγραμμα FastQ χρησιμοποιήθηκε για ποιοτικό έλεγχο των sequence reads.

2.5.3 Trinity

Το trinity είναι ένα από τα δύο προγράμματα που χρησιμοποιήσαμε για τη συναρμολόγιση των φιλτραρισμένων reads σε μεγαλύτερα κομμάτια, δηλαδή contigs.

2.5.4 Velvet – MetaVelvet

Τα χρησιμοποιήσαμε και αυτά για τη συναρμολόγιση των φιλτραρισμένων reads σε μεγαλύτερα κομμάτια, δηλαδή contigs, για να δούμε αν υπάρχει κάποια διαφορά στα αποτελέσματα που θα μας δώσουν σε σχέση με το Tinity.

Οι αλγόριθμοί όπως έχουμε αναφέρει, χρησιμοποιούν γραφήματα de Bruijn.

Ο ορισμός του μήκους των k- mers είναι πολύ σημαντικό βήμα για να λειοτυργήσει σωστά το πρόγραμμα. Το εργαλείο velveth βοηθά στην κατασκευή του κατάλληλου dataset για το εργαλείο velvetg και υποδεικνύει στο σύστημα τι αντιπροσωπεύει το κάθε αρχείο. Το εργαλείο velvetg είναι η καρδιά του Velvet και φτιάχνει το γράφημα de Bruijn. Αυτό στην περίπτωση της μεταγονιδιωματικής ανάλυσης είναι ένα συνολικό γράφημα.

Το MetaVelvet, αφού ρυθμιστεί με κατάλληλο μήκος k- mer, σπάζει αυτό το γράφημα που έχει δημιουργηθεί από το σύνολο των reads σε επιμέρους υπο-γραφήματα με βάση το γονιδίωμα του κάθε είδους (το διαφορετικό είδος είναι η κάθε

διαφορετική κορυφή στα διαγράμματα κανονικής κατανομής που δημιουργούνται). Στη συνέχεια φτιάχνει contigs με βάση το δεδομένο υπογράφημα κάθε φορά.

2.5.5 Prodigal

Το πρόγραμμα Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) είναι το πρόγραμμα εύρεσης γονιδίων που χρησιμοποιήσαμε στην εργασία μας.

2.5.6 Phylosift

Το πρόγραμμα PhyloSift χρησιμοποιήθηκε για την ταξινομική ανάλυση χρησιμοποιώντας απευθείας τα δεδομένα της μεταγονιδιωματικής ανάλυσης, όπως έχουν προκύψει από το Prodigal.

Για την οπτική παρουσίαση της ταξινομικής σύνοψης και τη διερευνητική ανάλυση των δεδομένων παράγονται τα διαγράμματα Krona, δηλαδή zoomable διαγράμματα πίτας.

2.5.7 Archaeopteryx

Το Archaeopteryx χρησιμοποιήθηκε για την οπτικοποίηση, ανάλυση και editing των φυλογενετικών δέντρων (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx), με τα δεδομένα που πήραμε από το Phylosift.

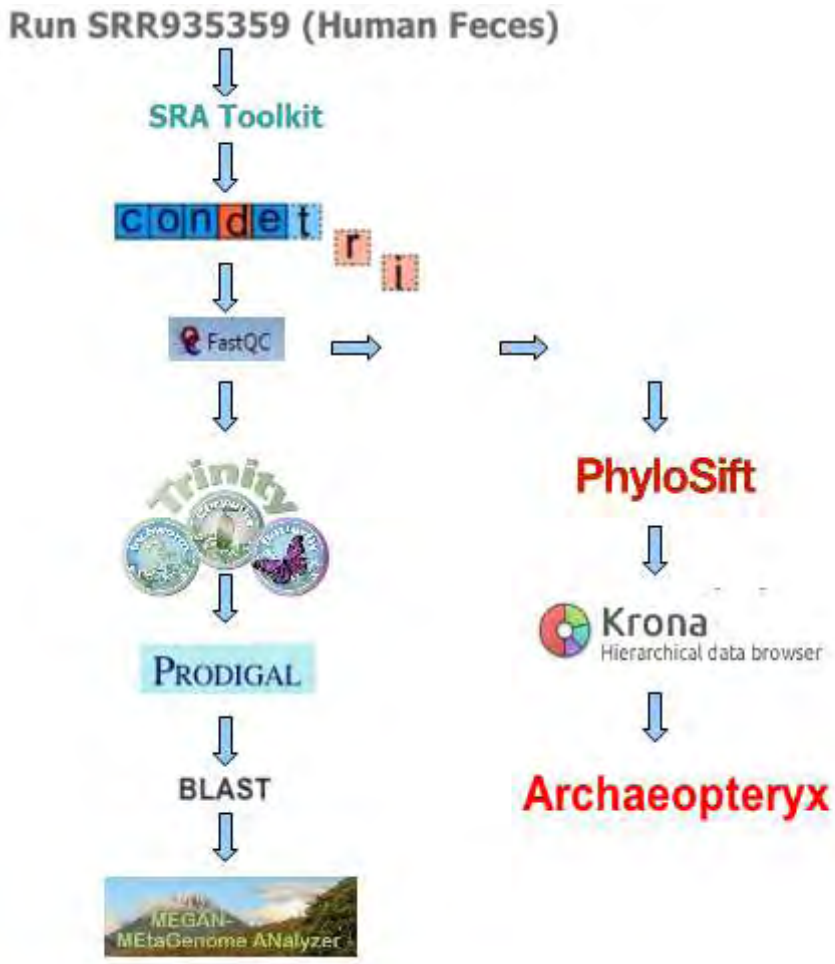
2.5.8 Blast

Blast έγινε με την database του NCBI – nt (nucleotide) και με την database του NCBI – nr (non redundant) με τις αλληλουχίες που πρέκυψαν από το Prodigal. Ήταν απαραίτητο για το πρόγραμμα MEGAN.

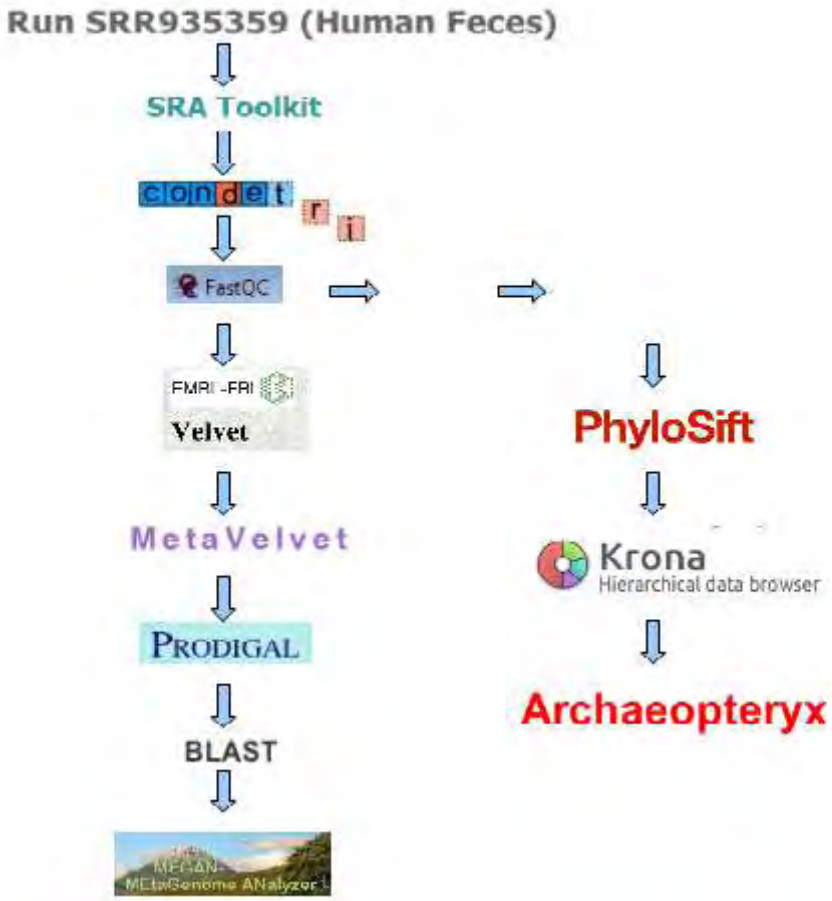
2.5.9 MEGAN

Το MEGAN χρησιμοποιήθηκε για την ταξινομική ανάλυση των δεδομένων μας, μετά από το Blast. Το πρόγραμμα αυτό επιτρέπει επίσης, μια λειτουργική ανάλυση χρησιμοποιώντας την ταξινόμηση του SEED, μια ανάλυση KEGG, επιχειρώντας και την οπτική σύγκριση μιας συλλογής από διαφορετικά σύνολα δεδομένων.

2.6 Συνοπτικά τα βήματα που ακολουθήθηκαν στην παρούσα εργασία



Εικόνα 33. Διαδικασία ανάλυσης της μεταγονιδιωματικής αλληλουχίας με το πρόγραμμα Trinity για την κατασκευή των contigs



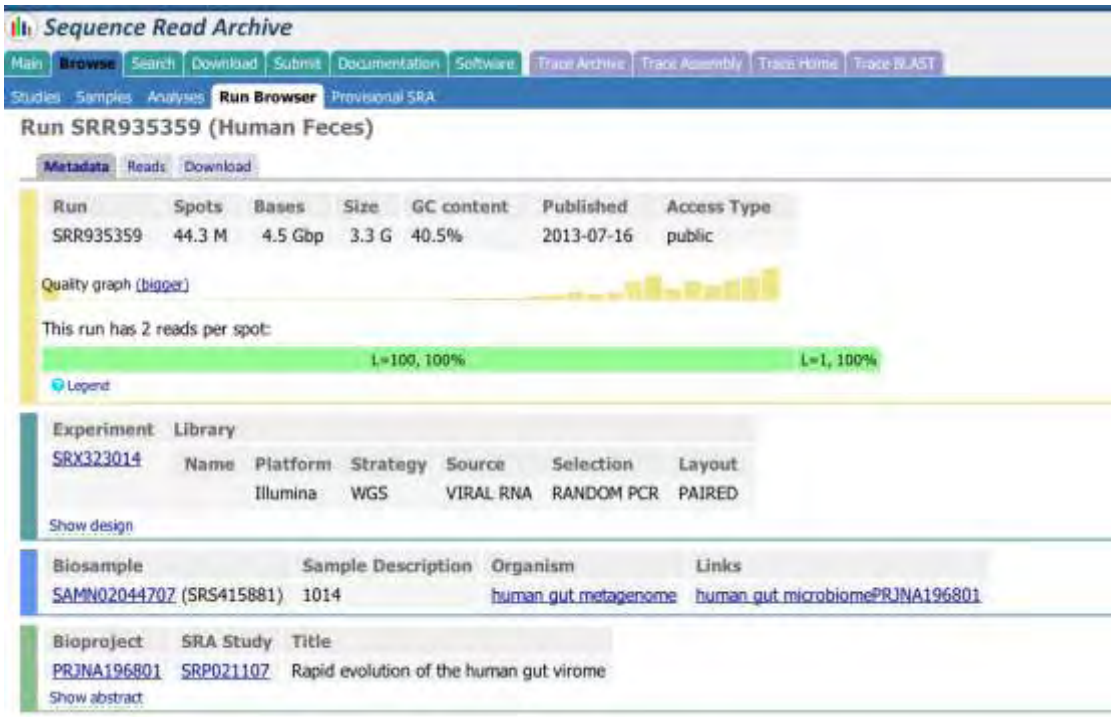
Εικόνα 34. Διαδικασία ανάλυσης της μεταγονιδιωματικής αλληλουχίας με τα προγράμματα Velvet, Metavelvet για την κατασκευή των contigs

Η ίδια διαδικασία ακολουθήθηκε και για το άλλο δείγμα, αυτό της ενεργής λυματολάσπης.

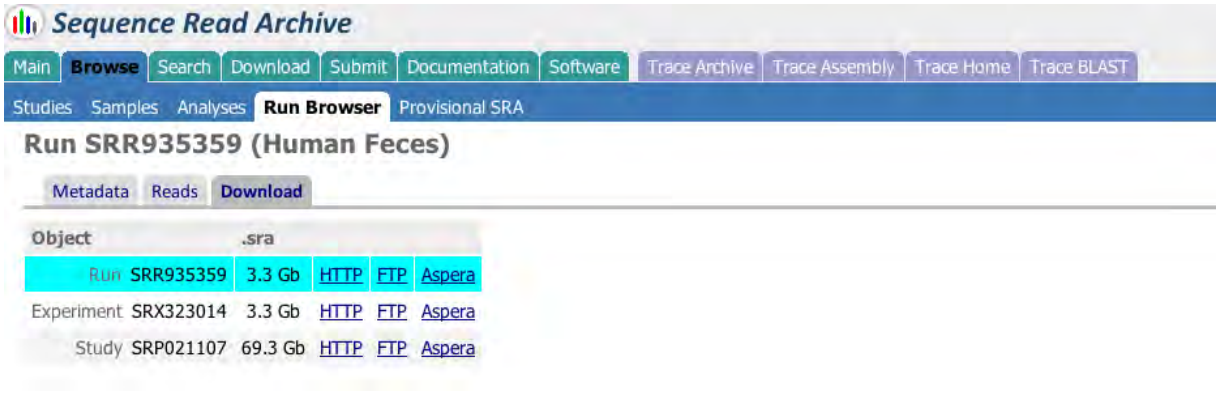
3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 SRA Toolkit

Σαν πρώτο βήμα κατεβάζουμε από το SRA το **SRR935359** (http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR935359), που είναι ανθρώπινα κόπρανα από την εργασία για την ταχεία εξέλιξη του ανθρώπινου εντερικού virome εικόνες 35, 36).



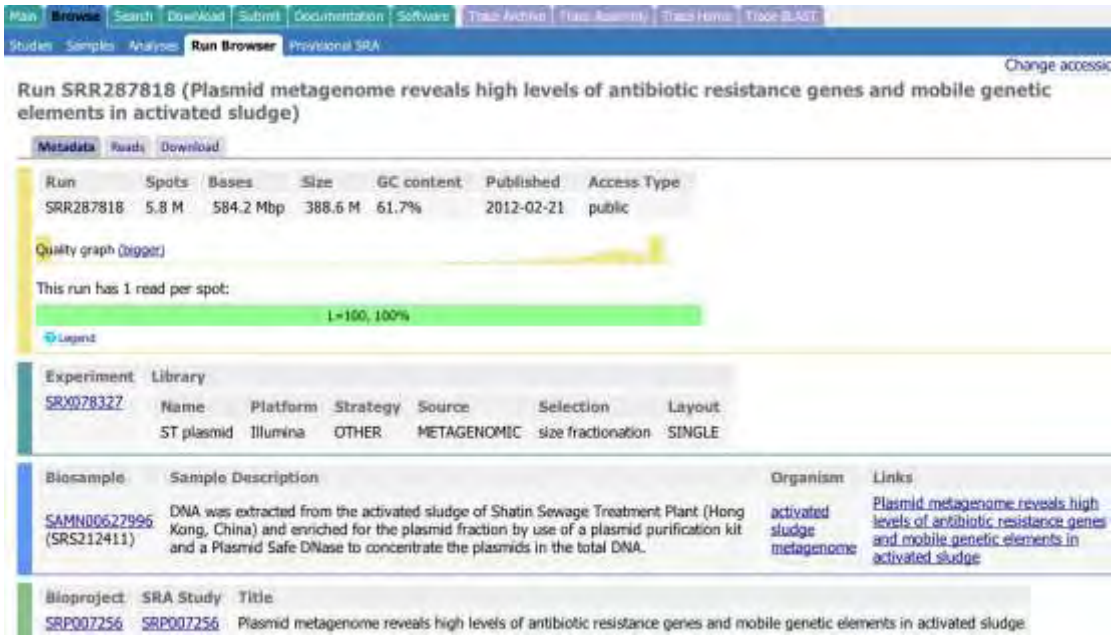
Εικόνα 35. Το SRR935359 στην ιστοσελίδα του SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR935359)



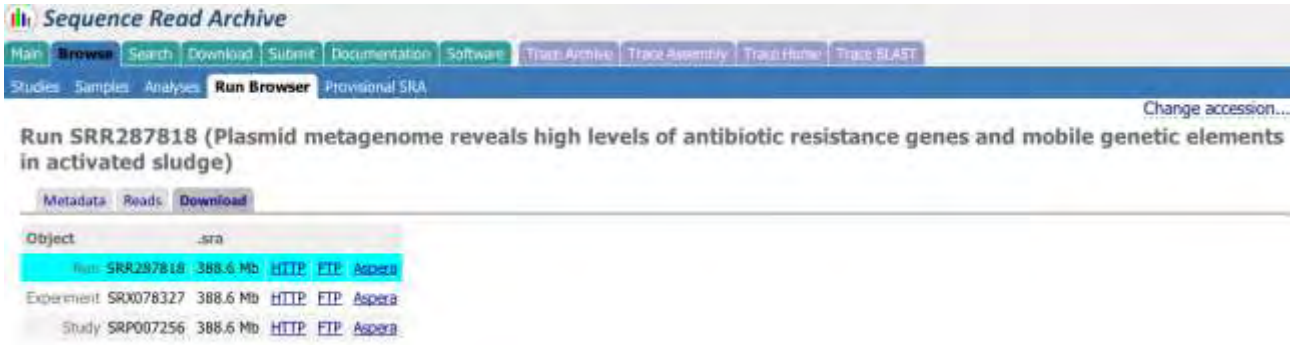
Εικόνα 36. Το SRR935359 έτοιμο για download

(http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR935359)

και το **SRR287818** (ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP007/SRP007256), που είναι η ενεργός λυματολάσπη από την εργασία για τη μελέτη του πλασμιδιακού της μεταγονιδιώματος (εικόνες 37, 38).



Εικόνα 37. Το SRR287818 στην ιστοσελίδα του SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR287818)



Εικόνα 38. Το SRR287818 έτοιμο για download (http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR287818)

Για να γίνει αυτό πρέπει πρώτα να κατεβάσουμε το SRA-toolkit (December 2 2013, version 2.3.4-2 release), το οποίο ξezipράρουμε και αποθηκεύουμε. Αρχικά,

εκτελούμε το εργαλείο διαμόρφωσης Java σύμφωνα με τις οδηγίες του http://eutils.ncbi.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std , γράφοντας την εντολή, ενώ βρισκόμαστε στο bin:

java -jar sratoolkit.jar

Αυτό το εργαλείο ρυθμίζει την περιοχή λήψης για τα ληφθέντα αρχεία. Εφόσον έχουμε ανακτήσει τα αρχεία που θέλουμε, δηδ τα SRR935359.sra και SRR287818.sra και και τα έχουμε αποθηκεύσει στο φάκελο bin, χρησιμοποιούμε την παρακάτω εντολή, ενώ πάλι βρισκόμαστε στο directory bin για να αποκτήσουν τα αρχεία μας fastq μορφή:

./fastq-dump –M 50 --split-3 SRR935359.sra

και

./fastq-dump –M 50 -- split-3 SRR287818.sra

Με το --split-3 SRR_.sra κανονικά όταν τα reads είναι paired- end δημιουργούνται στο bin δύο αρχεία με ονομασίες SRR_1.fastq και SRR_2.fastq, τα οποία έχουν χωρισμένες τις paired- end αλληλουχίες τους (δηδ.η μία αλληλουχία του ζεύγους στο ένα αρχείο σε μορφή fastq και η άλλη στο άλλο). Το –M 50 (minReadLen) είναι το φίλτρο που βάζουμε εμείς για να είναι το μήκος της κάθε ακολουθίας >= του 50 (δηδ. φιλτράρουμε με το μήκος των ακολουθιών). Στη δική μας περίπτωση παίρνουμε στο bin ένα αρχείο για το καθένα SRR935359.fastq (15.4 GB) και SRR287818.fastq (1.9 GB) γιατί όπως φαίνονται γραμμένα στα αρχεία, οι paired- end αλληλουχίες τους αναγνωρίζονται σαν ξεχωριστές αλληλουχίες (σαν single -end) και όχι σαν paired- end.

