ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
Τομέας IV:
Σύνθεσης και Ανάπτυξης Βιομηχανικών Διαδικασιών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη βιοπληροφορικών εργαλείων για την ανάλυση μεταγονιδιωματικών δεδομένων»

ΤΖΑΝΟΥ ΕΛΕΝΗ

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ:

ΜΑΜΜΑ ΔΙΟΜΗ

ΑΘΗΝΑ 2024

Τζάνου Ελένη
Ανάπτυξη βιοπληροφορικών εργαλείων για την ανάλυση μεταγονιδιωματικών δεδομένων
Διπλωματική Εργασία, Σχολή Χημικών Μηχανικών
Τομέας IV: Σύνθεσης και Ανάπτυξης Βιομηχανικών Διαδικασιών
Εθνικό Μετσόβιο Πολυτεχνείο
Επιβλέπουσα Καθηγήτρια: Μαμμά Διομή

Στοιχεία επικοινωνίας:
e-mail: elenatz11@outlook.com

Περίληψη

Η ραγδαία πρόοδος της τεχνολογίας αλληλούχισης νέας γενιάς έχει επιτρέψει την πραγματοποίηση πληθώρας ερευνών που σχετίζονται με την πιθανή συσχέτιση του ανθρώπινου μικροβιώματος με νευροψυχιατρικές διαταραχές, όπως αυτή της σχιζοφρένειας. Η μεταγονιδιωματική αλληλούχιση αμπλικονίων που προέρχονται από φυλογενετικούς δείκτες, ιδιαίτερα του 16S rRNA γονιδίου, επιτρέπει την ταξινομική ανάλυση και τον προσδιορισμό της βακτηριακής ποικιλομορφίας βιολογικών δειγμάτων χωρίς την ανάγκη μεθόδων καλλιέργειας. Η βιοπληροφορική επεξεργασία των δεδομένων αλληλούχισης αμπλικονίων, η οποία είναι απαραίτητη λόγω του όγκου, της πολυπλοκότητας και των σφαλμάτων που παρέχουν, περιλαμβάνει κυρίως μεθόδους ομαδοποίησης (OTUs) ή αποθορυβοποίησης (ASVs). Παρά τον κοινό τους στόχο, οι συγκεκριμένες μέθοδοι διαφέρουν ως προς τη λογική και την εφαρμογή τους, απαιτώντας την επικύρωση της συμβατότητας της βιολογικής ερμηνείας των αποτελεσμάτων τους για την έγκυρη και αξιόπιστη διερεύνηση βακτηριακής σύνθεσης βιολογικών δειγμάτων.

Η παρούσα Διπλωματική Εργασία αποσκοπεί στη βιοπληροφορική ανάλυση δεδομένων αλληλούχισης του 16S rRNA γονιδίου που προέρχονται από δείγματα αίματος ατόμων με σχιζοφρένεια εφαρμόζοντας μεθόδους βασισμένες σε ASVs και OTUs προκειμένου να ερευνηθούν τυχόν διαφορές στην βιολογική ερμηνεία των αποτελεσμάτων τους. Τα διαθέσιμα δεδομένα έχουν προκύψει από την δειγματοληψία αίματος είκοσι ασθενών με σχιζοφρένεια σε δύο χρονικές στιγμές: i) κατά την εφμάνιση του πρώτου ψυχωσικού επεισοδίου και ii) μετά από ένα μήνα αντιψυχωσικής φαρμακευτικής χορήγησης. Επιπλέον, στην ανάλυση συμπεριλήφθηκαν και τρία δείγματα αρνητικού ελέγχου. Τα δείγματα επεξεργαστήκαν για την αλληλούχιση της περιοχής V3-V4 του 16S rRNA γονιδίου και την εισαγωγή τους στην πλατφόρμα Illumina, από την οποία παράχθηκαν fastq αρχεία με paired-end 2 x 250 bp αναγνώσματα. Η βιοπληροφορική επεξεργασία των δεδομένων πραγματοποιήθηκε στην πλατφόρμα QIIME2, από την οποία κατασκευάστηκαν δύο ροές επεξεργασίας βασιζόμενες σε ASV μεθόδους (DADA2 και Deblur) και μία αντίστοιχη σε OTU κλειστής αναφοράς (VSEARCH) με τουλάχιστον 97% ομοιότητα με τη βάση δεδομένων SILVA. Το στάδιο προεπεξεργασίας των επεξεργαστικών ροών περιλαμβάνει την αφαίρεση μη βιολογικών αλληλουχιών, την επιλογή αποκοπής των τελικών αλληλουχιών των paired-end αναγνωσμάτων (μόνο για DADA2), την συγχώνευση των paired-end αναγνωσμάτων, το φιλτράρισμα των paired-end (για DADA2) και συγχωνευμένων (για Deblur/VSEARCH) αναγνωσμάτων με βάση την ποιότητα, την περικοπή των συγχωνευμένων αναγνωσμάτων σε ίσο μήκος (μόνο για Deblur), την αφαίρεση χιμαιρικών αλληλουχιών καθώς και την παραγωγή ASVs/OTUs. Τα δεδομένα οδηγήθηκαν στην διαδικασία ταξινομικής ανάθεσης χρησιμοποιώντας τον προ-εκπαιδευμένο ταξινομητή Naïve Bayes με τη βάση δεδομένων SILVA. Στην συνέχεια, αφαιρέθηκαν οι επιμολύνσεις, συμπεριλαμβανομένου των μη-στοχευόμενων, των σημαντικά χαμηλής σχετικής αφθονίας (<0,002%) και των βιολογικά μη αναμενόμενων ταξινομικών κατηγοριών αντίστοιχα. Οι ροές επεξεργασίας ολοκληρώθηκαν με την ανάλυση ποικιλομορφίας, κατά την οποία κατασκευάστηκαν καμπύλες αραίωσης για τον προσδιορισμό βάθους αλληλούχισης και υπολογίστηκαν δείκτες εντροπίας Shannon για την εκτίμηση της α-ποικιλομορφίας των δειγμάτων. Επίσης, επιχειρήθηκε η βέλτιστη επιλογή τιμών βασικών παραμέτρων που σχετίζονται με την συγχώνευση, το φιλτράρισμα βάση ποιότητας και τα σημεία αποκοπής αναγνωσμάτων, εξετάζοντας τη συμπεριφορά των δεδομένων έως και την ταξινομική ανάθεση στα διάφορα παραμετρικά σενάρια.

Τα αποτελέσματα φανέρωσαν ότι οι διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής ($Overlap_{min}$=10, 20 και 30) κατά την συγχώνευση των paired-end αναγνωσμάτων επέφεραν παρόμοιο αντίκτυπο στα δεδομένα αλληλούχισης και στις ταξινομικές πληροφορίες τους. Η αφαίρεση των τελικών αλληλουχιών των paired-end αναγνωσμάτων στον DADA2 (trim@=f:220 r:225) έναντι της επιλογής μη-αποκοπής αυτών (no-trim) και η διατήρηση μεγαλύτερο μήκους αναγνωσμάτων στον Deblur (trim@=380 έναντι του trim@=250) βελτίωσαν την αποδοτικότητα των ASVs. Οι διαφορετικές προσεγγίσεις ποιοτικού φιλτραρίσματος, με τον DADA2 να αξιολογεί τα paired-end αναγνώσματα με βάση το μέγιστο ποσοστό αναμενόμενων σφαλμάτων ($e.e_{max}$=0.5, 1.5 και 2.5) και ο Deblur/VSEARCH αντίστοιχα την ελάχιστη βαθμολογία ποιότητας PHRED ανά βάση των συγχωνευμένων αναγνωσμάτων (Q_{min}= 20, 22 και 26), οδήγισαν σε παρόμοια αποτελέσματα, με την αύξηση της αξιοπιστίας των τελικών αποτελεσμάτων να επιφέρει την απώλεια όγκου αναγνωσμάτων και ταξινομικών πληροφοριών. Η επιλογή των τελικών τιμών παραμέτρων (DADA2:[$Overlap_{min}$=10, $e.e_{max}$=1.5, no-trim], Deblur:[$Overlap_{min}$=10, Q_{min}= 22, trim@=380] και VSEARCH:[$Overlap_{min}$=10, Q_{min}= 22]) βασίστηκε κυρίως στην βέλτιστη αποκόμιση αξιόπιστων διατηρητέων αναγνωσμάτων και ταξινομικών μονάδων όσο αφορά την ποιότητα και τον αριθμό τους. Επίσης, κατά την επιλογή αυτή, έγινε η προσπάθεια εφαρμογής παρόμοιων παραμετρικών σεναρίων μεταξύ των ροών επεξεργασίας για την αποτελεσματικότερη σύγκρισή τους.

Κατά την σύγκριση των επεξεργαστικών ροών, αναδείχθηκε η σημαντική απώλεια πρωτογενών δεδομένων από την αποθορυβοποίηση Deblur, εντοπίστηκε ενός σημαντικός όγκος μη-ταξινομημένων ASVs από την αποθορυβοποίηση DADA2, και προσδιορίστηκε πολύ μεγάλος αριθμός OTUs από την ομαδοποίηση VSEARCH. Η συγκριτική ταξινομική ανάλυση των μεθόδων ASVs/OTUs φανέρωσε αυξημένη ταξινομική ομοιότητα μέχρι και σε επίπεδο γένους μεταξύ των συνόλων δεδομένων, με το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών να οδηγεί σε ακόμα πιο παρόμοια αποτελέσματα, και ανέδειξε την αδυναμία των ASVs/OTUs στην ταξινόμηση σε επίπεδο είδους. Η αφαίρεση των βακτηριακών ταξινομικών μονάδων γένους *Lactobacillus*, των οποίων η παρουσία τους στα βιολογικά δείγματα χαρακτηρίστηκε ως αποτέλεσμα επιμόλυνσης, είχε έντονο αντίκτυπο στον όγκο των διατηρητέων αναγνωσμάτων, το οποίο ήταν παρόμοιο μεταξύ των τριών διαφορετικών επεξεργαστικών ροών. Οι ταξινομικές σχετικές αφθονίες των δειγμάτων δεν έδειξαν διαφορές μεταξύ των μεθόδων. Επίσης, η βακτηριακή σύνθεση σε επίπεδο φυλής του αίματος σχιζοφρενών παρουσίασε αύξηση των *Firmicutes* σε σύγκριση με υγιή άτομα βάσει βιβλιογραφίας. Οι καμπύλες αραίωσης των τριών επεξεργαστικών ροών φανέρωσαν διαφορές, με τον VSEARCH να δείχνει συνεχή αύξηση των παρατηρούμενων OTUs συναρτήσει του βάθους αλληλούχισης. Όσο αναφορά στην α-ποικιλομορφίας, και στις τρεις επεξεργαστικές ροές δεν παρατηρήθηκαν στατιστικά σημαντικές διαφορές στα δείγματα αίματος και αρνητικού ελέγχου. Ο VSEARCH οδήγησε σε υψηλότερες τιμές δείκτη Shannon, υποδεικνύοντας μια πιθανή υπερεκτίμηση της βακτηριακής ποικιλομορφίας.

Η παρούσα εργασία, σε συμφωνία με την υπάρχουσα σχετική βιβλιογραφία, υποστηρίζει την προτίμηση των μεθόδων αποθορυβοποίησης, και πιο συγκεκριμένα του DADA2, έναντι της ομαδοποίησης για την ανάλυση του 16S rRNA γονιδίου λόγω της υπολογιστικής αποδοτικότητας, της ανεξαρτησίας του από βάση δεδομένων και της ευελιξίας της επιλογής του μήκους των paired-end αναγνωσμάτων.

Λέξεις κλειδί: 16S rRNA γονίδιο, ομαδοποίηση, αποθορυβοποίηση, ASVs, OTUs, DADA2, Deblur, VSEARCH, QIIME2, μικροβίωμα του αίματος, σχιζοφρένεια

Abstract

Development of bioinformatic tools for the analysis of metagenomic data

The advances in next-generation sequencing technology have led to an great number of studies related to the investigation of the human microbiome and its potential association with neuropsychiatric disorders, such as schizophrenia. The amplicon metagenomic sequencing of phylogenetic markers, particularly the 16S rRNA gene, allows for the taxonomic analysis and identification of bacterial diversity of biological samples without the need for culture methods. The bioinformatic analysis of amplicon sequencing data that is required due to the volume, complexity and errors that they provide, mainly involves clustering (OTUs) or denoising (ASVs) methods. Despite their common goal, these methods differ in their logic and application, requiring validation of the compatibility in the biological interpretation of the results obtained by both methods for valid and reliable investigation of bacterial composition in biological samples.

This Diploma Thesis aims in the bioinformatic analysis of 16S rRNA gene sequencing data from blood samples of individuals with schizophrenia using both ASV and OTU based methods, in order to investigate any differences in the biological interpretation of the results. The available data have been obtained from blood sampling of twenty patients with schizophrenia at two timepoints: i) at the presentation of the first episode psychosis and ii) after one month of antipsychotic medication. In addition, three negative control samples were included in the analysis. The samples were processed for the V3-V4 region sequencing of the 16S rRNA gene and were imported into an Illumina platform, from which fastq files with paired-end 2 x 250 bp reads were generated. The bioinformatic processing of these data was performed on the QIIME2 platform, from which three workflows were constructed: two workflows based on ASV methods (DADA2 and Deblur) and one workflow based on OTU closed reference method (VSEARCH) with at least 97% similarity threshold to the SILVA database. The preprocessing stage of the workflows includes removing non-biological sequences, trimming the final sequences of paired-end reads (for DADA2 only), merging paired-end reads, quality filtering of paired-end (for DADA2) and merged (for Deblur/VSEARCH) reads, truncation of merged reads to equal length (for Deblur only), removing chimeric sequences and generating ASVs/OTUs. The data were taken through the process of taxonomic assignment using the pre-trained Naïve Bayes classifier with the SILVA database. Then, contaminants were removed, including non-targeted, significantly low relative abundance (<0.002%) and biologically unexpected taxa. The workflows were completed with diversity analysis, during which rarefaction curves were constructed to determine sequencing depth and Shannon entropy indices were calculated to estimate the a-diversity of the samples. An attempt was also made to optimally select values of key parameters related to merging, quality filtering and read cut-off points, by examining the behaviour of the data up to the taxonomic assignment on the different parametric scenarios.

The results revealed that different values of minimum overlap length ($Overlap_{min}$=10, 20 and 30) when merging paired-end reads had a similar impact on the sequencing data and their taxonomic information. Trimming the final sequences of paired-end reads in DADA2 (trim@=f:220 r:225) versus choosing not to trim them (no-trim) and retaining longer read lengths in Deblur (trim@=380 versus trim@=250) improved the efficiency of ASVs. The different quality filtering approaches, with DADA2 evaluating paired-end reads based on the maximum expected error rate ($e.e_{max}$=0.5, 1.5 and 2.5) and Deblur/VSEARCH on the

minimum quality PHRED score per base of merged reads (Qmin= 20, 22 and 26), led to similar results, with the increase in reliability of the final results led in a loss of read volume and taxonomic information. The choice of final parameter values (DADA2:[$Overlap_{min}$=10, $e.e_{max}$=1.5, no-trim], Deblur:[$Overlap_{min}$=10, Q_{min}= 22, trim@=380] and VSEARCH:[$Overlap_{min}$=10, Q_{min}= 22]) was mainly based on the optimal obtaining of reliable retained reads and taxa, in terms of their quality and number. Also, during this choice, an attempt was made to apply similar parametric scenarios between the workflows, for their more effective comparison.

When comparing the workflows, a significant loss of raw-data was highlighted by Deblur denoising, a significant volume of unclassified ASVs was identified by DADA2 denoising, and a very large number of OTUs was produced by VSEARCH clustering. Comparative taxonomic analysis of the ASVs/OTUs based methods revealed increased taxonomic similarity up to genus level between the compared datasets, with low relative abundance taxa filtering leading to even more similar results, and highlighted the weakness of ASVs/OTUs taxonomic classification at species-level. The removal of bacterial taxa of the genus *Lactobacillus*, whose presence in biological samples was characterized as a result of contamination, had a strong impact on the volume of retained reads, which was similar between the three different workflows. The relative taxonomic compositions of the samples showed no differences between methods. Also, in this study the blood bacterial composition at phylum level of schizophrenic patients showed an increase in *Firmicutes* compared to blood bacterial composition of healthy subjects reported in the literature. Rarefaction curves of the three workflows revealed differences, with VSEARCH showing a consistent increase in observed OTUs in correlation to the sequencing depth. Conserning a-diversity analysis, in all three workflows no statistically significant differences were observed in the blood and negative control samples. Additionally, VSEARCH resulted in higher Shannon index values, indicating a possible overestimation of bacterial diversity.

This work, in agreement with the existing relevant literature, supports the preference of denoising methods, and more specifically of DADA2, over clustering methods for 16S rRNA gene analysis due to its computational efficiency, database independency and flexibility in choosing the length of paired-end reads.

Key words: 16S rRNA gene, clustering, denoising, ASVs, OTUs, DADA2, Deblur, VSEARCH, QIIME2, blood microbiome, schizophrenia

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος – Ευχαριστίες

Η παρούσα Διπλωματική Εργασία εκπονήθηκε στα πλαίσια της ολοκλήρωσης του κύκλου σπουδών μου στη Σχολή Χημικών Μηχανικών ΕΜΠ, υπό την επίβλεψη της καθηγήτριας ΕΜΠ Διομή Μαμμά, και αποτελεί μία ανάλυση που υποστηρίχθηκε από το τμήμα της Βιοπληροφορικής του τομέα της Σύνθεσης και Ανάπτυξης Βιομηχανικών Διαδικασιών.

Στο εκτεταμένο βασίλειο της επιστημονικής έρευνας, το πεδίο της ανάλυσης μικροβιώματος καταλαμβάνει έναν τεράστιο χώρο, οποίος έχει αναδείξει την πολυπλοκότητα των μικροβιακών κοινοτήτων και τον τρόπο με τον οποίο διαμορφώνουν και συμβάλλουν αυτές τον περίπλοκο κόσμο μας. Κατά την πλοήγηση σε αυτήν την αχαρτογράφητη περιοχή, φανερώθηκε η συμβολή του κλάδου της Βιοπληροφορικής, καθώς μέσω αυτού προσφέρεται η δυνατότητα της αποκάλυψης του αόρατου και του ξετυλίγματος των μυστηρίων που κωδικοποιούνται στο μικροβιακό DNA, και όχι μόνο. Αυτός ο κλάδος αποτελεί τον γάμο της Βιολογίας και της Πληροφορικής, ένας συνδυασμός που επεκτείνει τα όρια της παραδοσιακής Βιολογίας, εγκαινιάζοντας μια εποχή όπου η υπολογιστική ικανότητα των αλγορίθμων και της μηχανικής μάθησης φωτίζει τις σκιές της βιολογικής πολυπλοκότητας. Η γοητεία της Βιοπληροφορικής δεν έγκειται μόνο στο γνωσιακό περιεχόμενο και ενδιαφέρον, αλλά και στην ανιδιοτελή υποστήριξη που προσφέρεται από την επιστημονική κοινότητα που την απαρτίζουν. Μια κοινότητα που δεσμεύεται όχι μόνο από κοινή γνώση αλλά και από ένα ήθος συνεργασίας, όπου οι ερωτήσεις βρίσκουν απαντήσεις, οι συμβουλές προσφέρονται γενναιόδωρα και το πνεύμα της ανοιχτής επιστήμης κυριαρχεί. Η συμβίωση μεταξύ βιοπληροφορικής και ανάλυσης μικροβιώματος ξετυλίγεται ως αφήγηση συλλογικής ανάπτυξης, όπου τα δωρεάν δημοσιευμένα εργαλεία και μεθοδολογίες δίνουν τη δυνατότητα στους ερευνητές να περιηγηθούν στην πολυπλοκότητα των μικροβιωμάτων με επιδεξιότητα και ακρίβεια. Ωστόσο, στον κατάλογο της ελληνικής βιβλιογραφίας, οι αναφορές σε αυτόν τον δυναμικό κλάδο ήταν συγκριτικά ελλιπείς. Έτσι, εκτός από την διεκπεραίωση του βασικού στόχου της παρούσας εργασίας, δημιουργήθηκε μια ευκαιρία να εξερευνήσω, να διαλευκάνω και να γεφυρώσω αυτό το χάσμα μέσω της εξοικείωσης ορολογιών και της κατανόησης μεθόδων σε ότι αφορά το πεδίο του μικροβιώματος, της τεχνολογίας αλληλούχισης και εξειδικευμένα της ανάλυσης δεδομένων αλληλούχισης του 16S rRNA γονιδίου. Προτείνω ανεπιφύλακτα την ενσωμάτωση όλο και περισσότερων βιοπληροφορικών γνώσεων και πρακτικών σε όποιον σπουδάζει και εντριβεί στην κλάδο της Βιοτεχνολογίας, διότι οι δυνατότητες που προσφέρουν στην έρευνα υπόσχονται όχι μόνο την επιστημονική, αλλά και την προσωπική πρόοδο.

Καθώς σκέφτομαι την ακαδημαϊκή μου πορεία στην Σχολή Χημικών Μηχανικών, η οποία δεν θα μπορούσε να ολοκληρωθεί με πιο ολιστικό και εκπαιδευτικό τρόπο, νιώθω υποχρεωμένη να εκφράσω τη βαθύτατη ευγνωμοσύνη μου σε όσους έπαιξαν κομβικούς ρόλους σε αυτό το ταξίδι. Η διεκπεραίωση της παρούσας διπλωματικής εργασίας δεν είναι μόνο ένα προσωπικό επίτευγμα αλλά μια συλλογική προσπάθεια που κατέστη δυνατή με την υποστήριξη και τη συμβολή πολλών ανθρώπων. Αναγνωρίζοντας τους ανεκτίμητους ρόλους τους, εκφράζω τις ειλικρινείς ευχαριστίες μου στους ακόλουθους ανθρώπους που συνέβαλαν καθοριστικά στην πραγματοποίηση αυτής της προσπάθειας.

Πρωτίστως, θα ήθελα να ευχαριστήσω την Δρ. Ελένη Λουτράρη (Ίδρυμα Εντατικής Θεραπείας και Επείγουσας Ιατρικής Θώρακος του Νοσοκομείου Ευαγγελισμού) και την Μεταδιδάκτορα Μαριάνθη Λογοθέτη που μου εμπιστεύτηκαν ευαίσθητα δεδομένα και υλικό,

όπου χωρίς αυτά, δεν θα μπορούσε να υποστηριχτεί η παρούσα Διπλωματική Εργασία. Φυσικά, είμαι βαθιά ευγνώμων και για την επιστημονική οξυδέρκεια και τεχνογνωσία που μου παρείχαν. Επιπλέον, η καθοδήγηση και η αμέριστη υποστήριξή της Μαριάνθης καθόλη τη διάρκεια διεκπεραίωσης της παρούσας μελέτης αποτέλεσε ύψιστης σημασίας για την ολοκλήρωση της εργασίας, και για αυτόν τον λόγο την ευχαριστώ ιδιαιτέρως από τα βάθη της καρδιάς μου.

Φυσικά, θα ήθελα να εκφράσω την εγκάρδια ευγνωμοσύνη μου και στην επιβλέπουσα καθηγήτριά μου, την κ. Μαμμά, που όχι μόνο μου εμπιστεύτηκε την ευκαιρία να διεξαγάγω το συγκεκριμένο θέμα Διπλωματικής Εργασίας, αλλά και που μου στάθηκε ως κάτι περισσότερο από μέντορας κατά την ολοκλήρωση των σπουδών μου. Η ακλόνητη δέσμευσή της να μεταδώσει γνώσεις υπήρξε ανεκτίμητη, αλλά εξίσου σημαντική ήταν η ενσυναίσθηση και η προσωπική υποστήριξή της. Η καθοδήγησή της επεκτάθηκε πέρα από το ακαδημαϊκό πεδίο, προσφέροντας γνώσεις για την πλοήγηση στις προκλήσεις και την προώθηση της προσωπικής ανάπτυξης. Είμαι ειλικρινά ευγνώμων για την εμπιστοσύνη, την ενθάρρυνση και τη φροντίδα που μου παρείχε.

Επιπλέον, θα ήθελα να ευχαριστήσω βαθύτατα τον Υποψήφιο Διδάκτορα Θωμά Γκέκα, για την γενναιοδωρία του να μοιραστεί τις γνώσεις τους, ειδικά στο κλάδο της Πληροφορικής, οι οποίες εμπλούτισαν σημαντικά την κατανόησή πολύπλοκων, και ίσως και απλών, υπολογιστικών πτυχών, συμβάλλοντας εξαιρετικά στην ποιότητα της δουλειάς μου. Ακόμη, ευχαριστώ θερμά τον Υποψήφιο Διδάκτορα Παναγιώτη Αγιουτάντη για τις χρήσιμες συμβουλές του, οι οποίες έπαιξαν καθοριστικό ρόλο στην επιτυχία της παρούσας Διπλωματικής Εργασίας.

Ολοκληρώνοντας, δεν θα μπορούσα να παραλείψω τον προσωπικό μου κύκλο στις ευχαριστίες μου, που όμως λόγω πρακτικότητας δεν θα αναφέρω ονόματα, διότι αν ξεκινούσα την απαρίθμηση του αντίκτυπου του κάθε ατόμου, η λίστα θα ήταν εκτενής και βαθιά προσωπική. Άλλωστε, ποιος θα μπορούσε να με σταματήσει στο να γράψω άλλες 200 σελίδες ακόμη; Κανένας είναι η απάντηση. Για αυτό λοιπόν, εν συντομία, θα ήθελα να εκφράσω την ειλικρινής ευγνωμοσύνη μου στις στενές παιδικές και ακαδημαϊκές φιλίες μου, και φυσικά στην οικογένειά μου για την αμέριστη υποστήριξή τους σε όλη αυτή την ακαδημαϊκή διαδρομή. Η ενθάρρυνση, η κατανόηση και η υπομονή τους ήταν και θα είναι οι πυλώνες της ζωής μου. Νιώθω πραγματικά τυχερή που έχω ένα τέτοιο υποστηρικτικό δίκτυο και που μου συμπαραστέκεται δίπλα μου καθημερινά, και στις επιτυχίες αλλά και στις προκλήσεις. Η πίστη τους σε μένα υπήρξε κινητήριος δύναμη και είμαι βαθιά ευγνώμων για την ανιδιοτελή αγάπη τους.

Έλενα Τζάνου

Ιούλιος, 2024

1 Θεωρητικό Υπόβαθρο

1.1 Μικροβίωμα

1.1.1 Εισαγωγή - Από τον μικροοργανισμό στο μικροβίωμα

Μέχρι και τον 17ο αιώνα, η ανθρωπότητα χρησιμοποιούσε βιοτεχνολογικές μεθόδους στον τομέα των τροφίμων, αξιοποιώντας μικροοργανισμούς, χωρίς όμως να γνωρίζει την ύπαρξή τους,, καθώς ο μικρόκοσμος δεν είναι ορατός με το γυμνό μάτι. Η ανάπτυξη των μικροσκοπίων ήταν αυτή που επέτρεψε την ανακάλυψη ενός νέου, άγνωστου κόσμου και οδήγησε στην ταυτοποίηση των μικροοργανισμών. Συγκεκριμένα, ο Ολλανδός επιστήμονας Άντονι βαν Λέβενχουκ (Antony van Leeuwenhoek), παγκοσμίως αναγνωρισμένος ως ο πατέρας της μικροβιολογίας, είναι ο πρώτος άνθρωπος που ανακάλυψε τον ασύλληπτο κόσμο των «μικρόζωων» (animalcules), χρησιμοποιώντας ένα αυτοδημιούργητο μικροσκόπιο (N. Lane, 2015). Στην δημοσίευσή του το 1677 μ.Χ., ανέφερε χαρακτηριστικά ότι παρατήρησε «μικρά ζωάκια να κινούνται πολύ όμορφα» σε δείγματα νερού που προέρχονται από τις βροχές, το χιόνι, τα πηγάδια και τη θάλασσα. Διεξάγοντας δημιουργικά πειράματα, εξερευνώντας και χειραγωγώντας το μικροσκοπικό του σύμπαν με οδηγό μόνο την επιστημονική του περιέργεια, ανακάλυψε τόσο τα πρώτιστα όσο και τα βακτήρια.

Όμως, μέχρι και το 1882 μ.Χ., ο επιστημονικός κόσμος δεν ήταν έτοιμος να αποδεχτεί την πρωτοπόρα έννοια των μικροοργανισμών. Εκείνη την εποχή, ο Γερμανός γιατρός Ρόμπερτ Κοχ (Robert Koch), ανακαλύπτοντας τον βάκιλο της φυματίωσης του ανθρώπου και το μικρόβιο της χολέρας, κατάφερε να συσχετίσει την προέλευση πολλών ασθενειών στον άνθρωπο και στα ζώα με μικροβιακές επιμολύνσεις (Lakhtakia, 2014). Η συσχέτιση αυτή οδήγησε στην επισημοποίηση του όρου "μικροοργανισμός".

Παρότι η ανάπτυξη της έννοιας της παθογένειας ήταν ένα σημαντικό ορόσημο στη μικροβιολογία, έρευνες τον περασμένο αιώνα έδειξαν ότι οι μικροοργανισμοί δεν σχετίζονται μόνο με ασθένειες ή παθογένεια. Μάλιστα, στις αρχές του 20ου αιώνα, διαπιστώθηκε ότι η συντριπτική πλειοψηφία των μικροβίων είναι απαραίτητη για τη λειτουργία του οικοσυστήματος. Για πρώτη φορά, αναφέρεται ότι οι μικροοργανισμοί υπάρχουν παντού στο φυσικό περιβάλλον, και ότι συσχετίζονται με ξενιστές έχοντας ευεργετικές επιδράσεις σε αυτούς (Berg et al., 2020). Έτσι, στην οικολογία, καθώς και στην μικροβιολογία, εισήχθηκε για πρώτη φορά η έννοια της μικροχλωρίδας.

Πλησιάζοντας στην σύγχρονη εποχή, η ιδέα ότι οι μικροοργανισμοί υπάρχουν ως μεμονωμένα κύτταρα αρχίζει να αλλάζει, καθώς γίνεται ολοένα και πιο προφανές ότι τα μικρόβια εμφανίζονται μέσα σε πολύπλοκα συγκροτήματα, στα οποία οι αλληλεπιδράσεις και η επικοινωνία αυτών είναι κρίσιμες για τη δυναμική των πληθυσμών τους και για τις λειτουργικές τους δραστηριότητες. Η ανακάλυψη του DNA, ο ορισμός του κεντρικού δόγματος της βιολογίας, η ανάπτυξη βιοτεχνολογικών εφαρμογών και η υπολογιστική ισχύς επέτρεψαν στις αρχές του 21ου αιώνα την ανάπτυξη της τεχνολογίας αλληλούχισης, η οποία εξελίσσεται ακόμα μέχρι και σήμερα, με αποτέλεσμα να προκύψουν δεδομένα τα οποία έχουν επισημάνει τόσο την πανταχού παρουσία μικροβιακών κοινοτήτων σε συνδυασμό με ανώτερους οργανισμούς όσο και τους κρίσιμους ρόλους των μικροβίων στην υγεία του ανθρώπου, των ζώων και των φυτών.

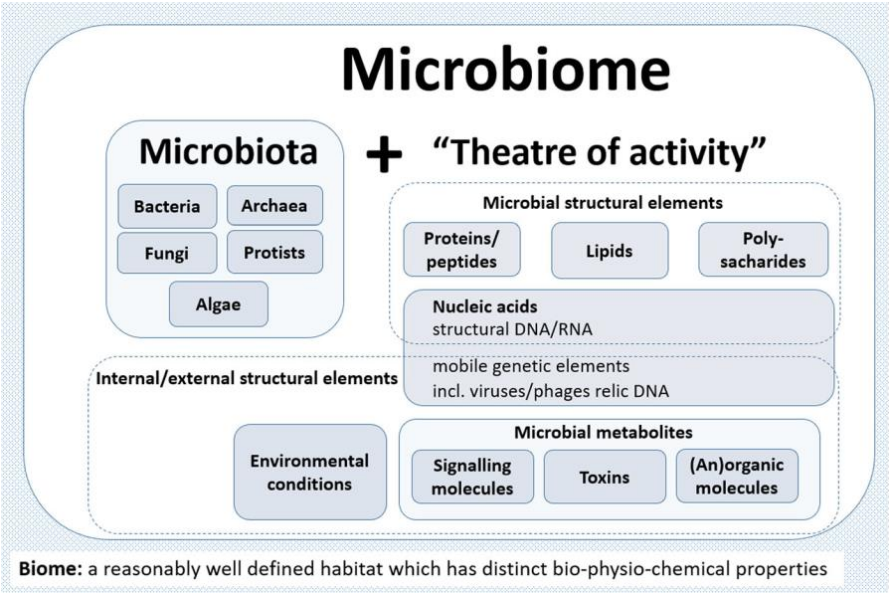
Αυτές οι νέες δυνατότητες έχουν φέρει την επανάσταση στη μικροβιακή οικολογία, λόγω του ότι η ανάλυση υψηλής απόδοσης γονιδιωμάτων και μεταγονιδιωμάτων που έχει

προσφέρει η τεχνολογία αλληλούχισης επόμενης γενιάς (Next Generation Sequencing, η αλλιώς NGS) παρέχει αποτελεσματικές μεθόδους για την διερεύνηση του λειτουργικού δυναμικού μεμονωμένων μικροοργανισμών καθώς και ολόκληρων κοινοτήτων στο φυσικό τους περιβάλλον. Όμως, οι βασικοί ορισμοί των μικροοργανισμών και της μικροβιακής κοινότητας δεν επαρκούν στην περιγραφή της πολυπλοκότητας αυτών των νέων δεδομένων που αφορούν άμεσα αυτά τα συστήματα, εισάγοντας τις τελευταίες δεκαετίες έναν νέο όρο στον κλάδο της μικροβιολογίας, αυτόν του μικροβιώματος.

1.1.2 Ορισμός μικροβιώματος

Η αναγκαιότητα της δημιουργίας ενός νέου όρου και η ταχεία ανάπτυξη του ενδιαφέροντος αυτού, που καλύπτει πολλούς διαφορετικούς τομείς τα τελευταία 30 χρόνια, έχει φέρει ως αποτέλεσμα την έλλειψη της σαφήνειας ενός κοινού αποδεκτού ορισμού του όρου μικροβιώματος. Μόλις το 2019 πραγματοποιήθηκε μια διεθνής συνάντηση στα πλαίσια του έργου Microbiome Support με στόχο τη θέσπιση διεθνών ερευνητικών προτύπων σε ότι αφορά με αυτό το θέμα (Berg et al., 2020). Στο συγκεκριμένο συνέδριο προτάθηκε ένας ορισμός του μικροβιώματος, όπου σύμφωνα με αυτήν την πρόταση, το μικροβίωμα (microbiome) ορίζεται ως μια χαρακτηριστική μικροβιακή κοινότητα που καταλαμβάνει ένα καλά καθορισμένο οικολογικό χώρο και έχει ξεχωριστές φυσικοχημικές ιδιότητες. Το μικροβίωμα δεν αναφέρεται μόνο στους μικροοργανισμούς που εμπλέκονται αλλά περιλαμβάνει και την «σκηνή δραστηριότητάς» τους (theatre of activity) **(Σχήμα 1.1**), που οδηγεί στο σχηματισμό συγκεκριμένων οικολογικών θώκων (ecological niche). Πιο αναλυτικά, η μικροχλωρίδα (microflora/microbiota) ορίζεται ως η ομάδα ζωντανών μικροοργανισμών που ανήκουν σε διαφορετικά βασίλεια, όπως τα βακτήρια, τα αρχαία, οι μύκητες, τα φύκια και τα πρώτιστα. Οι φάγοι, οι ιοί, τα πλασμίδια, και το ελεύθερο DNA συνήθως δεν θεωρούνται ζωντανοί μικροοργανισμοί και συνεπώς δεν ανήκουν στη μικροχλωρίδα. Παρόλα αυτά, τα παραπάνω στοιχεία έχουν την δυνατότητα να επηρεάσουν μια μικροβιακή κοινότητα, με αποτέλεσμα να αποτελούν μέρος ενός μικροβιώματος. Η κατηγορία στην οποία ανήκουν τα στοιχεία αυτά ονομάζεται «σκηνή δραστηριότητας», το οποίο περιλαμβάνει και τα μικροβιακά δομικά στοιχεία (πρωτεΐνες, λιπίδια, πολυσακχαρίτες, DNA), τους μεταβολίτες (τοξίνες, άλλα οργανικά και ανόργανα μόρια), τα κινητά γενετικά στοιχεία και το νεκρό DNA ενσωματωμένο στις περιβαλλοντικές συνθήκες του βιότοπου.

Το μικροβίωμα αποτελεί ένα δυναμικό και διαδραστικό μικρο-οικοσύστημα, που είναι επιρρεπές σε αλλαγές στον χρόνο και στον χώρο, και είναι πάντα ενσωματωμένο σε ένα μακρο-οικοσύστημα (πχ. ευκαριωτικοί ξενιστές). Οι μικροβιακές αλληλεπιδράσεις σε ένα τέτοιο σύστημα αποτελούν βάση για την λειτουργία και την εξελικτική δυναμική μικροβιακών κοινοτήτων από τις οποίες αποτελείται. Επίσης, οι αλληλεπιδράσεις ξενιστή-μικροβιώματος διαμορφώνουν την αμοιβαία φυσική κατάσταση, τον φαινότυπο και τον μεταβολισμό τους, ενισχύοντας τη θεωρία της συνεξέλιξης των μικροβίων και του ξενιστή. Σε μια κατάσταση ασθένειας, το ολοβίωμα χαρακτηρίζεται ως δυσβίωση (παθοβίωμα), σε αντίθεση με την ενβίωση που αναφέρεται σε μια κατάσταση ισορροπημένης αλληλεπίδρασης ξενιστή-μικροβιώματος. Ο διαχωρισμός των μικροοργανισμών σε ευεργετικούς, παθογόνους και ουδέτερους, σύμφωνα με της αλληλεπιδράσεις των μικροβίων με τον ξενιστή, βασίζεται σε μια καθαρά ανθρωποκεντρική οπτική (Berg et al., 2020).



Σχήμα 1.1 Απεικόνιση της σύνθεσης του όρου «μικροβίωμα», το οποίο περιέχει τόσο τη μικροχλωρίδα (κοινότητα μικροοργανισμών) όσο και τη «σκηνή δραστηριότητάς» τους (δομικά στοιχεία, μεταβολίτες και τις περιβαλλοντικές συνθήκες). (Berg et al., 2020)

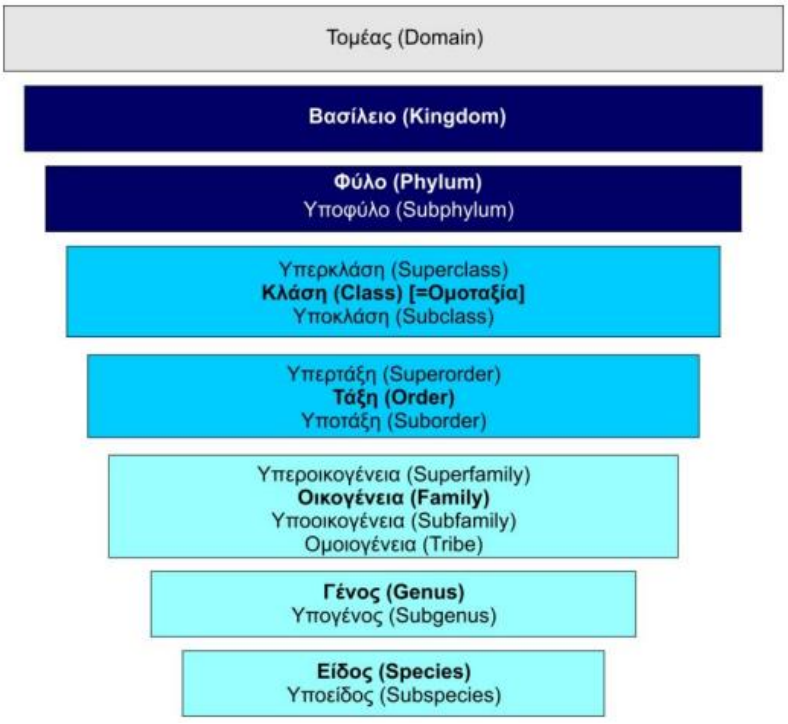
1.1.3 Μικροβιολογία συστημάτων

Προκειμένου να αναλυθεί σφαιρικά ένα μικροβίωμα, η διερεύνηση και η περιγραφή των ζωντανών μικροοργανισμών που το απαρτίζουν αποτελεί σημαντικό στάδιο για την επίτευξη αυτού. Η συστηματική μικροβιολογία, ένα πεδίο που περιλαμβάνει την ταξινόμηση και την φυλογένεση, παίζει καθοριστικό ρόλο στην διαφάνεια την ταυτοτήτων των μικροοργανισμών. Η συστημική βιολογία περιλαμβάνει τη μελέτη της ποικιλομορφίας των οργανισμών και των σχέσεων μεταξύ τους. Συνδέει τη φυλογένεση με την ταξινόμηση, την επιστήμη στην οποία οι οργανισμοί χαρακτηρίζονται, ονομάζονται και ταξινομούνται σύμφωνα με καθορισμένα κριτήρια. Η μικροβιολογία συστημάτων είναι ο κλάδος της μικροβιολογίας που ασχολείται με την περιγραφή μικροοργανισμών χρησιμοποιώντας μια πολύπλευρη προσέγγιση που αξιολογεί φαινοτυπικές, γονοτυπικές και φυλογενετικές πληροφορίες, με σκοπό την περιγραφή των χαρακτηριστικών τους και τη διερεύνηση της εξελικτικής τους ιστορίας. Οι φαινοτυπικές αναλύσεις εξετάζουν τα μορφολογικά, μεταβολικά, φυσιολογικά και χημικά χαρακτηριστικά του κυττάρου, ενώ οι γονοτυπικές αναλύσεις λαμβάνουν υπόψη τα χαρακτηριστικά του γονιδιώματός του. Αυτά τα δύο είδη αναλύσεων χρησιμοποιούνται για την κατηγοριοποίηση μικροοργανισμών με βάση τις ομοιότητες τους και συμπληρώνονται από φυλογενετικές αναλύσεις, οι οποίες επιδιώκουν την τοποθέτηση μικροοργανισμών σε ένα εξελικτικό πλαίσιο (Madigan et al., 2021).

Αν και η προσπάθεια περιγραφής ενός μικροοργανισμού στην σύγχρονη εποχή ξεπερνά την απλή αναγνώρισή του με το όνομα και απαιτεί την τοποθέτησή του σε ένα ευρύτερο πεδίο, η ονομασία των μικροοργανισμών είναι το θεμέλιο της συστηματικής μικροβιολογίας. Η ονοματολογία των μικροοργανισμών, αλλά και των οργανισμών γενικά, ακολουθεί το διωνυμικό σύστημα που επινόησε ο Σουηδός γιατρός και βοτανολόγος Carl Linnaeus, στο οποίο δίνονται στα είδη ονόματα γένους και επίθετα ειδών. Τα ονόματα είναι στα λατινικά ή λατινοποιημένα ελληνικά, αναγραφόμενα με πεζούς πλάγιους χαρακτήρες, όπου το όνομα γένους είναι ουσιαστικό και το επίθετο είδους είναι ένα επίθετο που περιγράφει κάποια βασική ιδιότητα του μικροοργανισμού ή του βιότοπου από τον οποίο

προήλθε. Τα ονόματα πολλές φορές προέρχονται από το επίθετο ενός επιστήμονα που έχει κάνει σημαντικές συνεισφορές στη μελέτη μιας συγκεκριμένης ομάδας μικροοργανισμών. Ένα κλασσικό παράδειγμα ονοματολογίας μικροοργανισμού αποτελεί του βακτηρίου *Escherichia coli*. Σε αυτήν την περίπτωση, η ονομασία του γένους είναι τιμητική για τον παιδίατρο και βακτηριολόγο Theodor Eshcerich, ο οποίος ανακάλυψε το βακτήριο (Shulman et al., 2007). Το δεύτερο συνθετικό περιγράφει την περιοχή που συνήθως αποικίζεται από το συγκεκριμένο βακτήριο, δηλαδή το έντερο.

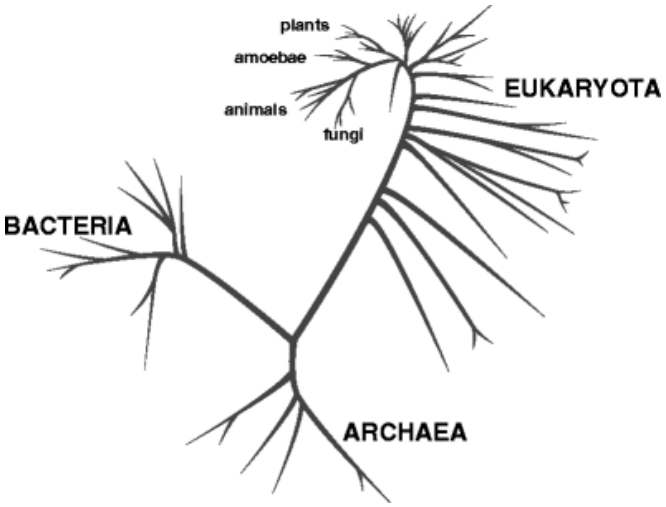
Η ταξινόμηση αποτελεί την κατηγοριοποίηση των μικροοργανισμών σε ομάδες παρόμοιων μικροοργανισμών. Τα μεμονωμένα μικροβιακά στελέχη ταξινομούνται σε είδη, τα είδη ταξινομούνται σε γένη, τα γένη σε οικογένειες, οι οικογένειες σε τάξεις, οι τάξεις σε ομοταξίες, οι ομοταξίες σε φύλα και τα φύλα σε τομείς (**Σχήμα 1.2**). Ταξινομώντας τους μικροοργανισμούς σε ομάδες, ορίζεται συνολικά ο φυσικός μικροβιακός κόσμος και καθίσταται δυνατή η αποτελεσματική επικοινωνία για τη μελέτη όλων των πτυχών συγκεκριμένων μικροοργανισμών, συμπεριλαμβανομένης της συμπεριφοράς, της οικολογίας, της φυσιολογίας και της παθογένειας τους (Madigan et al., 2021). Τα είδη είναι οι θεμελιώδεις μονάδες της βιολογικής ποικιλομορφίας και ο τρόπος με τον οποίο διακρίνονται και ταξινομούνται τα είδη στη μικροβιολογία επηρεάζει σε μεγάλο βαθμό την ικανότητα περιγραφής και αξιολόγησης της ποικιλότητας του μικροβιακού κόσμου ή, ακόμα πιο συγκεκριμένα, ενός μικροβιώματος. Παρόλα αυτά, η περιγραφή ενός μικροβιώματος σε πιο ψηλά ταξινομικά επίπεδα μπορεί να αποδειχθεί επίσης χρήσιμη, ειδικά σε ομάδες μικροβιωμάτων που έχουν μεγάλο αριθμό ειδών (Wakita et al., 2018).



Σχήμα 1.2 Οι βαθμίδες του ιεραρχικού συστήματος ταξινόμησης του Λινναίου. Με έντονα γράμματα οι συνηθέστερα χρησιμοποιούμενες. (Kostopoulos, 2015)

Η φυλογένεση, μία έννοια που εισήλθε γενικά στην συστηματική μικροβιολογία πιο πρόσφατα σε σύγκριση με την ταξινόμηση, επιτρέπει την ανίχνευση της εξελικτικής ιστορίας των μικροοργανισμών. Μέσω της φυλογενετικής ανάλυσης, η οποία εφαρμόζεται κατά κόρων σε δεδομένα μοριακής αλληλούχισης, προσφέρεται η δυνατότητα κατασκευής φυλογενετικών δέντρων και κατά επέκταση απεικόνισης των πρότυπων διακλάδωσης της

ζωής (**Σχήμα 1.3**). Έτσι, παρέχονται γνώσεις για την κοινή καταγωγή και τις γενετικές συνδέσεις μεταξύ διάφορων μικροοργανισμών.



Σχήμα 1.3 Απεικόνιση φυλογένεσης μεγάλης κλίμακας, που υποδεικνύει τις σχέσεις μεταξύ μεγάλων ομάδων οργανισμών. (Thanukos, 2009)

Τα τελευταία χρόνια, η διαδικασία ταξινόμησης μικροοργανισμών έχει ενσωματώσει σε μεγάλο βαθμό τα φυλογενετικά χαρακτηριστικά που παρέχουν αυτά. Επιπλέον, αν και είναι διακριτά πεδία, η ταξινόμηση και η φυλογένεση συμπλέκονται στην προσπάθεια διαλεύκανσης περίπλοκων μικροβιωμάτων. Σε συνδυασμό, τα δύο πεδία προσφέρουν μια ολιστική άποψη της μικροβιακής ποικιλομορφίας, υπερβαίνοντας την απλή καταλογογράφηση των ειδών.

1.1.4 Ανθρώπινο μικροβίωμα

Το ανθρώπινο σώμα περιέχει έναν τεράστιο αριθμό μικροβίων, που περιλαμβάνει βακτήρια, αρχαία, ιούς και ευκαρυώτες, από τα οποία εξαρτάται άμεσα η ίδια η ύπαρξή του. Γενικά, εκτιμάται ότι ο αριθμός των ανθρώπινων κυττάρων σε ένα μόνο άτομο είναι της τάξεως των 10 τρισεκατομμυρίων (10^{13}), ενώ σύμφωνα με μια εκτίμηση του αριθμού βακτηρίων που φέρει ένα ανθρώπινο σώμα καταλήγει σε μια αναλογία 1,3:1, που σημαίνει ότι για κάθε τρία ανθρώπινα κύτταρα υπάρχουν περίπου τέσσερα αντίστοιχα βακτηριακά. Ωστόσο, αυτές οι εκτιμήσεις δεν λαμβάνουν υπόψη τους μύκητες, τους ιούς και τους φάγους που υπάρχουν σε διάφορα σημεία του σώματος, οι οποίοι, στην περίπτωση των ιών και των φάγων, θα μπορούσαν να ισοδυναμούν με τις βακτηριακές εκτιμήσεις ή, πιο πιθανό, θα μπορούσαν να υπερβαίνουν τον αριθμό τους τουλάχιστον κατά μια τάξη μεγέθους (Gilbert et al., 2018). Παρά το γεγονός ότι είναι 1.000 φορές μικρότερα από τα ανθρώπινα κύτταρα, τα βακτήρια αποτελούν περίπου το 2% της μάζας ενός ενήλικου ανθρώπινου σώματος (1,5 kg), δηλαδή περίπου ισοδύναμο με το μέγεθος ενός ανθρώπινου εγκεφάλου (Moeller et al., 2016). Επιπλέον, έχουν δημοσιευτεί έρευνες στις οποίες το ανθρώπινο μικροβίωμα έχει θεωρηθεί ως το «τελευταίο ανθρώπινο όργανο», διότι εκτός του ότι κληρονομείται εύκολα στο νεογέννητο, όπως κάθε άλλο όργανο, το μικροβίωμα έχει χαρακτηριστικά φυσιολογίας και παθολογίας. Επίσης, η συλλογική υγεία του μικροβιώματος μπορεί να καταστραφεί όταν αλλοιωθεί η πληθυσμιακή δομή του, οδηγώντας σε διάφορες αλλαγές στον άνθρωπο από τη σύλληψη μέχρι και το θάνατο (Baquero & Nombela, 2012).

Συνεξελισσόμενο με τον ξενιστή, το μικροβίωμα έχει συμβάλει στην διαμόρφωση των φαινότυπων των προγόνων του ανθρώπου. Η αντιστοιχία των φυλογενετικών δέντρων

της εντερικής βακτηριακής μικροχλωρίδας και του ανθρώπου (Ochman et al., 2010) αποδεικνύει τη συνεξέλιξη ξενιστή-μικροβιώματος και συνεπάγεται με τη μετάδοση μικροβίων εντός του είδους και μεταξύ των γενεών. Μέσω της διαδικασίας της φυσικής επιλογής, οι μεταλλάξεις οδηγούν σε εξελικτικές προσαρμογές στις περιβαλλοντικές συνθήκες, και γνωρίζοντας ότι τα ανθρώπινα περιβάλλοντα έχουν αλλάξει δραματικά κατά τη διάρκεια της ανθρώπινης εξέλιξης, οι διατροφικές αλλαγές και η έκθεση στον λιμό υπήρξαν σημαντικές πιέσεις. Όμως, ενώ υπάρχουν στοιχεία προσαρμοστικών χαρακτηριστικών επιβίωσης στην πείνα στο ανθρώπινο γονιδίωμα (Hancock et al., 2010), οι προσαρμογές του ανθρώπινου μικροβιώματος που προσφέρουν χαρακτηριστικά εξοικονόμησης ενέργειας για τον ξενιστή παραμένουν ακόμα άγνωστες (Dominguez-Bello et al., 2019).

Θεωρητικά, μαζί με τα μικροβιακά του μέλη, ο άνθρωπος ανέπτυξε ένα ανοσοποιητικό σύστημα του οποίου ένας από τους πολλούς βασικούς ρόλους αποτελεί η προσπάθεια καταπολέμησης του μικροβιακού αποικισμού στο εσωτερικό του σώματός του. Αποτελεί δεδομένο ότι το ανοσοποιητικό σύστημα του ανθρώπου εξελίχθηκε με πολύπλοκους μηχανισμούς αναγνώρισης και καταστροφής εισβαλόντων παθογόνων ή μη μικροβίων σε «απαγορευμένα» σημεία του σώματος, περιορίζοντας τη μικροχλωρίδα κυρίως στα φυσικά εξωτερικά άκρα και κοιλότητες του σώματος του. Τα σημεία αυτά περιλαμβάνουν το στόμα, τις ρινικές κοιλότητες, το λαιμό, το στομάχι, το έντερο, τις ουρογεννητικές οδούς και το δέρμα. Έτσι, το σύνολο των μικροοργανισμών αποτελεί η διεπαφή του ανθρώπινου σώματος και του εξωτερικού περιβάλλοντος. Αυτό σημαίνει ότι σε όλες τις αλληλεπιδράσεις μεταξύ του σώματος και των διάφορων περιβαλλοντικών συνθηκών, όπως η διατροφή και το ηλιακό φως, υπάρχει ένα προστατευτικό πέπλο μικροοργανισμών, των οποίων οι λειτουργίες παίζουν σημαντικό ρόλο στη διατήρηση της γενικής υγείας και ευεξίας του ανθρώπου (Dominguez-Bello et al., 2019). Βέβαια, ο αποικισμός και η προσαρμογή μικροβίων πάνω στα διάφορα σημεία του ανθρώπινου σώματος εξαρτώνται επίσης και από τις χαρακτηριστικές περιβαλλοντικές συνθήκες των σημείων αυτών. Για παράδειγμα, τα προαιρετικά αναερόβια είναι περισσότερο κυρίαρχα στη γαστρεντερική οδό, ενώ τα αυστηρά αερόβια κατοικούν στην αναπνευστική οδό, τη ρινική κοιλότητα και την επιφάνεια του δέρματος. Επιπλέον, το ανθρώπινο μικροβίωμα εξελίσσεται συνεχώς ως απόκριση στους παράγοντες που επηρεάζουν τον ξενιστή. Παράγοντες όπως η ηλικία, η διατροφή, ο τρόπος ζωής, η γεωγραφική τοποθεσία, οι ορμονικές αλλαγές, τα κληρονομικά γονίδια και η υποκείμενη νόσος είναι σημαντικοί καθοριστικοί παράγοντες του ανθρώπινου μικροβιώματος σε κάθε δεδομένη χρονική στιγμή (Aggarwal et al., 2023).

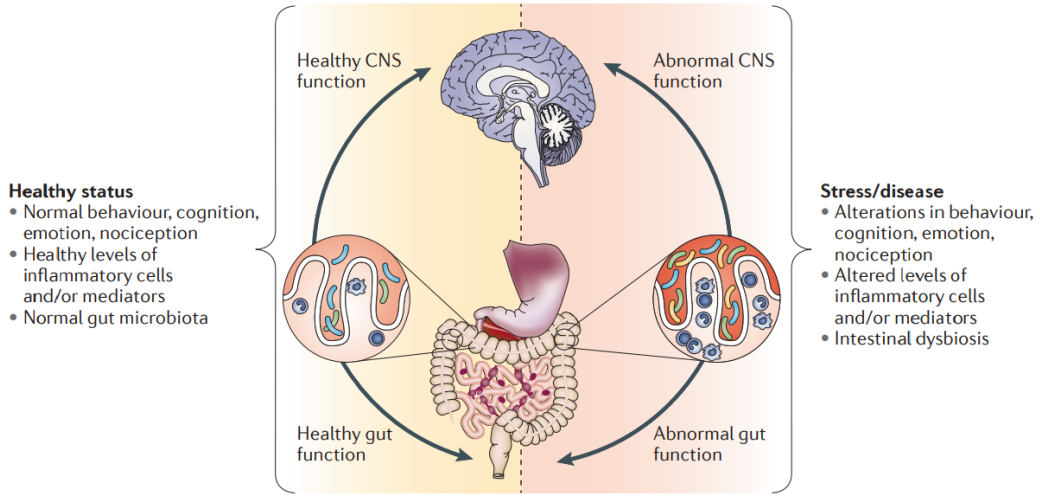
Ωστόσο, η ανάπτυξη της τεχνολογίας ανίχνευσης μικροβιακού DNA επέτρεψε την κατάρριψη του τοπογραφικού περιορισμού, ανακαλύπτοντας ότι και το εσωτερικό του ανθρώπινου σώματος περιέχει μικροβιακές κοινότητες, οι οποίες δεν επιφέρουν απαραίτητα παθογένεια. Ιστοί που κάποτε θεωρούνταν στείροι, όπως ο εγκέφαλος, ο μαστός, ο πλακούντας και το ουροποιητικό σύστημα, φιλοξενούν επίσης μοναδικές βακτηριακές κοινότητες (Aagaard et al., 2014; Branton et al., 2013; Urbaniak et al., 2014; Wolfe et al., 2012). Μάλιστα, μια πρόσφατη μελέτη έδειξε σε περισσότερα από 1500 δείγματα όγκου, συμπεριλαμβανομένων των παρακείμενων φυσιολογικών ιστών, ένα πλούσιο μικροβίωμα που αποτελείται από 528 διαφορετικά βακτηριακά είδη (Nejman et al., 2020). Παρόλα αυτά, ενώ ο ρόλος των μικροβιωμάτων που βρίσκονται στην διεπαφή του ανθρώπινου σώματος με το εξωτερικό περιβάλλον έχει ερμηνευτεί και αποδειχθεί, ο ρόλος του μικροβιώματος στο εσωτερικό του σώματος παραμένει ακόμα αινιγματικός.

1.1.4.1 Μικροβίωμα του εντέρου

Αξιοσημείωτη είναι η αναφορά του εντερικού μικροβιώματος, καθώς και της λειτουργίας του, εφόσον αποτελεί τη μικροβιακή κοινότητα με τη μεγαλύτερη συγκέντρωση στον άνθρωπο και περιλαμβάνει περισσότερα από 1.000 διαφορετικά βακτηριακά είδη (Lloyd-Price et al., 2016). Ωστόσο, κυριαρχούν τέσσερα είδη σε επίπεδο φυλών: οι *Bacteroidetes*, οι *Firmicutes*, τα *Actinobacteria* και τα *Proteobacteria*, με τους *Bacteroidetes* και *Firmicutes* να αποτελούν τα πιο άφθονα είδη σε ένα φυσιολογικό υγιές έντερο. Ένας τεράστιος αριθμός ειδών κυριαρχεί γενικά στα χαμηλότερα ταξινομικά επίπεδα, ωστόσο η σύνθεση αυτών των μικροβίων ποικίλλει σημαντικά, οδηγώντας σε έντονη διακύμανση μεταξύ των ατόμων.

Η συμβιωτική μικροβιακή κοινότητα του εντέρου έχει αποδειχθεί εξαιρετικά κρίσιμη για την εκπαίδευση, ρύθμιση και ανάπτυξη του ανοσοποιητικού συστήματος, για τον μεταβολισμό ορμονών, την πρόσληψη θρεπτικών συστατικών και για την προστασία από την εισβολή παθογόνων μικροοργανισμών (Valdes et al., 2018), η οποία αποτελεί μια βασική υπόθεση ως προς το γιατί οι μεταμοσχεύσεις κοπράνων είναι τόσο αποτελεσματικές στη θεραπεία λοιμώξεων με *Clostridium difficile* (Bagdasarian et al., 2015). Η αναγνώριση της σημασίας του εντερικού μικροβιώματος έχει συμβάλει στη δημιουργία πολλών εθνικών και διεθνών ερευνητικών προγραμμάτων που είναι αφιερωμένα στην εξέταση του ανθρώπινου μικροβιώματος με μεγάλη λεπτομέρεια (Madigan et al., 2021). Όμως, μελέτες έχουν δείξει ότι οι αλλαγές στο ανοσοποιητικό μπορεί να συνδέονται άμεσα με μια δυσβιοτική χλωρίδα του εντέρου. Επίσης, απειλητικές για τη ζωή ασθένειες, όπως ο καρκίνος, οι καρδιαγγειακές παθήσεις, η φλεγμονώδης νόσος του εντέρου και βακτηριακές λοιμώξεις που δύσκολα θεραπεύονται λόγω αντοχής στα αντιβιοτικά, έχουν επίσης συνδεθεί με την εντερική δυσβίωση (Morgan & Huttenhower, 2012; Pascal et al., 2018). Συλλογικά, αυτές οι έρευνες, εκτός του ότι φανερώνουν τον κρίσιμο ρόλο του μικροβιώματος του εντέρου ως προς τον ευεξία του ανθρώπινου σώματος, έχουν δείξει ότι αφενός η μικροβιακή ποικιλομορφία μεταξύ των ατόμων είναι τόσο μεγάλη που δεν υπάρχει το ίδιο μικροβιακό είδος σε όλα τα άτομα, αλλά αφετέρου, η αφθονία ενός βακτηρίου δεν αντικατοπτρίζει απαραίτητα τη λειτουργική του σημασία, καθώς οι παραλλαγές τόσο στα κοινά όσο και στα σπάνια ταξινομικά είδη συνδέονται όλο και περισσότερο με την υγεία και τη λειτουργία του ξενιστή (Lloyd-Price et al., 2016).

Ταυτόχρονα, η ερευνητική ένταση στον τομέα αυτόν έχει φέρει στην επιφάνεια δεδομένα που συσχετίζουν το εντερικό μικροβίωμα όχι μόνο με την σωματική, αλλά και την ψυχική ευεξία του ανθρώπου. Συγκεκριμένα, υπάρχουν σοβαρές ενδείξεις ότι η εντερική μικροβιακή κοινότητα μπορεί να επηρεάσει τη φυσιολογική ανάπτυξη και λειτουργία του εγκεφάλου, και κατ'επέκταση την νευροφυσιολογία και συμπεριφορά του ανθρώπου (**Σχήμα 1.4**) (Mayer et al., 2014; Socała et al., 2021). Αυτό το φαινόμενο υποστηρίζεται από την θεωρία της ύπαρξης του «άξονα εγκεφάλου-εντέρου» (ή αλλιώς «άξονας εντέρου-εγκεφάλου»), ένας όρος του οποίου η πρώτη επίσημη εμφάνιση γίνεται τη δεκαετία του 1980, με την ανάπτυξη της τεχνολογίας απεικόνισης του εγκεφάλου (Cryan et al., 2019).



Σχήμα 1.4 Επίδραση ενός υγιούς και δυσμενούς μικροβιώματος του εντέρου στον άξονα εντέρου-εγκεφάλου. (Doré et al., 2013)

Ο όρος αυτός, που αποτελεί την αμφίδρομη βιοχημική επικοινωνία μεταξύ της γαστρεντερικής οδού και του κεντρικού νευρικού συστήματος, έγινε πιο ξεκάθαρος το 2004 με την ανακάλυψη της εξασθενημένης απόκρισης στο στρες σε ποντίκια χωρίς μικροβιακό φορτίο στο έντερο (Sudo et al., 2004), ωθώντας την ερευνητική κοινότητα να στραφεί στη μελέτη της επιρροής του εντερικού μικροβιώματος στον άξονα εγκεφάλου-εντέρου και στον συμπεριφορικό και ψυχιατρικό αντίκτυπο αυτού στον ξενιστή. Έρευνες έχουν φέρει στην επιφάνεια τις παραλλαγές στο εντερικό μικροβίωμα και την επίδραση σε διάφορες διαταραχές του κεντρικού νευρικού συστήματος, συμπεριλαμβανομένων, ενδεικτικά, του άγχους, των καταθλιπτικών διαταραχών, της σχιζοφρένειας, του αυτισμού, της νόσου Alzheimer και του Parkinson (Cryan et al., 2019). Επίσης, έχουν έρθει πρόσφατα στο προσκήνιο θεραπευτικές παρεμβάσεις για την αντιμετώπιση της δυσβίωσης ή της διαταραχής του εντερικού μικροβιώματος και τον μετριασμό των επιπτώσεών της στον άξονα εγκεφάλου-εντέρου, καθώς διαλευκαίνεται όλο και περισσότερο αυτή η μοναδική σχέση. Ως αποτέλεσμα, έχουν γίνει έρευνες σχετικά με τη χρήση προβιοτικών στη θεραπεία του άγχους και της κατάθλιψης τόσο ως αυτόνομη θεραπεία όσο και ως συμπλήρωμα σε κοινά συνταγογραφούμενα φάρμακα (Clapp et al., 2017). Επιπλέον, υπάρχει ένα αυξανόμενο ενδιαφέρον για την ικανότητα του μικροβιώματος του εντέρου να επηρεάζει τον μεταβολισμό των φαρμάκων, επηρεάζοντας έτσι την ποσότητα του ενεργού φαρμάκου που διατίθεται στον ξενιστή (Swanson, 2015).

Συνεπώς, είναι προφανής ο λόγος που η αμφίδρομη σύνδεση μεταξύ του εγκεφάλου, του εντέρου και του μικροβιώματος έχει έρθει στο προσκήνιο της ιατρικής ερευνητικής κοινότητας τα τελευταία χρόνια. Ο αυξανόμενος όγκος αποδεικτικών στοιχείων που τεκμηριώνουν αυτή τη σύνδεση υποδηλώνει ότι είναι ένας πολύτιμος τομέας για μελλοντική ιατρική και διατροφική πρακτική και έρευνα (Gebrayel et al., 2022). Παρόλα αυτά, το μικροβίωμα του εντέρου δεν αποτελεί το μοναδικό θέμα μελέτης στον τομέα αυτόν, διότι η εξειδικευμένη εστίαση δεν απαντά σε όλα τα ερωτήματα που δημιουργούνται γύρω από το μικροβίωμα και τον τρόπο που επηρεάζει τον άνθρωπο, με αποτέλεσμα να στραφεί η έρευνα σε μικροβιακές κοινότητες των οποίων ο ρόλος είναι μέχρι και σήμερα σε μεγάλο βαθμό ασαφής.

1.1.4.2 Μικροβίωμα του αίματος

Στον άνθρωπο και σε πολλά ζώα, η μεταφορά των θρεπτικών ουσιών στα κύτταρα των ιστών και η απομάκρυνση των αχρήστων από αυτά γίνεται από το κυκλοφορικό σύστημα, το οποίο αποτελείται από την καρδιά, τα αιμοφόρα αγγεία και το αίμα που κυκλοφορεί μέσα σ' αυτά (Saladin, 2011). Εστιάζοντας στο ανθρώπινο αίμα, αυτό αποτελεί ένα υγρό το οποίο περιλαμβάνει το πλάσμα (54%), τα ερυθρά αιμοσφαίρια (45%), τα λευκά αιμοσφαίρια (<1%) και έναν μεταβλητό αριθμό αιμοπεταλίων (<1%), ανάλογα με την κατάσταση της υγείας του ατόμου. Ενώ τα ερυθρά αιμοσφαίρια είναι υπεύθυνα κυρίως για τη μεταφορά οξυγόνου, τα λευκά αιμοσφαίρια χρησιμεύουν ως ένα εξαιρετικά αποτελεσματικό σύστημα επιτήρησης εισβαλλόντων μικροβίων στο αίμα. Η κύρια λειτουργία των αιμοπεταλίων είναι να αντιδρούν στην αιμορραγία από τραυματισμό αιμοφόρων αγγείων μέσω της πήξης (Castillo et al., 2019). Συνεπώς, το αίμα είναι ένα ρευστό μέσο που μεταφέρει και διατηρεί τα πιο βασικά, αλλά αναμφισβήτητα πιο ουσιαστικά, στοιχεία της ζωής. Χρειάστηκε πολύ καιρό και έντονη διαμάχη για να καταρριφθεί η προκατάληψη της σύνδεσης ευεξίας και αποστειρωμένου αίματος και να πραγματοποιηθεί η αποδοχή του όρου «υγιές μικροβίωμα του αίματος». Βέβαια, αν και η φυλογενετική ποικιλομορφία των βακτηρίων που εντοπίζεται στο υγιές αίμα αποτελεί περιορισμένη γνώση, οι σύγχρονες έρευνες υποστηρίζουν την ύπαρξη αυτού του είδους μικροβιώματος, χωρίς περιθώρια αμφισβήτησης. Ενώ ο ρόλος που αποτελεί είναι ακόμα και σήμερα αινιγματικός, μελέτες ενισχύουν όλο και περισσότερο την ιδέα ότι το υγιές μικροβίωμα του ανθρώπινου αίματος είναι αποτέλεσμα συνεχής μικροβιακής μετανάστευσης από εμπλουτισμένα μικροβιώματα του ανθρώπινου αίματος στο κυκλοφορικό σύστημα (Castillo et al., 2019).