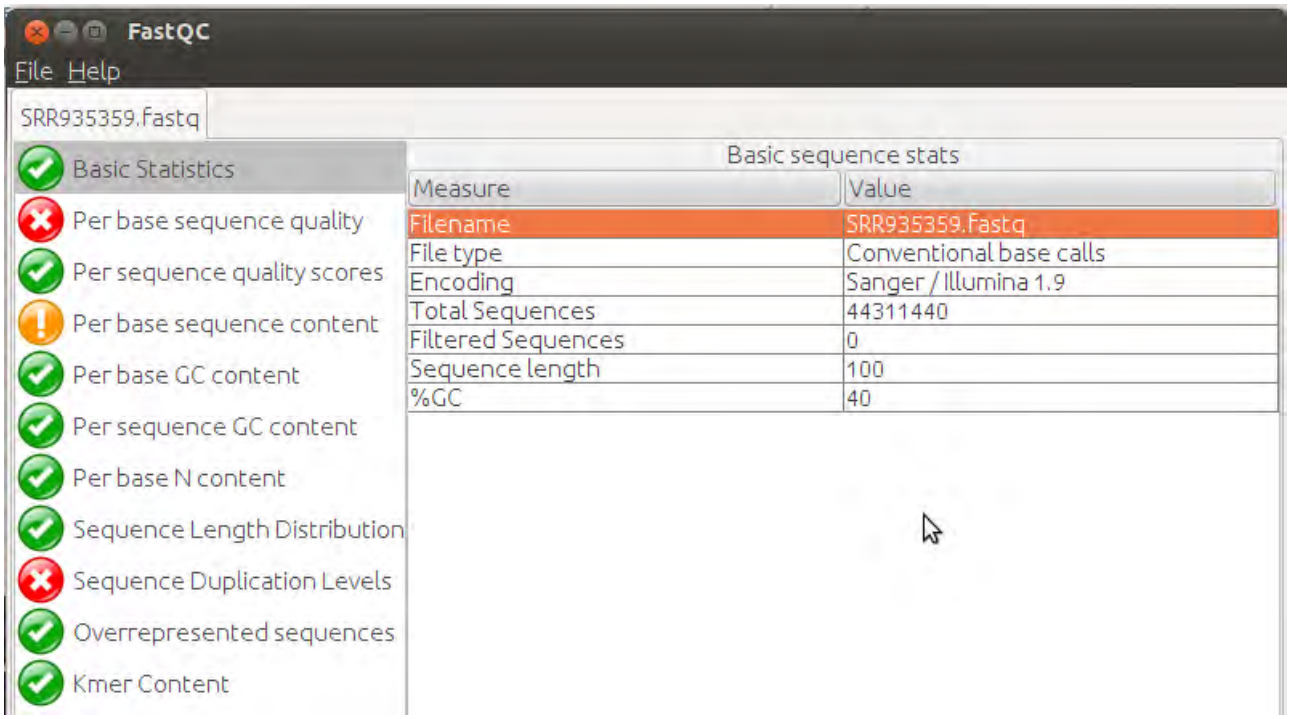
3.2 Condetri και FastQC

Κατεβάζουμε το condetri_v2.2.pl (http://code.google.com/p/condetri/downloads/detail?name=condetri_v2.2.pl), το

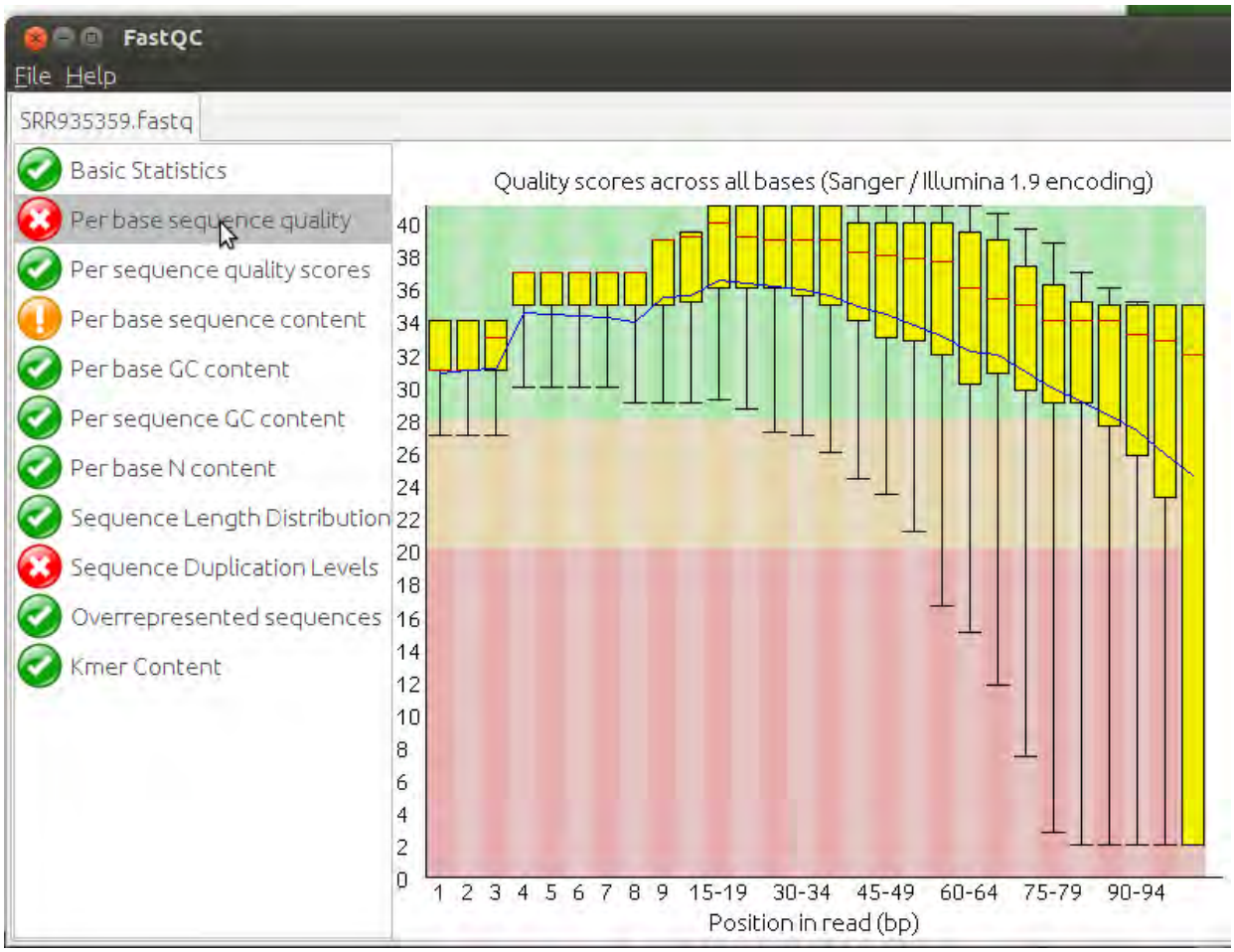
ξezipράρουμε και το αποθηκεύουμε. Αποθηκεύουμε επίσης εκεί τα fastq αρχεία μας SRR935359.fastq και SRR287818.fastq.

Πριν τρέξουμε το Condetri, ανοίγουμε τα δύο παραπάνω αρχεία μας με το FastQC και παρατηρούμε τα βασικά στατιστικά που περιγράφουν το κάθε δείγμα καθώς και την ποιότητα των βάσεων.

Για το SRR935359.fastq βλέπουμε στην εικόνα 39 ότι έχει 44,311,440 αλληλουχίες και σχετικά καλής ποιότητας reads (εικόνα 40).

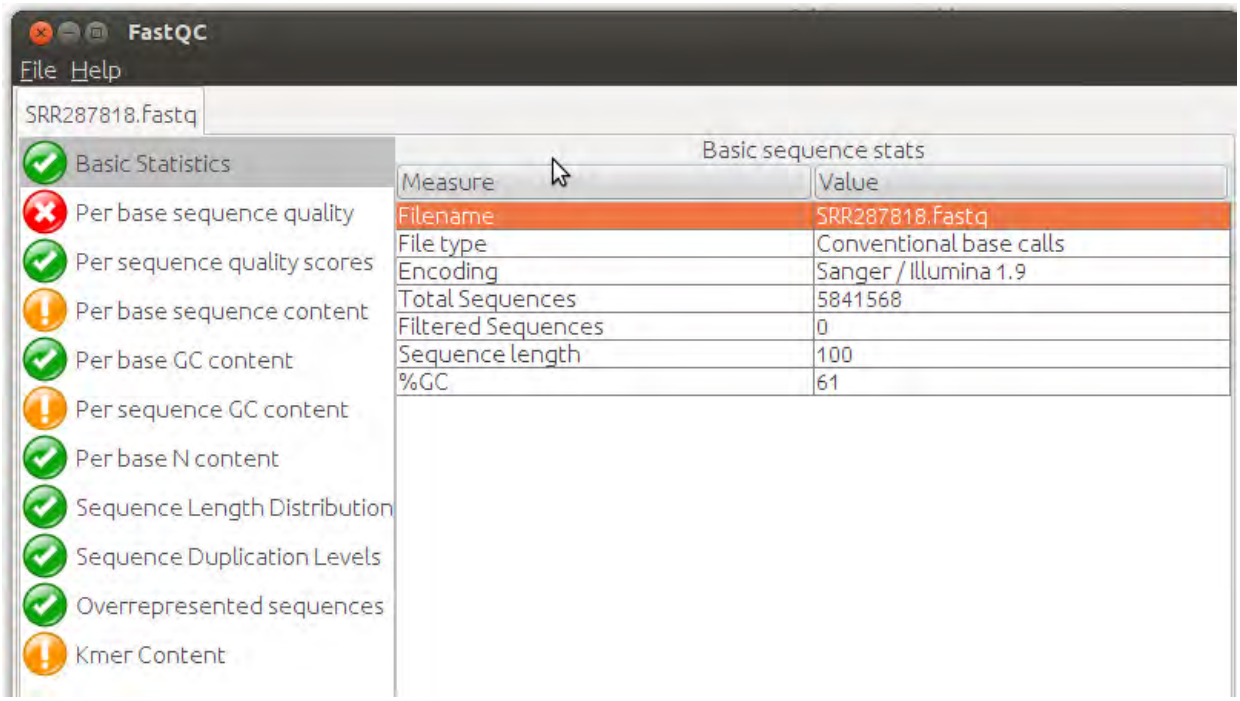


Εικόνα 39. Τα χαρακτηριστικά της SRR935359 πριν το trimming

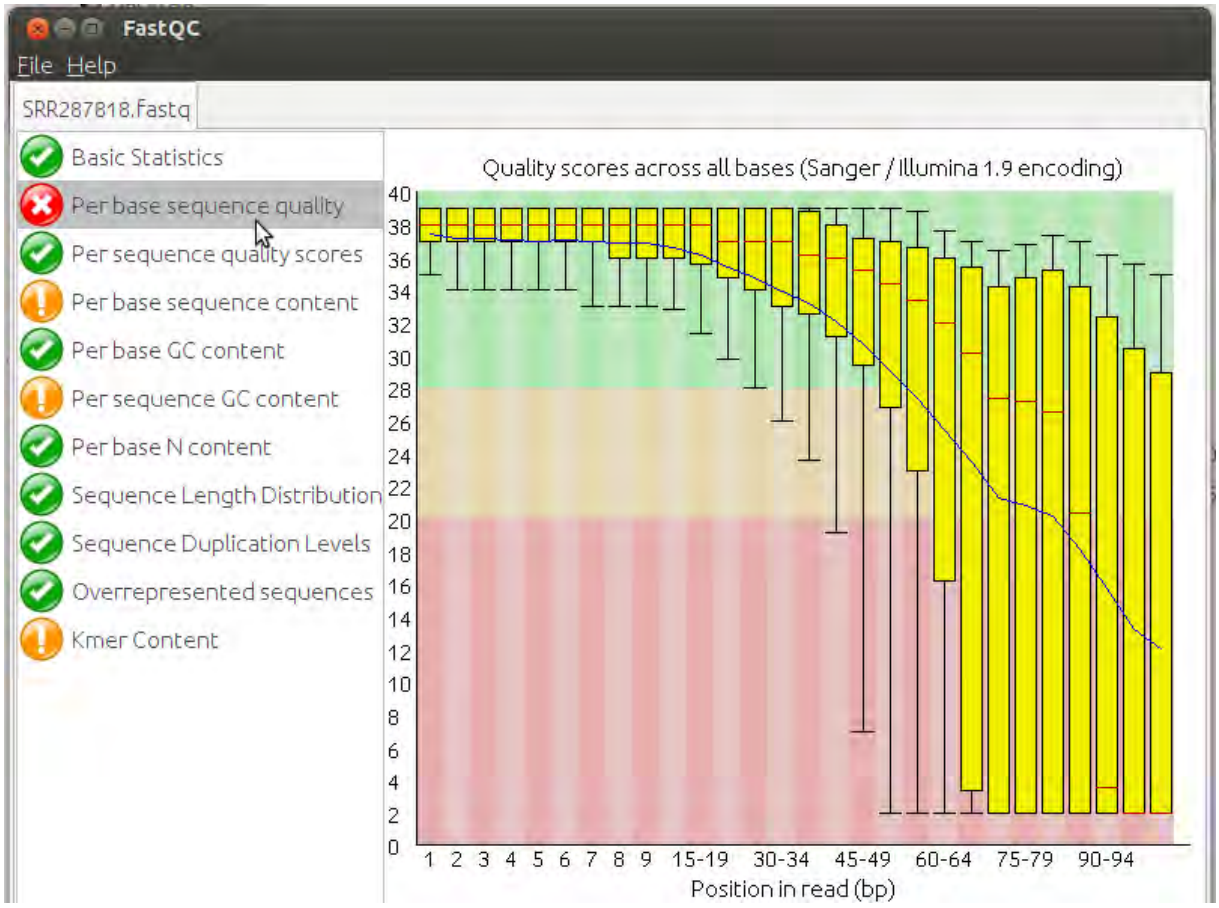


Εικόνα 40. Η ποιότητα των reads της SRR935359 πριν το trimming

Για το SRR287818.fastq βλέπουμε στην εικόνα 41 ότι έχει 5,841,568 αλληλουχίες και σχετικά κακής ποιότητας reads (εικόνα 42).



Εικόνα 41. Τα χαρακτηριστικά της SRR287818 πριν το trimming



Εικόνα 42. Η ποιότητα των reads της SRR287818 πριν το trimming

Ανοίγουμε το terminal απο το condetri για να αρχίσει το trimming και βάζουμε την εντολή:

./condetri_v2.2.pl –fastq1=SRR935359.fastq -prefix=SRR935359 -hq=25 -lq=13 -frac=0.8 -minlen=35 -mh=5 -ml=1 -sc=33 –rmN; ./condetri_v2.2.pl –fastq1=SRR287818.fastq -prefix=SRR287818 -hq=25 -lq=13 -frac=0.8 -minlen=35 -mh=5 -ml=1 -sc=33 –rmN;

Το πρόγραμμα τρέχει για περίπου 4h. Το -fastq=SRR_.fastq είναι το FASTQ file που έχουμε ήδη αποθηκεύσει στο φάκελο του Condetri. Αν είχαμε paired- end αλληλουχίες θα έπρεπε να βάλουμε στην εντολή μετά το πρώτο file και δεύτερο –fastq2=”όνομα αρχείου”.fastq. Το prefix=SRR_ είναι το όνομα του output αρχείου με τις trimmαρισμένες αλληλουχίες. Το hq είναι το hiqh quality threshold, το lq το low quality threshold, το frac το fraction του read που θα πρέπει να υπερβαίνει το hq, το minlen όπως είπαμε και πιο πάνω είναι το minimum μήκος read που επιτρέπεται, το mh σημαίνει ότι όταν ο αριθμός που του δόθηκε φταστεί απο τις διαδοχικές βάσεις με hq, σταματάει το trimming, το ml είναι ο maximum αριθμός των βάσεων με lq που επιτρέπονται μετά από ένα stretch των hq βάσεων από το 3'- άκρο και τέλος το sc είναι το scoring table της Illumina (ASCII-sc, συνήθως είναι 64 για for Illumina/Solexa version από 1.8 και πάνω και 33 για το Sanger standard).

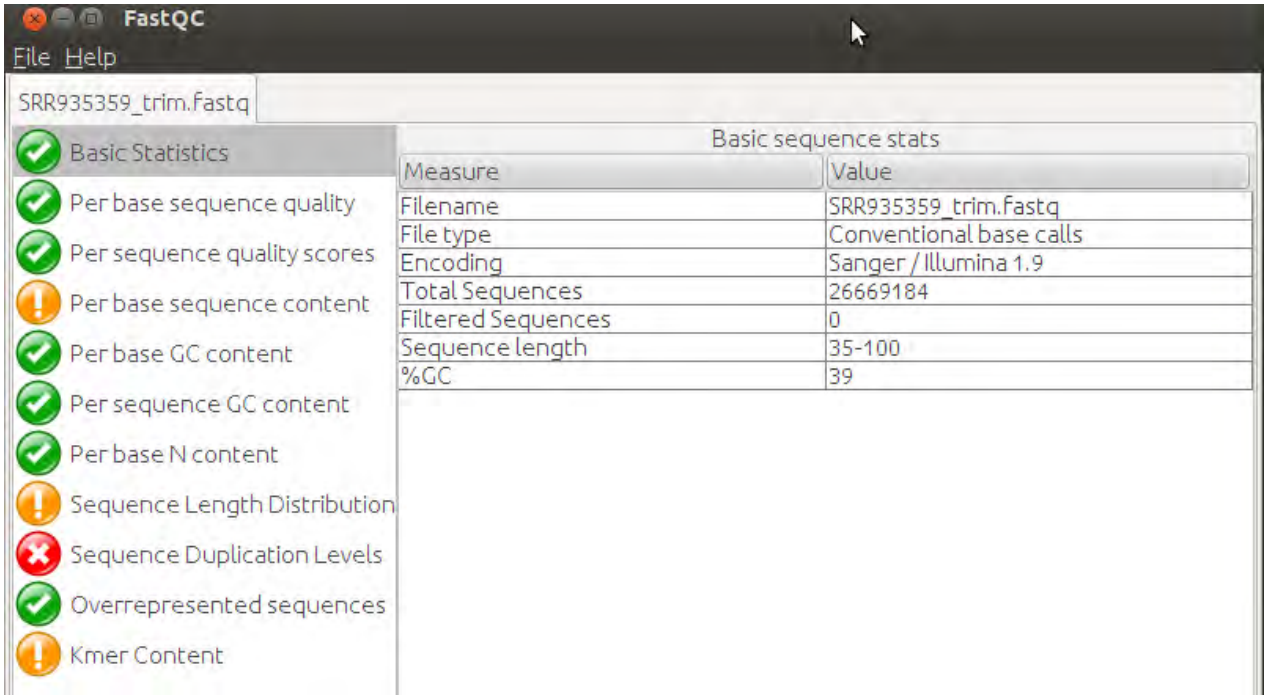
Μετά το τρέξιμο της εντολής, δημιουργούνται δύο αρχεία SRR935359_trim.fastq και SRR287818_trim.fastq με τα trimmαρισμένα reads σε μορφή FASTQ (αν είχαμε paired- end reads θα δημιουργούνταν ακόμα ένα αρχείο και ένα τρίτο με αυτά που έμειναν unpaired) και δύο ακόμη SRR935359.stats και SRR287818.stats. Τα δύο τελευταία έχουν στήλες οι οποίες με τη σειρά δείχνουν:

Το prefix (δηδ το ονομα του αρχειου), τον αριθμό των reads στο/α αρχικό/ά file(s), τον αριθμό των paired- end reads μετά το trimming, τον αριθμό των βάσεων σε ζευγάρια μετά το trimming, τον αριθμό των unpaired reads μετά το trimming και τον αριθμό των unpaired βάσεων μετά το trimming.

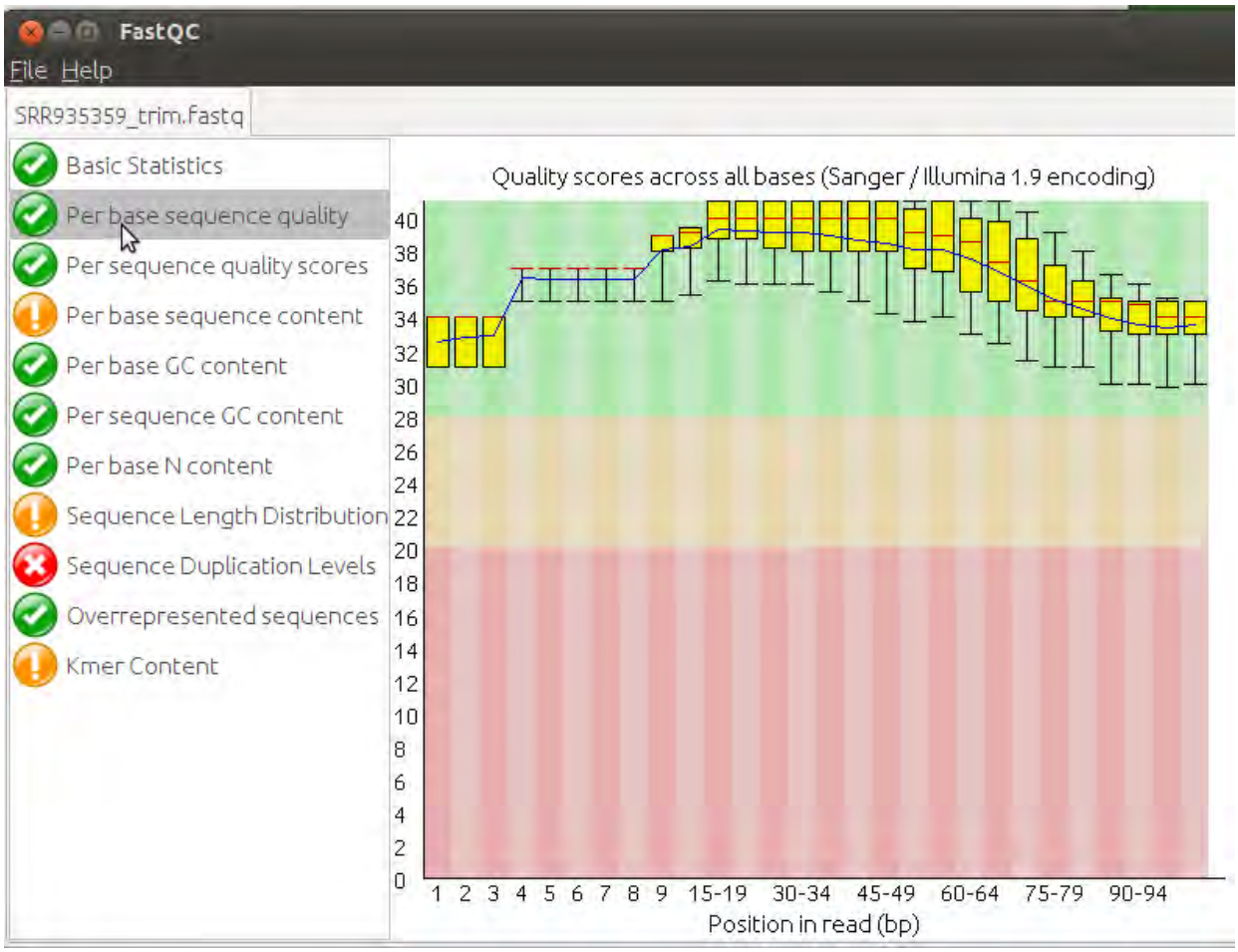
Το πρόγραμμα FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/INSTALL.txt) μπορούμε αν θέλουμε να το τρέξουμε και από το terminal ενώ είμαστε στο FastQC και πατώντας την εντολή:

/fastqc SRR_.fastq

Αλλιώς ανοίγουμε το fastqc από το φάκελο και του λέμε Open➔όποιο αρχείο θέλουμε. Το αρχείο SRR935359_trim.fastq, έχει τώρα 26,669,184 αλληλουχίες (εικόνα 43) και πολύ καλής ποιότητας reads (εικόνα 44).