<u>Εξέλιξη και διαμόρφωση της έννοιας «υγιές μικροβίωμα του αίματος»</u>

Για πολλά χρόνια, και μέχρι τα μέσα του 20ου αιώνα, το αίμα θεωρούνταν ότι είναι ένα αποστειρωμένο περιβάλλον και ότι η εισβολή ή παρουσία μικροβίων σε αυτό σχετιζόταν μόνο με λοιμώδη νοσήματα (Velmurugan et al., 2020), με το συνηθέστερο να είναι η σήψη του αίματος (Hajj et al., 2018). Τα ευρήματα του Ρόμπερτ Κοχ συνέβαλαν σημαντικά σε αυτήν την θεωρία, με ένα από αυτά να αποτελεί η διαπίστωση ότι το αίμα βοοειδών που είχαν μολυνθεί με την νόσο του άνθρακα (anthrax) περιείχε μεγάλο αριθμό κυττάρων του είδους *Bacillus anthracis* και η απόδειξη της μετάδοσης της νόσου από το ένα ζώο στο άλλο με την έγχυση μικρού όγκου αίματος από το μολυσμένο ζώο σε ένα υγιές (Blevins & Bronze, 2010). Πλησιάζοντας την σύγχρονη εποχή, γίνεται αποδεχτή από την επιστημονική κοινότητα η παροδική ύπαρξη μικροβιακού φορτίου στο αίμα σε «υγιή» (δηλαδή απαλλαγμένα από λοιμώδης νόσημα) άτομα λόγω μικροβιακής μεταφοράς από τον συνήθη τόπο αποικισμού τους, όπως η γαστρεντερική οδός, η στοματική κοιλότητα και το δέρμα, στο κυκλοφορικό σύστημα (Potgieter et al., 2015). Άλλωστε, πολλές μελέτες αναφέρουν και αποδεικνύουν την παρουσία μικροβιακού φορτίου και μικροβιακών μεταβολιτών (παράγωγα μικροβιακού μεταβολισμού) στο αίμα, χωρίς την σύνδεση παθογένειας ή παροδικής επιμόλυνσης, με αποτέλεσμα να έχει προκαλέσει μεγάλο ενδιαφέρον η προοπτική ύπαρξης ενός «υγιούς» μικροβιώματος στο αίμα (Castillo et al., 2019). Παράλληλα, δεν προκαλεί έκπληξη το γεγονός ότι η ιδέα αυτή έχει δεχτεί, και ακόμα δέχεται, έντονη κριτική και αμφισβήτηση λόγω της παραδοσιακής αντίληψης του αποστειρωμένου περιβάλλοντος, πόσο μάλλον όταν η παρουσία ακόμη και ενός βακτηριακού κυττάρου ανά χιλιοστόλιτρο αίματος στην βακτηριαιμία ή σηψαιμία μπορεί να αποτελέσει απειλητική για τη ζωή (Murray & Witebsky, 2014).

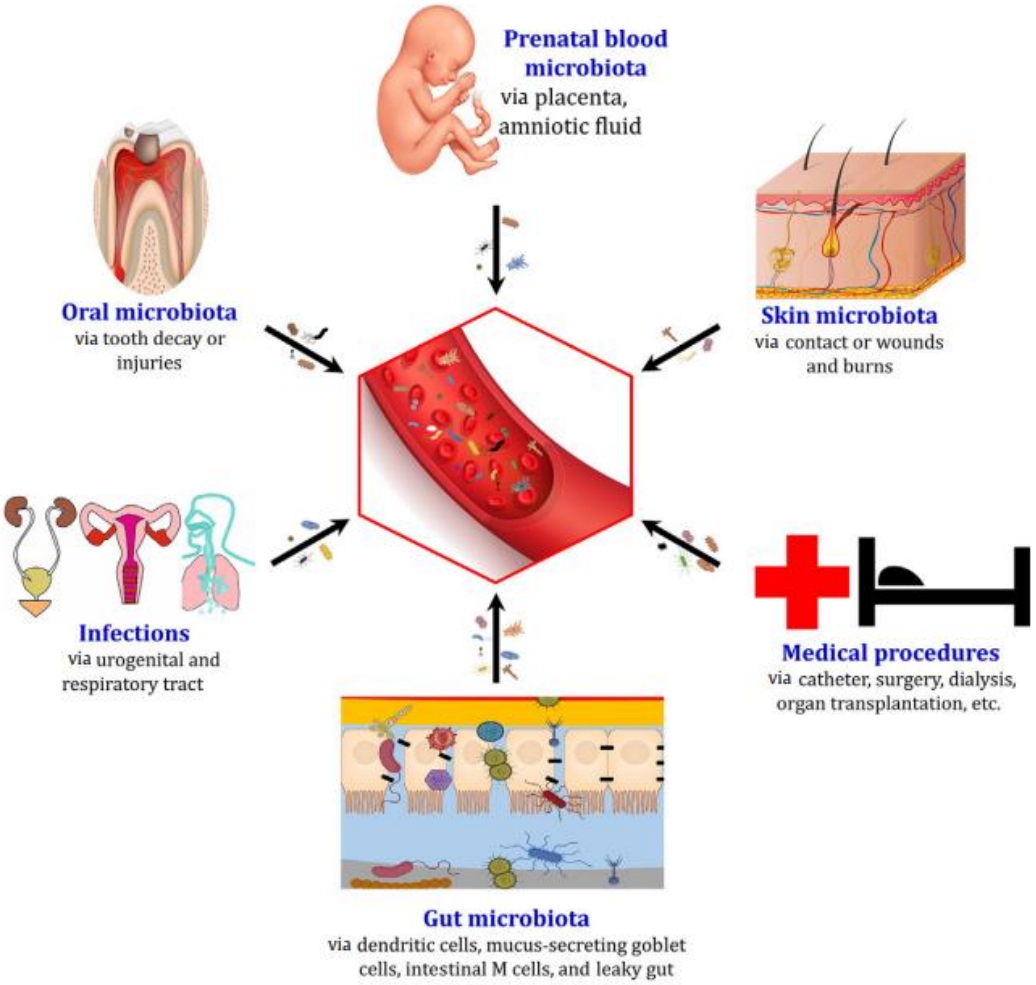
Η διαμάχη σχετικά με την φυσιολογική ύπαρξη «ξένων» κυττάρων στο ανθρώπινο αίμα εκτείνεται πίσω στα τέλη της δεκαετίας του 1960, όπου αναφέρεται για πρώτη φορά η πιθανή παρουσία μεταβολικά ενεργών βακτηρίων στο αίμα υγιών ανθρώπων (Tedeschi et al., 1969) και, σχεδόν μια δεκαετία αργότερα, διαπιστώνεται βακτηριακή ανάπτυξη με τεχνικές καλλιέργειας σε αντίστοιχα δείγματα (Domingue & Schlegel, 1977). Το 2001 εντοπίζεται για πρώτη φορά βακτηριακό γενετικό υλικό στο αίμα υγιών ατόμων με την μέθοδο PCR (Nikkari et al., 2001), και ένα χρόνο μετά επιτυγχάνεται η απεικόνιση ενός βακτηρίου με ιδιαίτερο σχηματισμό στο μικροσκόπιο (McLaughlin et al., 2002), γεγονός που δέχτηκε έντονη κριτική, με τον ισχυρισμό ότι στην πραγματικότητα το μικροσκόπιο δεν παρουσίασε τίποτα περισσότερο από σωματίδια που προέρχονται από την αποσύνθεση ερυθροκυττάρων (Martel et al., 2017; Mitchell et al., 2016). Μέσα στα επόμενα χρόνια, όπου εγκαταλείπεται σιγά σιγά η κλασσική μέθοδος καλλιέργειας και εφαρμόζονται όλο και περισσότερο καινοτόμες αναλυτικές τεχνολογίες, ακολουθούν δεκάδες έρευνες που ενισχύουν την ιδέα της ύπαρξης ενός υγιούς μικροβιώματος στο αίμα (Castillo et al., 2019), στις οποίες εκτός από βακτηριακό, εντοπίζεται γενετικό υλικό ιών (Moustafa et al., 2017), αρχαίων (Dinakaran et al., 2014) και μυκήτων (Panaiotov et al., 2018). Επιπλέον, μελέτες που χαρακτηρίζουν το μικροβιακό προφίλ του αίματος σε άτομα με ασθένεια, έχουν επίσης ανιχνεύσει γενετικό υλικό στις υγιείς ομάδες ελέγχου τους (Q. Li et al., 2018; Qiu et al., 2019).

Εκτός από το *status quo* της ιδεολογίας του αποστειρωμένου αίματος, μεθοδολογικές τεχνικές έχουν παρεμποδίσει την έρευνα του μικροβιώματος του ανθρώπινου αίματος. Πολλοί μικροοργανισμοί που βρίσκονται φυσικά στο ανθρώπινο αίμα μπορεί στην πραγματικότητα να βρίσκονται σε αδρανή κατάσταση, με αποτέλεσμα οι μέθοδοι που βασίζονται στην καλλιέργεια αυτών να μην μπορούν να χρησιμοποιηθούν αξιόπιστα για την υποστήριξη της ύπαρξης του μικροβιώματος (Potgieter et al., 2015). Επιπλέον, η συγκέντρωση βακτηριακού γενετικού υλικού στο αίμα είναι τυπικά πολύ χαμηλή, περιπλέκοντας έτσι δραματικά την εξαγωγή και τον προσδιορισμό αυτού κατά την εφαρμογή ευαίσθητων τεχνικών, όπως αυτή της PCR και της αλληλούχισης (Païssé et al., 2016). Παράλληλα, παρατηρείται μειωμένη η εφαρμογή αυστηρών πειραματικών ελέγχων στις έρευνες, όπως χρήση δειγμάτων ελέγχου (controls), τα οποία είναι απαραίτητα κατά τη μελέτη μικροβιωμάτων χαμηλής βιομάζας που είναι επιρρεπή σε μόλυνση από εξωτερικές πηγές. Το γεγονός αυτό έχει προβληματίσει πολύ την επιστημονική κοινότητα, διότι έρευνες έχουν φέρει στο προσκήνιο την επίδραση που ασκούν οι επιμολύνσεις που προέρχονται από αντιδραστήρια και εργασιακά περιβάλλοντα κατά την εφαρμογή τεχνικών αλληλούχισης. Βέβαια, παρότι έχουν ανιχνευθεί πάνω από 90 διαφορετικά μικρόβια σε επίπεδο γένους σε αντιδραστήρια απομόνωσης DNA και προετοιμασίας δειγμάτων προς αλληλούχιση (Lauder et al., 2016; Salter et al., 2014), οι μελέτες που έχουν συμπεριλάβει δείγματα πειραματικού ελέγχου παρουσιάζουν έντονες διαφορές μεταξύ αυτών και δειγμάτων αίματος από υγιή άτομα, τόσο στην ποσότητα του μικροβιακού γενετικού υλικού όσο και στη μικροβιακή ταξινομική σύνθεση (Dinakaran et al., 2014; Moriyama et al., 2008; Païssé et al., 2016; Traykova et al., 2017). Παρόλα αυτά, επί του παρόντος δεν υπάρχει συγκεκριμένο και αξιόπιστο μέσο διαπίστωσης για το εάν το μικροβιακό DNA και το RNA που εντοπίζονται στο υγιές ανθρώπινο αίμα αντιπροσωπεύουν είτε ζωντανά, είτε νεκρά, είτε ενεργά ή μη βακτήρια (Castillo et al., 2019).

<u>Προέλευση και πύλες εισόδου βακτηρίων στο αίμα</u>

Μέχρι σήμερα, παραμένει αμφιλεγόμενο εάν τα βακτήρια που βρίσκονται στο αίμα εκμεταλλεύονται μια βιώσιμη οικολογικά θέση ή αν απλώς εγκαθίστανται παροδικά στο

αίμα. Ορισμένοι ερευνητές προτείνουν ότι η παρουσία βακτηρίων στο αίμα αποδίδεται σε μεγάλο βαθμό στη μετατόπιση τους από άλλα σημεία του σώματος (**Σχήμα 1.5**). Επίσης, εφόσον η κληρονομική μετάδοση του μικροβιώματος είναι ένα σχεδόν καθολικό φαινόμενο στο ζωικό βασίλειο (Funkhouser & Bordenstein, 2013), όπως για τον ανθρώπινο εντερικό μικροβίωμα, έτσι και τα βακτήρια του αίματος θα μπορούσαν επίσης να έχουν μητρική προέλευση.



Σχήμα 1.5 Προέλευση και πύλες εισόδου μικροβίων στο αίμα. (Velmurugan et al., 2020)

Αν και το εμβρυϊκό και το μητρικό αίμα δεν αναμιγνύεται κατά τη διάρκεια της κύησης, τα βακτήρια θα μπορούσαν να αποικίσουν στο κυκλοφορικό σύστημα του εμβρύου ακόμη και πριν τον τοκετό. Η εφαρμογή της τεχνολογίας αλληλούχισης σε δείγματα αίματος υγιών νεογνών που γεννήθηκαν με καισαρική τομή φανέρωσε την έντονη παρουσία μικροοργανισμών, ενώ στην ίδια έρευνα πραγματοποιήθηκε και η απομόνωση βακτηριακού DNA από τους ομφάλιους λώρους αυτών των νεογέννητων (Jiménez et al., 2005). Έτσι, ενώ πιστεύεται ευρέως ότι το πρώτο εμβόλιο ενός νεογνού προέρχεται από την επαφή με το μικροβίωμα του κόλπου, των κοπράνων ή του δέρματος της μητέρας κατά τη διάρκεια του τοκετού και ότι αυτό το μικροβίωμα του βρέφους στη συνέχεια εμπλουτίζεται μέσω του θηλασμού (Azad et al., 2013; Biasucci et al., 2008; Dominguez-Bello et al., 2010; Penders et al., 2006), ορισμένα στοιχεία υποδεικνύουν την ύπαρξη ενός εμβρυϊκού μικροβιώματος στη μήτρα και τον εμπλουτισμό αυτού του αρχικού συνόλου μικροοργανισμών μετά τη γέννηση (Romano-Keeler & Weitkamp, 2015). Διάφορες ερευνητικές ομάδες έχουν προτείνει την παρουσία βακτηρίων στον πλακούντα, στον αμνιακό σάκο, στις εμβρυϊκές μεμβράνες και στο

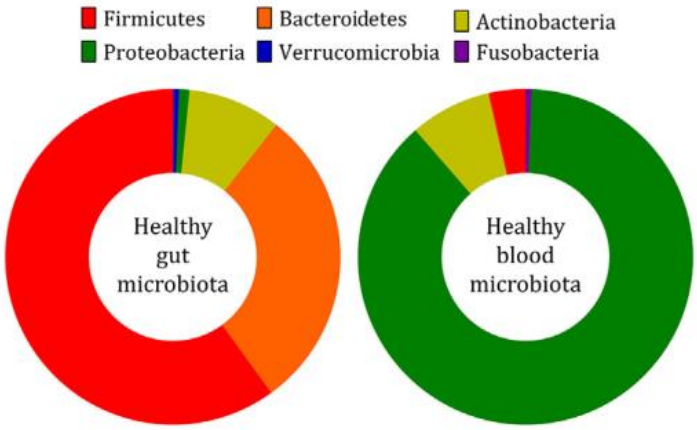
μηκώνιο, ευνοώντας αυτή την υπόθεση (Castillo et al., 2019). Η παρουσία μικροοργανισμών στο αίμα νεογέννητων θα μπορούσε να προέρχεται από άλλα σημεία του σώματος του εμβρύου κατά την διάρκεια της κύησης, όπως το έντερο και την στοματική κοιλότητα. Η υπόθεση αυτή υποστηρίζεται μόνο από ευρήματα που πραγματεύονται την μετατόπιση βακτηρίων στο αίμα από άλλα σημεία του σώματος μετά την γέννα του ανθρώπου, τα οποία θα αναλυθούν στην συνέχεια. Ενώ οι μηχανισμοί που εμπλέκονται στην επακόλουθη μετάδοση βακτηρίων από άλλα σημεία του σώματος στο έμβρυο είναι άγνωστοι, μια πιθανή οδός εισόδου των βακτηρίων στο έμβρυο θα μπορούσε να περιλαμβάνει την κατάποση αμνιακού υγρού κατά τη διάρκεια της κύησης (Romano-Keeler & Weitkamp, 2015).

Μία άλλη πιθανή πηγή μικροοργανισμών στο ανθρώπινο αίμα είναι μέσω μετατόπισης από πηγές με εμπλουτισμένο μικροβίωμα ως αποτέλεσμα τραυματισμών, λοιμώξεων και χαλάρωσης φραγμών. Πράγματι, διάφορες μελέτες έχουν καταλήξει σε συσχέτιση του διαβήτη (Sato et al., 2014), της κίρρωσης (Traykova et al., 2017), διάφορων καρδιαγγειακών παθήσεων (Amar et al., 2013; Dinakaran et al., 2014) και αιματολογικών διαταραχών (Manzo & Bhatt, 2015) με τη μετατόπιση βακτηρίων από την εντερική οδό, κυρίως μέσω του εντερικού επιθηλιακού βλεννογόνου. Ακόμα κι αν η επιθηλιακή μεμβράνη του εντέρου δεν βρίσκεται σε κίνδυνο, άλλοι μηχανισμοί θα μπορούσαν να διευκολύνουν την είσοδο των εντερικών βακτηρίων στο κυκλοφορικό σύστημα. Τα δενδριτικά κύτταρα, τα κυλικοειδή κύτταρα που εκκρίνουν εντερική βλέννα και τα λεμφοειδή κύτταρα που σχετίζονται με τον εντερικό βλεννογόνο δρουν ως μεταβιβαστές βακτηρίων από το έντερο στο αίμα υπό φυσιολογικές συνθήκες (Castillo et al., 2019). Αντίστοιχα, έχει προταθεί ότι βακτήρια που προέρχονται από μικροβίωμα του δέρματος (Cogen et al., 2008; Kowarsky et al., 2017) και του στόματος (Bahrani-Mougeot et al., 2008; Forner et al., 2006) θα μπορούσαν επίσης να διαχέονται στο αίμα όταν διαταράσσονται τα φράγματα μεταξύ αυτών των περιβαλλόντων και του κυκλοφορικού συστήματος. Η σύγκριση δεδομένων μικροβιώματος από υγιές αίμα με δεδομένα που προέρχονται από διάφορα υγιή σημεία του ανθρώπινου σώματος έδειξε ότι το μικροβίωμα του αίματος μοιάζει πολύ με το μικροβίωμα του δέρματος και του στόματος, ενώ διαφέρει ουσιαστικά από το εντερικό μικροβίωμα (Whittle et al., 2018).

Είναι προφανές ότι η ακριβής προέλευση του υγιούς μικροβιώματος του αίματος δεν έχει διευκρινιστεί πλήρως. Θα μπορούσε, υποθετικά, να προέρχεται από τη μητέρα πριν από τη γέννηση ή από τη μετατόπιση μικροοργανισμών που προέρχονται από άλλες πηγές μετά τη γέννηση και κατά τη διάρκεια του κανονικού κύκλου ζωής του ανθρώπου. Όπως το ανθρώπινο εντερικό μικροβίωμα, έτσι και το μικροβίωμα του αίματος μπορεί να επηρεαστεί από την ηλικία, τις διατροφικές και άλλες καθημερινές συνήθειες (πχ. κάπνισμα), και το ανοσοποιητικό σύστημα του ξενιστή. Επίσης, μπορεί κάλλιστα να περιλαμβάνει ένα προσαρμοστικό μικροοικολογικό σύστημα που είναι επιρρεπές σε περιβαλλοντικές επιρροές και έκθεση σε νέα μικροβιακά είδη (Castillo et al., 2019).

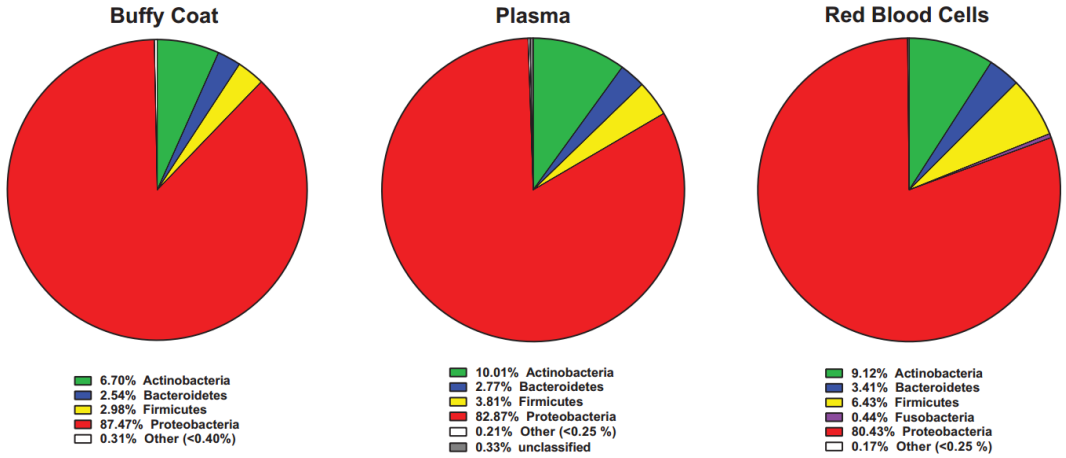
Βακτηριακή σύνθεση στο υγιές αίμα

Παρά το γεγονός ότι η ύπαρξη μικροβιώματος στο αίμα σε υγιή άτομα φαίνεται να υποστηρίζεται από πρόσφατες μελέτες, η γνώση της φυλογενετικής ποικιλομορφίας των βακτηρίων στο αίμα παραμένει περιορισμένη. Σε αντίθεση με τον κυρίαρχο βακτηριακό φύλο που παρατηρείται τυπικά στο ανθρώπινο μικροβίωμα του εντέρου (δηλαδή, *Firmicutes* και *Bacteroidetes*), το υγιές ανθρώπινο μικροβίωμα του αίματος φαίνεται να κυριαρχείται από *Proteobacteria*, και στην συνέχεια από *Actinobacteria, Firmicutes* και *Bacteroidetes* (**Σχήμα 1.6**) (Castillo et al., 2019; Panaiotov et al., 2021).



Σχήμα 1.6 Διαγράμματα πίτας που αντιπροσωπεύουν τη διακύμανση της βακτηριακής ποικιλότητας σε επίπεδο φύλων στο έντερο και το αίμα υγιών ατόμων. (Velmurugan et al., 2020)

Μάλιστα, μία πρόσφατη μελέτη, στην οποία απομονώθηκαν τα 3 βασικά στοιχεία του αίματος από δείγματα που προήλθαν από υγιής ανθρώπους, φανέρωσε μεγάλη βακτηριακή ποικιλομορφία στα ερυθροκύτταρα, ενώ πολύ λιγότερη αντίστοιχα στο πλάσμα και στην λευκοκυτταρική στοιβάδα, η οποία ουσιαστικά αποτελεί το ρευστό που περιέχει μεγάλο ποσοστό των αιμοπεταλίων και λευκών αιμοσφαιρίων (**Σχήμα 1.7**). Όμως, αξιοσημείωτες αποτελούν οι ποσοστιαίες κατανομές του βακτηριακού DNA σε επίπεδο φύλων στα τρία κλάσματα του αίματος, οι οποίες αποκαλύπτουν παρόμοια αποτελέσματα. Συγκεκριμένα, τα *Proteobacteria* παρουσιάζονται κατά μέσο όρο μεταξύ 80,4 και 87,4% και στα 3 κλάσματα, ενώ ακολουθούν τα *Actinobacteria* μεταξύ 6,7 και 10,0%, οι *Firmicutes* μεταξύ 3,0 και 6,4% και τα *Bacteroidetes* μεταξύ 2,5 και 3,4%. Σε επίπεδο ομοταξίας, τα *Fusobacteria* και τα *Flavobacteria* παρατηρούνται σε μεγάλη αφθονία, τα οποία εντοπίζονται κυρίως στα ερυθροκύτταρα, ενώ επίσης αναφέρονται πολλά μέλη της ομοταξίας *Clostridia* να κυριαρχούν στο πλάσμα (Païssé et al., 2016).



Σχήμα 1.7 Διαγράμματα πίτας που αντιπροσωπεύουν τις μέσες σχετικές αναλογίες των βακτηριακών φύλων στα τρία διαφορετικά κλάσματα του αίματος: την λευκοκυτταρική στοιβάδα, το πλάσμα και τα ερυθροκύτταρα.(Païssé et al., 2016)

1.1.5 Συσχέτιση μικροβιώματος και σχιζοφρένιας

Η σχιζοφρένεια είναι μια σύνθετη και ετερογενής νευροψυχιατρική διαταραχή, η οποία αποτελεί μια από τις πιο δύσκολα αντιμετωπίσιμες ψυχιατρικές νόσους. Τα

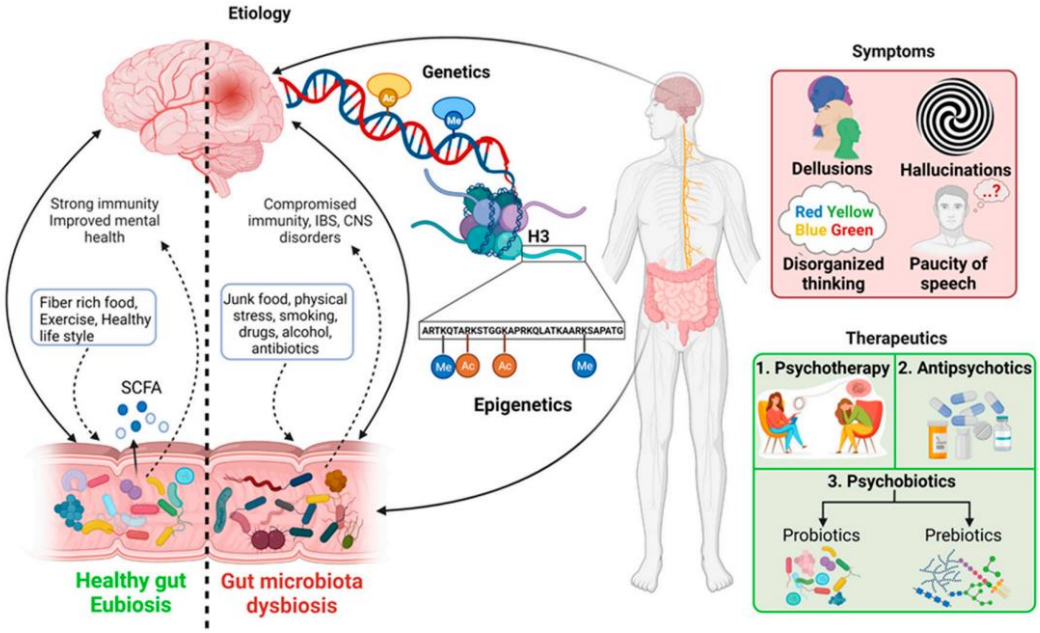
συμπτώματα που χαρακτηρίζουν την ασθένεια κατηγοριοποιούνται σε δύο είδη, στα θετικά και στα αρνητικά συμπτώματα αντίστοιχα (McCutcheon et al., 2020). Τα θετικά συμπτώματα αντικατοπτρίζουν την παραμόρφωση της φυσιολογικής λειτουργίας του εγκεφάλου, συμπεριλαμβανομένου ενδεικτικά των παραληρητικών ιδεών, ψευδαισθήσεων και αποδιοργανωμένων σκέψεων και ομιλιών, ενώ τα αρνητικά συμπτώματα αναφέρονται στη μείωση ή απουσία φυσιολογικών συμπεριφορών. Παραδείγματα αρνητικών συμπτωμάτων είναι η απάθεια, η ανηδονία, η αλογία και η κοινωνική απόσυρση. Η διάγνωση της διαταραχής βασίζεται σε κλινική εκτίμηση των προαναφερόμενων συμπτωμάτων, τα οποία εμφανίζονται συνήθως στην νεαρή ενήλικη ζωή, ενώ συχνά μια πρόδρομη περίοδος προηγείται του πρώτου ψυχωσικού επεισοδίου. Η σχιζοφρένεια έχει επιπολασμό περίπου 1% παγκοσμίως και αντιπροσωπεύει μια τεράστια επιβάρυνση για την υγειονομική περίθαλψη, με το ετήσιο σχετικό κόστος στις Ηνωμένες Πολιτείες να υπολογίζεται σε περισσότερα από 150 δισεκατομμύρια δολάρια (Cloutier et al., 2016). Η διαταραχή σχετίζεται επίσης με μειωμένο προσδόκιμο ζωής, συγκεκριμένα περίπου 15 χρόνια λιγότερα από τον γενικό πληθυσμό, ενώ παράλληλα εκτιμάται ένα ποσοστό κινδύνου θανάτου από αυτοκτονία της τάξεως 5-10% κατά τη διάρκεια της ζωής ενός ασθενή (Hjorthøj et al., 2017).

Η μέθοδος θεραπείας για την αντιμετώπιση της σχιζοφρένειας στην σημερινή εποχή είναι η χορήγηση αντιψυχωσικών φαρμάκων σε συνδυασμό συνήθως με ψυχιατρική παρακολούθηση (McCutcheon et al., 2020). Η βασική λειτουργία πολλών αντιψυχωσικών φαρμάκων είναι να εμποδίζουν ορισμένους υποδοχείς ντοπαμίνης στον εγκέφαλο, δεδομένου ότι έχει παρατηρηθεί μία συσχέτιση μεταξύ των ψυχωσικών επεισοδίων και της αυξημένης απελευθέρωσης ντοπαμίνης. Πράλληλα, οι ψυχολογικές θεραπείες βοηθούν τον ασθενή στην αντιμετώπιση του προκατειλημμένου γνωσιακού σχήματος της διαταραχής του και στην επανεκτίμηση των ψυχωσικών συμπτωμάτων του. Η χορήγηση φαρμάκων καθώς και η ψυχολογική υποστήριξη έχουν αποδειχθεί εξαιρετικά αποτελεσματικές μέθοδοι για την σημαντική μείωση εμφάνισης θετικών συμπτωμάτων, δηλαδή των παραληρηματικών ιδεών και ψευδαισθήσεων.

Ο βασικός λόγος για τον οποίο η σχιζοφρένεια αποτελεί μία από τις πιο δύσκολα αντιμετωπίσιμες διαταραχές του νευρικού συστήματος είναι η έλλειψη κατανόησης της αιτίας εμφάνισης της (McCutcheon et al., 2020). Η πρώτη ερευνητική προσπάθεια για τον προσδιορισμό των αιτιών εκδήλωσης της ασθένειας ήταν επικεντρωμένη στην μελέτη γενετικών παραγόντων. Τα βασικά ευρήματα της προσέγγισης αυτής αποτελούν το αυξημένο ποσοστό κινδύνου κληρονομικότητας της διαταραχής στους απογόνους (6.5%) και οι γενετικές παραλλαγές, οι οποίες σχετίζονται με πολύ αυξημένο κίνδυνο σχιζοφρένειας (30-40%) και περιλαμβάνουν τη διαγραφή ή τον διπλασιασμό συγκεκριμένων τμημάτων του γονιδιώματος. Ωστόσο, τα προαναφερόμενα δεδομένα ερμηνεύουν ένα πολύ μικρό ποσοστό περιπτώσεων σχιζοφρένειας (4%), οδηγώντας την επιστημονική κοινότητα στην αναγνώριση του ρόλου μη-γενετικών παραγόντων. Αν και η σχιζοφρένεια εμφανίζεται συνήθως στην μετεφηβεία, πολλά αποδεικτικά στοιχεία φανερώνουν ότι η παθογένειά της ξεκινά πολύ πιο νωρίς στη νευροανάπτυξη. Αυτά τα στοιχεία περιλαμβάνουν αυξημένα ποσοστά προγεννητικών αντιξοοτήτων, όπως οι μητρικές λοιμώξεις, ο λιμός κατά τη διάρκεια της εγκυμοσύνης και οι μαιευτικές επιπλοκές, συμπεριλαμβανομένου του πρόωρου τοκετού και της προεκλαμψίας. Υπάρχουν επίσης στοιχεία που συνάδουν με την διαταραγμένη πρώιμη νευροανάπτυξη, όπως η αλλοιωμένη ανάπτυξη και οι ήπιες γνωστικές και κινητικές βλάβες στην παιδική ηλικία. Παρόλο που τα δεδομένα αυτά δεν δίνουν ακριβής απάντηση σε ότι αφορά για την βασική αιτία τα σχιζοφρένειας, η συνολική εικόνα υποδεικνύει ότι οι πρώιμοι

περιβαλλοντικοί παράγοντες, σε συνδυασμό, ενδεχομένως, με αυτών των γενετικών, μπορούν να διαταράξουν την ανάπτυξη του εγκεφάλου.

Με την ανερχόμενη θεωρία του άξονα εγκεφάλου-εντέρου και τον κρίσιμο ρόλο του ανθρώπινου μικροβιώματος, τα τελευταία 20 χρόνια η επιστημονική κοινότητα στρέφει την προσοχή της στην συσχέτιση της σχιζοφρένειας και του μικροβιώματος, με έμφαση αυτό του εντέρου (Munawar et al., 2021). Τα ευρήματα αυτής της συσχέτισης έχουν φανερώσει τις παραλλαγές στην σύνθεση και στην λειτουργία του μικροβιώματος του εντέρου ασθενών με σχιζοφρένεια σε σύγκριση με αυτό υγειών, οι οποίες διαφοροποιούν αντίστοιχα την επικοινωνία μεταξύ εντέρου και εγκεφάλου μέσω του άξονα εγκεφάλου-εντέρο (**Σχήμα 1.8**). Μελέτες σε πειραματόζωα, έχουν δείξει ότι το μικροβίωμα του εντέρου είναι ζωτικής σημασίας για την ανάπτυξη και λειτουργία των νευρικών, ανοσολογικών και ενδοκρινικών συστημάτων, τα οποία συχνά εμφανίζουν δυσλειτουργίες σε ασθενείς με σχιζοφρένεια. Μάλιστα, μία ανάλυση του βακτηριακού προφίλ του εντερικού μικροβιώματος σε δείγματα από κόπρανα 58 σχιζοφρενών, στους οποίους είχε χορηγηθεί ένα ή περισσότερα αντιψυχωσικά φάρμακα, πέντε ασθενών χωρίς καμία φαρμακευτική αγωγή και 69 υγιών, υπέδειξε ότι το μικροβίωμα μπορεί να μεταβάλει τη νευρική βιοχημεία και τη νευρική λειτουργία με τρόπους που σχετίζονται με την παθοφυσιολογία της σχιζοφρένειας (Zheng et al., 2019).



Σχήμα 1.8 Τα συμπτώματα, οι εμπλεκόμενοι παράγοντες και οι τρέχουσες θεραπείες στη σχιζοφρένεια. Ο συνδυασμός γενετικών, περιβαλλοντικών παραγόντων, συμπεριλαμβανομένου του μικροβιώματος του εντέρου, έχει ως αποτέλεσμα την εκδήλωση της νόσου. Η σχιζοφρένεια περιλαμβάνει ποικίλα συμπτώματα με περιορισμένες θεραπευτικές επιλογές. Στην αριστερή πλευρά του σχήματος, τα συμπαγή βέλη υποδεικνύουν την πιθανή αιτία της σχιζοφρένειας και τα διακεκομμένα βέλη αντιπροσωπεύουν την αμφίδρομη σχέση του μικροβιώματος του εντέρου τόσο σε υγιές εντερικό μικροβίωμα όσο και σε κατάσταση δυσβίωσης.(Munawar et al., 2021)

Παρόλα αυτά, τα αποτελέσματα ενός σημαντικού αριθμού ερευνών πάνω στην σχιζοφρένεια ως προς την βακτηριακή σύνθεση του εντέρου είναι αντικρουόμενα (Szeligowski et al., 2020). Σε επίπεδο φύλων, τα *Proteobacteria* και τα *Firmicutes* βρέθηκαν τόσο σημαντικά αυξημένα όσο και μειωμένα στη σχιζοφρένεια, ένα φαινόμενο που παρουσιάζεται και στα βακτήρια της ομοταξίας *Clostridia*. Ίσως ένα από τα σταθερά ευρήματα είναι μια σημαντική αύξηση των *Lactobacilli* στη σχιζοφρένεια και σε άτομα με αυξημένο κίνδυνο σχιζοφρένειας, η οποία συσχετίστηκε ακόμη και με τη σοβαρότητα των

συμπτωμάτων. Βέβαια, το γεγονός αυτό έχει μπερδέψει την επιστημονική κοινότητα, δεδομένου ότι οι γαλακτοβάκιλλοι είναι κοινά συστατικά των προβιοτικών και θεωρείται ότι προσφέρουν σημαντικό όφελος στην ψυχική υγεία. Επιπλέον, σημαντική εύρεση αποτελεί η αύξηση των *Firmicutes* και η μείωση *Bacteroidetes* σε εντερικό μικροβίωμα πειραματόζωων και ανθρώπων μετά από την χορήγηση αντιψυχωσικών φαρμάκων. Εξερευνώντας και άλλα σημεία του σώματος, μέσω της σύγκρισης του μικροβιώματος από τον στοματοφάρυγγα μεταξύ ασθενών με σχιζοφρένεια και υγιών ατόμων, διαπιστώθηκε στους ασθενείς αυξημένη συγκέντρωση γαλακτοβακίλλων, ένα φαινόμενο που συνάδει με αυτό του μικροβιώματος του εντέρου.

Ένα από τα πιο χαρακτηριστικά φαινόμενα της δυσβίωσης του άξονα εντέρου-εγκεφάλου αποτελεί η βακτηριακή μετατόπιση της μικροχλωρίδας από το έντερο στο κυκλοφορικό σύστημα (Munawar et al., 2021). Συγκεκριμένα, σε μελέτες που χρησιμοποιούνται ζωικά μοντέλα, η αλλοιωμένη μικροβιακή ποικιλομορφία του εντέρου (δυσβίωση του εντέρου) έχει συσχετιστεί έντονα με την αυξημένη διαπερατότητα του φραγμού του εντέρου (διαρρέον έντερο – leaky gut) και τη μετατόπιση βακτηριακών αντιγόνων όπως οι λιποπολυσακχαρίτες ενδοτοξίνης (LPS) στην κυκλοφορία του αίματος, η οποία είναι γνωστή ως «ενδοτοξαιμία». Η μικροβιακή μετατόπιση έχει ως αποτέλεσμα νευρολογική απόπτωση, πιθανώς οδηγώντας σε ανοσοδιαμεσολαβούμενη ανάπτυξη σχιζοφρένειας. Επιπλέον, οι LPS προκαλούν αυτοανοσία στους περιφερειακούς μηχανισμούς ανοχής. Αυτοί οι μηχανισμοί είναι τα σημεία ελέγχου για τη ρύθμιση των Τ-κυττάρων και της ανοσοκαταστολής που προκαλείται από αυτά. Με την παρουσία τέτοιων μεταβολιτών, ο αιματοεγκεφαλικός φραγμός (BBB) πιθανώς καταστρέφεται, με αποτέλεσμα να παρουσιάζεται νευροφλεγμονή.

Ενώ υπάρχουν έντονες ενδείξεις ότι το μικροβίωμα του αίματος παίζει σημαντικό ρόλο και ενδεχομένως είναι ο συνδετικός κρίκος του άξονα εγκεφάλου-εντέρου, ο αριθμός ερευνών που σχετίζουν το μικροβιακό προφίλ του αίματος με την σχιζοφρένεια είναι δραματικά μικρός. Από αυτές τις μελέτες, αξιοσημείωτο αποτελεί το εύρημα του αυξημένου δείκτη βακτηριακής μετατόπισης CD14 στο πλάσμα στα άτομα με σχιζοφρένεια σε σύγκριση με υγιή άτομα (Severance et al., 2013). Επιπλέον, μία ανάλυση της μικροβιακής κοινότητας σε ολικό αίμα, έδειξε αυξημένη μικροβιακή ποικιλομορφία στη σχιζοφρένεια, γεγονός που μπορεί να αποτελεί κομμάτι της αιτιολογίας της ασθένειας, ή ένα δευτερογενές επακόλουθο της κατάστασης της ασθένειας (Olde Loohuis et al., 2018). Συγκεκριμένα, παρατηρήθηκαν δύο κατηγορίες βακτηρίων, οι *Planctomycetes* και τα *Thermotogae*, τα οποία ήταν παρόντα σε σημαντικά περισσότερα δείγματα σχιζοφρένειας σε σύγκριση με τις άλλες εξεταζόμενες ομάδες ατόμων (πάσχοντες διπολικής διαταραχής, αμυατροφικής πλευρικής σκλήρυνσης καθώς και υγιή άτομα). Στην ίδια μελέτη, παρατηρήθηκε αντίστροφη συσχέτιση μεταξύ της αυξημένης μικροβιακής ποικιλομορφίας και του πληθυσμού των Τ-κυττάρων στην ομάδα ελέγχου, ένα εύρημα το οποίο μπορεί να συνδέεται με τη συσχέτιση του ανοσοποιητικού συστήματος με το μικροβίωμα στο αίμα όπως έχει προαναφερθεί.

Συνεπώς, είναι προφανής η ανάγκη περαιτέρω εξέτασης και ανάλυσης της συσχέτισης του μικροβιώματος στο αίμα και της σχιζοφρένειας, όχι μόνο για την καλύτερη κατανόηση της λειτουργίας του άξονα εγκεφάλου-εντέρου στην νόσο, αλλά και την ενδεχομένως ανακάλυψη ενός βιοδείκτη στο αίμα που συσχετίζεται με την εμφάνιση αυτής της νευρικής διαταραχής. Μία τέτοια ανακάλυψη θα ήταν μεγάλης αξίας, διότι τα πιο αξιοποιήσιμα μοντέλα για την ταυτοποίηση των βιοδεικτών και για τη μελέτη των αντιδράσεων φαρμακοαπόκρισης in vitro είναι τα περιφερικά δείγματα αίματος, καθώς έχουν

παρατηρηθεί μεταβολές που αφορούν το κεντρικό νευρικό σύστημα να αντανακλώνται στο αίμα (Reay et al., 2022). Επίσης, έχει εντοπιστεί ένας μεγάλος αριθμός γονιδίων που εκφράζονται τόσο στο αίμα όσο και στον προμετωπιαίο φλοιό του εγκεφάλου, ενώ, παράλληλα, έχει αποδειχθεί ότι περίπου τα μισά από τα γονίδια που έχουν συσχετιστεί με τη σχιζοφρένεια εκφράζονται και στο αίμα (Wagh et al., 2021). Επιπλέον, τα κύτταρα του αίματος είναι καθιερωμένο κυτταρικό μοντέλο στις ψυχιατρικές ασθένειες, επειδή παρουσιάζουν ομοιότητες με τα κύτταρα του εγκεφάλου (π.χ. πρωτέωμα, ορμονολογικές ομοιότητες, σηματοδότικά μονοπάτια) και έχουν τα πλεονεκτήματα της εύκολης προσβασιμότητας και καλλιέργειας, σε αντίθεση με την βιοψία εγκέφαλου ζωντανών ασθενών (Reay et al., 2022). Άρα, η συλλογή πληροφοριών που συσχετίζουν το αίμα με την σχιζοφρένεια, όπως το μικροβιακό τους φορτίο και διάφορες περιεχόμενες πρωτεΐνες, μπορεί να αποδειχτεί χρήσιμη για την εξαγωγή μοριακών υπογραφών που σχετίζονται με τον συγκεκριμένο τύπο ασθένειας, ή που σχετίζονται με τις διαφοροποιήσεις στην απόκριση αντιψυχωσικού φαρμάκου.

1.2 Μεταγονιδιωματική

1.2.1 Εισαγωγή

Είναι ευρέως αποδεκτό ότι η μικροβιακή ποικιλομορφία είναι τεράστια. Η συντριπτική πλειονότητα αυτών των οργανισμών, περίπου το 99%, δεν μπορούν εύκολα να καλλιεργηθούν και επομένως δεν μπορούν να μελετηθούν χρησιμοποιώντας παραδοσιακές προσεγγίσεις. Η κατανόηση ορισμένων βασικών αρχών της μοριακής βιολογίας, αρχές που διέπουν τις δραστηριότητες όλων των κυττάρων ανεξαρτήτως του τομέα στον οποίο ανήκουν, αποτελεί αναγκαία προϋπόθεση προκειμένου να μελετηθεί επιτυχώς ένα μικροβίωμα.

Η κάθε βιολογική λειτουργία ενός ζωντανού οργανισμού (μονοκύτταρου ή πολυκύτταρου, προκαρυωτικού η ευκαρυωτικού) καθορίζεται από την γενετική πληροφορία, δηλαδή από ένα πλήθος «οδηγιών», το οποίο είναι αποθηκευμένο μέσα στο κύτταρο που τον αρπατίζει. Η λειτουργική μονάδα της γενετικής πληροφορίας είναι το γονίδιο, και τα γονίδια αποτελούν μέρη χρωμοσωμάτων ή άλλων μεγάλων μορίων, γνωστά ως γενετικά στοιχεία. Το συνολικό άθροισμα των γενετικών στοιχείων που καθορίζει έναν οργανισμό ονομάζεται γενετικό υλικό ή γονιδίωμα. Η γενετική πληροφορία είναι ενσωματωμένη σε βιολογικά μακρομόρια , το δεοξυριβονουκλεϊκό οξύ (DeoxyriboNucleic Acid - DNA) και το ριβονουκλεϊκό οξύ (RiboNucleic Acid - RNA), όπου το DNA μεταφέρει το γενετικό προσχέδιο του κυττάρου ενώ το RNA, που παράγεται στο στάδιο της μεταγραφής, φέρει ένα αντίγραφο αυτού του προσχέδιου. Μια συγκεκριμένη μορφή RNA, που ονομάζεται αγγελιαφόρο RNA ή mRNA, μετατρέπεται κατά την διαδικασία της μετάφρασης σε καθορισμένες αλληλουχίες αμινοξέων, τις πρωτεΐνες, οι οποίες αποτελούν τα δομικά μέρη και λειτουργικά στοιχεία του κυττάρου. Συλλογικά, τα νουκλεϊκά οξέα (DNA και RNA) και οι πρωτεΐνες ονομάζονται πληροφορικά μακρομόρια, και αποτελούν όλα τα χαρακτηριστικά που καθιστούν ένα κύτταρο ικανό να επιβιώνει στο περιβάλλον του. Η επιβίωση μπορεί να περιλαμβάνει από την σύνθεση μεμονωμένων χημικών ενώσεων που χρειάζεται το ίδιο το κύτταρο έως την οργάνωση πολύπλοκων δικτύων αλληλεπίδρασης των ενώσεων αυτών, τα οποία ενεργοποιούνται για την εύρυθμη κυτταρική λειτουργία, την απόκρισή σε εξωτερικά ερεθίσματα και τον κυτταρικό πολλαπλασιασμό (Madigan et al., 2021).

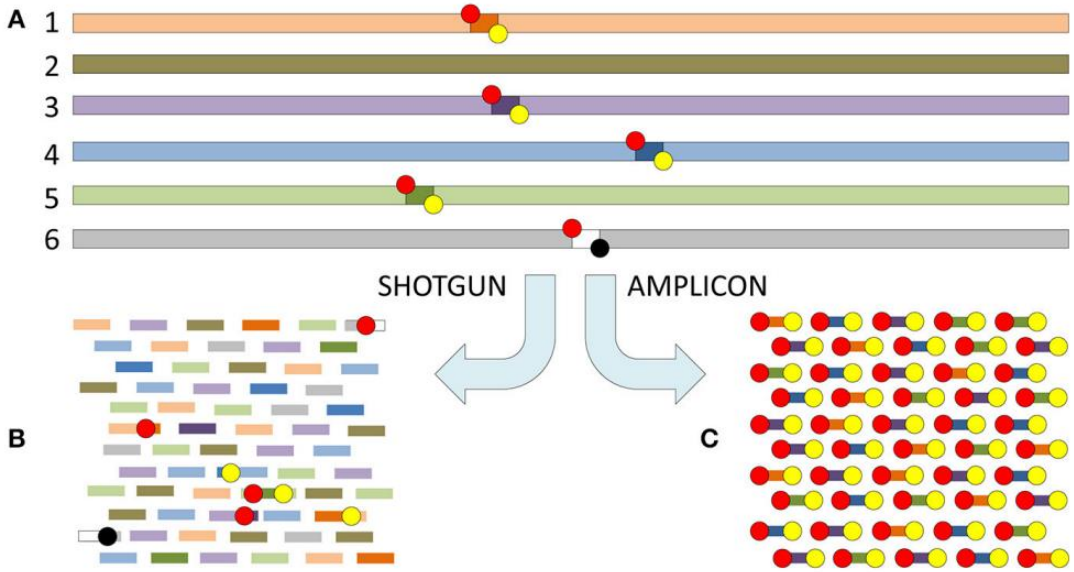
Η ολική αποκρυπτογράφηση των μορίων που αποτελούν ένα κύτταρο, ιστό ή οργανισμό επιτυγχάνεται με την εφαρμογή «-ομικών» τεχνολογιών, οι οποίες αποσκοπούν κυρίως στην καθολική ανίχνευση γονιδίων, mRNA, πρωτεϊνών και μεταβολιτών ενός μεμονωμένου οργανισμού. Αντίστοιχα, οι τεχνολογίες που στοχεύουν την διεξοδική μελέτη των προαναφερόμενων μορίων ονομάζονται γενωμική, μεταγραφωμική, πρωτεωμική και μεταβολομική. Συνεπώς, τα θεμέλια των «-ομικών» τεχνολογιών υπάγονται στην μελέτη των διαδοχικών αλληλουχιών των αζωτούχων βάσεων στα νουκλεϊκά οξέα και των αμινοξέων στις πρωτεΐνες, δηλαδή των χαρακτηριστικών που ελέγχουν την δομή και λειτουργία ενός κυττάρου. Όμως, το βασικό πόρισμα της ροής της γενετικής πληροφορίας είναι ότι τα χαρακτηριστικά αυτά εξαρτώνται αποκλειστικά από το γονιδίωμα του κυτταρικού οργανισμού, καθιστώντας, έτσι, τον κλάδο της γενωμικής μία ύψιστης σημασίας τεχνολογία για την κατανόηση της ταυτότητας, λειτουργίας και εξελικτικής πορείας ενός οργανισμού. Συγκεκριμένα, η γενωμική, ή αλλιώς γονιδιωματική, ορίζεται ως η ανάλυση του γονιδιώματος ή συγκεκριμένων γονιδίων ενός ζωντανού οργανισμού. Η επέκταση του ορισμού αυτού σε έναν μικροβιακό πληθυσμό ονομάζεται μεταγενωμική (metagenomics) ή αλλιώς μεταγονιδιωματική, και αναφέρεται στη μελέτη του σύνολο των γονιδιωμάτων ή συγκεκριμένων γονιδίων διαφορετικών μικροοργανισμών, ανεξαρτήτως με το αν βρίσκονται σε αφθονία μέσα στο εξεταζόμενο δείγμα ή όχι. Αντίστοιχα, η μεταγονιδιωματική προσφέρει την δυνατότητα ανάλυσης του DNA των μικροβιακών πληθυσμών ενός μικροβιώματος, λαμβάνοντας σημαντικές πληροφορίες για την ταυτότητα, την αφθονία και την γενετική πληροφορία των οργανισμών που περιέχει. Η μεταγονιδιωματική μελέτη ενός μικροβιώματος επιτυγχάνεται με την πλήρη αποκωδικοποίηση του μεταγονιδιώματός του, δηλαδή την καταγραφή της αλληλουχίας και των γονιδίων που περιέχει, μία διαδικασία που ονομάζεται αλληλούχιση.

Η μεταγονιδιωματική περιλαμβάνει όλο το φάσμα τεχνικών και τεχνολογιών που χρησιμοποιούνται για την ανάλυση του συνολικού γονιδιωματικού περιεχομένου, ή αλλιώς μεταγονιδιώματος, των μικροοργανισμών που ζουν σε ένα συγκεκριμένο περιβάλλον, χωρίς την καλλιέργεια αυτών (Madigan et al., 2021). Οι τεχνικές αυτές, που προέρχονται από μία πληθώρα διαφορετικών επιστημονικών τομέων όπως η χημεία, η συνθετική βιολογία, η βιοτεχνολογία και η βιοπληροφορική, έχουν γνωρίσει ραγδαία εξέλιξη τα τελευταία χρόνια και οι εφαρμογές τους πληθαίνουν τόσο στην βιομηχανία (Coughlan et al., 2015) όσο και στους κλάδους υγείας (Chiu & Miller, 2019). Η κινητήρια δύναμη πίσω από αυτήν την εξέλιξη είναι η ανάπτυξη καινούριων τεχνολογιών γονιδιακής αλληλούχισης (αλληλούχιση νέας γενιάς - next generation sequencing) οι οποίες επιτρέπουν την ψηφιακή καταγραφή των γενετικών πληροφοριών με πολύ μεγαλύτερη απόδοση από ότι στο παρελθόν. Ταυτόχρονα, λόγου του υπεραυξημένου όγκου δεδομένων που παράγουν αυτές οι τεχνολογίες και της αναγκαίας υπολογιστικής ανάλυσης αυτών, έχουν αναπτυχθεί και διαδοθεί διάφορα βιοπληροφορικά εργαλεία, τα οποία προσφέρουν υπολογιστικά και στατιστικά μοντέλα με σκοπό την μελέτη μικροβιακών κοινοτήτων (Almeida & De Martinis, 2019).

1.2.2 Μεθοδολογίες μεταγονιδιωματικής αλληλούχισης

Οι προσεγγίσεις της μεταγονιδιωματικής ταξινομούνται σε δύο ομάδες - στην ολική μεταγονιδιωματική και στην στοχευμένη μεταγονιδιωματική, οι οποίες βασίζονται σε στρατηγικές τυχαίας και επιλεκτικής αλληλούχισης, αντίστοιχα (**Σχήμα 1.9**). Ο όρος ολική μεταγονιδιωματική αναφέρεται κυρίως στην μαζική αλληλούχιση μεταγονιδιωμάτων (shotgun metagenomics), ενώ η στοχευμένη μεταγονιδιωματική αναφέρεται στην

αλληλούχιση μεμονωμένων γονιδίων ή χαρακτηριστικών αλληλουχιών (amplicon metagenomics) για κάθε είδος μικροοργανισμού που εμπεριέχεται στο εξεταζόμενο δείγμα (Bharti & Grimm, 2021). Και οι δύο προσεγγίσεις αποτελούν τεχνικές μελέτης μικροβιωμάτων, οι οποίες βοηθούν στον ποιοτικό και ποσοτικό προσδιορισμό των διαφορετικών μικροοργανισμών σε τέτοιες κοινότητες, ξεπερνώντας την ανάγκη της απομόνωσης και καλλιέργειας αυτών. Παρόλα αυτά, η κάθε προσέγγιση έχει τα δικά της προτερήματα, ενώ παράλληλα και οι δύο έχουν ιδιαιτερότητες, οι οποίες παίζουν πολύ μεγάλο ρόλο στα τελικά αποτελέσματα της εκάστοτε ανάλυσης.



Σχήμα 1.9 Η διαφορά μεταξύ της μαζικής μεταγονιδιωματικής αλληλούχισης και στοχευμένης μεταγονιδιωματικής αλληλούχισης αμπλικονίου. **(A)** Έξι γονιδιώματα, εμφανιζόμενα με διαφορετικά χρώματα, προέρχονται από 6 διαφορετικούς μικροοργανισμούς. Τα 5 από αυτά (1 και 3-6) περιέχουν μία κοινή περιοχή στην αριστερή πλευρά, εικονιζόμενη με μία κόκκινη τελεία, εκ των οποίων τα 4 από αυτά παρέχουν και μία ακόμα κοινή περιοχή στην δεξιά πλευρά, εικονιζόμενη με κίτρινη τελεία. **(B)** Στην μαζική μεταγονιδιωματική αλληλούχιση, προκύπτουν θραύσματα αλληλουχιών από τα 6 γονιδιώματα με τυχαίο μοτίβο. **(C)** Τα θραύσματα αλληλουχιών της στοχευμένης μεταγονιδιωματικής αλληλούχισης αμπλικονίου προκύπτουν από τις περιοχές των γονιδιωμάτων που οριοθετούνται από ένα συντηρημένο αριστερό και δεξιό μοτίβο (κόκκινες και κίτρινες τελείες). (Sekse et al., 2017)

1.2.2.1 Shotgun metagenomics

Στην μέθοδο μαζικής μεταγονιδιωματικής αλληλούχισης, το DNA εξάγεται από όλα τα κύτταρα που εμπεριέχονται στο εξεταζόμενο δείγμα, τεμαχίζεται σε μικρά θραύσματα, τα οποία είτε ενισχύονται με την μέθοδο PCR, ή εισάγονται κατευθείαν σε μηχάνημα αλληλούχισης. Με αυτόν τον τρόπο, καταγράφεται όλο το μεταγονιδίωμα, συμπεριλαμβανομένου ενδεχομένως και το γονιδιώματος του ξενιστή, παρέχοντας γενετικές πληροφορίες που αφορούν τους γονιδιωματικούς δεσμούς μεταξύ λειτουργίας και φυλογένεσης των καλλιεργησίμων και μη οργανισμών, το εξελικτικό προφίλ της λειτουργίας και δομής της συνολικής κοινότητας, καθώς και τους δυνητικά νέους βιοκαταλύτες ή ένζυμα (Bharti & Grimm, 2021). Οι πληροφορίες αυτές μπορούν επίσης να συμπληρωθούν με μετα-μεταγραφικές ή μετα-πρωτεωμικές προσεγγίσεις για την περιγραφή εκφραζόμενων λειτουργιών. Η μεταγονιδιωματική αποτελεί, επίσης, ένα ισχυρό εργαλείο για τη δημιουργία νέων υποθέσεων της μικροβιακής λειτουργίας ή για τον εντοπισμό εμπλεκόμενων

μεταβολικών μικροβιακών λειτουργιών (Thomas et al., 2012). Παρά αυτά τα πλεονεκτήματα, τα δεδομένα μεταγονιδιωματικής αλληλούχισης παρουσιάζουν αρκετές προκλήσεις, με τις βασικότερες να αποτελούν το μεγάλο κόστος της μεθόδου και η εκτεταμένη ανάλυση των παραγόμενων δεδομένων (Ranjan et al., 2016).

1.2.2.2 Amplicon metagenomics

Στην μέθοδο της μεταγονιδιωματικής αλληλούχισης αμπλικονίων, εξάγεται όλο το DNA του εξεταζόμενου δείγματος και ενισχύεται η στοχευόμενη περιοχή αλληλουχίας με την μέθοδο PCR χρησιμοποιώντας τους κατάλληλους εκκινητές (Liu et al., 2021), Στην συνέχεια, τα παραγόμενα αμπλικόνια προεπεξεργάζονται κατάλληλα για την αλληλούχισή τους σε πλατφόρμα αλληλούχισης. Συνήθως, η στοχευόμενη περιοχή αλληλουχίας αποτελεί περιοχή γνωστού γονιδίου, του οποίου το περιεχόμενο αποτελείται από πολλαπλές συντηρημένες και μεταβλητές περιοχές αλληλουχιών μεταξύ των μικροβιακών ειδών. Συνεπώς, η αλληλούχιση μίας περιοχής που περιέχει συντηρημένες και μεταβλητές αλληλουχίες παρέχει εξαιρετικό γενετικό υλικό για ταξινομικό και φυλογενετικό προφίλ του αμπλικονίου, με αποτέλεσμα την ταυτοποίηση του κυττάρου από το οποίο προήλθε.

Τα γονίδια που παρέχουν χαρακτηριστικά φυλογενετικών δεικτών είναι αυτά που εκφράζουν το ριβοσωμικό RNA, με τις περιοχές 16S, 18S και ITS να αποτελούν τις πιο μελετημένες και συχνά στοχευόμενες περιοχές για την εφαρμογή μεταγονιδιωματικής αλληλούχισης αμπλικονίου. Η αλληλούχιση του 16S rRNA γονιδίου χρησιμοποιείται ευρέως σε φυλογενετικές και ταξινομικές μελέτες βακτηρίων και αρχαίων. Η αλληλούχιση του 18S rRNA γονιδίου μπορεί να εφαρμοστεί σε ανάλυση μικροβιακών ευκαρυωτικών κυττάρων, όπως τους μύκητες και τα πρώτιστα. Η αλληλούχιση των εσωτερικών μεταγραφόμενων διαχωριστών (internal transcribed spacers – ITS) είναι επίσης μια προτιμώμενη μέθοδος μελέτης πληθυσμού μυκήτων, η οποία έχει μεγάλη σημασία για τη μοριακή φυλογενετική ανάλυση των μυκήτων σε επίπεδο γένους και είδους.

Η εφαρμογή της μεθόδου της μεταγονιδιωματικής αλληλούχισης αμπλικονίων παρουσιάζεται συχνότερα σε μελέτες μικροβιώματος σε σχέση με αυτή της μαζικής αλληλούχισης μεταγονιδιωμάτων (Bharti & Grimm, 2021). Αυτό οφείλεται στο γεγονός ότι η μέθοδος μπορεί να ελέγξει αποτελεσματικά την παρουσία συγκεκριμένου μικροοργανισμού ή παραλλαγές αυτού στο εξεταζόμενο δείγμα και να περιγράψει, καθώς και να συγκρίνει την ποικιλομορφία πολλών πολύπλοκων περιβαλλόντων. Επίσης, η ευελιξία των αμπλικονίων παρέχει γρήγορες, έγκυρες και οικονομικά αποδοτικές λύσεις για την κάλυψη των αναγκών διαφορετικών ερευνητικών έργων. Επιπλέον, αποφεύγοντας την αλληλούχιση του γονιδιώματος του ξενιστή, η οποία είναι αναπόφευκτή διαδικασία στην περίπτωση της μαζικής μεταγονιδιωματικής αλληλούχισης, η στοχευμένη αλληλούχιση αποτελεί ιδανική μέθοδος ανάλυσης μικροβιώματος με χαμηλή βιομάζα. Παράλληλα, παράγεται μικρότερος όγκος δεδομένων, με αποτέλεσμα την λιγότερο επιβαρυντική ανάλυσή τους. Η μεγαλύτερη αδυναμία της στοχευμένης αλληλούχισης αποτελεί το βάθος της ανάλυσης, καθώς εκτός του ότι οι φυλογενετικοί δείκτες περιορίζονται στην ταξινόμηση των μικροοργανισμών κυρίως μέχρι και σε επίπεδο γένους, δεν προσφέρουν επίσης πληροφορίες σχετικά με την λειτουργικότητά τους μεταξύ αυτών και του περιβάλλοντος από το οποίο προέρχονται. Βέβαια, υπάρχουν διαθέσιμα βιοπληροφορικά εργαλεία τα οποία χρησιμοποιούν ταξινομικά δεδομένα για την πρόβλεψη μεταβολικής λειτουργίας (Langille et al., 2013), ξεπερνώντας εν μέρει ένα από τα βασικά μειονεκτήματα.

1.2.3 Τεχνικές Αλληλούχισης Νέας Γενιάς (Next Generation Sequencing – NGS)

Με τα έμπρακτα αποτελέσματα του Προγράμματος Χαρτογράφησης του Ανθρώπινου Γονιδιώματος, του θεμελιώδους προγράμματος από διεθνείς ερευνητικές ομάδες που κράτησε από το 1990 έως το 2003, η επιστημονική κοινότητα εφάρμοσε για πρώτη φορά αναλύσεις μεταγονιδιωματικής αλληλούχισης στην αρχή της χιλιετίας (Jünemann et al., 2017). Αν και αποτέλεσαν πρωτοπόρα έργα, σκοπός των οποίων ήταν η διερεύνηση ενδιαιτημάτων που προηγουμένως θεωρούνταν σχεδόν απρόσιτα λόγω των ακραίων συνθηκών ή της πλούσιας ποικιλομορφίας τους, η διεξαγωγή τους ήταν ακόμα δαπανηρή και χρονοβόρα εκείνη την εποχή. Αυτό οφείλεται στο γεγονός ότι, μέχρι τότε, ο προσδιορισμός των αλληλουχιών του DNA βασιζόταν στις κλασσικές μεθόδους αλληλούχισης Sanger, δηλαδή στην άμεση κλωνοποίηση του περιβαλλοντικού DNA σε συνδυασμό με την τριχοειδή ηλεκτροφόρηση.

Ωστόσο, η έλευση των τεχνολογιών Αλληλούχισης Νέας Γενιάς (NGS) λίγο αργότερα έφερε επανάσταση στην αγορά και στην έρευνα με βάση την αλληλούχιση, οδηγώντας έτσι σε δραματική πτώση στο κόστος προσδιορισμού αλληλουχιών ενώ ταυτόχρονα απλοποιούσε όλη την διαδικασία, από την προετοιμασία του δείγματος μέχρι και τα παραγόμενα αποτελέσματα. Το πραγματικό πλεονέκτημα σε σχέση με την παραδοσιακή αλληλούχισης Sanger είναι ότι τα όργανα NGS μπορούν να πραγματοποιήσουν παράλληλα εκατομμύρια αντιδράσεις προσδιορισμού αλληλουχίας με πολύ υψηλή απόδοση και είναι σε θέση να αποκρυπτογραφήσουν όλο το DNA που περιέχεται σε ένα μείγμα σε μία προσπάθεια προσδιορισμού αλληλουχίας. Αυτό άνοιξε το δρόμο για την έρευνα με βάση την αλληλούχιση όλων των περιβαλλόντων και ειδών μικροοργανισμών, κυρίως εκείνων που είναι αδύνατο να αναπτυχθούν σε καθαρή καλλιέργεια, όπως για παράδειγμα των συμβιωτικών ή των απολιθωμένων υπολειμμάτων αρχαίων οργανισμών.

Η μέθοδος NGS αποτελεί ένα από τα μεγαλύτερα άλματα της μοριακής βιολογίας των τελευταίων 20 ετών και έχει βρει εφαρμογή σε πολυάριθμους κλάδους, όπως αυτοί της υγείας, της μικροβιολογίας, της βοτανικής και της βιομηχανίας τροφίμων. Με τη δραματική πτώση του κόστους και του χρόνου αλληλούχισης, πλέον η NGS είναι μια προσιτή αλλά και πρακτική μέθοδος για την διερεύνηση μικροβιωμάτων.

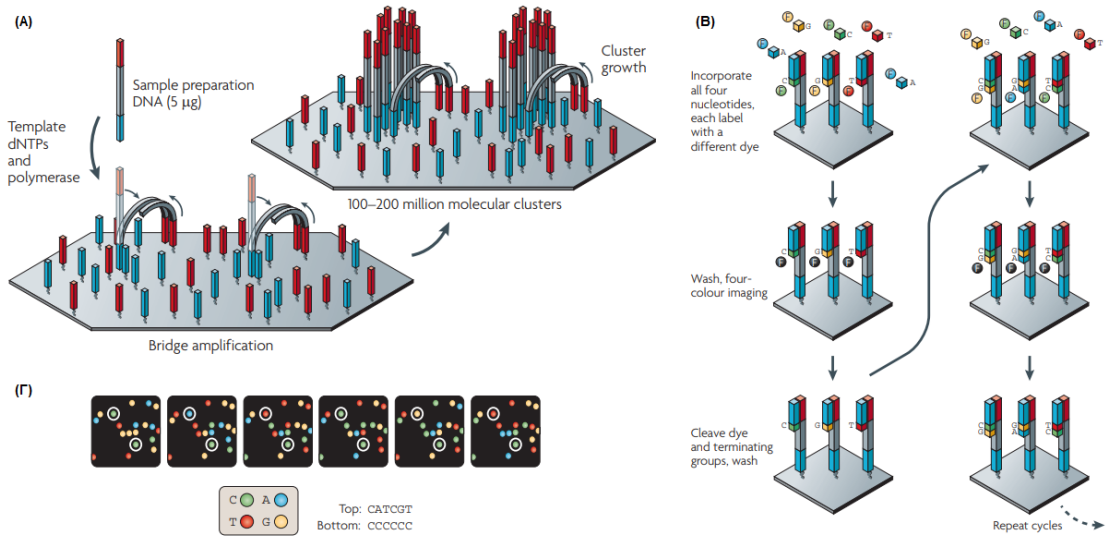
Γενικά, η μεταγονιδιωματική αλληλούχιση με την μέθοδο NGS περιλαμβάνει 3 βασικά βήματα: (1) τον κατακερματισμό ή την ενίσχυση του στοχευόμενου DNA, ανάλογα ποια μεταγονιδιωματική μέθοδο εφαρμόζεται (shotgun και amplicon αντίστοιχα), (2) την προετοιμασία μεταγονιδιωματικής βιβλιοθήκης, μια διαδικασία στην οποία τα διάφορα τμήματα του DNA τροποποιούνται έτσι ώστε κάθε τμήμα DNA να εξειδικεύεται και να μπορεί να χρησιμοποιηθεί για την εισαγωγή του σε πλατφόρμα αλληλούχισης, την παράλληλη αλληλούχισή του με άλλα θραύσματα DNA και την αναγνώριση του αντίστοιχου δείγματος από το οποίο προήλθε, και (3) την αλληλούχιση, στην οποία η μεταγονιδιωματική βιβλιοθήκη φορτώνεται σε μια πλατφόρμα αλληλούχισης και επιτυγχάνεται έτσι ο προσδιορισμός των αλληλουχιών των θραυσμάτων DNA, με τα παραγόμενα δεδομένα που προκύπτουν να ονομάζονται αναγνώσεις αλληλουχιών ή αναγνώσματα (reads).

Στο διάστημα της εικοσαετίας, οι πιο διαδεδομένες πλατφόρμες NGS που αναπτύχθηκαν αποτελούν αυτές τις δεύτερης γενιάς, δηλαδή οι Roche/454 Pyrosequencing (Margulies et al., 2005), SOLiD (Pandey et al., 2008), Ion Torrent (Rothberg et al., 2011) και Illumina/Solexa (Bentley et al., 2008) των οποίων η διαφορά βασίζεται στην τεχνική αλληλούχισης που εφαρμόζουν. Στην προσπάθεια ανάπτυξης πιο σύγχρονων πλατφόρμων,

αναπτύχθηκαν οι μεθοδολογίες μονομοριακής αλληλούχισης (single-molecule DNA sequencing) (Gupta, 2008) με πρωτεργάτες τις εταιρείες Helicos (Check Hayden, 2009) και PacBio (Au et al., 2012). Αυτές οι τεχνολογίες ονομάστηκαν τεχνολογίες αλληλούχισης τρίτης γενιάς, αλλά η αξιοποίησή τους από την επιστημονική κοινότητα ήταν σχετικά περιορισμένη έως πρόσφατα εξ' αιτίας της μεγαλύτερης προδιάθεσης που έχουν σε λάθη κατά την καταγραφή των βάσεων σε σχέση με τις τεχνολογίες δεύτερης γενιάς (Bleidorn, 2016).

1.2.3.1 Illumina/Solexa

Η πιο επιτυχημένη τεχνική αλληλούχισης νέας γενιάς είναι αυτή της εταιρίας Illumina, χρησιμοποιώντας την τεχνολογία που αναπτύχθηκε για πρώτη φορά από τη Solexa και τη Lynx Therapeutics (Slatko et al., 2018). Η Illumina εφαρμόζει εντελώς νέες μεθόδους, τόσο για το στάδιο ποσοτικής ενίσχυσης των θραυσμάτων DNA, όσο και για τον τελικό προσδιορισμό των βάσεων που προστίθενται κατά την αντίδραση αλληλούχισης. Η μέθοδος αυτή χρησιμοποιεί την ακινητοποίηση όλων των θραυσμάτων σε στερεό υπόστρωμα, όπου ακολουθεί ποσοτική ενίσχυση γέφυρας (bridge amplification) (**Σχήμα 1.10**). Κατά την αντίδραση αυτή ο πολυμερισμός συμβαίνει με ταυτόχρονη σύνδεση και των δύο άκρων κάθε θραύσματος σε γειτονικές αλληλουχίες-προσαρμογείς με τελικό αποτέλεσμα τον σχηματισμό ενός συσσωματώματος αλληλουχιών-κλώνων ανά θραύσμα. Ο εντοπισμός των βάσεων των θραυσμάτων ανά συσσωμάτωμα γίνεται με τη χρήση ειδικά κατασκευασμένων τριφωσφορικών δεοξυνουκλεοτιδίων (dNTPs), ώστε να σταματάνε την αντίδραση πολυμερισμού σε κάθε κύκλο, καθώς και τη χρήση ιχνηθετημένων με τέσσερις διαφορετικές φωσφορίζουσες ομάδων ώστε να είναι δυνατή η ταυτόχρονη καταγραφή τους. Οι αντιδράσεις επαναλαμβάνονται για N+1 γύρους, όπου N είναι ο αριθμός του μήκους της αλληλουχίας που επιθυμείται να προσδιοριστεί , και η προσθήκη μίας ακόμα βάσης οφείλεται στον καλύτερο προσδιορισμό της βάσης στην θέση N.



Σχήμα 1.10 Μέθοδος αλληλούχισης Solexa/Illumina. **(Α)** Η ποσοτική ενίσχυση θραυσμάτων στερεάς φάσης (solid-phase amplification), η οποία αποτελείται από δύο στάδια. Το πρώτο στάδιο αφορά την ακινητοποίηση των μονόκλωνων θραυσμάτων (γκρι χρώμα) σε στερεό υπόστρωμα χάρις τις αλληλουχίες-προσαρμογείς (κόκκινο και μπλε χρώμα) που έχουν συνδεθεί στα άκρα τους και δρουν και ως εκκινητές. Η ακινητοποίηση των θραυσμάτων γίνεται παράλληλα με την προσθήκη DNA πολυμεράσης και dNTPs ώστε να αρχίσει η αντίδραση πολυμερισμού των συμπληρωματικών τους κλώνων. Στο δεύτερο στάδιο γίνεται ποσοτική ενίσχυση γέφυρας (bridge amplification) του κάθε θραύσματος με την κάθε αλυσίδα-κλώνο να προσκολλάται σε κάποιο γειτονικό ολιγονουκλεοτίδιο-

προσαρμογέα κατά τη διάρκεια της αντίδρασης πολυμερισμού. Κατά αυτόν τον τρόπο δημιουργούνται συσσωματώματα (clusters) κλώνων από κάθε θραύσμα πάνω στο στερεό υπόστρωμα που δίνουν ισχυρότερο σήμα κατά την αντίδραση αλληλούχισης. **(Β)** Η αντίδραση επιμήκυνσης συμπληρωματικών αλυσίδων χρησιμοποιεί ειδικά κατασκευασμένα dNTPs συνδεδεμένα με μία ομάδα -O-αζιδομεθυλίου (-O- N3) στο άκρο τους καθώς και ιχνηθετημένα με τέσσερις διαφορετικές φωσφορίζουσες ομάδες F. Η ομάδα αζιδομεθυλίου σταματά την αντίδραση πολυμερισμού μετά την προσθήκη τους και η φωσφορίζουσες ομάδες επιτρέπουν την καταγραφή σήματος εικόνας για κάθε συσσωμάτωμα θραυσμάτων-κλώνων. Την καταγραφή ακολουθεί ενζυμικό κόψιμο και απομάκρυνση των φωσφοριζουσών ομάδων και των αζιδομεθυλίων αφήνοντας μία ομάδα -OH στην 3' θέση των ελεύθερων νουκλεοτιδίων, ώστε να μπορέσει να συνεχιστεί η αντίδραση πολυμερισμού με την προσθήκη dNTPs στον επόμενο κύκλο αντιδράσεων. **(Γ)** Ο προσδιορισμός των βάσεων των θραυσμάτων γίνεται με διαδοχική καταγραφή εικόνων για κάθε συσσωμάτωμα. Τα συσσωματώματα των οποίων η αλληλουχία εξετάζεται παραπάνω είναι σημειωμένα με άσπρο κύκλο. (Metzker, 2010)

Σήμερα, η Illumina υποστηρίζει μια τεράστια ποικιλία πρωτοκόλλων αλληλούχισης, συμπεριλαμβανομένων των μεθόδων της μεταγονιδιωματικής αλληλούχισης (Slatko et al., 2018). Οι διάφορες πλατφόρμες αλληλούχισης που προσφέρει η Illumina, όπως τα μοντέλα MiniSeq, MiSeq, NextSeq, NovaSeq και HiSeq, παρέχουν διαφορετικά επίπεδα απόδοσης και μήκη αναγνωσμάτων. Αυτό οφείλεται στο γεγονός ότι οι πλατφόρμες είναι ικανές για την αλληλούχιση μονού άκρου (single-end) ή ζεύγους άκρων (paired-end). Κατά την αλληλούχιση single-end, η αλληλουχία προσδιορίζεται από το ένα της άκρο, ενώ στην paired-end αλληλούχιση προσδιορίζεται και από τα δύο άκρα της, με αποτέλεσμα να προκύπτουν δυο αναγνώσματα του ίδιο θραύσματος DNA. Από την μέθοδο single-end προκύπτει μεγάλος όγκος δεδομένων γρήγορα και με χαμηλό, σχετικά, οικονομικό κόστος, ενώ στην μέθοδο paired-end δημιουργούνται υψηλότερης ποιότητας δεδομένα αλληλούχισης και παράλληλα προσδιορίζεται μεγαλύτερο μήκος αλληλουχιών.

Συνολικά, οι πλατφόρμες Illumina παρέχουν εξαιρετικά καλά χαρακτηριστικά, όσον αφορά την κάλυψη μήκους αλληλουχιών (150 έως 600 bp), την παραγωγή μεγάλου όγκου δεδομένων (1 έως 3000 Gb), την ταχύτητα εκτέλεσης αλληλούχισης (4 έως 56 ώρες) καθώς και το χαμηλότερο κόστος εφαρμογής στην αγορά (Hu et al., 2021). Ταυτόχρονα, η χρήση βιοπληροφορικών εργαλείων για την επεξεργασία των παραγόμενων δεδομένων τους είναι εύκολα προσβάσιμη μέσω του διαδικτύου. Το βασικότερο πλεονέκτημα της τεχνολογίας Illumina είναι το χαμηλό ποσοστό σφάλματος υποκατάστασης (0.1%), το οποίο την καθιστά την πιο αξιόπιστη τεχνολογία αλληλούχισης ανά βάση στην αγορά. Ένα από τα κύρια μειονεκτήματα των πλατφόρμων είναι ο αυστηρός έλεγχος φόρτωσης των δειγμάτων, επειδή η υπερφόρτωση μπορεί να οδηγήσει σε επικαλυπτόμενες συστάδες και κατά επέκταση στην κακή ποιότητα δεδομένων αλληλούχισης

1.2.4 Βιοπληροφορική

Τα πλεονεκτήματα των νέων τεχνολογιών έγιναν εμφανή εξ' αρχής, καθώς οι πολύ υψηλές αποδόσεις τους συνδυάζονταν με σημαντική μείωση του κόστους κάθε πειράματος. Παρ' όλα αυτά, η εξέλιξη των τεχνικών αλληλούχισης σε επίπεδο πειραματικής διαδικασίας, δεν σήμαινε την βελτιστοποίηση εξ ολοκλήρου της αναλυτικής μεθόδου (Y. Li & Chen, 2014). Τα μηχανήματα αλληλούχισης δεύτερης γενιάς παρήγαγαν δεδομένα που περιελάμβαναν αναγνωσμένες αλληλουχίες μικρότερου μεγέθους σε σχέση με τις κλασσικές τεχνικές. Οι αναγνωσμένες αλληλουχίες των ~1000 βάσεων που προέκυπταν από την αλληλούχιση Sanger είχαν πλέον αντικατασταθεί από αντίστοιχες αλληλουχίες μικρότερες των κατά μέσο όρο 100 βάσεων. Ταυτόχρονα, η υπερκάλυψη του μειονεκτήματος αυτού, μέσω της υψηλής απόδοσης των συσκευών αλληλούχισης νέας γενιάς, οδήγησε στην αύξηση του πλήθους των αναγνωσμένων αλληλουχιών κατά πολλές τάξεις μεγέθους με τον τελικό

τους αριθμό να αγγίζει πλέον δεκάδες εκατομμύρια ανά πείραμα. Αυτό είχε ως αποτέλεσμα τα δεδομένα αλληλούχισης που προκύπτουν να είναι πολύ μεγαλύτερου όγκου και πολυπλοκότητας, καθιστώντας έτσι την ανάλυσή τους ένα εξαιρετικά δυσεπίλυτο πρόβλημα.

Ο διεπιστημονικός τομέας που αναπτύσσει μεθόδους και εργαλεία λογισμικού για την διαχείριση και την κατανόηση βιολογικών δεδομένων, ιδίως όταν τα σύνολα δεδομένων είναι μεγάλα και πολύπλοκα, είναι η βιοπληροφορική (Gauthier et al., 2019). Ο τομέας αυτός χρησιμοποιείται για τις *in silico*[1] αναλύσεις βιολογικών ερωτημάτων και περιλαμβάνει την απόκτηση, αποθήκευση, επεξεργασία και εικονοποίηση βιολογικών δεδομένων με τη χρήση υπολογιστικών και στατιστικών τεχνικών (Mirzaei, 2020). Στην μεταγονιδιωματική, τα δύο πεδία βιοπληροφορικής που εφαρμόζονται είναι η συγκριτική και η λειτουργική βιοπληροφορική (Alves et al., 2018). Η συγκριτική βιοπληροφορική δίνει έμφαση στις πληροφορίες που μπορούν να συλλεχθούν συγκρίνοντας τα μεταγονιδιώματα μεταξύ τους και αξιοποιώντας τις εξελικτικές πληροφορίες. Η λειτουργική βιοπληροφορική χρησιμοποιεί πληροφορίες για να βγάλει συμπεράσματα σχετικά με τη λειτουργία των συστατικών ενός δεδομένου μεταγονιδιώματος. Γνωρίζοντας της δύο μεθόδους μεταγονιδιωματικής αλληλούχισης (shotgun και amplicon), αυτά τα δύο πεδία μπορεί είναι συμπληρωματικά και σε κάποιο βαθμό να αλληλοκαλύπτονται, επομένως η μεταξύ τους διάκριση δεν είναι πολύ αυστηρή.

Ο κορμός της βιοπληροφορικής ανάλυσης δεδομένων που έχουν προκύψει από shotgun ή amplicon μεταγονιδιωματική αλληλούχιση καθιερώθηκε την προηγούμενη δεκαετία (Liu et al., 2021). Μακροσκοπικά, οι ροές εργασίας βιοπληροφορικής ανάλυσης και των δύο μεθόδων έχουν κοινή λογική. Αρχικά, τα δεδομένα αλληλούχισης υπόκεινται σε μία διαδικασία προεπεξεργασίας, στην οποία τα δεδομένα οργανώνονται και «καθαρίζονται» από αλληλουχίες χαμηλής ποιότητας. Στην συνέχεια, τα αναγνώσματα επεξεργάζονται κατάλληλα με σκοπό να προκύψουν νέα δεδομένα, των οποίων το περιεχόμενο τους αντιπροσωπεύουν όσο πιο κοντά γίνεται την αρχική μεταγονιδιωματική μορφής του. Έπειτα, ακολουθεί η έκφραση των δεδομένων σε βιολογικές πληροφορίες, όπως τον προσδιορισμό της φυλογενετικής και λειτουργικής σημασίας τους, χρησιμοποιώντας κυρίως ενημερωμένες βάσεις δεδομένων. Τέλος, εφαρμόζονται στατιστικές αναλύσεις, με σκοπό την εις βάθος διερεύνηση της συνολικής εικόνας των δεδομένων.

Όμως, σε μικροσκοπικό επίπεδο, οι βιοπληροφορικές αναλύσεις των δύο μεθόδων διαφέρουν σημαντικά, ειδικά στο στάδιο επεξεργασίας που αφορά την προσπάθεια της μετάφρασης των δεδομένων στην αρχική βιολογική του πληροφορία (Oulas et al., 2015). Συγκεκριμένα, στην ανάλυση αμπλικονίων, η μετατροπή των καθαρών δεδομένων σε βιολογικές πληροφορίες απαιτεί την επιλογή των αναγνωσμάτων ως αντιπροσωπευτικές αλληλουχίες ενός είδους. Αντίστοιχα, στην shotgun μεταγονιδιωματική απαιτείται η διερεύνηση των φυλογενετικών δεικτών στα δεδομένα καθώς και η συναρμολόγηση των καθαρών αναγνωσμάτων σε μεγάλα, πολύπλοκα σύνολα δεδομένων μεταγονιδιώματος, ονομαζόμενα ως contigs. Σε σύγκριση με αυτή του αμπλικονίου, η shotgun μεταγονιδιωματική ανάλυση μπορεί να παρέχει άμεσα λειτουργικά προφίλ γονιδίων και να φτάσει σε πολύ μεγαλύτερης ευκρίνειας ταξινομική ανάλυση. Ωστόσο, λόγω του μεγάλου όγκου δεδομένων απαιτείται μεγάλος όγκος υπολογιστικών πόρων για την εκτέλεση της ανάλυσης.

[1] Πείραμα που εκτελείται σε υπολογιστή ή μέσω υπολογιστικής προσομοίωσης

[2] Το S αναφέρεται σε μονάδες Svedberg, ένας συντελεστής καθίζησης που παρέχει ένα μέτρο του

Συνεπώς, ανεξαρτήτως της μεθόδου, η μεταγονιδιωματική ανάλυση απαιτεί την διαδοχική επεξεργασία των δεδομένων κατά την οποία πρέπει να ληφθούν αποφάσεις για την επιλογή κατάλληλων βιοπληροφορικών εργαλείων. Η χρήση ενός συγκεκριμένου εργαλείου βασίζεται στον στόχο του εκάστοτε επεξεργαστικού βήματος και πάντα πρέπει να λαμβάνονται υπόψη οι δυνατότητες και οι ιδιαιτερότητες αυτού. Για την απλοποίηση της περιπλοκότητας του σχεδιασμού της υπολογιστικής διαδικασίας έχουν αναπτυχθεί αυτοματοποιημένες ροές διεργασίας (pipelines) (Liu et al., 2021). Σε μία τυπική μεταγονιδιωματική ανάλυση, χρησιμοποιούνται πολλά pipelines ενωμένα σε σειρά, δημιουργώντας έτσι μία συνολική ροή επεξεργαστικής ανάλυσης (workflow). Έτσι, η έξοδος ενός pipeline γίνεται είσοδος του επόμενου στην αλληλουχία επεξεργασίας των δεδομένων. Ως αποτέλεσμα αυτού, ο χρήστης ενός pipeline λαμβάνει τα τελικά αποτελέσματα έχοντας μόνο εισάγει τα αρχικά δεδομένα από τη διαδικασία αλληλούχισης και επιλέξει τις κατάλληλες παραμέτρους λειτουργίας των επί μέρους εργαλείων. Συνήθως, τα pipelines έχουν αναπτυχθεί σε ένα προγραμματιστικό περιβάλλον, όπως η R (Calle, 2019), αλλά έχουν δημιουργηθεί επίσης ειδικές πλατφόρμες μεταγονιδιωματικής ανάλυσης, όπως το Mothur (Schloss et al., 2009) και το QIIME 2 (Bolyen et al., 2019), στις οποίες έχουν συλλεχθεί και ενσωματωθεί διάφορα διαθέσιμα pipelines με σκοπό την απλοποίηση και την ταχύτερη ανάλυση δεδομένων.

Ωστόσο, οι μέθοδοι και τα πρότυπα βιοπληροφορικής ανάλυσης μικροβιώματος εξελίσσονται ραγδαία τα τελευταία χρόνια, προσφέροντας την δυνατότητα νέων ευρημάτων που αφορούν τις δομές και τις λειτουργίες μιας μικροβιακής κοινότητας. Παράλληλα, αυτές οι νέες εξελίξεις έχουν καταστήσει δύσκολη την επιλογή των κατάλληλων λογισμικών και pipelines για την επιστημονική κοινότητα, ειδικά για τα άτομα που δεν διαθέτουν το απαραίτητο υπόβαθρο βασικών βιοπληροφορικών γνώσεων. Συνεπώς, σε μία μεταγονιδωματική ανάλυση απαιτείται η καλή κατανόηση της προέλευσης και μορφολογίας των δεδομένων προς επεξεργασία, την έρευνα αναβαθμισμένων πρωτοκόλλων ανάλυσης τέτοιων δεδομένων, την μελέτη των δυνατοτήτων και τον τρόπο χρήσης των διαθέσιμων pipelines, την επιλογή των κατάλληλων pipelines και την μεταξύ τους σύνδεση, με αποτέλεσμα να οργανωθεί και κατασκευαστεί ένα υπολογιστικό workflow (Galloway-Peña & Hanson, 2020).

1.3 Αλληλούχιση του 16S rRNA Γονιδίου

1.3.1 Εισαγωγή

Μεταξύ των διάφορων μεθόδων ανάλυσης μικροβιώματος που βασίζονται στις τεχνολογίες αλληλούχισης νέας γενιας, η αλληλούχιση αμπλικονίων του 16S ριβοσωμικού RNA γονιδίου έχει αποδειχθεί μία αξιόπιστη και αποτελεσματική μέθοδος για τον προσδιορισμό της βακτηριακής σύνθεσης ενός μικροβιώματος (Matsuo et al., 2021). Το βακτηριακό γονίδιο 16S rRNA περιέχει εννέα μεταβλητές περιοχές (V1 έως V9) που διαχωρίζονται από εξαιρετικά διατηρημένες αλληλουχίες εντός των διαφορετικών ειδών βακτηρίων. Για την ταυτοποίηση των βακτηρίων, το γονίδιο 16S rRNA ενισχύεται πρώτα με αλυσιδωτή αντίδραση πολυμεράσης (PCR) με εκκινητές που προσδένονται στις διατηρημένες περιοχές και στη συνέχεια προσδιορίζεται η αλληλουχία τους μέσω της χρήσης της τεχνολογίας αλληλούχισης. Τα δεδομένα αλληλούχισης υποβάλλονται σε βιοπληροφορική ανάλυση, στην οποία χρησιμοποιούνται οι μεταβλητές περιοχές για τη διάκριση μεταξύ των βακτηριακών ταξινομήσεων. Στις παρακάτω ενότητες αναφέρονται αναλυτικά τα στάδια μιας

τέτοιας ανάλυσης, ξεκινώντας από την κατανόηση του ρόλου του γονιδίου 16S rRNA και τις φυλογενετικές του πληροφορίες καθώς και την αξιοποίηση αυτού στον σχεδιασμό μιας έρευνας μικροβιώματος.

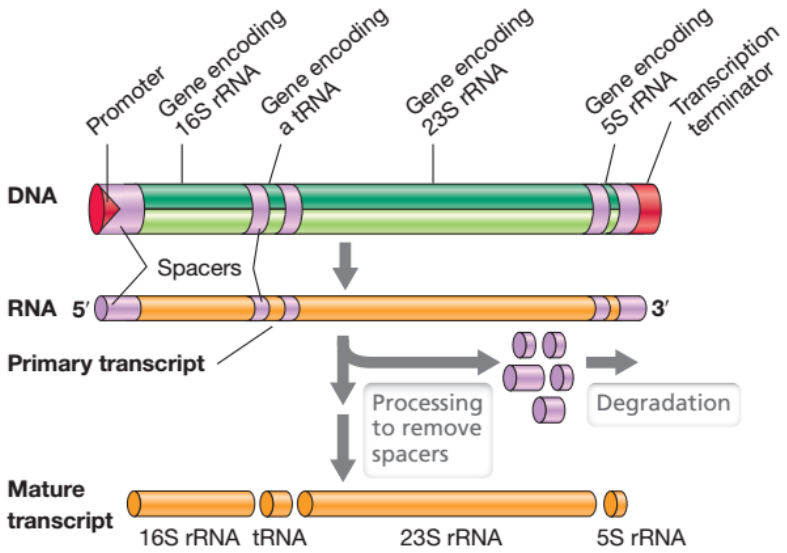
1.3.2 Το 16S rRNA γονίδιο και οι εφαρμογές/περιορισμοί της αλληλούχισής του

Τα ριβοσώματα είναι μεγάλα σύμπλοκα πρωτεϊνών και RNA στα οποία πραγματοποιείται η βιοσύνθεση των πρωτεϊνών του κυττάρου. Ένα κύτταρο μπορεί να έχει πολλές χιλιάδες ριβοσώματα και ο αριθμός τους αυξάνεται με υψηλότερους ρυθμούς ανάπτυξης. Το ριβόσωμα είναι μια εξαιρετικά δυναμική μοριακή δομή αποτελούμενο από δύο υπομονάδες, οι οποίες εναλλάσσονται και διαχωρίζονται κατά τη διάρκεια της μεταφραστικής διαδικασίας και παράλληλα αλληλεπιδρούν με πολλές άλλες πρωτεΐνες.

Τα ριβοσώματα των βακτηρίων και των αρχαίων αποτελούνται από τις ριβοσωμικές υπομονάδες 30S και 50S, οι οποίες παράγουν ανέπαφα τα ριβοσώματα 70S[2]. Κάθε ριβοσωμική υπομονάδα περιέχει συγκεκριμένα ριβοσωμικά RNA (rRNAs) και ριβοσωματικές πρωτεΐνες. Η υπομονάδα 30S περιέχει το 16S rRNA και 21 πρωτεΐνες και η υπομονάδα 50S περιέχει τα 5S και 23S rRNAs και 31 πρωτεΐνες. Στα βακτήρια, η γενετική πληροφορία που κωδικοποιεί τα rRNAs είναι οργανωμένη σε μεταγραφικές μονάδες, δηλαδή σε τμήματα DNA που το καθένα οριοθετείται από μία θέση έναρξης και μία θέση τερματισμού και μεταγράφεται σε ένα RNA. Σε μία ενιαία μεταγραφική μονάδα περιλαμβάνονται τα γονίδια που μεταγράφουν τα 16S, 23S και 5S rRNA καθώς και ένα tRNA (**Σχήμα 1.11**). Μετά την διαδικασία της μεταγραφής, το ενιαίο τμήμα RNA επεξεργάζεται από ειδικές πρωτεΐνες, οι οποίες τεμαχίζουν το τμήμα για να σχηματίσουν τα μεμονωμένα rRNA και tRNA. Στο γονιδίωμα των βακτηρίων, ο αριθμός των μεταγραφικών μονάδων που κωδικοποιούν τα rRNA του κυττάρου κυμαίνεται από 5 έως 10. Παραδείγματος χάρη, η *Escherichia coli* περιέχει επτά μεταγραφικές μονάδες rRNA (Madigan et al., 2021).

Όσον αφορά για το 16S rRNA, οι βιολογικές του λειτουργίες παίζουν πολύ σημαντικό ρόλο στην δομή του ριβοσώματος των βακτηρίων και στην διαδικασία της πρωτεϊνοσύνθεσης (Church et al., 2020). Δομικά, το 16S έχει το ρόλο της σκαλωσιάς στην υπομονάδα 30S του ριβοσώματος, ακινητοποιώντας έτσι τις ριβοσωμικές πρωτεΐνες που εμπεριέχονται σε αυτήν. Επιπλέον, το 16S αλληλεπιδρά με το 23S rRNA, συμβάλλοντας στην ενοποίηση των δύο υπομονάδων 50S και 30S του ριβοσώματος. Ένας ακόμα βασικός ρόλος του 16S rRNA είναι η χρήση του 3' άκρου του για την σύνδεση του με το κωδικόνιο έναρξης AUG του mRNA, με αποτέλεσμα να σχετίζεται με την έναρξη της πρωτεϊνικής σύνθεσης.

[2] Το S αναφέρεται σε μονάδες Svedberg, ένας συντελεστής καθίζησης που παρέχει ένα μέτρο του μεγέθους των σωματιδίων με βάση τον ρυθμό καθίζησής τους.

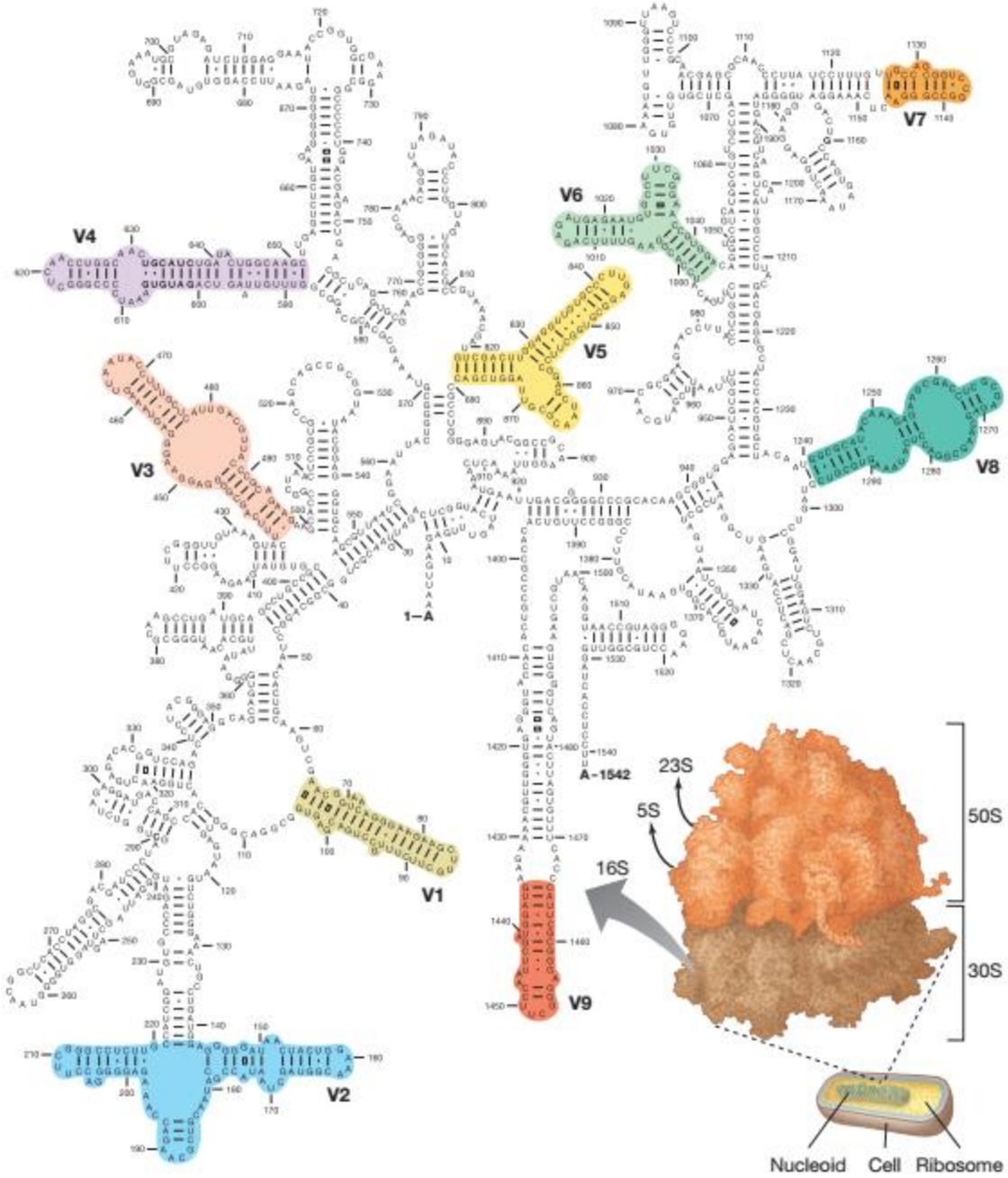


Σχήμα 1.11 Μια μεταγραφική μονάδα ριβοσωμικού RNA από βακτήρια και η επακόλουθη επεξεργασία του. Στα βακτήρια, όλες οι μεταγραφικές μονάδες rRNA έχουν τα γονίδια των 16S rRNA, 23S rRNA και 5S rRNA σε σειρά. Στη συγκεκριμένη μεταγραφική μονάδα, ο διαχωρισμός μεταξύ των γονιδίων rRNA 16S και 23S περιέχει ένα γονίδιο tRNA. Σε άλλες μεταγραφικές μονάδες αυτή η περιοχή μπορεί να περιέχει περισσότερα από ένα γονίδια tRNA. Συχνά ένα ή περισσότερα γονίδια tRNA ακολουθούν επίσης το γονίδιο 5S rRNA και μεταγράφονται. (Madigan et al., 2021)

Λόγω της λειτουργικής σταθερότητάς του 16S rRNA, το γονίδιο που το κωδικοποιεί έχει συντηρηθεί κατά ένα μεγάλο βαθμό στην εξελικτική πορεία των βακτηρίων. Έτσι, το 1985 προτάθηκε για πρώτη φορά η συστηματική σύγκριση των αλληλουχιών του 16S rRNA για προκαρυώτες, καθώς και του 18S rRNA για ευκαρυώτες, από διαφορετικούς μικροοργανισμούς με σκοπό την εκτίμηση των εξελικτικών σχέσεων μεταξύ των μικροοργανισμών και την κατασκευή των φυλογενετικών τους δέντρων (D. J. Lane et al., 1985). Πράγματι, έχει αποδειχτεί ότι οι αλληλουχίες γονιδίου 16S rRNA περιέχουν υπερμεταβλητές περιοχές που μπορούν να παρέχουν ειδικές χαρακτηριστικές αλληλουχίες για κάθε είδος, με αποτέλεσμα να φανούν χρήσιμες για την ταυτοποίηση βακτηρίων και κατά επέκταση την εξελικτική του πορεία. Κατά μήκος του γονιδίου του 16S, που είναι κατά μέσο όρο 1500 bp, περιλαμβάνονται 9 υπερμεταβλητές περιοχές (V1-V9), που κυμαίνονται από περίπου 30 έως 100 bp (**Σχήμα 1.12**) (Vargas-Albores et al., 2017). Ο βαθμός διατήρησης ποικίλλει ευρέως μεταξύ υπερμεταβλητών περιοχών, με τις πιο διατηρημένες περιοχές να συσχετίζονται με ταξινομικά υψηλότερα επίπεδα και τις λιγότερο διατηρημένες περιοχές σε χαμηλότερα επίπεδα, όπως το γένος και το είδος (B. Yang et al., 2016).

Συνεπώς, το γονίδιο 16S rRNA παρέχει τρία βασικά χαρακτηριστικά: (i) Το γονίδιο υπάρχει σε όλα τα βακτήρια. (ii) Το γονίδιο έχει διατηρηθεί επαρκώς ώστε να είναι εύκολα αναγνωρίσιμο, ενώ ταυτόχρονα είναι αρκετά διαφορετικό ώστε να επιτρέπει την αναγνώριση διαφορετικών βακτηριακών ειδών (iii) Τέλος, το γονίδιο αποτελείται από συνυφασμένα συντηρημένα και μεταβλητά πεδία. Αυτά τα χαρακτηριστικά καθιστούν το γονίδιο έναν ισχυρό φυλογενετικό δείκτη για την ταξινομική ανάλυση βακτηρίων (Gołębiewski & Tretyn, 2020). Επίσης, λόγω των συνυφασμένων συντηρημένων περιοχών, επιτρέπεται ο σχεδιασμός καθολικών εκκινητών για την εφαρμογή PCR, και κατά επέκταση η μεταγονιδιωματικη αλληλούχιση αμπλικονίου του γονιδίου αυτού για την μελέτη του βακτηριακού πληθυσμού ενός εξεταζόμενου δείγματος (Matsuo et al., 2021). Δεδομένου ότι ο μεγαλύτερος μικροβιακός πληθυσμός που αποικίζει στο ανθρώπινο σώμα είναι αυτός των βακτηρίων, η εφαρμογή της αλληλούχισης του 16S rRNA γονιδίου έχει αποτελέσει την πιο διαδεδομένη

μέθοδο ανάλυσης μικροβιωμάτων που προέρχονται από διάφορα σημεία του σώματος (Ames et al., 2017).



Σχήμα 1.12 Ριβοσωμικό RNA (rRNA). Πρωτογενής και δευτερογενής δομή του 16S rRNA από Escherichia coli. Στα βακτήρια, το μόριο αποτελείται από διατηρημένες και μεταβλητές περιοχές (V1–V9). Οι κατά προσέγγιση θέσεις των μεταβλητών περιοχών υποδεικνύονται με χρώμα. Η δομή του ριβοσώματος των βακτηρίων 70S αποτελείται από υπομονάδες 30S και 50S. Το 16S rRNA είναι μέρος της υπομονάδας 30S ενώ τα 5S και 23S rRNA είναι μέρη της υπομονάδας 50S. (Madigan et al., 2021)

Κατά τη διεξαγωγή μιας μελέτης προσδιορισμού της αλληλουχίας του γονιδίου 16S rRNA, μία ή περισσότερες υπερμεταβλητές περιοχές ενισχύονται με την μέθοδο PCR χρησιμοποιώντας εκκινητές ευρείας περιοχής που ο καθένας συνδέεται σε μια διατηρημένη περιοχή. Στη συνέχεια, τα παραγόμενα αμπλικόνια εισάγονται σε μια πλατφόρμα αλληλούχισης για την καταγραφή των αλληλουχιών αυτών των περιοχών. Οι πληροφορίες σε αυτές τις περιοχές χρησιμοποιούνται για τη μελέτη της ταξινομικής σύνθεσης και

ποικιλομορφίας που υπάρχει στο δείγμα. Ανάλογα με την εφαρμογή, μερικές φορές μπορούν να γίνουν κατάλληλες ταξινομικές αναλύσεις από μεταβλητά θραύσματα του 16S rRNA τόσο μικρά όσο 100 bp, καθιστώντας τις δημοφιλείς και προσιτές πλατφόρμες σύντομης αλληλούχισης (π.χ. Illumina) κατάλληλες για ανάλυση μικροβιώματος. Ενώ είναι δυνατός ο προσδιορισμός της αλληλουχίας ολόκληρου του γονιδίου 16S rRNA, αυτό απαιτεί μεγαλύτερες χρονικές και κοστολογικές επενδύσεις, γεγονός που μπορεί να υπονομεύσει τα πλεονεκτήματα αυτής της προσέγγισης. Όμως, ο προσδιορισμός αλληλουχίας μεταξύ των ολόκληρων γονιδίων rRNA, έχει αποδειχθεί χρήσιμος για τον ταυτοποίηση βακτηριακών ειδών και στελεχών (Johnson et al., 2019).

Όταν επιλέγεται η αλληλούχιση συγκεκριμένων περιοχών του γονιδίου 16S rRNA για την ανάλυση ενός μικροβιώματος, είναι αναγκαία η αναφορά και η κατανόηση των περιορισμών αυτής της προσέγγισης. Αρχικά, αυτή η μέθοδος δεν μπορεί να περιγράψει το μεταβολικό δυναμικό, τη δραστηριότητα μιας μικροβιακής κοινότητας καθώς και το αν τα βακτήρια είναι ζωντανά η νεκρά (Weinroth et al., 2022). Αυτό οφείλεται στο γεγονός ότι το γονίδιο 16S rRNA είναι ένα γονίδιο που βρίσκεται σε όλα τα προκαρυωτικά και οποιαδήποτε μεταβλητότητα αλληλουχίας υποδηλώνει μόνο φυλογενετική απόκλιση, η οποία δεν συσχετίζεται αναγκαστικά με τα μεταβολικά χαρακτηριστικά των βακτηρίων, την επιρροή τους σε νοσήματα και την αντίσταση τους στα αντιβιοτικά. Επίσης, η παρουσία του γονιδίου δεν υποδηλώνει την κατάσταση στην οποία βρίσκεται ο μικροοργανισμός, διότι ακόμα και μετά τον θάνατό του υπάρχει περίπτωση το γονιδίωμά του να έχει παραμείνει άθικτο. Επιπλέον, η αλληλούχιση μίας συγκεκριμένης περιοχής του 16S rRNA μπορεί να μην περιέχει επαρκή μεταβλητότητα στις αλληλουχίες της για τη διάκριση ειδών ή στελεχών (Johnson et al., 2019).

Ένας περαιτέρω περιορισμός αυτής της προσέγγισης είναι ο αυθαίρετος αριθμός αμπλικονίων που προσδιορίζεται από την αλληλούχιση, με αποτέλεσμα να μην ανακλάται το συνολικό απόλυτο μικροβιακό φορτίο ή η απόλυτη αφθονία (Gloor et al., 2017). Η μέτρηση του συνολικού βακτηριακού φορτίου θα απαιτούσε ποσοτικές μεθόδους, όπως η ποσοτική αλυσιδωτή αντίδραση πολυμεράσης (qPCR), και δεν μπορεί να ληφθεί μόνο με την αλληλούχιση του γονιδίου 16S rRNA. Όμως, τα δεδομένα που παράγονται μπορούν να παρέχουν χρήσιμες πληροφορίες σχετικά με τα παρόντα μικρόβια και την σχετική αφθονία τους. Αυτό επιτρέπει την εκτίμηση της ποικιλομορφίας ενός μόνο δείγματος, αλλά και τη σύγκριση ομοιοτήτων και διαφορών μεταξύ πολλών δειγμάτων.

Τέλος, ένα γνωστό μειονέκτημα αυτής της προσέγγισης είναι ότι οι διαφορετικοί μικροοργανισμοί έχουν διαφορετικούς αριθμούς αντιγράφων του γονιδίου 16S rRNA στο γονιδίωμά τους, επομένως η σχετική αφθονία επηρεάζεται από αυτούς που έχουν υψηλότερους αριθμούς αντιγράφων των γονιδίων 16S rRNA στο γονιδίωμά τους. Ενώ αυτός είναι ένας γνωστός περιορισμός, τα εργαλεία για τη διόρθωση αυτού του ζητήματος δεν έχουν παράγει συνεπή αποτελέσματα και η κοινή πρακτική είναι να μην προσαρμόζεται ο αριθμός αντιγράφων του γονιδίου 16S rRNA ανά βακτήριο (Louca et al., 2018).

1.3.3 Σχεδιασμός έρευνας, δειγματοληψία και συλλογή μεταδεδομένων

Όταν πρόκειται για την ανάλυση μικροβιωμάτων, ένας σχολαστικός σχεδιασμός της έρευνας είναι απαραίτητος για την απόκτηση εγκύρων και ουσιαστικών αποτελεσμάτων. Αν και δεν υπάρχει μία τελική άποψη σχετικά με το ποιες είναι οι βέλτιστες πρακτικές μιας έρευνας μικροβιώματος (Berg et al., 2020), υπάρχουν κάποιες ερευνητικές προσεγγίσεις που

χρησιμοποιούνται για την κλινική έρευνα μικροβιωμάτων. Οι πιο δημοφιλείς προσεγγίσεις περιλαμβάνουν τις συγχρονικές έρευνες (cross-sectional studies), τις έρευνες κοόρτης (case-control studies), τις διαχρονικές έρευνες (longitudinal studies) και τις τυχαιοποιημένες δοκιμές ελέγχου (randomized controlled trials, RCT) (**Πίνακας 1.1**) (Qian et al., 2020). Φυσικά, η επιλογή της ερευνητικής προσέγγισης, ή ακόμα και ο συνδυασμός αυτών, εξαρτάται αποκλειστικά από το ερευνητικό ερώτημα που επιθυμείται να λυθεί.

Πίνακας 1.1 Οι πιο δημοφιλείς ερευνητικές προσεγγίσεις ανάλυσης μικροβιώματος και η κύρια μεθοδολογία τους.

Τύπος Ερευνητικής Προσέγγισης	**Μεθοδολογία**
Συγχρονική έρευνα	Ανάλυση μικροβιώματος ενός πληθυσμού ή αντιπροσωπευτικού υποσυνόλου σε μία συγκεκριμένη χρονική στιγμή
Έρευνα κοόρτης	Ανάλυση και σύγκριση μικροβιωμάτων από δύο ομάδες ανθρώπων που διαφέρουν ως προς μία έκβαση, η οποία συνήθως είναι η παρουσία ή απουσία μίας νόσου ή πάθησης
Διαχρονική έρευνα	Επαναλαμβανόμενη ανάλυση μικροβιώματος ενός πληθυσμού ή αντιπροσωπευτικού υποσυνόλου σε σύντομες ή μεγάλες χρονικές περιόδου
Τυχαιοποιημένη δοκιμή ελέγχου	Ανάλυση μικροβιώματος ενός πληθυσμού για την αξιολόγηση της αποτελεσματικότητας μίας συγκεκριμένης παρέμβασης

Αρχικά, λαμβάνοντας υπόψη το είδος του μικροβιώματος προς ανάλυση, την ομάδα ή τις πολλαπλές ομάδες ανθρώπων και την επιθυμητή συσχέτιση αυτών των δύο αντικειμένων, το πρώτο βήμα της έρευνας είναι η συλλογή των συμμετεχόντων. Ο καθορισμός συγκεκριμένων κριτηρίων ένταξης και αποκλεισμού (inclusive and exclusive criteria) επιτρέπει την καλύτερη αντιστοίχιση διαφορετικών ομάδων και περιορίζει παράγοντες που επηρεάζουν την κατάσταση του μικροβιώματος, όπως η ηλικία, το φύλο, το BMI, η διατροφή, η φαρμακευτική αγωγή, η εθνικότητα και η γεωγραφική περιοχή (Qian et al., 2020). Επιπλέον, όταν ένας ερευνητής σχεδιάζει ένα πείραμα, είναι σημαντικό να εκτιμήσει το μέγεθος του δείγματος, και κατά επέκταση τον αριθμό των εθελοντών. Ένα κατάλληλο μέγεθος δείγματος επιτρέπει σε μια έρευνα μικροβιώματος να διακρίνει τις διαφορές μεταξύ των ομάδων και να εξοικονομήσει πόρους και χρόνο.

Στην συνέχεια, ακολουθεί η διαδικασία της δειγματοληψίας, η συντήρηση των δειγμάτων και η αποθήκευση τους. Γενικά, η επιλογή του κατάλληλου σημείου από το οποίο θα πραγματοποιηθεί η δειγματοληψία στις μελέτες μικροβιώματος είναι ένα ουσιαστικό, και συνήθως πολύπλοκο, βήμα της έρευνας και του πειραματικού σχεδιασμού (Weinroth et al., 2022). Ο προσδιορισμός του σημείου δειγματοληψίας πρέπει να βασίζεται στην υπόθεση ή/και στους στόχους της έρευνας, καθώς και στη γνώση σχετικά με τη φυσιολογία του ξενιστή ή/και του ιστού-στόχου. Ταυτόχρονα, πρέπει να ληφθεί υπόψη και ο χρόνος στον οποίο πραγματοποιείται η δειγματοληψία, ειδικά όταν πρόκειται για τον σχεδιασμό μίας διαχρονικής έρευνας. Η συλλογή των δειγμάτων απαιτεί τη διατήρηση της στειρότητας, επομένως, όλα τα δοχεία και τα εργαλεία θα πρέπει να αποστειρώνονται πριν από τη χρήση ή να είναι αποστειρωμένα και μιας χρήσης (Gołębiewski & Tretyn, 2020).

Φυσικό επακόλουθο της δειγματοληψίας είναι η διαδικασία συντήρησης και αποθήκευσης των δειγμάτων, εφόσον η συλλογή δειγμάτων σχεδόν ποτέ δεν γίνεται ταυτόχρονα από όλους τους συμμετέχοντες. Οι μέθοδοι διατήρησης και αποθήκευσης δειγμάτων συνήθως προσαρμόζονται στην πειραματική μέθοδο και τον τύπο του δείγματος, ενώ η πιο ευέλικτη μέθοδος είναι η απευθείας κατάψυξη των δειγμάτων (Qian et al., 2020).

Ωστόσο, πολλές φορές τα δείγματα συλλέγονται εκτός κλινικών/ερευνητικών εγκαταστάσεων (π.χ. στην κατοικία των συμμετεχόντων) όπου υπό αυτές τις συνθήκες, η χρήση ειδικών κιτ συντήρησης είναι μια εναλλακτική λύση. Είναι πολύ σημαντικό οι μέθοδοι διατήρησης και αποθήκευσης των δειγμάτων να είναι συνεπείς σε όλα τα δείγματα για να ελαχιστοποιηθούν οι πιθανές παραλλαγές που θα μπορούσαν να προκαλέσουν εσφαλμένα αποτελέσματα.

Ένα από τα κρίσιμα βήματα του αρχικού σταδίου μίας έρευνας μικροβιώματος είναι η συλλογή και η καταγραφή μεταδεδομένων. Ο όρος «μεταδεδομένα» (metadata) χρησιμοποιείται για να υποδείξει όλα τα περιγραφικά δεδομένα που χαρακτηρίζουν τα βιολογικά δείγματα που συλλέγονται (Weinroth et al., 2022). Από την αρχή έως και την ολοκλήρωση της έρευνας, είναι πολύ σημαντικό να συλλεχθούν όσο το δυνατόν περισσότερες πληροφορίες και λεπτομέρειες που περιγράφουν τα δείγματα. Οι πληροφορίες συμπεριλαμβάνουν χαρακτηριστικά που αφορούν την προέλευση και τον τρόπο συλλογής των δειγμάτων, την προετοιμασία αυτών, καθώς και άλλες παρατηρήσεις που έγιναν σε όλη τη μελέτη. Στην ιδανική περίπτωση, τα μεταδεδομένα θα πρέπει να παρέχουν όλες τις απαραίτητες πληροφορίες για την επανάληψη μιας έρευνας ή για την εκ νέου δειγματοληψία σε μεταγενέστερα στάδια, αφενός, και θα πρέπει να επιτρέπουν στους ερευνητές να επαναχρησιμοποιούν τα δεδομένα σε ένα ευρύτερο πλαίσιο αφετέρου (Cernava et al., 2022). Μόλις συλλεχθούν τα μεταδεδομένα, οι πληροφορίες θα πρέπει να αποθηκεύονται κατάλληλα σε ψηφιακή μορφή.

Ο λόγος για τον οποίο η συλλογή έγκυρων μεταδεδομένων αποτελεί σημαντικό αρχικό στάδιο της έρευνας είναι η βελτίωση των μεταγενέστερων βιοπληροφορικών αναλύσεων. Οι αναλύσεις αυτές περιλαμβάνουν την χρήση των βιοπληροφορικών εργαλειών/pipelines για την επεξεργασία των δεδομένων αλληλούχισης, τον προσδιορισμό της ποικιλομορφίας των δειγμάτων, την ταξινομική ανάθεση καθώς και τις στατιστικές αναλύσεις. Συγκεκριμένα, η χρήση βιοπληροφορικών εργαλείων απαιτεί την συνέπεια των πληροφοριών που συμπεριλαμβάνονται στα μεταδεδομένα, καθώς και την μορφή αυτών, για την κατάλληλη εισαγωγή δεδομένων, καθιστώντας έτσι τα ακριβή μεταδεδομένα σημαντικά για τις αναλύσεις. Επιπλέον, τα pipelines ενσωματώνουν συχνά στατιστικές αναλύσεις, γεγονός που καθιστά την οργάνωση μεταδεδομένων αναπόσπαστο βήμα για την πραγματοποίηση πολλαπλών συγκρίσεων μεταξύ των διαφόρων μεταβλητών. Συνεπώς, τα ακριβή μεταδεδομένα μπορούν να βελτιώσουν τις παρατηρήσεις των δεδομένων ή συγκεκριμένες ομάδες αυτών (Weinroth et al., 2022).

1.3.4 Απομόνωση DNA, επιλογή υπερμεταβλητής περιοχής και ενίσχυση με PCR

Για να ληφθεί αντιπροσωπευτικό και ποιοτικό μεταγονιδίωμα για την ενίσχυση του γονιδίου 16S rRNA και στη συνέχεια για την αλληλούχισή του, είναι αναγκαία η αποτελεσματική διάσπαση των κυτταρικών τοιχωμάτων ή μεμβρανών όλων των βακτηρίων που εμπεριέχονται στα εξεταζόμενα δείγματα. Αυτό επιτυγχάνεται με την διαδικασία της κυτταρικής λύσης, κατά την οποία το κυτταρικό τοίχωμα και/ή η μεμβράνη διασπάται ή καταστρέφεται για την απελευθέρωση του εσωτερικού κυτταρικού υλικού, όπως το DNA, το RNA, οι πρωτεΐνες ή τα οργανίδια του κυττάρου. Επί του παρόντος, υπάρχουν πολλές μεθόδοι κυτταρικής λύσης, συμπεριλαμβανομένης της μηχανικής, φυσικής, χημικής και ενζυμικής (Shehadul Islam et al., 2017). Για την επιλογή της βέλτιστης μεθόδου κυτταρικής λύσης, πρέπει να ληφθεί υπόψη το κατά πόσο θα είναι αποτελεσματική σε όλους τους τύπους κυττάρων εντός του δείγματος αλλά και το κατά πόσο είναι εφαρμόσιμη σε όλους τους

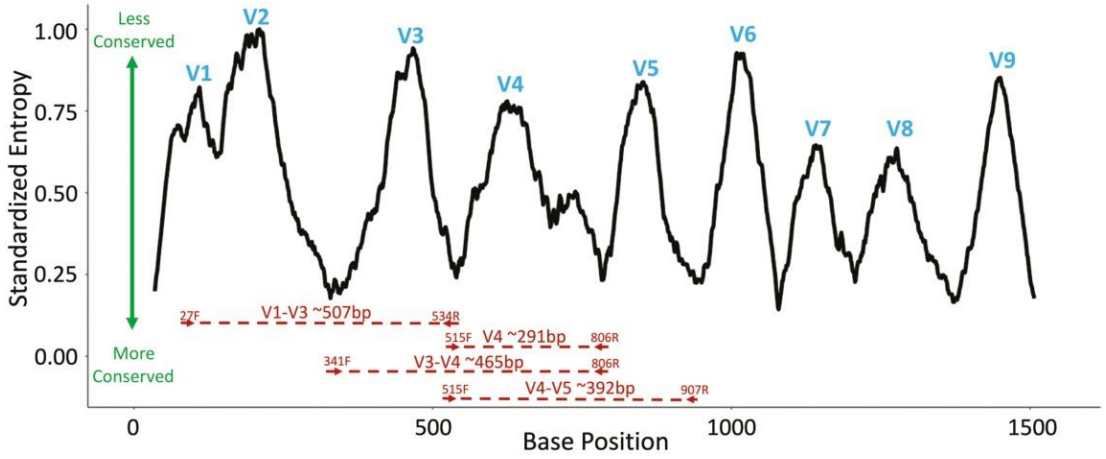
τύπους δειγμάτων της τρέχουσας έρευνας. Επειδή ορισμένα κύτταρα είναι πιο ανθεκτικά στην κυτταρική λύση, όπως τα θετικά κατά Gram βακτήρια και τα βακτηριακά ενδοσπόρια, μία επιλεγμένη τεχνική λύσης θα μπορούσε να οδηγήσει σε απομονωμένο μεταγονιδιωματικό DNA που δεν αντιπροσωπεύει την πραγματική ποικιλομορφία ή σύνθεση του μικροβιώματος. Ως εκ τούτου, η μηχανική λύση έχει αποδειχθεί ότι βελτιώνει την απόδοση απομόνωσης DNA, αυξάνει την παρατηρούμενη βακτηριακή ποικιλομορφία και βελτιώνει την ανάκτηση του μεταγονιδιωματικού DNA από θετικά κατά Gram βακτήρια και ενδοσπόρια. Επιπλέον, για να υπάρχει αντικειμενική σύγκριση διάφορων τύπων δειγμάτων, είναι σημαντικό να καθοριστεί μία μεθοδολογία κυτταρικής λύσης, είτε πρόκειται για μία μόνο μέθοδο είτε για τον συνδυασμό αυτών, η οποία να είναι αποτελεσματική σε όλους τους τύπους δειγμάτων. Με την ολοκλήρωση της διαδικασίας της κυτταρικής λύσης, χρησιμοποιούνται ειδικά κιτ απομόνωσης νουκλεϊκών οξέων για τον διαχωρισμό του DNA από τα υπόλοιπα συστατικά των κυττάρων.

Το επόμενο βήμα της έρευνας αποτελεί την ενίσχυση μίας συγκεκριμένης περιοχής του γονιδίου 16S rRNA με την μέθοδο PCR. Σε αυτό το στάδιο, πρέπει να παρθεί η απόφαση για το ποια ή ποιες υπερμεταβλητές περιοχές από τις συνολικά εννιά θα χρησιμοποιηθούν για την ταξινομική ανάθεση των βακτηρίων. Η επιλογή μιας συγκεκριμένης περιοχής εξαρτάται από δύο βασικούς παράγοντες, το βάθος της ταξινομικής ανάλυσης και την διαθέσιμη οικονομική υποστήριξη. Κατά κανόνα, οι πιο συντηρητικές περιοχές είναι χρήσιμες για τον προσδιορισμό των ταξινομικών κατηγοριών υψηλότερου επιπέδου, ενώ οι πιο μεταβαλλόμενες μπορούν να βοηθήσουν στην αναγνώριση γένους ή είδους (Bukin et al., 2019). Επιπλέον, η επιλογή ενός συνδυασμού υπερμεταβλητών περιοχών, δηλαδή η τάση ενίσχυσης μεγαλύτερου μήκους του γονιδίου, επιτρέπει επίσης την καλύτερη και πιο ποιοτική ταξινομική ανάλυση (Fuks et al., 2018). Βέβαια, η αύξηση του μήκους του αμπλικονίου που επιθυμείται να αλληλουχιθεί συνεπάγεται με την αύξηση του κόστους των αναλώσιμων και της πλατφόρμας αλληλούχισης προς χρήση.

Αξιοσημείωτο είναι το γεγονός ότι έρευνες έχουν δείξει την σημαντική αλλαγή της αντιληπτής δομής μιας μικροβιακής κοινότητας συναρτήσει μιας συγκεκριμένης στοχευόμενης υπερμεταβλητής περιοχής (Teng et al., 2018). Ως εκ τούτου, πρέπει να δοθεί βάση στην αξιολόγηση της επιλογής υπερμεταβλητών περιοχών, έτσι ώστε να ελαχιστοποιηθούν οι παραμορφώσεις και οι αντικρούσεις στην ανάλυση και σύγκριση μικροβιωμάτων. Πολλές φορές, η αξιολόγηση των βέλτιστων περιοχών είναι ασαφής. Σε αυτές τις περιπτώσεις προτείνεται η επιλογή μίας υπερμεταβλητής περιοχής που έχει χρησιμοποιηθεί σε παρόμοιες δημοσιευμένες έρευνες, προσφέροντας την δυνατότητα σύγκρισης των αποτελεσμάτων (Weinroth et al., 2022). Μέχρι και σήμερα, οι πιο συχνές επιλογές υπερμεταβλητών περιοχών είναι η περιοχή V1-V3, η V3-V4, η V4 και η V4-V5 (**Σχήμα 1.13**).

Εφόσον αποφασιστεί η περιοχή στην οποία θα πραγματοποιηθεί η ανάλυση, ακολουθεί ο σχεδιασμός της αλυσιδωτής αντίδρασης πολυμεράσης (PCR). Η εφαρμογή αυτής της αντίδρασης επιτρέπει την πολλαπλή αντιγραφή της ενδιαφερόμενης γενετικής περιοχής από όλο το μεταγονιδίωμα και την αποκλειστική αλληλούχισή της σε μεταγενέστερες αναλύσεις. Σε μία τέτοια αντίδραση, εκτός από το πρότυπο DNA (template), τα συστατικά που έχουν πρωταγωνιστικό ρόλο είναι το ζεύγος εκκινητών και η DNA πολυμεράση (Kadri, 2019). Όπως και στο πρώτο στάδιο της ροής της γενετικής πληροφορίας, έτσι και στην PCR ο ρόλος του ενζύμου DNA πολυμεράση είναι η σύνθεση νέων κλώνων DNA με βάση την συμπληρωματικότητα. Όμως, ακόμα και μετά τον διαχωρισμό του

δίκλωνου DNA που εφαρμόζεται στην αρχή της αντίδρασης, το ένζυμο είναι ανίκανο να ξεκινήσει από μόνο του την αντιγραφή. Επειδή η DNA πολυμεράση μπορεί να προσθέσει ένα νουκλεοτίδιο μόνο σε μια προϋπάρχουσα ομάδα 3'-OH, χρειάζεται έναν εκκινητή, δηλαδή ένα συνθετικό DNA ολιγονουκλεοτίδιο που είναι συμπληρωματικό με μία από τις αλυσίδες του δίκλωνου DNA, στον οποίο μπορεί να προσδεθεί και να προσθέσει το πρώτο νουκλεοτίδιο. Αυτή η απαίτηση καθιστά δυνατή την οριοθέτηση μιας συγκεκριμένης περιοχής αλληλουχίας του προτύπου DNA για την ενίσχυσή της.



Σχήμα 1.13 Απεικόνιση συντηρημένων και υπερμεταβλητών περιοχών του γονιδίου 16S rRNA. Οι υπερμεταβλητές περιοχές επισημαίνονται με μπλε χρώμα και οι διατηρημένες περιοχές υποδεικνύονται με χαμηλή εντροπία. Με κόκκινο χρώμα, επισημαίνονται τα τέσσερα συνηθέστερα χρησιμοποιούμενα ζεύγη εκκινητών. (Weinroth et al., 2022)

Συνεπώς, ο σχεδιασμός των κατάλληλων εκκινητών αποτελεί το κλειδί για την αντιγραφή των περιοχών προς ενδιαφέροντος του γονιδίου 16S rRNA και η ακρίβεια της αλληλούχισής τους εξαρτάται σε μεγάλο βαθμό από την επιλογή αυτών (Sambo et al., 2018). Κατά βάσει, χρησιμοποιώντας τις συντηρημένες περιοχές, σχεδιάζονται ζεύγη εκκινητών ευρέως φάσματος, οι οποίοι με τη σειρά τους μπορούν να χρησιμοποιηθούν για την απομόνωση των υπερμεταβλητών περιοχών. Ένα ζεύγος εκκινητών αποτελείται από έναν εμπρόσθιο (forward) και έναν ανάστροφο (reverse) εκκινητή, όπου ο πρώτος προορίζεται να ταιριάζει με την αρχή της κύριας αλληλουχίας του βακτηριακού 16S, ενώ ο δεύτερος πρέπει να ταιριάζει με την αντιπαράλληλη αλληλουχία. Δεδομένου ότι οι αλληλουχίες των συντηρημένων περιοχών του γονιδίου 16S δεν είναι πανομοιότυπες σε διαφορετικά είδη βακτηρίων, χρησιμοποιούνται εκφυλισμένοι εκκινητές. Ένα σύνολο εκκινητών ονομάζεται εκφυλισμένο όταν χρησιμοποιείται ως μείγμα ολιγονουκλεοτιδικών μορίων που περιέχουν διαφορετικά νουκλεοτίδια σε καθορισμένες θέσεις. Ένα ζεύγος εκφυλισμένων εκκινητών μπορεί φυσικά να επεκταθεί σε ένα σύνολο μη εκφυλισμένων ζευγών εκκινητών, των οποίων τα στοιχεία λαμβάνονται με την ανάθεση όλων των πιθανών συνδυασμών βάσεων στα εκφυλισμένα νουκλεοτίδια του αρχικού ζεύγους.

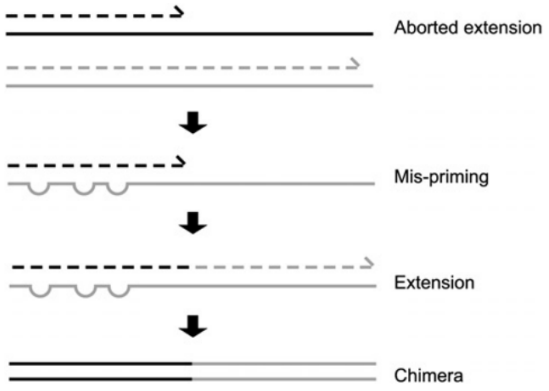
Η ανάπτυξη των ζευγών εκκινητών για την ενίσχυση των γονιδίων 16S rRNA διεξάγεται *in silico* χρησιμοποιώντας βάσεις δεδομένων ως σημεία αναφοράς. Τα βασικά χαρακτηριστικά ενός ζεύγους εκκινητών για την ενίσχυση της επιθυμητής περιοχής του γονιδίου 16S rRNA είναι: (i) η μέγιστη πειραματική αποτελεσματικότητα και αποκλειστικότητα, όσον αφορά την ενίσχυση της επιλεγμένης περιοχής αλληλουχίας και όχι άλλες, (ii) η μέγιστη κάλυψη, δηλαδή το ζεύγος εκκινητών να ταιριάζει με όλα τα θραύσματα όλων των βακτηριακών αλληλουχιών 16S από διαφορετικά είδη, και (iii) η ελάχιστη μεταβλητότητα, όσον αφορά τον αριθμού συνδυασμών που μπορούν να προσφέρουν οι

εκφυλισμένοι εκκινητές (Sambo et al., 2018). Το τελευταίο χαρακτηριστικό αναφέρεται στην ελαχιστοποίηση της πιθανότητας να ενισχυθούν περισσότερο κάποιες βακτηριακές αλληλουχίες σε σχέση με άλλες λόγω του ότι καλύπτονται από περισσότερους συνδυασμούς εκκινητών, παραποιώντας έτσι την πραγματική σχετική αφθονία των διαφορετικών ειδών.

Στο πέρασμα των προηγούμενων χρόνων, έχουν σχεδιαστεί πολλά ζεύγη εκκινητών για την μελέτη τις ποικιλομορφίας συγκεκριμένων ταξινομικών ομάδων και έχουν γίνει πολλές προσπάθειες να αναπτυχθούν «καθολικά» ζεύγη του γονιδίου 16S για την κάλυψη σχεδόν ολόκληρης της ποικιλομορφίας μίας φυσικής μικροβιακής κοινότητας (Fadeev et al., 2021). Ένα από τα πιο ευρέως χρησιμοποιούμενα ζεύγη εκκινητών για τη διερεύνηση της βακτηριακής ποικιλότητας σε διάφορα περιβάλλοντα είναι το 341F/785R, οι οποίοι στοχεύουν τις υπερμεταβλητές περιοχές V3-V4 του γονιδίου 16S rRNA. Ένα εναλλακτικό ζεύγος εκκινητών είναι το 515F-Y/926R που στοχεύουν τις υπερμεταβλητές περιοχές V4-V5, το οποίο είναι επίσης σε θέση να συλλάβει την ποικιλομορφία των αρχαίων κοινοτήτων.

Ακόμα και αν έχουν εφευρεθεί εξαιρετικές τεχνικές κυτταρικής λύσης, απομόνωσης DNA και τεχνικές PCR για την μεταγενέστερη αλληλούχιση του γονιδίου 16S, αυτές οι βασικές εφαρμογές ανάλυσης μικροβιώματος εμπεριέχουν πολλούς παράγοντες που μπορούν να εισάγουν σημαντικά σφάλματα στα τελικά αποτελέσματα (Dos Santos et al., 2019), ειδικά όταν τα αρχικά δείγματα έχουν χαμηλή βιομάζα. Η ανικανότητα της πλήρης απομόνωσης του μεταγονιδιώματος, οι επιμολύνσεις, η εξειδίκευση των εκκινητών, η παρέμβαση μη-στοχευμένου DNA και η προτιμώμενη ενίσχυση ορισμένων αλληλουχιών είναι κάποιοι από αυτούς τους παράγοντες.

Ένα από τα μεγαλύτερα σφάλματα της PCR που συμβαίνει σε δείγματα με μεικτά πρότυπα DNA είναι ο σχηματισμός χιμαιρικών, δηλαδή τεχνικών αλληλουχιών που σχηματίζονται από την εσφαλμένη ένωση δύο ή περισσότερων βιολογικών αλληλουχιών (Haas et al., 2011). Αυτό οφείλεται στις ελλιπείς αντιγραφές ενός κλώνου κατά τη διάρκεια της PCR οι οποίες επιτρέπουν στους επόμενους κύκλους να χρησιμοποιούν έναν μερικώς εκτεταμένο κλώνο για να συνδεθούν με το πρότυπο μιας διαφορετικής, αλλά παρόμοιας αλληλουχίας (**Σχήμα 1.14**). Ο μερικώς εκτεινόμενος κλώνος δρα στη συνέχεια ως εκκινητής για να επεκταθεί και να σχηματίσει μια χιμαιρική αλληλουχία. Μόλις δημιουργηθεί, η χιμαιρική αλληλουχία στη συνέχεια ενισχύεται περαιτέρω σε επόμενους κύκλους. Το τελικό αποτέλεσμα είναι ένα τεχνούργημα της PCR που δεν αντιπροσωπεύει μια αλληλουχία που υπάρχει στη φύση. Αν και υπάρχουν βιοπληροφορικά εργαλεία που εντοπίζουν και αφαιρούν τέτοιες αλληλουχίες, είναι προφανές ότι τα χιμαιρικά επιφέρουν σημαντικό σφάλμα στην ποιότητα και στην σχετική αφθονία των αναγνωσμάτων μετά την διαδικασία αλληλούχισης.



Σχήμα 1.14 Σχηματισμός χιμαιρικών αλληλουχιών κατά την ενίσχυση PCR. Ένα προϊόν επέκτασης που έχει απορριφθεί από έναν προηγούμενο κύκλο PCR μπορεί να λειτουργήσει ως εκκινητής σε

έναν επόμενο κύκλο PCR. Εάν αυτό το προϊόν επέκτασης που έχει απορριφθεί χρησιμοποιηθεί ως εκκινητής για την σύνθεση διαφορετικού θραυσμάτος DNA από το αρχικό, σχηματίζεται μια χιμαιρική αλληλουχία. (Haas et al., 2011)

1.3.5 Επιλογή πλατφόρμας αλληλούχισης νέας γενιάς και προετοιμασία βιβλιοθήκης

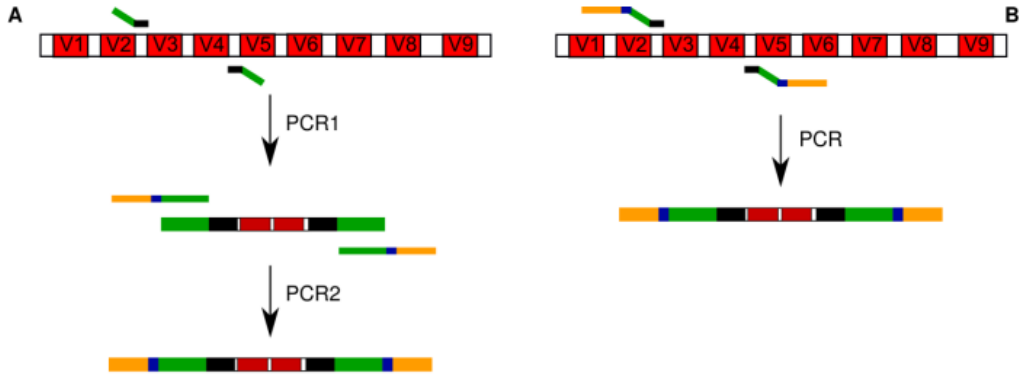
Οι πρόσφατες εξελίξεις στην τεχνολογίας αλληλούχισης Illumina, οι οποίες επέτρεψαν την αύξηση του μήκους αναγνώσματος εώς και 300 bp για single-end ή 500 bp για paired-end αναγνώσματα, έχουν παρακινήσει την επιστημονική κοινότητα να χρησιμοποιεί όλο και περισσότερο τις διαθέσιμες πλατφόρμες της. Αυτό οφείλεται στο γεγονός ότι γίνεται δυνατή η αλληλούχιση σε παραπάνω από μία υπερμεταβλητή περιοχή, προσθέτοντας έτσι περισσότερες φυλογενετικές πληροφορίες στην έρευνα, οι οποίες οδηγουν στην επίλυση ταξινομικών ασαφειών.

Όσον αφορά τον αριθμό των παραγόμενων αναγνωσμάτων, έρευνες έχουν υποδείξει ότι 10,000 με 15,000 αναγνώσματα ανά δείγμα επιτρέπουν έναν ικανοποιητικό χαρακτηρισμό μίας βακτηριακής κοινότητας (Bukin et al., 2019). Το κόστος της αλληλούχισης ποικίλλει σημαντικά και εξαρτάται από την τεχνολογία και την πλατφόρμα. Γενικά, το κόστος ανά βάση της Illumina είναι ελαφρώς χαμηλότερο από αυτό της IonTorrent, ενώ τα ποσοστά σφάλματος μεταξύ αυτών των δύο είναι περίπου 0,1% και 0,5% αντίστοιχα. Βέβαια, και στις δύο τεχνολογίες υπάρχει αρνητική συσχέτιση ποσότητας και ποιότητας δεδομένων, όπου κατά μήκος του αναγνώσματος η ποιότητα μειώνεται λόγω της ασυγκράτητης φάσης (not keeping phase), στην οποία η σύνθεση ορισμένων μορίων καθυστερεί ή ξεπερνά την πλειοψηφία (Tan et al., 2019). Επιπλέον, για την Illumina παρατηρείται μια αρνητική συσχέτιση μεταξύ της πυκνότητας των συστάδων και της μέσης ποιότητας των αναγνωσμάτων, η οποία είναι ιδιαίτερα έντονη όταν είναι χαμηλή η ποικιλομορφία των εισαγόμενων αμπλικονίων.