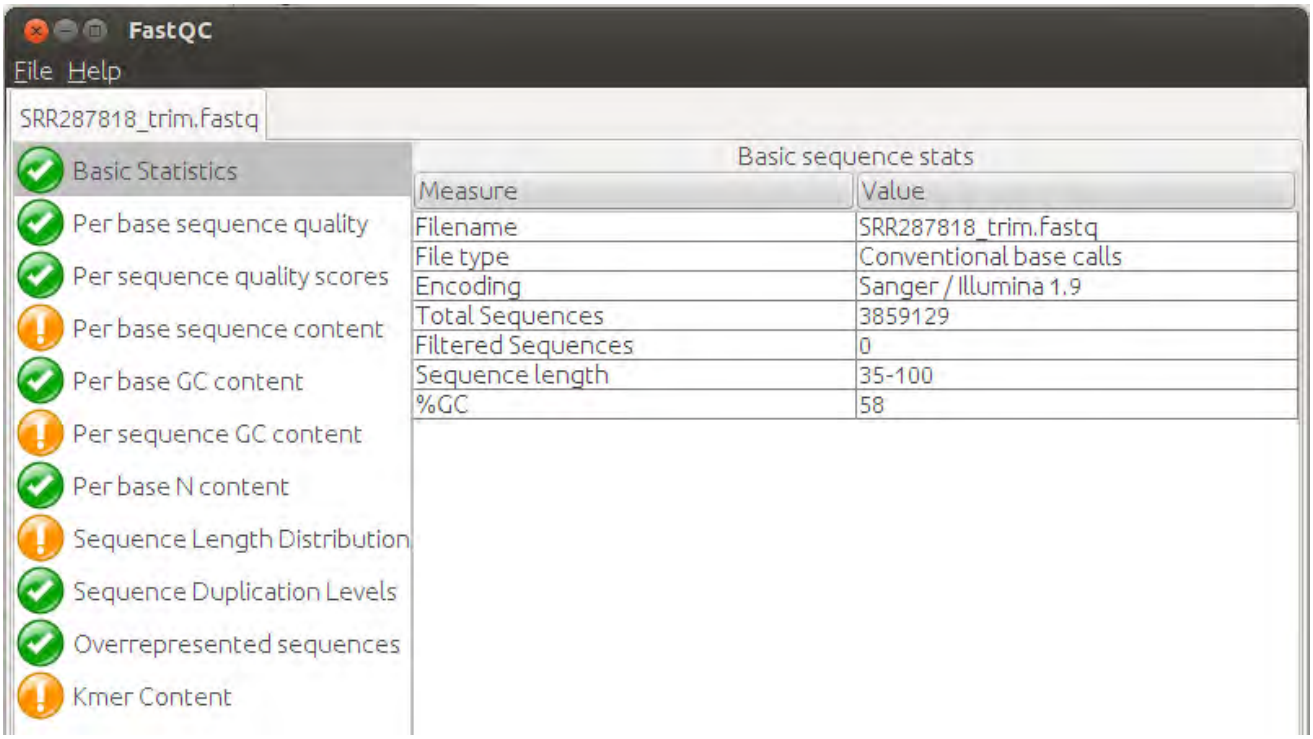


Εικόνα 43. Τα χαρακτηριστικά της SRR935359 μετά το trimming

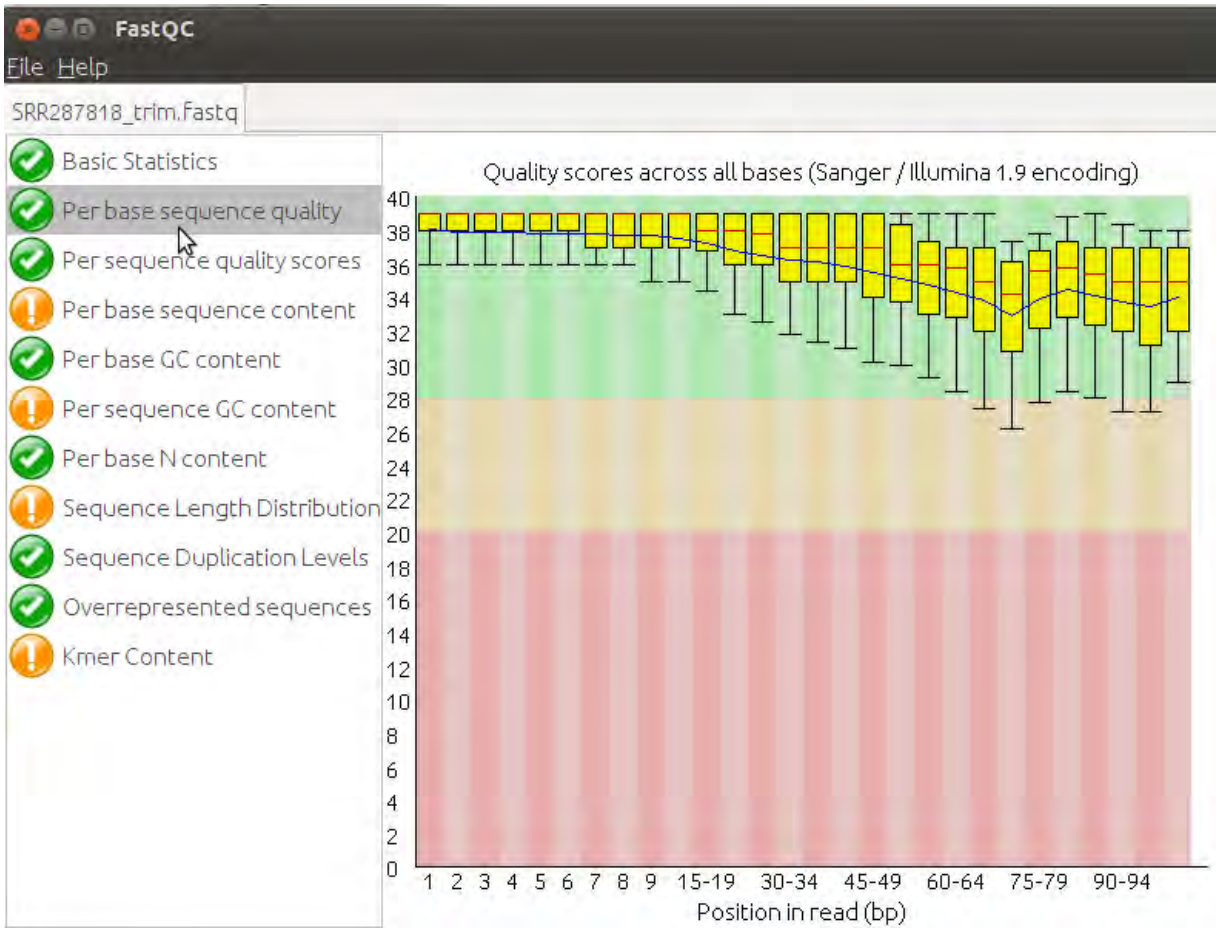


Εικόνα 44. Η ποιότητα των reads της SRR935359 μετά το trimming.

Το αρχείο SRR287818_trim.fastq, έχει τώρα 3,859,129 αλληλουχίες (εικόνα 45) και επίσης πολύ καλής ποιότητας reads (εικόνα 46).



Εικόνα 45. Τα χαρακτηριστικά της SRR287818 μετά το trimming



Εικόνα 46. Η ποιότητα των reads της SRR287818 μετά το trimming.

3.3 Trinity

Έχουμε κατεβάσει, ξεzipράρει και αποθηκεύσει το πρόγραμμα Trinity (trinityrnaseq_r20131110) από το http://trinityrnaseq.sourceforge.net/ . Αντιγράφουμε τα αρχεία SRR935359_trim.fastq και SRR287818_trim.fastq στο φάκελο trinityrnaseq_r20131110 για να κάνουμε την de novo συναρμολόγιση και να πάρουμε τα contigs. Από εκεί ανοίγουμε το terminal και γράφουμε την εντολή:

./Trinity.pl --seqType fq --JM 90G --single SRR935359_trim.fastq --CPU 12 --out SRR935359_run

και μόλις ολοκληρωθεί,

./Trinity.pl --seqType fq --JM 90G --single SRR287818_trim.fastq --CPU 12 --out SRR287818_run

Το πρόγραμμα τρέχει για 10-15 min. Όσον αφορά στις παραμέτρους σύμφωνα με το http://trinityrnaseq.sourceforge.net/#running_trinity , το --seqType fq το πληκτρολογούμε επειδή έχουμε τα αρχεία μας σε fastq μορφή, το --JM 90G είναι το Jellyfish Memory, δηδ. ο αριθμός των GB της μνήμης του συστήματος που χρησιμοποιούνται απο το jellyfish του προγράμματος για να μετρήσει τα k- mers, το --single SRR935359_trim.fastq σημαίνει οτι είναι single –end και το όνομα του αχείου (αν είχαμε paired- end θα είχαμε και δύο αρχεία και αντί για –single θα έπρεπε να βάλουμε --left SRR_.fastq και --right SRR_.fastq), το --CPU 12 είναι ο αριθμός των CPUs (default=2) και το --out SRR935359_run είναι ο φάκελος που δημιουργείται.

Εμείς πήραμε δύο φακέλους τους SRR935359_run και SRR287818_run. Μέσα στον κάθε φάκελο που δημιουργήθηκαν τα αρχεία Trinity.fasta, αυτά που περιέχουν τα contigs σε fasta μορφή. Στη συνέχεια για να κάνουμε το gene prediction με το πρόγραμμα Prodigal, αποθηκεύουμε τα δύο αρχεία Trinity.fasta στο φάκελο Prodigal και τα μετονομάζουμε SRR935359_Trinity.fasta και SRR287818_Trinity.fasta. Το SRR935359_Trinity.fasta έχει 11,577 αλληλουχίες, δηδ. contigs, με μέσο όρο μήκους 16,719 b, μεγαλύτερο μήκος contig 33,237 b και

μικρότερο 201 b και το SRR287818_Trinity.fasta έχει 25,567 αλληλουχίες, δηδ. contigs με μέσο όρο μήκους 521.85 b, μεγαλύτερο μήκος contig 33,247 b και μικρότερο 201 b.

3.4 Velvet- MetaVelvet

Κατεβάζουμε το πρόγραμμα Velvet (version 1.2.10) από http://www.ebi.ac.uk/~zerbino/velvet/ το ξεzipράρουμε και το αποθηκεύουμε πριν το τρέξουμε. Πρώτα ορίζουμε από το directory του velvet_1.2.10, όπου βρισκόμαστε, σύμφωνα με τις οδηγίες του http://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf , το μήκος των k-mers στο 29 με την εντολή:

make 'MAXKMERLENGTH=29'

Έχουν δημιουργηθεί τα velveth και velvetg. Εν τω μεταξύ κατεβάζουμε το πρόγραμμα MetaVelvet 1.2.02 (http://metavelvet.dna.bio.keio.ac.jp/src/), το ξεzipράρουμε και το αποθηκεύουμε. Και εδώ, σύμφωνα με τις οδηγίες του http://metavelvet.dna.bio.keio.ac.jp/ , ορίζουμε από το directory MetaVelvet-1.2.02.το μήκος των k-mers στο 29 με την εντολή:

make 'MAXKMERLENGTH=29'

Έχει δημιουργηθεί το meta-velvetg. Το hash length το ορίζουμε στο 29 γιατί οι αλληλουχίες μας είναι single- end.

Το velveth βοηθά στη δημιουργία ενός dataset για το velvetg υποδεικνύοντας ποιό sequence file αντιπροσωπεύει. Ουσιαστικά σαν input του δίνουμε read sequences σε κάποια μορφή (π.χ. fasta ή fastq κ.α.) και αυτό δημιουργεί ένα k-mer hash πίνακα. Το velvetg είναι ο πυρήνας του προγράμματος Velvet και φτάχνει το γράφημα de Bruijn. Το meta-velvetg είναι αυτό που σπάζει το γράφημα σε επιμέρους υπογραφήματα de Bruijn για κάθε είδος (αν είναι δυνατό) που βρίσκεται στο μεταγονιδίωμα. Εκτός από τα παραπάνω τρία απαραίτητη στη διαδικασία της συναρμολόγησης είναι και η zlib library.

Στη συνέχεια εκτελούμε το velveth για να εισάγουμε τις read sequences SRR935359_trim.fastq και SRR287818_trim.fastq και να δημιουργήσουμε το k-mer

hash table. Αυτό γίνεται ενώ είμαστε στο velvet_1.2.10, γράφοντας την εντολή:

./velveth SRR935359_dir 29 –fastq –short SRR935359_trim.fastq

και στη συνέχεια μόλις τέξει:

./velveth SRR287818_dir 29 –fastq –short SRR287818_trim.fastq

Το velveth τρέχει περίπου για 3min για το καθένα. Δημιουργούνται έτσι δύο φάκελοι στο velvet_1.2.10 με τα ονόματα SRR935359_dir και SRR287818_dir και μέσα στον καένα από σε αυτούς εμπεριέχονται και τα αρχεία Sequences και Roadmaps. Αμέσως μετά εκτελούμε το velvetg για την κατασκευή του γραφήματος de Bruijn. Αυτό γίνεται ενώ είμαστε στο velvet_1.2.10, γράφοντας την εντολή (με το expected coverage ρυθμισμένο στο auto για να μπορεί να λειτουργήσει στη συνέχεια το meta-velvetg):

./velvetg SRR935359_dir –exp_cov auto

και

./velvetg SRR287818_dir –exp_cov auto

Τρέχουν για 50 min και 3 min αντίστοιχα. Δημιουργούνται έτσι δύο φάκελοι οι SRR935359_dir και SRR287818_dir και μέσα στον καθένα βρίσκεται ένα αρχείο με την ονομασία Graph2. Τέλος, τρέχουμε το meta-velvetg δίνοντας από το MetaVelvet 1.2.02 στο οποίο βρισκόμαστε την εντολή:

./meta-velvetg ../velvet_1.2.10/ SRR935359_dir

και μετά

./meta-velvetg ../velvet_1.2.10/ SRR287818_dir

Αυτό επειδή τα SRR935359_dir και SRR287818_dir βρίσκονται στο φάκελο velvet_1.2.10. Συνεπώς πρέπει να δωθεί η σχετική διεύθυνση (ή και η πλήρης διεύθυνση μέχρι τα SRR935359_dir και SRR287818_dir), δηλαδή ../ που είναι ένα επίπεδο πάνω από το φάκελο MetaVelvet 1.2.02 και από αυτό το επίπεδο προς τα κάτω δηλαδη /velvet_1.2.10 για να πάει στον ομώνυμο φάκελο και SRR935359_dir ή SRR287818_dir ακόμη πιο κάτω στο επιθυμητό ομώνυμο αρχείο την κάθε φορά. Το πρόγραμμα έτρεξε 15 min και 3min αντίστοιχα. Μετά το τρέξιμο δημιουργείται το αρχείο meta-velvetg.contigs.fa στον κάθε φάκελο SRR935359_dir και SRR287818_dir. Σε αυτό το αρχείο FASTA είναι τα κύρια αποτελέσματα της συναρμολόγησης, δηλαδή τα contigs.

Στη συνέχεια για να κάνουμε το gene prediction με το πρόγραμμα Prodigal, αποθηκεύουμε τα δύο αρχεία meta-velvetg.contigs.fa στο φάκελο Prodigal και τα μετονομάζουμε SRR935359_ meta-velvetg.contigs.fa και SRR287818_ meta-velvetg.contigs.fa. Το SRR935359_ meta-velvetg.contigs.fa έχει 20,411 αλληλουχίες, δηδ. contigs, με μέσο όρο μήκους 302.32 b, μεγαλύτερο μήκος contig 18,887 b και μικρότερο 57 b και το SRR287818_ meta-velvetg.contigs.fa έχει 66,820 αλληλουχίες, δηδ. contigs με μέσο όρο μήκους 139.2 b, μεγαλύτερο μήκος contig 3,265 b και μικρότερο 57 b.

3.5 Prodigal

Με το πρόγραμμα αυτό γίνεται το gene prediction. Κατεβάζουμε το prodigal.v2_60.linux (http://code.google.com/p/prodigal/downloads/detail?name=prodigal.v2_60.linux), το οποίο είναι binary file, δηλαδή έχει μόνο το πρόγραμμα το οποίο μπορούμε να το τρέξουμε απευθείας χωρίς να ξezipράρουμε κ.λ.π. Αφού ανοίξουμε το terminal από το σημείο στο οποίο βρίσκεται το Prodigal, σύμφωνα με τις οδηγίες του http://code.google.com/p/prodigal/source/browse/README , (δημιουργούμε ένα φάκελο με το όνομα αυτό), πρέπει να το κάνουμε εκτελέσιμο, άρα πατάμε

```
chmod 755 prodigal.v2_60
```

3.5.1 Gene Prediction με τα αποτελέσματα του Trinity

Στη συνέχεια πρέπει να τρέξουμε το πρόγραμμα με τις κατάλληλες

παραμέτρους ώστε να δούμε το gene prediction, δηδ. τα γονίδια που βρίσκονται μέσα στα contigs που πήραμε από τα αρχεία SRR935359_Trinity.fasta και SRR287818_Trinity.fasta (τα οποία το έχουμε αντιγράψει μέσα στο φάκελο Prodigal). Η εντολή που δίνουμε είναι:

./prodigal.v2_60.linux –a trans_ SRR935359 –d nuc_ SRR935359 –f gbk –g 11 –i SRR935359_Trinity.fasta –o output_ SRR935359 –p meta –s start_ SRR935359

Το πρόγραμμα τρέχει για 2-3 min. Οι παράμετροι που χρησιμοποιούμε είναι με τη σειρά:

–a trans_ SRR935359 : είναι το αρχείο στο οποίο θα αποθηκευτούν οι πρωτεϊνικές αλληλουχίες σε fasta μορφή

–d nuc_ SRR935359 : είναι το αρχείο στο οποίο θα αποθηκευτούν οι νουκλεοτιδικές αλληλουχίες σε fasta μορφή

–f gbk : είναι το output format, το οποίο ουσιαστικά δείχνει τη μορφή του πίνακα στο αρχείο του output. Ο πίνακας μπορεί να έχει τη μορφή που έχει στη Genebank (gbk), αλλά και άλλες μορφές όπως αν πληκτρολογήσουμε αντί για gbk, ggf ή sco (simple coordinate)

–g 11 : είναι ο αριθμός του πίνακα μετάφρασης της Genebank (όλοι οι πίνακες βρίσκονται εδώ http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), ο 11 είναι ο standard microbial table

–i SRR935359_Trinity.fasta: είναι το αρχείο input από το οποίο θα αντλήσει τα δεδομένα το Prodigal για να τρέξει

–o output_ SRR935359 : είναι το αρχείο output που θα δημιουργηθεί μετά την ανάλυση και θα μας δείχνει τις CDS (οι κωδικοποιούμενες αλληλουχίες), τα όρια των γονιδίων, τα DEFINITION FEATURES κ.λ.π. όπως είναι η δομή της Genebank (Genebank – divider //, εδώ γιατί ορίσαμε την παράμετρο gbk)

–p meta : γιατί κάνουμε ανάλυση μεταγονιδιώματος

–s start_ SRR935359: είναι το αρχείο το οποίο δημιουργείται περιέχει τα πιθανά γονίδια με τα σκορ του καθενός, την αρχή και τέλος κάθε γονοδίου, το strand (+ ή -) από το οποίο προέρχεται κ.α.

Στη συνέχεια επαναλαμβάνουμε την εντολή για το SRR287818_Trinity.fasta

./prodigal.v2_60.linux –a trans_ SRR287818 –d nuc_ SRR287818 –f gbk –g 11 –i SRR287818_Trinity.fasta –o output_ SRR287818 –p meta –s start_ SRR287818

Οπότε μετά στο φάκελο του Prodigal έχουμε τα αρχεία :
nuc_ SRR935359 (με 31,175 νουκλεοτιδικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 524.9 b, με μεγαλύτερο μήκος 9,489 b και μικρότερο μήκος 60 b),
nuc_ SRR287818 (με 32,324 νουκλεοτιδικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 367.8 b, με μεγαλύτερο μήκος 4,929b και μικρότερο μήκος 60b),
output_ SRR935359,
output_ SRR287818,
start_ SRR935359,
start_ SRR287818,
trans_ SRR935359 (με 31,175 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 175 a, με μεγαλύτερο μήκος 3,163 a και μικρότερο μήκος 20 a) και
trans_ SRR287818 (με 32,324 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 122.6 a, με μεγαλύτερο μήκος 1,643 a και μικρότερο μήκος 20 a).