Η δημιουργία βιβλιοθήκης στις NGS τεχνολογίες, δηλαδή η κατάλληλη επεξεργασία των δειγμάτων για την εισαγωγή τους στις πλατφόρμες, απαιτεί την προσθήκη προσαρμογέων (adapters) στα άκρα των πρότυπων DNA για την ανάγνωσή του. Στην περίπτωση της Illumina, τα δύο ζεύγη προσαρμογέων ονομάζονται P5 και P7 (Gołębiewski & Tretyn, 2020). Ο βασικός ρόλος αυτών είναι η ακινητοποίηση των θραυσμάτων DNA εντός της πλατφόρμας, η οποία επιτρέπει την αλληλούχισή τους. Επιπλέον, η αλληλούχιση περισσότερων του ενός δειγμάτων σε μία εκτέλεση απαιτεί την προσθήκη μικρών ετικετών με αλληλουχίες στα θραύσματα, οι οποίες επιτρέπουν την αντιστοίχιση αυτών με τα δείγματα από τα οποία προήλθαν. Οι ετικέτες αυτές ονομάζονται Μοριακές Ταυτότητες (Molecular IDentifiers, MIDs) ή ραβδοκώδικες (barcodes). Ανάλογα με την τεχνολογία, οι ετικέτες μπορούν να αλληλουχηθούν μαζί με τα θραύσματα DNA (IonTorrent) ή να αναγνωριστούν ξεχωριστά (Illumina), γεγονός που προτιμάται καθώς επιτρέπει την αλληλούχιση μεγαλύτερων θραυσμάτων.

Και στις δύο τεχνολογίες αλληλούχισης, η προσθήκη των προσαρμογέων και ετικετών επιτυγχάνεται με την σύντηξή τους με τους επιλεγμένους εκκινητές μέσω ενός ή δύο κύκλων PCR (**Σχήμα 1.15**) (Gołębiewski & Tretyn, 2020). Η προετοιμασία της βιβλιοθήκης ενός κύκλου PCR απαιτεί τον σχεδιασμό εκκινητών που αποτελούνται, ξεκινώντας από το 5' άκρο, από τον προσαρμογέα, τον barcode, μια καθολική αλληλουχία και τον ειδικό εκκινητή για την στοχοποίηση της επιθυμητής περιοχής του γονιδίου. Στην προσέγγιση των δύο κύκλων PCR, χρειάζονται δύο σετ εκκινητών: έναν για την ενίσχυση του γονιδίου το οποίο περιλαμβάνει την καθολική αλληλουχία και τον ειδικό επιλεγμένο εκκινητή, καθώς και έναν

για την προσθήκη των προσαρμογέων και barcodes, που περιλαμβάνει τον προσαρμογέα, τον barcode και την καθολική αλληλουχία. Ο ρόλος της καθολικής αλληλουχίας και στις δύο περιπτώσεις είναι να γεφυρώσει το επιθυμητό βιολογικά γονίδιο με τις απαραίτητες τεχνικές αλληλουχίες για την αλληλούχισή του. Στην πραγματικότητα, λοιπόν, ανεξαρτήτως της προσέγγισης, η προετοιμασία της βιβλιοθήκης ξεκινάει πριν οποιασδήποτε εφαρμογή PCR, με το ζεύγος εκκινητών να σχεδιάζονται με τέτοιο τρόπο, ώστε να αποτελείται και από τις καθολικές αλληλουχίες και από τις αλληλουχίες για την πρόσδεσή τους στην επιθυμητή περιοχή τους γονιδίου. Για την ελαχιστοποίηση των σφαλμάτων της PCR, είναι προτιμητέα η προσέγγιση στην οποία εφαρμόζεται ένας κύκλος.



Σχήμα 1.15 Οι προσεγγίσεις προετοιμασίας βιβλιοθήκης, όπου **(A)** η μέθοδος δύο κύκλων PCR και **(B)** η μέθοδος ενός κύκλου PCR. Οι χρωματικές ενδείξεις για τους εκκινητές είναι οι εξής: μαύρο – καθολικοί εκκινητές για την ενίσχυση της επιθυμητής περιοχής του γονιδίου, πράσινο – καθολικές αλληλουχίες, μπλε – barcodes/MID και Πορτοκαλί – αλληλουχία προσαρμογέων. (Gołębiewski & Tretyn, 2020)

Αφού ολοκληρωθεί η κατασκευή της βιβλιοθήκης για κάθε δείγμα, αξιολογείται η ποσότητά της και υποβάλλονται όλες οι βιβλιοθήκες σε κανονικοποίηση (Hussing et al., 2018). Η ποσοτικοποίηση μιας βιβλιοθήκης είναι ένα από τα βασικά βήματα που διασφαλίζουν καλά αποτελέσματα αλληλούχισης και στις δύο τεχνολογίες. Όσο μεγαλύτερη η ποσότητα των αμπλικονίων, τόσο μειώνεται η πιθανότητα σφάλματος κατά την διάρκεια της αλληλούχισης. Επιπρόσθετα, είναι πολύ σημαντική και η αραίωση των βιβλιοθηκών σε ίση μοριακή συγκέντρωση, μια διαδικασία η οποία ονομάζεται κανονικοποίηση. Με αυτόν τον τρόπο διασφαλίζεται η ομοιόμορφη αλληλούχιση όλων των θραυσμάτων DNA.

1.3.6 Επιμολύνσεις και δείγματα ελέγχου

Όπως έχει ήδη αναφερθεί, τα αποτελέσματα μίας έρευνας μικροβιώματος με την ανάλυση του 16S rRNA γονιδίου θα μπορούσαν να επηρεαστούν από διάφορους παράγοντες, όπως οι μέθοδοι δειγματοληψίας, τα κιτ απομόνωσης DNA και οι μέθοδοι αλληλούχισης. Ένας ακόμα σημαντικός παράγοντας που είναι παρόν σε κάθε πειραματική εφαρμογή της ανάλυσης και που μπορεί να επιφέρει προβλήματα στην αξιολόγηση των αποτελεσμάτων αμπλικονίων είναι οι επιμολύνσεις (Hornung et al., 2019). Υπάρχουν δύο είδη επιμολύνσεων που πρέπει να ληφθούν υπόψη, η διασταυρούμενη επιμόλυνση μεταξύ των δειγμάτων και η εξωγενής επιμόλυνση από το περιβάλλον. Για την μέγιστη αποφυγή επιμόλυνσης, που μπορεί να συμβεί σε όλα τα πειραματικά στάδια, πρέπει να παρθούν διαφορετικά μέτρα αντιμετώπισης για το κάθε στάδιο.

Αρχικά, η δειγματοληψία και η μεταφορά των δειγμάτων εγκυμονούν αναπόφευκτα κίνδυνους επιμόλυνσης, είτε διασταυρούμενης μεταξύ των δειγμάτων είτε εξωγενούς φύσεως από τον εξοπλισμό δειγματοληψίας και τα φιαλίδια αποθήκευσης. Η αντιμετώπιση αυτών των κίνδυνων περιλαμβάνει την χρήση αποστειρωμένου εξοπλισμού μιας χρήσης ή την σχολαστική αποστείρωση (π.χ. φλόγα) μη αναλώσιμων εργαλείων. Στην διαδικασία της απομόνωσης του DNA και της δημιουργίας μεταγονιδιωματικής βιβλιοθήκης ο έλεγχος επιμόλυνσης γίνεται όλο και πιο δύσκολος. Υπάρχει περίπτωση τα κιτ απομόνωσης του DNA να είναι μολυσμένα, ενώ κατά την κατασκευή της βιβλιοθήκης, εκτός από τον κίνδυνο μολυσμένων αντιδραστηρίων, η πιθανότητα διασταυρούμενης επιμόλυνσης είναι μεγάλη κατά την ρύθμιση των αντιδράσεων (Salter et al., 2014). Επιπλέον, οι αναπηδήσεις ετικετών, οι μεταλλάξεις των barcodes και οι υπολειπόμενες αλληλουχίες από προηγούμενες εκτελέσεις μπορούν να προκαλέσουν εξίσου διασταυρούμενες επιμολύνσεις κατά την διάρκεια της αλληλούχισης (Gołębiewski & Tretyn, 2020). Σε αυτές τις περιπτώσεις, πρέπει να δοθεί έμφαση στον σχεδιασμό των barcodes για να ελαχιστοποιηθεί ο κίνδυνος αναπήδησης των ετικετών, ενώ υποσύστημα της πλατφόρμας αλληλούχισης πρέπει να απολυμαίνεται συστηματικά με χρήση διαλύματος υποχλωριώδους νατρίου για να αποτραπεί η αλληλούχιση προηγούμενων αλληλουχιών.

Σε κάθε περίπτωση, ο βέλτιστος τρόπος αντιμετώπισης επιμολύνσεων και αξιολόγησης αποτελεσμάτων είναι η συμπερίληψη αρνητικών και θετικών δειγμάτων ελέγχου κατά την προετοιμασία δειγμάτων για αλληλούχιση (Eisenhofer et al., 2019). Με αυτόν τον τρόπο, παρέχεται η δυνατότητα εντοπισμού οποιασδήποτε επιμόλυνσης στη ροή των πειραματικών εφαρμογών, καθώς και βελτίωσης της μεταγενέστερης βιοπληροφορικής ανάλυσης. Υπάρχουν δύο είδη αρνητικών δειγμάτων ελέγχου που χρησιμοποιούνται συχνά σε αναλύσεις μικροβιώματος: τα δείγματα ελέγχου απομόνωσης DNA και βιβλιοθήκης αντίστοιχα. Το πρώτο είδος χρησιμοποιείται κατά την απομόνωση του νουκλεϊκού οξέος και τυπικά αποτελείται από ένα ρυθμιστικό διάλυμα λύσης, απουσία οποιουδήποτε βιολογικού υλικού. Τα δείγματα ελέγχου βιβλιοθήκης, ονομαζόμενα επίσης ως δείγματα ελέγχου απουσίας πρότυπου DNA (no-template controls, NTCs) χρησιμοποιούνται κατά την ενίσχυση PCR των επιλεγμένων υπερμεταβλητών περιοχών και αποτελούνται από υπερκαθαρό νερό, αντί για πρότυπο DNA. Και τα δύο είδη δειγμάτων ελέγχου θα πρέπει να περιλαμβάνονται σε όλη την πειραματική ροή και να ακολουθούνται μαζί με τα βιολογικά δείγματα προς ανάλυση. Λόγω του ότι δεν υπάρχει σαφήνεια ως προς τον κατάλληλο αριθμό και τύπο αρνητικών δειγμάτων ελέγχου, είναι στο χέρι του ερευνητή να αποφασίσει τα καταλληλότερα μέτρα για τη διασφάλιση της εσωτερικής εγκυρότητας και εμπιστοσύνης των ευρημάτων της μελέτης (Gołębiewski & Tretyn, 2020).

Τα θετικά δείγματα ελέγχου είναι τα δείγματα που περιέχουν βακτηριακές κοινότητες από τις οποίες είναι γνωστή η ταξινομική περιεκτικότητα της, καθώς και η συγκέντρωσή της (Weinroth et al., 2022). Τα δείγματα αυτά, ονομαζόμενα επίσης ως πλαστές κοινότητες (mock communities), μπορούν να κατασκευαστούν στο εργαστήριο ή να προμηθευτούν από το εμπόριο. Η χρήση αυτών εξυπηρετεί στον εσωτερικό τεχνικό έλεγχο διατήρησης και επεξεργασίας δειγμάτων, στον έλεγχο της απόδοση της απομόνωσης DNA και στην αποτελεσματικότητα της αλληλούχισης. Με αυτόν τον τρόπο, επιτρέπεται η αξιολόγηση μεροληπτικών και ποσοστιαίων σφαλμάτων καθώς και της συνολικής ποιότητας, τόσο των πειραματικών όσο και των βιοπληροφορικών διαδικασιών.

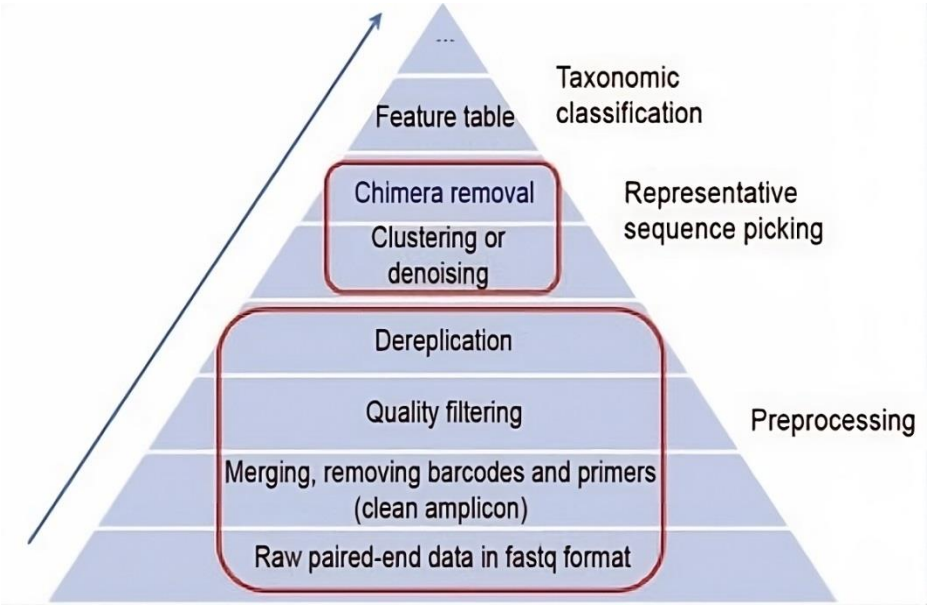
Συνεπώς, η χρήση δειγμάτων ελέγχου είναι σημαντική για τον χαρακτηρισμό μικροβιωμάτων, ειδικά όταν τα δείγματα έχουν χαμηλή βακτηριακή βιομάζα. Δυστυχώς,

μόνο το 30% των υπαρχόντων ερευνών έχει αναφέρει την χρήση αρνητικών δειγμάτων ελέγχου, και το 10% την χρήση θετικών αντίστοιχα (Hussing et al., 2018). Πολλές έρευνες έχουν διαπιστώσει τον ενδεχόμενο βακτηριακό αποικισμό σε δείγματα που έως τότε είχαν αναγνωριστεί ως στείρα (Eisenhofer et al., 2019). Ωστόσο, αυτά τα θετικά αποτελέσματα μπορεί να έχουν προκληθεί λόγω επιμολύνσεων, και η απουσία δειγμάτων ελέγχου σε αυτές τις έρευνες δεν βοηθάει στην εγκυρότητά τους. Επομένως, συνιστάται πολύ έντονα η χρήση αρνητικών και θετικών δειγμάτων ελέγχου σε αναλύσεις μικροβιώματος από δείγματα χαμηλής βιομάζας, όπως το αίμα, το αμνιακό υγρό, το εγκεφαλονωτιαίο υγρό και ο πλακούντας (Karstens et al., 2019).

1.4 Βιοπληροφορική Ανάλυση του 16S rRNA Γονιδίου

1.4.1 Εισαγωγή

Τα δεδομένα που προκύπτουν από τις πλατφόρμες αλληλούχισης περιλαμβάνουν εκατοντάδες χιλιάδες, ή ακόμα και εκατομμύρια αναγνώσματα, αποθηκευμένα σε ειδικά διαμορφωμένα αρχεία. Η βιοπληροφορική επεξεργασία των δεδομένων αλληλούχισης μίας υπερμεταβλητής περιοχής του 16S rRNA γονιδίου μπορεί να δώσει τις αρχικές αλληλουχίες από τις οποίες προήλθαν καθώς και τις ταξινομικές πληροφορίες για τον μικροβιακό πληθυσμό του δείγματος. Τα στάδια αυτής της επεξεργασίας περιλαμβάνουν την προεπεξεργασία των δεδομένων, από την οποία τα αναγνώσματα συγχωνεύονται στην περίπτωση που είναι paired-end και καθαρίζονται από χαμηλής ποιότητας αλληλουχίες και αλληλουχίες μη-βιολογικής προέλευσης, και την παραγωγή αντιπροσωπευτικών αλληλουχιών των αμπλικονίων, κατά την οποία τα αναγνώσματα οργανώνονται σε ομάδες είτε με την μέθοδο ομαδοποίησης είτε με αυτήν της αποθορυβοποίησης (**Σχήμα 1.16**). Επιπλέον, περιλαμβάνεται και η διαχείριση χιμαιρικών αλληλουχιών καθώς και αυτών που έχουν προκύψει από την επιμόλυνση των δειγμάτων. Τέλος, ακολουθεί η ταξινομική ανάλυση και την φυλογενετική συσχέτιση των τελικών αντιπροσωπευτικών αλληλουχιών. Οι διαδικασίες αυτές δίνουν σημαντικές πληροφορίες τόσο για την μικροβιακή σύνθεση των δειγμάτων όσο και για την σύγκριση της ποικιλομορφίας αυτών μέσω στατιστικών αναλύσεων. Δεδομένης της πολυπλοκότητας και της άμεσης εξάρτησης του κάθε προαναφερόμενου σταδίου ανάλυσης από τα αρχικά δεδομένα αλληλούχισης, γίνεται εμφανής η σημαντικότητα της εξασφάλισης της ορθότητας κάθε σταδίου στις παρακάτω υποενότητες.



Σχήμα 1.16 Τα βασικά στάδια της βιοπληροφορικής ανάλυσης δεδομένων αλληλούχισης αμπλικονίων για την εύρεση των ταξινομικών πληροφοριών τους. (Qian et al., 2020)