Πριν κάνουμε το BlastN μετονομάζουμε τα αρχεία nuc_ SRR935359 και nuc_ SRR287818 σε nuc_ SRR935359.fasta και nuc_ SRR287818.fasta χρησιμοποιώντας την εντολή:

mv nuc_ SRR935359 nuc_ SRR935359.fasta

και

mv nuc_ SRR287818 nuc_ SRR287818.fasta

Τέλος αποθηκεύουμε τα αρχεία που μετονομάστηκαν στο φάκελο bin του φακέλου ncbi-blast-2.2.27+ του φακέλου Blast.

3.5.2 Gene Prediction με τα αποτελέσματα του MetaVelvet

Πρέπει να τρέξουμε το πρόγραμμα με τις κατάλληλες παραμέτρους ώστε να δούμε το gene prediction, δηδ. τα γονίδια που βρίσκονται μέσα στα contigs που πήραμε από τα αρχεία SRR935359_ meta-velvetg.contigs.fa και SRR287818_ meta-velvetg.contigs.fa τα οποία έχουμε ήδη αντιγράψει στο φάκελο Prodigal. Η εντολή που δίνουμε είναι:

./prodigal.v2_60.linux –a trans_ SRR935359_2 –d nuc_ SRR935359_2 –f gbk –g 11 –i SRR935359_ meta-velvetg.contigs.fa –o output_ SRR935359_2 –p meta –s start_ SRR935359_2

και μετά

./prodigal.v2_60.linux –a trans_ SRR287818_2 –d nuc_ SRR287818_2 –f gbk –g 11 –i SRR287818_ meta-velvetg.contigs.fa –o output_ SRR287818_2 –p meta –s start_ SRR287818_2

Το πρόγραμμα έτρεξε για 3-5 min για την κάθε εντολή. Οπότε μετά στο φάκελο του Prodigal έχουμε τα αρχεία:

nuc_ SRR935359_2 (με 21,331 νουκλεοτιδικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 253 b, με μεγαλύτερο μήκος 7,116 b και μικρότερο μήκος 60 b),

nuc_ SRR287818_2 (με 60,730 νουκλεοτιδικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 134.9 b, με μεγαλύτερο μήκος 1,938 b και μικρότερο μήκος 60 b),

output_ SRR935359_2,

output_ SRR287818_2,

start_ SRR935359_2,

start_ SRR287818_2,

trans_ SRR935359_2 (με 21,331 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες

σε γονίδια, με μέσο όρο μήκους 84.35 a, με μεγαλύτερο μήκος 2,372 a και μικρότερο μήκος 20 a) και

trans_ SRR287818_2 (με 60,730 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια, με μέσο όρο μήκους 45 a, με μεγαλύτερο μήκος 2,372 a και μικρότερο μήκος 20 a).

Πριν κάνουμε το BlastN μετονομάζουμε τα αρχεία nuc_ SRR935359_2 και nuc_ SRR287818_2 σε nuc_ SRR935359_2.fasta και nuc_ SRR287818_2.fasta χρησιμοποιώντας την εντολή:

mv nuc_ SRR935359_2 nuc_ SRR935359_2.fasta

και

mv nuc_ SRR287818_2 nuc_ SRR287818_2.fasta

Τέλος αποθηκεύουμε τα αρχεία που μετονομάστηκαν στο φάκελο bin του φακέλου ncbi-blast-2.2.27+ του φακέλου Blast.

3.6 Blast

Από το φάκελο Blast/ncbi-blast-2.2.27+/bin φτάνουμε στο φάκελο bin όπου βρίσκεται το πρόγραμμα blastn. Κατεβάζουμε από την database του NCBI – nt (nucleotide) το φάκελο nt.gz (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/). Το ξεzippάρουμε και αποθηκεύουμε τα αρχεία του στο bin. Πρέπει να φτιάξουμε σε αυτό το σημείο την database ώστε να είναι στη σωστή μορφή για να τρέξει το blast. Αυτό γίνεται με την εντολή:

./makeblastdb –in nt –out BlastN_db2_nt –dbtype nucl

Οπότε δημιουργούνται αρχεία BlastN_db2_nt με διάφορες καταλήξεις (.nsq, nin κ.α.). Αυτό σημαίνει ότι φτιάχτηκε η βάση δεδομένων. Για να κάνουμε BlastN ανοίγουμε το terminal και πληκτρολογούμε:

./blastn –db BlastN_db2_nt –query nuc_ SRR935359.fasta
–out blastn_ SRR935359 –evalue 1e-10 –outfmt 0

και αμέσως μετά το τρέξιμο του πρώτου:

./blastn –db BlastN_db2_nt –query nuc_ SRR287818.fasta
–out blastn_ SRR287818 –evalue 1e-10 –outfmt 0

Ξανακάνουμε blastn με άλλο output format:

./blastn –db BlastN_db2_nt –query nuc_ SRR935359.fasta
–out blastn1_ SRR935359 –evalue 1e-10 –outfmt “6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos qframe sframe” –max_target_seqs 1 –num_threads 12

και αμέσως μετά το τρέξιμο του πρώτου:

./blastn –db BlastN_db2_nt –query nuc_ SRR287818.fasta
–out blastn1_ SRR287818 –evalue 1e-10 –outfmt “6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos qframe sframe” –max_target_seqs 1 –num_threads 12

Το BlastN έκανε να τρέξει περίπου 40 min, 50 min, 5 min και 10 min αντίστοιχα με τη σειρά των εντολων. Ουσιαστικά οι παράμετροι που θέσαμε σημαίνουν με τη σειρά:

-db BlastN_db2_nt: σαν database να χρησιμοποιεί αυτή που δημιουργήσαμε στο προηγούμενο βήμα από τη μορφή fasta

–query nuc_ SRR935359.fasta: σαν ακολουθίες επερώτησης χρησιμοποιούμε αυτές από το gene prediction που κάναμε με το Prodigal και τις αποθηκεύσαμε στο αρχείο που φαίνεται σε fasta μορφή

–out blastn1_ SRR935359: το αρχείο που δημιουργείται μετά το blast με τα αποτελέσματά του

–evalue 1e-10 : ο αριθμός των hits που μπορεί κανείς να "περιμένει" για να δεί κατά τύχη όταν ψάχνει σε μία βάση δεδομένων συγκεκριμένου μεγέθους. Μειώνεται εκθετικά καθώς το σκορ (S) των match αυξάνεται. Ουσιαστικά, περιγράφει τον τυχαίο θόρυβο. Για παράδειγμα, μια τιμή του E ίση με 1 που αντιστοιχεί σε ένα χτύπημα μπορεί να ερμηνευθεί ότι σε μια βάση δεδομένων συγκεκριμένου μεγέθους θα μπορούσε κανείς να περιμένει να δεί 1 match με παρόμοιο σκορ απλώς κατά τύχη. Όσο χαμηλότερο είναι το e-value, ή όσο πιο κοντά στο μηδέν, τόσο πιο "σημαντικό" το match είναι. Το evalue μπορεί επίσης να χρησιμοποιηθεί ως ένας βολικός τρόπος για τη δημιουργία ένος κατωφλίου σημαντικότητας για την αναφορά των αποτελεσμάτων.

–outfmt “6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos qframe sframe” : ορίζουμε έτσι πώς θέλουμε να φαίνονται τα αποτελέσματα του blast όταν θα ανοίξουμε το αρχείο blastp_test1, ουσιαστικά θέτουμε τι θα έχει η κάθε στήλη με τη σειρά

–max_target_seqs 1 : πρέπει να οριστεί ούτως ώστε να παίρουμε με κάθε αλληλουχία query ένα αποτέλεσμα στο blast, αυτό με τη μεγαλύτερη ομοιότητα και όχι όλα τα ομόλογα που μπορεί αυτό να βει.

–num_threads 12 : Δείχνει τον αριθμό των threads που χρησιμοποιεί ο υπολογιστής για να τρέξει το blast (εδώ χρησιμοποίησε τα 12 από τα 16 που έχει).

Κάνουμε αυτή τη φορά BlastN για τα δεδομένα από το MetaVelvet. Η βάση δεδομένων BlastN_db2_nt είναι ήδη φτιαγμένη από πριν. Για να κάνουμε BlastN ανοίγουμε το terminal και πληκτρολογούμε:

./blastn –db BlastN_db2_nt –query nuc_ SRR935359_2.fasta
–out blastn_ SRR935359_2 –evalue 1e-10 –outfmt 0

και αμέσως μετά το τρέξιμο του πρώτου:

./blastn –db BlastN_db2_nt –query nuc_ SRR287818_2.fasta
–out blastn_ SRR287818_2 –evalue 1e-10 –outfmt 0

Ξανακάνουμε blastn με άλλο output format:

./blastn –db BlastN_db2_nt –query nuc_ SRR935359_2.fasta
–out blastn1_ SRR935359_2 –evalue 1e-10 –outfmt “6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos qframe sframe” –max_target_seqs 1 –num_threads 12

και αμέσως μετά το τρέξιμο του πρώτου:

./blastn –db BlastN_db2_nt –query nuc_ SRR287818_2.fasta
–out blastn1_ SRR287818_2 –evalue 1e-10 –outfmt “6 qacc sacc evalue qstart qend sstart send bitscore qlen slen length pident ppos qframe sframe” –max_target_seqs 1 –num_threads 12

Το BlastN έκανε να τρέξει περίπου 10 min, 33 min, 3 min και 16 min αντίστοιχα με τη σειρά των εντολων. Ουσιαστικά οι παράμετροι που θέσαμε είναι ίδιοι με παραπάνω.

3.7 PhyloSift

Κατεβάζουμε το PhyloSift και συγκεκριμένα τη version 1.0 (phylosift_v1.0.0_02) από εδώ http://phylosift.wordpress.com/, ξezippάρουμε και αποθηκεύουμε. Μέσα στο φάκελο του Phylosift αποθηκεύουμε τις fastq αλληλουχίες μας SRR935359_trim.fastq και SRR287818_trim.fastq.

Για να τρέξουμε το πρόγραμμα ανοίγουμε το terminal από το σημείο που το αποθηκεύσαμε, σύμφωνα με το http://phylosift.wordpress.com/tutorials/running-phylosift/illumina-tutorial/ , και πατάμε την εντολή:

cd phylosift_v1.0.0_02

Αυτό για να οδηγηθούμε μέσα στο directory του προγράμματος. Στη συνέχεια εκτελούμε τις εντολές:

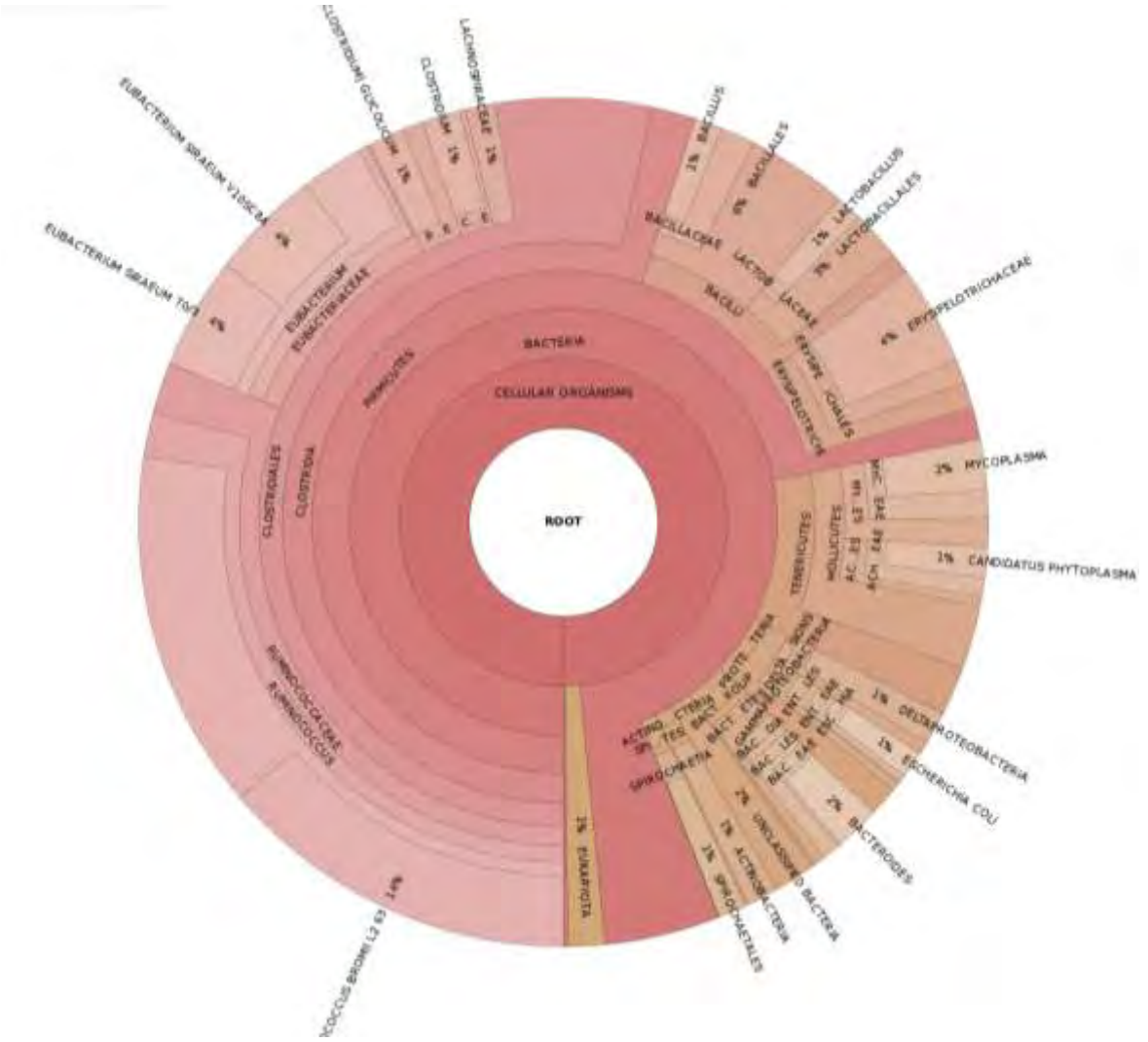
./phylosift all --f --debug --output phylosift_SRR935359

και

./phylosift all SRR287818_trim.fastq --f --debug --output phylosift_SRR287818

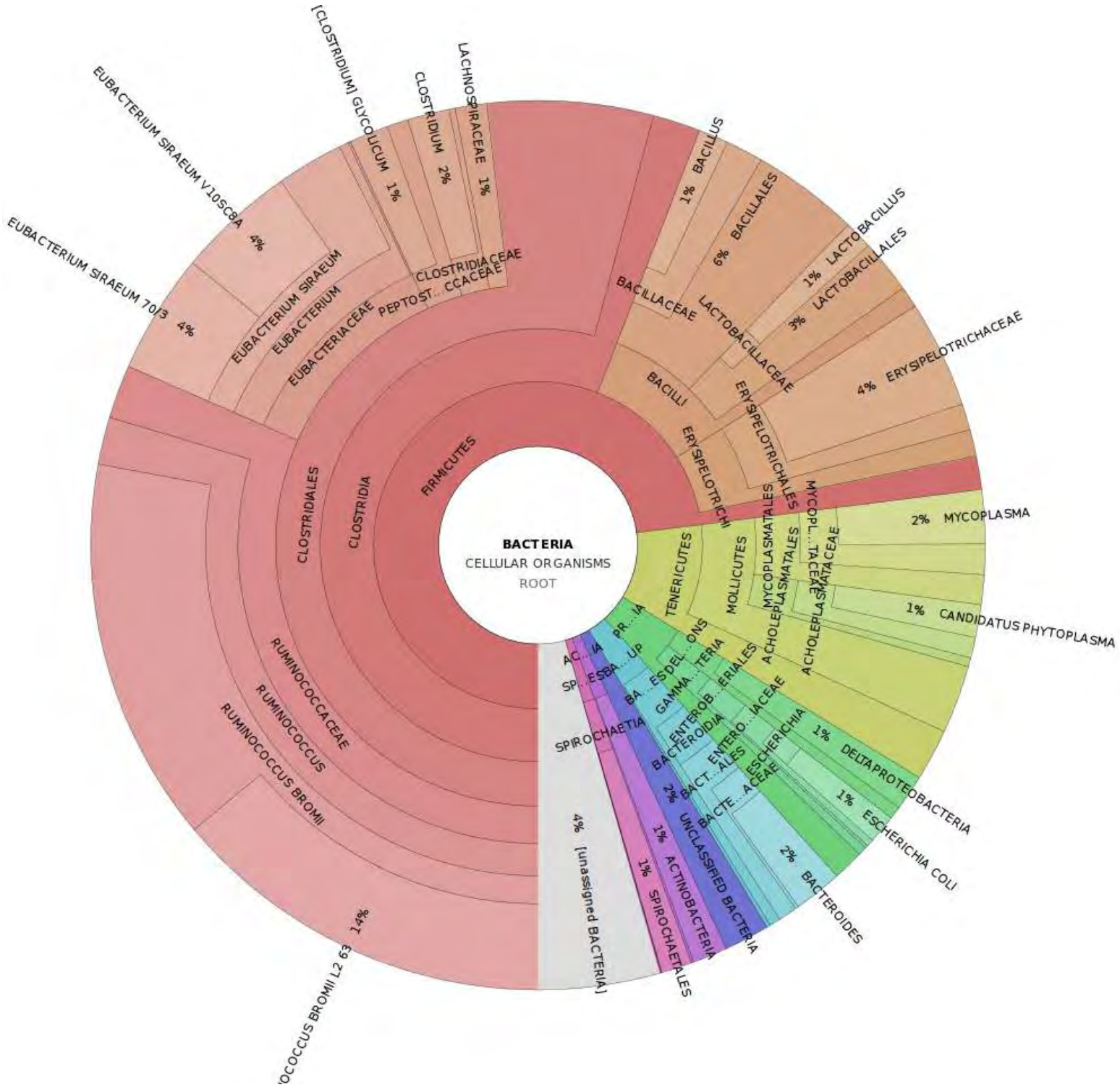
Το πρόγραμμα για να τρέξει έκανε περίπου 52 h το SRR935359_trim.fastq και το SRR287818_trim.fastq 14 h 30 min. Το all σημαίνει ότι θα πρέπει να τρέξει όλα τα στάδια της φυλογενετικής ανάλυσης των δεδομένων της μεταγενομικής αλληλουχίας. Το όνομα του αρχείου είναι το input file και στη συγκεκριμένη περίπτωση είναι single – end, το output είναι το όνομα του φακέλου που δημιουργείται. Μέσα εκεί βρίσκονται αρχεία και φάκελοι, όπως ο treeDir, ο οποίος περιέχει ένα αρχείο για τον κάθε marker (όλοι έχουν κατέβει προηγουμένως και είναι αποθηκευμένοι στο /home_directory/share/phylosift/markers/ και /home_directory/share/phylosift/ncbi/), που αποτελούν στην πραγματικότητα δέντρα. Το αρχείο SRR935359_trim.fastq.html (ή ανίστοιχα SRR287818_trim.fastq.html), του output φακέλου οδηγεί στον krona viewer (επιτρέπει την εξερεύνηση των δεδομένων, αφού οι χρήστες με ένα κλικ στην ομάδα που επιθυμούν μπορούν να επεκτείνουν την ταξινόμηση και να προβάλουν τις πληροφορίες σε ολοένα και μικρότερη κλίμακα).

Τα αποτελέσματα του SRR935359_trim.fastq.html στον krona viewer:



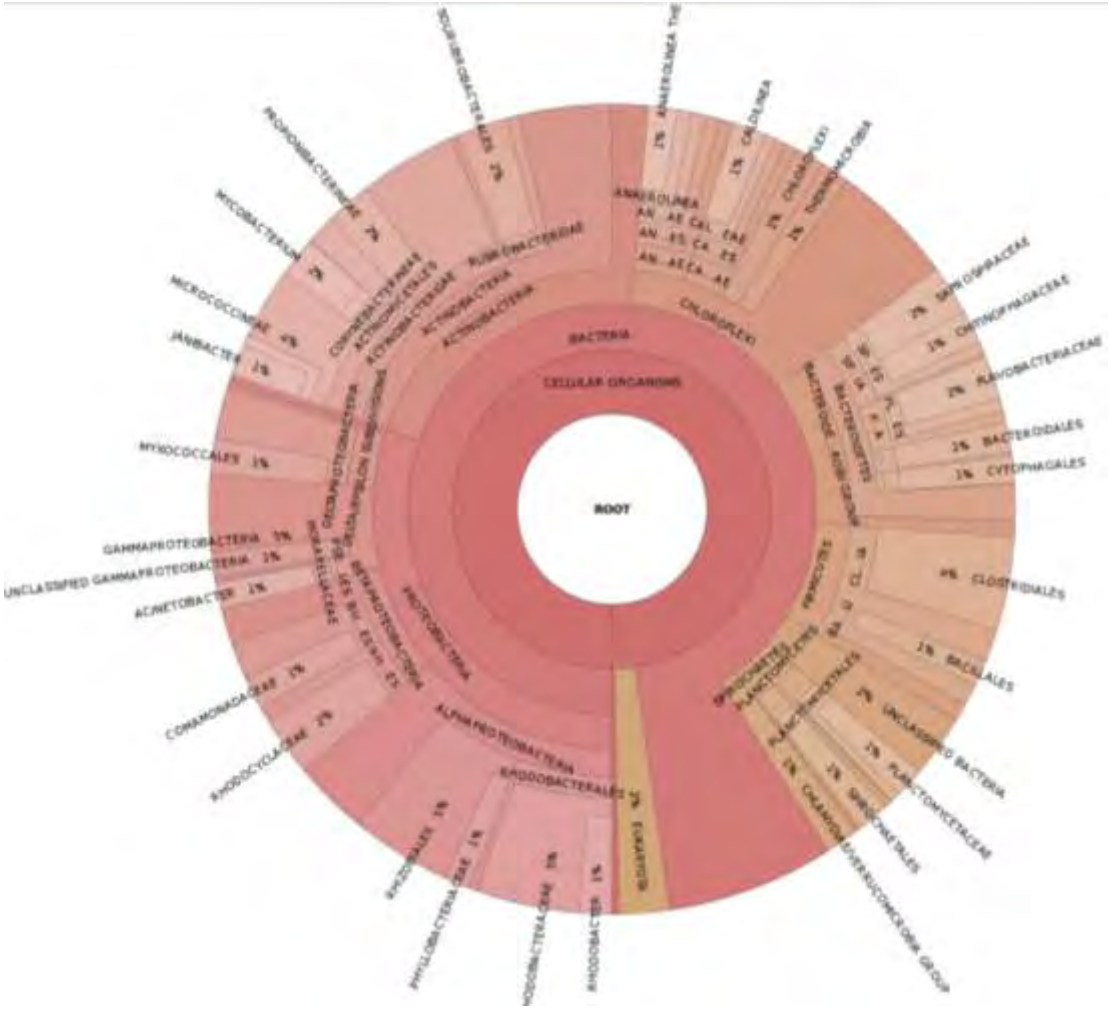
Εικόνα 47. Η οπτικοποίηση με τον Krona viewer για την ταξινόμηση του SRR935359

Εικόνα 48. Η οπτικοποίηση με τον Krona viewer για την ταξινόμηση του SRR935359, αρχίζοντας από το επίπεδο των Cellular organisms



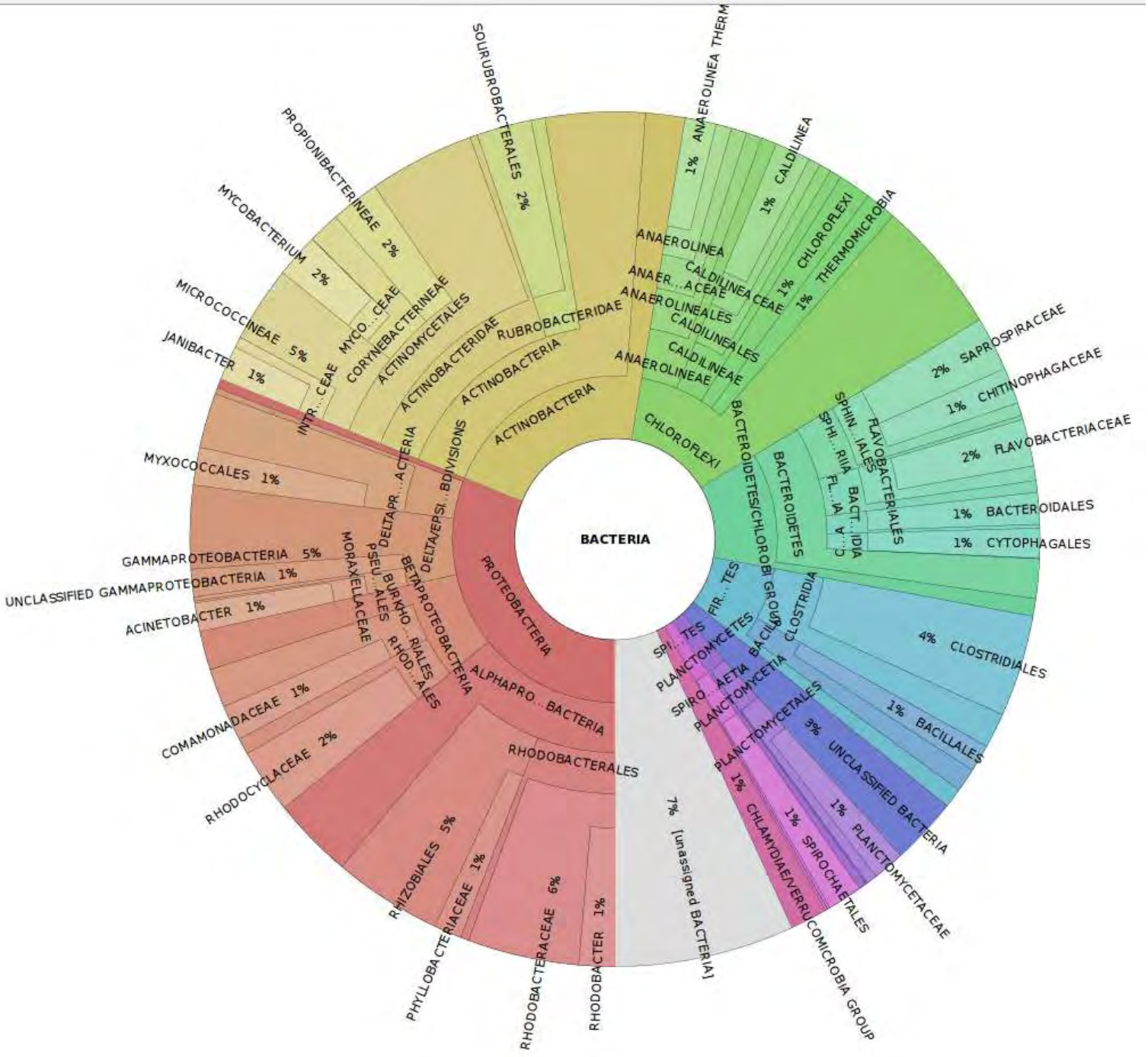
Εικόνα 49. Η οπτικοποίηση με τον Krona viewer για την ταξινόμηση του SRR935359, αρχίζοντας από ένα επίπεδο ακόμη πιο μέσα, το επίπεδο των βακτηρίων

Τα αποτελέσματα του SRR287818_trim.fastq.html στον krona viewer:



Εικόνα 50. Η οπτικοποίηση με τον Krona viewer για την ταξινόμηση του SRR287818

Εικόνα 51. Η οπτικοποίηση με τον Krona viewer για την ταξινόμηση του SRR287818, αρχίζοντας από το επίπεδο των Cellular organisms



Εικόνα 52. Η οπτικοποίηση με τον Krona viewer για την ταξινόμηση του SRR287818, αρχίζοντας από ένα επίπεδο ακόμη πιο μέσα, το επίπεδο των βακτηρίων

Στους φακέλους που δημιουργούνται βρίσκονται και τα αρχεία SRR935359_trim.fastq.xml και SRR287818_trim.fastq.xml. Αυτά για να οπτικοποιηθούν χρειάζονται τη βοήθεια ενός tree viewer του Archaeopteryx. Αυτό το

πρόγραμμα το κατεβάζουμε από εδώ: https://sites.google.com/site/cmzmasek/home/software/archaeopteryx . Αυτό που κατεβαίνει έχει την ονομασία forester_1028.jar και το αποθηκεύουμε χωρίς extraction σε ένα φάκελο που δημιουργούμε μέσα στο Phylogeny και τον ονομάζουμε Archaeopteryx2. Μαζί πρέπει να κατεβάσουμε και το αρχείο _aptx_configuration_file.txt από το ίδιο site. Για να φορτώσουμε όμως ένα δέντρο από το treeDir του output φακέλου (εδώ phylosift_SRR935359 και phylosift_SRR287818) στο Archaeopteryx, πρέπει να χρησιμοποιήσουμε το guppy software kit, το οποίο είχε ήδη κατέβει με το PhyloSift και βρίσκεται στο φάκελο bin του προγράμματος. Το guppy μετατρέπει την κατάληξη των αρχείων του output φακέλου από .jplace σε .xml ή .tre. Για να γίνει αυτό αντιγράφουμε τα αρχεία από το phylosift_SRR935359 και phylosift_SRR287818 στο bin που βρίσκεται το guppy, ας πάρουμε για παράδειγμα το V771.codon.updated.sub1.3.jplace το V771.codon.updated.sub1.1.jplace από τους δύο φακέλους αντίστοιχα και από εκεί ανοίγουμε το terminal και πατάμε τη εντολή:

./guppy fat V771.codon.updated.sub1.3.jplace

και μετά

./guppy fat V771.codon.updated.sub1.1.jplace

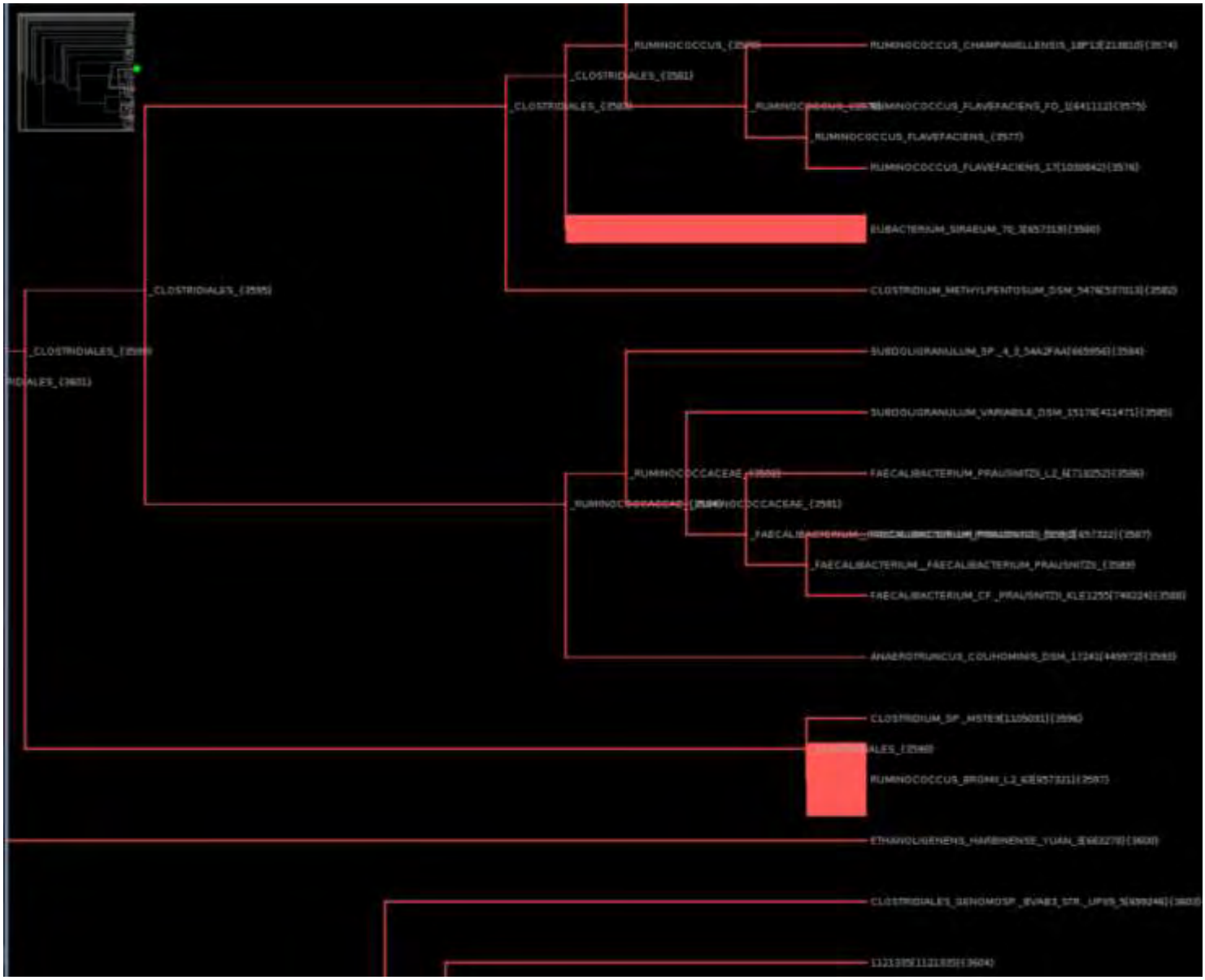
Αυτόματα δημιουργείται στο ίδιο σημείο στο bin ένα αρχείο με ίδιο όνομα αλλά με την κατάληξη .xml (δηδ. V771.codon.updated.sub1.3.xml και V771.codon.updated.sub1.1.xml). Για να το δούμε στο Archaeopteryx πηγαίνουμε και ανοίγουμε το terminal από το φάκελο Archaeopteryx2 και πατάμε την εντολή για να τρέξει το πρόγραμμα Archaeopteryx:

java –jar forester_1028.jar

Έτσι ανοίγει το Archaeopteryx για να δούμε το δέντρο. Για να φορτώσουμε το δέντρο πατάμε File ➔ Read Tree from File και φορτώνουμε το δέντρο που μας ενδιαφέρει (εδώ το V771.codon.updated.sub1.1.xml ή ακόμη και τα

SRR935359_trim.fastq.xml και SRR287818_trim.fastq.xml, που δεν χρειάζονται μετατροπή). Πρέπει να έχουμε τσεκάρει τις πλαϊνές επιλογές "Colorize Branches" και "Use Branch – Widths" για να δούμε σε ποιούς οργανισμούς χτυπάει (συνήθως είναι uncultured, γιατί δύο οργανισμοί για να μπούν στο ίδιο είδος πρέπει να έχουν ομοιότητα > από 97%, αλλιώς μπαίνουν σε άλλο είδος και επειδή δεν ξέρουμε συνήθως σε ποιό καταχωρούνται ως uncultured).

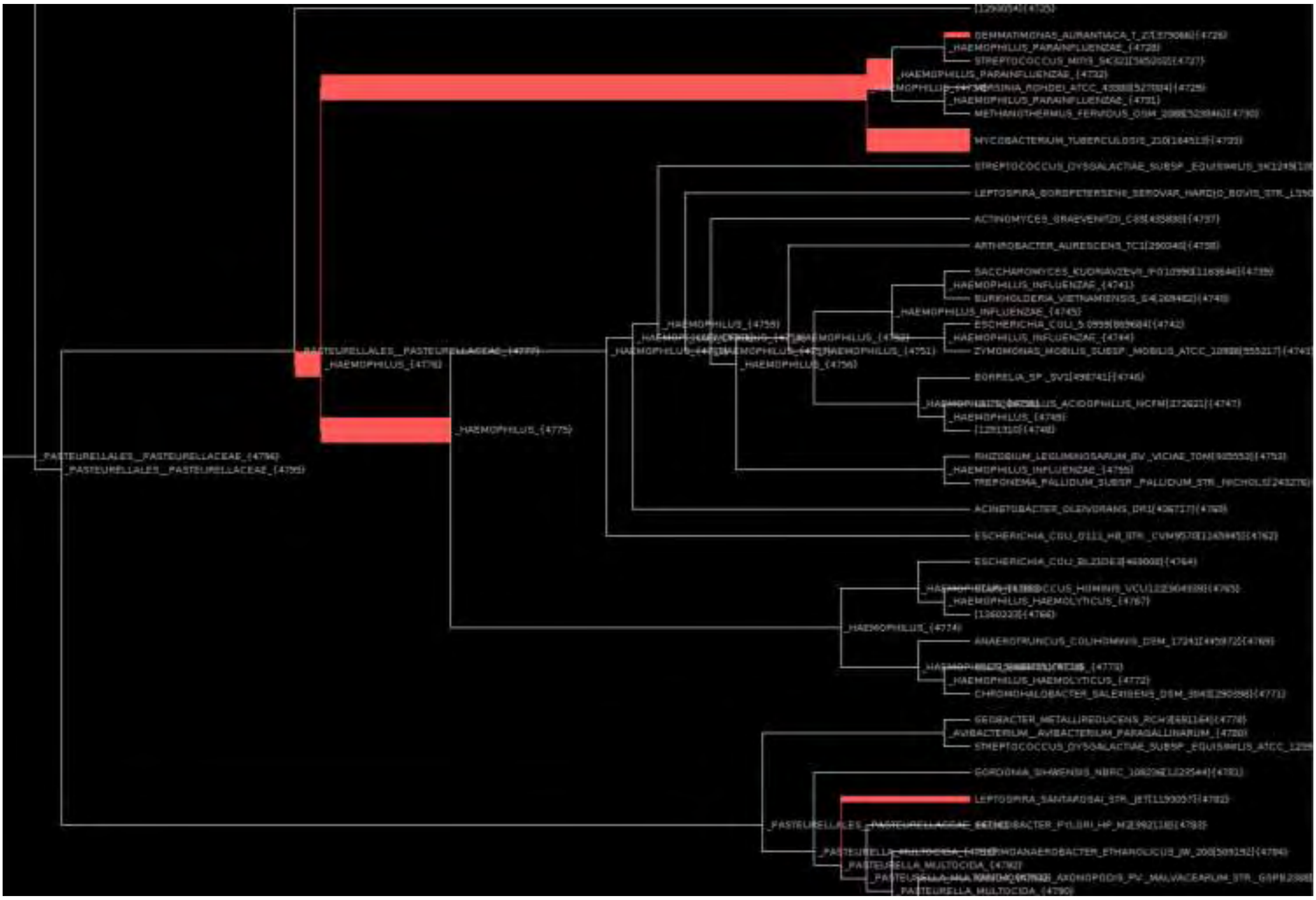
Για το αρχείο SRR935359_trim.fastq.xml εμφανίζεται η ταξονομική απεικόνιση με τη βοήθεια του Archaeopteryx



Εικόνα 53. Zoom in στην ταξονομική απεικόνιση του SRR935359 για να φανούν

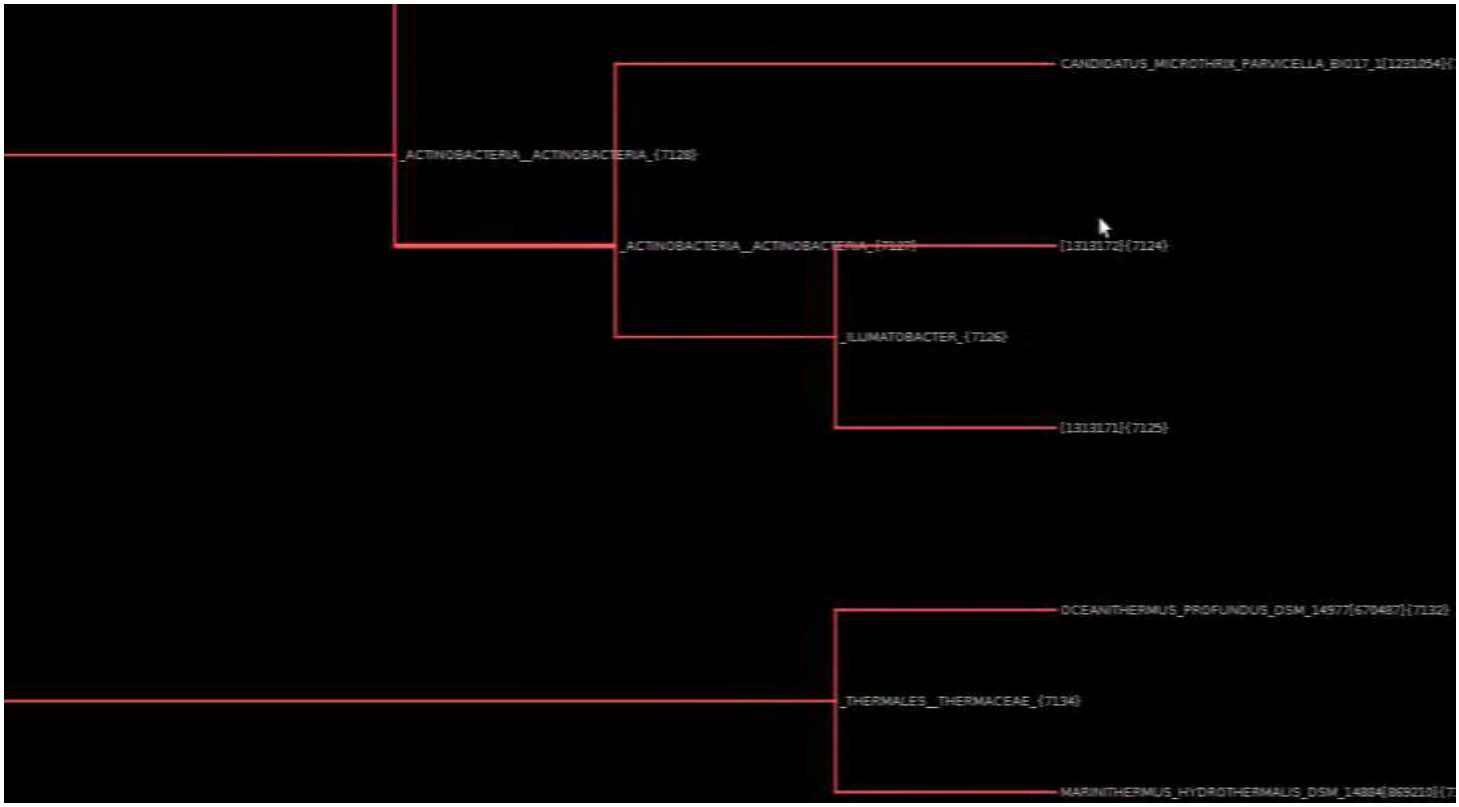
καλύτερα οι αντιθέσεις μεταξύ των κλαδιών με περισσότερα reads (έντονη γραμμή) και αυτών με λιγότερα