1.4.2 Πρωτογενή δεδομένα και προεπεξεργασία

Τα δεδομένα που προκύπτουν από τις πλατφόρμες αλληλούχισης είναι συνήθως αρχεία κειμένου της μορφής FASTQ ή FASTA (Hosseini et al., 2016). Σε ένα αρχείο FASTQ, η καταχώρηση ενός αναγνώσματος περιλαμβάνει τέσσερα πεδία, τον τίτλο του αναγνώσματος, την αλληλουχία του, μια προαιρετική επανάληψη του τίτλου και τις πληροφορίες σχετικά με την ποιότητά του (**Σχήμα 1.17**). Η μορφολογία του αρχείου FASTA είναι πανομοιότυπη με αυτής του FASTQ, με την βασικά διαφορά τους να έγκειται στο ότι στα αρχεία FASTA δεν περιλαμβάνεται η προαιρετική επανάληψη του τίτλου και κυρίως οι πληροφορίες ποιότητας.

```
Identifier      @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence        TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign        +
Quality scores  hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier      @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence        GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign        +
Quality scores  hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

Σχήμα 1.17 Παράδειγμα μορφολογίας ενός αρχείου FASTQ που περιλαμβάνει την καταχώρηση δύο αναγνωσμάτων.(Hosseini et al., 2016)

Ο τίτλος αρχίζει με το σύμβολο @ και ακολουθείται από ένα αναγνωριστικό αλληλουχίας και μία προαιρετική περιγραφή, όπως το όνομα οργάνου αλληλούχισης, το μήκος του αναγνώσματος κ.λπ. Η αλληλουχία του αναγνώσματος περιέχει την αλληλουχία των βάσεων που προέκυψε από την αλληλούχιση του θραύσματος DNA, των οποίων ο συμβολισμός τους είναι A, C, G και T. Στην περίπτωση που η πλατφόρμα αποτύχει να προσδιορίσει μία βάση, η κατοχύρωση της γίνεται με διφορούμενα σύμβολα, όπως το N. Η τρίτη γραμμή κειμένου ξεκινά με το σύμβολο + και στην συνέχεια υπάρχει είτε κενό είτε η επανάληψη του τίτλου αναγνώσματος, με το @ να απουσιάζει. Στο τελευταίο πεδίο

παραθέτονται οι βαθμολογίες ποιότητας σε κωδικοποιημένη μορφή ASCII [3] για κάθε νουκλεοτιδική βάση. Η διαβάθμιση ακολουθεί το μοντέλο βαθμολογίας ποιότητας PHRED, σύμφωνα με το οποίο μία βαθμολογία ποιότητας PHRED Q σχετίζεται λογαριθμικά με την πιθανότητα εσφαλμένης καταγραφής της βάσης P ως εξής:

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

Τα σύνολα συμβόλων ASCII που χρησιμοποιούνται για την κωδικοποίηση της βαθμολογίας ποιότητας είναι τα [33:73] ή [64:104], όπου οι τιμές των συμβόλων είναι ίσες με Q+33 ή Q+64 αντίστοιχα (Malysa et al., 2015).

Όταν σε μία ανάλυση του 16S rRNA γονιδίου χρησιμοποιούνται περισσότερα από ένα δείγματα προς μελέτη, εισάγονται όλα μαζί για αλληλούχιση με σκοπό την εξοικονόμηση χρημάτων. Έτσι, όλα τα αμπλικόνια αλληλουχούνται ταυτόχρονα και η καταχώρηση των αναγνωσμάτων γίνεται τυχαία. Μερικές πλατφόρμες αλληλούχισης παρέχουν ενσωματωμένα λογισμικά τα οποία αυτόματα διαχωρίζουν και οργανώνουν τα πρωτογενή αναγνώσματα με βάση τα barcodes που περιέχουν, τα αφαιρούν από τα αναγνώσματα και στην συνέχεια τα αποθηκεύονται σε ξεχωριστά για το κάθε δείγμα FASTQ αρχεία (Wilkins et al., 2021). Υπάρχουν όμως περιπτώσεις που δεν είναι δυνατή η εφαρμογή αυτού του λογισμικού ή η ίδια η πλατφόρμα να μην διαθέτει αυτό το εργαλείο. Τα δεδομένα που προκύπτουν σε αυτές τις περιπτώσεις είναι πολυπλεγμένα (multiplexed) και είναι αναγκαία η εφαρμογή ενός εξωτερικού λογισμικού ή pipeline που να παρέχει την δυνατότητα διαχωρισμού των αναγνωσμάτων. Η διαδικασία διαχωρισμού, ονομαζόμενη επίσης ως αποπολυπλεξία (demultiplex), επιτυγχάνεται με βάση της αλληλουχίες των barcode που περιέχουν τα αναγνώσματα σε συνδυασμό με ένα αρχείο μεταδεδομένων, για την αντιστοίχιση των barcodes με τα αρχικά δείγματα. Αφού διαχωριστούν και οργανωθούν σε συστάδες τα αναγνώσματα με βάση τα barcodes, αποθηκεύονται ξεχωριστά και κατάλληλα σε μορφή FASTQ ή FASTA. Μόνο εάν τα δεδομένα είναι αποπολυπλεγμένα μπορεί να εφαρμοστεί οποιαδήποτε περεταίρω επεξεργασία και ανάλυση στα αναγνώσματα.

Η πρώτη παρεμβατική ενέργεια που εφαρμόζεται στα αναγνώσματα που προέρχονται από δεδομένα αλληλούχισης του 16S γονιδίου είναι η αποκοπή και αφαίρεση αλληλουχιών μη-βιολογικής προέλευσης. Οι αλληλουχίες αυτές μπορούν να προέρχονται από τον προσαρμογέα, από τα barcodes ή από του καθολικούς εκκινητές. Οι πλατφόρμες αλληλούχισης που παράγουν αυτόματα αποπολυπλεγμένα δεδομένα λογικά οι προσασμογείς και τα barcodes δεν περιέχονται στα αναγνώσματα γιατί αφαιρούνται επίσης αυτόματα. Όμως, ακόμα και σε αυτήν την περίπτωση, είναι αναγκαία η αφαίρεση των αλληλουχιών του εκκινητή. Έχουν αναπτυχθεί κατάλληλα επεξεργαστικά εργαλεία για τον αποκλειστικό εντοπισμό ανεπιθύμητων αλληλουχιών και την αποκοπή τους, με την συνοδεία φυσικά των κατάλληλα διαμορφωμένων μεταδεδομένων. Το cutadapt (Martin, 2011) είναι ένα από πιο δημοφιλές εργαλεία για αυτόν το σκοπό, το οποίο υποστηρίζει επίσης και την διαδικασία της αποπολυπλεξίας.

Επιπλέον, εάν τα δεδομένα αλληλούχισης έχουν προκύψει από την τεχνική paired-end της πλατφόρμας Illumina, τα παραγόμενα αναγνώσματα είναι υπό την μορφή μικρο - αναγνωσμάτων. Τα μικρο-αναγνώσματα έχουν προέλθει από την αλληλούχιση των

[3] Ο «Αμερικάνικος Πρότυπος Κώδικας για Ανταλλαγή Πληροφοριών» (American Standard Code for Information Interchange, ASCII) είναι ένα κωδικοποιημένο σύνολο λατινικών χαρακτήρων για την αναπαράσταση κειμένου κυρίως στους υπολογιστές και σε συσκευές τηλεπικοινωνίας.

εμπρόσθιας και ανάστροφης περιοχής του κάθε στοχευμένου θραύσματος DNA. Ιδανικά, κατά τον σχεδιασμό της έρευνας όταν αποφασίζεται η χρήση αυτής της τεχνικής, το μήκος της στοχευόμενης περιοχής του 16S γονιδίου προτιμάται να είναι προσεγγιστικά μικρότερο από το συνολικό μήκος των παραγόμενων μικρο-αναγνωσμάτων. Εάν το μέγεθος του στοχευόμενου θραύσματος DNA είναι μικρότερο από το διπλάσιο του μήκους του μικρο-αναγνώσματος ενός άκρου, δηλαδή εάν υπάρχει επικάλυψη, τα αντίστοιχα ζεύγη μικρο-αναγνωσμάτων μπορούν να συγχωνευθούν σε ένα θραύσμα/ανάγνωσμα. Με τη συγχώνευση paired-end αναγνωσμάτων, η επικαλυπτόμενη περιοχή μεταξύ τους μπορεί επίσης να αναπτυχθεί για τη διόρθωση σφαλμάτων αλληλουχίας και πιθανώς να αποδώσει αλληλουχίες υψηλότερης ποιότητας.

Η ένωση paired-end αναγνωσμάτων αποτελεί εξίσου ένα από τα σημαντικά πρώτα βήματα επεξεργασίας. Ως εκ τούτου, η ακρίβειά της είναι κρίσιμη για όλες τις μεταγενέστερες αναλύσεις. Συνήθως, τα εμπρόσθια (forward) και ανάστροφα (reverse) αναγνώσματα είναι αποθηκευμένα σε ξεχωριστά αρχεία σε ένα φάκελο που αντιστοιχεί στο δείγμα από τα οποία προήλθαν. Φυσικά, η κατοχύρωση των δεδομένων μπορεί να έχει προκύψει είτε από την ίδια την πλατφόρμα αλληλούχισης είτε από την διαδικασία dimultiplex. Ο αλγόριθμός που χρησιμοποιείται συχνά για την συγχώνευση paired-end αναγνωσμάτων είναι ο PEAR (Zhang et al., 2014), ο οποίος έχει ενσωματωθεί σε πολλά βιοπληροφορικά εργαλεία και λογισμικά για τον ίδια σκοπό. Η σειρά με την οποία πρέπει να γίνει η αφαίρεση μη βιολογικών αλληλουχιών και η συγχώνευση των paired-end αναγνωσμάτων δεν είναι αυστηρή, αν και προτιμάται γενικά τα δεδομένα να απαλλάσσονται από μη βιολογικές πληροφορίες όσο πιο νωρίς γίνεται (Dacey & Chain, 2021), ενδεχομένως για την ελάφρυνση του όγκου δεδομένων στις μεταγενέστερες επεξεργασίες.

Μία σημαντική διαδικασία στη βιοπληροφορική ανάλυση του 16S είναι ο ποιοτικός έλεγχος των αναγνωσμάτων. Εκτός από το ποσοστιαίο σφάλμα αλληλούχισης μίας βάσης, που για τις πλατφόρμες Illumina είναι 0,1%, η ποιότητα των βάσεων είναι συνήθως λίγο πιο χαμηλή στις πρώτες βάσεις κάθε αναγνώσματος και αρκετά πιο χαμηλή στις τελευταίες βάσεις (Tan et al., 2019). Η παράβλεψη της ύπαρξης αυτών μπορεί να είναι ζημιογόνα για οποιαδήποτε μεταγονιδιωματική ανάλυση, καθώς μπορεί να προσθέσει αναξιόπιστες και δυνητικά τυχαίες αλληλουχίες στο σύνολο δεδομένων και να οδηγήσει στην εσφαλμένη ερμηνεία αυτών (Del Fabbro et al., 2013). Συνεπώς, η εξέταση της γενικής ποιότητας των δεδομένων αποτελεί απαραίτητο βήμα για την αξιολόγηση της αξιοπιστίας των αναγνωσμάτων, και κατά επέκταση την εξασφάλιση έγκυρων μεταγενέστερων αποτελεσμάτων (Bokulich et al., 2013). Δεδομένου αυτού, η ανάπτυξη τεχνικών και επεξεργαστικών εργαλείων για την αντιμετώπιση σφαλμάτων αλληλούχισης ήταν αναπόφευκτη. Γενικά, οι τεχνικές βελτίωσης ποιότητας αναγνωσμάτων περιλαμβάνουν την μερική αποκοπή ενός ή και των δύο άκρων τους, την αξιολόγηση της γενικής ποιότητας τους για την επιλογή διατήρησης ή απόρριψής τους και την συγχώνευση των paired-end.

Η λογική της μερικής αποκοπής των άκρων κάθε αναγνώσματος είναι αρκετά απλή. Αφαιρώντας τα άκρα του αναγνώσματος που τείνουν να περιέχουν βάσεις με χαμηλότερη ποιότητα σε σχέση με το κεντρικό του κομμάτι, αυτομάτως αυξάνεται η γενική ποιότητά του. Η διαδικασία της αφαίρεσης των αλληλουχιών των εκκινητών από τα αναγνώσματα συνήθως εξασφαλίζει ταυτόχρονα και την αποκοπή της χαμηλής ποιότητας βάσεων του αρχικού άκρου. Στην αφαίρεση του τελικού άκρου του αναγνώσματος πρέπει να ληφθούν υπόψη κάποια σημαντικά χαρακτηριστικά. Αρχικά, στην περίπτωση των single-end αναγνωσμάτων, η αποκοπή είναι αρκετά ελαστική και η μόνη παράμετρος που πρέπει να ληφθεί υπόψη είναι

το τελικό μήκους του αναγνώσματος. Αυτό οφείλεται στο ότι η διατήρηση του όσο τον δυνατόν γίνεται μέγιστου μήκους αναγνώσματος ισοδυναμεί με την αποκόμιση περισσότερων ταξινομικών και φυλογενετικών πληροφοριών στις περαιτέρω αναλύσεις (Del Fabbro et al., 2013).

Η ίδια λογική εμπίπτει και στα paired-end αναγνώσματα, αλλά ο ερευνητής καλείται να αποφασίσει εάν θα αποκόψει πριν ή μετά την συγχώνευση των μικρο-αναγνωσμάτων. Στην περίπτωση της προ-συγχώνευσή τους, η μερική αποκοπή των τελικών άκρων των μικρο-αναγνωσμάτων αποδεδειγμένα βελτιώνει σημαντικά τα μεταγενέστερα αποτελέσματα ως προς την αξιοπιστίας τους (Mohsen et al., 2019). Όμως, ο μόνος τρόπος για να συμβεί αυτό είναι η εξασφάλιση του επαρκούς μήκους της επικαλυπτόμενης περιοχής, έτσι ώστε τα paired-end αναγνώσματα να ενωθούν αποτελεσματικά στην συνέχεια. Η απουσία αυτής της εξασφάλισης επιφέρει την αποτυχία της συγχώνευσης των αναγνωσμάτων και, κατά συνέπεια, την απώλεια μεγάλου όγκου αναγνωσμάτων από τα αρχικά στάδια της ανάλυσης. Μία πολύ καλή λύση αυτού του προβλήματος είναι η συγχώνευση των paired-end αναγνωσμάτων πριν την εξέταση της ποιότητας τους. Με αυτόν τον τρόπο εξασφαλίζεται η απαραίτητη επικαλυπτόμενη περιοχή και ταυτόχρονα βελτιώνεται η αξιοπιστία και ποιότητά αυτής της περιοχής (Zhang et al., 2014). Αυτό οφείλεται στο γεγονός ότι οι τιμές ποιότητας των βάσεων της επικαλυπτόμενης περιοχής υποβάλλονται σε επεξεργασία κατά την συγχώνευση, στην οποία προσδιορίζονται και προβλέπονται βελτιωμένες τελικές τιμές ποιότητας λόγω επιτυχής αντιστοιχίας. Ωστόσο, η παρουσία χαμηλής ποιότητας βάσεων στο τελικό άκρων των μικρο-αναγνωσμάτων, που ισοδυναμεί με την παρουσία ενδεχόμενων εσφαλμένων βάσεων στην επικαλυπτόμενη περιοχή, μπορεί να επιφέρει εξίσου την αποτυχία συγχώνευσης, οδηγώντας και πάλι στην απώλεια αναγνωσμάτων (Mohsen et al., 2019).

Μια εξίσου καλή τεχνική ελέγχου ποιότητας είναι η αξιολόγηση της συνολικής ποιότητας του κάθε αναγνώσματος. Η αποκοπή των δύο άκρων και η συγχώνευση των αναγνωσμάτων προσφέρουν σίγουρα την βελτίωση της ποιότητάς του, αλλά δεν εξασφαλίζουν την γενική καλή ποιότητα του τελικού αναγνώσματος. Η απλή μέθοδος αξιολόγησης ποιότητας του συνολικού αναγνώσματος είναι ο προσδιορισμός του αριθμού βάσεων που έχουν μικρότερη από μια επιθυμητή τιμή βαθμολογία ποιότητας Q, και η αφαίρεση του αναγνώσματος από τα δεδομένα αν ο αριθμός αυτών των βάσεων θεωρείται μεγάλος (Bokulich et al., 2013). Γενικά, είναι κοινώς αποδεκτό ότι οι τιμές βαθμολογίας ποιότητας $Q \geq 20$ χαρακτηρίζουν τις βάσεις επαρκώς αξιόπιστες για την διατήρησή τους (Pfeifer, 2017). Συνεπώς, θέτοντας την ελάχιστη τιμή ποιότητας βάσης $Q_{min}=20$ είναι μία καλή στρατηγική για να προσδιοριστεί πόσες κακής ποιότητας βάσεις περιέχει ένα ανάγνωσμα. Εάν είναι σημαντικά μεγάλος αυτός ο αριθμός, ακόμα και μετά την αποκοπή των άκρων του, προτείνεται η εξ ολοκλήρου αφαίρεση αυτού του αναγνώσματος.

Μία πιο αποτελεσματική, και ενδεχομένως αντικειμενική, προσέγγιση αξιολόγησης γενικής ποιότητας αναγνώσματος είναι ο προσδιορισμός του αναμενόμενου αριθμού σφαλμάτων (expected errors) σε κάθε ανάγνωσμα (Edgar & Flyvbjerg, 2015). Με την υπόθεση ότι τα σφάλματα σε διαφορετικές θέσεις συμβαίνουν ανεξάρτητα, ο αναμενόμενος αριθμός σφαλμάτων e.e. ορίζεται ως ο μέσος αριθμός σφαλμάτων E που θα παρατηρούνταν σε μια μεγάλη συλλογή αλληλουχιών n, όπου το ποσοστό σφάλματος σε κάθε θέση p_i δίνεται από τη βαθμολογία ποιότητάς της Q_i, και προσδιορίζεται ως εξής:

$$e.e. = \frac{E}{n} = \frac{1}{n}\sum_{n}^{i} p_i = \frac{1}{n}\sum_{n}^{i} 10^{-Q_i/10}$$

Ο τρόπος φιλτραρίσματος χαμηλής ποιότητας αναγνωσμάτων είναι η επιβολή μίας μέγιστης τιμής $e.e._{max}$ στον μέσο αναμενόμενο αριθμό σφαλμάτων, έτσι ώστε τα αναγνώσματα με υψηλό e.e. να απορριφθούν. Γενικά, οι τιμές e.e. ≤ 3% που προκύπτουν από τον προσδιορισμό της ποιότητας του αναγνώσματος θεωρούνται καλές και, κατά συνέπεια, εξασφαλίζουν την αξιοπιστία του αναγνώσματος.

Εν' ολίγοις, ο ποιοτικός έλεγχος είναι ένα αρκετά περίπλοκο στάδιο της ανάλυσης δεδομένων αλληλούχισης του 16S rRNA γονιδίου, στο οποίο πρέπει να παρθούν πολλές αποφάσεις για την μεγιστοποίηση της αξιοπιστίας των αποτελεσμάτων. Η απουσία ενός καθολικού πρωτόκολλου δεν ευνοεί την κατάσταση, αλλά τουλάχιστον κάποιες βασικές προσεγγίσεις, όπως η αποκοπή των άκρων και η εκτίμηση της συνολικής ποιότητας αναγνωσμάτων, αδιαμφισβήτητα θα πρέπει να εφαρμόζονται στην ροή επεξεργασίας. Η επιλογή βασικών παραμέτρων, δηλαδή το μήκος αποκοπής και το όριο ποιότητας, εξαρτάται από τα δεδομένα που έχουν προκύψει από την αλληλούχιση, όπως το μήκος αναγνωσμάτων και το ποσοστό σφάλματος της πλατφόρμας, καθώς και την ελαστικότητα της έρευνας, όπου αν είναι κλινικής σημασίας θα πρέπει να επιβληθούν πολύ πιο αυστηρά όρια ποιότητας. Τα πιο δημοφιλή εργαλεία για τον ποιοτικό έλεγχο δεδομένων είναι τα λογισμικά FastQC (Andrews, 2010) και Trimmomanic (Bolger et al., 2014), καθώς και ο αλγόριθμος q-score (Bokulich et al., 2013) ο οποίος είναι διαθέσιμος στο λογισμικό QIIME2 (Bolyen et al., 2019).

Τέλος, για την μείωση του όγκου τους, τα δεδομένα υποβάλλονται σε διαδικασία διαγραφής πανομοιοτήτων αναγνωσμάτων και στην αντιστοίχιση όλων ένα μοναδικό ανάγνωσμα συνοδευόμενο από την συχνότητα με την οποία παρουσιαζόταν πριν την δεδομένη επεξεργασία (dereplication). Με αυτόν τον τρόπο, μειώνεται η απαιτούμενη μνήμη και ο χρόνος επεξεργασίας των δεδομένων στις μεταγενέστερες αναλύσεις. Ο αλγόριθμος VSEARCH (Rognes et al., 2016) παρέχει την συνάρτηση *derep_* για αυτόν τον σκοπό. Όμως, είναι σημαντικό αυτό το βήμα να εφαρμοστεί αφού ολοκληρωθεί ο ποιοτικός έλεγχος των μεμονωμένων αναγνωσμάτων, διότι εξαλείφονται οι βαθμολογίες ποιότητας βάσεων με αυτόν τον τρόπο.

1.4.3 Μέθοδοι παρασκευής αντιπροσωπευτικών παραλλαγών αμπλικονίων – Ομαδοποίηση και Αποθορυβοποίηση

Στην ανάλυση του γονιδίου 16S rRNA, ο προσδιορισμός της προέλευσης μίας συγκεκριμένης περιοχής αλληλουχιών προκύπτει από την αντιστοίχηση γνωστής προέλευσης αλληλουχιών και ο στόχος είναι να προσδιοριστεί η ταξινομική προέλευσή της με βάση έναν δυνητικά μικρό αριθμό παραλλαγών σε σχέση με παρόμοια είδη. Όμως, με την ενίσχυση PCR και την αλληλούχιση στοχευμένης περιοχής, τα παραγόμενα δεδομένα συνοδεύονται από χιμαιρικές αλληλουχίες καθώς και από ψευδές μοναδικές παραλλαγές νουκλεοτιδίων (single nucleotide variants, SNVs) στα αναγνώσματα που έχουν προκύψει από την εσφαλμένη αλληλούχιση. Ενώ το στάδιο του ποιοτικού ελέγχου βοηθάει σε ένα βαθμό στην αποκόμιση καθαρών δεδομένων, δεν επαρκεί για την διόρθωση σφαλμάτων και την απομάκρυνση μη-βιολογικών αναγνωσμάτων που προκύπτουν από αυτές τις εφαρμογές. Αυτό μπορεί να οδηγήσει στην εσφαλμένη απόδοση της ανάλυσης, που σημαίνει είτε στην ανίχνευση ενός παρόμοιου, αλλά εσφαλμένου βακτηρίου, είτε στην ψευδή ανακάλυψη ενός νέου.

Για την ελαχιστοποίηση των προαναφερόμενων επιπτώσεων, έχουν αναπτυχτεί πολλά βιοπληροφορικά εργαλεία τα οποία στοχεύουν στην εύρεση των αντιπροσωπευτικών

παραλλαγών αμπλικονίων (representative amplicon sequences), δηλαδή αλληλουχιών βιολογικής προέλευσης (Liu et al., 2021). Τα εργαλεία αυτά κατηγοριοποιούνται σε δύο μεθόδους, σε αυτήν της συσταδοποίησης/ομαδοποίησης (clustering) και της αποθορυβοποίησης (denoising). Αν και οι λειτουργίες αυτών έχουν κοινό σκοπό, η λογική τους, ο τρόπος εφαρμογής τους καθώς και τα παραγόμενα δεδομένα που προκύπτουν από αυτές τις μεθόδους διαφέρουν σημαντικά. Η κάθε τεχνική επεξεργασίας αμπλικονίων του 16S γονιδίου συνοδεύεται από τις δικά της πλεονεκτήματα και μειονεκτήματα και η επιλογή μίας συγκεκριμένης βιοπληροφορικής στρατηγικής εξαρτάται από την προσέγγιση της ολικής ανάλυσης.

1.4.3.1 Ομαδοποίηση

Οι πρώτες προσεγγίσεις για την ελαχιστοποίηση ψευδών αποτελεσμάτων αλληλούχισης είναι αυτές της ομαδοποίησης. Η τεχνική της ομαδοποίησης βασίζεται στην ιδέα ότι οι πολύ συγγενικοί μικροοργανισμοί παρέχουν παρόμοιες αλληλουχίες στο φυλογενετικό τους γονίδιο. Έτσι, εάν το ποσοστό ομοιότητας των αλληλουχιών συγκεκριμένων αναγνωσμάτων είναι πολύ μεγάλο, τα αναγνώσματα αυτά συγκεντρώνονται σε μία συστάδα στην οποία μοιράζονται μια συναινετική αλληλουχία αναγνώσματος. Στην προσπάθεια ομαδοποίησης εκατομμύριων αναγνωσμάτων, δημιουργούνται χιλιάδες συστάδες όπου η κάθε μία αντιπροσωπεύει την αφθονία μίας συγκεκριμένης παραλλαγής αμπλικονίου. Μία συστάδα ονομάζεται λειτουργική ταξινομική μονάδα (operating taxonomic unit, OTU) και υπάρχουν τρείς διαφορετικές μέθοδοι παραγωγής της, η ομαδοποίηση χωρίς δεδομένα αναφοράς, η ομαδοποίηση βάσει δεδομένων αναφοράς και ο συνδυασμός αυτών των δύο. Στον κλάδο της έρευνας μικροβιώματος, ανεξαρτήτως της μεθόδου ομαδοποίησης, έχει προκύψει μια σύμβαση όπου η ομαδοποίηση των αναγνωσμάτων σε OTUs επιτυγχάνεται χρησιμοποιώντας ένα ελάχιστο όριο ομοιότητας 97% ή μέγιστο όριο διαφοροποίησης 3%.

Η πιο εύκολα κατανοήσιμη και ταυτόχρονα πιο υπολογιστικά περίπλοκη μέθοδος παραγωγής OTUs είναι η συσταδοποίηση χωρίς την χρήση δεδομένων αναφοράς, ονομαζόμενη επίσης ως *de novo* ομαδοποίηση. Με αυτή την μέθοδο, οι αλληλουχίες των αναγνωσμάτων συγκρίνονται μεταξύ τους, και μετά από μια σειρά εφαρμογής επαναλαμβανόμενων αλγόριθμων ομαδοποίησης με ένα καθορισμένο όριο ομοιότητας, δημιουργούνται εκ νέου (*de novo*) συστάδες OTUs. Άρα, στην *de novo* ομαδοποίηση δεν απαιτείται βάση δεδομένων για την δημιουργία συμπλεγμάτων OTUs διότι προκύπτουν από την εσωτερική παρακολούθηση των δεδομένων αλληλούχισης. Ένα βασικό μειονέκτημα της μεθόδου είναι η βαριά υπολογιστική επεξεργασία η οποία δεν μπορεί να εφαρμοστεί παράλληλα, με αποτέλεσμα να συνοδεύεται από μεγάλο χρόνο επεξεργασίας για μεγάλα δεδομένα αλληλουχισης. Όταν ο όγκος των αναγνωσμάτων με ψευδείς αλληλουχίες είναι μεγάλος, η μέθοδος οδηγεί στην παραγωγή μεγάλου αριθμού μοναδικών OTUs με αποτέλεσμα να καταλαμβάνονται μεγάλες ποσότητες μνήμης και να εκτίνεται ο χρόνος επεξεργασίας κατά την εφαρμογή αυτής της μεθόδου. Επιπλέον, η de novo ομαδοποίηση πρέπει να επαναλαμβάνεται κάθε φορά που προστίθενται ή αφαιρούνται δεδομένα από την ανάλυση, διότι ένα ανάγνωσμα μπορεί να ομαδοποιείται διαφορετικά ανάλογα με το ποιες άλλες αλληλουχίες ανιχνεύθηκαν στη μελέτη.

Μία πιο υπολογιστικά αποδοτική μέθοδος παραγωγής OTUs είναι η ομαδοποίηση βάσει δεδομένων αναφοράς, ή αλλιώς κλειστής αναφοράς (closed-reference clustering). Όπως υπονοείται από την ονομασία της, η μέθοδος χρησιμοποιεί μία βάση δεδομένων αλληλουχιών από στοχευμένα γονίδια, των οποίων η ταξινομική προέλευσή τους είναι γνωστή, με σκοπό την σύγκριση αυτών με τα δεδομένα αλληλούχισης προς μελέτη. Με αυτόν

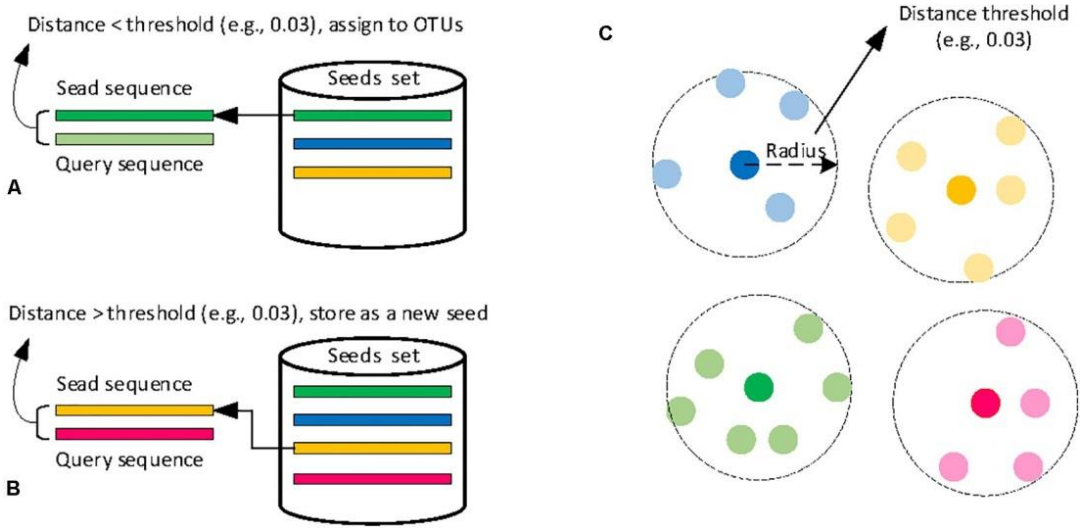
τον τρόπο ελαχιστοποιούνται τα σφάλματα αλληλουχιών διότι ένας μικρός αριθμός ψευδών SNVs είναι απίθανο να αλλάξει την τελική συναινετική αλληλουχία που αντιπροσωπεύει ένα ολόκληρο OTU. Επιπλέον, εάν ένα ανάγνωσμα παρέχει επαρκή σφάλματα στην αλληλουχία του και αποτύχει να ομαδοποιηθεί λόγω αυτού, η μέθοδος κλειστής αναφοράς θα αφαιρέσει το ανάγνωσμα εξ ολοκλήρου, μειώνοντας έτσι και τον χρόνο εκτέλεσής τους. Εκτός του ότι είναι υπολογιστικά γρήγορη, η μέθοδος επιτρέπει την εύκολη σύγκριση αποτελεσμάτων μεταξύ ερευνών που χρησιμοποιήθηκαν αντίστοιχες βάσεις δεδομένων αναφοράς. Παράλληλα, μπορεί να επιτρέψει την ταχεία ενσωμάτωση νέων δεδομένων στη μελέτη χωρίς να είναι απαραίτητη η εκ νέου ανάλυση των προηγούμενων αποτελεσμάτων.

Παρόλα αυτά, η μέθοδος της ομαδοποίησης κλειστής αναφοράς φέρει το μειονέκτημα ότι εξαρτάται πλήρως από τις αλληλουχίες που παρέχουν οι βάσεις δεδομένων και επομένως υπόκειται σε τυχόν σφάλματα ή μεροληψίες (biases) αυτών. Κατά την κατασκευή των δεδομένων αναφοράς, οι αλληλουχίες αναφοράς που επιλέγονται πρέπει να είναι λιγότερο από 97% παρόμοιες μεταξύ τους σε όλο το μήκος τους. Ωστόσο, οι πιο συχνά χρησιμοποιούμενες μεταβλητές περιοχές εντός του 16S rRNA γονιδίου δεν εξελίσσονται με τον ίδιο ρυθμό σε σχέση με αυτό του πλήρους μήκους. Έτσι, ένα ανάγνωσμα που αντιπροσωπεύει ένα θραύσμα του γονιδίου μπορεί να είναι περισσότερο από 97% παρόμοιο με πολλαπλές αλληλουχίες αναφοράς (Westcott & Schloss, 2015). Επιπλέον, υπάρχει η περίπτωση επιλεγμένη βάση δεδομένων να μην αντικατοπτρίζει επαρκώς τη βιοποικιλότητα του δείγματος προς ανάλυση. Εάν το δείγμα είναι ασυνήθιστο ή προέρχεται από μία εντελώς νέα πηγή, η βάση δεδομένων είναι πιθανό να μην περιλαμβάνει τις κατάλληλες αλληλουχίες για την αντιστοίχησή τους με τα δεδομένα αλληλούχισης. Έτσι, εάν ένας σημαντικός αριθμός αναγνωσμάτων παρέχει πρωτότυπες αλληλουχίες που δεν έχουν προκύψει από σφάλμα, τότε δεν θα μπορέσει να αντιστοιχιστεί σε ένα OTU, χάνοντας σημαντικό υλικό.

Ένα ακόμα πρόβλημα της προσέγγισης κλειστής αναφοράς είναι η ευκρίνεια των OTUs, διότι δύο αναγνώσματα μπορεί να είναι κατά 97% παρόμοια με μία αλληλουχία αναφοράς, αλλά μπορεί να είναι μόνο κατά 94% παρόμοια μεταξύ τους. Αυτό σημαίνει ότι υπάρχει κίνδυνος ομαδοποίησης πολλαπλών παρόμοιων ειδών σε ένα ενιαίο OTU, με τις ατομικές τους ταυτοποιήσεις να χάνονται στην περίληψη μίας ομάδας. Για την ελαχιστοποίηση της απώλειας της ποικιλομορφίας κατά την ομαδοποίηση, έχει δοκιμαστεί η αύξηση του ποσοστού ομοιότητας, τείνοντας στο 100%. Όμως, με αυτήν την εφαρμογή, αυξάνεται ο κίνδυνος αναγνώρισης εσφαλμένων αλληλουχιών ως νέα είδη, δίνοντας έτσι ψευδή αυξημένη ποικιλομορφία.

Η τρίτη μέθοδος ομαδοποίησης είναι ένας συνδυασμός των προαναφερόμενων και ονομάζεται ομαδοποίηση ανοιχτής αναφοράς (open-reference clustering). Αυτή η μέθοδος περιλαμβάνει την εφαρμογή ομαδοποίησης κλειστής αναφοράς που ακολουθείται από την *de novo* ομαδοποίηση για τα αναγνώσματα που δεν κατάφεραν να ομαδοποιηθούν βασιζόμενα σε αλληλουχίες αναφοράς. Θεωρητικά, η μέθοδος ομαδοποίησης ανοιχτής αναφοράς εκμεταλλεύεται τα δυνατά χαρακτηριστικά τόσο της κλειστής αναφοράς όσο και της *de novo* ομαδοποίησης. Πρακτικά, όμως, ο τρόπος των δύο μεθόδων με τον οποίο κατασκευάζονται οι συστάδες είναι σημαντικά διαφορετικός και συνεπώς ο συνδυασμός διαφορετικά ορισμένων OTUs μπορεί να δημιουργήσει προβλήματα.

Υπάρχουν διάφοροι αλγόριθμοι ομαδοποίησης και παραγωγής OTUs, με τους περισσότερους από αυτούς να εμπίπτουν στην κατηγορία των άπληστων αλγόριθμων [4] που ακολουθούν την μέθοδο ευρετικής ομαδοποίησης (heuristic clustering) (Bhat et al., 2019). Η άπληστη ευρετική ομαδοποίηση είναι μια μέθοδος διαμερισματικής ομαδοποίησης που λειτουργεί σε ένα συγκεκριμένο επίπεδο κάθε φορά (Wei et al., 2021). Η άπληστη ομαδοποίηση λειτουργεί επιλέγοντας πρώτα ένα τυχαίο ανάγνωσμα εισόδου ως αντιπροσωπευτικό, ονομαζόμενο επίσης ως σπόρος, ή αλλιώς φύτρα/δείκτης (seed), και στη συνέχεια κάθε επόμενο ανάγνωσμα εισόδου συγκρίνεται με το υπάρχον σύνολο σπόρων (**Σχήμα 1.18**). Εάν η αλληλουχία του αναγνώσματος εισόδου ταιριάζει με την αλληλουχία ενός από τους σπόρους σε ένα προκαθορισμένο επίπεδο ομοιότητας, θα προστεθεί στην συστάδα που αντιπροσωπεύεται από αυτόν τον σπόρο. Διαφορετικά, θα ληφθεί ως νέος σπόρος. Παραδείγματα αλγορίθμων αυτής της κατηγορίας είναι τα UPARSE (Edgar, 2013), USEARCH (Edgar, 2010) και VSEARCH (Rognes et al., 2016). Το VSEARSH είναι ένα ευέλικτο και δωρεάν βιοπληροφορικό εργαλείο παραγωγής OTUs που έχει σχεδιαστεί ως εναλλακτική λύση του USEARCH, του οποίου η χρήση γίνεται μόνο με πληρωμή.



Σχήμα 1.18 Διαγραμματική απεικόνιση μίας τυπικής ευρετικής μεθόδου ομαδοποίησης. (Α) Η ανάθεση αναγνώσματος σε ένα ήδη υπάρχων σπόρο (seed), (Β) η δημιουργία νέου σπόρου και (Γ) τα παραγόμενα OTUs. (Wei et al., 2021)

1.4.3.2 Αποθορυβοποίηση

Ενώ οι προσεγγίσεις ομαδοποίησης προσπαθούν να κατατάσσουν παρόμοιες αλληλουχίες αναγνωσμάτων σε μια αφηρημένη συναινετική αλληλουχία, ελαχιστοποιώντας έτσι την επίδραση τυχόν σφαλμάτων αλληλουχίας, οι προσεγγίσεις αποθορυβοποίησης έχουν εντελώς αντίθετη κατεύθυνση αντιμετώπισης σφαλμάτων. Ο σκοπός των αλγόριθμων αποθορυβοποίησης είναι η ανίχνευση αναγνωσμάτων που παρέχουν λανθασμένες αλληλουχίες και στην συνέχεια η προσπάθεια συγχώνευσης αυτών με τα αναγνώσματα που παρέχουν την σωστή «μητρική» αλληλουχία (Antich et al., 2021). Τα αντιπροσωπευτικά αναγνώσματα που προκύπτουν από αυτήν την μέθοδο ονομάζονται Παραλλαγές Αλληλουχιών Αμπλικονίων (Amplicon Sequence Variants, ASVs). Υπάρχουν διάφορες τεχνικές αποθορυβοποίησης που η κάθε μία έχει προκύψει από την ανάπτυξη ενός αλγόριθμου για την εκτέλεσή της. Στην πραγματικότητα, οι διάφοροι αλγόριθμοι

[4] Ένας άπληστος αλγόριθμος (greedy algorithm) είναι κάθε αλγόριθμος που ακολουθεί την ευρετική επίλυση προβλημάτων της τοπικής βέλτιστης επιλογής σε κάθε στάδιο.

αποθορυβοποίησης έχουν ονομάσει τα παράγωγά τους με διαφορετικές ορολογίες, όπως λειτουργικές ταξινομικές υπομονάδες (sub-OTUs), OTUs μηδενικής ακτίνας (sero-radius OTUs, ZOTUS) ή συγκεκριμένες/ακριβής παραλλαγές αλληλουχιών (Exact Sequence Variants, ESVs). Όμως, δεδομένου ότι είναι όλα ισοδύναμα και ο όρος ASV χρησιμοποιείται πολύ περισσότερο από τις υπόλοιπες ορολογίες, από εδώ και στο εξής τα παράγωγα των αλγόριθμοι αποθορυβοποίησης θα αναφέρονται με αυτόν τον τρόπο.

Οι μέθοδοι παραγωγής ASV συνάγουν τις βιολογικές αλληλουχίες στο δείγμα πριν από την εισαγωγή των σφαλμάτων της ενίσχυσης PCR και της αλληλούχισης και διακρίνουν παραλλαγές αλληλουχιών που διαφέρουν μόλις κατά ένα νουκλεοτίδιο. Ο τρόπος διάκρισης βιολογικών αλληλουχιών από ψευδείς αλληλουχίες βασίζεται εν μέρει στην προσδοκία ότι οι βιολογικές αλληλουχίες είναι πιο πιθανό να παρατηρηθούν επανειλημμένα σε αντίθεση με τις αλληλουχίες που περιέχουν σφάλματα. Συνεπώς, η αποθορυβοποίηση δεν μπορεί να εκτελεστεί ανεξάρτητα σε κάθε ανάγνωση. Η μικρότερη μονάδα δεδομένων από την οποία μπορούν να παραχθούν ASVs είναι τα δεδομένα αλληλούχισης ενός δείγματος. Τα δεδομένα αυτά συνδυάζονται σε ένα μοντέλο σφάλματος, επιτρέποντας τη σύγκριση παρόμοιων αναγνωσμάτων για τον προσδιορισμό της πιθανότητας ενός δεδομένου αναγνώσματος σε μια δεδομένη συχνότητα να μην οφείλεται σε σφάλμα αλληλούχισης. Μέσω στατιστικού ελέγχου, δημιουργείται ένα p-value για κάθε ακριβή αλληλουχία, όπου η μηδενική υπόθεση είναι ισοδύναμη με την ακριβή αλληλουχία να είναι συνέπεια του σφάλματος αλληλούχισης. Μετά από αυτόν τον υπολογισμό, τα αναγνώσματα φιλτράρονται σύμφωνα με ένα κατώτερο όριο εμπιστοσύνης, αφήνοντας πίσω μια συλλογή από ακριβείς αλληλουχίες με μια καθορισμένη στατιστική εμπιστοσύνη.

Εφόσον τα ASVs δεν δημιουργούνται από μεθόδους ομαδοποίησης και δεν βασίζονται σε δεδομένα αναφοράς, οι αλληλουχίες που προκύπτουν παρέχουν μεγάλη ακρίβεια και τα αποτελέσματα αυτών μπορούν εύκολα να συγκριθούν μεταξύ ερευνών μικροβιώματος που χρησιμοποιούν την ίδια στοχευμένη περιοχή του 16S rRNA γονιδίου. Επιπλέον, μια δεδομένη στοχευμένη αλληλουχία θα πρέπει πάντα να δημιουργεί το ίδιο ASV. Έτσι, ένα δεδομένο ASV, ως ακριβής αλληλουχία, μπορεί να συγκριθεί με δεδομένα αναφοράς με πολύ μεγαλύτερη απόδοση. Ταυτόχρονα, επιτρέπεται η ακριβέστερη ταυτοποίηση μέχρι και το επίπεδο είδους. Οι πιο δημοφιλείς τεχνικές, και αντίστοιχα αλγόριθμοι, αποθορυβοποίησης είναι αυτές του Διαχωριστίκού Αλγόριθμου Αποθορυβοποίησης Αμπλικονίων (Divisive Amplicon Denoising Algorithm 2, DADA2) (Callahan et al., 2016) και του Deblur (Amir et al., 2017). Οι βασικές διαφορές αυτών των αλγόριθμων εντοπίζονται στο μοντέλο σφάλματος που χρησιμοποιείται και στο μέγεθος του δείγματος αναγνωσμάτων που συλλέγεται για την κατασκευή του μοντέλου.

Ο αλγόριθμος του DADA2 χρησιμοποιεί ένα παραμετρικό μοντέλο το οποίο βασίζεται στην αφθονία των αναγνωσμάτων και στο κατά πόσο διαφέρουν αυτά με τις αλληλουχίες που εμφανίζονται λιγότερο συχνά, ενώ παράλληλα συμπεριλαμβάνει τις βαθμολογίες ποιότητας των βάσεων που παρέχουν τα αναγνώσματα. Αυτό το μοντέλο σφάλματος ποσοτικοποιεί τον ρυθμό λji με τον οποίο ένα αμπλικόνιο παράχθηκε από την πλατφόρμα αλληλούχισης και προσδιορίστηκε με την αλληλουχία i από τα δείγματα j συναρτήσει της σύνθεσης και της γενικής ποιότητας της αλληλουχίας. Στη συνέχεια, ένα ακόμα μοντέλο για τον αριθμό των επαναλαμβανόμενων παρατηρήσεων της αλληλουχίας i, παραμετροποιημένο από τον ρυθμό λji, χρησιμοποιείται για τον υπολογισμό ενός p-value με μηδενική υπόθεση ότι ο αριθμός των αναγνωσμάτων του αμπλικονίου της αλληλουχίας i, δηλαδή η αφθονία του, είναι συνεπής με το μοντέλο σφάλματος. Με βάση αυτές τις τιμές και

με ένα όριο πιθανότητας, ο αλγόριθμος αποφασίζει αν πρέπει να εκχωρηθούν μετρήσεις από ένα λιγότερο άφθονο, "προερχόμενο από λάθος" ανάγνωσμα σε μία πιο άφθονη, πραγματική αλληλουχία. Για την παρασκευή του παραμετρικού μοντέλου, ο αλγόριθμος επιλέγει αναγνώσματα από όλα τα δείγματα.

Σε αντίθεση με τον DADA2, ο Deblur εκτελεί την αποθορυβοποίηση ξεχωριστά για κάθε δείγμα. Ο τρόπος με τον οποίο επιτυγχάνει αυτό είναι η σύγκριση των αποστάσεων Hamming από ανάγνωσμα σε ανάγνωσμα με ένα προφίλ άνω ορίου σφάλματος σε συνδυασμό με έναν greedy αλγόριθμο. Αρχικά, οι αλληλουχίες ταξινομούνται κατά αφθονία και, στη συνέχεια, από την πιο άφθονη αλληλουχία, ο αριθμός των προβλεπόμενων αναγνωσμάτων που προέρχονται από σφάλματα αφαιρείται από τα γειτονικά αναγνώσματα βάσει την απόσταση Hamming. Η απόσταση Hamming μεταξύ δύο σειρών συμβόλων ίσου μήκους, όπου στην συγκεκριμένη περίπτωση αποτελούν δύο αναγνώσματα, είναι ο αριθμός των θέσεων στις οποίες τα αντίστοιχα σύμβολα, δηλαδή οι βάσεις, είναι διαφορετικά. Με άλλα λόγια, ο αλγόριθμος υπολογίζει τον ελάχιστο αριθμό αντικαταστάσεων που απαιτούνται για την αλλαγή της μιας συμβολοσειράς στην άλλη ή τον ελάχιστο αριθμό σφαλμάτων που θα μπορούσαν να έχουν μετατρέψει τη μία συμβολοσειρά στην άλλη. Κάθε αλληλουχία της οποίας η αφθονία πέφτει στο 0 κατά τη διάρκεια αυτής της αφαίρεσης, καταργείται από την λίστα έγκυρων αλληλουχιών. Συνεπώς, αλληλουχίες που δεν λαμβάνονται υπόψη ως έγκυρες (δηλαδή ο θόρυβος) αφαιρούνται. Μετά την εφαρμογή του Deblur, διατηρούνται μόνο τα αναγνώσματα που πιθανότατα παρουσιάστηκαν στο όργανο αλληλούχισης.

1.4.4 Εντοπισμός και διαχείριση χιμαιρικών αλλληλουχιών

Μέχρι και το στάδιο της αποθορυβοποίησης ή ομαδοποίησης για την παραγωγή ASVs ή OTUs αντίστοιχα, ο ποιοτικός έλεγχος στοχεύει κατά κύριο λόγο στην διαχείριση των παραγόμενων αναγνωσμάτων και σφαλμάτων που έχουν προκύψει από τα όργανα αλληλούχισης. Παρόλα αυτά, μετά από όλη την προαναφερόμενη επεξεργασία, υπάρχει μεγάλη πιθανότητα τα δεδομένα αλληλούχισης να εμπεριέχουν πειραματικά σφάλματα που δεν έχουν ακόμα αντιμετωπιστεί. Μία από τις κατηγορίες πειραματικών σφαλμάτων που προκύπτουν από την ανάλυση του 16S rRNA γονιδίου είναι οι χιμαιρικές αλληλουχίες (Mysara et al., 2015). Όπως έχει ήδη προαναφερθεί, οι χιμαιρικές αλληλουχίες αποτελούν παράγωγα της ενίσχυσης PCR των επιθυμητών περιοχών του φυλογενετικού δείκτη. Ο σχηματισμός χιμαιρικών αλληλουχιών συμπεριλαμβάνει την ενίσχυση πρόωρου τερματισμένου αμπλικονίου σε διαφορετικό θραύσμα βακτηριακού DNA τις ίδια εκχύλισης και την αντιγραφή του μέχρι και την ολοκλήρωση των επόμενων κύκλων PCR. Συνήθως, οι χιμαιρικές αλληλουχίες αποτελούνται από δύο φυλογενετικά διακριτές μητρικές αλληλουχίες (Chen et al., 2015), με μελέτες να υποδεικνύουν ποσοστά σχηματισμού χιμαιρικών άνω του 30% κατά τη διάρκεια ενίσχυσης αλληλουχιών με PCR από κλωνοποιημένα γονίδια 16S ή από μικτό βακτηριακό γονιδιωμα (Haas et al., 2011).

Αυτές οι αλληλουχίες μπορούν να προκαλέσουν επιπλοκές σε μεταγενέστερες αναλύσεις, όπως στην αναγνώριση και την ταξινόμηση των βακτηριακών ταξινομικών κατηγοριών και στην μελέτη των φυλογενετικών τους σχέσεων (Chen et al., 2015). Ένας σημαντικός αριθμός χιμαιρικών αλληλουχιών έχει εντοπιστεί σε βάσεις δεδομένων και παρά τις εκτεταμένες προσπάθειες επιμέλειας αυτών, πολλές ακόμα εξακολουθούν να υπάρχουν σε αυτές (Bhavesh Tiwarekar et al., 2023). Επιπλέον, η συμπερίληψη τέτοιων αλληλουχίων σε μεταγενέστερες αναλύσεις φέρει ως αποτελέσματα την τεχνητή αύξηση της ποικιλομορφίας

βιολογικών δειγμάτων, επειδή ερμηνεύονται ψευδώς ως μοναδικές αλληλουχίες που αντιπροσωπεύουν νέα είδη (Mysara et al., 2015). Παρόλο που οι ρυθμοί σχηματισμού χιμαιρικών αλληλουχιών μπορούν να μειωθούν πειραματικά, με τις περισσότερες προσπάθειες να έχουν κατευθυνθεί προς τη βελτίωση των βημάτων ενίσχυσης PCR και/ή προετοιμασίας δείγματος DNA, καμία μέθοδος δεν έχει αποδειχθεί για την πλήρη εξάλειψη αυτών των τεχνουργημάτων (Haas et al., 2011). Ως εκ τούτου, η ικανότητα αναγνώρισης χιμαιρικών αλληλουχιών κατά την βιοπληροφορική ανάλυση δεδομένων αλληλούχισης του 16S rRNA γονιδίου είναι κρίσιμη για την ορθή εξαγωγή συμπερασμάτων του προφίλ μικροβιακών κοινοτήτων. Αν και η ανίχνευση χιμαιρικών αλληλουχιών αμπλικονίων του 16S rRNA γονιδίου είναι ιδιαίτερα δύσκολη, καθώς οι αλληλουχίες είναι σύντομες και πολύ παρόμοιες (de la Cuesta-Zuluaga & Escobar, 2016), υπάρχουν αρκετοί διαθέσιμοι αλγόριθμοι που έχουν σχεδιαστεί για αυτόν τον σκοπό. Γενικά, διακρίνονται δύο κατηγορίες εργαλείων ανίχνευσης χιμαιρικών αλληλουχιών. Η πρώτη περιλαμβάνει εργαλεία που χρησιμοποιούν μεθόδους αναφοράς (reference-based), ενώ η δεύτερη, και μάλιστα πιο σύγχρονη, αφορά τα εργαλεία που εφαρμόζουν *de novo* μεθόδους (Haas et al., 2011).

Η μέθοδος που βασίζεται σε αναφορά κυρίως εξετάζει τις αλληλουχίες αναγνωσμάτων που ενδεχομένως αποτελούν χιμαιρικές έναντι μιας επιμελημένης βάσης δεδομένων αναφοράς χωρίς χιμαιρικές αλληλουχίες (Haas et al., 2011). Αυτή η προσέγγιση έχει εφαρμοστεί στην πρώτη γενιά εργαλείων ανίχνευσης χιμαιρικών αναγνωσμάτων, όπως το Pintail (Ashelford et al., 2005) και το Bellerophon (Huber et al., 2004). Μια σημαντική βελτίωση της προσέγγισης που βασίζεται σε αναφορά επιτεύχθηκε μέσω του εργαλείου ChimeraSlayer (Haas et al., 2011), το οποίο χρησιμοποιεί το 30% του άκρου κάθε αναγνώσματος ως βάση για την αναζήτησή του σε ένα συνόλου δεδομένων αναφοράς και την εύρεση της πλησιέστερης θυγατρικής του αλληλουχίας, αν υπάρχει. Το UCHIME (Edgar et al., 2011) αποτελεί ένα ακόμη εργαλείο ανίχνευσης χιμαιρικών αλληλουχιών, το οποίο βασίστηκε στην υλοποίηση του ChimeraSlayer (Haas et al., 2011), με την μόνη διαφορά να είναι ότι τα αναγνώσματα χωρίζονται σε τέσσερα τμήματα και στην συνέχεια αυτά τα τμήματα ξεχωριστά αναζητούνται αντίστοιχα σε μια βάση δεδομένων αναφοράς. Η *de novo* μέθοδος ανίχνευσης χιμαιρικών αλληλουχιών βασίζεται στην υπόθεση ότι οι μητρικές οποιασδήποτε χιμαιρικής αλληλουχίας έχουν περάσει τουλάχιστον έναν περισσότερο κύκλο ενίσχυσης PCR από τις χιμαιρικές αλληλουχίες. Αυτό σημαίνει ότι οι πιο σχετικά άφθονες αλληλουχίες σε ένα σύνολο δεδομένων αλληλούχισης είναι απίθανο να είναι χιμαιρικές, και επομένως, μπορούν να χρησιμοποιηθούν ως βάση αναφοράς (de la Cuesta-Zuluaga & Escobar, 2016). Παραδείγματα εργαλείων που βασίζονται στην de novo μέθοδο συμπεριλαμβάνουν το *de novo* UCHIME (Edgar et al., 2011) και το *de novo* ChimeraSlayer (Haas et al., 2011). Η χρήση του προαναφερόμενων μεθόδων στην ανάλυση δεδομένων αλληλούχισης ολοκληρώνεται συνήθως με την απόρριψη των χιμαιρικών αναγνωσμάτων από το σύνολο δεδομένων και δεν χρησιμοποιούνται σε μεταγενέστερες αναλύσεις.

Παρόλο που η αντικειμενική σύγκριση μεταξύ διαφορετικών εργαλείων ανίχνευσης χιμαιρικών αλληλουχιών είναι δύσκολη, καθώς ο καθένας από τους αλγόριθμους βασίζεται στα δικά του δεδομένα ελέγχου, είναι δυνατή η διάκριση της βέλτιστης επιλογής μεθόδου ανάλογα την προέλευση των δεδομένων αλληλούχισης (Mysara et al., 2015). Σε καταστάσεις που αφορούν δεδομένα που προέρχονται από δείγματα καλά μελετημένων περιβάλλοντων, οι προσεγγίσεις που βασίζονται σε αναφορά είναι πολύ αποτελεσματικές στη διάκριση μεταξύ χιμαιρικών και μη αλληλουχιών (μητρικές). Αντίθετα, οι αλγόριθμοι που εφαρμόζουν *de novo* μεθόδους έχουν το πλεονέκτημα της ικανότητας να ανιχνεύσουν χιμαιρικές αλληλουχίες ακόμα και αν τα βιολογικά δείγματα προς μελέτη παρέχουν μικροβιακή κοινότητα που μέχρι

τώρα δεν έχει μελετηθεί και περιγραφτεί καλά. Τα εργαλεία που εξαρτώνται από βάση αναφοράς βασίζονται σε συλλογές δεδομένων που συνήθως περιέχουν μόνο αλληλουχίες γονιδίων από καλλιεργημένα βακτήρια και δεν αναμένεται να έχουν την ίδια καλή απόδοση σε δείγματα που περιέχουν αλληλουχίες από ακόμη μη καλλιεργήσιμους οργανισμούς (de la Cuesta-Zuluaga & Escobar, 2016). Επομένως, προτείνεται η χρήση αλγορίθμων που εφαρμόζουν την *de novo* μέθοδο προκειμένου να ελαχιστοποιηθεί η διόγκωση της ποικιλομορφίας που προκαλείται από τις χιμαιρικές αλληλουχίες.

1.4.5 Ταξινομική και φυλογενετική ανάλυση

Η ταξινομική και η φυλογενετική ανάλυση αποτελούν τα πιο κρίσιμα βιοπληροφορικά στάδια σε ότι αφορά τις μελέτες μικροβιακών κοινοτήτων που βασίζονται στην αλληλούχιση του 16s rRNA γονιδίου. Αξιοποιώντας τις πληροφορίες που παρέχει ο φυλογενετικός δείκτης, γίνεται δυνατή η ταξινομική κατηγοριοποίηση (taxonomic classification) των αλληλουχιών του, η οποία περιλαμβάνει την ανάθεση αυτών σε γνωστές ταξινομικές ομάδες. Παράλληλα, μέσω της φυλογενετικής ανάλυσής του, παρέχεται μια βαθύτερη κατανόηση των εξελικτικών σχέσεων μεταξύ των διαφορετικών μικροοργανισμών που βρίσκονται στα δείγματα προς μελέτη. Με αυτόν τον τρόπο, επιτρέπεται ο προσδιορισμός των ειδών, της σχετικής αφθονίας καθώς και της εξελικτικής πορείας των βακτηρίων, των οποίων το γενετικό τους υλικό ήταν παρόν στο εξεταζόμενο δείγμα μικροβιώματος. Συνεπώς, ολοκληρώνοντας μια σειρά από βαριές υπολογιστικές διαδικασίες για τον ποιοτικό έλεγχο των πρωτογενών δεδομένων αλληλούχισης, το επόμενο βήμα της βιοπληροφορικής ανάλυσης του 16S rRNA γονιδίου είναι η διερεύνηση των ταξινομικών και φυλογενετικών πληροφοριών που παρέχουν τα ASVs ή OTUs.

Ανεξαρτήτως της μεθόδου κατασκευής τους, η ταξινομική ανάθεση των αντιπροσωπευτικών αμπλικονίων σε γνωστές ταξινομικές ομάδες, όπως τα είδη, τα γένη, τις οικογένειες, τις τάξεις, τις ομοταξίες και τα φύλα, επιτυγχάνεται με την σύγκριση των αλληλουχιών τους με μία βάση δεδομένων αναφοράς η οποία παρέχει αλληλουχίες με γνωστή ταξινομική προέλευση (Gao et al., 2017). Όμως, η εύρεση της πλησιέστερης αντιστοίχισης μίας άγνωστης αλληλουχίας με μία δεδομένη δεν επαρκεί για την ταυτοποίηση της. Αυτό οφείλεται στο γεγονός ότι υπάρχει περίπτωση δύο άγνωστες αλληλουχίες διαφορετικής ταξινομικής προέλευσης να αντιστοιχηθούν εξίσου καλά με την ίδια αλληλουχία από την βάση δεδομένων. Επομένως, χρειάζεται να προσδιοριστεί η πλησιέστερη ταξινομική συσχέτιση των αντιπροσωπευτικών αλληλουχιών με τις αλληλουχίες που παρέχονται από μια βάση δεδομένων με κάποιο βαθμό εμπιστοσύνης ή συναίνεσης. Για την επίτευξη αυτού, έχουν αναπτυχθεί διάφορα υπολογιστικά εργαλεία που εφαρμόζουν τεχνικές μηχανικής μάθησης (machine learning), γνωστά ως ταξινομικοί ταξινομητές (taxonomy classifiers), όπου ανάλογα την μέθοδο ταξινόμησης που χρησιμοποιούν, μπορούν να κατηγοριοποιηθούν ευρέως σε ταξινομητές που βασίζονται σε μεθόδους στοίχισης (alignment) και σε μεθόδους απαλλαγμένες από στοίχιση (alignment-free).

Μεταξύ αυτών, η πιο διαδεδομένη μέθοδος ταξινόμησης των αντιπροσωπευτικών αναγνωσμάτων αποτελεί η στοίχισή τους με αλληλουχίες γνωστής ταξινομικής προέλευσης που παρέχονται από μια βάση δεδομένων αναφοράς. Στόχος της μεθόδου στοίχισης είναι ο εντοπισμός όμοιων και διαφορετικών περιοχών μεταξύ αυτών. Στην συνέχεια, ακολουθεί η ταξινομική ανάθεση της άγνωστης αλληλουχίας σε μία αλληλουχία αναφοράς από την οποία προκύπτει η βέλτιστη στοίχιση, δηλαδή το βέλτιστο επίπεδο ομοιότητας τους. Παραδείγματα

βιοπληροφορικών εργαλείων που βασίζονται σε αυτήν την μέθοδο είναι το BLAST (Basic Local Alignment Search Tool) (Camacho et al., 2009) και το MEGAN (MEtaGenome ANalyzer) (Mitra et al., 2011). Ο λόγος για τον οποίο η μέθοδος στοίχισης παραμένει δημοφιλής και ευρέως χρησιμοποιούμενη προσέγγιση για την ταξινόμηση των αλληλουχιών του 16S rRNA γονιδίου είναι ότι μπορεί να παρέχει υψηλής ακρίβειας αποτελέσματα. Ωστόσο, η μέθοδος συνοδεύεται από υψηλή υπολογιστική πολυπλοκότητα για την ταξινόμηση δεδομένων αλληλούχισης νέας γενιάς, δεδομένου ότι είναι μεγάλος ο όγκος τους (Zielezinski et al., 2017).

Για την αντιμετώπιση αυτού του μειονεκτήματος, έχουν προταθεί άλλες προσεγγίσεις ταξινομικής ταξινόμησης, οι οποίες δεν βασίζονται στην στοίχιση αλληλουχιών. Αντίθετα, αυτές οι μέθοδοι χρησιμοποιούν συνήθως μια ποικιλία αριθμητικών χαρακτηριστικών που καταγράφουν διαφορετικές πτυχές της σύνθεσης ή της δομής της άγνωστης αλληλουχίας και στη συνέχεια εφαρμόζουν αλγόριθμους στατιστικής με σκοπό την ταξινόμησή της σε μια συγκεκριμένη ταξινομική ομάδα. Η πιο δημοφιλής μεταξύ τέτοιων μεθόδων βασίζεται στην ανάλυση συχνότητας των μοτίβων μικρών αλληλουχιών νουκλεοτιδίων μήκους k (Saha et al., 2019).

Τα τελευταία χρόνια, η χρήση μεθόδων στοίχισης ή χωρίς στοίχιση σε συνδυασμό με μεθόδων μηχανικής μάθησης γίνεται όλο και περισσότερο δημοφιλής στον τομέα της βιοπληροφορικής. Στο πλαίσιο της ταξινόμησης αλληλουχιών του 16S rRNA γονιδίου, μια κοινή προσέγγιση που βασίζεται στη μηχανική μάθηση είναι η εποπτευόμενη μάθηση, η οποία περιλαμβάνει την εκπαίδευση ενός μοντέλου σε ένα σύνολο επισημασμένων δεδομένων αλληλούχισης για την πρόβλεψη της ταξινόμησης νέων, άγνωστων αλληλουχιών με βάση τα χαρακτηριστικά τους, όπως η ομοιότητα, το μήκος ή η σύνθεση τους, καθώς και άλλες πληροφορίες στοίχισης ή μη-στοίχισης που προέρχονται από αυτές (Greener et al., 2022). Από την άλλη πλευρά, η μη-εποπτευόμενη μάθηση χρησιμοποιείται για τον εντοπισμό μοτίβων ή συστάδων εντός μη-επισημασμένων δεδομένων, δηλαδή ένα σύνολο αλληλουχιών που δεν έχουν ακόμη ταξινομηθεί στις ταξινομικές τους ομάδες, με σκοπό την πρόβλεψη της ταξινόμησης μιας νέας αλληλουχίας σε αυτά τα δεδομένα. Αυτή η τεχνική μπορεί να φανεί ιδιαίτερα χρήσιμη για τον προσδιορισμό νέων ταξινομικών ομάδων. Παραδείγματα ταξινομητών που βασίζονται σε τεχνικές μηχανικής μάθησης συμπεριλαμβάνουν τον αφελή Bayesian ταξινομητή (Naïve Bayes), την μηχανή διανυσματικής υποστήριξης (Support Vector Machine), τα δέντρα αποφάσεων (Decision Trees), τα οποία μπορούν τα επεκταθούν και στην εφαρμογή τυχαίων δασών (Random Forest), και τα νευρωνικά δίκτυα (Neural Networks), που εμπίπτουν στην κατηγορία της βαθιάς μάθησης (Mathieu et al., 2022). Εξειδικευμένα βιοπληροφορικά εργαλεία που χρησιμοποιούν τους προαναφερόμενους ταξινομητές για την ανάθεση ταξινομικών κατηγοριών των γονιδίων 16S rRNA αποτελούν οι SILVA (Quast et al., 2013) και RDP (Ribosomal Database Project) (Wang et al., 2007) ταξινομητές. Επιπλέον, αν και δεν χρησιμοποιείται συνήθως ως ειδικό εργαλείο για την ταξινομική ανάθεση των αλληλουχιών του 16S rRNA γονιδίου, το Scikit-learn (Pedregosa et al., 2011), ή αλλιώς sklearn, είναι μια ισχυρή βιβλιοθήκη μηχανικής μάθησης στην Python το οποίο παρέχει ένα ευρύ φάσμα αλγορίθμων και εργαλείων για την κατασκευή και την αξιολόγηση μοντέλων μηχανικής μάθησης. Εάν ο ερευνητής έχει την τεχνογνωσία και τους πόρους για την εφαρμογή του, το Scikit-learn μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλες βιβλιοθήκες της Python και με προσαρμοσμένα scripts για την ανάπτυξη ταξινομητών που βασίζονται σε μηχανική μάθηση για ταξινομική ανάθεση αλληλουχιών.

Παρά τη διαθεσιμότητα αυτών των ταξινομικών εργαλείων, η ταξινόμηση σε επίπεδο είδους των αλληλουχιών του 16S rRNA γονιδίου εξακολουθεί να παραμένει μια σοβαρή πρόκληση για τους ερευνητές του μικροβιώματος (Gao et al., 2017). Επιπλέον, αυτές οι τεχνικές είναι σε μεγάλο βαθμό ευαίσθητες στην ποιότητα και την πληρότητα των δεδομένων εκπαίδευσης, και κατά επέκταση των βάσεων δεδομένων, που χρησιμοποιούνται για την εφαρμογή τους. Λαμβάνοντας υπόψη ότι οι βάσεις δεδομένων χρησιμοποιούνται για την μετατροπή αλληλουχιών σε ευανάγνωστα ονόματα βακτηρίων, η επιλογή μίας αξιόπιστης βάσης δεδομένων αποτελεί ύψιστης σημασίας στο στάδιο της ταξινόμησης των αλληλουχιών του 16S rRNA γονιδίου.

Γενικά, υπάρχουν αρκετές διαθέσιμες βάσεις δεδομένων που παρέχουν ταξινομικές πληροφορίες του 16S rRNA γονιδίου, συμπεριλαμβανομένων των Greengenes (McDonald et al., 2012), RDP (Wang et al., 2007) και SILVA (Quast et al., 2013), οι οποίες είναι και οι πιο διαδεδομένες. Επί του παρόντος, η βάση δεδομένων SILVA είναι η πιο ολοκληρωμένη, παρέχοντας 436.680 ταξινομημένες αλληλουχίες του γονιδίου 16S rRNA. Ωστόσο, η ταξινόμηση σε επίπεδο είδους δεν επιτυγχάνεται σε μεγάλο βαθμό σε αυτήν τη βάση δεδομένων, διότι οι ταξινομήσεις ειδών εμπίπτουν συχνά στην κατηγορία του «μη καλλιεργημένου» (uncultured) ή του «μεταγονιδιώματος» (metagenome) (Trego et al., 2022). Αντίθετα, έχουν επίσης αναπτυχθεί ειδικές βάσεις δεδομένων οι οποίες είναι κατάλληλες προς χρήση ανάλογα από που έχουν προέλθει τα δεδομένα προς ανάλυση, με σκοπό την βελτίωση της ταξινομικής ευκρίνειας σε επίπεδο είδους. Παραδείγματα τέτοιων βάσεων δεδομένων αποτελούν η MiDAS (Dueholm et al., 2022), η οποία είναι ειδικά διαμορφωμένη για μικροβιώματα επεξεργασίας λυμάτων, η TaxAss (Rohwer et al., 2018), η οποία κατασκευάστηκε αντίστοιχα για μικροβιώματα γλυκού νερού, και η RefSoil (Choi et al., 2017), η οποία είναι αντίστοιχα για μικροβιώματα εδάφους.

Ολοκληρώνοντας το στάδιο της ταξινομική ανάθεσης του 16S rRNA γονιδίου, παρέχεται η δυνατότητα διερεύνησης των ταξινομικών πληροφοριών που εμπεριέχουν τα δείγματα προς ανάλυση. Αυτό επιτυγχάνεται με την δημιουργία οπτικών αναπαραστάσεων της βακτηριακής σύνθεσης των δειγμάτων μέσω γραφημάτων απλών ή στοιβαγμένων ράβδων και πίτας. Οι συγκεκριμένοι μέθοδοι απεικόνισης μπορούν να δείξουν τη σχετική αφθονία διαφορετικών ταξινομικών κατηγοριών σε διάφορα ταξινομικά επίπεδα (π.χ. φυλή, οικογένεια, κ.λπ.) που παρέχουν τα δείγματα, είτε το καθένα ξεχωριστά είτε σε ομάδες αυτών. Συνήθως, για την απλοποιημένη περιγραφή της βακτηριακής σύνθεσης των μικροβιωμάτων, οι ταξινομικές κατηγορίες παρουσιάζονται σε επίπεδο φυλής ή γένους στα προαναφερόμενα γραφήματα (Liu et al., 2021).

Καθορίζοντας τη σημασία της ταξινομικής ανάθεσης ως θεμελιώδης βήμα στη βιοπληροφορική ανάλυση των δεδομένων του 16S rRNA γονιδίου, η προσοχή στρέφεται στο επόμενο κρίσιμο στάδιο: τη φυλογενετική ανάλυση. Ενώ η ταξινομική ανάλυση παρέχει πληροφορίες για την ταυτότητα των βακτηριακών κοινοτήτων, η φυλογενετική ανάλυση εμβαθύνει στις εξελικτικές σχέσεις μεταξύ αυτών των μικροοργανισμών. Στο πλαίσιο της ανάλυσης 16S rRNA γονιδίων, η ιδέα της φυλογενετικής ανάλυσης είναι η χρήση των αλληλουχιών για την εκχώρηση συμπερασμάτων σχετικά με την εξελικτική απόσταση μεταξύ των διαφορετικών βακτηριακών ταυτοτήτων. Ο απώτερος σκοπός αυτού του βιοπληροφορικού σταδίου είναι η κατασκευή ενός φυλογενετικού δέντρου, δηλαδή ενός δικτύου στο οποίο το μήκος των κλάδων αντικατοπτρίζει την εξελικτική απόσταση μεταξύ των διαφορετικών αλληλουχιών, με τους μεγαλύτερους κλάδους να υποδηλώνουν

μεγαλύτερες αποστάσεις, και οι κόμβοι αντιπροσωπεύουν τον κοινό πρόγονο (Z. Yang & Rannala, 2012).

Συνήθως, ο τρόπος με τον οποίο επιτυγχάνεται η διερεύνηση των εξελικτικών σχέσεων διάφορων μικροοργανισμών είναι μέσω της μεθόδου πολλαπλής στοίχισης αλληλουχιών (Multiple Sequence Alignment – MSA), μια υπολογιστική τεχνική που ευθυγραμμίζει μαζικά αλληλουχίες γονιδίων από διάφορους μικροοργανισμούς μεταξύ τους. Με την μαζική στοίχιση των αλληλουχιών 16S rRNA γονιδίων, εντοπίζονται παρόμοιες και διαφορετικές περιοχές αυτών (Z. Yang & Rannala, 2012). Οι παρόμοιες, ή αλλιώς διατηρημένες, περιοχές υποδηλώνουν εξελικτική σταθερότητα, ενώ οι διαφορετικές, η αλλιώς αποκλίνουσες, περιοχές φανερώνουν στις γενετικές παραλλαγές. Έτσι, μέσω της στοίχισης, αναγνωρίζονται οι ομόλογες θέσεις εντός των γονιδιακών αλληλουχιών, οι οποίες υποδεικνύουν τις εξελικτικές συνδέσεις μεταξύ των βακτηρίων και προσδιορίζουν τον κοινό πρόγονο από το οποίο προήλθαν. Υπάρχουν διάφορα μαθηματικά μοντέλα που αξιοποιούν τις πληροφορίες που έχουν προκύψει από την διαδικασία στοίχισης για την κατασκευή φυλογενετικών δέντρων, συμπεριλαμβανομένων της μέγιστης πιθανοφάνειας (maximum likelihood) και την Μπεϋζιανή Συμπερασματολογία (Bayesian Inference). Παραδείγματα βιοπληροφορικών εργαλείων που εφαρμόζουν την μέθοδο πολλαπλής στοίχισης αλληλουχιών για την ανάλυση 16S rRNA γονιδίων είναι το Clustal Omega (Sievers & Higgins, 2018), το MAFFT (Katoh et al., 2002) και το MUSCLE (Edgar, 2004), ενώ διαθέσιμα εργαλεία για την κατασκευή φυλογενετικών δέντρων συμπεριλαμβάνουν το FastTree 2 (Price et al., 2010), το RAxML (Stamatakis, 2014) και το IQ-TREE (Nguyen et al., 2015). Το φυλογενετικό δέντρο που προκύπτει μπορεί να οπτικοποιηθεί χρησιμοποιώντας διάφορα εργαλεία λογισμικού, όπως το GraPhlAn (Asnicar et al., 2015) και το iTOL (Letunic & Bork, 2019).

Η φυλογενετική ανάλυση μπορεί να είναι υπολογιστικά εντατική, ειδικά για μεγάλα σύνολα δεδομένων, ή εξαιρετικά αποκλίνουσες αλληλουχίες (Z. Yang & Rannala, 2012). Παρόλα αυτά, μέσω της στοίχισης, μπορεί να παρέχει μία λεπτομερή εικόνα μιας μικροβιακής κοινότητας. Επιπλέον, μπορεί να χρησιμοποιηθεί για την αναγνώριση νέων ειδών και για τον προσδιορισμό της ταξινόμησης άγνωστων οργανισμών με βάση τις αλληλουχίες γονιδίου 16S rRNA τους, στην περίπτωση που μία αλληλουχία δεν είναι κατοχυρωμένη σε μια από τις διαθέσιμες βάσεις δεδομένων ταξινόμησης. Βέβαια, είναι αναγκαίο να σημειωθεί ότι η βασική λειτουργικότητα ενός φυλογενετικού δέντρου είναι στην εκμαίευση φυλογενετικών πληροφοριών που μπορούν να παρέχουν δεδομένα αλληλούχισης 16S rRNA γονιδίων για τον προσδιορισμό της βακτηριακής ποικιλομορφίας βιολογικών δειγμάτων (Trego et al., 2022). Η ανάλυση ποικιλομορφίας αποτελεί ένα μεταγενέστερο βιοπληροφορικό στάδιο, του οποίου ο ορισμός και ο ρόλος αναφέρεται σε παρακάτω υποενότητα.

1.4.6 Προσδιορισμός και αφαίρεση επιμολύνσεων

Ακόμα και στις πιο στείρες συνθήκες, η μη σκόπιμη εισαγωγή βακτηρίων κατά την δειγματοληψία, την εξαγωγή DNA και την ενίσχυση PCR στα δείγματα προς μελέτη αποτελεί πολλές φορές αναπόφευκτο συμβάν, και όπως η παραγωγή χιμαιρικών αλληλουχιών, χαρακτηρίζεται ως ένα βασικό πειραματικό σφάλμα της ανάλυσης του 16S rRNA γονιδίου (Mokhtari & Ridenhour, 2022). Οι επιμολύνσεις που εισάγονται στα δεδομένα αλληλούχισης χαρακτηρίζονται ως σημαντικό εμπόδιο στην ικανότητα να χαρακτηριστούν με ακρίβεια βακτηριακές κοινότητες σε περιβαλλοντικά και βιολογικά δείγματα, ειδικά εκείνων με

χαμηλή μικροβιακή βιομάζα, όπως αυτά που προέρχονται από την ουροδόχο κύστη και το αίμα (Karstens et al., 2019). Μολυσματικά βακτήρια που εισάγονται στο δείγμα πριν από την ενίσχυση PCR μπορούν να κυριαρχήσουν στη σύνθεση δειγμάτων χαμηλής μικροβιακής βιομάζας, όπου σε ακραίες περιπτώσεις μπορούν να αποτελούν πάνω από το 80% του δείγματος, υπονομεύοντας έτσι μια ανάλυση του 16S rRNA γονιδίου. Επιπλέον, οι επιμολύνσεις μπορούν να επηρεάσουν τα βιολογικά συμπεράσματα μιας μελέτης, όπως να διογκώσουν την ποικιλομορφία των δειγμάτων, να παραμορφώσουν την σχετική αφθονία των πραγματικών μικροβίων του περιβάλλοντος και να παραποιήσουν τις διαφορές μεταξύ κλινικών ομάδων.

Η πιθανότητα να προκύψουν επιμολύνσεις σε δεδομένα αλληλούχισης αμπλικονίων είναι πολύ υψηλή. Ως εκ τούτου, είναι απαραίτητο να εντοπιστούν, να ελαχιστοποιηθούν και να φιλτραριστούν μολυσματικές αλληλουχίες κατά την βιοπληροφορική ανάλυση δεδομένων αλληλούχισης για την αποφυγή εισαγωγή μεροληψιών στα τελικά αποτελέσματα. Επί του παρόντος, υπάρχουν δύο βασικές κατηγορίες προσεγγίσεων ελέγχου ή εξάλειψης πηγών επιμολύνσεων. Η πρώτη κατηγορία αφορά τις πειραματικές παρεμβάσεις σε συνδυασμό με την βιοπληροφορική διαχείριση αυτών για τον καθαρισμό των δεδομένων από επιμολύνσεις, ενώ η δεύτερη κατηγορία περιλαμβάνει στρατηγικές που χρησιμοποιούν μόνο την δύναμη της βιοπληροφορικής και των στατιστικών μεθόδων (Mokhtari & Ridenhour, 2022).