Ταξονομική απεικόνιση για έναν τυχαίο marker, τον V771, του SRR935359



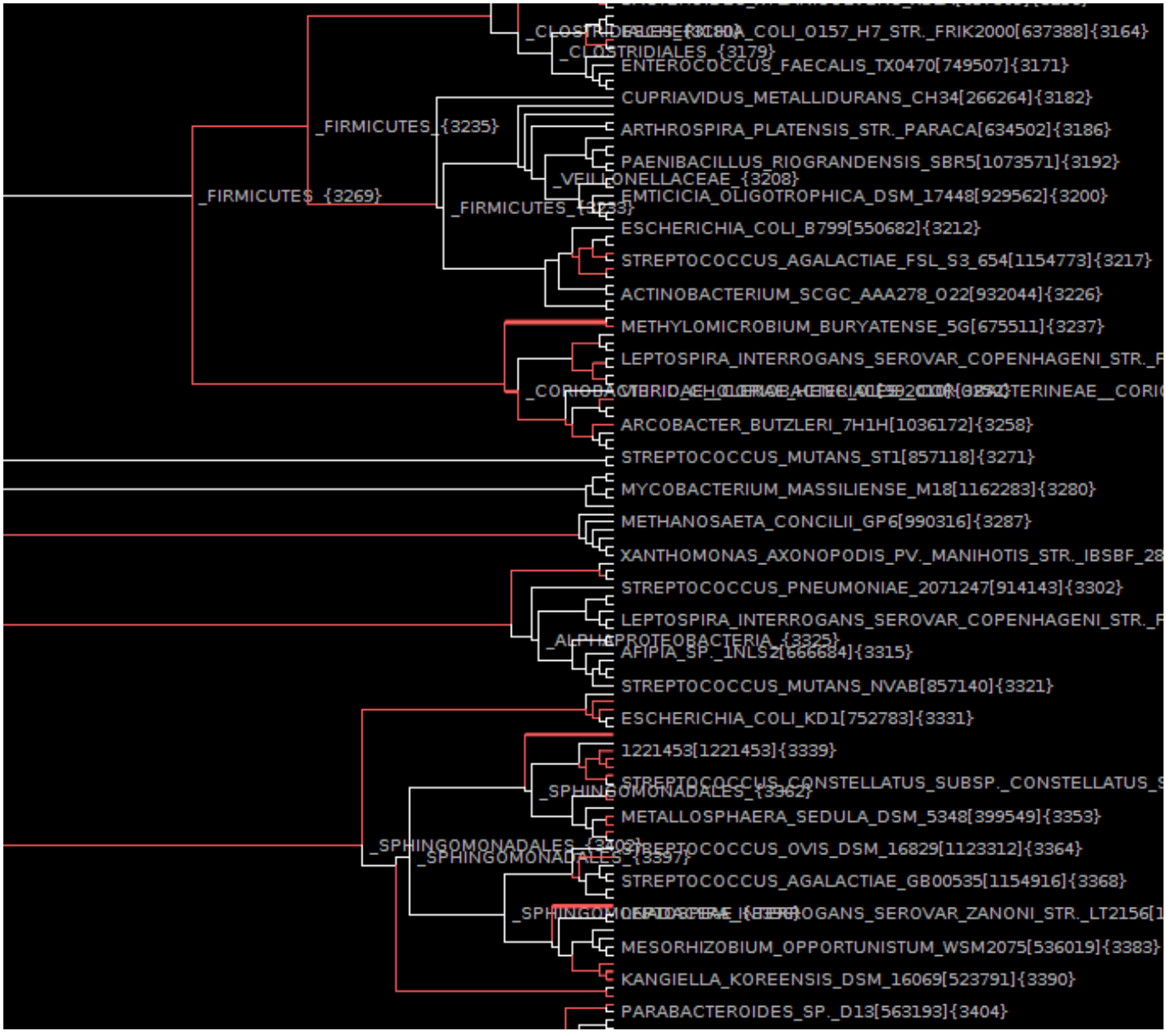
Εικόνα 54. Zoom in στην ταξονομική απεικόνιση του marker V771 του SRR935359 για να φανούν καλύτερα οι αντιθέσεις μεταξύ των κλαδιών με περισσότερα reads (έντονη γραμμή) και αυτών με λιγότερα

Για το SRR287818_trim.fastq.xml εμφανίζεται η ταξινομική απεικόνιση με τη βοήθεια του Archaeopteryx



Εικόνα 55. Zoom in στην ταξινομική απεικόνιση του SRR287818 για να φανούν καλύτερα οι αντιθέσεις μεταξύ των κλαδιών με περισσότερα reads (έντονη γραμμή) και αυτών με λιγότερα

Ταξονομική απεικόνιση για έναν τυχαίο marker, τον V771, του SRR287818.



Εικόνα 56. Zoom in στην ταξονομική απεικόνιση του marker V771 του SRR287818 για να φανούν καλύτερα οι αντιθέσεις μεταξύ των κλαδιών με περισσότερα reads (έντονη γραμμή) και αυτών με λιγότερα

Στις τέσσερις παραπάνω εικόνες φαίνεται η fat visualization που βγάζει τα δέντρα με χρωματιστά άκρα και πιο χοντρά ανάλογα με τον αριθμό των reads που είναι τοποθετημένα στο κάθε κλαδί (επιλεγμένα τα: colourize branches και use branch-widths).

3.8 MEGAN

Κατεβάζουμε το MEGAN_unix_4_70_4.sh (http://ab.inf.uni-tuebingen.de/data/software/megan4/download/welcome.html) και πρέπει πρώτα να το κάνουμε εκτελέσιμο ανοίγοντας το terminal από το σημείο που έχει κατέβει το αρχείο, σύμφωνα με το http://ab.inf.uni-tuebingen.de/data/software/megan4/download/manual.pdf , γράφοντας την εντολή:

chmod +x MEGAN_unix_4_70_4.sh

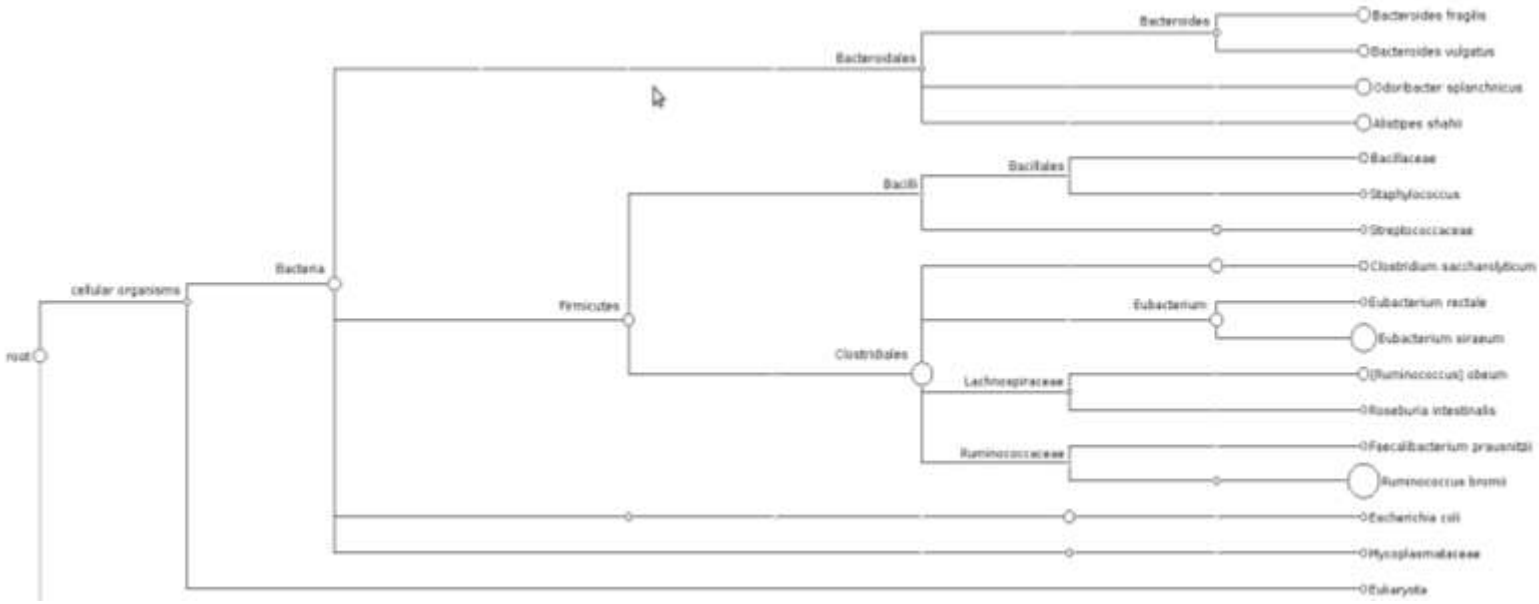
Στη συνέχεια τρέχουμε το πρόγραμμα: ./ MEGAN_unix_4_70_4.sh
και κάνουμε εγκατάσταση του προγράμματος. Αυτό στη συνέχεια, αφού μου ζητήσει το license ανοίγει φορτώνοντας το αρχείο της NCBI taxonomy και δείχνοντάς το δεύτερο επίπεδο της taxonomy σαν ένα rooted tree.

Το MEGAN χρησιμοποιεί τις δικές του μορφές αρχείων για να αποθηκεύσει τα δεδομένα που περιγράφουν το αποτέλεσμα που υπολογίστηκε (από BLASTX, BLASTP ή BLASTN) από μια σύγκριση αλληλουχίας μεταξύ DNA reads και βάσης δεδομένων με αλληλουχίες αναφοράς. Τα αρχεία θα πρέπει να είναι σε RMA format (RMA2 για το MEGAN4 που χρησιμοποιούμε εμείς), δηλαδή να έχουν την κατάληξη .rma (read match archive) που ουσιαστικά είναι ένα binary format. Στο site του MEGAN, από όπου κατεβάσαμε το πρόγραμμα, υπάρχουν κάποια παραδείγματα τέτοιων αρχείων .rma που μπορεί να κατεβάσει ο καθένας για να δει πως λειτουργεί το πρόγραμμα.

Για να φορτώσουμε στο πρόγραμμα αυτό ένα output αρχείο από BLAST θα πρέπει το Blast που θα έχουμε ήδη τρέξει να το έχουμε ρυθμίσει έτσι ώστε η παράμετρος –outfmt να έχει την τιμή 0 ή 7. Πριν φορτώσουμε τα αρχεία που θέλουμε στο megan, πρέπει να τους προσθέσουμε την κατάληξη .blastn, γιατί απαιτείται από το πρόγραμμα αλλιώς δεν μπορεί να τις φορτώσει. Αυτό γίνεται με μετονομασία.

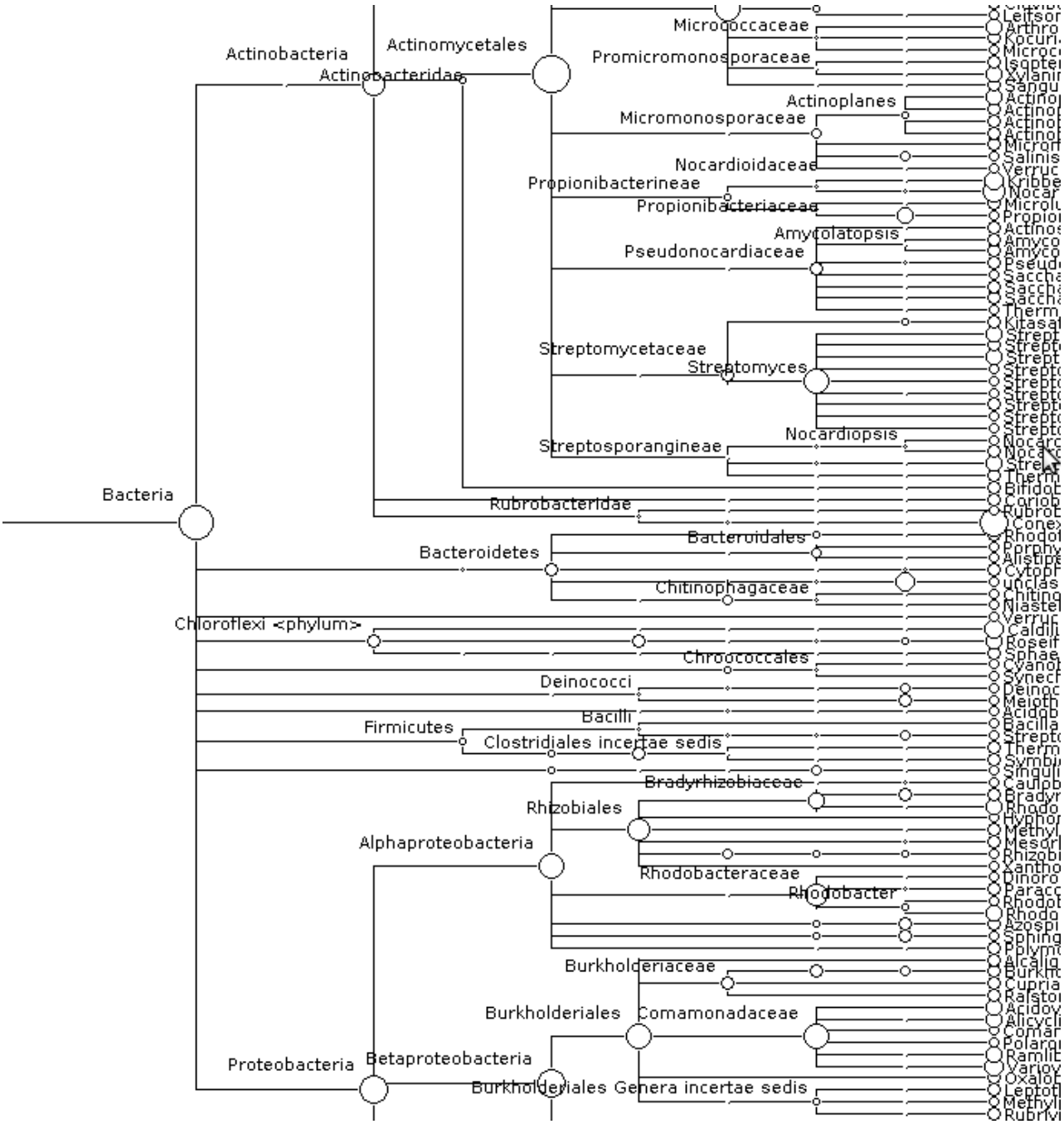
Συνεπώς τώρα έχουμε τα αρχεία blastn_SRR935359.blastn και blastn_SRR287818.blastn.

Ανοίγουμε το Megan από το φάκελο megan. Από το File ➔ Import from BLAST δίνουμε το αρχείο blastn_SRR935359.blastn, στο επόμενο βήμα τη fasta μορφή των reads, δηδ. nuc_ SRR935359.fasta και δημιουργεί, αφού μας ρωτήσει πού, το αρχείο blastn_ SRR935359.rma (εμείς το αποθηκεύσαμε στο bin του ncbi-blast-2.2.27+). Μετά από αυτό, μας εμφανίζει την taxonomy από το σύνολο των δεδομένων της αλληλουχίας που του δώσαμε (αφού το έχουμε κάνει collapse από το Tree➔Collapse At Taxonomic Level➔Species).



Εικόνα 57. Τα αποτελέσματα του MEGAN για την ταξινόμηση του SRR935359 όπου φαίνονται τα κύρια φύλα της βακτηριακής κοινότητας (Bacteroides και Firmicutes)

Ανοίγουμε το Megan από το φάκελο megan. Από το File ➔ Import from BLAST δίνουμε το αρχείο blastn_SRR287818.blastn, στο επόμενο βήμα τη fasta μορφή των reads, δηδ. nuc_ SRR287818.fasta και δημιουργεί, αφού μας ρωτήσει πού, το αρχείο blastn_ SRR287818.rma (εμείς το αποθηκεύσαμε στο bin του ncbi-blast-2.2.27+). Μετά από αυτό, μας εμφανίζει την taxonomy από το σύνολο των δεδομένων της αλληλουχίας που του δώσαμε (αφού πρώτα το έχουμε κάνει collapse από το Tree➔Collapse At Taxonomic Level➔Kingdom και αμέσως μετά collapse από το Tree➔Collapse At Taxonomic Level➔Species).



Εικόνα 58. Τα αποτελέσματα του MEGAN για την ταξινόμηση του SRR287818 όπου

φαίνονται τα κύρια φύλα της βακτηριακής κοινότητας (Actinobacteria, Chloroflexi, Proteobacteria, Bacteroidetes και Firmicutes)

Στο MEGAN οπτικοποιήθηκε και η ταξινομική ανάλυση των αρχείων εκείνων που είχαν προέλθει από την επεξεργασία με το πρόγραμμα MetaVelvet και διαπιστώσαμε ότι φέρουν τα ίδια αποτελέσματα με τις παραπάνω.

3.9 Συνοπτικός πίνακας των προγραμμάτων βιοπηροφορικής και των αποτελεσμάτων τους

Προγράμματα	**Διάρκεια**		**Output**	
	SRR935359 44,311,440 raw reads	**SRR287818 5,841,568 raw reads**	**SRR935359**	**SRR287818**
Condetri	4h		SRR935359_trim.fastq	SRR287818_trim.fastq
			26,669,184 reads	3,859,129 reads
Trinity	15min	10min	SRR935359_Trinity.fasta	SRR287818_Trinity.fasta
			11,577 contigs	25,567 contigs
			Μέσος όρος μήκους: 16,719 b	Μέσος όρος μήκους: 521.85 b
			μεγαλύτερο μήκος contig 33,237 b	μεγαλύτερο μήκος contig 33,247
			μικρότερο 201 b	μικρότερο 201 b
MetaVelvet	15 min	3 min	SRR935359_ meta-velvetg.contigs.fa	SRR287818_ meta-velvetg.contigs.fa
			20,411 contigs	66,820 contigs
			Μέσος όρος μήκους: 302.32 b	Μέσος όρος μήκους: 139.2 b

			μεγαλύτερο μήκος contig 18,887 b	μεγαλύτερο μήκος contig 3,265 b
			μικρότερο 57 b	μικρότερο 57 b
Prodigal From Trinity	3 min	2 min	trans_ SRR935359	trans_ SRR287818
			31,175 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια	32,324 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια
			Μέσος όρος μήκους: 175 a	Μέσος όρος μήκους: 122.6 a
			μεγαλύτερο μήκος 3,163 a	μεγαλύτερο μήκος 1,643 a
			μικρότερο μήκος 20 a	μικρότερο μήκος 20 a
Prodigal From MetaVelvet	5 min	3 min	trans_ SRR935359_2	trans_ SRR287818_2
			21,331 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια	60,730 πρωτεϊνικές αλληλουχίες αντιστοιχισμένες σε γονίδια
			Μέσος όρος μήκους: 84.35 a	Μέσος όρος μήκους: 45 a
			μεγαλύτερο μήκος 2,372 a	μεγαλύτερο μήκος 2,372 a
			μικρότερο μήκος 20 a	μικρότερο μήκος 20 a
Phylosift	52h	14h 30min	SRR935359_trim.fastq.html	SRR287818_trim.fastq.html
			SRR935359_trim.fastq.xml	SRR287818_trim.fastq.xml

Blast from Trinity	40 min	50 min	blastn_SRR935359.blastn	blastn_ SRR287818.blastn
Blast from MetaVelvet	10 min	33 min	blastn_SRR935359_2.blastn	blastn_SRR287818_2.blastn

Πίνακας 3. Απεικόνιση των προραμμάτων και των αποτελεσμάτων τους καθώς και του χρόνου της διαδικασίας που ακολουθήθηκε στην παρούσα εργασία

4. ΣΥΖΗΤΗΣΗ

4.1 Η μεταγονιδιωματική μέχρι σήμερα

Η μεταγονιδιωματική χρησιμοποιείται για τη μελέτη του ολικού γονιδιωματικού DNA που έχει αποκτηθεί από μικροοργανισμούς του περιβάλλοντος, οι οποίοι δεν μπορούν να καλλιεργηθούν στο εργαστήριο και οι οποίοι αποτελούν την πλειοψηφία των μικροοργανισμών πάνω στη Γη. Η συσσώρευση των γνώσεων που παίρνουμε από τις αλληλουχίες αυτών των μικροοργανισμών έχει επεκτείνει σε μεγάλο βαθμό την αντίληψή μας για το δυναμικό των μικροβιακών οργανισμών στη φύση και για το αντίκτυπό τους στο περιβάλλον και την ανθρώπινη υγεία, καθιστώντας τη μεταγονιδιωματική έναν από τους πιο γρήγορα ανερχόμενους επιστημονικούς κλάδους (Harris et al., 2012).

Η συνολική εικόνα που σχηματίζεται με την προσέγγιση αυτή επιτρέπει την ακριβή εξαγωγή συμπερασμάτων για φυλογενετικές σχέσεις. Ένα από τα πιο σημαντικά πλεονεκτήματα της NGS αλληλούχισης είναι ο πλούτος των πληροφοριών που μπορεί να παράγει. Μία αλληλούχιση σε βάθος καθιστά δυνατή την ανίχνευση ακόμα και μελών σε μικρές συγκεντρώσεις, εντός πολύπλοκων πληθυσμών.