Οι πειραματικές προσπάθειες ελέγχου επιμολύνσεων αφορούν κυρίως την συμπερίληψη αρνητικών/τυφλών ή θετικών δειγμάτων ελέγχου για κάθε παρτίδα βιολογικών δειγμάτων, την χρήση τους σε ολόκληρη την πειραματική διαδικασία αλληλούχισης και την βιοπληροφορική ανάλυση τους σε σχέση με τα δείγματα προς μελέτη (Eisenhofer et al., 2019). Αν και δεν υπάρχει ακριβής διαδικασία για τον τρόπο αντιμετώπισης των αλληλουχιών που εμφανίζονται στα τεχνικά δείγματα, έχουν προταθεί τρόποι διαχείρισης αυτών των δειγμάτων. Μια προσέγγιση είναι η σύγκριση της σχετικής ταξινομικής αφθονίας ή ποικιλομορφίας των πραγματικών βιολογικών δειγμάτων με τα δείγματα ελέγχου (de la Cuesta-Zuluaga & Escobar, 2016). Έτσι, εάν τα πραγματικά δείγματα παρέχουν παρόμοιες ταξινομικές αφθονίες ή γενικά ποικιλομορφίες με τα δείγματα ελέγχου, τότε υπάρχουν σοβαρές ενδείξεις επιμόλυνσης κατά την πειραματική επεξεργασία των δειγμάτων. Μία άλλη πρόταση είναι η αφαίρεση των επιπρόσθετων αλληλουχιών που εντοπίστηκαν στα δείγματα ελέγχου από τα βιολογικά δείγματα (Karstens et al., 2019). Ωστόσο, αυτή η προσέγγιση μπορεί να είναι πολύ αυστηρή και να οδηγήσει στην αφαίρεση βακτηριακών ταξινομήσεων που στην πραγματικότητα είναι βιολογικά συσχετιζόμενα με τα δείγματα προς ανάλυση.

Ένα από τα πλεονεκτήματα της αλληλούχισης δειγμάτων ελέγχου είναι η ικανότητα ανίχνευσης και ποσοτικοποίησης των επιμολύνσεων καθώς και τον προσδιορισμό της προέλευσής τους (Mokhtari & Ridenhour, 2022). Παρόλα αυτά, η συμπερίληψη κατάλληλων αρνητικών ή θετικών δειγμάτων ελέγχου δεν είναι πάντα εύκολη και η εγκυρότητά τους για μια συγκεκριμένη έρευνα μικροβιώματος δεν μπορεί πάντα να εξασφαλιστεί και ενδέχεται να μην είναι διαθέσιμα τυποποιημένα πρωτόκολλα για το σχεδιασμό αυτών των δειγμάτων. Επιπλέον, κανένα δείγμα ελέγχου δεν είναι ικανό να εξαλείψει πλήρως τις ενδεχόμενες επιμολύνσεις εύκολα και αξιόπιστα σε όλες τις περιπτώσεις. Αυτό οφείλεται στο γεγονός ότι υπάρχουν διάφορα είδη επιμολύνσεων που θα μπορούσαν να προκύψουν στα δεδομένα αλληλούχισης που δεν αντικατοπτρίζονται ξεκάθαρα στα δείγματα ελέγχου.

Δεδομένου ότι η αξιοποίηση των δειγμάτων ελέγχου δεν επαρκεί στην ολική αφαίρεση των επιμολύνσεων που μπορούν να προκύψουν στα δεδομένα αλληλούχισης, έχουν προταθεί αρκετές βιοπληροφορικές τεχνικές για την αντικειμενική απομάκρυνση των

μολυσματικών στοιχείων (Mokhtari & Ridenhour, 2022). Η πιο εύκολη μέθοδος αφαίρεσης πιθανών επιμολύνσεων είναι η απομάκρυνση των μοναδικών ή πολύ χαμηλής σχετικής αφθονίας ASVs/OTUs ή ταξινομικών μονάδων από τα δεδομένα αλληλούχισης. Πράγματι, είναι πολύ συνηθισμένο να παρουσιάζονται σπάνιες ταξινομικές κατηγορίες των οποίων η παρουσία στα περισσότερα δείγματα είναι μηδενική (Cao et al., 2020). Μελέτες ποιοτικού ελέγχου δεδομένων μικροβιώματος υποδεικνύουν ότι πολλές σπάνιες ταξινομικές μονάδες προκαλούνται από τεχνουργήματα/σφάλματα ή/και από επιμολύνσεις. Για αυτόν τον λόγο, παρατηρείται συχνά στις αναλύσεις του 16S rRNA γονιδίου η εφαρμογή φίλτρου αφθονίας, στην οποία αφαιρούνται αλληλουχίες ή ταξινομικές μονάδες με συνολικό αριθμό αναγνωσμάτων ή ποσοστό μικρότερο από ένα εμπειρικό όριο σε όλα τα δείγματα (Karstens et al., 2019), με το όριο να είναι συνήθως ≤1% (Reitmeier et al., 2021). Ωστόσο, αυτή η μέθοδος προϋποθέτει ότι όλες οι μικροβιακές επιμολύνσεις έχουν χαμηλή σχετική αφθονία, κάτι που μπορεί να μην ισχύει ιδιαίτερα για δείγματα χαμηλής μικροβιακής μάζας (Karstens et al., 2019). Έτσι, αφαιρούνται επίσης όλες οι μη μολυσματικές αλληλουχίες κάτω από αυτό το όριο, οδηγώντας στην απώλεια σημαντικών πληροφοριών. Επιπλέον, δεν υπάρχει σαφή όριο αφθονίας, καθώς εμπειρικά υπάρχει ομοφωνία στο ποσοστό ≤1%, και αυθαίρετα έχουν επιλεχτεί σε διάφορες έρευνες μικρότερα ποσοστά.

Αναγνωρίζοντας αυτά τα μειονεκτήματα, οι δημιουργοί του decontam (Davis et al., 2018), ενός ανοιχτού κώδικα πακέτου της R, κατασκεύασαν ένα βιοπληροφορικό εργαλείο για τον εντοπισμό και την αφαίρεση εξωγενής επιμόλυνσης από τα δεδομένα αλληλούχισης χρησιμοποιώντας στατιστικά μοντέλα. Το decontam εφαρμόζει δύο απλές μεθόδους *de novo* ταξινόμησης που βασίζονται σε ευρέως αναπαραγόμενα χαρακτηριστικά επιμόλυνσης: (α) Οι αλληλουχίες από μολυσματικές ταξινομικές κατηγορίες είναι πιθανό να έχουν συχνότητες που συσχετίζονται αντιστρόφως με τη συγκέντρωση του DNA δείγματος και (β) οι αλληλουχίες από μολυσματικές ταξινομικές κατηγορίες είναι πιθανό να έχουν υψηλότερο επιπολασμό στα δείγματα ελέγχου από ότι στα πραγματικά βιολογικά δείγματα. Για την εφαρμογή αυτού του εργαλείου απαιτούνται τα μεταδεδομένα ποσοτικοποίησης DNA που στις περισσότερες περιπτώσεις είναι εγγενή στην προετοιμασία των δειγμάτων προς αλληλούχιση ή τα δεδομένα αρνητικών δειγμάτων ελέγχου. Αυτή η προσέγγιση είναι στενά συνδεδεμένη αλλά δεν ταυτίζεται με την απλή μέθοδο φιλτραρίσματος. Ένας σημαντικός περιορισμός του decontam είναι η υπόθεση ότι οι επιμολύνσεις και οι πραγματικές βιολογικές αλληλουχίες των δειγμάτων είναι διακριτές μεταξύ τους, παραβιάζοντας έτσι την περίπτωση της διασταυρούμενης επιμόλυνσης λόγω της αλληλούχισης συγκεντρωμένων δειγμάτων (Mokhtari & Ridenhour, 2022).

Μια ακόμη κατηγορία επιμολύνσεων που δεν αντιμετωπίζονται από τις προαναφερόμενες μεθόδους είναι οι ταξινομικές κατηγορίες που χαρακτηρίζονται βιολογικά μη αναμενόμενες για τα δείγματα που αναλύονται (Salter et al., 2014). Μία καλή ένδειξη παρουσίας τέτοιων στοιχείων είναι η σχετική αφθονία τους στα τελικά αποτελέσματα. Γενικά, η παρουσία αλληλουχιών που έχουν προκύψει από επιμολύνσεις σε δεδομένα αλληλούχισης είναι μεγαλύτερη σε δείγματα χαμηλής βιομάζας, όπως του αίματος και του πνεύμονα, σε σχέση με δείγματα υψηλής βιομάζας, όπως του εντέρου. Αυτό σημαίνει ότι υπάρχει μεγάλη πιθανότητα σε δείγματα χαμηλού μικροβιακού φορτίου να κυριαρχήσει μολυσματικό DNA, και κατά επέκταση να εμφανιστεί αυτό ταξινομημένο με έντονα υψηλή σχετική αφθονία σε κάποια, αν όχι όλα, δείγματα. Συνεπώς, εάν ένας ερευνητής συναντήσει μια τέτοια περίπτωση στα δεδομένα αλληλούχισής του, καλείται να προσδιορίσει και να αφαιρέσει χειροκίνητα αυτές τις αλληλουχίες, που όμως έχουν ήδη αναγνωριστεί ως

επιμολύνσεις σε δημοσιευμένες βάσεις δεδομένων ή βιβλιογραφικές αναφορές (Karstens et al., 2019).

Η τελευταία κατηγορία επιμολύνσεων που μπορούν να εντοπιστούν σε δεδομένα αλληλούχισης φυλογενετικών αμπλικονίων είναι οι αλληλουχίες που δεν αντικατοπτρίζουν την ανάλυση που εκτελείται. Λόγω της φύσης του 16S rRNA γονιδίου, είναι πιθανό να ενισχυθούν και ταξινομηθούν αλληλουχίες που προέρχονται από μιτοχόνδρια, χλωροπλάστες και ευκαρυωτικά κύτταρα (de la Cuesta-Zuluaga & Escobar, 2016). Αυτά τα στοιχεία θα πρέπει να αφαιρεθούν χειροκίνητα από τα δεδομένα ανάλυσης, καθώς και οι βακτηριακές ή οι αλληλουχίες αρχαίων σύμφωνα με το εύρος της μελέτης και τους εκκινητές που χρησιμοποιούνται στην έρευνα. Επιπλέον, εάν η ανάλυση αφορά αυστηρά την ταξινομική διερεύνηση βιολογικών δειγμάτων, οι αλληλουχίες που δεν κατάφεραν να ταξινομηθούν σε τουλάχιστον επίπεδο φυλής ή γενικά (unassigned) θα πρέπει να αφαιρεθούν από τα δεδομένα.

1.4.7 Ανάλυση ποικιλομορφίας

Απώτερος σκοπός μίας μεταγονιδιωματικής έρευνας, ιδιαίτερα στο πεδίο της ανάλυσης του 16S rRNA γονιδίου, είναι η διερεύνηση και κατανόηση των μικροβιακών συστημάτων καθώς και της συσχέτισή τους με το περιβάλλον από το οποίο συλλέχθηκαν. Ωστόσο, η ταξινομική ανάλυση, δηλαδή η απλή αναγνώριση και η σχετική αφθονία των βακτηρίων που εμπεριέχουν τα δείγματα προς μελέτη, δεν διαφωτίζει επαρκώς την πολυπλοκότητα των μικροβιωμάτων και την επιρροή τους από περιβαλλοντικούς παράγοντες. Συνεπώς, είναι αναγκαία μια ακόμη βιοπληροφορική επεξεργασία των δεδομένων, δεδομένα που έχουν προκύψει ύστερα από έναν εκτενή ποιοτικό έλεγχο, ώστε να επιλυθούν ερωτήματα που αφορούν την πολυπλοκότητα, την κατανομή και την δομή μικροβιακών κοινοτήτων. Το βιοπληροφορικό στάδιο σε μία ανάλυση του 16S rRNA γονιδίου που παρέχει τέτοιου είδους πληροφορίες είναι η ανάλυση ποικιλομορφίας, ή αλλιώς ποικιλότητας. Αυτό το στάδιο στοχεύει στον ποσοτικό προσδιορισμό και χαρακτηρισμό της ποικιλότητας βακτηριακών κοινοτήτων μέσω την μετατροπής των αλληλουχιών που παρέχουν τα ASVs ή OTUs και την συχνότητα με την οποία παρουσιάζονται στα δεδομένα σε κατάλληλες αριθμητικές μετρήσεις ποικιλομορφίας (diversity metrics) (Regueira-Iglesias et al., 2023). Όμως, η μοναδικότητα ενός βακτηριακού είδους και η πολυπλοκότητα ενός μικροβιώματος που αποτυπώνεται στα δεδομένα αλληλούχισης παρουσιάζουν μεγάλες προκλήσεις στην ανάλυση ποικιλομορφίας και ερμηνεία των αποτελεσμάτων. Για αυτόν τον λόγο, απαιτούνται ειδικές τεχνικές για την ανάλυση τέτοιων πολύπλοκων δεδομένων, οι οποίες αναπτύσσονται ακόμα και σήμερα, παρά την έντονη δραστηριότητα που υπάρχει στον κλάδο στην μικροβιωματικής έρευνας.

Η ανάλυση ποικιλομορφίας περιλαμβάνει τέσσερα βασικά βήματα, τα οποία σε σειρά είναι η αποτίμηση του βάθους αλληλούχισης (sequencing depth), ο υπολογισμός των μετρήσεων ποικιλομορφίας, η κατασκευή κατάλληλων διαγραμμάτων για την οπτικοποίηση των αποτελεσμάτων και την εφαρμογή στατιστικού ελέγχου, όπου αυτή επιθυμείται. Παρακάτω, αναφέρεται η κύρια μεθοδολογία για την ανάλυση ποικιλομορφίας μικροβιωμάτων από δεδομένα αλληλούχισης του γονιδίου 16S rRNA, αυτή της α-ποικιλομορφίας (alpha diversity). Ωστόσο, πριν από οποιαδήποτε διερεύνηση, είναι πολύ σημαντική η σωστή επιλογή των δειγμάτων καθώς και της ποσότητα δεδομένων που θα χρησιμοποιηθούν για περαιτέρω ανάλυση.

Το πρώτο και βασικό ζήτημα που πρέπει να αντιμετωπιστεί στην ανάλυση ποικιλομορφίας είναι η επιλογή του βάθους αναγνωσμάτων των δεδομένων αλληλούχισης.

Αυτό σημαίνει ότι πρέπει να επιλεγεί ένας συγκεκριμένος αριθμός αναγνωσμάτων που να καλύπτει την ποικιλομορφία όλων των δειγμάτων. Όμως, ο αριθμός αναγνωσμάτων μπορεί να ποικίλει μεταξύ των δειγμάτων λόγω των διακυμάνσεων στη συλλογή δειγμάτων, την αποθήκευση τους, την εξαγωγή DNA και την προετοιμασία της βιβλιοθήκης (Weinroth et al., 2022). Είναι σημαντικό να ληφθούν υπόψη αυτές οι διαφορές, έτσι ώστε να υπάρχει αντικειμενικότητα στην ερμηνεία μικροβιακών συνθέσεων και στην σύγκριση αυτών μεταξύ των δειγμάτων ή/και πειραμάτων. Επομένως, η σωστή κανονικοποίηση, η οποία χαρακτηρίζεται ως μια διαδικασία μετασχηματισμού για τη διόρθωση της μεταβλητότητας του μεγέθους των δεδομένων κάθε δείγματος, είναι κρίσιμη για τη διασφάλιση της εγκυρότητας μίας ανάλυσης ποικιλομορφίας.

Η βασική μέθοδος κανονικοποίησης είναι η αραίωση των δεδομένων (rarefaction), κατά την οποία αξιολογείται η επάρκεια των αλληλουχιών στα δείγματα και εκτιμάται ένας αναμενόμενος αριθμός παρατηρούμενων ASVs ή OTUs σε ένα δεδομένο βάθος αλληλούχισης (Regueira-Iglesias et al., 2023). Ο τρόπος με τον οποίο επιτυγχάνεται αυτό είναι μέσω της τυχαίας επιλογής υποσυνόλων αλληλουχιών από κάθε δείγμα, παράγοντας έτσι υποδείγματα. Η διαδικασία αυτή επαναλαμβάνεται, αυξάνοντας σταδιακά τον αριθμό αλληλουχιών που παρέχει το κάθε υπόδειγμα, μέχρι να ολοκληρωθεί η συλλογή όλων των αναγνωσμάτων που εμπεριέχουν τα δείγματα. Στην συνέχεια, κατασκευάζονται οι καμπύλες αραίωσης (rarefaction curves), οι οποίες αναπαριστώνται μέσω της γραφικής παράστασης του αριθμού των παρατηρούμενων ASVs ή OTUs έναντι του βάθους αλληλούχισης, δηλαδή του αριθμού αναγνωσμάτων υποδειγματοληψίας. Καθώς αυξάνεται σταδιακά το βάθος δειγματοληψίας, ο αριθμός των παρατηρούμενων ASVs ή OTUs σε κάθε δείγμα αυξάνεται έως ότου σταθεροποιηθεί. Η σταθεροποίηση του αριθμού των παρατηρούμενων ASVs ή OTUs διαπιστώνεται όταν οι καμπύλες αραίωσης επιπεδοποιηθούν, η αλλιώς όταν φτάσουν χαρακτηριστικά σε πλατώ. Το πλατώ μιας τέτοιας καμπύλης υποδηλώνει ότι η επιπλέον προσθήκη αναγνωσμάτων κατά την υποδειγματοληψία είναι απίθανο να αποκαλύψει πολλά περισσότερα ASVs ή OTUs, υποδεικνύοντας ότι το βάθος δειγματοληψίας αλληλουχιών στο οποίο εμφανίζεται αυτό είναι επαρκής.

Συνεπώς, η κατασκευή του διαγράμματος των καμπύλων αραίωσης του κάθε δείγματος αποτελεί σημαντικό βήμα για την κανονικοποίηση των δεδομένων, διότι παρέχουν πληροφορίες για το εάν ένα συγκεκριμένο βάθος αλληλούχισης είναι επαρκές για να συλλάβει την ποικιλομορφία που υπάρχει σε ένα δείγμα και παράλληλα βοηθούν στη λήψη αποφάσεων σχετικά με το εάν απαιτείται η προσθήκη αλληλουχιών για να αποκαλυφθεί η πραγματική ποικιλομορφία των δειγμάτων (Regueira-Iglesias et al., 2023). Η επιλογή του βάθος αλληλούχισης μέσω της διαδικασίας αραίωσης περιλαμβάνει έναν συμβιβασμό μεταξύ της διατήρησης όσο το δυνατόν περισσότερων δεδομένων και της επίτευξης ενός κοινού βάθους αλληλούχισης για την ανάλυσης ποικιλομορφίας. Είναι ζωτικής σημασίας να επιτευχθεί μια ισορροπία, καθώς η πολύ έντονη αραίωση μπορεί να οδηγήσει σε απώλεια δεδομένων, ενώ η πολύ μικρή μπορεί να προκαλέσει σφάλματα λόγω της ανικανότητας παρατήρησης των πληροφοριών που παρέχουν τα δείγματα. Δεδομένου ότι είναι αναγκαία η επιλογή ενός κοινού βάθους αλληλούχισης μεταξύ των διαφορετικών δειγμάτων, μέσω του διαγράμματος επιλέγεται ο κοινός αριθμός αναγνωσμάτων στο σημείο που στα περισσότερα δείγματα οι καμπύλες αραίωσης πιάνουν σε πλατώ. Όμως, λόγω των έντονων διακυμάνσεων του αριθμού αναγνωσμάτων παρέχουν συνήθως τα δείγματα, υπάρχει μεγάλη πιθανότητα το σημείο αυτό να ξεπερνάει τον αριθμό αναγνωσμάτων που εμπεριέχουν κάποια δείγματα, με αποτέλεσμα να απορρίπτονται συνολικά τα δεδομένα τους στην περαιτέρω ανάλυση ποικιλομορφίας (Weinroth et al., 2022). Έτσι, το βάθος αλληλούχισης επιλέγεται συνήθως με

βάση τον μικρότερο αριθμό αναγνωσμάτων που παρατηρείται στο σύνολο των δειγμάτων για να αποφευχθεί η απώλεια δεδομένων.

Εφόσον επιλεχθεί το κατάλληλο βάθος αλληλούχισης των δεδομένων, η διαδικασία κανονικοποίησης ολοκληρώνεται με τον μετασχηματισμό των αφθονιών που παρέχουν τα παρατηρούμενα ASVs ή OTUs με σκοπό να διασφαλιστεί ότι τα συνολικά αναγνώσματα του κάθε δείγματος γίνονται αντικειμενικά συγκρίσιμα στις περεταίρω αναλύσεις ποικιλομορφίας (Regueira-Iglesias et al., 2023). Αρχικά, προσδιορίζονται οι σχετικές αφθονίες των παρατηρούμενων ASVs ή OTUs του κάθε δείγματος μέσω των λόγων αφθονίας. Οι λόγοι αφθονίας υπολογίζονται διαιρώντας την αφθονία του κάθε ASVs ή OTUs σε εάν δείγμα με τον συνολικό αριθμό αναγνωσμάτων σε αυτό το δείγμα. Έπειτα, οι λόγοι αφθονίας κλιμακώνονται πολλαπλασιάζοντάς τους με έναν σταθερό παράγοντα, όπως τη διάμεσο ή τον μέσο όρο των συνολικών αναγνώσεων σε όλα τα δείγματα. Οι κλιμακωμένες σχετικές αφθονίες χρησιμοποιούνται στη συνέχεια ως κανονικοποιημένα δεδομένα για τον υπολογισμό των μετρήσεων ποικιλομορφίας.

Οι βασικές μετρήσεις ποικιλομορφίας για την περιγραφή της μικροβιακής σύνθεσης ενός μικροβιώματος χρησιμοποιώντας δεδομένα αλληλούχισης από το γονίδιο του 16S rRNA είναι αυτές της άλφα ποικιλομορφίας, οι οποίες δεν παρέχουν πληροφορίες για τις αλλαγές στην αφθονία συγκεκριμένων μικροβιακών ειδών, αλλά επιτρέπουν την παρατήρηση μιας ευρύτερης αλλαγής ή διαφοράς στην σύνθεση των δειγμάτων (Qian et al., 2020). Συγκεκριμένα, η α-ποικιλομορφία αναφέρεται στην περιγραφή της ποικιλομορφίας μέσα σε ένα δείγμα. Οι μετρήσεις της α-ποικιλομορφίας αντικατοπτρίζουν τον πλούτο (richness) ή την ομοιομορφία (eveness) ενός μικροβιακού δείγματος ή στοχεύουν να αντικατοπτρίζουν έναν συνδυασμό και των δύο αυτών χαρακτηριστικών. Ο πλούτος αναφέρεται στον συνολικό αριθμό των ειδών ή των ASVs/OTUs σε ένα δείγμα, ενώ η ομοιομορφία αναφέρεται στην σχετική αφθονία αυτών, δηλαδή στον αριθμό των αναγνωσμάτων που αντιπροσωπεύουν τα είδη βακτηρίων ή ASVs/OTUs. Τα δεδομένα που χρησιμοποιούνται για τον υπολογισμό αυτών των μετρήσεων έχουν συνήθως τη μορφή ενός πίνακα, όπου οι σειρές αντιπροσωπεύουν τα δείγματα και οι στήλες αντιπροσωπεύουν τις ταξινομικές μονάδες ή ASVs/OTUs που παρατηρούνται στο προεπιλεγμένο βάθος αλληλούχισης. Κάθε κελί στον πίνακα περιέχει την αφθονία ως κανονικοποιημένη τιμή μιας συγκεκριμένης ταξινομικής μονάδας ή ASV/OTU σε ένα συγκεκριμένο δείγμα (Xia et al., 2018).

Υπάρχουν διάφοροι δείκτες α-ποικιλομορφίας που χρησιμοποιούνται συχνά στην μικροβιακή έρευνα, συμπεριλαμβανομένου του δείκτη Chao 1 (Chao, 1984) και του δείκτη Simpson (SIMPSON, 1949). Μεταξύ αυτών, ο πιο διαδεδομένος αποτελεί ο δείκτης εντροπίας Shannon (Xia et al., 2018). Ο δείκτης εντροπίας Shannon (H') (Xia et al., 2018), γνωστός και ως δείκτης Shannon-Wiener, είναι ένα μέτρο ποικιλομορφίας που λαμβάνει υπόψη τόσο τον πλούτο όσο και την ομοιομορφία ενός δείγματος. Ο τύπος για τον υπολογισμό του δείκτη ποικιλομορφίας Shannon έχει ως εξής:

$$H' = -\sum_{i=1}^{S} p_i \cdot \ln(p_i)$$

όπου S είναι ο αριθμό των διαφορετικών ταξινομικών μονάδων ή ASVs/OTUs και pi είναι το ποσοστό των αναγνωσμάτων που ανήκουν στο i είδος ή ASV/OTU σε σχέση με τον συνολικό αριθμό αναγνωσμάτων. Ο δείκτης εντροπίας Shannon παρέχει μια θετική αριθμητική τιμή που αντιπροσωπεύει ολικά την ποικιλομορφία ενός δείγματος, ενώ δίνει μεγαλύτερη βαρύτητα στα σπάνια είδη ή ASV/OTU, κάτι που σημαίνει ότι η τιμή αυξάνεται

όταν αυξάνεται ο αριθμός των σπάνιων ειδών. Συνεπώς, όσο υψηλότερη είναι η τιμή του, τόσο πιο άφθονη είναι η α-ποικιλομορφία ενός δείγματος.

Ολοκληρώνοντας τον υπολογισμό των δεικτών ποικιλομορφίας που επιθυμούνται, το επόμενο βήμα της ανάλυσης ποικιλομορφίας είναι η οπτικοποίηση των αποτελεσμάτων μέσω της κατασκευής κατάλληλων γραφημάτων. Η σημασία της γραφικής απεικόνισης των μετρήσεων ποικιλομορφίας έγκειται στην ικανότητα μετατροπής πολύπλοκων αριθμητικών δεδομένων σε ουσιαστικές πληροφορίες, βοηθώντας στην εξερεύνηση της δυναμικής των μικροβιακών κοινοτήτων, στην ερμηνεία μοτίβων και στην αποτελεσματική μετάδοση των ευρημάτων, τόσο στο εξειδικευμένο όσο και στο ευρύ κοινό (Peeters et al., 2021). Τα γραφήματα επιτρέπουν την σύγκριση των δεικτών ποικιλομορφίας μεταξύ των διαφορετικών ομάδων δειγμάτων, συνθηκών ή χρονικών σημείων για την κατανόηση του τρόπου με τον οποίο οι μικροβιακές κοινότητες ανταποκρίνονται σε περιβαλλοντικές αλλαγές, πειραματικές διαταραχές ή χρονικές παραλλαγές. Με αυτόν τον τρόπο επιτυγχάνεται η ανακάλυψη των πληροφοριών που αναζητούνται για την απάντηση των ερωτημάτων ή υποθέσεων μιας έρευνας και παράλληλα παρέχεται μια βάση για τη δημιουργία νέων υποθέσεων. Επιπλέον, οι γραφικές απεικονίσεις βοηθούν στον έλεγχο των δεδομένων διευκολύνοντας τον εντοπισμό ακραίων ή ακανόνιστων μοτίβων. Η οπτική επιθεώρηση είναι ιδιαίτερα πολύτιμη στις αναλύσεις εκτεταμένων συνόλων δεδομένων, διασφαλίζοντας την ακεραιότητα των δεικτών ποικιλομορφίας.

Η οπτικοποίηση των αποτελεσμάτων ποικιλομορφίας περιλαμβάνει την επιλογή κατάλληλων τεχνικών απεικόνισης με βάση την φύση των δεδομένων και των στόχων της έρευνας που διεξάγεται (Peeters et al., 2021). Είναι αναγκαία η αποσαφήνιση των ερευνητικών ερωτημάτων ή υποθέσεων που επιθυμούνται να αντιμετωπιστούν μέσω των γραφικών απεικονίσεων, διότι με βάση αυτά επιτρέπεται η ομαδοποίηση δειγμάτων ανάλογα με το περιβάλλον από τα οποίο προέρχονται, τον χρόνο και τις συνθήκες στις οποίες πραγματοποιήθηκε η δειγματοληψία. Οι δείκτες της α-ποικιλομορφίας συχνά απεικονίζονται μέσω θηκογραμμάτων (box plots) (Peeters et al., 2021). Αυτά τα γραφήματα είναι χρήσιμα επειδή παρέχουν μια συμπαγή περίληψη της κατανομής των τιμών της α-ποικιλομορφίας σε διαφορετικές ομάδες δειγμάτων, επιτρέποντας παράλληλα την οπτική σύγκριση των τιμών αυτών μεταξύ των ομάδων.

Στο πεδίο την μικροβιωματικής έρευνας, εάν οι διαγραμματικές απεικονίσεις δεν συνοδεύονται από στατιστικούς ελέγχους, η εγκυρότητα των αποτελεσμάτων σχετικά με την βακτηριακή ποικιλομορφία των δειγμάτων παραμένει περιορισμένη. Ο στατιστικός έλεγχος διαδραματίζει κρίσιμο ρόλο στην αυστηρή αξιολόγηση της σημαντικότητας των παρατηρούμενων διαφορών στην άλφα και βήτα ποικιλομορφία των δειγμάτων και η απουσία αυτού μπορεί να φέρει τον κίνδυνο παρερμηνείας των τυχαίων παραλλαγών ως σημαντικών βιολογικών διαφορών και κατά επέκταση να οδηγήσει σε λανθασμένα συμπεράσματα (Pan, 2021). Ως εκ τούτου, ο συνδυασμός των διορατικών απεικονίσεων και ισχυρών στατιστικών αναλύσεων αποτελεί υψίστης σημασίας, διασφαλίζοντας ότι η εξερεύνηση και η ερμηνεία της βακτηριακής ποικιλομορφίας στηρίζονται σε αυστηρό επιστημονικό έλεγχο και ενθαρρύνοντας μια πιο ολοκληρωμένη κατανόηση της περίπλοκης δυναμικής των μικροβιακών κοινοτήτων και την αναπαραγωγή της μικροβιωματικής έρευνας.

Στην επιδίωξη μιας βαθύτερης κατανόησης των μικροβιακών συνθέσεων, η επιλογή των κατάλληλων στατιστικών μεθόδων ή μοντέλων χαρακτηρίζεται κρίσιμη για την εκτέλεση μίας ολοκληρωμένης ανάλυσης ποικιλομορφίας (Pan, 2021). Οι κλασικοί στατιστικοί

έλεγχοι, αν και είναι πολύτιμοι σε πολλά πλαίσια, μπορεί να αποδειχθούν ανεπαρκείς για πολύπλοκη φύση των δεδομένων μικροβιακής ποικιλομορφίας που προέρχονται από την αλληλούχιση του γονιδίου 16S rRNA. Τα εγγενή χαρακτηριστικά των συνόλων μικροβιωματικών δεδομένων, όπως τα μη κανονικά κατανεμημένα και συνθετικά δεδομένα, θέτουν προκλήσεις που οι κλασικοί στατιστικοί έλεγχοι ενδέχεται να μην αντιμετωπίζουν αποτελεσματικά και υπάρχει πιθανότητα η χρήση αυτών να φέρει παραπλανητικά ή μη ερμηνεύσιμα αποτελέσματα. Για αυτόν τον λόγο, είναι αναγκαία η εφαρμογή εξειδικευμένων στατιστικών μεθόδων, εκ των οποίων η επιλογή τους εξαρτάται από τις προϋποθέσεις τους, την φύση των δεδομένων ποικιλομορφίας και τον πειραματικό σχεδιασμό της έρευνας.

Για την ανάλυση των δεικτών της α-ποικιλομορφίας, οι Mann-Whitney U (McKnight & Najab, 2010) και Kruskal-Wallis (Kruskal & Wallis, 1952) έλεγχοι χαρακτηρίζονται κατάλληλοι στατιστικά εργαλεία για σύγκριση της ποικιλομορφίας δύο ή παραπάνω ανεξάρτητων ομάδων δειγμάτων διότι μπορούν να χειριστούν μη-κανονικές κατανομές και παρέχουν ισχυρά αποτελέσματα. Πιο αναλυτικά, ο Mann-Whitney U (McKnight & Najab, 2010) έλεγχος, γνωστός και ως έλεγχος αθροίσματος διατάξεων του Wilcoxon, είναι ένας μη παραμετρικός στατιστικός έλεγχος που χρησιμοποιείται για να εξεταστεί εάν υπάρχει στατιστικά σημαντική διαφορά στις μετρήσεις ποικιλομορφίας μεταξύ δύο ανεξάρτητων ομάδων δειγμάτων ή πειραματικών συνθηκών. Η εφαρμογή του Mann-Whitney U ελέγχου ακολουθεί μια απλή διαδικασία, ξεκινώντας με τη διατύπωση μηδενικής και εναλλακτικής υπόθεσης, επιβεβαιώνοντας την απουσία ή την παρουσία σημαντικών διαφορών στους υπολογισμούς της α-ποικιλομορφίας. Εφόσον καταταχτούν οι τιμές ποικιλομορφίας από την χαμηλότερη στην υψηλότερη και από τις δύο ομάδες και δημιουργηθεί η κατανομή Mann-Whitney U, υπολογίζεται ο στατιστικός έλεγχος (U) από τον εξής τύπο:

$$U = R_1 - \frac{n_1 \cdot (n_1 + 1)}{2}$$

όπου R_1 είναι το είναι το άθροισμα των κατατάξεων της πρώτης ομάδας και n_1 είναι ο αριθμός των συνολικών μετρήσεων της πρώτης ομάδας. Στην συνέχεια, ο στατιστικός έλεγχος συγκρίνεται με κρίσιμες τιμές από την Mann-Whitney U κατανομή ή χρησιμοποιείται για τον υπολογισμό της p-τιμής (p-value). Εάν η τιμή p είναι κάτω από το επιλεγμένο επίπεδο σημαντικότητας (π.χ. 0,05), η μηδενική υπόθεση απορρίπτεται.

Ο Kruskal-Wallis (Kruskal & Wallis, 1952) έλεγχος είναι ένας μη-παραμετρικός στατιστικός έλεγχος που χρησιμοποιείται για να προσδιοριστεί εάν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ δύο ή περισσότερων ανεξάρτητων ομάδων. Είναι μια επέκταση του Mann-Whitney U ελέγχου για περισσότερες από δύο ομάδες, ενώ η λογική του ορισμού των υποθέσεων και η διαχείριση των τιμών α-ποικιλομορφίας είναι παρόμοια. Για τον υπολογισμό του ελέγχου Kruskal-Wallis (H), ο τύπος που χρησιμοποιείται είναι πιο περίπλοκος και έχει ως εξής:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

όπου το N είναι ο συνολικός αριθμός των μετρήσεων, το k είναι ο αριθμός των ομάδων, το R_i είναι ο αριθμός των κατατάξεων για την ομάδα i και το n_i είναι ο αριθμός των μετρήσεων στης ομάδας i. Στην συνέχεια, ακολουθεί ο προσδιορισμός της σημαντικότητας p για να αξιολογηθεί εάν υπάρχουν σημαντικές διαφορές στις τιμές α-ποικιλομορφίας μεταξύ πολλαπλών ομάδων ή συνθηκών. Στην περίπτωση που ο έλεγχος υποδεικνύει σημαντική

διαφορά, μπορούν να πραγματοποιηθούν περεταίρω έλεγχοι για να προσδιοριστούν ποιες συγκεκριμένες ομάδες διαφέρουν μεταξύ τους.

1.4.8 QIIME2

Για την ανάλυση μικροβιώματος από δεδομένα αλληλούχισης του 16S rRNA γονιδίου, έχει αναπτυχθεί μια σειρά από εξαιρετικά ισχυρά εργαλεία λογισμικών ανοιχτού κώδικα (open source), συμπεριλαμβανομένου του mother (Schloss et al., 2009), του phyloseq (McMurdie & Holmes, 2013) και άλλα συναφή εργαλεία που διατίθενται μέσω του Bioconductor (Marini et al., 2020). Ένα ακόμα ευρέως χρησιμοποιούμενο λογισμικό είναι το QIIME 2 (Bolyen et al., 2019), το οποίο είναι μια πλήρως ανασχεδιασμένη πλατφόρμα βιοπληροφορικής ανάλυσης μικροβιώματος που βασίζεται στη δημοφιλή πλατφόρμα QIIME (Caporaso et al., 2010), την οποία και έχει αντικαταστήσει.

Το αρχικό QIIME (ακρωνύμιο του Quantitative Insights Into Microbial Ecology), που στην σύγχρονη εποχή αναφέρεται ως QIIME 1, δημοσιεύτηκε το 2010. Λόγω της έντονης ανάγκης διαχείρισης δεδομένων αλληλούχισης αμπλικονίων εκείνη την περίοδο, το QIIME 1 κατάφερε να βρεθεί στο επίκεντρο του κλάδου της έρευνας μικροβιώματος, συγκεντρώνοντας πάνω από 30.000 αναφορές σε διάφορα επιστημονικά περιοδικά την τελευταία δεκαετία (Google Scholar, 2023). Το συγκεκριμένο λογισμικό προσφέρει στους χρήστες της την δυνατότητα μετατροπής ακατέργαστων δεδομένων αλληλούχισης που παράγονται από διάφορες πλατφόρμες αλληλούχισης, συμπεριλαμβανομένου και της Illumina, σε ποιοτικά γραφήματα και στατιστικές αναλύσεις. Συγκεκριμένα, το QIIME 1 αποτελείται από μία συλλογή βιοπληροφορικών εργαλείων για την προεπεξεργασία των αναγνωσμάτων, την κατασκευή OTUs, την ταξινομική ανάθεση αυτών, την φυλογενετική ανακατασκευή και την ανάλυση και οπτικοποίηση της ποικιλομορφίας των δειγμάτων. Μερικά από αυτά τα εργαλεία έχουν κατασκευαστεί από τους ίδιους ιδρυτές του QIIME 1 στην Python, ενώ τα υπόλοιπα έχουν ενσωματωθεί κατάλληλα στον κώδικα του λογισμικού, όπως το USEARCH (Edgar, 2010).

Έχοντας το προνόμιο μίας οργανωμένης πλατφόρμας ανάλυσης αλληλουχιών αμπλικονίων σε συνδυασμό με την ευέλικτη χρήση σημαντικών βιοπληροφορικών εργαλείων, το QIIME 1 απέκτησε γρήγορα έναν σημαντικό αριθμό χρηστών. Το φαινόμενο αυτό οδήγησε στην ανάπτυξη δημόσιων διαδικτυακών συζητήσεων (forums) και εκπαιδευτικών σεμιναρίων για την συστηματική καθοδήγηση των χρηστών του λογισμικού. Η τακτική αλληλεπίδραση μεταξύ των χρηστών και της ομάδας διαχείρισης της πλατφόρμας αποτέλεσε ζωτικής σημασίας για την βελτίωση του QIIME 1. Όμως, οι μικρο-αναβαθμίσεις της πλατφόρμας δεν επίλυαν κάποια θεμελιώδη προβλήματα που αντιμετώπιζαν οι χρήστες. Για παράδειγμα, οι χρήστες του QIIME 1, πολλοί από τους οποίους δεν ήταν εκπαιδευμένοι αναλυτές δεδομένων αλληλούχισης, δυσκολεύονταν να κατανοήσουν και να αναφέρουν αξιόπιστα τις συχνά πολύπλοκες βιοπληροφορικές ροές εργασίας (workflows). Αυτό έκανε δύσκολη την αναπαραγωγή της έρευνάς τους και η έλλειψη αυτοματοποιημένης καταγραφής όλων των βημάτων στη ροή εργασιών συχνά αποτελούσε εμπόδιο στην παροχή τεχνικής υποστήριξης στους χρήστες. Επιπλέον, η αποκλειστική αξιοποίηση του λογισμικού μέσω γραμμή εντολών παρεμπόδιζε τα άτομα που δεν ήταν εξοικειωμένα με προγραμματιστικές διεπαφές (interfaces). Επίσης, λόγω της έντονης δραστηριότητας στον κάδο της ανάλυσης δεδομένων αλληλούχισης, έχουν αναβαθμιστεί και δημιουργηθεί πολλά σύγχρονα

βιοπληροφορικά εργαλεία, τα οποία όμως δεν κατάφεραν να ενσωματωθούν στο QIIME 1 μετά την ίδρυσή του.

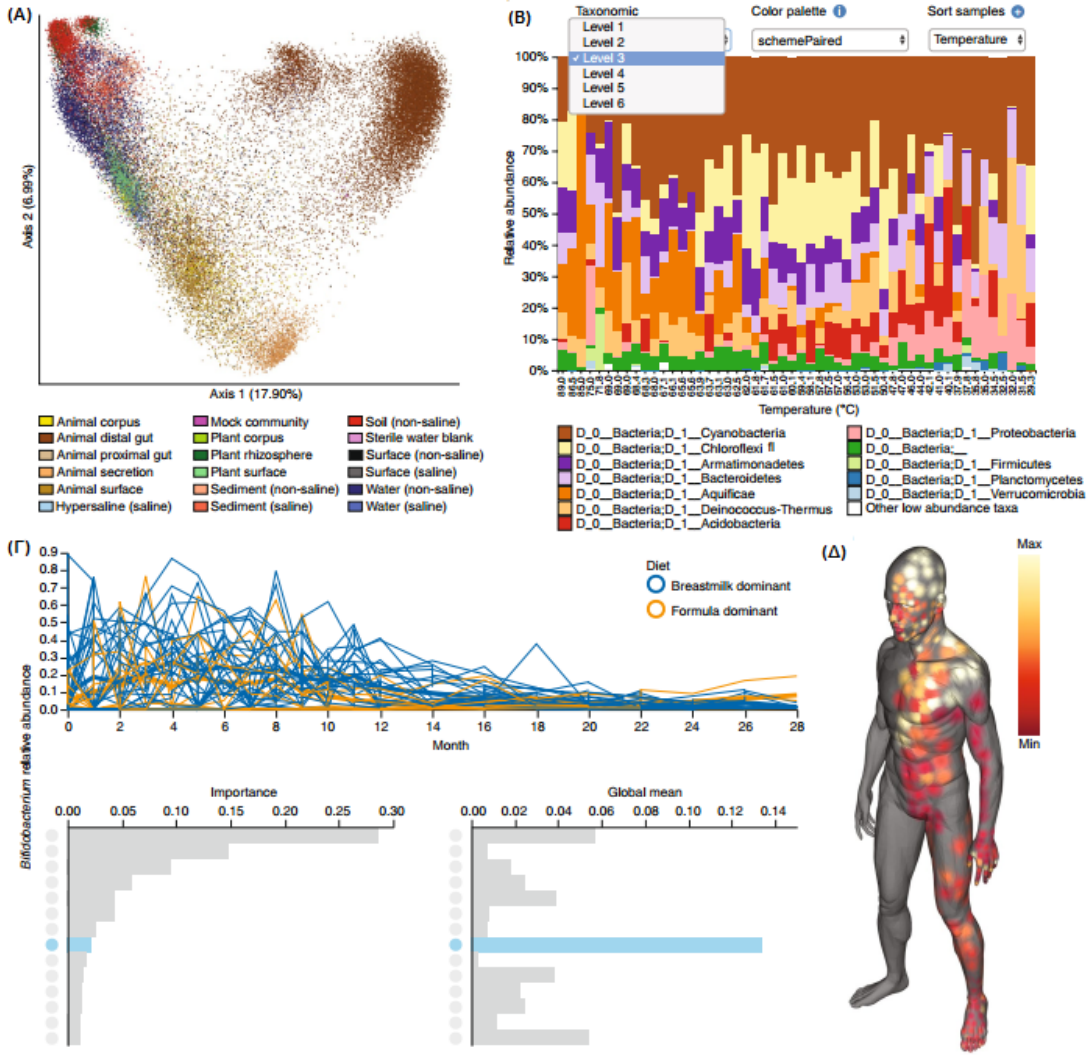
Συνεπώς, για την εξ' ολοκλήρου κάλυψη των αναγκών των χρηστών και την αντιμετώπιση των ελλείψεων του QIIME 1, οι διαχειριστές του λογισμικού ανασχεδίασαν εκ νέου την πλατφόρμα, γεγονός που οδήγησε στην έκδοση του QIIME 2 το έτος 2019. Κατά βάσει, το QIIME 2 αποτελεί η πλατφόρμα η οποία διατηρεί τα βασικά χαρακτηριστικά που έκαναν το QIIME 1 ένα ισχυρό και ευρέως χρησιμοποιούμενο βιοπληροφορικό εργαλείο, ενώ παράλληλα παρέχει νέα χαρακτηριστικά που οδηγούν στην ολοκληρωμένη και σύγχρονη ανάλυση μικροβιώματος. Τα ανανεωμένα χαρακτηριστικά συμπεριλαμβάνουν την κατασκευή διαδραστικών χωροχρονικών διαγραμμάτων για την εύκολη αλληλεπίδραση και ερμηνεία των παραγόμενων δεδομένων (**Σχήμα 1.19**), την υποστήριξη της μεταβολομικής και shotgun μεταγονιδιωματικής ανάλυσης, καθώς και την αυτοματοποιημένη παρακολούθηση προέλευσης δεδομένων αλληλούχισης, δηλαδή του workflow, για τη διασφάλιση της διαφάνειας και αναπαραγωγισιμότητας των δεδομένων μικροβιώματος.

Επιπλέον, το QIIME 2 παρέχει όλα τα εργαλεία επεξεργασίας και διαγραμματοποίησης δεδομένων υπό την μορφή πρόσθετων (plug-ins). Γενικά, το plug-in είναι το λογισμικό που προσφέρει την δυνατότητα εισαγωγής νέων χαρακτηριστικών και λειτουργιών σε ένα κεντρικό πρόγραμμα χωρίς να αλλάζει το ίδιο το πρόγραμμα υποδοχής. Αυτό σημαίνει ότι τα βιοπληροφορικά εργαλεία που παρέχονται από την πλατφόρμα είναι σε ένα βαθμό ανεξάρτητα από τον κορμό του κώδικα του κεντρικού προγράμματος, ενώ παράλληλα είναι ενσωματωμένα με τέτοιον τρόπο έτσι ώστε η αξιοποίησή και η αναβάθμισή τους να είναι ευέλικτη και μη παρεμβατική στην βασική λειτουργία του QIIME 2. Δεδομένου αυτού, οι ιδρυτές της πλατφόρμας, οι εξωτερικοί προγραμματιστές βιοπληροφορικων εργαλείων ανάλυσης μικροβιώματος που επιθυμούν να έχουν πρόσβαση στις μεθόδους τους μέσω του QIIME 2, καθώς και οποιοσδήποτε χρήστης της πλατφόρμας μπορούν να αναπτύξουν, να εισάγουν και να διαδώσουν εύκολα μία καινούρια μέθοδο με την μορφή plug-in στο κεντρικό πρόγραμμα. Συνεπώς, με την εύκολη εγκατάσταση σύγχρονων μεθόδων στο QIIME 2, αναβαθμίζονται διάφορες πτυχές αναλύσεων μικροβιώματος, όπως την βελτίωση της ταξινομικής ανάθεσης μέσω της κατασκευής ASVs αντί των OTUs.

Ένα ακόμα προνομιακό χαρακτηριστικό του QIIME 2 είναι ο ίδιος ο σχεδιασμός του λογισμικού που επιτρέπει την πρόσβαση στις ίδιες μεθόδους μέσω διαφορετικών τύπων διεπαφής, συμπεριλαμβανομένου της γραμμής εντολών και του προγραμματιστικού/διαδικτυακού. Από το 2021, το QIIME 2 υποστηρίζει επίσης το Galaxy (Jalili et al., 2020), μια δημόσια πλατφόρμα για την επεξεργασία μεγάλων συνόλων δεδομένων σε μια ισχυρή διαδικτυακή υποδομή. Μέσω του δημοφιλούς γραφικού περιβάλλοντος του Galaxy, οι ερευνητές μικροβιώματος έχουν πλέον πλήρη πρόσβαση στη λειτουργικότητα του QIIME 2, χωρίς να απαιτείται η χρήση της γραμμής εντολών ή προγραμματιστική εμπειρία.

Τέλος, πίσω από το QIIME 2 κρύβεται μια μεγάλη ομάδα τεχνικής υποστήριξης για την άμεση επίλυση προβλημάτων ή αποριών σχετικών με την πλατφόρμα που μπορεί να συναντήσει ο κάθε χρήστης. Μέσω της δημόσιας διαδικτυακής ιστοσελίδας της πλατφόρμας (https://qiime2.org/), οι χρήστες έχουν πρόσβαση σε αναλυτικές οδηγίες για τον τρόπο χρήσης του QIIME 2 καθώς και μεθόδους και τεχνικές ανάλυσης μικροβιώματος που προτείνεται ανάλογα με τα δεδομένα προς επεξεργασία. Επιπρόσθετα, η ιστοσελίδα προσφέρει και την άμεση επικοινωνία με την ομάδα υποστήριξης για την επίλυση ενδεχομένως πιο εξειδικευμένων θεμάτων σχετικά με την πλατφόρμα αλλά και γενικά για ότι

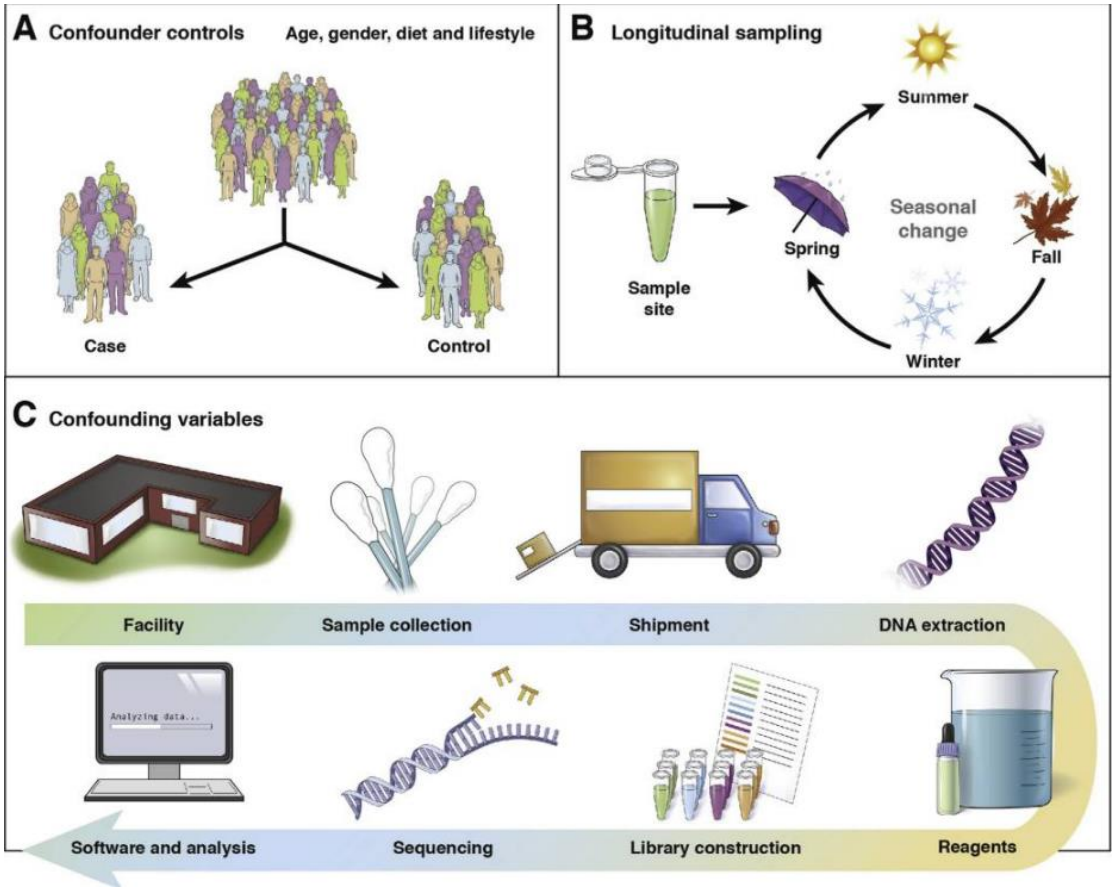
αφορά την έρευνα μικροβιώματος. Αποτέλεσμα όλων των προαναφερόμενων χαρακτηριστικών είναι η καλλιέργεια μίας ποικιλόμορφης, χωρίς αποκλεισμούς, κοινότητα επιστημόνων, συμπεριλαμβανομένου και των μηχανικών λογισμικού, στατιστικολόγων, εκπαιδευτικών, φοιτητών και άλλων ενδιαφερόμενων της επιστήμης του μικροβιώματος, η οποία μοιράζεται ανοιχτά μεθόδους, δεδομένα και γνώση για την προώθηση της έρευνας μικροβιώματος.



Σχήμα 1.19 Απεικόνιση 4 διαφορετικών διαδραστικών διαγραμμάτων που προσφέρονται από το QIIME 2. **(Α)** Διάγραμμα διασποράς 37.680 δειγμάτων με τα χρώματα να αντιπροσωπεύουν τον τύπο δείγματος, δείχνοντας την επεκτασιμότητα του QIIME 2. **(Β)** Διαδραστικό διάγραμμα ταξινομικής σύνθεσης που επιτρέπει την οπτικοποίηση της μικροβιακής σύνθεσης δειγμάτων σε διάφορα ταξινομικά επίπεδα. **(Γ)** Γραφική παράσταση μεταβλητότητας αφθονίας ενός συγκεκριμένου μικροβίου σε βιολογικά δείγματα με την πάροδο του χρόνου. Τα γραφήματα ράβδων κατατάσσουν τη σημαντικότητα (προγνωστική ισχύς για το χρονικό σημείο) και τη μέση αφθονία όλων των μικροβιακών χαρακτηριστικών, προσφέροντας μια απεικόνιση για περαιτέρω ανάλυση μεταβλητότητας. **(Δ)** Μοριακή χαρτογράφηση της επιφάνειας του ανθρώπινου δέρματος. Οι έγχρωμες κηλίδες αντιπροσωπεύουν την αφθονία του μικρομορίου συστατικού (sodium laureth sulfate) στο ανθρώπινο δέρμα. Τα δείγματα δεδομένων μπορούν να οπτικοποιηθούν σε τρισδιάστατα μοντέλα, υποστηρίζοντας έτσι την ανακάλυψη χωρικών μοτίβων. (Bolyen et al., 2019, p. 2)

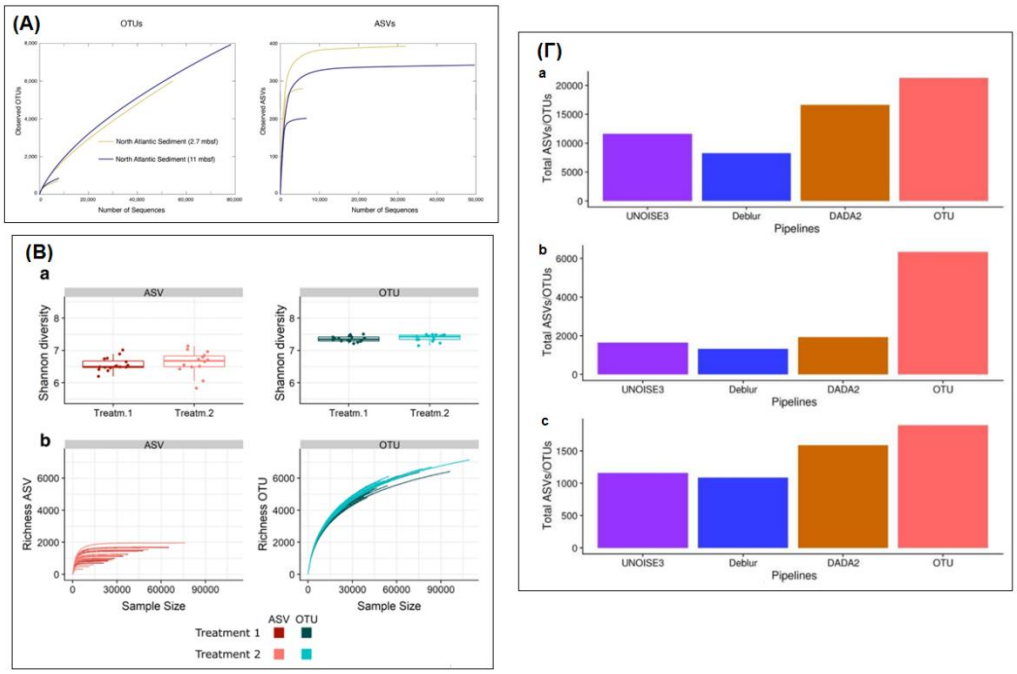
1.4.9 OTUs vs ASVs

Όταν πρόκειται για την διεξαγωγή μίας κλινικής έρευνας μικροβιώματος, είναι ξεκάθαρο σε αυτό το σημείο ότι κάθε βήμα και επεξεργαστικό στάδιο, συμπεριλαμβανομένου της δειγματοληψίας, της πειραματικής επεξεργασίας αυτών, της διαδικασίας αλληλούχισης καθώς και της βιοπληροφορικής ανάλυσης των δεδομένων, αποτελούν παράγοντες που μπορούν να προκαλέσουν διαφοροποιήσεις στα τελικά αποτελέσματα και στην βιολογική ερμηνεία δειγμάτων (**Σχήμα 1.20**) (Allaband et al., 2019). Ακόμα και αν όλα τα βήματα εκτελεστούν μεθοδικά και βάση πρωτοκόλλων, το τελευταίο στάδιο που αφορά την βιοπληροφορική επεξεργασία είναι αυτό που εν τέλει αποκαλύπτει το προφίλ και την δυναμική ενός μικροβιώματος. Στο πλαίσιο της ανάλυσης του 16S rRNA γονιδίου, η επιλογή των κατάλληλων βιοπληροφορικών εργαλείων και η σωστή χρήση τους καθορίζουν την προσπάθεια αποκόμισης έγκυρων και όσο το δυνατόν βέλτιστων πληροφοριών μίας βακτηριακής κοινότητας. Μία από τις πιο έντονες διαμάχες τις επιστημονικής κοινότητας για την διαχείριση δεδομένων αλληλούχισης αμπλικονίων είναι η επιλογή ομαδοποίησης αυτών σε λειτουργικές ταξινομικές μονάδες (OTUs) ή αποθορυβοποίησης αυτών σε ακριβές παραλλαγές αμπλικονίων (ASVs).



Σχήμα 1.20 Η διεξαγωγή μίας κλινικής έρευνας μικροβιώματος απαιτεί ιδιαίτερη προσοχή σε πολλούς παράγοντες. **(A)** Η διαστρωμάτωση από πιθανούς συγχυτικούς παράγοντες (π.χ. ηλικία, φύλο, διατροφή, παράγοντες τρόπου ζωής και φάρμακα) μπορεί να βοηθήσει στην επίλυση διαφορών στη μικροχλωρίδα μεταξύ ομάδων ενδιαφέροντος που διαφορετικά θα μπορούσαν να καλυφθούν από ένα συγχυτικό αποτέλεσμα. **(B)** Οι διαχρονικές μελέτες είναι ιδιαίτερα ισχυρά επειδή ελέγχουν συγχυτικούς παράγοντες και επιτρέπουν την αξιολόγηση της σταθερότητας της κοινότητας. **(C)** Για όλες τις μελέτες, η τυποποίηση των τεχνικών παραγόντων και η επεξεργασία δειγμάτων είναι ουσιαστικής σημασίας για τον έλεγχο της διαφοροποίησης που εισάγεται σε κάθε βήμα της διαδικασίας: κιτ αντιδραστηρίων, εκκινητών, αποθήκευση δειγμάτων και άλλοι παράγοντες. (Allaband et al., 2019)

Με την αυξανόμενη δημοτικότητα των προσεγγίσεων αποθορυβοποίησης, πολλές μελέτες έχουν συγκρίνει τα αποτελέσματα δεδομένων αλληλούχισης αμπλικονίων που έχουν προκύψει από την μέθοδο ομαδοποίησης και την μέθοδο αποθορυβοποίησης. Οι έρευνες που βασίζονται σε πρότυπα δείγματα βακτηριακών κοινοτήτων δείχνουν ότι οι μέθοδοι αποθορυβοποίησης έχουν υψηλότερη ευαισθησία στην ανίχνευση βακτηριακών ειδών, ακόμα και σε επίπεδο στελέχους (Caruso et al., 2019; Needham et al., 2017; Prodan et al., 2020; Xue et al., 2018). Ωστόσο, μελέτες που χρησιμοποίησαν σύνολα δεδομένων αλληλούχισης τα οποία έχουν προκύψει από φυσικά δείγματα για την σύγκριση των μεθόδων, όπως δείγματα θαλάσσιου ιζήματος (Kerrigan & D'Hondt, 2022), υπόγειων υδάτων (Kerrigan & D'Hondt, 2022), ιζήματος και νερού ποταμών (Kerrigan & D'Hondt, 2022), χώματος (Joos et al., 2020; Nearing et al., 2018), ριζόσφαιρας (Joos et al., 2020) και διάφορων αβιοτικών περιβαλλόντων (Glassman & Martiny, 2018) καθώς και ανθρώπινου μητρικού γάλακτος (Moossavi et al., 2020) και εντερικών μικροβιωμάτων από ανθρώπους (Nearing et al., 2018; Prodan et al., 2020), ποντίκια (Nearing et al., 2018) και γαρίδες (García-López et al., 2021), βρήκαν παρόμοια βιολογικά αποτελέσματα ως προς την ταξινομική σύνθεση. Από αυτές τις μελέτες, η κύρια αδυναμία των προσεγγίσεων που βασίζονται στην ομαδοποίηση των αναγνωσμάτων σε OTUs φαίνεται να είναι η ακριβής ανίχνευση της άλφα ποικιλομορφίας, καθώς οι μέθοδοι ομαδοποίησης συχνά υπερεκτιμούν τον βακτηριακό πλούτο σε σύγκριση με τις μεθόδους αποθορυβοποίησης (**Σχήμα 1.21**). Δεδομένου αυτού του βαθμού ομοιότητας, η ταυτόχρονη χρήση και των δύο προσεγγίσεων OTU και ASV μπορεί να είναι χρήσιμη για την ενίσχυση γενικών συμπερασμάτων σχετικά με την ποικιλομορφία και τη σύνθεση ενός μικροβιώματος, και οι διαφορές μεταξύ των αποτελεσμάτων των δύο προσεγγίσεων συμβάλλουν για την επισήμανση σημείων όπου απαιτείται λεπτομερέστερος έλεγχος.



Σχήμα 1.21 (A) Καμπύλες βακτηριακής ταξινομικής αφθονίας λειτουργικών ταξινομικών μονάδων (OTUs, αριστερά) και παραλλαγών αλληλουχίας αμπλικονίου (ASVs, δεξιά) για δύο δείγματα θαλάσσιων ιζημάτων τόσο στον πλήρη αριθμό των αλληλουχιών τους όσο και σε ένα υποδειγματοληπτικό σύνολο δεδομένων 10.000 αλληλουχιών. (Kerrigan & D'Hondt, 2022) **(B)** Οι δείκτες ποικιλομορφίας Shannon **(a)** και ο πλούτος έναντι του βάθους αλληλούχισης **(b)** των μεθόδων παραγωγής ASVs και OTUs σε σύνολο δεδομένων βακτηριακού χώματος. Τα δείγματα (n=16) έχουν υποστεί σε δύο διαφορετικές επεξεργασίες. (Joos et al., 2020) **(Γ)** Ο συνολικός αριθμός ASVs/OTUs που προέκυψε από διάφορες μεθόδους ομαδοποίησης και αποθορυβοποίησης σε

δεδομένα **(α)** χώματος, **(b)** ποντικών και **(c)** ανθρώπινου εντερικού μικροβιώματος. (Nearing et al., 2018)

1.5 Σκοπός Διπλωματικής Εργασίας

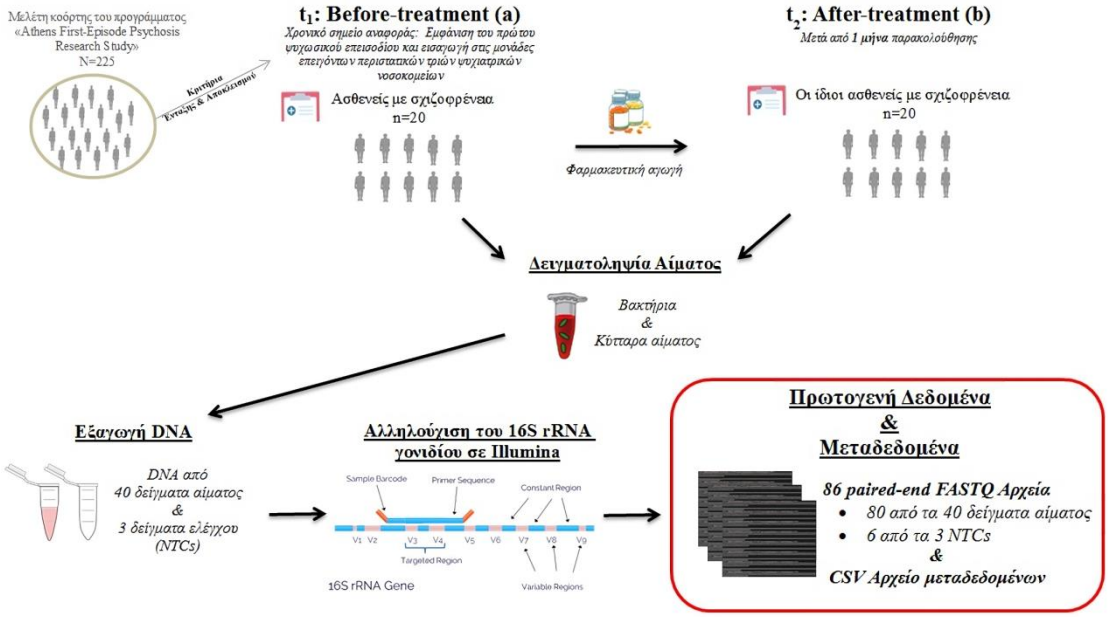
Ενώ υπάρχουν έντονες ενδείξεις ότι το μικροβίωμα του αίματος παίζει σημαντικό ρόλο και ενδεχομένως είναι ο συνδετικός κρίκος του άξονα εγκεφάλου-εντέρου, ο αριθμός ερευνών που σχετίζουν το μικροβιακό προφίλ του αίματος με την σχιζοφρένεια είναι μικρός. Ταυτόχρονα, με την πληθώρα διαθέσιμων βιοπληροφορικών εργαλείων, η ανάλυση δεδομένων αλληλούχισης έχει γίνει σε μεγάλο βαθμό περίπλοκη και η επιλογή των κατάλληλων εργαλείων επεξεργασίας αποτελεί υψίστης σημασίας για την ερμηνεία κλινικών αποτελεσμάτων. Ειδικά στο πλαίσιο της ανάλυσης δεδομένων αλληλούχισης του 16S rRNA γονιδίου, το δίλημμα μεταξύ της χρήσης μεθόδων ομαδοποίησης ή αποθορυβοποίησης για ανάλυση αμπλικονίων υπογραμμίζει την ανάγκη για την τυποποίηση των ροών επεξεργασίας τέτοιων δεδομένων.

Η παρούσα διπλωματική εργασία μελετά το μικροβίωμα του αίματος στη σχιζοφρένεια στοχεύοντας στην ανάλυση δεδομένων αλληλούχισης του 16S rRNA γονιδίου από δείγματα αίματος σχιζοφρενών με την χρήση τριών διαφορετικών βιοπληροφορικων εργαλείων της διαδεδομένης πλατφόρμας QIIME2. Τα εργαλεία αυτά βασίζονται στην ομαδοποίηση αναγνωσμάτων σε OTUs (VSEARCH) και στην αποθορυβοποίηση αναγνωσμάτων σε ASVs (DADA2 και Deblur). Ο κύριος σκοπός είναι η διαπίστωση της ομοιότητας ή μη του ταξινομικού προφίλ και της άλφα ποικιλομορφίας των βιολογικών δειγμάτων που προκύπτουν από τις διαφορετικές μεθόδους ομαδοποίησης και αποθορυβοποίησης αμπλικονίων. Παράλληλα, εξετάζεται η συμπεριφορά των δεδομένων αλληλούχισης, όπως ο αριθμός διατηρητέων αναγνωσμάτων στα διάφορα επεξεργαστικά στάδια και στις διάφορες τιμές παραμέτρων που απαιτούνται να εφαρμοστούν για τον συνολικό ποιοτικό έλεγχο των δεδομένων αλληλούχισης του 16S rRNA γονιδίου.

2 Μεθοδολογία

2.1 Δεδομένα

Η προέλευση όλων των δεδομένων και των πληροφοριών τους για την πραγματοποίηση της παρούσας διπλωματικής εργασίας αποτελεί το Εργαστήριο Βιοτεχνολογίας της Σχολής Χημικών Μηχανικών του ΕΜΠ σε συνεργασία με την Α' Ψυχιατρική Κλινική του ΕΚΠΑ του Αιγινήτειου Νοσοκομείου και τα Ερευνητικά Εργαστήρια Γ. Λιβανός και Μ. Σίμου της Α' Κλινικής Εντατικής Θεραπείας του ΕΚΠΑ του νοσοκομείου «Ο Ευαγγελισμός» στο πλαίσιο του ερευνητικού προγράμματος με τίτλο «Προοπτική μελέτη του μικροβιώματος στο αίμα ασθενών με πρώτο επεισόδιο σχιζοφρένειας υποδεικνύει βιοδείκτες ανταπόκρισης στην θεραπεία». Το συγκεκριμένο πρόγραμμα πρόκειται για μια διαχρονική έρευνα, στην οποία μελετάται το μικροβίωμα του αίματος σε άτομα με σχιζοφρένεια που παρουσίασαν το πρώτο επεισόδιο ψύχωσης καθώς και τις μεταβολές αυτού μετά από ένα μήνα αντιψυχωσικής θεραπείας (**Σχήμα 2.1**). Στις παρακάτω υποενότητες αναφέρονται ο πληθυσμός που συμμετείχε στην έρευνα, οι πληροφορίες που σχετίζονται με την δειγματοληψία του αίματος από τους συμμετέχοντες, τον τρόπο προετοιμασίας των δειγμάτων και την αλληλούχισή τους. Τέλος, γίνεται αναφορά των πρωτογενών δεδομένων και μεταδεδομένων που έχουν προκύψει από τα προαναφερόμενα, τα οποία εν τέλει αποτελούν τα αρχεία αξιοποίησης και επεξεργασίας της παρούσας μελέτης.



Σχήμα 2.1 Διάγραμμα ροής της προέλευσης των πρωτογενών δεδομένων.

2.1.1 Συμμετέχοντες

Οι εθελοντές της παρούσας εργασίας αποτελούν μια υποομάδα ενός μεγαλύτερου πληθυσμού, ο οποίος συλλέχθηκε από μια μελέτη κοόρτης του προγράμματος «Athens First-Episode Psychosis Research Study» (Erzin et al., 2023; Xenaki et al., 2020, 2022). Ο σκοπός του προγράμματος είναι η διερεύνηση μεταξύ των πολλαπλών γενετικών, περιβαλλοντικών και νευρομεταβολικών παραγόντων κινδύνου εμφάνισης ψυχωσικών διαταραχών. Η έρευνα πραγματοποιήθηκε μέσω της κλινικής διαχείρισης ασθενών οι οποίοι παρουσίασαν το Πρώτο

Ψυχωσικό Επεισόδιο (First Episode Psychosis – FEP) και είχαν ελάχιστη ή μηδενική έκθεση σε αντιψυχωσική θεραπεία στο παρελθόν. Για την δημιουργία της υποομάδας που περιλαμβάνει η παρούσα μελέτη τέθηκαν συγκεκριμένα κριτήρια ένταξης και αποκλεισμού. Τα κριτήρια ένταξης ήταν η παρουσία πρώτου ψυχωσικού επεισοδίου σύμφωνα με τη διαγνωστική συνέντευξη για ψύχωση (World Health Organization, 1992), η ελάχιστη έκθεση σε αντιψυχωσικά πριν την έναρξη του προγράμματος (< 2 βδομάδες) και η διάγνωση των ασθενών με σχιζοφρένεια σύμφωνα με τα ICD-10 (World Health Organization, 1992), DSM-IV-TR (GUZE, 1995) και DSM-5 (American Psychiatric Association, 2013). Τα γενικά κριτήρια αποκλεισμού έλαβαν υπόψη τις οξείες και χρόνιες γενικές ιατρικές παθήσεις που συνδέονται με αντιφλεγμονώδη ή/και αντιβιοτική θεραπεία, τις ψυχωσικές διαταραχές λόγω άλλης ιατρικής πάθησης ή οξείας δηλητηρίασης, τον δείκτη ευφυΐας (IQ < 70), τις αναπτυξιακές διαταραχές και την συγγένεια με εγγεγραμμένο συμμετέχοντα. Από τις προϋποθέσεις ένταξης και αποκλεισμού, οι συμμετέχοντες στην παρούσα εργασία κατέληξαν να είναι 20 ασθενείς διαγνωσμένοι με σχιζοφρένεια που παρουσίασαν το πρώτο ψυχωσικό επεισόδιο και προσήλθαν στις μονάδες επειγόντων περιστατικών τριών ψυχιατρικών νοσοκομείων στην Αθήνα (414 Στρατιωτικό Νοσοκομείο, Πανεπιστημιακό Νοσοκομείο «Αττικόν» και Γενικό Νοσοκομείο «Σισμανόγλειο»). Όλα τα άτομα είναι άντρες ηλικίας μεταξύ 20 και 39 ετών και κατά την έναρξη του προγράμματος είτε δεν είχαν λάβει φαρμακευτική αγωγή είτε είχαν ελάχιστη έκθεση σε αντιψυχωσικά. Η διεξαγωγή της μελέτης έχει λάβει έγκριση και έχει χορηγηθεί από την Επιτροπή Ιατρικής Βιοηθικής του Αιγινήτειου Νοσοκομείου. Ο κάθε συμμετέχοντας παρείχε μια υπογεγραμμένη ενημερωμένη συγκατάθεση (informed consent) που περιλάμβανε όλες τις απαραίτητες λεπτομέρειες για τις διαδικασίες, τους πιθανούς κινδύνους και τα οφέλη της μελέτης.

2.1.2 Δειγματοληψία του αίματος και αλληλούχιση του 16S rRNA γονιδίου

Κατά την έναρξη του προγράμματος (t_1) και μετά από έναν μήνα παρακολούθησης και φαρμακευτικής αγωγής (t_2), συλλέχθηκαν δείγματα αίματος από όλους τους ασθενείς για την διεξαγωγή τυπικών εξετάσεων αίματος. Το κάθε δείγμα διαμοιράστηκε σε δύο δείγματα, από τα οποία το ένα από αυτά χρησιμοποιήθηκε για την προετοιμασία του για αλληλούχιση. Τα δείγματα αυτά, που είναι συνολικά 40, συντηρήθηκαν στους -20C^{o}, ωσότου να αξιοποιηθούν για αλληλούχιση. Για την αποφυγή επιμολύνσεων κατά την διάρκεια της πειραματικής διαδικασίας, η διαχείρηση και επεξεργασία των δειγμάτων πραγματοποιήθηκε σε κατάλληλες συνθήκες αποστείρωσης. Επιπλέον, για τον έλεγχο τυχόν επιμολύνσεων κατά την πειραματική διαδικασία της αλληλούχισης, συμπεριλήφθηκαν 3 δείγματα ελέγχου απουσίας DNA (no template controls, NTC), τα οποία περιείχαν μόνο υπερκαθαρό νερό. Τα δείγματα αυτά υποβλήθηκαν στην ίδια επεξεργασία με τα βιολογικά δείγματα, από την εξαγωγή DNA μέχρι και την αλληλούχιση.

Για την απομόνωση του γενετικού υλικού από τα δείγματα χρησιμοποιήθηκε το κιτ NucleoSpin® Blood σύμφωνα με τις οδηγίες του κατασκευαστή, με ορισμένες τροποποιήσεις για τη συν-απομόνωση του βακτηριακού DNA. Για την αλληλούχιση του 16S rRNA γονιδίου, στοχοποιήθηκε και ενισχύθηκε με την μέθοδο PCR η περιοχή V3-V4 του γονιδίου χρησιμοποιώντας τον Εμπρόσθιο Εκκινητή 341F 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG και Ανάστροφο Εκκινητή 805R 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC . Τα παραγόμενα αμπλικόνια καθαρίστηκαν, υποβλήθηκαν σε παρασκευή βιβλιοθήκης και

αλληλουχήθηκαν σύμφωνα με Πρωτόκολλο Παρασκευής Βιβλιοθήκης 16S Μεταγονιδιωματικής Αλληλούχισης Illumina MiSeq (Illumina MiSeq 16S Metagenomic Sequencing Library Preparation Protocol). Η διαδικασία της αλληλούχισης διεξάχθηκε στο Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών (ΙΙΒΕΑΑ) σε πλατφόρμα αλληλούχισης Illumina MiSeqPE250, για την παραγωγή paired-end 2 x 250 bp αναγνωσμάτων.

2.1.3 Πρωτογενή δεδομένα και μεταδεδομένα

Από την 16S rRNA Illumina αλληλούχιση στο ΙΙΒΕΑΑ λήφθηκαν συνολικά 86 διαχωρισμένα (dimultiplexed) paired-end fastq αρχεία με βαθμό μετατόπισης PHRED 33, από τα οποία μία δυάδα αρχείων αντιστοιχούν σε ένα αρχικό δείγμα. Συγκεκριμένα, τα 80 έχουν προκύψει από την αλληλούχιση 40 δειγμάτων αίματος και αντίστοιχα τα υπόλοιπα 6 από 3 δείγματα ελέγχου NTCs. Τα δείγματα αίματος προέρχονται από τους 20 διαγνωσμένους με σχιζοφρένεια ασθενείς σε δύο διαφορετικές χρονικές στιγμές, πριν (t_1= before treatment = α) και μετά (t_2= after treatment = b) από την θεραπεία ενός μήνα αντιψυχωσικής φαρμακευτικής αγωγής. Εκτός από τα fastq αρχεία, λήφθηκε και ένα αρχείο .csv, το οποίο περιέχει όλες τις πληροφορίες των αρχείων σχετικά με την προέλευση τους, δηλαδή τα μεταδεδομένα. Το αρχείο αυτό παρέχει την ονομασία των δειγμάτων και την τοπική και χρονική προέλευσή τους.

2.2 Βασικά Υπολογιστικά Εργαλεία

Για την εκπόνηση της παρούσας διπλωματικής εργασίας απαιτούνταν σημαντική επεξεργαστική ισχύς η οποία δεν παρέχεται από κοινούς οικιακούς υπολογιστές. Για τον λόγο αυτό, χρησιμοποιήθηκε ο Server της Σχολής Χημικών Μηχανικών, ο οποίος διαθέτει 64 πυρήνες επεξεργασίας, μνήμη RAM 256 gigabyte (GB) και χρησιμοποιεί την CentOS 8.0 του λειτουργικού συστήματος Linux.

Τα βασικά επεξεργαστικά εργαλεία ανάλυσης αναγνωσμάτων και μικροβιώματος που αξιοποιήθηκαν στον ίδιο Server αποτελούν το QIIME2 (Bolyen et al., 2019) έκδοσης 2023.9 για την βιοπληροφορική ανάλυση και διαγραμματική απεικόνιση των αποτελεσμάτων, καθώς και το InteractiveVenn (Heberle et al., 2015), ένα διαδικτυακό εργαλείο για την ανάλυση συνόλων μέσω διαγραμμάτων Venn. Η γλώσσα προγραμματισμού που αξιοποιήθηκε ήταν η BASH για την χρήση της γραμμής εντολών του Linux. Η εκτέλεση της βιοπληροφορικής ανάλυσης πραγματοποιήθηκε εξ ολοκλήρου μέσω terminal.

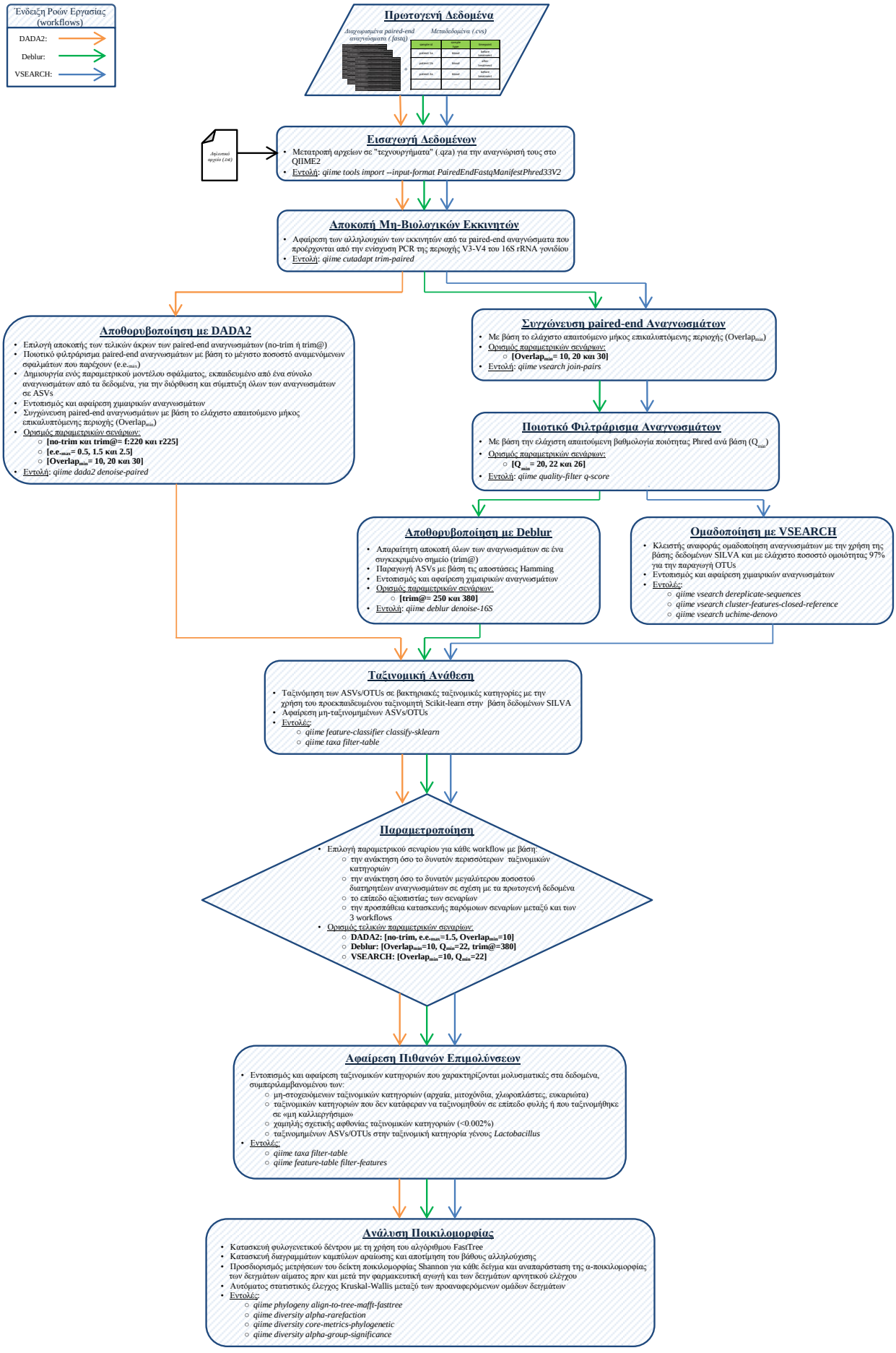
2.3 Υπολογιστική Διαδικασία

Η συλλογιστική πορεία διεκπεραίωσης της παρούσας διπλωματικής εργασίας έτσι ώστε να επιτευχθεί ο απώτερος σκοπός της ακολουθεί κυρίως τις δυνατότητες που παρέχει η πλατφόρμα QIIME2 (https://docs.qiime2.org/2023.9/). Για την διερεύνηση των διαφορών στην βιολογική ερμηνεία πρωτογενών δεδομένων αλληλούχισης του 16S rRNA γονιδίου κατά την εφαρμογή διαφορετικών εργαλείων αποθορυβοποίησης και ομαδοποίησης αναγνωσμάτων, χρησιμοποιούνται δύο εργαλεία αποθορυβοποίησης (DADA2 και Deblur) και ένα ομαδοποίησης (VSEARCH) που βρίσκονται διαθέσιμα στην πλατφόρμα. Συνεπώς, η

υπολογιστική διαδικασία αποτελείται από τρείς βασικές ροές επεξεργασίας (workflows), οι οποίες βασίζονται στα απαραίτητα βήματα επεξεργασίας για την εφαρμογή των προαναφερόμενων εργαλείων αποθορυβοποίησης και ομαδοποίησης (**Σχήμα 2.2**). Τα βήματα επεξεργασίας δεδομένων και οι σειρά τους είναι εμπνευσμένα από τις οδηγίες που παρέχονται από το πρόγραμμα QIIME2 (Estaki et al., 2020), τις οδηγίες των κατασκευαστών των βασικών βιοπληροφορικών εργαλείων (Amir et al., 2017; Bokulich et al., 2013; Callahan et al., 2016; Martin, 2011; Rognes et al., 2016; Zhang et al., 2014), τις δημοσιευμένες έρευνες που αφορούν για τις προτάσεις ανάλυσης δεδομένων αλληλούχισης του 16S rRNA γονιδίου γενικά (de la Cuesta-Zuluaga & Escobar, 2016; Edgar & Flyvbjerg, 2015; Liu et al., 2021; Mohsen et al., 2019; Qian et al., 2020; Reitmeier et al., 2021; Trego et al., 2022) και αυτών που αφορούν εξειδικευμένα αντίστοιχα για δείγματα με χαμηλό μικροβιακό φορτίο (Karstens et al., 2019; Salter et al., 2014) καθώς και αυτών που αφορούν στην σύγκριση αποτελεσμάτων κατά την εφαρμογή εργαλείων αποθορυβοποίησης και ομαδοποίησης (García-López et al., 2021; Joos et al., 2020; Kerrigan & D'Hondt, 2022; Nearing et al., 2018).

Επιπλέον, επειδή πρόκειται για ανάλυση πρωτογενών δεδομένων αλληλούχισης που έχουν προκύψει από άγνωστου βακτηριακού περιεχομένου βιολογικά δείγματα και όχι από πρότυπα δείγματα γνωστής βακτηριακής κοινότητας, είναι αναγκαία η προσαρμογή κάποιων βασικών παραμέτρων στα διαθέσιμα δεδομένα που χρησιμοποιούνται στην παρούσα εργασία. Έτσι, επιλέγονται τρεις βασικοί παράμετροι να εξεταστούν ως προς την επιρροή που προσδίδουν στα δεδομένα. Για την επεξεργαστική ροή του DADA2 είναι η επιλογή αποκοπής των paired-end αναγνωσμάτων, το μέγιστο ποσοστό αναμενόμενων σφαλμάτων για το ποιοτικό φιλτράρισμα των δεδομένων καθώς και το ελάχιστο μήκος επικαλυπτόμενης περιοχής για την συγχώνευση των αναγνωσμάτων. Για την ροή επεξεργασίας του Deblur, οι τρεις παράμετροι που εξετάζονται είναι εξίσου το ελάχιστο μήκος επικαλυπτόμενης περιοχής για την συγχώνευση των αναγνωσμάτων, η ελάχιστη βαθμολογία ποιότητας των βάσεων που απαιτείται να παρέχουν τα αναγνώσματα κατά το ποιοτικό φιλτράρισμα των αναγνωσμάτων καθώς και το σημείο αποκοπής των αναγνωσμάτων έτσι ώστε κατά την αποθορυβοποίηση αυτών να παρέχουν όλα τα ίδιο μήκος. Για την επεξεργαστική ροή του VSEARCH, οι παράμετροι είναι δύο που εξετάζονται, αυτή του ελάχιστου μήκους επικαλυπτόμενης περιοχής για την συγχώνευση των αναγνωσμάτων, και αυτή της ελάχιστης βαθμολογίας ποιότητας των βάσεων που απαιτείται να παρέχουν τα αναγνώσματα κατά το ποιοτικό φιλτράρισμα αυτών. Οι τιμές που επιλέγονται να εξεταστούν καθώς και η επιλογή των βέλτιστων για την περαιτέρω ανάλυση των δεδομένων περιγράφονται αναλυτικά παρακάτω στην υποενότητα.

Τέλος, για την απεικόνιση των αποτελεσμάτων που προκύπτουν μετά από κάθε στάδιο επεξεργασίας δεδομένων χρησιμοποιείται η εξειδικευμένη ιστοσελίδα της πλατφόρμας για την οπτικοποίηση δεδομένων (https://view.qiime2.org/). Σημειώνεται ότι για τον προσδιορισμό του μήκους των paired-end αναγνωσμάτων σε bp και των τιμών βαθμολογίας ποιότητας θέσης Phred των αναγνωσμάτων επιλέγεται από το QIIME2 αυτόματα ένα τυχαίο δείγμα 10000 αναγνωσμάτων από το σύνολο δεδομένων.



Σχήμα 2.2 Διάγραμμα ροής της υπολογιστικής διαδικασίας των τριών διαφορετικών επεξεργαστικών ροών που εκτελούνται στην παρούσα διπλωματική εργασία στην πλατφόρμα QIIME2.

2.3.1 Εισαγωγή πρωτογενών δεδομένων

Για την επεξεργασία τους στο QIIME 2, τα δεδομένα εισόδου (input) είναι αναγκαίο να αποθηκευτούν σε μορφή «τεχνουργημάτων» (artifacts της μορφής .qza). Το QIIME2 υποστηρίζει αυτήν την διαδικασία εφαρμόζοντας την μέθοδο qiime tools import, η οποία διαθέτει διάφορες αυτόματες ή μη συναρτήσεις εισαγωγής δεδομένων. Ο αυτόματος τρόπος εισαγωγής των δεδομένων εξαρτάται από το είδος και την μορφή αυτών, καλύπτοντας αρχεία FASTQ, τα οποία η μορφολογία τους ακολουθούν το πρωτόκολλο EMP και Casava 1.8, και αρχεία FASTA που δεν διαθέτουν βαθμολογίες ποιότητας αλληλουχιών. Σε περίπτωση που οι παραπάνω περιπτώσεις δεδομένων δεν αντιπροσωπεύουν τα αρχικά δεδομένα του χρήστη, δίνεται μία εναλλακτική μη αυτόματη μέθοδος εισαγωγής αρχείων στο QIIME2, στην οποία απαιτείται η κατασκευή ενός «δηλωτικού» (manifest) αρχείου και στην συνέχεια τον ορισμό του είδους και μορφής δεδομένων αλληλούχισης.

Πιο αναλυτικά, το manifest αποτελεί αρχείο TSV (tab-separated values) της μορφής .tsv ή .txt που περιέχει 2 στήλες, όπου στην πρώτη ορίζεται η ταυτότητα του κάθε αρχείου πρωτογενούς δεδομένου, ή αλλιώς των δειγμάτων, και στη δεύτερη καταγράφεται το «απόλυτο όνομα διαδρομής» (absolute file path) του ίδιου αρχείου. Συνεπώς, ο ρόλος του αρχείου αυτού είναι η αντιστοίχιση του ονόματος του κάθε δείγματος με τα αρχεία δεδομένων, υποδεικνύοντας την διαδρομή που θα χρειαστεί να ακολουθήσει το πρόγραμμα για την εύρεση των δεδομένων στον υπολογιστή. Στην συνέχεια, απαιτείται ο ορισμός του είδους των αρχείων, με το QIIME2 να προσφέρει τέσσερεις διαφορετικές παραλλαγές εισαγωγής δεδομένων FASTQ χρησιμοποιώντας το manifest για τον σκοπό αυτό. Οι παραλλαγές εξαρτώνται από το αν τα αναγνώσματα είναι paired-end ή single-end και από το αν η βαθμολογία ποιότητας θέσης αναγνωσμάτων αποκωδικοποιείται με βαθμό μετατόπισης PHRED 33 ή 64.

Στην συγκεκριμένη ανάλυση, τα δεδομένα αλληλούχισης δεν αντιπροσωπεύουν τις περιπτώσεις δεδομένων αυτόματης εισαγωγής και αποθήκευσης τους σε .qza στο QIIME2. Συνεπώς, για την επίτευξη της εισαγωγής των παραπάνω πρωτογενών δεδομένων εφαρμόζεται η μη αυτόματη μέθοδος, όπου αρχικά κατασκευάζεται ένα αρχείο manifest και στην συνέχεια χρησιμοποιείται η εντολή εισαγωγής δεδομένων qiime tools import --input-format PairedEndFastqManifestPhred33V2, με σκοπό την εισαγωγή αρχείων FASTQ, τα οποία περιέχουν μικρο-αναγνώσματα paired-end με βαθμό μετατόπισης PHRED 33. Με αυτή την διαδικασία, εισάγονται τα 86 διαθέσιμα αρχεία FASTQ και ένα αρχείο manifest και καταλήγει ένα αρχείο artifact που περιέχει όλο το περιεχόμενο των πρωτογενών δεδομένων.

2.3.2 Αποκοπή μη-βιολογικών εκκινητών

Η πρώτη ουσιαστική επεξεργασία των πρωτογενών δεδομένων αποτελεί η αφαίρεση αλληλουχιών μη-βιολογικής προέλευσης από τα paired-end αναγνώσματα. Συγκεκριμένα, , τα αναγνώσματα λήφθηκαν από την πλατφόρμα Illumina με ένα επιπλέον κομμάτι αλληλουχιών στην αρχή που προέρχεται από του εκκινητές που χρησιμοποιήθηκαν για την ενίσχυση και αλληλούχιση της περιοχής V3-V4 του 16S rRNA γονιδίου. Η αποκοπή αυτών των αλληλουχιών από τα αναγνώσματα είναι αναγκαία έτσι ώστε η περαιτέρω επεξεργασία των δεδομένων να αφορά αποκλειστικά για αναγνώσματα που έχουν βιολογική προέλευση.

Το QIIME2 υποστηρίζει τέτοιου είδους εφαρμογές χρησιμοποιώντας το πρόσθετο (plugin) cutadapt, το οποίο βασίζεται στον τρόπο λειτουργίας του αλγόριθμου cutadapt

(Martin, 2011). Ο συγκεκριμένος αλγόριθμος προσφέρει έναν σημαντικό αριθμό μεθόδων προεπεξεργασίας δεδομένων αλληλούχισης, όπως τον διαχωρισμό αναγνωσμάτων και τον έλεγχο του μήκους και ποιότητας αυτών. Παρόλα αυτά, η πιο βασική λειτουργία που προσφέρει ο αλγόριθμος στο QIIME2 είναι η δυνατότητα εντοπισμού και αφαίρεσης οποιασδήποτε ανεπιθύμητης σειράς αλληλουχίας ορισμένη από τον χρήστη από όλα τα δεδομένα αλληλούχισης.

Για αυτό το σκοπό, στην παρούσα ανάλυση εφαρμόζεται η εντολή qiime cutadapt trim-paired, που αφορά την αποκοπή αλληλουχιών από διαχωρισμένα (demultiplexed) paired-end δεδομένα αλληλούχισης. Οι σειρές αλληλουχιών, οι οποίες δόθηκαν από την πλατφόρμα Illumina που χρησιμοποιήθηκε και πρέπει να αφαιρεθούν, είναι για τα εμπρόσθια αναγνώσματα F: CCTACGGGNGGCWGCAG 3’ και για τα ανάστροφα R: GACTACHVGGGTATCTAATCC 3’. Εκτός από τον ορισμό των ανεπιθύμητων αλληλουχιών, οι υπόλοιποι παράμετροι παραμένουν στην επιπρόσθετη κατάσταση, με αποτέλεσμα να μην υποβληθούν τα δεδομένα σε περαιτέρω επεξεργασία εκτός από την διαδικασία της αποκοπής. Έτσι, σε αυτό το υπολογιστικό στάδιο χρησιμοποιείται το αρχείο artifact των πρωτογενών δεδομένων και προκύπτει ένα νέο αρχείο artifact στο οποίο απουσιάζουν τα τμήματα αλληλουχίας που στοχοποιήθηκαν σε όλα τα αναγνώσματα.

2.3.3 Συγχώνευση paired-end αναγνωσμάτων

Στην παρούσα εργασία, η πλατφόρμα Illumina που χρησιμοποιήθηκε για την λήψη δεδομένων δημιούργησε ζεύγη αναγνωσμάτων με μήκος 2x250 bp από αλληλούχιση των εμπρόσθιων και ανάστροφων κλώνων κάθε θραύσματος DNA που στοχοποιήθηκε, δηλαδή της περιοχής V3-V4 του 16S rRNA γονιδίου, του οποίου το μήκος του έχει προσδιοριστεί να είναι κατά μέσο όρο 458 bp με εύρος 433-483 bp (Vargas-Albores et al., 2017). Εφόσον το μέγεθος του θραύσματος DNA που μελετάται είναι μικρότερο από το διπλάσιο του μήκους των μικρο-αναγνωσμάτων, υπάρχει μια περιοχή επικάλυψης, με αποτέλεσμα να υπάρχει δυνατότητα αντιστοίχισης και συγχώνευσης των paired-end αναγνωσμάτων σε ένα θραύσμα. Η εφαρμογή της συγχώνευσης των paired-end αναγνωσμάτων προηγείται συνήθως από αυτή του φιλτραρίσματος με βάση την ποιότητά τους, διότι κατά την διαδικασία της ένωσης, η επικαλυπτόμενη περιοχή μεταξύ των μικρο-αναγνωσμάτων διορθώνεται από τυχόν σφάλματα αλληλουχιών, αποδίδοντας πιθανώς αναγνώσματα υψηλότερης ποιότητας. Συνεπώς, η ένωση paired-end αναγνωσμάτων αποτελεί το πρώτο βήμα βαριάς επεξεργασίας των δεδομένων και, κατά επέκταση, η ακρίβειά του είναι κρίσιμη για όλες τις μεταγενέστερες αναλύσεις.

Στο QIIME2, η ένωση paired-end αναγνωσμάτων υποστηρίζεται από τον αλγόριθμο VSEARCH (Rognes et al., 2016) χρησιμοποιώντας τη συνάρτηση merge_pairs. Η μέθοδος ένωσης paired-end αναγνωσμάτων βασίζεται στον τρόπο λειτουργίας του αλγόριθμου PEAR (Zhang et al., 2014), ο οποίος υπολογίζει γρήγορα και αξιόπιστα τη βέλτιστη ευθυγράμμιση της επικαλυπτόμενης περιοχής αλληλουχιών των εμπρόσθιων και ανάστροφων αναγνωσμάτων. Πιο αναλυτικά, ο αλγόριθμος συγχωνεύει και βαθμολογεί όλες τις πιθανές επικαλύψεις για κάθε αντιστοιχία paired-end αναγνωσμάτων με μια βαθμολογία συναρμολόγησης (assembly score - AS). Η βαθμολογία υπολογίζεται μέσω ενός πίνακα βαθμολόγησης, ο οποίος προσδίδει στις αναντιστοιχίες μια αρνητική τιμή β ενώ προσδίδει στις αντιστοιχίες μια θετική τιμή α, λαμβάνοντας υπόψη και τις βαθμολογίες ποιότητας των βάσεων. Αφού προσδιοριστούν οι επικαλύψεις με τις υψηλότερες βαθμολογίες συναρμολόγησης, διεξάγεται ένας στατιστικός έλεγχος, αξιολογώντας τη στατιστική

σημαντικότητα των συγχωνευμένων αναγνωσμάτων. Για τον έλεγχο της αξιοπιστίας των συγχωνευμένων αναγνωσμάτων, υπολογίζεται μια τιμή *p-value* με μηδενική υπόθεση τα δύο μικρο-αναγνώσματα που ενώθηκαν να είναι ανεξάρτητα μεταξύ τους. Με τον όρο ανεξάρτητα εννοείται ότι οποιαδήποτε επικάλυψη μεταξύ των δύο αναγνωσμάτων είναι καθαρά τυχαία σύμπτωση. Εάν τα συγχωνευμένα αναγνώσματα δεν περάσουν αυτόν τον έλεγχο ή το μήκος επικάλυψης είναι μικρότερο από ένα όριο που καθορίζεται από τον χρήστη, το ζεύγος των αναγνωσμάτων δεν θα συγχωνευθεί. Διαφορετικά, ο αλγόριθμος διατηρεί το συγχωνευμένο τμήμα, υπολογίζοντας τις νέες βαθμολογίες ποιότητας και διορθώνοντας τα τυχόν σφάλματα των βάσεων των επικαλυπτόμενων περιοχών χρησιμοποιώντας τις βαθμολογίες ποιότητας Phred. Τα μικρο-αναγνώσματα που δεν κατάφεραν να συγχωνευθούν απορρίπτονται στο τέλος της διαδικασίας αυτόματα.

Για την εφαρμογή της μεθόδου δεν απαιτείται η προεπεξεργασία των εισαγόμενων δεδομένων ή συγκεκριμένο μήκος αναγνωσμάτων, ενώ παράλληλα η πλειοψηφία των ελεγχόμενων παραμέτρων που αφορούν στον τρόπο λειτουργίας του merge_pairs προτείνεται από τον κατασκευαστή του αλγόριθμου να παραμείνουν στην επιπρόσθετη κατάστασή τους. Η μοναδική παράμετρος που είναι αναγκαίο να οριστεί από τον χρήστη αποτελεί το ελάχιστο μήκος επικαλυπτόμενης περιοχής ($Overlap_{min}$) των paired-end αναγνωσμάτων. Η παράμετρος αυτή εξαρτάται από το αναμενόμενο κατά προσέγγιση μήκος αλληλουχιών που στοχοποιήθηκε και την πλατφόρμα Illumina που χρησιμοποιήθηκε στην έρευνα, προσδιορίζοντας έτσι την μέγιστη επικάλυψη που θα μπορούσαν να έχουν τα αναγνώσματα.

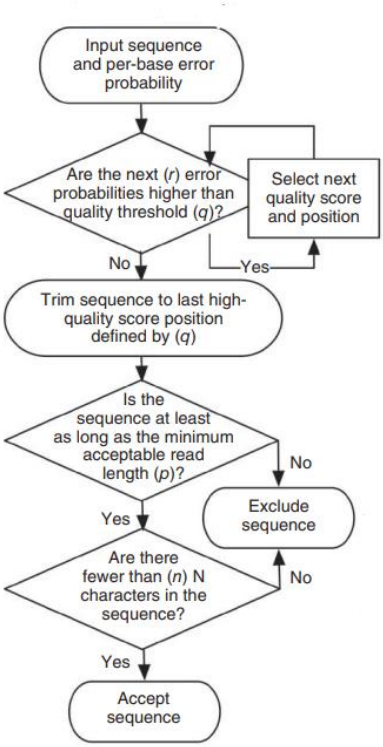
Στην παρούσα ανάλυση, το μήκος της επικαλυπτόμενης περιοχής εκτιμάται να είναι κατά μέσο όρο 42 bp με εύρος 17-67 bp. Χρησιμοποιώντας την εντολή qiime vsearch join-pairs, εφαρμόζονται 3 διαφορετικές τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, $Overlap_{min}$ = 10, 20 και 30, με σκοπό την σύγκριση των αποτελεσμάτων και την εύρεση των βέλτιστων παραμέτρων για τα διαθέσιμα δεδομένα. Έτσι, σε αυτό το υπολογιστικό στάδιο εισάγεται το αρχείο artifact των αποκομένων πρωτογενών δεδομένων και προκύπτουν τρία νέα αρχεία artifacts που αντιστοιχούν στις τρεις διαφορετικές τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής. Το περιεχόμενο αυτών των αρχείων αποτελεί τα συγχωνευμένα paired-end αναγνώσματα που κατάφεραν να διατηρηθούν με τις προαναφερόμενες προϋποθέσεις, και έχουν τις προδιαγραφές να χαρακτηρίζονται από εδώ και στο εξής ως single-end αναγνώσματα.

2.3.4 Ποιοτικό φιλτράρισμα αναγνωσμάτων

Δεδομένου των σφαλμάτων αλληλούχισης που δημιουργούνται από τις ταχέως εξελισσόμενες πλατφόρμες αλληλούχισης, όπως η Illumina, το φιλτράρισμα ποιότητας αποτελεί αναπόσπαστο μέρος της ανάλυσης δεδομένων αλληλούχισης υψηλής απόδοσης. Με αυτόν τον τρόπο, αφαιρούνται χαμηλής ποιότητας αναγνώσματα που ενδεχομένως θα μπορούσαν να υπερεκτιμήσουν τη μικροβιακή ποικιλομορφία του αρχικού βιολογικού δείγματος. Οι πλατφόρμες Illumina παράγουν μια βαθμολογία ποιότητας Phred για κάθε νουκλεοτίδιο, η οποία σχετίζεται με την πιθανότητα λανθασμένης τοποθέτησης του ίδιο νουκλεοτιδίου στο ανάγνωσμα.

Η διαδικασία αφαίρεσης και ελέγχου αναγνωσμάτων χαμηλής ποιότητας υποστηρίζεται από το QIIME2 χρησιμοποιώντας τη μέθοδο q-score (Bokulich et al., 2013) του πρόσθετου quality-filter. Η συγκεκριμένη μέθοδος βασίζεται σε μία αλγοριθμική διαδικασία η οποία χρησιμοποιεί τη βαθμολογία Phred και τις παραμέτρους που καθορίζονται

από το χρήστη με σκοπό την αφαίρεση αλληλουχιών ή ολόκληρων αναγνωσμάτων που δεν πληρούν την επιθυμητή ποιότητα. Οι παράμετροι που ορίζονται από τον χρήστη αποτελούν το ελάχιστο μήκος αναγνώσματος υψηλής ποιότητας (p), ο μέγιστος αριθμός διαδοχικών βάσεων χαμηλής ποιότητας (r), ο μέγιστος αριθμός διφορούμενων βάσεων, που συνήθως κωδικοποιούνται ως N, (n) και η ελάχιστη βαθμολογία ποιότητας Phred (q). Η στρατηγική του αλγόριθμου λειτουργεί κατά βάση ανά νουκλεοτίδιο, περικόπτοντας τα αναγνώσματα στη θέση όπου η ποιότητά τους αρχίζει να πέφτει (**Σχήμα 2.3**). Πιο αναλυτικά, εάν η βαθμολογία ποιότητας των (r) διαδοχικών βάσεων του αναγνώσματος είναι υψηλότερη από (q), ελέγχεται η ποιότητα της επόμενης διαδοχικής σειράς βάσεων. Διαφορετικά, το ανάγνωσμα περικόπτεται στην τελευταία θέση βαθμολογίας υψηλής ποιότητας που ορίζεται από το (q). Στην συνέχεια, προσδιορίζεται το τελικό μήκος του αναγνώσματος, απορρίπτοντάς το στην περίπτωση που είναι μικρότερο από το ελάχιστο αποδεκτό μήκος αναγνώσματος (p). Τέλος, ελέγχεται ο αριθμός N χαρακτήρων του αναγνώσματος, όπου εάν και μόνο προσδιοριστεί μικρότερος από (n), το ανάγνωσμα γίνεται αποδεκτό.



Σχήμα 2.3 Ροή διαδικασίας ποιοτικού φιλτραρίσματος της q-score μεθόδου, όπου (p) το ελάχιστο μήκος ανάγνωσης υψηλής ποιότητας, (r) ο μέγιστος αριθμός διαδοχικών βάσεων χαμηλής ποιότητας, (n) ο μέγιστος αριθμός διφορούμενων βάσεων, που συνήθως κωδικοποιούνται ως N, και (q) η ελάχιστη βαθμολογία ποιότητας Phred. (Bokulich et al., 2013)

Ο συντάκτης του αλγόριθμου συνιστά τις τυπικές τιμές για αυτές τις παραμέτρους, όπου r=3, p=75%, q=3 και n=0, οι οποίες είναι οι προεπιλεγμένες τιμές και στο QIIME2. Παρόλα αυτά, για την καλύτερη ανάλυση μικροβιακών κοινοτήτων, ο κατασκευαστής προτείνει την αύξηση της ελάχιστης ποιότητας Phred (q>=20). Έτσι, διατηρούνται τα πιο αξιόπιστα αναγνώσματα, τα οποία τείνουν να αντιπροσωπεύουν περισσότερο το περιεχόμενο του μεταγονιδιώματος του αρχικού βιολογικού δείγματος.

Συνεπώς, στην παρούσα ανάλυση χρησιμοποιήθηκε η εντολή qiime quality-filter q-score, αφήνοντας τις τιμές των παραμέτρων στην επιπρόσθετη κατάστασή τους, με εξαίρεση

την τιμή της ελάχιστής βαθμολογίας ποιότητας (q). Για την σύγκριση των αποτελεσμάτων και την εύρεση των βέλτιστων παραμέτρων, εφαρμόζεται 3 διαφορετικές τιμές ελάχιστης ποιότητας, όπου $q=Q_{min}=20$, 22 και 26. Τα αρχεία που χρησιμοποιήθηκαν σε αυτό το στάδιο ανάλυσης είναι αυτά που παράχθηκαν από το στάδιο της συγχώνευσης αναγνωσμάτων. Συνολικά, από τα 3 artifacts συγχωνευμένων αναγνωσμάτων, προέκυψαν 9 artifacts, των οποίων το περιεχόμενό τους αποτελεί συγχωνευμένα υψηλής ποιότητας αναγνώσματα με ελάχιστο αριθμό επικάλυψης $Overlap_{min}$= 10, 20 και 30 και ελάχιστη τιμή ποιότητας Q_{min}=20, 22 και 26.

2.3.5 Αποθορυβοποίηση και ομαδοποίηση αναγνωσμάτων

Για την ελαχιστοποίηση των επιπτώσεων των σφαλμάτων που προκύπτουν από την αλληλούχιση του 16S rRNA γονιδίου και την διαχείριση του μεγέθους των δεδομένων, έχουν αναπτυχθεί δύο στρατηγικές για την διερεύνηση αντιπροσωπευτικών βιολογικών αλληλουχιών αμπλικονίων, αυτή της αφαίρεσης θορύβου (denoising), ή αλλιώς αποθορυβοποίησης, και της ομαδοποίησης (clustering). Αντίστοιχα, οι δύο αυτές μέθοδοι παράγουν Λειτουργικές Ταξινομικές Μονάδες (Operational Taxonomic Units, OTUs) και Παραλλαγές Αλληλουχιών Αμπλικονίων (Amplicon Sequence Variants, ASVs). Το λογισμικό QIIME2 διαθέτει δύο pipelines για την αποθορυβοποίηση των δεδομένων, το DADA2 (Callahan et al., 2016) και το Deblur (Amir et al., 2017), ενώ διαθέτει ένα pipeline για την ομαδοποίησή τους, το VSEARCH (Rognes et al., 2016).

2.3.5.1 DADA2

Το pipeline του DADA2, εκτός από το στάδιο τις αποθορυβοποίησης, έχει ενσωματωμένες και τις διαδικασίες του ποιοτικού ελέγχου και της συγχώνευσης των paired-end αναγνωσμάτων. Συγκεκριμένα, η πρώτη λειτουργία του pipeline κατά την εισαγωγή των δεδομένων είναι ο ποιοτικός τους έλεγχος, στον οποίο δίνεται η επιλογή αποκοπής των άκρων που έχουν χαμηλή ποιότητα και στην συνέχεια φιλτράρονται τα αναγνώσματα που γενικά έχουν χαμηλή ποιότητα χρησιμοποιώντας μια μέγιστη τιμή ποσοστού αναμενόμενων σφαλμάτων. Στην συνέχεια, αφού συλλεχθούντα πανομοιότυπα αναγνώσματα (dereplication), τα μικρο-αναγνώσματα υποβάλλονται στην διαδικασία αποθορυβοποίησης, κατά την οποία δημιουργείται ένα παραμετρικό μοντέλο σφάλματος που εκπαιδεύεται από ένα σύνολο αναγνωσμάτων με σκοπό την διόρθωση και σύμπτυξη των εσφαλμένων αλληλουχιών σε ASVs. Ύστερα, το pipeline συμπεριλαμβάνει και την διαδικασία εντοπισμού χιμαιρικών αλληλουχιών, όπου τα ASVs με σχετικά χαμηλή αφθονία συγκρίνονται με τα υπόλοιπα ASVs και ελέγχεται η ομοιότητά τους. Εάν τα "υποπτα" ASVs παρέχουν αλληλουχίες που μπορούν να ανακατασκευαστούν με τον ίδιο ακριβώς τρόπο συνδυάζοντας ένα αριστερό τμήμα και ένα δεξιό τμήμα από δύο διαφορετικά ASVs που παρουσιάζονται με πολύ μεγαλύτερη αφθονία, τότε τα ύποπτα ASVs αναγνωρίζονται ως χιμαιρικά και τα αναγνώσματα αφαιρούνται από τα δεδομένα. Τέλος, ακολουθεί η συγχώνευση των αναγνωσμάτων με προκαθορισμένο ελάχιστο απαιτούμενο μήκος επικάλυψης. Ο λόγος που η εφαρμογή της αποθορυβοποίησης γίνεται πριν την συγχώνευση των paired-end αναγνωσμάτων είναι για την αποφυγή της μετατροπής των βαθμολογίων ποιότητας των βάσεων της επικαλυπτόμενης περιοχής, η οποία, σύμφωνα με τους συγγραφείς του αλγορίθμου, προκαλεί έντονη διακύμανση στην συνολική ποιότητα των αναγνωσμάτων με αποτέλεσμα να παρεμβαίνει αρνητικά στην λειτουργία του αλγόριθμου.

Εφόσον το pipeline περιλαμβάνει τον ποιοτικό έλεγχο και την συγχώνευση των αναγνωσμάτων, τα δεδομένα εισόδου που χρησιμοποιούνται είναι αυτά που προέκυψαν μετά από την αφαίρεση των μη-βιολογικών αλληλουχιών. Για τον προσδιορισμό των βέλτιστων παραμέτρων για τα συγκεκριμένα δεδομένα, ο αλγόριθμος εφαρμόζεται με α) 2 διαφορετικές παραμέτρους αποκοπής άκρων, β) 3 διαφορετικές ανώτατες ποσοστιαίες τιμές αναμενόμενων σφαλμάτων και γ) 3 διαφορετικές τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής. Για την παράμετρο της αποκοπής, στην μία περίπτωση επιλέγεται να μην εφαρμοστεί περαιτέρω αποκοπή στα αναγνώσματα α.1) (no trim), ενώ στην δεύτερη περίπτωση επιλέγεται η αφαίρεση του τελικού (δεξιού) άκρου και των δύο μικρο-αναγνωσμάτων α.2) (trim@). Το σημείο στο οποίο γίνεται η αποκοπή είναι στην $220^{η}$ βάση για τα εμπρόσθια αναγνώσματα και στην $225^{η}$ αντίστοιχα για τα ανάστροφα. Ο λόγος που επιλέγονται αυτά τα σημεία αποκοπής είναι τα ελάχιστα μήκη που παρέχουν τα αντίστοιχα αναγνώσματα μετά την αποκοπή μη-βιολογικών αλληλουχιών. Με αυτόν τον τρόπο, αξιοποιούνται όλα τα αναγνώσματα σε μεταγενέστερες αναλύσεις. Δεν επιλέγεται η μεγαλύτερη αποκοπή του τελικού άκρου διότι υπάρχει κίνδυνος της ολικής αφαίρεσης της επικαλυπτόμενης περιοχής. Επιπλέον, δεν επιλέγεται η αποκοπή του αριστερού άκρου διότι η αφαίρεση αυτού έχει ήδη επιτευχθεί στο προηγούμενο στάδιο της αποκοπής μη-βιολογικών αλληλουχιών.

Για την παράμετρο των ποσοστιαίων αναμενόμενων σφαλμάτων, επιλέγονται οι ανώτερες τιμές $e.e._{max}$= 2.5, 1.5 και 0.5, όπου, σύμφωνα με την προαναφερόμενη εξίσωση, αντιστοιχούν στις ελάχιστες τιμές βαθμολογίας ποιότητας των βάσεων που θα πρέπει να παρέχουν κατά προσέγγιση τα αναγνώσματα: Q_{min}= 20 ή 22 ή 26. Όσον αφορά στην τιμή του ελάχιστου μήκους επικαλυπτόμενης περιοχής, επιλέγονται οι τιμές $Overlap_{min}$= 10, 20 και 30. Όλες οι υπόλοιποι παράμετροι που αφορούν την εκτέλεση του αλγόριθμου παραμένουν στις τιμές που προτείνονται από συγγραφείς των αλγορίθμων.

Συνεπώς, για την αποθορυβοποίηση των δεδομένων με την χρήση του αλγόριθμου DADA2, χρησιμοποιείται η εντολή qiime dada2 denoise-paired, με τα δεδομένα εισόδου να είναι το artifact που παράχθηκε από την διαδικασία αφαίρεσης των εκκινητών. Τα δεδομένα εξόδου που προέκυψαν είναι 18 αρχεία artifacts, των οποίων το περιεχόμενό τους αποτελείται από ASVs και την συχνότητα με την οποία παρουσιάζονται στα δεδομένα για κάθε σενάριο διαφορετικών τιμών παραμέτρων.

2.3.5.2 Deblur

Λόγω της φύσης του ίδιου του αλγόριθμου, το pipeline του Deblur εκτός από την διαδικασία αποθορυβοποίησης περιλαμβάνει και την αποκοπή των εισαγόμενων αναγνωσμάτων σε ένα συγκεκριμένο μήκος, προκειμένου να προσδιοριστούν επιτυχώς οι αποστάσεις Hamming. Επιπλέον, συμπεριλαμβάνεται η διαδικασία της απαλοιφής πανομοιότυπων αναγνωσμάτων (dereplication), καθώς και του εντοπισμού χιμαιρικών αλληλουχιών πριν την ολοκλήρωση της επεξεργασίας. Για τον εντοπισμό των χιμαιρικών αλληλουχιών, ο Deblur εφαρμόζει μια διαδικασία επικύρωσης ότι οι αλληλουχίες των παραγόμενων ASVs είναι βιολογικής προέλευσης χρησιμοποιώντας μια βάση δεδομένων. Η βάση δεδομένων που χρησιμοποιείται είναι η Greengenes, και η επικύρωση μια αλληλουχίας επιτυγχάνεται με την τουλάχιστον 88% ομοιότητάς της με μία αντίστοιχη υπάρχουσα των δεδομένων αναφοράς. Εάν αποτύχει η αντιστοίχηση, το παραγόμενο ASV θεωρείται αποτέλεσμα σφάλματος και αφαιρείται από τα δεδομένα. Μία ακόμα λειτουργία εντοπισμού χιμαιρικών αλληλουχιών που προσφέρει το pipeline είναι ο προσδιορισμός ASVs με πολύ

χαμηλή αφθονία, και η διατήρηση εκείνων που η αλληλουχία τους εμφανίζεται τουλάχιστον 10 φορές συνολικά σε όλα τα δείγματα.

Συνεπώς, τα δεδομένα εισόδου πρέπει να έχουν προεπεξεργαστεί κατάλληλα ως προς την συγχώνευση των paired-end αναγνωσμάτων και τον ποιοτικό τους έλεγχο. Για αυτόν τον λόγο, τα αρχεία που χρησιμοποιούνται για την εκτέλεση του Deblur είναι τα 9 artifacts που προέκυψαν από τον προαναφερόμενο ποιοτικό έλεγχο, των οποίων το περιεχόμενό του αποτελεί συγχωνευμένα υψηλής ποιότητας αναγνώσματα με ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής $Overlap_{min}$= 10, 20 και 30 και ελάχιστη τιμή βαθμολογία ποιότητας ανά βάση Q_{min}=20, 22 και 26. Όσον αφορά στην παράμετρο αποκοπής, επιλέγεται να εφαρμοστούν δύο διαφορετικά σημεία αποκοπής. Στην μία περίπτωση τα αναγνώσματα αποκόπτονται στην $250^{η}$ βάση, ενώ στην δεύτερη περίπτωση στην $380^{η}$ βάση. Το κριτήριο επιλογής των δύο αυτών σημείων είναι η προσπάθεια να συμπεριληφθούν όσον το δυνατόν περισσότερα αναγνώσματα στην διαδικασία της αποθορυβοποίησης, διότι τα αναγνώσματα που έχουν μικρότερα μήκη απορρίπτονται αυτόματα και δεν μπορούν να λάβουν μέρος σε καμία περαιτέρω ανάλυση.

Επομένως, για την αποθορυβοποίηση των δεδομένων αλληλούχισης με την χρήση του αλγόριθμου Deblur, εφαρμόστηκε η εντολή qiime deblur denoise-16S με δύο διαφορετικά σημεία αποκοπής, τα οποία είναι trim@= 250 και 380. Τα δεδομένα εισόδου είναι αυτά που παράχθηκαν ακριβώς μετά τον ποιοτικό έλεγχο, δηλαδή 9 artifacts, και τα συνολικά δεδομένα εξόδου προέκυψαν ίσα με 18 artifact, των οποίων το περιεχόμενό αφορά ASVs και την συχνότητα με την οποία παρουσιάζονται στα δεδομένα.

2.3.5.3 VSEARCH

Για την ομαδοποίηση των δεδομένων αλληλούχισης χρησιμοποιείται το pipeline του VSEARCH που παρέχει την μέθοδο ομαδοποίησης κλειστής αναφοράς, δηλαδή με την χρήση μίας βάσης δεδομένων. Το pipeline που παρέχει το QIIME2 συμπεριλαμβάνει μόνο την διαδικασία ομαδοποίησης, κατά την οποία είναι αναγκαίο να δοθεί η βάση δεδομένων καθώς και το ελάχιστο ποσοστό ομοιότητας που πρέπει να έχουν οι αλληλουχίες για την αντιστοίχησή τους με τις αλληλουχίες αναφοράς. Εάν υπάρχει επιτυχής αντιστοίχηση, δημιουργούνται νέα OTUs ή ενσωματώνονται σε ένα ήδη υπάρχον. Συνεπώς, τα δεδομένα εισόδου πρέπει να έχουν υποβληθεί σε μία κατάλληλη προεπεξεργαστική διαδικασία έτσι ώστε τα παραγόμενα OTUs να είναι όσο τον δυνατόν υψηλότερης ακρίβειας γίνεται. Επιπλέον, τα δεδομένα εξόδου θα πρέπει να υποβληθούν σε μία ακόμα επεξεργασία που αφορά τον έλεγχο και διαχείριση χιμαιρικών αλληλουχιών και OTUs.

Εφόσον τα αναγνώσματα εισόδου πρέπει να έχουν ήδη υποβληθεί στην διαδικασία προεπεξεργασίας, χρησιμοποιούνται τα δεδομένα που προέκυψαν από την συγχώνευση και τον ποιοτικό έλεγχο με τις διάφορες τιμές παραμέτρων ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας, δηλαδή 9 artifacts. Όσον αφορά στην αποκοπή των αναγνωσμάτων, για την εφαρμογή του VSEARCH δεν είναι απαραίτητο τα αναγνώσματα να είναι σε ένα συγκεκριμένο μήκος για την παραγωγή OTUs. Δεδομένου αυτού και του γεγονότος ότι δεν υπάρχει διαθέσιμο ανεξάρτητο εργαλείο στο QIIME2 που να εκτελεί αυτήν την διαδικασία, τα αναγνώσματα δεν υποβάλλονται σε αποκοπή συγκεκριμένου μήκους. Επίσης, τα δεδομένα εισόδου θα πρέπει να υποστούν την διαδικασία συλλογής πανομοιότυπων αναγνωσμάτων (dereplication), χρησιμοποιώντας την εντολή qiime vsearch dereplicate-sequences. Στην συνέχεια, δίνεται η εντολή qiime vsearch cluster-

features-closed-reference, για την εκτέλεση της ομαδοποίησης των αναγνωσμάτων σε OTUs. Η βάση δεδομένων που χρησιμοποιήθηκε είναι η SILVA, η οποία παρέχει τις πιο αναβαθμισμένες αλληλουχίες αναφοράς στοχευμένων γονιδίων. Επιπλέον, ορίζεται ως ελάχιστο ποσοστό ομοιότητας για την ομαδοποίησης των αλληλουχιών η τιμή 97%. Από τα παραγόμενα δεδομένα, εξετάζεται η ύπαρξη χιμαιρικών OTUs χρησιμοποιώντας την εντολή qiime vsearch uchime-denovo, η οποία ακολουθεί την *de novo* μέθοδο ανίχνευσης χιμαιρικών αλληλουχιών όπου τα πιο σχετικά άφθονα OTUs χρησιμοποιούνται ως βάση δεδομένων και συγκρίνονται με τα λιγότερα άφθονα αντίστοιχα. Εφόσον και αν εντοπιστούν αυτά τα στοιχεία, η διαδικασία ολοκληρώνεται με την αφαίρεση των χιμαιρικών OTUs από τα δεδομένα.

Επομένως, από όλη την προαναφερόμενη διαδικασία, προέκυψαν 9 διαφορετικά αρχεία artifacts, όπου το καθένα περιέχει τα παραγόμενα OTUs και την συχνότητα με την οποία παρουσιάζονται στα δεδομένα με διαφορετικές παραμετρικές τιμές ελάχιστης επικάλυψης και ποιότητας. Επιπλέον, είναι απαλλαγμένα από χιμαιρικές αλληλουχίες.

2.3.6 Ταξινομική ανάθεση

Αφού τα δεδομένα αλληλούχισης σε αυτό το σημείο δεν χρειάζονται περαιτέρω επεξεργασία ως προς την ποιότητά τους, το επόμενο βήμα είναι η διερεύνηση των ταξινομικών πληροφοριών τους. Αυτή η διαδικασία επιτυγχάνεται με την ταξινομική ανάθεση των παραγόμενων ASVs και OTUs. Το pipeline που παρέχει το QIIME2 για αυτόν το σκοπό είναι του Scikit-learn, χρησιμοποιώντας την εντολή qiime feature-classifier classify-sklearn. Γενικά, το Scikit-learn είναι ένα πακέτο μηχανικής μάθησης στην Python το οποίο παρέχει ένα ευρύ φάσμα αλγορίθμων και εργαλείων για την κατασκευή και την αξιολόγηση μοντέλων μηχανικής μάθησης. Το συγκεκριμένο πακέτο είναι ενσωματωμένο και στην πλατφόρμα QIIME2, από το οποίο χρησιμοποιείται ο αλγόριθμος Naïve Bayes για την ταξινομική ανάθεση των ASVs και OTUs. Ο αλγόριθμος αυτός εμπίπτει στην κατηγορία των ταξινομητών εποπτευόμενης μάθησης, που σημαίνει ότι χρησιμοποιεί μία βάση δεδομένων για να προβλέψει την ταξινομική προέλευση μιας αλληλουχίας. Έτσι, ανάλογα με το ποσοστό εμπιστοσύνης που προκύπτει για κάθε αντιστοίχηση, τα ASVs και OTUs κατηγοριοποιούνται στα διάφορα ταξινομικά επίπεδα. Όσο αυξάνεται το ποσοστό εμπιστοσύνης αντιστοίχησης, τόσο αυξάνεται και η ευκρίνεια της ταξινομικής προέλευσης, που συνεπάγεται με την ταυτοποίηση της βακτηριακής προέλευσης η οποία φτάνει μέχρι και το επίπεδο του είδους.

Φυσικά, η ακρίβεια της ταξινομικής ανάθεσης εξαρτάται άμεσα από την επιλογή των δεδομένων αναφοράς. Για αυτόν το λόγο, στην παρούσα εργασία χρησιμοποιείται η βάση δεδομένων της SILVA, η οποία παρέχεται με μεγάλη αξιοπιστία από την πλατφόρμα του QIIME2. Επιπλέον, το QIIME2 παρέχει την επιλογή να χρησιμοποιηθεί ο αλγόριθμος που είναι ήδη εκπαιδευμένος στην βάση δεδομένων της SILVA, με αποτέλεσμα να μην είναι αναγκαία η περαιτέρω εκπαίδευση του ταξινομητή, εξοικονομώντας έτσι υπολογιστικούς πόρους και χρόνο στην διαδικασία της ταξινομική ανάλυσης των δεδομένων.

Τα δεδομένα εισόδου είναι όλα τα προαναφερόμενα artifacts που εμπεριέχουν τα παραγόμενα ASVs και OTUs στις διάφορες τιμές παραμέτρων επικάλυψης, αποκοπής και ποιοτικού ελέγχου, δηλαδή συνολικά 45 artifacts. Τα δεδομένα εξόδου προέκυψαν συνολικά ίσα με 45 artifacts που παρέχουν την αντιστοίχηση των ASVs και OTUs με τις διάφορες ταξινομικές κατηγορίες. Επίσης, χρησιμοποιείται η εντολή qiime taxa filter-table για την

αφαίρεση ASVs/OTUs που δεν κατάφεραν να ταξινομηθούν σε καμία κατηγορία της βάσης δεδομένων της SILVA (unassigned).

2.3.7 Επιλογή τελικών τιμών παραμέτρων

Για την επιλογή των βέλτιστων τιμών παραμέτρων για κάθε workflow, τα βασικά αποτελέσματα που προέκυψαν μετά την ολοκλήρωση της διαδικασίας της ταξινομικής ανάθεσης συγκεντρώνονται και συγκρίνονται μεταξύ τους ξεχωριστά για το καθένα workflow. Τα αποτελέσματα αυτά αποτελούν ο αριθμός των διατηρητέων ταξινομημένων αναγνωσμάτων, ο αριθμός των ταξινομημένων ASVs και OTUs και ο αριθμός των βακτηριακών ταξινομήσεων. Είναι σημαντικό να σημειωθεί ότι για τα pipelines Deblur και VSEARCH λαμβάνεται υπόψη όλη η επεξεργαστική ροή (workflow) που απαιτείται για την εφαρμογή τους. Αυτό οφείλεται στο ότι η εφαρμογή αυτών των pipelines χρειάζεται μια σειρά επεξεργασιών των δεδομένων όπως, την συγχώνευση και το ποιοτικό φιλτράρισμα των αναγνωσμάτων, η οποία δεν πραγματοποιείται αυτόματα σε αντίθεση με την εφαρμογή του DADA2.

Η επιλογή των τελικών τιμών παραμέτρων βασίζεται στην ανάκτηση όσον τον δυνατόν περισσότερης ταξινομικής πληροφορίας γίνεται από τα πρωτογενή δεδομένα. Έτσι, δίνεται περισσότερη έμφαση στους αυξημένους αριθμούς βακτηριακών ταξινομικών μονάδων και διατηρητέων αναγνωσμάτων, με αποτέλεσμα την ανάκτηση της μεγαλύτερης ακρίβειας της ταξινομικής προέλευσης των αλληλουχιών καθώς και της σχετικής τους αφθονίας στο αρχικό δείγμα. Ο αριθμός των παραγόμενων ASVs και OTUs χρησιμοποιείται κυρίως για την αξιολόγηση της απόδοσης αυτών σε μεταγενέστερο στάδιο , όπου θεωρητικά όσο ο αριθμός τους τείνει να είναι ίσος με τον αριθμό των βακτηριακών ταξινομήσεων, τόσο μεγαλύτερη θα είναι και η ταξινομική τους ευκρίνεια. Παράλληλα, λαμβάνεται υπόψη το επίπεδο αξιοπιστίας των ίδιων παραμέτρων, καθώς και το γεγονός ότι οι τελικώς επιλεγμένες τιμές παραμέτρων και των τριών workflows να είναι όσο τον δυνατόν γίνεται παρόμοιοι για να είναι πιο αντικειμενική η περαιτέρω σύγκρισή τους.

2.3.8 Αφαίρεση πιθανών επιμολύνσεων

Έχοντας επιλέξει τα κυρίαρχα σύνολα δεδομένων για την τελική ανάλυση των δεδομένων, γίνεται δυνατή εις βάθος εξέταση των αποτελεσμάτων της ταξινομικής ανάθεσης στα διαφορετικά workflows. Σε αυτό το στάδιο, είναι αναμενόμενο να παρατηρηθούν στοιχεία στα δεδομένα που θεωρητικά μπορούν να χαρακτηριστούν αποτέλεσμα επιμόλυνσης δεδομένου ότι γενικά η πιθανότητα να προκύψουν επιμολύνσεις σε δεδομένα αλληλούχισης αμπλικονίων είναι πολύ υψηλή. Ως εκ τούτου, είναι απαραίτητο να εντοπιστούν, να ελαχιστοποιηθούν και να φιλτραριστούν μολυσματικές αλληλουχίες κατά την βιοπληροφορική ανάλυση δεδομένων αλληλούχισης για την αποφυγή εισαγωγή μεροληψιών στα τελικά αποτελέσματα. Στην παρούσα εργασία, παρατηρήθηκαν 3 κατηγορίες επιμολύνσεων στα δεδομένα, συμπεριλαμβανομένου αυτής των μη-στοχευόμενων ταξινομικών κατηγοριών, των σημαντικά χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών και μιας σημαντικά υψηλής σχετικής αφθονίας και βιολογικά μη αναμενόμενης ταξινομικής κατηγορίας.

Αρχικά, πραγματοποιήθηκε η αφαίρεση των μη-στοχευόμενων ταξινομικών κατηγοριών χρησιμοποιώντας η εντολή qiime taxa filter-table, κατά την οποία αφαιρέθηκαν

από τα δεδομένα τα ASVs/OTUs που ταξινομήθηκαν σε αρχαία (archaea), μιτοχόνδρια (mitochondria), χλωροπλάστες (chloroplast) και ευκαριώτα (eukaryota). Τα στοιχεία αυτά λογικά προέκυψαν λόγω της φύσης του 16S rRNA γονιδίου όπου κατέληξαν να ενισχυθούν οι αλληλουχίες αυτών κατά την PCR. Εφόσον οι προαναφερόμενες ταξινομικές κατηγορίες δεν αντικατοπτρίζουν την ανάλυση που εκτελείται, αφαιρούνται και από τα 3 σύνολα δεδομένων ανάλυσης. Επιπλέον, χρησιμοποιώντας την ίδια εντολή, φιλτράρονται εξίσου οι ταξινομικές κατηγορίες που δεν κατάφεραν να ταξινομηθούν σε τουλάχιστον επίπεδο φυλής ή που ταξινομήθηκαν στην κατηγορία του «μη καλλιεργημένου» (uncultured). Ο λόγος για τον οποίο αυτά τα στοιχεία εντοπίστηκαν και στα 3 σύνολα δεδομένων βασίζεται στην αδυναμία της βάσης δεδομένων SILVA να παρέχει της απαραίτητες ταξινομικές κατηγορίες για να χαρακτηρίσει όλα τα ASVs/OTUs.

Οι επόμενες ταξινομικές κατηγορίες που αφαιρούνται από τα σύνολα δεδομένων είναι αυτές που παρέχουν σημαντικά χαμηλή σχετική αφθονία ή αντιπροσωπεύονται από μοναδικά ASVs/OTUs στο κάθε σύνολο εφαρμόζοντας την εντολή qiime feature-table filter-features. Αυτά τα στοιχεία, που κατάφεραν να ταξινομηθούν, ενδέχεται να μην αντιπροσωπεύουν την πραγματική βιολογική ποικιλότητα, αλλά να έχουν προκύψει από σφάλματα της ενίσχυσης PCR, της αλληλούχισης ή και σε κάποιες περιπτώσεις λόγω επιμόλυνσης και η συγκράτησή τους μπορεί να οδηγήσει σε εσφαλμένα αυξημένη ποικιλομορφία στις περαιτέρω αναλύσεις (Cao et al., 2020). Έτσι, αφαιρούνται οι ταξινομικές κατηγορίες που παρέχουν σχετική αφθονία μικρότερη από 0,002% στο κάθε σύνολο δεδομένων ξεχωριστά. Για την διερεύνηση της επίδρασης του φιλτραρίσματος με βάση την σχετική αφθονία στα δεδομένα που προέκυψαν από την εφαρμογή των διαφορετικών workflows, κατασκευάζονται επιπλέον διαγράμματα ράβδων και Venn πριν και μετά το φιλτράρισμα. Ο σκοπός αυτών των διαγραμμάτων είναι για την εξέταση του αριθμού και των κοινών ταξινομικών κατηγοριών στα διαφορετικά ταξινομικά επίπεδα αντίστοιχα. Επιπρόσθετα, για τον έλεγχο της ταξινομικής σύνθεσης των βιολογικών δειγμάτων, κατασκευάζονται διαγράμματα ράβδων που απεικονίζουν την σχετική αφθονία των βακτηριακών ταξινομικών κατηγοριών σε επίπεδο φυλής σε κάθε δείγμα και για τα τρία σύνολα δεδομένων, χρησιμοποιώντας την εντολή qiime taxa barplot. Τέλος, με την εφαρμογή της εντολής qiime taxa filter-table, αφαιρούνται οι ταξινομικές κατηγορίες και από τα τρία σύνολα δεδομένων που θεωρούνται μη αναμενόμενες για τα δείγματα αίματος και χαρακτηρίζονται αποτέλεσμα επιμόλυνσης βάση βιβλιογραφίας.

2.3.9 Ανάλυση ποικιλομορφίας

Σκοπός του βιοπληροφορικού σταδίου της ανάλυσης ποικιλομορφίας είναι η διερεύνηση και η σύγκριση της α-ποικιλομορφίας μεταξύ των δειγμάτων. Για την επίτευξη αυτού του σκοπού, στην παρούσα εργασία επιχειρείται ο προσδιορισμός των δεικτών α-ποικιλομορφίας Shannon, ένα ποσοτικό μέτρο του πλούτου και της ομοιομορφίας μιας βακτηριακής κοινότητας, για κάθε δείγμα του κάθε συνόλου δεδομένων που προέκυψαν από την εφαρμογή των 3 διαφορετικών workflows. Η πλατφόρμα του QIIME2 παρέχει την δυνατότητα υπολογισμού αυτού του δείκτη μέσω της εντολής qiime diversity core-metrics-phylogenetic. Σημειώνεται ότι η συγκεκριμένη εντολή προσδιορίζει ταυτόχρονα και άλλες μετρήσεις ποικιλομορφίας, συμπεριλαμβανομένου των αποστάσεων β-ποικιλομορφίας, αλλά η παρούσα διπλωματική εργασία επικεντρώνεται στις μετρήσεις που αφορούν μόνο στο δείκτη του Shannon. Η εφαρμογή της εντολής υπολογισμού δεικτών ποικιλομορφίας απαιτεί ως δεδομένα εισόδου το artifact που εμπεριέχει τα ASVs/OTUs καθώς και τον αριθμό των

αναγνωσμάτων που τα συνοδεύουν, ένα artifact που εμπεριέχει τις φυλογενετικές σχέσεις των αλληλουχιών που παρέχουν τα ASVs/OTUs και μία τιμή ορισμένη από την χρήστη που αφορά το βάθος δειγματοληψίας, ή αλλιώς αλληλούχισης. Συνεπώς, πριν τον προσδιορισμό των δεικτών Shannon, η εντολή προϋποθέτει να κατασκευαστεί ένα φυλογενετικό δέντρο από τα παραγόμενα ASVs/OTUs, το οποίο δεν χρησιμοποιείται για τον υπολογισμό των μετρήσεων του δείκτη Shannon, αλλά για τις άλλες εξαγόμενες μετρήσεις. Επίσης, πρέπει να οριστεί και ο αριθμός αναγνωσμάτων που θα πρέπει να παρέχει το κάθε δείγμα για το κάθε διαφορετικό σύνολο δεδομένων των 3 workflows.

Προκειμένου να κατασκευαστεί το φυλογενετικό δέντρο του κάθε συνόλου δεδομένων, είναι αναγκαίο να προσδιοριστούν οι φυλογενετικές και συγγενικές σχέσεις των αλληλουχιών που ταξινομήθηκαν. Η διαδικασία αυτή επιτυγχάνεται χρησιμοποιώντας την εντολή qiime phylogeny align-to-tree-mafft-fasttree, η οποία αξιοποιεί τον αλγόριθμο FastTree (Price et al., 2010) του οποίου η βασική λειτουργία είναι η ευθυγράμμιση των αλληλουχιών και η δημιουργία φυλογενετικού δέντρου βάσει των κοινών αντιστοιχίσεων ευθυγράμμισης. Τα δεδομένα εισόδου είναι τα artifacts που παρέχουν τις αλληλουχίες των ASVs και OTUs και τα δεδομένα εξόδου είναι οι φυλογενετικές πληροφορίες αυτών των αλληλουχιών μέσω φυλογενετικού δέντρου.

Δεδομένου ότι ο αριθμός αναγνωσμάτων για κάθε δείγμα ποικίλει, πρέπει να επιλεχθεί το κατάλληλο βάθος αλληλούχισης για το κάθε σύνολο δεδομένων έτσι ώστε να είναι όσο τον δυνατόν αντικειμενικές και αξιόπιστες οι μετρήσεις ποικιλομορφίας και στατιστικοί έλεγχοι που θα εφαρμοστούν στην συνέχεια. Η διαδικασία αυτή επιτυγχάνεται με την κατασκευή καμπυλών αραίωσης (rarefaction curves), οι οποίες αποτελούν μια διαγραμματική απεικόνιση του αριθμού των παρατηρούμενων ASVs/OTUs ως συνάρτηση των αριθμών των αναγνωσμάτων/αλληλουχιών που τα αντιπροσωπεύουν για κάθε δείγμα. Με αυτόν τον τρόπο, δίνεται η δυνατότητα του προσδιορισμού του βάθους αλληλούχισης που απαιτεί το κάθε δείγμα προκειμένου να καλυφθεί επαρκώς η ποικιλομορφία τους. Συνεπώς, για την εξέταση των καμπύλων αραίωσης, χρησιμοποιείται η εντολή qiime diversity alpha-rarefaction, κατά την οποία απαιτούνται ως δεδομένα εισόδου τα αρχεία των ASVs/OTUs καθώς και των φυλογενετικών δέντρων που παράχθηκαν σε προηγούμενο στάδιο για το κάθε σύνολο δεδομένων. Το βάθος αλληλούχισης (d), δηλαδή ο αριθμός των αναγνωσμάτων που εκτιμάται ότι καλύπτει όλη την ποικιλομορφία των δειγμάτων για το κάθε σύνολο, ορίζεται για το DADA2 ίσο με d_{DADA}=22157, για το Deblur ίσο με d_{Deblur}=8459 και για το VSEARCH ίσο με $d_{VSEARCH}$=24059.

Στην συνέχεια, υπολογίζονται οι μετρήσεις α-ποικιλομορφίας Shannon σε κάθε δείγμα για το κάθε σύνολο δεδομένων εφαρμόζοντας την προαναφερόμενη εντολή και κατασκευάζεται ένα θηκόγραμμα αντίστοιχα για τις τρείς βασικές ομάδες δειγμάτων, δηλαδή των δειγμάτων αίματος πριν (before_treatment) και μετά (after_treatment) την χορήγηση αντιψυχωσικών φαρμάκων και τα δείγματα αρνητικού ελέγχου (controls), χρησιμοποιώντας την εντολή qiime diversity alpha-group-significance. Με την χρήση της ίδιας εντολής, εκτελείται αυτόματα και ο στατιστικός έλεγχος Kruskal-Wallis μεταξύ των τιμών τις α-ποικιλομορφίας των 3 αυτών ομάδων για να εξεταστούν τυχόν σημαντικές διαφορές.

3 Αποτελέσματα

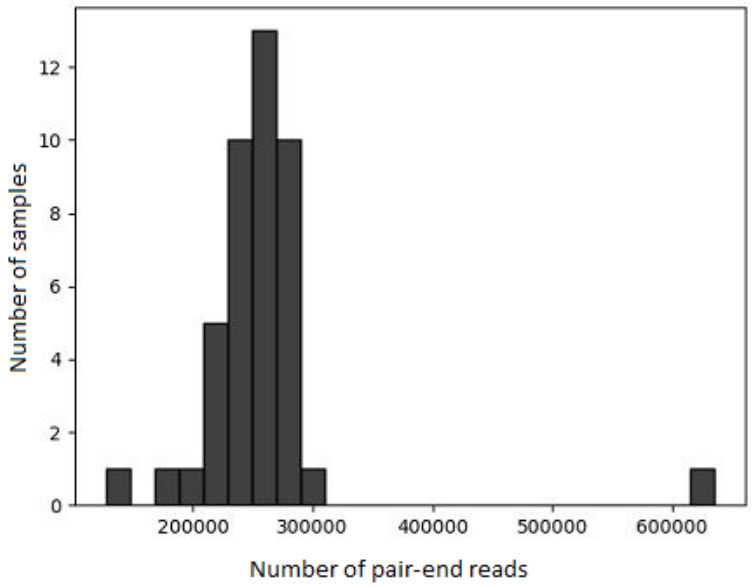
3.1 Βασικά Χαρακτηριστικά Πρωτογενών Δεδομένων

Με την εισαγωγή των πρωτογενών δεδομένων στο QIIME2, δίνεται η δυνατότητα να μελετηθούν τα βασικά χαρακτηριστικά αυτών, τα οποία περιλαμβάνουν τον συνολικό αριθμό αναγνωσμάτων που παρέχει το κάθε δείγμα, μία προσεγγιστική τιμή του μήκους των paired-end αναγνωσμάτων σε bp, καθώς και μια διαγραμματική προσέγγιση των τιμών βαθμολογίας ποιότητας θέσης των αναγνωσμάτων. Υπενθυμίζεται ότι για τις προσεγγιστικές τιμές του μήκους των αναγνωσμάτων και της βαθμολογίας ποιότητας θέσης επιλέχθηκε αυτόματα από το πρόγραμμα ένα τυχαίο δείγμα 10000 αναγνωσμάτων από τα συνολικά. Επιπλέον, ο αριθμός paired-end αναγνωσμάτων ταυτίζεται με τον αριθμό εμπρόσθιων (forward) και ανάστροφων (reverse) μικρο-αναγνωσμάτων αντίστοιχα (**Παράρτημα 1**).

Ο συνολικός αριθμός των παραγόμενων paired-end αναγνωσμάτων είναι ίσος με 11148151 (**Πίνακας 3.1**), με το κάθε δείγμα να παρέχει κατά μέσο όρο 259259 αναγνώσματα. Ο ελάχιστος αριθμός paired-end αναγνωσμάτων που παρουσιάζεται στα δείγματα είναι ίσος με 126881 και ο μέγιστος 635568, που σημαίνει ότι τα δείγματα παρέχουν κατά πολύ μεγαλύτερο αριθμό παραγόμενων αναγνωσμάτων από τον προτεινόμενο για την ικανοποιητική περιγραφή της βακτηριακής σύστασής του (Bukin et al., 2019). Ιδανικά, στο τέλος της βιοπληροφορικής επεξεργασίας θα ήταν επιθυμητό ο αριθμός διατηρητέων αναγνωσμάτων να καταλήξουν εξίσου εντός του ορίου των 10000 με 15000 αναγνωσμάτων (Bukin et al., 2019). Τα περισσότερα δείγματα παρέχουν αριθμό αναγνωσμάτων αρκετά κοντά στο μέσο όρο, με εξαίρεση αυτών των ακραίων τιμών (**Σχήμα 3.1**). Συγκεκριμένα, το δείγμα με τον ελάχιστο αριθμό paired-end αναγνωσμάτων, που αντιστοιχεί σε δείγμα ελέγχου (NTC), παρέχει σημαντικά χαμηλότερο σε σχέση με τα υπόλοιπα δείγματα, ενώ το δείγμα με τον μέγιστο αριθμό, που αντιστοιχεί σε δείγμα αίματος του πρώτου ασθενή, παρουσιάζει ακόμα πιο έντονη διαφορά (**Παράρτημα 1**).

Πίνακας 3.1 Γενική περιγραφή αριθμού πρωτογενών αναγνωσμάτων που παρέχουν συνολικά τα δείγματα.

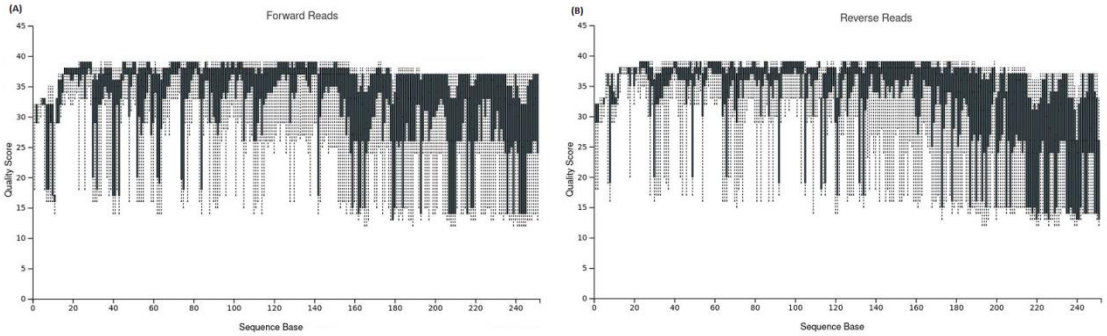
	Αριθμός paired-end αναγνωσμάτων
Ελάχιστος	126,881
Μέσος όρος	259,259
Μέγιστος	635,568
Συνολικός	11,148,151



Σχήμα 3.1 Ιστόγραμμα συχνότητας δειγμάτων που παρέχουν μία δεδομένη τιμή αριθμού paired-end αναγνωσμάτων.

Το μήκος της συντριπτικής πλειοψηφίας των paired-end αναγνωσμάτων των πρωτογενών δεδομένων (91%) είναι ίσο με 251 bp, που είναι ισοδύναμο με το μέγιστο μήκος που είναι ικανή να αλληλουχίσει η πλατφόρμα Illumina MiSeqPE250, η οποία παράγει paired-end 2 x 250 bp αναγνώσματα και η προθήκη μίας τελικής βάσης στα μικρο-αναγνώσματα οφείλετε στην τεχνολογίας της (**Πίνακας 3.2**). Τα δεδομένα φέρουν και αναγνώσματα μικρότερου μήκους, αλλά δεν απέχουν πολύ από το προαναφερόμενο μέγιστο μήκος. Το ελάχιστο μήκος των εμπρόσθιων αναγνωσμάτων είναι 237 bp και των ανάστροφων αναγνωσμάτων 246 bp αντίστοιχα. Τα ελάχιστα μήκη αντιστοιχούν στο 2% του συνολικού πληθυσμού των αναγνωσμάτων. Το μήκος των υπόλοιπων αναγνωσμάτων (7%) προέκυψε ίσο με 250 bp.

Η ποιότητα των πρωτογενών αναγνωσμάτων παρουσιάζεται αρκετά ικανοποιητική. Κατά προσέγγιση, η βαθμολογία ποιότητας Phred ανά βάση εκτιμάται να είναι ίσο με 20, ενώ μικρότερες τιμές εκτιμούνται να είναι μεγαλύτερες από 10 (**Σχήμα 3.2**). Αυτό σημαίνει ότι το ποσοστό σφάλματος ανά βάση είναι περίπου 0.01%, ενώ στην χειρότερη περίπτωση είναι 0.1%. Αναμενόμενη είναι η παρατήρηση των βαθμολογιών βάσεων στα άκρα των αναγνωσμάτων που τείνουν να παρέχουν χαμηλότερη βαθμολογία ποιότητας σε σχέση με το κεντρικό του κομμάτι, με το φαινόμενο να παρουσιάζεται πιο έντονα στο δεξί άκρο (Tan et al., 2019). Το γεγονός ότι η ποιότητα των βάσεων των πρωτογενών δεδομένων είναι ικανοποιητική δεν αναιρεί την ανάγκη του ποιοτικού ελέγχου σε μεταγενέστερες αναλύσεις, διότι το ποσοστό 0.1% πιθανότητας να έχει προκύψει μια βάση λόγω σφάλματος είναι αρκετά σημαντικό και επιβλαβές στην έκταση των $11 \cdot 10^6$ αναγνωσμάτων. Όμως, είναι εξίσου σημαντικό να αναφερθεί ότι η πλατφόρμα Illumina παρείχε δεδομένα αλληλούχισης με προσεγγιστικά πάνω από 99% αξιοπιστία.



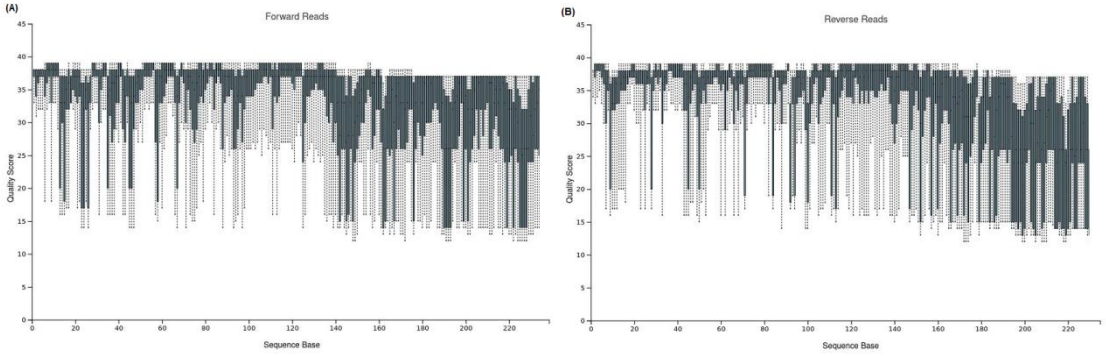
Σχήμα 3.2 Θηκογράμματα βαθμολογίας ποιότητας ανά θέση βάσης των πρωτογενών **(A)** εμπρόσθιων και **(B)** ανάστροφων αναγνωσμάτων αντίστοιχα.

3.2 Αποκοπή Μη-Βιολογικών Εκκινητών

Η διαδικασία αποκοπής των μη-βιολογικών αλληλουχιών που προέρχονται από τους εκκινητές οδήγησε στην μείωση του μήκους των αναγνωσμάτων καθώς και στην βελτίωση της ποιότητάς τους, ενώ δεν παρατηρήθηκε μεταβολή όσον αφορά στον συνολικό αριθμό αναγνωσμάτων. Δεδομένου ότι τα μήκη των αλληλουχιών που αφαιρέθηκαν είναι 17 bp και 21 bp για τα εμπρόσθια και ανάστροφα αναγνώσματα αντίστοιχα, παρατηρείται αντίστοιχη μεταβολή του συνολικού μήκους αναγνωσμάτων κατά την αφαίρεση αυτών (**Πίνακας 3.2**). Η αποκοπή των μη–βιολογικών αλληλουχιών από τα paired-end αναγνώσματα οδήγησε και στην αφαίρεση των αρχικών βάσεων που είχαν χαμηλότερη ποιότητα σε σχέση με αυτά που βρίσκονται στη κεντρική περιοχή. Για αυτόν τον λόγο, παρατηρείται βελτίωση της βαθμολογίας ποιότητας που παρέχουν οι βάσεις στο αριστερό άκρο των αναγνωσμάτων (**Σχήμα 3.3**).

Πίνακας 3.2 Το μήκος των εμπρόσθιων και ανάστροφων αναγνωσμάτων πριν και μετά την αφαίρεση μη-βιολογικών αλληλουχιών

Ποσοστιαίο πλήθος αναγνωσμάτων	**Μήκος αναγνωσμάτων σε bp**			
	Εμπρόσθια Αναγνώσματα		Ανάστροφα Αναγνώσματα	
	Πρωτογενή	Μετά την αφαίρεση	Πρωτογενή	Μετά την αφαίρεση
2%	237	220	246	225
7%	250	233	250	229
91%	251	234	251	230



Σχήμα 3.3 Θηκογράμματα βαθμολογίας ποιότητας ανά θέση βάσης των **(A)** εμπρόσθιων και **(B)** ανάστροφων αναγνωσμάτων μετά την αποκοπή των μη-βιολογικών αλληλουχιών. Παρατηρείται βελτίωση της ποιότητας στο αριστερό άκρο των paired-end αναγνωσμάτων.

3.3 Συγχώνευση paired- end Αναγνωσμάτων

Τα αποτελέσματα των συγχωνευμένων paired-end αναγνωσμάτων στις διάφορες τιμές του ελάχιστου μήκους επικαλυπτόμενης περιοχής ως προς τον αριθμό των διατηρητέων αναγνωσμάτων, την ποιότητα αυτών καθώς και το προσεγγιστικό μήκος τους δεν παρουσιάζουν ιδιαίτερες μεταβολές μεταξύ τους. Παρόλα αυτά, η συνολική εικόνα των δεδομένων που προέκυψαν από την επεξεργασία τους παρουσιάζει σημαντική διαφορά σε σχέση με αυτή των πρωτογενών δεδομένων. Υπενθυμίζεται ότι τα αποτελέσματα της παρούσας υποενότητας έχουν προκύψει από το επεξεργαστικό βήμα των workflows του Deblur και VSEARCH (**Σχήμα 2.2**), δηλαδή ύστερα της διαδικασίας αποκοπής των μη-βιολογικών αλληλουχιών (2.3.2).

Πιο αναλυτικά, παρατηρήθηκε ότι ανακτήθηκε περίπου το 79% των αναγνωσμάτων κατά την ολοκλήρωση της διαδικασίας, που σημαίνει ότι περίπου το 21% των paired-end αναγνωσμάτων δεν παρείχε τις απαραίτητες προδιαγραφές για την συγχώνευση τους (**Πίνακας 3.3**). Αυτό μπορεί να οφείλεται στο γεγονός ότι τα paired-end αναγνώσματα που απορρίφθηκαν προέρχονται από την αλληλούχιση αμπλικονίων των οποίων το μήκος τους ήταν μεγαλύτερο από 500 bp, και συνεπώς η πλατφόρμα αλληλούχισης δεν κατάφερε να προσδιορίσει τις αλληλουχίες της επικαλυπτόμενης περιοχής. Επιπλέον, υπάρχει η πιθανότητα τα άκρα αυτών των αναγνωσμάτων να παρέχουν σε ένα μεγάλο βαθμό εσφαλμένες βάσεις, με αποτέλεσμα να μην καταφέρουν να βρουν τα ζεύγη τους κατά την συγχώνευσή τους. Αναπόφευκτη είναι η ελαφρώς μικρότερη ανάκτηση των διατηρητέων αναγνωσμάτων καθώς αυξάνεται η τιμή του ελάχιστου απαιτούμενου μήκους επικαλυπτόμενης περιοχής.

Πίνακας 3.3 Γενική περιγραφή αριθμού διατηρητέων αναγνωσμάτων που παρέχουν συνολικά τα δείγματα στις διάφορες τιμές ελάχιστου απαιτούμενου μήκους επικαλυπτόμενης περιοχής. Στην παρένθεση αναγράφεται το ποσοστό διατηρητέων αναγνωσμάτων σε σχέση με τον συνολικό αριθμό πρωτογενών αναγνωσμάτων.

	Αριθμός διατηρητέων αναγνωσμάτων		
$Overlap_{min}$	**10**	**20**	**30**
Ελάχιστος	117272	117271	117271
Μέσος όρος	205986	205832	205398
Μέγιστος	576207	575660	575630
Συνολικός	8857412 ***(79,45%)***	8850783 ***(79,39%)***	8832154 ***(79,23%)***

Ο μέσος όρος του αριθμού αναγνωσμάτων καθώς και οι ακραίες τιμές των δειγμάτων μεταξύ των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρουσιάζουν ιδιαίτερες μεταβολές. Όμως, σε σχέση με την αρχική εικόνα των δεδομένων, παρουσιάζεται αισθητή μείωση στον αριθμό αναγνωσμάτων ανά δείγμα (**Παράρτημα 1**), η οποία έχει φέρει σαν αποτέλεσμα την μείωση των διακυμάνσεων του αριθμού αναγνωσμάτων μεταξύ αυτών (**Παράρτημα 2**). Εξαίρεση αποτελεί το δείγμα που παρέχει τον μέγιστο αριθμό διατηρητέων αναγνωσμάτων, το οποίο παρουσιάζει ακόμα έντονη διαφορά σε σχέση με τα υπόλοιπα δείγματα.

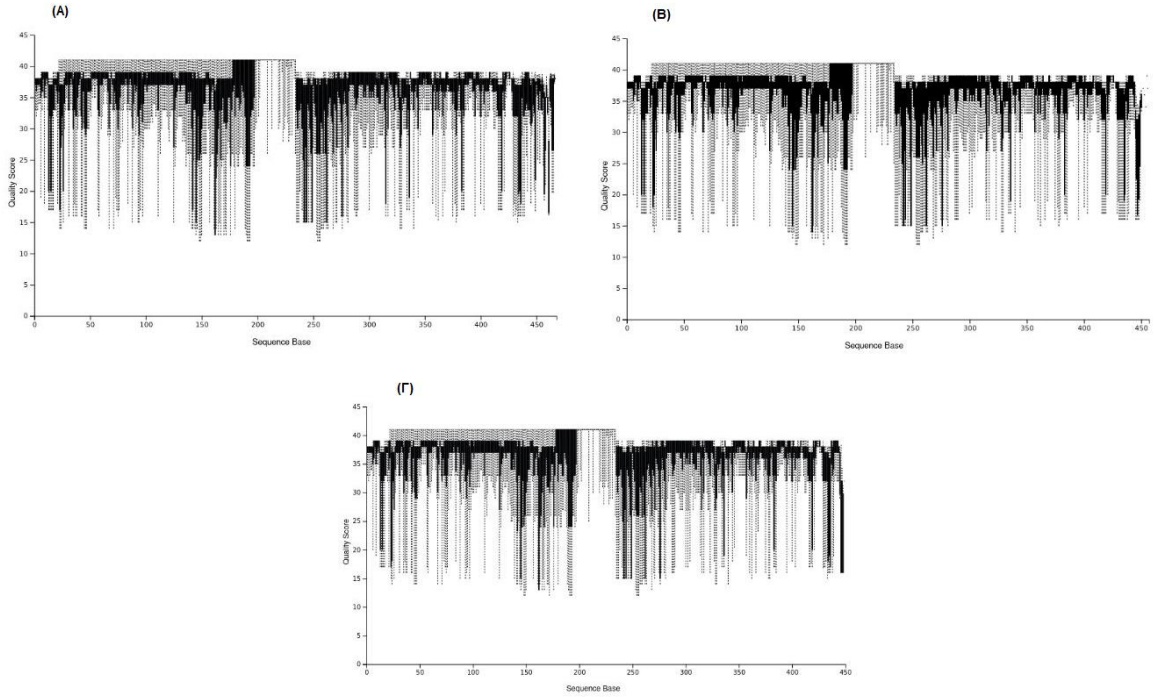
Όσον αφορά στο μήκος των συγχωνευμένων διατηρητέων αναγνωσμάτων, δεν παρατηρούνται έντονες διαφορές μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής ($Overlap_{min}$) (**Πίνακας 3.4**). Η πλειοψηφία των συγχωνευμένων αναγνωσμάτων προκύπτει να έχουν μήκος 427 bp, που αντικατοπτρίζει ικανοποιητικά το μήκος των περιοχών V3-V4 του γονιδίου 16S rRNA βάσει βιβλιογραφίας (Vargas-Albores et al., 2017). Εξίσου ικανοποιητικό προσεγγιστικό μήκος αποτελούν οι τιμές 405 και 407 bp που παρέχουν ένα μικρό ποσοστό των συγχωνευμένων αναγνωσμάτων. Όμως, περίπου το 10% αυτών παρουσιάζουν σημαντικά μικρότερο μήκος, κατά προσέγγιση 251 bp, που δηλώνει ότι η επικαλυπτόμενη περιοχή αυτών των αναγνωσμάτων ήταν όλο το μήκος των paired-end μικρο-αναγνωσμάτων. Αν και υπάρχει περίπτωση να αποτελούν αναγνώσματα που προέρχονται από την αλληλούχιση του 16S rRNA γονιδίου, κατά πάσα πιθανότητα είναι αποτέλεσμα σφαλμάτων του σταδίου της κατασκευής μεταγονιδιωματικής βιβλιοθήκης, και ενδεχομένως είναι χιμαιρικές αλληλουχίες.

Πίνακας 3.4 Το προσεγγιστικό μήκος των συγχωνευμένων αναγνωσμάτων στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής.

Αθροιστική ποσοστιαία κατανομή διατηρητέων αναγνωσμάτων	**Μήκος διατηρητέων αναγνωσμάτων (bp) για:**		
	$Overlap_{min}$ =10	$Overlap_{min}$ =20	$Overlap_{min}$ =30
2%	247	247	247
9%	251	251	251
25%	405	407	407
50%	427	427	427
75%	427	427	427
91%	427	427	427
98%	428	428	428

Όσον αφορά στην ποιότητα των αναγνωσμάτων, δεν παρατηρείται μεταβολή στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής (**Σχήμα 3.4**). Παρόλα αυτά, παρατηρείται έντονη διαφορά στην συνολική εικόνα της ποιότητας των αναγνωσμάτων κατά την ολοκλήρωση της διαδικασίας συγχώνευσης. Και τα δύο άκρα των συγχωνευμένων αναγνωσμάτων παρέχουν πια εξαιρετικά ψηλή ποιότητα βάσεων, ενώ παρατηρείται έντονη διακύμανση στην κεντρική περιοχή τους. Μάλιστα, η ποιότητα των βάσεων της επικαλυπτόμενης περιοχής (200-240 bp) χαρακτηρίζεται άψογη, με την ελάχιστη τιμή βαθμολογίας Phred να είναι ίση με 25, που σημαίνει ότι οι βάσεις έχουν ποσοστό αξιοπιστίας άνω του 99,99%. Το φαινόμενο αυτό οφείλεται στο ότι η απόδοση της συγχώνευσης των αναγνωσμάτων ήταν εξαιρετικά μεγάλη, και η αποτελεσματική συστοιχία των βάσεων οδήγησε στην αύξηση των Phred βαθμολογιών τους. Συνεπώς, ο μετασχηματισμός προς αύξηση της ποιότητας αυτής της περιοχής δηλώνει την σημαντικά υψηλή αξιοπιστία της. Φυσικά, οι περιοχές που βρίσκονται γύρω από την επικαλυπτόμενη περιοχή παρέχουν πολύ

χαμηλότερες βαθμολογίες ποιότητας σε σχέση με αυτές που μετασχηματίστηκαν, διότι παρέμειναν στην ίδια κατάσταση προ συγχώνευσης.



Σχήμα 3.4 Θηκογράμματα βαθμολογίας ποιότητας ανά θέση βάσης των συγχωνευμένων αναγνωσμάτων για τις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, όπου **(Α)** $Overlap_{min}$=10, **(Β)** $Overlap_{min}$=20 και **(Γ)** $Overlap_{min}$=30. Δεν παρατηρούνται σημαντικές αλλαγές της ποιότητας των συγχωνευμένων αναγνωσμάτων στις διαφορετικές τιμές $Overlap_{min}$.

3.4 Φιλτράρισμα Χαμηλής Ποιότητας Αναγνωσμάτων

Υπενθυμίζεται ότι τα αποτελέσματα της παρούσας υποενότητας έχουν προκύψει από το επεξεργαστικό βήμα των workflows του Deblur και VSEARCH, δηλαδή ύστερα της διαδικασίας συγχώνευσης paired-end αναγνωσμάτων (2.3.3). Τα δεδομένα που προέκυψαν μετά την διαδικασία αφαίρεσης χαμηλής ποιότητας συγχωνευμένων αναγνωσμάτων στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας Phred ανά βάση παρουσιάζουν σχετικές διαφοροποιήσεις μεταξύ τους, με την διακύμανση του ποσοστιαίου αριθμού διατηρητέων αναγνωσμάτων να είναι περίπου 40% με 63% (**Πίνακας 3.5**). Οι διαφορές των αποτελεσμάτων μεταξύ των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής χαρακτηρίζονται αμελητέες, όπου κατά την αύξηση αυτού του απαιτούμενου μήκους επιφέρει προσεγγιστικά την μείωση των διατηρητέων αναγνωσμάτων της τάξεως του 0.02% έως 0.16%. Όμως κατά την αύξηση της ελάχιστης απαιτούμενης βαθμολογίας ποιότητας ανά βάση παρατηρείται αισθητά η μείωση των συνολικά διατηρητέων αναγνωσμάτων. Ενώ μεταξύ των αποτελεσμάτων των Q_{min}=20 και 22 δεν έχουν σημαντικές διαφορές, όπου παρουσιάζεται διαφοροποίηση περίπου 0.5%, τα δεδομένα της παραμέτρου Q_{min}=26 παρουσιάζει προσεγγιστικά 20% λιγότερα διατηρητέα αναγνώσματα σε σχέση με τα υπόλοιπα.

Πίνακας 3.5 Γενική περιγραφή αριθμού διατηρητέων αναγνωσμάτων των συνολικών δειγμάτων στις διαφορές τιμές ελάχιστης βαθμολογίας ποιότητας και ελάχιστου μήκους επικαλυπτόμενης περιοχής. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός διατηρητέων αναγνωσμάτων για $Overlap_{min}$=10 και:		
Q_{min}	**20**	**22**	**26**
Ελάχιστος	91941	91346	46067
Μέσος όρος	164956	163578	103812
Μέγιστος	454022	450639	281793
Συνολικός	7093128 ***(63,63%)***	7033867 ***(63,09%)***	4463950 ***(40,04%)***
	Αριθμός διατηρητέων αναγνωσμάτων για $Overlap_{min}$=20 και:		
Q_{min}	**20**	**22**	**26**
Ελάχιστος	91941	91346	46067
Μέσος όρος	164847	163469	103803
Μέγιστος	453580	450198	281778
Συνολικός	7088449 ***(63,58%)***	7029209 ***(63,05%)***	4463568 ***(40,02%)***
	Αριθμός διατηρητέων αναγνωσμάτων για $Overlap_{min}$=30 και:		
Q_{min}	**20**	**22**	**26**
Ελάχιστος	91941	91346	46067
Μέσος όρος	164556	163180	103756
Μέγιστος	453559	450177	281778
Συνολικός	7075916 ***(63,47%)***	7016748 ***(62,94%)***	4461532 ***(40,02%)***

Ο μέσος όρος του αριθμού αναγνωσμάτων των δειγμάτων καθώς και οι ακραίες τιμές αυτών παρουσιάζουν αντίστοιχες μεταβολές με αυτές του συνολικού αριθμού διατηρητέων αναγνωσμάτων. Μεταξύ των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής και συγκεκριμένα των τιμών ελάχιστης βαθμολογία ποιότητας Q_{min}=20 και 22 δεν παρουσιάζονται ιδιαίτερες διαφορές τα αποτελέσματά τους, ενώ τα αποτελέσματα της τιμής Q_{min}=26 παρουσιάζουν την πιο έντονη μείωση των διατηρητέων αναγνωσμάτων ανά δείγμα (**Παράρτημα 3**). Η συνολική εικόνα των αποτελεσμάτων παρουσιάζει την αισθητή μείωση στον αριθμό αναγνωσμάτων ανά δείγμα, η οποία έχει φέρει σαν αποτέλεσμα την ακόμα πιο έντονη μείωση των διακυμάνσεων του αριθμού αναγνωσμάτων μεταξύ αυτών (**Παράρτημα 4**). Εξαίρεση συνεχίζει να αποτελεί το δείγμα που παρέχει τον μέγιστο αριθμό διατηρητέων αναγνωσμάτων, το οποίο παρουσιάζει ακόμα έντονη διαφορά σε σχέση με τα υπόλοιπα.

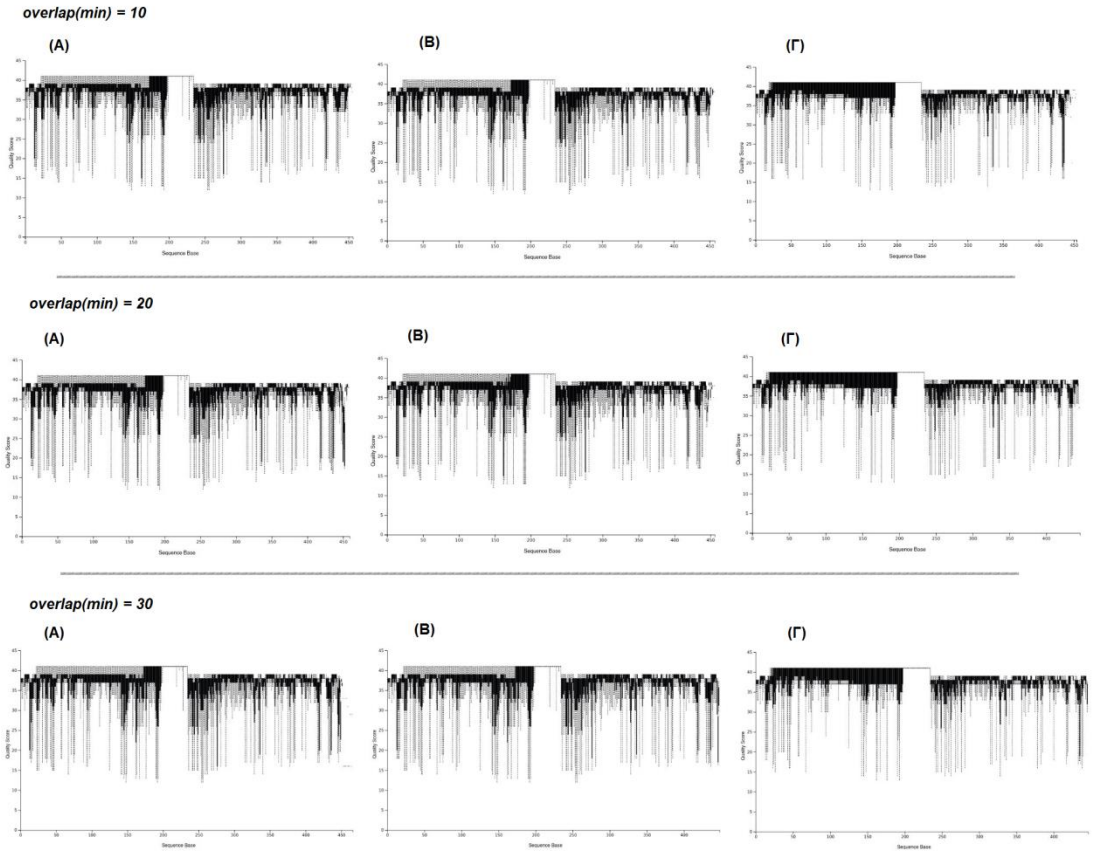
Όσον αφορά στο μήκος των διατηρητέων αναγνωσμάτων, μεταξύ των αποτελεσμάτων στις διάφερες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρουσιάζονται διαφορές (**Πίνακας 3.6**). Αντίστοιχα, τα δεδομένα των τιμών ελάχιστης βαθμολογία ποιότητας Q_{min}=20 και 22 δεν παρουσιάζονται διαφορές ενώ τα αποτελέσματα του Q_{min}=26 φαίνεται να έχουν μεγαλύτερο ποσοστό αναγνωσμάτων που παρέχουν μήκος 250 bp. Αυτό σημαίνει ότι η επιπλέον αύξηση της απαιτούμενης ποιότητας των αναγνωσμάτων επιφέρει την μείωση αναγνωσμάτων με μήκος 400 bp και άνω, απορρίπτοντας έτσι υλικό που το οποίο όμως αντικατοπτρίζει περισσότερο την στοχευόμενη περιοχή V3-V4 του 16S rRNA γονιδίου.

Η συνολική ποιότητα των αναγνωσμάτων παρουσιάζει αναμενόμενα βελτίωση μετά την διαδικασία φιλτραρίσματος χαμηλής ποιότητας αναγνωσμάτων (**Σχήμα 3.5**). Μεταξύ των

διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρουσιάζονται σημαντικές διαφορές καθώς και μεταξύ των τιμών ελάχιστης βαθμολογίας ποιότητας Q_{min}=20 και 22. Η ποιότητα των αναγνωσμάτων που έχουν υποστεί πιο αυστηρό ποιοτικό έλεγχο με Q_{min}=26 παρουσιάζουν πολύ μεγαλύτερη αξιοπιστία, με την πλειοψηφία αυτών να έχουν ποιότητα ανά βάση πάνω από 35, και κατά επέκταση πάνω από 99,999% αξιοπιστία.

Πίνακας 3.6 Το προσεγγιστικό μήκος των συγχωνευμένων και ποιοτικά φιλτραρισμένων αναγνωσμάτων στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογία ποιότητας.

Αθροιστική ποσοστιαία κατανομή αριθμού αναγνωσμάτων	**Μήκος διατηρητέων αναγνωσμάτων (bp)**								
	$Overlap_{min}$=30			**$Overlap_{min}$=30**			**$Overlap_{min}$=30**		