Η μεταγονιδιωματική ανάλυση έχει εφαρμοστεί τόσο σε περιβαλλοντικές μελέτες όσο και στην έρευνα βιολογικών δεικτών, ενώ έχει γίνει πρόοδος σε πολλούς τομείς όπως στην ταξινομική κατάταξη των οργανισμών χρησιμοποιώντας το 16S ριβοσωμικό RNA (με αλληλούχιση της κατάλληλα επιλεγμένης περιοχής του γονιδίου). Επίσης, έχει επιτευχθεί η καθιέρωση καταλόγου των μικροβιακών γονιδίων και η μελέτη τους κατευθείαν από το φυσικό περιβάλλον (έδαφος ή νερό), γεγονός που επιτρέπει την ανίχνευση περισσότερων νέων μικροβίων και των γονιδίων τους από τη μικροβιακή κοινότητα. Ακόμη, γίνονται τέτοιου είδους αναλύσεις και για το ανθρώπινο μικροβίωμα σχετίζοντάς το με την εμφάνιση διάφορων ασθενειών.

4.2 Εντοπισμός μεταγονιδιώματος

Ένα προς ανάλυση μεταγονιδίωμα μπορεί να εφαρμοστεί είτε στο 16S rRNA, είτε σε πλασμίδια, είτε στο ολικό μικροβιακό DNA, είτε σε ιϊκό υλικό.

4.2.1 Μεταγονιδίωμα με βάση το 16S ριβοσωμικό RNA

Η εποχή της μεταγονιδιωματικής εγκαινιάστηκε από μελέτες οι οποίες χρησιμοποιούσαν το 16S rRNA ως φυλογενετικό δείκτη των μικροβιακών taxa (Pace, 1997). Το γονίδιο 16S εμφανίζεται σε όλους τους οργανισμούς, εκτός των ιών και αντιπροσωπεύει το 80% του ολικού βακτηριακού RNA. Δομικά αποτελείται από συντηρημένες αλλά και μεταβλητές περιοχές. Με το να εστιάζουμε σε ένα μικρό κομμάτι του μικροβιακού γονιδιώματος, όπως το 16S rRNA, κατεβαίνει το κόστος της αλληλούχισης σημαντικά. Αυτή η προσέγγιση ήταν αρκετά αποτελεσματική στην παρακολούθηση της διακύμανσης των πληθυσμών (Caporaso et al., 2012). Παρ' όλα αυτά, δεν υπάρχει σταθερή σχέση μεταξύ της διατήρησης του 16S rRNA και του υπόλοιπου βακτηριακού γονιδιώματος. Σε πολλά βακτήρια συμβαίνει οριζόντια μεταφορά των λειτουργικών γονιδίων ή ακόμη σημαντικές γονιδιακές ανακατατάξεις, οι οποίες μπορεί να μην μαρτυρώνται μόνο από τη μελέτη της 16S rRNA περιοχής (Altermann, 2012). Επίσης, τα βακτήρια μπορεί να υποστούν οριζόντια μεταφορά ακόμα και ολόκληρων των 16S γονιδίων (Asai et al., 1999). Συνεπώς, η ταυτοποίηση με ανιχνευτές βασισμένους στο γονίδιο 16S ή η ανάλυση που είναι βασισμένη σε κομμάτια της αλληλουχίας του 16S μπορεί να οδηγήσουν σε λανθασμένη αναγνώριση, καθώς ο δείκτης αυτός μπορεί να αντιπροσωπεύει μια δομή που προέκυψε από τέτοιου είδους οριζόντια μεταφορά (Schouls et al., 2003).

4.2.2 Μεταγονιδίωμα που εντοπίζεται στα μικροβιακά πλασμίδια (Plasmidomics)

Τα πλασμίδια συχνά χρησιμεύουν ως μεσολαβητές μεταφοράς γονιδίων και ως μια ισχυρή εξελικτική δύναμη για το μικροβιακό περιβάλλον. Το πλασμιδίωμα έχει σε πολύ υψηλό βαθμό στοιχεία μωσαϊκισμού και δυνατότητα μεταφοράς ανάμεσα στα phyla. Αυτά τα γνωρίσματα των πλασμιδίων οδήγησαν στη μελετη των μεταγονιδιωμάτων τους και παρατηρήθηκε ότι μπορούν να προσδώσουν πλεονεκτήματα στους ξενιστές και ότι παίζουν ρόλο στην προσαρμογή στο περιβάλλον (Brown Kav et al., 2012).

4.2.3 Ιικό μεταγονιδίωμα

Στην ιολογία η NGS έχει γίνει παντοδύναμο εργαλείο που μπορεί να ανιχνεύσει, να αναγνωρίσει και να ποσοτικοποιήσει καινούργιους ιούς σε ένα μόνο

στάδιο (Dunowska et al., 2012). Μάλιστα, έχει αποδειχθεί ότι είναι ευαίσθητη ως μέθοδος για την ανίχνευση μολυσματικών παραγόντων που συνδέονται με τους ανθρώπινους ιστούς. Σε μέτριο βάθος αλληλούχισης, ικά μεταγραφήματα μπορούν να ανιχνευθούν σε συχνότητες κάτω από 1 σε 1,000,000 *(Moore et al.,* 2011). Μία από τις ευτυχείς συνέπειες της αλληλούχισης σε βάθος, είναι η συμπτωματική αλληλούχιση ικού DNA ή RNA, που οδήγησε στην ανακάλυψη αρκετών νέων ιών (Li et al., 2011).

4.3 Μεταγονιδιώματα με μεγάλο ενδιαφέρον

Η μεταγονιδιωματική ως τεχνολογία μας έχει επιτρέψει να κατανοήσουμε καλύτερα τι συμβαίνει με την ανθρώπινη εντερική χλωρίδα καθώς επίσεις και το τι συμβαίνει σε διάφορα περιβάλλοντα.

4.3.1 Ανθρώπινο εντερικό μεταγονιδίωμα

Οι άνθρωποι φέρουν 10 φορές περισσότερα βακτηριακά κύτταρα από ότι τα δικά τους και 100 φορές περισσότερα βακτηριακά γονίδια απ'ότι τα κληρονομούμενά τους *(Kuczynski et al., 2012)*. Η επίπτωση στην ανθρώπινη υγεία είναι τόσο μεγάλη ώστε το ανθρώπινο μικροβίωμα μπορεί να θεωρηθεί ένα επιπλέον όργανο. Οι αλληλεπιδράσεις ξενιστή- γονιδίων – μικροβίων αποτελούν καθοριστικούς παράγοντες για την ανάπτυξη ορισμένων πολυπαραγοντικών χρόνιων παθήσεων *(Vaarala et al., 2008)*. Σχεδόν 100 τρισεκατομμύρια βακτήρια ζούν και επιβιώνουν μέσα στον ανθρώπινο γαστρεντερικό σωλήνα, ενώ τα αρχαία, οι μύκητες και οι ιοί αντιπροσωπεύουν μικρότερο κομμάτι αυτής της μικροβιακής κοινότητας. Οι διαταραχές των πληθυσμών της μικροχλωρίδας του εντέρου έχουν δειχθεί να σχετίζονται με ασθένειες όπως η παχυσαρκία, ο διαβήτης, και η φλεγμονώδης νόσος του εντέρου.

The gene catalogue of the human gut microbiome[48]	
Number of genes in the study	3,300,000
Number of genes shared by less than 20% of individuals	2,376,000
Core genes (shared by at least 50% of individuals)	294,000
Average number of genes carried per individual	536,000
Number of shared genes (with at least one other individual)	204,000
Bacteria	99.1%
Archaea	0.8%
Eukaryotic or viral origin	0.1%
Number of species present in each individual	160

Εικόνα 59. Κατάλογος των γονιδίων που βρέθηκαν στο ανθρώπινο εντερικό μικροβίωμα (Qin et al., 2010)

Αν και το μικροβίωμα του εντέρου έχει λάβει πολλή προσοχή, υπάρχει ένας αυξανόμενος αριθμός μελετών και σε άλλες ανθρώπινες μικροχλωρίδες, όπως αυτές του κόλπου και του δέρματος. Στο Human Microbiome Project (http://commonfund.nih.gov/hmp/index) μελετήθηκαν πολλές διαφορετικές θέσεις στο ανθρώπινο σώμα, συμπεριλαμβανομένων των ρινικών διόδων, της στοματικής κοιλότητας, του δέρματος, του γαστρεντερικού σωλήνα, και της ουρογεννητικής οδού. Αυτή η μελέτη έχει αυξήσει την επίγνωση της πολυπλοκότητας και της σημασίας του μεταγονιδιώματος στην ανθρώπινη υγεία.

4.3.2 Μικροβίωμα και περιβάλλον

Έχει εκτιμηθεί ότι υπάρχουν μεταξύ 100.000 και 1.000.000 διαφορετικά είδη μικροβίων ανά γραμμάριο εδάφους *(Curtis and Sloan,* 2005), αλλά μόνο ένα μικρό μέρος τους μπορεί να καλλιεργηθεί στο εργαστήριο.

Η βιοαποκατάσταση είναι μια φιλική προς το περιβάλλον και οικονομικά ανταγωνιστική στρατηγική για την εξάλειψη των ξενοβιοτικών ενώσεων από τα μολυσμένα περιβάλλοντα. Η NGS παρέχει ζωτικής σημασίας γνώσεις για τους μοριακούς και βιολογικούς μηχανισμούς που εμπλέκονται στην βιοαποικοδόμηση. Αυτή η προσέγγιση αναμένεται πως θα βελτιώσει τις μικροβιακές βιοθεραπευτικές στρατηγικές, παρακολουθώντας την πρόοδό τους και καθορίζοντας την επιτυχία τους *(Desai et al.,* 2010).

Τα μικροβιακά ένζυμα έχουν πολλές γνωστές εφαρμογές ως βιολογικοί καταλύτες. Ωστόσο, μόνο λίγα χρησιμοποιούνται αυτή τη στιγμή για βιοκατάλυση,

παρά την πληθώρα των μεταγονιδιωματικών ακολουθιών που είναι διαθέσιμες σε βάσεις δεδομένων (Fernández-Arrojo et al., 2010).

4.3.3 Ιοί & Βακτηριοφάγοι

Η αλληλούχιση ρουτίνας του γενετικού υλικού των ιών έχει οδηγήσει σε ένα μεγάλο αριθμό ιικών γονιδιωμάτων, όπου αποκαλύπτεται η αξιοσημείωτη μεταβλητότητα των ιών. Ο υψηλός βαθμός μετάλλαξης των RNA ιών προκύπτει από τις επιρρεπείς σε σφάλματα πολυμεράσες και τις περιορισμένες λειτουργίες της RNA επιδιόρθωσης. Έχει δειχθεί από πρόσφατες εργασίες πως η υψηλή μεταλλακτικότητα των ιών είναι απαραίτητη για την προσαρμοστική εξέλιξη και την ικανότητά τους να προκαλούν ασθένειες. Η αλληλούχιση των ιικών mRNAs μπορεί να προσφέρει πλούτο πληροφοριών σχετικά με την ιική δραστηριότητα, καθώς και τους μηχανισμούς δράσης τους. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν αργότερα για το σχολιασμό του ιϊκού γονιδιώματος. Η ιϊκή μεταγονιδιωματική επεκτείνει την τρέχουσα γνώση μας για τις αλληλεπιδράσεις ιού - ξενιστή αποκαλύπτοντας γονίδια που μπορούν να μεταχειριστούν τους ξενιστές με διάφορους και συχνά μη αναμενόμενους τρόπους. Η δυνατότητα ανακάλυψης των ιών μέσω της ιϊκής μεταγονιδιωματικής μπορεί να βοηθήσει ένα ευρύ φάσμα επιστημονικών κλάδων, συμπεριλαμβανομένων της εξελικτικής βιολογίας, της επιτήρησης των παθογόνων παραγόντων, και της βιοτεχνολογίας (Hann et al., 2010). Εκτός από τη βελτίωση της ανίχνευσης των ιών που προκαλούν ασθένειες, οι γονιδιωματικές μέθοδοι αποκάλυψαν τη σημασία των ιών και σε υγιή άτομα. Το ανθρώπινο ιϊκό μεταγονιδίωμα υποδηλώνει ότι οι άνθρωποι είναι σχεδόν συνεχώς εκτεθειμένοι σε ιούς, οι οποίοι ενδέχεται και να προκαλέσουν συμπτώματα. Έτσι προκύπτει ότι το ιϊκό μεταγονιδίωμα είναι ένα σημαντικό συστατικό του περιβάλλοντος που μπορεί να αλληλεπιδράσει με τα γενετικά χαρακτηριστικά του ξενιστή για να συμβάλλει στην εμφάνιση ασθενειών.

Οι ιοί των προκαρυωτών, δηλαδή οι φάγοι, είναι μία από τις μεγαλύτερες δεξαμενές της ανεξερεύνητης γενετικής ποικιλομορφίας στη Γη. Οι βακτηριοφάγοι αποτελούν την απόλυτη πλειοψηφία του συνόλου των οργανισμών στη βιόσφαιρα. Η γενετική ποικιλότητα του πληθυσμού είναι πολύ υψηλή. Συχνά η οριζόντια μεταφορά γενετικού υλικού έχει ως αποτέλεσμα τον μωσαϊκισμό. Οι βακτηριοφάγοι παίζουν σημαντικό ρόλο στη διαμόρφωση βακτηριακών πληθυσμών. Με την έλευση της

NGS, τα επόμενα χρόνια η εξερεύνηση του γονιδιώματος των φάγων υπόσχεται να δώσει αποτελέσματα που θα είναι ιδιαίτερα αποκαλυπτικά.

Τα γονιδιώματα των φάγων είναι πολύ μικρότερα από ότι τα κυτταρικά των ξενιστών τους. Ως αποτέλεσμα, η ενίσχυση του DNA είναι συνήθως ένα αναγκαίο βήμα πριν την αλληλούχιση του μεταγονιδιώματος, η οποία ενισχύει το προϊόν σημαντικά (McDaniel et al., 2014). Παρ'όλους τους περιορισμούς που υπάρχουν, το ιικό μεταγονιδίωμα του θαλάσιου περιβάλλοντος έδωσε μια γενική εικόνα για το θαλάσσιο ιικό κόσμο (Sharon et al., 2011). Ως εκ τούτου, το μεγαλύτερο μέρος της γνώσης μας σχετικά με τα πλήρη γονιδιώματα των θαλάσσιων φάγων προέρχεται από καλλιεργημένους εκπροσώπους, απομονωμένους μόνο λόγω της επιτυχίας στην καλλιέργεια των ξενιστών (Rappé et al., 2002) (*Zhao et al.,* 2013).

4.3.4 Θαλάσσιο περιβάλλον

Τα θαλάσσια περιβάλλοντα, συμπεριλαμβανομένου του υπεδάφους τους, περιέχουν συνολικά περίπου 3.67 $*10^{30}$ μικροοργανισμούς *(Whitman et al.,* 1998). Το 71% της επιφάνειας της γης καλύπτεται από ωκεανό. Αυτό το περιβάλλον αντιπροσωπεύει το 80% της ζωής πάνω στη γη καθώς επίσης και ένα τεράστιο απόθεμα της βιοποικιλότητας και της εκμεταλλεύσιμης μικροβιακής βιοτεχνολογίας (Kennedy et al., 2008).

Η κλωνοποίηση του περιβαλλοντικού DNA σε φοσμίδια χρησιμοποιείται επιτυχώς για τη μελέτη των κομματιών προκαρυωτικών γονιδιωμάτων ακαλλιέργητων μικροβίων και του θαλάσσιου περιβάλλοντος, προσπερνώντας εντελώς την καλλιέργεια του ξενιστή *(Ghai et al.,* 2010), ενώ αποτελεί μια εναλλακτική λύση για τη δημιουργία ολόκληρων γονιδιωμάτων των φάγων (Mizuno et al., 2013). Αυτό κατέστη εφικτό χάρη στην παρατήρηση ότι τα κομμάτια του μεταγονιδιωματικού DNA που εισάγονται στα φοσμίδια δείχνουν στα αποτελέσματά τους σημαντική εκπροσώπιση των γονιδιωματικών κομματιών των φάγων (DeLong et al., 2006). Στην πραγματικότητα, ένας αντιγραφόμενος φάγος κατά τη διάρκεια του λυτικού κύκλου του σε ένα κύτταρο, παρέχει μια φυσική ενίσχυση που θυμίζει εργαστηριακή κλωνοποίηση ή άλλες μεθόδους πολλαπλασιασμού του γενώματος (Dean et al., 2001). Παλαιότερα, τα μεταγονιδιωματικά φοσμίδια είχαν αποδειχθεί ότι μπορούν να δείξουν καταγωγές των θαλάσσιων φάγων όπως των κυανοφάγων (cyanophages) (Mizuno et al., 2013) και των SAR11 ιών (Zhao et al., 2013).

Σε μια πολλή πρόσφατη έρευνα Mizuno et al., έχει ανακαλυφθεί ότι τα κυτταρικά μεταγονιδιώματα άκρως παραγωγικών νερών των ωκεανών (με το μέγιστο βάθος της χλωροφύλλης) περιέχουν σημαντικές ποσότητες ιικού DNA που προέρχεται από κύτταρα που υφίστανται λυτικό κύκλο. Με το πλεονέκτημα αυτού του φαινομένου, χρησιμοποιήθηκαν τα μεταγονιδιωματικά φοσμίδια που περιείχαν το ιικό DNA των δειγμάτων της Μεσογείου με το μέγιστο βάθος της χλωροφύλλης. Αυτή η μέθοδος επέτρεψε την περιγραφή ολόκληρων των γονιδιωμάτων 208 νέων θαλάσσιων φάγων. Η ποικιλομορφία αυτών των γονιδιωμάτων ήταν αξιοσημείωτη, συνεισφέροντας 21 γονιδιωματικές ομάδες βακτηριοφάγων εκ των οποίων οι 10 είναι εντελώς νέες. Επιπλέον, κατάφεραν να αντιστοιχίσουν τους ξενιστές σε πολλούς από αυτούς. Αυτοί οι ξενιστές που προβλέφθηκαν αντιπροσωπεύουν μια μεγάλη ποικιλία σημαντικών θαλάσσιων μικροβίων (π.χ. μέλη των κλαδιών SAR11 και SAR116, Cyanobacteria αλλά και των πρόσφατα σχολιασμένων χαμηλών σε περιεχόμενο GC ακτινοβακτηρίων). Επίσης, ένα μεταγονιδίωμα κατασκευασμένο από το ίδιο ενδιαίτημα, έδειξε ότι πολλά από τα γονιδιώματα των νέων φάγων, εκπροσωπούνταν σε αφθονία. Άλλα διαθέσιμα ιικά μεταγονιδιώματα έδειξαν επίσης ότι ορισμένοι από τους νέους φάγους είναι κατανεμημένοι σε χαμηλά έως μέτρια βάθη νερών των ωκεανών. Τα διαθέσιμα γονιδιώματα επιτρέπουν μια άμεση προσέγγιση του πληθυσμού του ιογενούς μεταγονιδιώματος επιβεβαιώνοντας το μωσαϊκισμό των γονιδιωμάτων των φάγων (Mizuno et al., 2013).

4.4 Σχολιασμός αποτελεσμάτων της εργασίας

Στην παρούσα εργασία ακολουθήσαμε δύο οδούς για την ανάλυση των δεδομένων μας που διέφεραν στο πρόγραμμα με το οποίο κάναμε τη συναρμολόγιση σε contigs. Χρησιμοποιήσαμε γι’ αυτό το σκοπό το πρόγραμμα Trinity τη μια φορά και το MetaVelvet την άλλη. Η απόκλιση στα αποτελέσματά μας ήταν πολύ μικρή καθώς το Trinity μας έδωσε λιγότερα και μεγαλύτερα σε μήκος contigs, ενώ το MetaVelvet λίγο περισσότερα και μικρότερα σε μήκος. Η διαφορά τους ίσως οφείλεται στο γεγονός ότι το MetaVelvet εντοπίζει και αφαιρεί χιμαιρικές αλληλουχίες που μπορεί να δημιουργούνται ενώ το Trinity όχι. Πάντως τα ταξινομικά αποτελέσματα δεν είχαν διαφορές ως προς τα κυρίαρχα φύλα που προέκυψαν.

Τα ανθρώπινα κόπρανα περιέχουν τουλάχιστον 10^{-9} σωματίδια, που μοιάζουν με ιούς, ανά γαμμάριο. Πολλά από αυτά αναγνωρίζονται σαν ιοί που μολύνουν

βακτήρια (βακτηριοφάγοι), αλλά η πλειοψηφία τους παραμένει άγνωστη. Ακόμη και σήμερα, δείγματα ανθρώπινου εντερικού ιώματος δίνουν κυρίως καινούργιους ιούς και η μειοψηφία δίνει ORFs που έχουν ομολογία με γονίδια που έχουν ξαναμελετηθεί.

Οι βακτηριοφάγοι έχουν βιοϊατρική σημαντικότητα καθώς έχουν την ικανότητα να μεταφέρουν τα γονίδιά τους στον ξενιστή παρέχοντας έτσι αυξημένη παθογένεια, ανθεκτικότητα στα αντιβιοτικά, και ίσως νέα μεταβολική ικανότητα. Παρά τη σημασία τους, οι δυνάμεις διαφοροποίησης των γονιδιωμάτων βακτηριοφάγων του ανθρώπινου ξενιστή δεν έχουν μελετηθεί με κάθε λεπτομέρεια. Οι άνθρωποι δείχνουν σημαντική διαφοροποίηση από άτομο σε άτομο όσον αφορά τη εντερική βακτηριακή τους χλωρίδα. Οι μεγάλες διαφορές στους πληθυσμούς των φάγων, όμως, ανάμεσα στα άτομα μπορούν να επηρεαστούν και από την ιϊκή εξέλιξη που συμβαίνει μέσα στον ξενιστή.

Η συγκεκριμένη εργασία από την οποία κατεβάσαμε το μεταγονιδίωμα, διεξάχθηκε για τη μελέτη της προέλευσης και της φύσης των ιϊκών πληυθσμών που υπάρχουν στο ανθρώπινο έντερο, με δείγμα από τα κόπρανα ενός ατόμου, καθώς και για το χαρακτηρισμό των βακτηρίων του ξενιστή (Minot et al., 2013).

Τα απότελέσματα τη δικής μας εργασίας συνέπεσαν με τα αποτελέσματα της πηγής όσον αφορά στα βακτήρια του ξενιστή. Και εμείς όπως και οι Minot et al., 2013, βρήκαμε χρησιμοποιώντας τα προγράμματα Phylosift και MEGAN πως τα μέλη των φύλων Bacteroides και Firmicute είναι τα πιο άφθονα μέλη της κοινότητας.

Για το δεύτερο μεταγονιδίωμα με το οποίο ασχοληθήκαμε, πρέπει να αναφέρουμε ότι η ευρεία χρήση αντιβιοτικών έχει ως αποτέλεσμα την εκρροή τους στο περιβάλλον, την ανάπτυξη αντοχής στα αντιβιοτικά μέσα σε νοσοκομειακά λύματα, στα αστικά λύματα και στην κοπριά ζωϊκής προέλευσης. Τα αστικά λύματα από τις αποχετεύσεις περιλαμβάνουν εντερικά βακτήρια που έχουν εκτεθεί προηγουμένως σε αντιβιοτικά, έτσι οι εγκαταστάσεις επεξεργασίας λυμάτων (sewage treatment plants - STPs) θεωρούνται ως σημαντικές δεξαμενές για γονίδια ανθεκτικότητας σε αντιβιοτικά (antibiotic resistance genes - ARGs). Γενετικά στοιχεία που μπορούν να μεταφερθούν π.χ. πλασμίδια, τρανσποζόνια κ.λ.π., συχνά εμπλέκονται στην οριζόντια μεταφορά των ARGs μέσα στα βακτήρια. Με τη βοήθεια των αλληλουχιών εισαγωγής ISs (insertion sequences), τα τρανσποζόνια μεταπηδούν τυχαία και περιστασιακά σε γονιδίωμα ή πλασμίδιο και αυτό έχει σαν αποτέλεσμα τη

μεταφορά γονιδίων προκαλώντας πολλαπλή ανθεκτικότητα. Ακόμη μπορεί τα integrons να ενσωματώσουν και να μεταφέρουν κασέττες γονιδίων ανθεκτικότητας στη μεταβλητή περιοχή τους και να διευκολύνουν τη μεταφορά των γονιδίων αυτών. Integrons που μεταφέρουν πολλαπλά ARGs συχνά εντοπίζονται σε περιβάλλον από STPs.

Η ενεργή λυματολάσπη και τα βιοφίλμ στα STPs χαρακτηρίζονται από υψηλή μικροβιακή πυκνότητα και ποικιλία, η οποία μπορεί να διευκολύνει την οριζόντια μεταφορά ενός ARG. Επιπλέον, η πλασμιδιακή ποικιλότητα είναι πολύ υψηλή στις μικροβιακές κοινότητες των STPs. Η μεταγονιδιωματική προσέγγιση είναι απαραίτητη για την κατανόηση των πλασμιδίων που υπάρχουν σε STP μικροοργανισμούς.

Στην εργασία από την οποία κατεβάσαμε το δεύτερο μεταγονιδίωμα, οι ερευνητές απομόνωσαν πλασμίδια από μικροβιακές κοινότητες ενεργής λυματολάσπης ενός STP στο Hong Kong, το αλληλούχισαν (high-throughput sequencing) και έκαναν μεταγονιδιωματική ανάλυση για τη μελέτη της ποικιλίας και της σχετικής αφθονίας των ARGs, των ϊκών παραγόντων ή virulent factors (VFs) και των γενετικών στοιχίων που μπορούν να μετατεθούν (Zhang et al., 2011).

Τα αποτελέσματα τη δικής μας ανάλυσης συνέπεσαν με τα αποτελέσματα της πηγής όσον αφορά στα βακτήρια του ξενιστή. Και εμείς όπως και οι Zhang et al., 2011, βρήκαμε χρησιμοποιώντας τα προγράμματα Phylosift και MEGAN πως η βακτηριακή κοινότητα στην ενεργή λυματολάσπη αποτελείται κυρίως από τα μέλη των φύλων Actinobacteria, Chloroflexi, Proteobacteria, Bacteroidetes και Firmicutes.

4.5 Μελλοντικές Προοπτικές

Κατά τη διάρκεια των τελευταίων ετών, η μεταγονιδιωματική έχει επιταχύνει την κατανόηση της μικροβιακής οικολογίας και εξέλιξης, χάρη στις τεχνικές προόδους των πλατφορμών αλληλούχισης και των πιο εξελιγμένων προγραμμάτων βιοπληροφορικής. Η μεταγονιδιωματική είναι αναμφίβολα ένα εξαιρετικό εργαλείο, που συμπληρώνει τις γνώσεις μας γύρω από τη βιόσφαιρα και συγκεκριμένα διερευνά τόσο το ερώτημα του «ποιος είναι εκεί έξω» όσο και του «τι κάνει» στη φύση. Είναι αλήθεια ότι η μελέτη της μεταγονιδιωματικής εξακολουθεί να αναγνωρίζεται από κάποιους μικροβιακούς οικολόγους ως ένα εργαλείο που είναι δύσκολο να αντιμετωπιστεί λόγω της πολυπλοκότητάς του και του τεράστιου όγκου

των δεδομένων. Ωστόσο, δεν είναι καθόλου απίθανο να κάνουμε μεταγονιδιωματική ανάλυση στο μέλλον σε ρυθμό τέτοιο, όπως κάνουμε και την PCR ρουτίνας και την ηλεκτροφότηση σήμερα στο εργαστήριο. Η τρέχουσα ανάπτυξη εργαλείων και βάσεων δεδομένων για μεταγονιδιωματικές μελέτες βρίσκεται ακόμη σε πολύ πρώιμη φάση και υπάρχουν ακόμη πολλές προκλήσεις που πρέπει να ξεπεραστούν. Δεν θα πάρει καιρό για να δούμε την εκρηκτική ανάπτυξη της μεταγονιδιωματικής.

5. ΒΙΒΛΙΟΓΡΦΙΑ

Altermann, E. (2012). Tracing lifestyle adaptation in prokaryotic genomes. Front. Microbiol. *3*, 48.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anselmetti, D. (2012). Nanopores: Tiny holes with great promise. Nat. Nanotechnol. *7*, 81–82.

Asai, T., Zaporojets, D., Squires, C., and Squires, C.L. (1999). An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc. Natl. Acad. Sci. U. S. A. *96*, 1971–1976.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. *13*, R122.

Brady, A., and Salzberg, S.L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods *6*, 673–676.

Brown Kav, A., Sasson, G., Jami, E., Doron-Faigenboim, A., Benhar, I., and Mizrahi, I. (2012). Insights into the bovine rumen plasmidome. Proc. Natl. Acad. Sci. U. S. A. *109*, 5452–5457.

Caporaso, J.G., Paszkiewicz, K., Field, D., Knight, R., and Gilbert, J.A. (2012). The Western English Channel contains a persistent microbial seed bank. ISME J. *6*, 1089–1093.

Curtis, T.P., and Sloan, W.T. (2005). Microbiology. Exploring microbial diversity--a vast below. Science *309*, 1331–1333.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., 4th, Bik, H.M., and Eisen, J.A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ *2*, e243.

Dean, F.B., Nelson, J.R., Giesler, T.L., and Lasken, R.S. (2001). Rapid amplification

of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. *11*, 1095–1099.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.-U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science *311*, 496–503.

Desai, C., Pathak, H., and Madamwar, D. (2010). Advances in molecular and "-omics" technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites. Bioresour. Technol. *101*, 1558–1569.

Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K., and Nattkemper, T.W. (2009). TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics *10*, 56.

Dunowska, M., Biggs, P.J., Zheng, T., and Perrott, M.R. (2012). Identification of a novel nidovirus associated with a neurological disease of the Australian brushtail possum (Trichosurus vulpecula). Vet. Microbiol. *156*, 418–424.

Fernández-Arrojo, L., Guazzaroni, M.-E., López-Cortés, N., Beloqui, A., and Ferrer, M. (2010). Metagenomic era for biocatalyst identification. Curr. Opin. Biotechnol. *21*, 725–733.

Flohr, H., and Breull, W. (1975). Effect of etafenone on total and regional myocardial blood flow. Arzneimittelforschung. *25*, 1400–1403.

Gerlach, W., and Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. *39*, e91.

Ghai, R., Martin-Cuadrado, A.-B., Molto, A.G., Heredia, I.G., Cabrera, R., Martin, J., Verdú, M., Deschamps, P., Moreira, D., López-García, P., et al. (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J. *4*, 1154–1166.

Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C.T., Brown, C.T., Desai, N., Eisen, J.A., Evers, D., Field, D., et al. (2010a). Meeting report: the terabase

metagenomics workshop and the vision of an Earth microbiome project. Stand. Genomic Sci. *3*, 243–248.

Gilbert, J.A., Meyer, F., Jansson, J., Gordon, J., Pace, N., Tiedje, J., Ley, R., Fierer, N., Field, D., Kyrpides, N., et al. (2010b). The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. Stand. Genomic Sci. *3*, 249–253.

Glass, E.M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb. Protoc. *2010*, pdb.prot5368.

Goll, J., Rusch, D.B., Tanenbaum, D.M., Thiagarajan, M., Li, K., Methé, B.A., and Yooseph, S. (2010). METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics. Bioinforma. Oxf. Engl. *26*, 2631–2632.

Gori, F., Folino, G., Jetten, M.S.M., and Marchiori, E. (2011). MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. Bioinforma. Oxf. Engl. *27*, 196–203.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. *33*, D121–124.

Handelsman, J. (2005). Sorting out metagenomes. Nat. Biotechnol. *23*, 38–39.

Hann, D.R., Gimenez-Ibanez, S., and Rathjen, J.P. (2010). Bacterial virulence effectors and their activities. Curr. Opin. Plant Biol. *13*, 388–393.

Harris, S.R., Clarke, I.N., Seth-Smith, H.M.B., Solomon, A.W., Cutcliffe, L.T., Marsh, P., Skilton, R.J., Holland, M.J., Mabey, D., Peeling, R.W., et al. (2012). Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. Nat. Genet. *44*, 413–419, S1.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S.,

Clark, D.S., Chen, F., Zhang, T., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science *331*, 463–467.

Hoff, K.J., Tech, M., Lingner, T., Daniel, R., Morgenstern, B., and Meinicke, P. (2008). Gene prediction in metagenomic fragments: a large scale machine learning approach. BMC Bioinformatics *9*, 217.

Hoff, K.J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res. *37*, W101–105.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res. *21*, 1552–1560.

Idury, R.M., and Waterman, M.S. (1995). A new algorithm for DNA sequence assembly. J. Comput. Biol. J. Comput. Mol. Cell Biol. *2*, 291–306.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Kelley, D.R., Liu, B., Delcher, A.L., Pop, M., and Salzberg, S.L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. *40*, e9.

Kennedy, J., Marchesi, J.R., and Dobson, A.D. (2008). Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. Microb. Cell Factories *7*, 27.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. *21*, 487–493.

Kim, M., Lee, K.-H., Yoon, S.-W., Kim, B.-S., Chun, J., and Yi, H. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. Genomics Inform. *11*, 102–113.

Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. *40*, D54–56.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinforma. Oxf. Engl. *25*, 2737–2738.

Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D., and Knight, R. (2012). Experimental and analytical tools for studying the human microbiome. Nat. Rev. Genet. *13*, 47–58.

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100–3108.

Lai, B., Ding, R., Li, Y., Duan, L., and Zhu, H. (2012). A de novo metagenomic assembly program for shotgun DNA reads. Bioinforma. Oxf. Engl. *28*, 1455–1462.

Laserson, J., Jojic, V., and Koller, D. (2011). Genovo: de novo assembly for metagenomes. J. Comput. Biol. J. Comput. Mol. Cell Biol. *18*, 429–443.

Lasken, R.S. (2009). Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem. Soc. Trans. *37*, 450–453.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: short oligonucleotide alignment program. Bioinforma. Oxf. Engl. *24*, 713–714.

Li, S.-C., Chan, W.-C., Lai, C.-H., Tsai, K.-W., Hsu, C.-N., Jou, Y.-S., Chen, H.-C., Chen, C.-H., and Lin, W.-C. (2011). UMARS: Un-MAppable Reads Solution. BMC Bioinformatics *12 Suppl 1*, S9.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

MacDonald, N.J., Parks, D.H., and Beiko, R.G. (2012). Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. *40*,

e111.

MacLean, D., Jones, J.D.G., and Studholme, D.J. (2009). Application of "next-generation" sequencing technologies to microbial genetics. Nat. Rev. Microbiol. *7*, 287–296.

Maitra, R.D., Kim, J., and Dunbar, W.B. (2012). Recent advances in nanopore sequencing. Electrophoresis *33*, 3418–3428.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature *470*, 198–203.

Markowitz, V.M., Chen, I.-M.A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., et al. (2012). IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. *40*, D123–129.

McDaniel, L.D., Rosario, K., Breitbart, M., and Paul, J.H. (2014). Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. Environ. Microbiol. *16*, 570–585.

McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods *4*, 63–72.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315–327.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. U. S. A. *110*, 12450–12455.

Mizuno, C.M., Rodriguez-Valera, F., Garcia-Heredia, I., Martin-Cuadrado, A.-B., and Ghai, R. (2013). Reconstruction of novel cyanobacterial siphovirus genomes from

Mediterranean metagenomic fosmids. Appl. Environ. Microbiol. *79*, 688–695.

Moore, R.A., Warren, R.L., Freeman, J.D., Gustavsen, J.A., Chénard, C., Friedman, J.M., Suttle, C.A., Zhao, Y., and Holt, R.A. (2011). The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PloS One *6*, e19838.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. *40*, e155.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. Bioinforma. Oxf. Engl. *25*, 1335–1337.

Niedringhaus, T.P., Milanova, D., Kerby, M.B., Snyder, M.P., and Barron, A.E. (2011). Landscape of next-generation sequencing technologies. Anal. Chem. *83*, 4327–4341.

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *15*, 387–396.

Osterman, A., and Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. Curr. Opin. Chem. Biol. *7*, 238–251.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. U. S. A. *96*, 2896–2901.

Pace, N.R. (1997). A molecular view of microbial diversity and the biosphere. Science *276*, 734–740.

Pareek, C.S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. J. Appl. Genet. *52*, 413–435.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2011). Meta-IDBA: a de

Novo assembler for metagenomic data. Bioinforma. Oxf. Engl. *27*, i94–101.

Prakash, T., and Taylor, T.D. (2012). Functional assignment of metagenomic data: challenges and applications. Brief. Bioinform. *13*, 711–727.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Rappé, M.S., Connon, S.A., Vergin, K.L., and Giovannoni, S.J. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature *418*, 630–633.

Rosen, G.L., Reichenberger, E.R., and Rosenfeld, A.M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinforma. Oxf. Engl. *27*, 127–129.

Schneider, G.F., and Dekker, C. (2012). DNA sequencing with nanopores. Nat. Biotechnol. *30*, 326–328.

Schouls, L.M., Schot, C.S., and Jacobs, J.A. (2003). Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J. Bacteriol. *185*, 7241–7246.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods *9*, 811–814.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics. PLoS Biol. *5*, e75.

Sharon, I., Battchikova, N., Aro, E.-M., Giglione, C., Meinnel, T., Glaser, F., Pinter, R.Y., Breitbart, M., Rohwer, F., and Béjà, O. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. ISME J. *5*, 1178–1190.

Shokralla, S., Spall, J.L., Gibson, J.F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. *21*, 1794–1805.

Smeds, L., and Künstner, A. (2011). ConDeTri--a content dependent read trimmer for

Illumina data. PloS One *6*, e26314.

Teeling, H., and Glöckner, F.O. (2012). Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. Brief. Bioinform. *13*, 728–742.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. *2*, 3.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.

Vaarala, O., Atkinson, M.A., and Neu, J. (2008). The "perfect storm" for type 1 diabetes: the complex interplay between intestinal microbiota, gut permeability, and mucosal immunity. Diabetes *57*, 2555–2562.

Venkatesan, B.M., and Bashir, R. (2011). Nanopore sensors for nucleic acid analysis. Nat. Nanotechnol. *6*, 615–624.

Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B.M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., et al. (2011). Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. ISME J. *5*, 918–928.

White, J.R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. *5*, e1000352.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U. S. A. *95*, 6578–6583.

Wilmes, P., and Bond, P.L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol. *14*, 92–97.

Wu, M., and Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. Genome Biol. *9*, R151.

Wu, D., Jospin, G., and Eisen, J.A. (2013). Systematic identification of gene families

for use as “markers” for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PloS One *8*, e77033.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, T., Zhang, X.-X., and Ye, L. (2011). Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. PloS One *6*, e26041.

Zhao, Y., Temperton, B., Thrash, J.C., Schwalbach, M.S., Vergin, K.L., Landry, Z.C., Ellisman, M., Deerinck, T., Sullivan, M.B., and Giovannoni, S.J. (2013). Abundant SAR11 viruses in the ocean. Nature *494*, 357–360.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010). The next-generation sequencing technology and application. Protein Cell *1*, 520–536.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. *38*, e132.

(2012). Bioinformatics for High Throughput Sequencing (New York, NY: Springer Science+Business Media, LLC).

http://www.genome.gov/sequencingcosts/

http://www.youtube.com/watch?v=nFfgWGFe0aA

http://www.youtube.com/watch?v=NHCJ8PtYCFc

http://www.youtube.com/ watch?v=GX6RSKh4J7E

http://www.illumina.com

http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

http://openwetware.org/wiki/Image:Libprep3.jpg

http://www.youtube.com/watch?v=77r5p8IBwJk&feature=related

http://www.youtube.com/watch?v=l99aKKHcxC4

http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/Semiconductor-Sequencing-Technology/Ion-Torrent-Technology-How-Does-It-Work.html

http://www.vincentabry.com/en/ion-proton-sequencer-decodes-human-genome-in-1-day-for-1000-dollars-1543

http://www.youtube.com/watch?v=yVf2295JqUg

https://www.youtube.com/watch?v=MxkYa9XCvBQ

http://www2.technologyreview.com/article/427677/nanopore-sequencing/

http://www.nature.com/scientificamerican/journal/v294/n1/box/scientificamerican0106-46_BX4.html

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement

http://en.wikipedia.org/wiki/ASCII

http://en.wikipedia.org/wiki/Phred_quality_score

http://openi.nlm.nih.gov/detailedresult.php?img=3096631_pone.0019534.g001&req=4

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

http://sourceforge.net/projects/amos/

http://sourceforge.net/projects/mira-assembler/

http://www.ebi.ac.uk/~zerbino/velvet/

http://metavelvet.dna.bio.keio.ac.jp/

Daniel Zerbino

Ewan Birney

European Bioinformatics Institute (EMBL-EBI)

http://trinityrnaseq.sourceforge.net/

Broad Institute

Hebrew University of Jerusalem

https://peerj.com/articles/243/

https://sites.google.com/site/cmzmasek/home/software/archaeopteryx

http://ab.inf.uni-tuebingen.de/software/megan

NCBI taxonomy

SEED

KEGG

http://orphelia.gobics.de/

http://prodigal.ornl.gov/

http://code.google.com/p/prodigal/downloads/detail?name=prodigal.v2_60.linux

https://img.jgi.doe.gov/cgi-bin/m/main.cgi

http://www.primer-e.com/

http://www.ncbi.nlm.nih.gov/

http://www.ncbi.nlm.nih.gov/gquery/

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=about

http://en.wikipedia.org/wiki/File:Example_1seq.pdf

http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR935359

ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP007/SRP007256

http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR287818

CentOS Linux 64 bit architecture

http://eutils.ncbi.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std

http://code.google.com/p/condetri/downloads/detail?name=condetri_v2.2.pl

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/INSTALL.txt

http://trinityrnaseq.sourceforge.net/#running_trinity

http://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf

http://metavelvet.dna.bio.keio.ac.jp/src/

http://code.google.com/p/prodigal/downloads/detail?name=prodigal.v2_60.linux

http://code.google.com/p/prodigal/source/browse/README

http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/

ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

http://phylosift.wordpress.com/

http://phylosift.wordpress.com/tutorials/running-phylosift/illumina-tutorial/

http://ab.inf.uni-tuebingen.de/data/software/megan4/download/welcome.html

http://ab.inf.uni-tuebingen.de/data/software/megan4/download/manual.pdf

http://commonfund.nih.gov/hmp/index