	Q_{min} =20	**Q_{min} =22**	**Q_{min} =26**	**Q_{min} =20**	**Q_{min} =22**	**Q_{min} =26**	**Q_{min} =20**	**Q_{min} =22**	**Q_{min} =26**
2%	247	247	241	247	247	244	247	247	241
9%	251	251	248	251	251	248	251	251	248
25%	385	402	251	386	374	251	374	385	251
50%	427	427	427	427	427	427	427	427	427
75%	427	427	427	427	427	427	427	427	427
91%	427	427	427	427	427	427	427	427	427
98%	428	428	427	428	428	427	428	428	427



Σχήμα 3.5 Θηκογράμματα βαθμολογίας ποιότητας ανά θέση βάσης των συγχωνευμένων και ποιοτικά φιλτραρισμένων αναγνωσμάτων για τις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας, όπου **(Α)** Q_{min}=20, **(Β)** Q_{min}=22 και **(Γ)** Q_{min}=26.

3.5 Παραγωγή ASVs και OTUs

Για την παραγωγή των αντιπροσωπευτικών αναγνωσμάτων υπενθυμίζεται ότι το pipeline αποθορυβοποίησης DADA2 έχει ενσωματωμένο όλο τον ποιοτικό έλεγχο όσον αφορά στην συγχώνευση και το ποιοτικό φιλτράρισμα αναγνωσμάτων (**Σχήμα 2.2**), και κατ' επέκταση τα δεδομένα εισόδου του ήταν αυτά που προέκυψαν μετά την αφαίρεση των μη-βιολογικών αλληλουχιών (2.3.2). Τα pipeline αποθορυβοποίησης Deblur και ομαδοποίησης VSEARCH απαιτούν δεδομένα εισόδου που έχουν υποστεί τον προαναφερόμενο ποιοτικό έλεγχο (2.3.3, 2.3.4). Παρακάτω, παρουσιάζονται τα αποτελέσματα μετά την εφαρμογή των pipelines DADA2, Deblur και VSEARCH στις διάφορες συνθήκες ποιοτικού φιλτραρίσματος, αποκοπής αναγνωσμάτων όπου αυτές εφαρμόζονται και στις διάφορες τιμές ελάχιστου απαιτούμενου μήκους επικαλυπτόμενης περιοχής. Τα αποτελέσματα περιλαμβάνουν τον αριθμό διατηρητέων αναγνωσμάτων, των παραγόμενων ASVs/OTUs και μια επισκόπηση της συχνότητας αυτών στο σύνολο των δεδομένων. Δεδομένου ότι αυτό το στάδιο επεξεργασίας μετατρέπει τα μοναδικά αναγνώσματα σε αντιπροσωπευτικά που συνοδεύονται από μια τιμή συχνότητας, δεν υπάρχει από εδώ και στο εξής η δυνατότητα επισκόπησης της ποιότητας των αναγνωσμάτων. Εκτός από τα τελικά αποτελέσματα του pipeline DADA2, από την πλατφόρμα QIIME2 παρέχεται η δυνατότητα επισκόπησης του αριθμού διατηρητέων αναγνωσμάτων μετά από κάθε βασικό βήμα επεξεργασίας δεδομένων αυτού του pipeline. Τα βήματα αυτά αφορούν την αποκοπή και το φιλτράρισμα με βάση την μέγιστη τιμή αναμενόμενων σφαλμάτων, την διαδικασία αποθορυβοποίησης, την συγχώνευση των paired-end αναγνωσμάτων και την αφαίρεση χιμαιρικών αλληλουχιών.

3.5.1 DADA2

Το πρώτο βήμα επεξεργασίας που εφαρμόζεται στο pipeline DADA2 είναι ο ποιοτικός έλεγχος των paired-end αναγνωσμάτων, ο οποίος περιλαμβάνει την αποκοπή των μικρο-αναγνωσμάτων στο επιθυμητό μήκος και το ποιοτικό φιλτράρισμα με βάση την τιμή του μέγιστου ποσοστού αναμενόμενων σφαλμάτων. Η πρώτη εικόνα αποτελεσμάτων φανερώνει την έντονη επιρροή της παραμέτρου των αναμενόμενων σφαλμάτων στον αριθμό διατηρητέων αναγνωσμάτων (**Παράρτημα 11**). Πιο αναλυτικά, η τιμή $e.e._{max}$=2,5 επιφέρει την μείωση των διατηρητέων αναγνωσμάτων κατά προσέγγιση στα 74% με 80%, η τιμή $e.e_{max}$=1,5 αντίστοιχα στα 54% με 58% και η τιμή $e.e._{max}$=0,5 αντίστοιχα στα 12%, ανάλογα τις συνθήκες αποκοπής των paired-end αναγνωσμάτων. Σε αντίθεση με την εφαρμογή διαφορετικής τιμής μέγιστων αναμενόμενων σφαλμάτων, η αποκοπή των paired-end αναγνωσμάτων δεν φαίνεται να παρουσιάζει σημαντικές αλλαγές στον αριθμό διατηρητέων αναγνωσμάτων. Συγκεκριμένα, τα αποτελέσματα παρουσιάζουν μια απόκλιση της τάξεως 0,5% έως 6%, με την μεγαλύτερη να παρουσιάζεται στα δεδομένα με την μικρότερη απαίτηση ποιοτικού φιλτραρίσματος, δηλαδή με $e.e._{max}$=2,5.

Το επόμενο στάδιο επεξεργασίας του DADA2 είναι η διαδικασία αποθορυβοποίησης των paired-end αναγνωσμάτων. Τα αποτελέσματα αυτής δεν παρουσιάζουν έντονη μεταβολή σε σχέση με τα αποτελέσματα του προηγούμενου σταδίου. Μάλιστα, η αποθορυβοποίηση των αναγνωσμάτων μετά το στάδιο βασικού ποιοτικού φιλτραρίσματος κοστίζει σε αριθμό αναγνωσμάτων ένα ποσοστό της τάξεως 0,09% εώς 0,92%, με το μεγαλύτερο ποσοστό να παρουσιάζεται στα δεδομένα με $e.e._{max}$=2,5 (**Παράρτημα 12**).

Το τελευταίο στάδιο επεξεργασίας του pipeline είναι η συγχώνευση των αποθορυβοποιημένων paired-end αναγνωσμάτων, το φιλτράρισμα πιθανών χιμαιρικών

αναγνωσμάτων και η παραγωγή των αντιπροσωπευτικών αναγνωσμάτων ASVs ταυτόχρονα. Με μία πρώτη ματιά, η επιλογή της τιμής ελάχιστου μήκους επικαλυπτόμενης περιοχής επιφέρει αμελητέα αλλαγή στα αποτελέσματα με σταθερές συνθήκες αποκοπής αναγνωσμάτων και τιμές ποσοστού αναμενόμενων σφαλμάτων (**Πίνακας 3.7**). Μάλιστα, το μέγιστο ποσοστό αναγνωσμάτων που απορρίπτεται κατά την αύξηση της τιμής ελάχιστου μήκους επικαλυπτόμενης περιοχής σε σχέση με τον αρχικό αριθμό αναγνωσμάτων είναι της τάξεως 0,4%. Παρόλα αυτά, το αντίκτυπο της συγχώνευσης και της αφαίρεσης χιμαιρικών αλληλουχιών εμφανίζεται πιο έντονα στα δεδομένα που έχουν μεγαλύτερη τιμή ποσοστού αναμενόμενων σφαλμάτων, με λίγο παραπάνω έμφαση στα δεδομένα χωρίς αποκοπή paired-end αναγνωσμάτων. Σε σχέση με τον αρχικό αριθμό αναγνωσμάτων, τα δεδομένα με $e.e._{max}$=2,5 μειώθηκαν κατά 10%-12%, με $e.e._{max}$=1,5 αντίστοιχα κατά 6%-8% και με $e.e._{max}$=0,5 αντίστοιχα περίπου κατά 1%. Δεδομένου ότι πραγματοποιούνται ταυτόχρονα δύο εφαρμογές σε αυτό το στάδιο, η αύξηση του αριθμού αναγνωσμάτων που απορρίπτονται κατά την αύξηση της τιμής των αναμενόντων σφαλμάτων μπορεί να οφείλεται στους εξής λόγους: (α) είτε στην αποτυχία συγχώνευσης των paired-end αναγνωσμάτων λόγω του ότι συντηρήθηκαν μέχρι εκείνο το στάδιο ζεύγη αναγνωσμάτων που φέρουν σφάλματα στο τελικό άκρο, (β) είτε στο γεγονός ότι διατηρήθηκε μεγαλύτερος αριθμός χιμαιρικών αλληλουχιών. Σημαντική παρατήρηση αποτελεί ότι όσο αυξάνεται η απαίτηση αξιοπιστίας των αναγνωσμάτων, το ποσοστό συγχώνευσης αναγνωσμάτων τείνει να είναι 100%. Σε όλα τα παραπάνω στάδια, παρατηρείτε αντίστοιχη μεταβολή του αριθμού αναγνωσμάτων ανά δείγμα σε σχέση με τον συνολικό αριθμό αναγνωσμάτων (**Παράρτημα 5**, **Παράρτημα 6**, **Παράρτημα 7**, **Παράρτημα 8**, **Παράρτημα 9**, **Παράρτημα 10**). Μεταξύ των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής, οι διαφορές ως προς τον αριθμό αναγνωσμάτων ανά δείγμα είναι αμελητέα και σε κάποιες περιπτώσεις μηδενική σε σταθερές συνθήκες αποκοπής και τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων. Επιπλέον, μεταξύ των αποτελεσμάτων που έχουν προκύψει από διαφορετικά σενάρια αποκοπής αναγνωσμάτων δεν παρατηρείται σημαντική διακύμανση στον αριθμό αναγνωσμάτων ανά δείγμα (**Παράρτημα 13**). Αναμενόμενο είναι το αντίκτυπο στον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα όσο αυξάνεται η αυστηρότητα ως προς την ποιότητα των αναγνωσμάτων, κάτι που παρουσιάστηκε αντίστοιχα και στον συνολικό αριθμό αναγνωσμάτων.

Πίνακας 3.7 Γενική περιγραφή αριθμού τελικών διατηρητέων συγχωνευμένων αναγνωσμάτων στο σύνολο των δειγμάτων που προέκυψαν από τον DADA2 στις διαφορές τιμές i) μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ii) ελάχιστου μήκους επικαλυπτόμενης περιοχής και iii) συνθήκες αποκοπής αναγνωσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Διατηρητέων Αναγνωσμάτων με $e.e._{max}$=2,5					
	No trim			**With trim**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	101103	101083	101083	95521	94811	94327
Μέσος Όρος	165166	164922	164591	154275	153711	153449
Μέγιστος	485628	485086	485086	484123	481297	481297
Συνολικός	7502100 ***(67,29%)***	7493079 ***(67,21%)***	7477953 ***(67,08%)***	7112481 ***(63,80%)***	7081005 ***(63,52%)***	7065359 ***(63,38%)***

	Αριθμός Διατηρητέων Αναγνωσμάτων με $e.e._{max}$=1,5					
	No trim			**With trim**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	78642	78642	78642	74522	74033	73736
Μέσος Όρος	122171	122136	122117	114974	114789	114620
Μέγιστος	383417	382922	382922	385584	384374	384374
Συνολικός	5632951 ***(50,53%)***	5626164 ***(50,47%)***	5615549 ***(50,37%)***	5415943 ***(48,58%)***	5400165 ***(48,44%)***	5389518 ***(48,34%)***
	Αριθμός Διατηρητέων Αναγνωσμάτων με $e.e._{max}$=0,5					
	No trim			**With trim**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	9818	9818	9818	10014	10014	10010
Μέσος Όρος	26078	26039	25978	25984	25981	25948
Μέγιστος	113045	112937	112937	117480	117269	117269
Συνολικός	1299788 ***(11,66%)***	1297475 ***(11,64%)***	1296367 ***(11,63%)***	1310260 ***(11,75%)***	1308298 ***(11,74%)***	1307449 ***(11,73%)***

Ο μέσος όρος του αριθμού διατηρητέων αναγνωσμάτων ανά δείγμα κυμαίνεται από 25948 έως 165166, ενώ οι ελάχιστες ακραίες τιμές αντίστοιχου αριθμού κυμαίνονται από 9818 έως 101103 (**Πίνακας 3.7**). Συνεπώς, ένας σημαντικός αριθμός δειγμάτων που επεξεργαστήκαν με την τιμή παραμέτρου $e.e._{max}$=0,5 προβλέπεται να παρέχει αριθμό αναγνωσμάτων μικρότερο από τον ελάχιστο αριθμό αναγνωσμάτων που συνιστάται να έχει το δείγμα για να αντικατοπτριστεί η πραγματική ποικιλομορφία αυτών. Σε κάθε παραμετρικό σενάριο, το δείγμα που έχει τον μεγαλύτερο αριθμό αναγνωσμάτων εξακολουθεί να αποτελεί έντονη ακραία τιμή στο πλήθος των δειγμάτων, με τον αριθμό να κυμαίνεται στα 117480 με 485628. Εξαιρώντας αυτό το δείγμα, η διακύμανση των υπολοίπων μειώνεται σε σχέση με την αρχική με την πιο έντονη να παρουσιάζεται στα δείγματα που επεξεργαστήκαν με τιμή παραμέτρου $e.e._{max}$=2,5.

Όσον αφορά στον αριθμό αντιπροσωπευτικών αναγνωσμάτων που παράχθηκαν, παρατηρείτε ότι επηρεάζεται σημαντικά από τις συνθήκες αποκοπής σε αντίθεση με τις διάφορες τιμές βαθμολογίας ποιότητας και μήκους επικαλυπτόμενης περιοχής (**Πίνακας 3.8**). Συγκεκριμένα, όταν επιλέγεται η μη αποκοπή των paired-end αναγνωσμάτων, ο αριθμός των παραγόμενων ASVs προκύπτει σχεδόν διπλάσιος σε κάθε υπολογιστικό σενάριο σε σχέση με τα αντίστοιχα αποτελέσματα που επιβλήθηκε η αποκοπή στα paired-end αναγνώσματα. Καθώς αυξάνεται η απαίτηση ως προς την αξιοπιστία των αναγνωσμάτων και το μήκος της επικαλυπτόμενης περιοχής, ο αριθμός των παραγόμενων ASVs μειώνεται. Βέβαια, μεταξύ των διάφορων τιμών του μήκους της επικαλυπτόμενης περιοχής δεν παρουσιάζονται σημαντικές διαφορές, ενώ μεταξύ των τιμών του ποσοστού αναμενόμενων σφαλμάτων παρουσιάζεται σημαντική μεταβολή των παραγόμενων ASVs όταν επιβάλλεται $e.e._{max}$=0,5.

Μια γενική επισκόπηση των συχνοτήτων που συνοδεύουν τα παραγόμενα ASVs φανερώνει τον μεγάλο όγκο μοναδικών και χαμηλής συχνότητας ASVs που παράχθηκαν σε κάθε παραμετρικό σενάριο (**Παράρτημα 15**). Μεταξύ των διάφορων τιμών μήκους

επικαλυπτόμενης περιοχής παρουσιάζεται αμελητέα διαφορά, ενώ αναμενόμενη είναι η έντονη μεταβολή μεταξύ των διαφόρων τιμών ποσοστού αναμενόμενων σφαλμάτων, η οποία παρουσιάζεται με παρόμοιο μοτίβο όπως και στον συνολικό αριθμό αναγνωσμάτων. Η επιλογή αποκοπής ή μη των paired-end αναγνωσμάτων επηρεάζει τις συχνότητες των ASVs, κάτι που παρατηρήθηκε αντίστοιχα και στον αριθμό των παραγόμενων ASVs. Όμως, σε αντίθεση με τον συνολικό αριθμό ASVs, η συχνότητά τους φαίνεται να αυξάνεται σημαντικά όταν επιλέγεται να αποκοπούν τα paired-end αναγνώσματα και ταυτόχρονα παρατηρείται μείωση του αριθμού μοναδικών ASVs (**Παράρτημα 14**).

Πίνακας 3.8 Αριθμός αντιπροσωπευτικών αναγνωσμάτων (ASVs) που παράχθηκαν από τον DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και συνθήκες αποκοπής.

Αριθμός Αντιπροσωπευτικών Αναγνωσμάτων - ASVs						
	No trim			**With trim**		
$Overlap_{min}$ / $e.e._{max}$	**10**	**20**	**30**	**10**	**20**	**30**
2,5	4,975	4,890	4,794	2,855	2,706	2,621
1,5	4,936	4,848	4,753	2,808	2,665	2,583
0,5	3,578	3,534	3,486	1,795	1,732	1,700

Δεδομένου ότι ο συνολικός αριθμός των αναγνωσμάτων δεν διαφέρει σημαντικά μεταξύ των δύο διαφορετικών σεναρίων αποκοπής, εντούτοις μπορεί να οφείλεται στο ότι η αποκοπή των paired-end αναγνωσμάτων οδήγησε στην μη διαφοροποίηση των παραγόμενων ASVs των οποίων η διαφοροποίηση τους παρουσιαζόταν στην περιοχή αλληλουχιών που αφαιρέθηκαν. Επίσης, πιθανολογείται και στο ότι ένα ποσοστό paired-end αναγνωσμάτων δεν συμπεριλήφθηκε στην επεξεργασία λόγω της αποκοπής, διότι ο DADA2 απορρίπτει αυτόματα τα αναγνώσματα που έχουν μικρότερο μήκος από το σημείο αποκοπής. Έτσι, τα απορριπτόμενα δεδομένα σε αυτό το σενάριο έχουν την ευκαιρία να συμπεριληφθούν όταν δεν επιλέγεται σημείο αποκοπής και ενδεχομένως να παραχθούν περισσότερα ASVs.

Το μήκος των αντιπροσωπευτικών αναγνωσμάτων ASVs παρουσιάζει αμελητέα αλλαγή μεταξύ των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής και μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ενώ υπάρχει έντονη διακύμανση μεταξύ των δύο συνθηκών αποκοπής (**Πίνακας 3.9**). Περίπου το 50% των ASVs που προέκυψαν χωρίς να έχουν αφαιρεθεί τα άκρα των paired-end αναγνωσμάτων έχουν μήκος <250 bp, ενώ στην περίπτωση αποκοπής όλα τα ASVs έχουν μήκος >230 bp. Συνεπώς, ενισχύεται η θεωρία του ότι η συμπερίληψη των αναγνωσμάτων που απορρίπτονται κατά την διαδικασία αποκοπής φέρει σαν αποτέλεσμα την παραγωγή μεγαλύτερου αριθμού ASVs. Όμως, είναι σημαντικό να σημειωθεί ότι ένα μήκος αναγνώσματος που κυμαίνεται στα 86 bp με 250 bp δεν αντικατοπτρίζει το μέσο μήκος της περιοχής V3-V4 του 16S rRNA (Vargas-Albores et al., 2017). Συνεπώς, στο σύνολο των αποτελεσμάτων υπάρχει περίπτωση ο μισός όγκος των αντιπροσωπευτικών αναγνωσμάτων να φέρουν εσφαλμένες αλληλουχίες.

Πίνακας 3.9 Το προσεγγιστικό μήκος των αντιπροσωπευτικών αναγνωσμάτων (ASVs) του DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και συνθήκες αποκοπής αναγνωσμάτων.

Μήκος Αντιπροσωπευτικών Αναγνωσμάτων - ASVs (bp)									
	No trim								
e.e.$_{max}$	**2,5**			**1,5**			**0,5**		
Overlap$_{min}$ / **Αθροιστική ποσοστιαία κατανομή**	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
2%	86	88	91	86	88	91	85	86	87
9%	134	136	138	132	135	137	117	122	124
25%	189	189	189	189	189	189	170	171	171
50%	263	260	257	262	260	257	243	243	243
75%	407	407	404	407	407	405	406	405	404
91%	427	427	427	427	427	427	427	427	427
98%	441	435	428	442	435	428	430	428	428
	With trim								
e.e.$_{max}$	**2,5**			**1,5**			**0,5**		
Overlap$_{min}$ / **Αθροιστική ποσοστιαία κατανομή**	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
2%	237	237	236	237	237	237	237	237	236
9%	248	248	247	248	248	247	247	247	247
25%	312	307	304	312	308	305	316	311	307
50%	404	402	402	404	402	402	407	405	404
75%	427	427	427	427	427	427	427	427	427
91%	431	427	427	430	427	427	428	427	427
98%	446	437	428	446	438	428	444	431	428

3.5.2 Deblur

Δεδομένου ότι ο προσδιορισμός των αποστάσεων Hamming των αναγνωσμάτων που πραγματοποιεί ο Deblur απαιτεί τα αναγνώσματα να έχουν όλα το ίδιο μήκος, επιλέγονται δύο σημεία αποκοπής με βάση τα αποτελέσματα του ποιοτικού φιλτραρίσματος. Αυτά τα σημεία είναι στην 250^{η} (trim@=250) και στην 380^{η} (trim@=380) βάση αντίστοιχα, με σκοπό την προσπάθεια συμπερίληψης όσον το δυνατόν μεγαλύτερου όγκου δεδομένων προς επεξεργασία, διότι τα αναγνώσματα που παρέχουν μικρότερο μήκος από τα σημεία αποκοπής απορρίπτονται αυτόματα.

Μια πρώτη εικόνα των αποτελεσμάτων του Deblur στις δύο διαφορετικές τιμές αποκοπής υποδεικνύει την αυστηρότητα του αλγόριθμου ως προς την διατήρηση αναγνωσμάτων (**Πίνακας 3.10**). Πιο συγκεκριμένα, το μέγιστο ποσοστό διατηρητέων αναγνωσμάτων προκύπτει 17% από το σύνολο δεδομένων με αποκοπή στην 250^{η} βάση και Q_{min}=20, ενώ αντίστοιχα το ελάχιστο ποσοστό προκύπτει ίσο με 9,48% στο σύνολο

δεδομένων με σημείο αποκοπής στην 380η βάση και Q_{min}=26. Επιπλέον, για σταθερές τιμές σημείου αποκοπής και ελάχιστης βαθμολογίας ποιότητας, παρατηρείται ότι η παράμετρος ελάχιστου μήκους επικαλυπτόμενης περιοχής κατέληξε να μην έχει καμία επιρροή στον συνολικό αριθμό αναγνωσμάτων. Όσον αφορά τις διάφορες τιμές ελάχιστης ποιότητας, μεταξύ των συνόλων δεδομένων Q_{min}=20 και 22 παρατηρείται αμελητέα διαφορά, ενώ για Q_{min}=26 παρουσιάζεται μεγαλύτερη επίπτωση ως προς τον αριθμό διατηρητέων αναγνωσμάτων. Η επιλογή του σημείου αποκοπής των αναγνωσμάτων οδηγεί σε διαφορετικό αριθμό διατηρητέων αναγνωσμάτων, όπου για σημείο αποκοπής στην 250η βάση να δίνει κατά προσέγγιση 3% με 4% παραπάνω αριθμό διατηρητέων αναγνωσμάτων σε σχέση με αυτόν των πρωτογενών δεδομένων.

Πίνακας 3.10 Γενική περιγραφή αριθμού τελικών διατηρητέων αναγνωσμάτων των συνολικών δειγμάτων που προέκυψαν από τον Deblur στις διάφορες τιμές i) ελάχιστης βαθμολογίας ποιότητας, ii) ελάχιστου μήκους επικαλυπτόμενης περιοχής και iii) σημείων αποκοπής αναγνωσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Διατηρητέων Αναγνωσμάτων με Q_{min}=20					
	trim@=250			**trim@=380**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	15933	15933	15933	12112	12112	12112
Μέσος Όρος	44077	44077	44077	32614	32614	32614
Μέγιστος	152680	152680	152680	109189	109189	109189
Συνολικός	1895317 ***(17.00%)***	1895317 ***(17.00%)***	1895317 ***(17.00%)***	1402418 ***(12,58%)***	1402418 ***(12,58%)***	1402418 ***(12,58%)***
	Αριθμός Διατηρητέων Αναγνωσμάτων με Q_{min}=22					
	trim@=250			**trim@=380**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	15907	15907	15907	12099	12099	12099
Μέσος Όρος	44027	44027	44027	32584	32584	32584
Μέγιστος	152541	152541	152541	109106	109106	109106
Συνολικός	1893144 ***(16.98%)***	1893144 ***(16.98%)***	1893144 ***(16.98%)***	1401094 ***(12.57%)***	1401094 ***(12.57%)***	1401094 ***(12.57%)***
	Αριθμός Διατηρητέων Αναγνωσμάτων με Q_{min}=26					
	trim@=250			**trim@=380**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	11307	11307	11307	9322	9322	9322
Μέσος Όρος	31477	31477	31477	24586	24586	24586
Μέγιστος	119524	119524	119524	88782	88782	88782
Συνολικός	1353500 ***(12.14%)***	1353500 ***(12.14%)***	1353500 ***(12.14%)***	1057206 ***(9.48%)***	1057206 ***(9.48%)***	1057206 ***(9.48%)***

Αντίστοιχα, παρατηρείται μείωση των διατηρητέων αναγνωσμάτων ανά δείγμα, όπου ανάλογα με το παραμετρικό σενάριο (**Παράρτημα 16**, **Παράρτημα 17**), τα δείγματα φέρουν κατά μέσο όρο 24586 με 44077 αναγνώσματα (**Πίνακας 3.10**). Με σταθερές τιμές ελάχιστης βαθμολογίας ποιότητας (Q_{min}) και σημείων αποκοπής (trim@), παρατηρείται εξίσου μηδενική διαφορά στον αριθμό αναγνωσμάτων ανά δείγμα στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής ($Overlap_{min}$). Επίσης, μεταξύ των αποτελεσμάτων που έχουν προκύψει από διαφορετικά σενάρια αποκοπής αναγνωσμάτων παρουσιάζεται αντίστοιχη διαφορά με τον συνολικό αριθμό αναγνωσμάτων, με τα δεδομένα αναγνωσμάτων μήκους 250

bp να φέρουν ελαφρώς μεγαλύτερο αριθμό. Καθώς αυξάνεται η παράμετρος ελάχιστης βαθμολογίας ποιότητας βάσεων μειώνεται ο αριθμός διατηρητέων αναγνωσμάτων ανά δείγμα.

Οι ελάχιστες ακραίες τιμές αριθμού αναγνωσμάτων που φέρουν τα δείγματα κυμαίνονται από 9322 έως 15933 (**Πίνακας 3.10**). Συνεπώς, ένας σημαντικός αριθμός δειγμάτων που επεξεργαστήκαν με Q_{min}=26 παρέχει αριθμό αναγνωσμάτων μικρότερο από τον αντίστοιχο ελάχιστο που συνιστάται να έχει το δείγμα για να αντικατοπτριστεί η πραγματική ποικιλομορφία αυτών (Bukin et al., 2019). Σε κάθε παραμετρικό σενάριο, το δείγμα που έχει τον μεγαλύτερο αριθμό αναγνωσμάτων εξακολουθεί να αποτελεί έντονη ακραία τιμή στο πλήθος των δειγμάτων, με τον αριθμό να κυμαίνεται στα 88782 με 152680. Εξαιρώντας αυτό το δείγμα, η διακύμανση των υπολοίπων μειώνεται σε σχέση με την αρχική με την πιο έντονη να παρουσιάζεται στα δείγματα που επεξεργαστήκαν με αποκοπή στην $380^{η}$ βάση (**Παράρτημα 18**).

Όσον αφορά στον αριθμό αντιπροσωπευτικών αναγνωσμάτων που παράχθηκαν, παρατηρείτε ότι επηρεάζεται σημαντικά από τα σημεία αποκοπής, με την επιλογή trim@=250 να οδηγεί σε μεγαλύτερο αριθμό παραγόμενων ASVs (**Πίνακας 3.11**). Αντίθετα, μεταξύ των διάφορων τιμών μήκους επικαλυπτόμενης περιοχής δεν παρουσιάζεται μεταβολή. Η αύξηση της τιμής παραμέτρου Q_{min} φέρει ως αποτέλεσμα την μείωση του αριθμού παραγόμενων ASVs, με το μεγαλύτερο αντίκτυπο ως προς την απώλεια αυτών να παρέχεται από την τιμή Q_{min}=26.

Πίνακας 3.11 Αριθμός αντιπροσωπευτικών αναγνωσμάτων (ASVs) που παράχθηκαν από τον Deblur στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, ελάχιστου μήκους επικαλυπτόμενης περιοχής και σημείων αποκοπής.

Αριθμός Αντιπροσωπευτικών Αναγνωσμάτων - ASVs						
	trim@=250			**trim@=380**		
$Overlap_{min}$ / Q_{min}	**10**	**20**	**30**	**10**	**20**	**30**
20	940	940	940	884	884	884
22	939	939	939	884	884	884
26	791	791	791	762	762	762

Μια γενική επισκόπηση των συχνοτήτων που συνοδεύουν τα παραγόμενα ASVs φανερώνει την απουσία μοναδικών ASVs σε όλα τα σύνολα διαφορετικών παραμετρικών σεναρίων (**Παράρτημα 19**, **Παράρτημα 20**). Αυτό οφείλεται στην λειτουργία του Deblur ως προς την διαχείριση χιμαιρικών αλληλουχιών, κατά την οποία επιβάλλεται η αυτόματη απόρριψη ASVs με συχνότητα παρατήρησης 10 ανά δείγμα. Η επιλογή τιμής μήκους επικαλυπτόμενης περιοχής παραμένει και σε αυτές τις μετρήσεις ένας παράγοντας που δεν επηρεάζει τα αποτελέσματα. Μεταξύ των δύο διαφορετικών σημείων αποκοπής, τα ASVs με μήκος 250 bp συνοδεύονται κατά μέσο όρο από μεγαλύτερη συχνότητα σε σύγκριση με αυτά που παρέχουν 380 bp (**Παράρτημα 19**). Επιπλέον, η αύξηση της ελάχιστης βαθμολογίας ποιότητας οδηγεί στην μείωση των συχνοτήτων των ASVs.

3.5.3 VSEARCH

Ο VSEARCH περιλαμβάνει την ομαδοποίηση σε λειτουργικές ταξινομικές μονάδες OTUs με 97% ομοιότητα χρησιμοποιώντας την βάση δεδομένων SILVA και το φιλτράρισμα χιμαιρικών αλληλουχιών. Όλες οι μετρήσεις των αποτελεσμάτων, συμπεριλαμβανομένου του

συνολικού αριθμού διατηρητέων αναγνωσμάτων και ανά δείγμα αντίστοιχα, της διακύμανσης των δειγμάτων, του αριθμού παραγόμενων OTUs και της συχνότητας με την οποία παρουσιάζονται στα δεδομένα, παρουσιάζονται με το ίδιο μοτίβο στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας με αυτό που προέκυψε στα αποτελέσματα του Deblur. Η επιλογή του ελάχιστου μήκους επικαλυπτόμενης περιοχής φαίνεται να μην επηρεάζει τις μετρήσεις, σε αντίθεση με αυτή της ελάχιστης βαθμολογίας ποιότητας, η οποία καθώς αυξάνεται, κάθε προαναφερόμενη μέτρηση μειώνεται.

Πιο αναλυτικά ως προς τις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, παρατηρείται ότι ο αριθμός διατηρητέων αναγνωσμάτων προκύπτει μικρότερος καθώς αυξάνεται το Q_{min} (**Πίνακας 3.12**). Η επιλογή μεταξύ των τιμών Q_{min}= 20 και 22 επηρεάζει αμελητέα το τελικό αποτέλεσμα, με ποσοστό διατηρητέων αναγνωσμάτων να προκύπτει αντίστοιχα 43,90% και 43,57%. Για τιμή Q_{min}= 26, το ποσοστό αναγνωσμάτων σε σχέση με τα πρωτογενή δεδομένα που συγκρατείται προκύπτει ίσος με 23,67%.

Πίνακας 3.12 Γενική περιγραφή αριθμού τελικών διατηρητέων αναγνωσμάτων στο σύνολο των δειγμάτων που προέκυψαν από τον VSEARCH στις διάφορες τιμές i) ελάχιστης βαθμολογίας ποιότητας και ii) ελάχιστου μήκους επικαλυπτόμενης περιοχής. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός διατηρητέων αναγνωσμάτων για **Q_{min}=20**		
$Overlap_{min}$	**10**	**20**	**30**
Ελάχιστος	40389	40389	40389
Μέσος όρος	113806	113806	113806
Μέγιστος	410488	410488	410488
Συνολικός	4893679 ***(43.90%)***	4893679 ***(43.90%)***	4893675 ***(43.90%)***
	Αριθμός διατηρητέων αναγνωσμάτων για **Q_{min}=22**		
$Overlap_{min}$	**10**	**20**	**30**
Ελάχιστος	40029	40029	40029
Μέσος όρος	112951	112951	112951
Μέγιστος	407800	407800	407800
Συνολικός	4856878 ***(43.57%)***	4856878 ***(43.57%)***	4856874 ***(43.57%)***
	Αριθμός διατηρητέων αναγνωσμάτων για **Q_{min}=26**		
$Overlap_{min}$	**10**	**20**	**30**
Ελάχιστος	22129	22129	22129
Μέσος όρος	61374	61374	61374
Μέγιστος	251040	251040	251040
Συνολικός	2639102 ***(23.67%)***	2639102 ***(23.67%)***	2639102 ***(23.67%)***

Ο αριθμός διατηρητέων αναγνωσμάτων ανά δείγμα κυμαίνεται κατά μέσο όρο 61374 με 113806, με τον μεγαλύτερο να το φέρει τα αποτελέσματα με Q_{min}=20 (**Πίνακας 3.12**). Οι ακραίες ελάχιστες τιμές αριθμού διατηρητέων αναγνωσμάτων κυμαίνονται από 22129 έως 40389. Επιπλέον, σε κάθε σύνολο αποτελεσμάτων, το δείγμα που έχει τον μεγαλύτερο αριθμό αναγνωσμάτων εξακολουθεί να αποτελεί έντονη ακραία τιμή στο πλήθος των δειγμάτων, με τον αριθμό να κυμαίνεται στα 251040 με 410488. Εξαιρώντας αυτό το δείγμα, η διακύμανση των υπολοίπων μειώνεται σε σχέση με την αρχική (**Παράρτημα 22**).

Όσον αφορά τον αριθμό λειτουργικών ταξινομικών μονάδων που παράχθηκαν, παρατηρείται μείωση αυτού καθώς αυξάνεται η τιμή ελάχιστης βαθμολογίας ποιότητας (**Πίνακας 3.13**). Για Q_{min}=20 και 22 δεν παρατηρείται σημαντική διαφορά, αλλά για Q_{min}=26 παρουσιάζεται αισθητή μείωση. Οι συχνότητες με τις οποίες παρουσιάζονται τα OTUs στα σύνολα δεδομένων παρουσιάζουν αντίστοιχη συμπεριφορά με τις υπόλοιπες μετρήσεις (**Παράρτημα 23**), με την σημαντική παρατήρηση να αποτελεί ο αυξημένος αριθμός παραγόμενων μοναδικών και χαμηλής αφθονίας OTUs (**Παράρτημα 24**).

Πίνακας 3.13 Αριθμός λειτουργικών ταξινομικών μονάδων (OTUs) που παράχθηκαν από τον VSEARCH στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας και ελάχιστου μήκους επικαλυπτόμενης περιοχής.

	Αριθμός Λειτουργικών Ταξινομικών Μονάδων - OTUs		
Q_{min}	**$Overlap_{min}$=10**	**$Overlap_{min}$=20**	**$Overlap_{min}$=30**
20	9570	9570	9570
22	9553	9553	9553
26	6438	6438	6438

3.6 Ταξινόμηση Ανάθεση Παραγόμενων ASVs και OTUs

Η ταξινόμηση των παραγόμενων ASVs και OTUs πραγματοποιήθηκε χρησιμοποιώντας την βάση δεδομένων SILVA. Παρακάτω, παρουσιάζεται ο αριθμός ταξινομικών μονάδων που προέκυψε μετά την διαδικασία ταξινόμησης. Ο αριθμός αυτός αντιπροσωπεύει τον αριθμό ταξινομήσεων, ή αλλιώς κατηγοριών, από τον οποίο τα ASVs και τα OTUs είχαν επιτυχία κατηγοριοποίησης σε βιολογική πληροφορία που παρέχεται από την βάση δεδομένων. Μεταξύ των διαφορετικών pipelines, η εικόνα των αποτελεσμάτων ως προς το ποσοστό ταξινόμησης των δεδομένων διαφέρει σημαντικά, ειδικά αυτή του DADA2 σε σχέση με την αντίστοιχη εικόνα των Deblur και VSEARCH. Μάλιστα, ένας μεγάλος όγκος δεδομένων που παράχθηκαν από τον DADA2 δεν κατάφερε να ταξινομηθεί, σε αντίθεση με τα αποτελέσματα του Deblur από τον οποία προέκυψε πλήρη ταξινόμηση δεδομένων. Ένας αμελητέος όγκος δεδομένων του VSEARCH δεν κατάφερε να ταξινομηθεί στην βάση δεδομένων.

Πιο αναλυτικά, για κάθε παραμετρικό σενάριο του DADA2, περίπου το 50% των ASVs που παράχθηκαν δεν κατάφερε να ταξινομηθεί στην βάση δεδομένων (**Πίνακας 3.14**). Το ποσοστό των μη-ανατεθειμένων ASVs κυμαίνεται από 36,88% έως 54,03%, με το μεγαλύτερο να παρουσιάζεται στα ASVs που προέκυψαν από την μη-αποκοπή των paired-end αναγνωσμάτων, $e.e_{max}$=2,5 και $Overlap_{min}$=10. Γενικά, παρατηρείται αύξηση του ποσοστό επιτυχίας ταξινόμησης των ASVs καθώς αυξάνεται το ελάχιστο μήκος επικαλυπτόμενης περιοχής και μειώνεται το μέγιστο ποσοστό αναμενόμενων σφαλμάτων. Επιπλέον, τα ASVs που προέκυψαν από την αποκοπή των τελικών άκρων των paired-end αναγνωσμάτων παρουσιάζουν αυξημένο ποσοστό επιτυχίας ταξινόμησης σε σχέση με τα αποτελέσματα χωρίς αποκοπή. Παρόλα αυτά, ο αριθμός των ταξινομημένων ASVs μεταξύ των διαφόρων παραμέτρων έχουν παρόμοιο μοτίβο με αυτόν των παραγόμενων ASVs.

Πίνακας 3.14 Αριθμός ταξινομημένων αντιπροσωπευτικών αναγνωσμάτων (ASVs) που παράχθηκαν από τον DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και συνθήκες αποκοπής paired-end αναγνωσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αρχικό αριθμό ASVs.

	Αριθμός Ταξινομημένων Αντιπροσωπευτικών Αναγνωσμάτων - ASVs					
	No trim			**With trim**		
$Overlap_{min}$ / $e.e._{max}$	**10**	**20**	**30**	**10**	**20**	**30**
2,5	2287 ***(45.97%)***	2261 ***(46.24%)***	2230 ***(46.52%)***	1550 ***(54.29%)***	1453 ***(53.70%)***	1431 ***(54.60%)***
1,5	2275 ***(46.09%)***	2249 ***(46.39%)***	2218 ***(46.67%)***	1535 ***(54.67%)***	1442 ***(54.11%)***	1422 ***(55.05%)***
0,5	1776 ***(49.64%)***	1753 ***(49.60%)***	1731 ***(49.66%)***	1125 ***(62.67%)***	1081 ***(62.41%)***	1073 ***(63.12%)***

Δεδομένου ότι τα μη ταξινομημένα ASVs αφαιρούνται από τα δεδομένα για την περαιτέρω επεξεργασία, τα αναγνώσματα που αντιπροσωπεύονται από αυτά τα ASVs φιλτράρονται εξίσου. Η επίπτωση του φιλτραρίσματος στην συνολική εικόνα των αποτελεσμάτων αποτελεί η ποσοστιαία μείωση του αριθμού αναγνωσμάτων της τάξεως του 2% με 14% σε σχέση με τα πρωτογενή δεδομένα (**Πίνακας 3.15**). Το μεγαλύτερο ποσοστό παρουσιάζεται στα δεδομένα χωρίς αποκοπή των paired-end αναγνωσμάτων, με $e.e_{max}$=2,5 και $Overlap_{min}$=10, ενώ καθώς αυξάνεται το $Overlap_{min}$ και μειώνεται το $e.e_{max}$ το ποσοστό αναγνωσμάτων που δεν κατάφερε ταξινομηθεί μειώνεται. Ενώ γενικά το μοτίβο των φιλτραρισμένων αναγνωσμάτων από μη-ταξινομημένα αναγνώσματα είναι παρόμοιο με αυτό των αρχικά παραγόμενων, στην περίπτωση των αποτελεσμάτων με $e.e_{max}$=0,5 και μη αποκοπή των τελικών άκρων προκύπτει μικρότερο ποσοστό διατηρητέων αναγνωσμάτων σε σχέση με τα αποτελέσματα που δεν εφαρμόστηκε η αποκοπή των τελικών άκρων paired-end αναγνωσμάτων.

Πίνακας 3.15 Γενική περιγραφή αριθμού ταξινομημένων αναγνωσμάτων των συνολικών δειγμάτων που προέκυψαν από τον DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους πεικαλυπτόμενης περιοχής και συνθήκες αποκοπής paired-end αναγνωσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Ταξινομημένων Αναγνωσμάτων με $e.e._{max}$=2,5					
	No trim			**With trim**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	51231	51066	50908	48782	48482	48332
Μέσος Όρος	136438	136376	136279	133292	132731	132686
Μέγιστος	460422	460422	460422	458539	456249	456249
Συνολικός	5866837 ***(52.63%)***	5864167 ***(52.60%)***	5860000 ***(52.56%)***	5731566 ***(51.41%)***	5707421 ***(51.20%)***	5705488 ***(51.18%)***
	Αριθμός Ταξινομημένων Αναγνωσμάτων με $e.e._{max}$=1,5					
	No trim			**With trim**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	38630	38497	38381	37061	36882	36772
Μέσος Όρος	101696	101643	101578	100556	100305	100274
Μέγιστος	362722	362722	362722	364382	363584	363584
Συνολικός	4372920 ***(39.23%)***	4370644 ***(39.21%)***	4367836 ***(39.18%)***	4323899 ***(38.79%)***	4313104 ***(38.69%)***	4311785 ***(38.68%)***

	Αριθμός Ταξινομημένων Αναγνωσμάτων με $e.e._{max}$=0,5					
	No trim			No trim		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστος	7249	7249	7219	7740	7740	7740
Μέσος Όρος	23625	23575	23568	24721	24691	24689
Μέγιστος	108024	107968	107968	111922	111763	111763
Συνολικός	1015889 ***(9.11%)***	1013723 ***(9.09%)***	1013413 ***(9.09%)***	1063019 ***(9.54%)***	1061729 ***(9.52%)***	1061638 ***(9.52%)***

Εστιάζοντας στα δείγματα ξεχωριστά, αρχικά παρατηρείται ότι τα αναγνώσματα των οποίων τα ASVs τους δεν έχουν ταξινομηθεί εντοπίστηκαν μόνο στα δείγματα αίματος και όχι στα δείγματα αρνητικού ελέγχου (**Παράρτημα 25**, **Παράρτημα 26**). Η αφαίρεση αυτών των αναγνωσμάτων από τα δείγματα προκάλεσε την μείωση των διατηρητέων αναγνωσμάτων στα δείγματα αίματος, με την ελάχιστη ακραία τιμή, που είναι ίση με 7219 αναγνώσματα, να την φέρει το παραμετρικό σενάριο χωρίς αποκοπή, $e.e._{max}$=0,5 και $Overlap_{min}$=30. Γενικά, ένας σημαντικός αριθμός δειγμάτων που επεξεργαστήκαν με τιμή $e.e._{max}$=0,5, παρέχει πια αριθμό αναγνωσμάτων μικρότερο από τον ελάχιστο αριθμό αναγνωσμάτων που συνιστάται να έχει το δείγμα για να αντικατοπτριστεί η πραγματική ποικιλομορφία αυτών. Σε κάθε παραμετρικό σενάριο, το δείγμα που έχει τον μεγαλύτερο αριθμό ταξινομημένων αναγνωσμάτων εξακολουθεί να αποτελεί έντονη ακραία τιμή στο πλήθος των δειγμάτων, με τον αριθμό να κυμαίνεται στα 107968 με 460422. Εξαιρώντας αυτό το δείγμα, η διακύμανση των υπολοίπων μειώνεται σε σχέση με την αρχική (**Παράρτημα 27**).

Η αφαίρεση των μη ταξινομημένων ASVs από τα δεδομένα, φέρνει σαν αποτέλεσμα την αύξηση του μέσου όρου των συχνοτήτων με τις οποίες παρουσιάζονται τα ταξινομημένα ASVs (**Παράρτημα 28**). Αυτό σημαίνει ότι η πλειοψηφία των ASVs που δεν ταξινομήθηκαν είτε είναι μοναδικά είτε είχαν πολύ χαμηλή αφθονία. Παρόλα αυτά, ακόμα έχει διατηρηθεί ένας μεγάλος αριθμός μοναδικών και χαμηλής συχνότητας ASVs (**Παράρτημα 29**).

Ο αριθμός των ταξινομικών κατηγοριών που έχει προκύψει δεν διαφέρει σημαντικά στις διάφορες τιμές σημείων αποκοπής, μέγιστου ποσοστού αναμενόμενων σφαλμάτων και ελάχιστου μήκους επικαλυπτόμενης περιοχής (**Πίνακας 3.16**). Τα αποτελέσματα χωρίς αποκοπή των τελικών άκρων των paired-end αναγνωσμάτων φέρουν μεγαλύτερο αριθμό ταξινομήσεων σε σύγκριση με αυτά που έχει εφαρμοστεί αποκοπή. Καθώς αυξάνεται η απαίτηση ως προς την ποιότητα των ASVs παρατηρείται μείωση του αριθμού ταξινομήσεων, με την πιο έντονη να την φέρνει η επιλογή $e.e._{max}$=0,5. Μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής παρουσιάζεται αμελητέα διαφορά, με τον μεγαλύτερο αριθμό ταξινομήσεων να τον φέρνει η επιλογή $Overlap_{min}$=10.

Πίνακας 3.16 Αριθμός ταξινομικών μονάδων (Taxa) που παράχθηκαν από την ταξινόμηση των ASVs του DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και συνθήκες αποκοπής paired-end αναγνωσμάτων.

	Αριθμός Ταξινομικών Μονάδων - Taxa					
	No trim			With trim		
$Overlap_{min}$ / **$e.e._{max}$**	**10**	**20**	**30**	**10**	**20**	**30**
2,5	404	403	403	399	399	399
1,5	403	402	402	396	395	395
0,5	360	360	360	348	348	348

Σε αντίθεση με του DADA2, τα ASVs του Deblur κατάφεραν να ταξινομηθούν όλα στην βάση δεδομένων SILVA. Αυτό σημαίνει ότι όλος ο όγκος δεδομένων που παράχθηκε από τον Deblur (**Παράρτημα 30**, **Παράρτημα 31)** ανεξάρτητα από την επιλογή τιμών των παραμέτρων, κατηγοριοποιείται σε γνωστή βιολογική πληροφορία. Στις διάφορες τιμές παραμέτρων, παρατηρείται ότι η αποκοπή στην 380^{η} βάση φέρει σαν αποτέλεσμα μεγαλύτερο αριθμό ταξινομήσεων σε σχέση με την αποκοπή στην 250^{η} βάση (**Πίνακας 3.17**). Για σταθερές τιμές ελάχιστης βαθμολογίας ποιότητας και σημείου αποκοπής, η επιλογή διαφορετικού ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν μεταβάλει τον αριθμό ταξινομήσεων. Εξίσου σταθερά αποτελέσματα παρουσιάζουν οι τιμές Q_{min}=20 και 22, ενώ για τιμή Q_{min}=26 ο αριθμός ταξινομήσεων μειώνεται.

Πίνακας 3.17 Αριθμός ταξινομικών μονάδων (Taxa) που παράχθηκαν από την ταξινόμηση των ASVs του Deblur στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, ελάχιστου μήκους επικαλυπτόμενης περιοχής και συνθήκες αποκοπής.

	Αριθμός Ταξινομικών Μονάδων - Taxa					
	trim@=250			**trim@=380**		
$Overlap_{min}$ / Q_{min}	**10**	**20**	**30**	**10**	**20**	**30**
20	349	349	349	378	378	378
22	349	349	349	378	378	378
26	321	321	321	345	345	345

Όσον αφορά τα αποτελέσματα ταξινόμησης των OTUs του VSEARCH, σχεδόν όλος ο όγκος δεδομένων που προέκυψε από το pipeline σε κάθε παραμετρικό σενάριο ταξινομήθηκε στην βάση δεδομένων, παρά μονό τρία μοναδικά OTUs. Δηλαδή, τρία μη-ταξινομημένα αναγνώσματα αφαιρέθηκαν από κάθε σύνολο δεδομένων, τα οποία εντοπίστηκαν σε δύο δείγματα αίματος (sample14b, sample12b) και σε ένα δείγμα αρνητικού ελέγχου (negativecontrol3) αντίστοιχα (**Παράρτημα 32**). Ο αριθμός ταξινομήσεων που προέκυψε φαίνεται να μην επηρεάζεται από τις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής (**Πίνακας 3.18**). Αντίθετα, η επιλογή ελάχιστης βαθμολογίας ποιότητας επηρεάζει αντιστρόφως ανάλογα τον αριθμό ταξινομήσεων.

Πίνακας 3.18 Αριθμός ταξινομικών μονάδων (taxa) που παράχθηκαν από την ταξινόμηση των OTUs του VSEARCH στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας και ελάχιστου μήκους επικαλυπτόμενης περιοχής.

	Αριθμός Ταξινομικών Μονάδων – taxa		
Q_{min}	**$Overlap_{min}$=10**	**$Overlap_{min}$=20**	**$Overlap_{min}$=30**
20	479	479	479
22	478	478	478
26	440	440	440

3.7 Επιλογή Τελικών Τιμών Παραμέτρων

Οι τρείς παράγοντες που εξετάζονται για την επιλογή τελικών τιμών παραμέτρων των αλγόριθμων είναι α) ο αριθμός των διατηρητέων αναγνωσμάτων που προέκυψαν μετά την εφαρμογή της ταξινομικής ανάθεσης, β) ο αριθμός των ταξινομημένων ASVs/OTUs καθώς και γ) ο αριθμός ταξινομικών μονάδων που προέκυψαν από αυτά (**Σχήμα 3.6**). Η συλλογιστική πορεία για την επιλογή των κατάλληλων τιμών παραμέτρων είναι η βέλτιστη αποκόμιση ταξινομημένων μονάδων από την βάση δεδομένων, που έχουν προκύψει με όσον

το δυνατόν μικρότερο αριθμό ASVs/OTUs, με τον μέγιστο δυνατό αριθμό διατηρητέων αναγνωσμάτων που χρειάστηκε για την παραγωγή τους, έτσι ώστε να είναι όσο πιο αντιπροσωπευτική η σχετική αφθονία τους. Παράλληλα, πρέπει να ληφθεί υπόψιν το επίπεδο αξιοπιστίας των ίδιων παραμέτρων, καθώς και το γεγονός ότι οι τελικώς επιλεγμένοι παράμετροι και των τριών επεξεργαστικών ροών να είναι όσο τον δυνατόν γίνεται παρόμοιοι για να είναι πιο αντικειμενική η περαιτέρω σύγκρισή τους. Για την καλύτερη κατανόηση των δεδομένων, εστιάζεται κάθε φορά μία συγκεκριμένη παράμετρος και περιγράφεται η αντίστοιχη επιρροή της και στα 3 workflows, ξεκινώντας από αυτήν με την λιγότερη επιρροή.

Αρχικά, η παράμετρος και των τριών workflow που παρουσιάζει την λιγότερη επιρροή στα συγκεκριμένα δεδομένα είναι αυτή του ελάχιστου απαιτούμενου μήκους επικαλυπτόμενης περιοχής ($Overlap_{min}$) για την συγχώνευση των paired-end αναγνωσμάτων. Συγκεκριμένα, για τα workflows των Deblur και VSEARCH οι διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής που έχουν επιλεχθεί δεν επηρεάζουν τα τρία διαφορετικά στοιχεία που μελετάται (διατηρητέα αναγνώσματα, ASVs/OTUs και ταξινομήσεις) όταν οι υπόλοιπες τιμές παραμέτρων παραμένουν σταθερές. Αμελητέα επιρροή παρατηρείται στα αποτελέσματα του DADA2, όπου καθώς αυξάνεται το απαιτούμενο μήκος επικαλυπτόμενης περιοχής, ο αριθμός των τριών στοιχείων ελαφρώς μειώνεται. Δεδομένου αυτού, επιλέγεται και για όλα τα workflows η τιμή ελάχιστου απαιτούμενου μήκους επικάλυψης των paired-end αναγνωσμάτων $Overlap_{min}=10$.

Η επόμενη παράμετρος που εξετάζεται είναι αυτής της αποκοπής των αναγνωσμάτων, εφόσον επηρεάζει τα αποτελέσματα μόνο των δύο εκ των τριών workflows, αυτό του DADA2 και του Deblur. Για την εφαρμογή του Deblur είναι αναγκαστική η αποκοπή των αναγνωσμάτων σε ένα συγκεκριμένο μήκος, σε αντίθεση με την εφαρμογή του DADA2. Συνεπώς, εστιάζεται αρχικά η παράμετρος αποκοπής στον Deblur.

Τα δύο σημεία που επιλέχθηκαν για να αποκοπούν τα συγχωνευμένα αναγνώσματα είναι στην $250^{η}$ βάση και στην $380^{η}$ βάση, που σημαίνει ότι τα αναγνώσματα στην συνέχεια έχουν μήκος 250 bp και 380 bp αντίστοιχα. Τα συγχωνευμένα αναγνώσματα που φέρουν μικρότερο μήκος από το σημείο αποκοπής απορρίπτονται αυτόματα. Η επιλογή του μήκους αναγνωσμάτων να είναι ίσο με 250 bp οδηγεί στην επεξεργασία μεγαλύτερου όγκου συγχωνευμένων αναγνωσμάτων και κατά επέκταση στην αύξηση των ταξινομημένων διατηρητέων αναγνωσμάτων σε αντίθεση με την επιλογή του μήκους 380 bp. Αντίστοιχη αύξηση παρατηρείται στον αριθμό ταξινομημένων ASVs. Παρόλα αυτά, ο αριθμός των ταξινομήσεων είναι μεγαλύτερος στην περίπτωση όπου το μήκος των αναγνωσμάτων είναι 380 bp, που σημαίνει ότι τα παραγόμενα ASVs παρέχουν μεγαλύτερη ευκρίνεια ως προς την βιολογική τους προέλευση. Για τον λόγο αυτό, επιλέγεται για τον Deblur το σημείο αποκοπής των συγχωνευμένων αναγνωσμάτων να είναι η $380^{η}$ βάση.

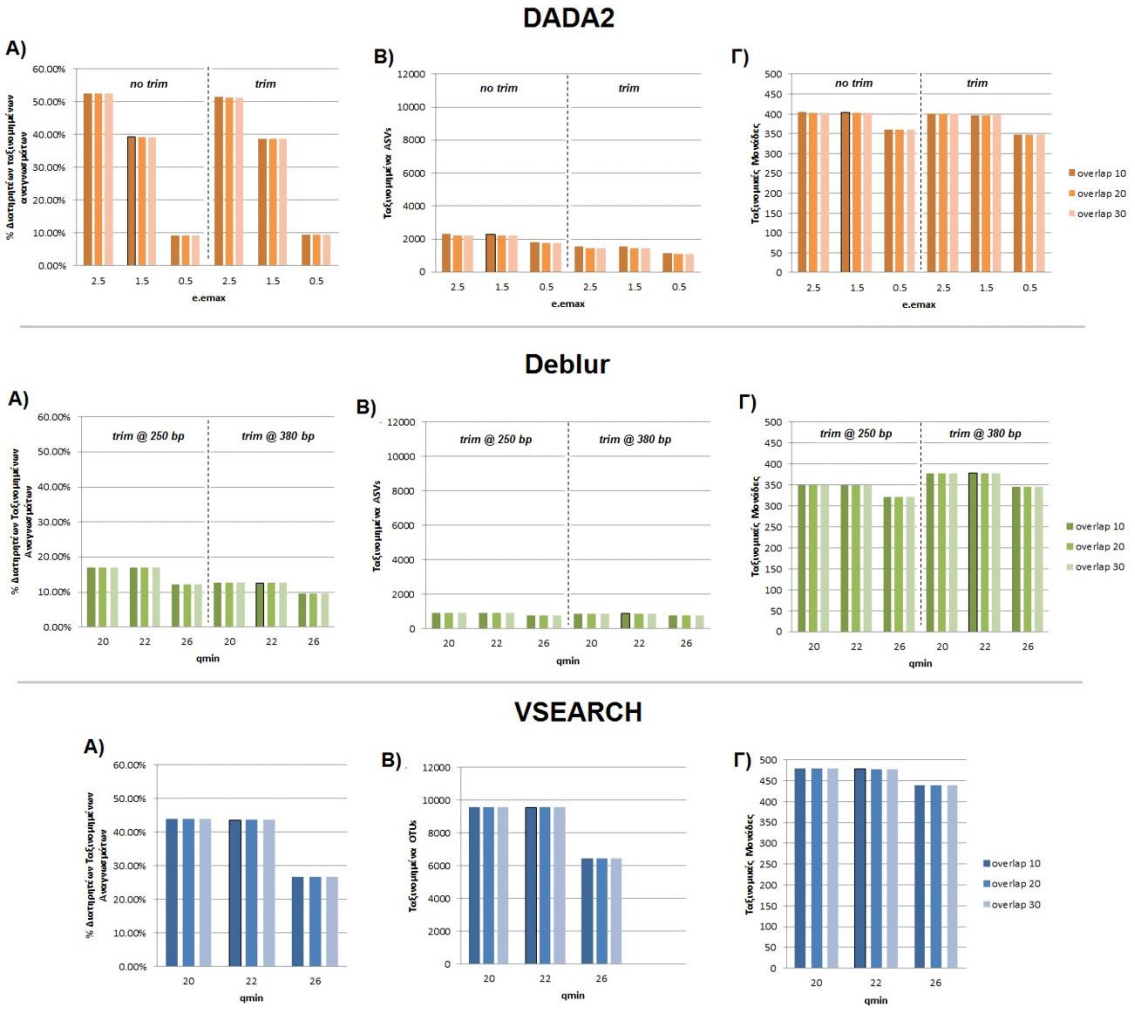
Πριν την εφαρμογή αποθορυβοποίησης του DADA2, παρέχεται η επιλογή αποκοπής των τελικών άκρων των paired-end αναγνωσμάτων, χωρίς όμως αυτή η επιλογή να είναι αναγκαία. Τα αποτελέσματα που προέκυψαν από την μη-αποκοπή paired-end αναγνωσμάτων παρέχουν ελαφρώς αυξημένο αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων καθώς και ταξινομημένων παραγόμενων ASVs σε σχέση με τα αποτελέσματα που προέκυψαν από την εφαρμογή αποκοπής. Επιπλέον, η μη-αποκοπή των μικρο-αναγνωσμάτων οδηγεί στην αντίστοιχη αύξηση του αριθμού ταξινομημένων μονάδων. Συνεπώς, επιλέγονται για την περαιτέρω ανάλυση τα δεδομένα του DADA2 που έχουν προκύψει χωρίς την αφαίρεση των τελικών άκρων των paired-end αναγνωσμάτων. Επίσης, με αυτήν την επιλογή, υπάρχει και μια συνοχή στα δεδομένα που παράγονται και από το workflow του VSEARCH.

Η παράμετρος της οποίας οι διάφορες τιμές οδηγεί σε αποτελέσματα που έχουν την μεγαλύτερη διακύμανση είναι αυτή που αφορά στην διαχείριση ποιότητας των αναγνωσμάτων. Στην περίπτωση του DADA2 η παράμετρος είναι το μέγιστο ποσοστό αναμενόμενων σφαλμάτων $e.e._{max}$ που επιτρέπεται να έχει ένα ανάγνωσμα, ενώ στον Deblur και στον VSEARCH η αντίστοιχη παράμετρος αποτελεί η ελάχιστη βαθμολογία ποιότητας q_{min} που απαιτείται να έχουν οι βάσεις που παρέχει ένα ανάγνωσμα. Ενώ αυτές οι δύο παράμετροι είναι διαφορετικές, καθώς και ο τρόπος διαχείρισης του ποιοτικού ελέγχου μεταξύ αυτών των pipelines είναι διαφορετικός, έχει παρθεί η παραδοχή ότι τα αναγνώσματα που παρέχουν λιγότερο ή ίσο με 2,5% αναμενόμενα σφάλματα, οι βάσεις τους έχουν κατά μέσο όρο βαθμολογία ποιότητας ίσο η μεγαλύτερο από 20. Αντίστοιχα, για 1,5% αναμενόμενων σφαλμάτων τα αναγνώσματα φέρουν βάσεις με βαθμολογία ποιότητας από 22 και πάνω, ενώ για 0,5% αντίστοιχα τα αναγνώσματα φέρουν βάσεις με βαθμολογία ποιότητας από 26 και πάνω.

Η προαναφερόμενη παραδοχή σε ένα βαθμό επιβεβαιώνεται στην συνολική εικόνα των αποτελεσμάτων. Με εξαίρεση το ποσοστό διατηρητέων αναγνωσμάτων, ο τρόπος που μεταβάλλονται τα αποτελέσματα του DADA2 κατά την μείωση της τιμής $e.e._{max}$, είναι παρόμοιος με τον τρόπο μεταβολής των αποτελεσμάτων του Deblur και VSEARCH καθώς αυξάνεται το Q_{min}. Συγκεκριμένα, μεταξύ των τιμών $e.e._{max}$= 2,5 και 1,5 παρατηρείται αμελητέα διαφορά στον αριθμό ταξινομημένων ASVs και ταξινομικών κατηγοριών του DADA2, κάτι που συμβαίνει και στα αποτελέσματα του Deblur και VSEARCH μεταξύ των τιμών Q_{min}= 20 και 22. Η μεγαλύτερη επίδραση σε αυτά τα αποτελέσματα παρατηρείται στην επιλογή τιμής $e.e._{max}$= 0,5 για τον DADA2 και αντίστοιχα στην τιμή Q_{min}= 26 για τον Deblur και τον VSEARCH, όπου ο αριθμός των ταξινομημένων ASVs/OTUs και ο αριθμός ταξινομήσεων μειώνονται σε αυτές τις περιπτώσεις αλλά όχι με πολύ ένταση.

Η μεταβολή του ποσοστού διατηρητέων ταξινομημένων αναγνωσμάτων στις διάφορες τιμές $e.e._{max}$ στα αποτελέσματα του DADA2 δεν παρουσιάζει την ίδια τάση με την αντίστοιχη μεταβολή στις διάφορες τιμές Q_{min} στα αποτελέσματα του Deblur και VSEARCH. Στην περίπτωση των Deblur και VSEARCH, στις διάφορες τιμές Q_{min} παρατηρείται η ίδια μεταβολή του ποσοστού διατηρητέων ταξινομημένων αναγνωσμάτων με την αντίστοιχη των ταξινομημένων ASVs/OTUs και ταξινομημένων μονάδων, δηλαδή χαρακτηρίζεται αμελητέα. Όμως, στην περίπτωση του DADA2, παρατηρείται σημαντική μείωση του ποσοστού διατηρητέων αναγνωσμάτων κατά τη μετάβαση από $e.e._{max}$= 2,5 σε 1,5, ενώ ακόμα πιο αισθητή μεταβολή του ποσοστού αυτού παρατηρείται στη μετάβαση από $e.e._{max}$= 1,5 σε 0,5.

Γενικά, ένας σημαντικός αριθμός δειγμάτων που επεξεργαστήκαν με τιμή $e.e._{max}$=0,5 για τον DADA2 και Q_{min}=26 για Deblur, παρέχει αριθμό διατηρητέων αναγνωσμάτων μικρότερο από τον ελάχιστο αριθμό αναγνωσμάτων που συνιστάται να έχει το δείγμα για να αντικατοπτριστεί η πραγματική ποικιλομορφία αυτών. Επιπλέον, αυτές οι τιμές παραμέτρων επιφέρουν την μείωση ταξινομημένων κατηγοριών και στα τρία pipelines, σε αντίθεση με τις άλλες τιμές. Συνεπώς, η επιλογή τιμών $e.e._{max}$=0,5 και Q_{min}=26 απορρίπτονται. Μεταξύ των υπόλοιπων τιμών $e.e._{max}$ και αντίστοιχων Q_{min}, ο αριθμός των ταξινομημένων ASVs/OTUs και ταξινομικών κατηγοριών δεν παρουσιάζει μεταβολές, ενώ μόνο στην περίπτωση του DADA2 παρατηρείται σημαντική μεταβολή στην αριθμό διατηρητέων αναγνωσμάτων. Δεδομένου ότι επιθυμείται να έχουν όσο το δυνατόν μεγαλύτερη αξιοπιστία τα τελικά αποτελέσματα ως προς την ποιότητά τους, επιλέγονται τα σύνολα δεδομένων που έχουν υποστεί ποιοτικό έλεγχο με τις τιμές $e.e._{max}$=1,5 και Q_{min}=22 για την περαιτέρω επεξεργασία τους.



Σχήμα 3.6 Διαγράμματα ράβδων που απεικονίζουν **(Α)** το ποσοστό διατηρητέων ταξινομημένων αναγνωσμάτων, **(Β)** τον αριθμό ταξινομημένων ASVs/OTUs και **(Γ)** τον αριθμό ταξινομημένων μονάδων που προέκυψαν από τα pipelines DADA2, Deblur και VSEARCH, στις διάφορες τιμές ελάχιστου απαιτούμενου μήκους επικαλυπτόμενης περιοχής paired-end αναγνωσμάτων ($Overlap_{min}$), συνθήκες αποκοπής αναγνωσμάτων και ποιοτικού ελέγχου ($e.e._{max}$ και Q_{min}). Τα σύνολα δεδομένων που επιλέγονται για περαιτέρω ανάλυση παρουσιάζονται στα διαγράμματα με έντονη γραφή στις ράβδους.Οι τελικές τιμές που επιλέγονται είναι για DADA2: [$Overlap_{min}$=10, no trim, $e.e._{max}$=1,5], για Deblur [$Overlap_{min}$=10, trim@= 380, Q_{min}=22] και για VSEARCH: [$Overlap_{min}$=10, Q_{min}=22].

Συνεπώς, τα σύνολα δεδομένων που συγκρατούνται για περαιτέρω ανάλυση για το κάθε pipeline είναι τα εξής:

- Για τον DADA2, επιλέχθηκε το σύνολο δεδομένων όπου δεν πραγματοποιήθηκε αποκοπή στα άκρα των paired-end αναγνωσμάτων (**no trim**), εφαρμόστηκε ποιοτικός έλεγχος με μέγιστο ποσοστό αναμενόμενων σφαλμάτων που μπορεί να παρέχει το μικρο-ανάγνωσμα ίσο με 1,5% (**$e.e._{max}$=1,5**) και η συγχώνευση αυτών επιτυγχάνθηκε με ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής των paired-end αναγνωσμάτων ίσο με 10 bp (**$Overlap_{min}$=10**).
- Για τον Deblur, επιλέχθηκε το σύνολο δεδομένων όπου η συγχώνευση των paired-end αναγνωσμάτων προέκυψε με ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής αυτών ίσο με 10 bp (**$Overlap_{min}$=10**), εφαρμόστηκε ποιοτικός έλεγχος με ελάχιστη τιμή βαθμολογίας ποιότητας Phred ανά βάση ίση με 22 (**Q_{min}=22**) και ορίστηκε το μήκος όλων των αναγνωσμάτων να είναι ίσο με 380 bp (**trim@=380 bp**).

- Για τον VSEARCH, επιλέχθηκε το σύνολο δεδομένων όπου η συγχώνευση των paired-end αναγνωσμάτων προέκυψε με ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής αυτών ίσο με 10 bp (**$Overlap_{min}$=10**) και εφαρμόστηκε ποιοτικός έλεγχος με ελάχιστη τιμή βαθμολογίας ποιότητας Phred ανά βάση ίση με 22 (**Q_{min}=22**).

3.8 Σύνοψη Επιλεγμένων Συνόλων Αποτελεσμάτων

Τα προεπιλεγμένα σύνολα δεδομένων έχουν προκύψει, ναι μεν με διαφορετικό τρόπο επεξεργασίας, αλλά με παρόμοιες συνθήκες ποιοτικού ελέγχου, συμπεριλαμβανομένου της συγχώνευσης και του φιλτραρίσματος χαμηλής ποιότητας αναγνωσμάτων.

Μεταξύ των τριών συνόλων, τον μεγαλύτερο αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων το φέρει το σύνολο δεδομένων που προέκυψε από τον VSEARCH, με ανάκτηση ύψους 43,57% σε σύγκριση με τον αριθμό αναγνωσμάτων των πρωτογενών δεδομένων (**Πίνακας 3.19**). Ακολουθεί το σύνολο δεδομένων του DADA2 με ποσοστό διατηρητέων αναγνωσμάτων ίσο με 39,23% και τέλος το σύνολο δεδομένων του Deblur με αντίστοιχο ποσοστό ίσο με 12,57%. Ο αριθμός αναγνωσμάτων που φέρουν τα δείγματα στα σύνολα των DADA2 και VSEARCH είναι αρκετά παρόμοιος, ενώ τα δείγματα του Deblur παρέχουν αρκετά χαμηλότερο αριθμό διατηρητέων αναγνωσμάτων (**Παράρτημα 33**). Και στα τρία σύνολα δεδομένων παρατηρείται ότι τα δείγματα έχουν παρόμοια ή σχεδόν ίδια διακύμανση μεταξύ τους και εντοπίζεται το δείγμα αίματος του πρώτου ασθενή να έχει σε κάθε περίπτωση πολύ μεγαλύτερο αριθμό αναγνωσμάτων σε σύγκριση με τα υπόλοιπα δείγματα (**Παράρτημα 34** (A)).

Πίνακας 3.19 Γενική περιγραφή αριθμού ταξινομημένων αναγνωσμάτων ανά δείγμα που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

Αριθμός Ταξινομημένων Αναγνωσμάτων			
	DADA2	**Deblur**	**VSEARCH**
Ελάχιστος	38630	12099	40029
Μέσος Όρος	101696	32584	112951
Μέγιστος	362722	109106	407800
Συνολικός	4372920 **(39.23%)**	1401094 **(12.57%)**	4856875 **(43.57%)**

Ο αριθμός των ταξινομημένων ASVs και OTUs παρουσιάζει σημαντική διαφορά μεταξύ των τριών συνόλων (**Πίνακας 3.20**). Ο ελάχιστος αριθμός εντοπίζεται στα αποτελέσματα του Deblur, ενώ σε σχέση με αυτόν, ο DADA2 παρέχει παραπάνω από το διπλάσιο αριθμό ASVs και αντίστοιχα ο VSEARCH φέρει παραπάνω από το δεκαπλάσιο αριθμό OTUs. Βέβαια, παρατηρώντας την συχνότητα των ASVs και των OTUs που φέρουν τα δεδομένα, ένας σημαντικός αριθμός αυτών που έχουν προκύψει από τον DADA2 και VSEARCH αντίστοιχα είναι μοναδικά ή με πολύ χαμηλή αφθονία (**Παράρτημα 34** (B), **Παράρτημα 35**). Αντίθετα, ο Deblur δεν παρέχει μοναδικά ASVs, παρόλα αυτά περιλαμβάνονται και σε αυτό το σύνολο δεδομένων ASVs με χαμηλή σχετική αφθονία (<0.002%).

Όσον αφορά στον αριθμό ταξινομικών μονάδων, παρατηρείται επίσης ότι ο μεγαλύτερος αριθμός προκύπτει από τον VSEARCH και ο μικρότερος αντίστοιχα από τον

Deblur (**Πίνακας 3.20**). Παρόλα αυτά, Ενώ οι διαφορές είναι αντίστοιχες με αυτές του αριθμού ASVs και OTUs, δεν παρουσιάζουν αντίστοιχη αναλογία. Σε σχέση με τον μικρότερο αριθμό ταξινομήσεων που τον φέρει ο Deblur, ο DADA2 και ο VSEARCH παρέχουν αυξημένο αριθμό ταξινομικών κατηγοριών προσεγγιστικά κατά 6% και 20% αντίστοιχα. Είναι σημαντικό να σημειωθεί ότι οι ταξινομικές κατηγορίες του κάθε συνόλου δεδομένων μπορεί να περιλαμβάνουν στοιχεία που δεν αφορούν τον ταξινομικό χαρακτηρισμό βακτηρίων, γεγονός που παρουσιάζεται στα αποτελέσματα της επόμενης υποενότητας.

Πίνακας 3.20 Αριθμός ταξινομημένων αντιπροσωπευτικών αναγνωσμάτων (ASVs), λειτουργικών ταξινομικών μονάδων (OTUs) και ταξινομικών μονάδων (Taxa) που παράχθηκαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH.

pipeline	Αριθμός Ταξινομημένων ASVs/OTUs	Αριθμός Ταξινομικών Μονάδων
DADA2	2275	403
Deblur	884	378
VSEARCH	9550	478

3.9 Αφαίρεση Μη-Στοχευόμενων Δεδομένων

Δεδομένου ότι ένας από τους στόχους της παρούσας εργασίας είναι η ανάλυση της βακτηριακής κοινότητας από δεδομένα του 16S rRNA γονιδίου, είναι απαραίτητη η αφαίρεση των ASVs και OTUs από τα δεδομένα που ταξινομήθηκαν σε ταξινομικές κατηγορίες που δεν αποτελούν βακτηριακή ταυτότητα. Τα στοιχεία προς αφαίρεση είναι αυτά που ταξινομήθηκαν σε αρχαία, μιτοχόνδρια, χλωροπλάστες και ευκαριώτα. Επιπλέον, τα ASVs και OTUs που δεν κατάφεραν να ταξινομηθούν σε τουλάχιστον επίπεδο φυλής καθώς και αυτών που ταξινομήθηκαν στην κατηγορία του «μη καλλιεργημένου» (uncultured) αφαιρούνται εξίσου από όλα τα δεδομένα, διότι δεν προσφέρουν χαρακτηριστική πληροφορία για την προέλευση των αλληλουχιών τους και ενδεχομένως η συγκράτησή τους να αυξήσει το ρίσκο να παρουσιαστεί εσφαλμένη αυξημένη ποικιλομορφία σε περαιτέρω στάδιο ανάλυσης.

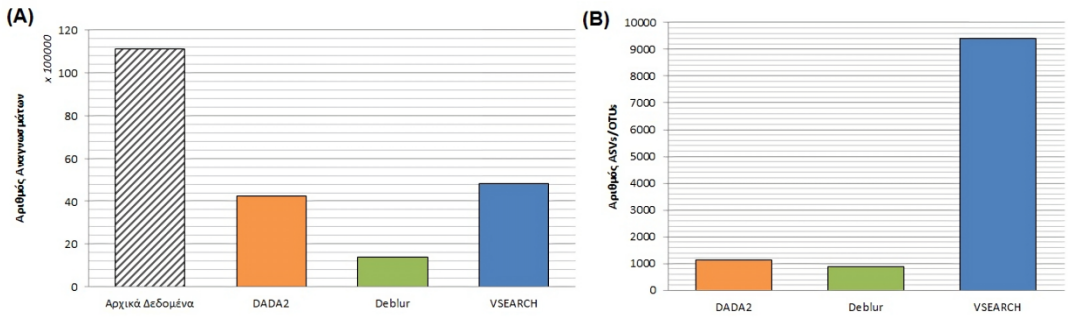
Με εξαίρεση στον αριθμό ASVs του DADA2, η επίπτωση της αφαίρεσης των προαναφερόμενων στοιχείων δεν δεν είναι μεγάλη στο σύνολο των δεδομένων (**Πίνακας 3.21**). Στα δεδομένα του Deblur εντοπίστηκαν μόνο 14 ASVs που χαρακτηρίζονται μη-στοχευόμενα στοιχεία, τα οποία αντιπροσώπευαν λιγότερο από 0,01% των συνολικών αρχικών αναγνωσμάτων. Ακολουθούν τα αποτελέσματα του VSEARCH, όπου προσδιορίστηκε περίπου το 2% των OTUs να αποτελεί μη-στοχευόμενα δεδομένα. Αυτά τα OTUs εκπροσωπούσαν προσεγγιστικά μόνο το 0,3% αναγνωσμάτων σε σχέση με τον αρχικό συνολικό αριθμό αναγνωσμάτων. Στην περίπτωση του DADA2, παρατηρείται ότι λιγότερο από το 50% των ταξινομημένων ASVs που παράχθηκαν κατάφερε να ταξινομηθεί σε τουλάχιστον επίπεδο φυλής βακτήριο. Παρόλα αυτά, δεν παρατηρείται σημαντική μεταβολή στον αριθμό διατηρητέων αναγνωσμάτων κατά την αφαίρεση των μη-στοχευόμενων στοιχείων, όπου μόνο το 1,3% αυτών να απορρίπτεται σε αυτό το στάδιο.

Πίνακας 3.21 Γενική περιγραφή αριθμού ταξινομημένων αναγνωσμάτων στο σύνολο των δειγμάτων, ο αριθμός ταξινομημένων αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) και η γενική περιγραφή των συχνοτήτων τους που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση μη στοχευόμενων αποτελεσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Καθαρών Ταξινομημένων Αναγνωσμάτων		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
Ελάχιστος	35796	12099	40029
Μέσος Όρος	98331	32273	112175
Μέγιστος	359280	108177	404121
Συνολικός	4228215 ***(37.93%)***	1387723 ***(12.45%)***	4823528 ***(43.27%)***
	Αριθμός Καθαρών Ταξινομημένων ASVs/OTUs		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
	1132	870	9393
	Συχνότητα Καθαρών Ταξινομημένων ASVs/OTUs		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
Ελάχιστος	1	10	1
Μέσος Όρος	3735	1595	514
Μέγιστος	618921	266201	462980

Από την επισκόπηση των συχνοτήτων που παρέχουν διατηρητέα ASVs και OTUs, καθώς και από τα αντίστοιχα διαγράμματα (**Πίνακας 3.21**, **Παράρτημα 36** (B)), επιβεβαιώνεται ότι και στα τρία σύνολα δεδομένων τα στοιχεία που φιλτραρίστηκαν αποτελούσαν μοναδικά ASVs και OTUs ή χαμηλής αφθονίας αντίστοιχα. Ωστόσο, είναι σημαντικό να σημειωθεί ότι ακόμα υπάρχουν στα δεδομένα μοναδικά ή χαμηλής αφθονίας ASVs και OTUs. Τα ιστογράμματα αριθμού αναγνωσμάτων ανά δείγμα δεν φανερώνουν σημαντική αλλαγή στην διακύμανση του αριθμού αναγνωσμάτων των δειγμάτων κατά την αφαίρεση των μη στοχευόμενων δεδομένων (**Παράρτημα 36** (A)).

Σε αυτό το στάδιο ανάλυσης, όπου έχει πραγματοποιηθεί η αφαίρεση μη-στοχευόμενων αλληλουχιών, τα τρία σύνολα δεδομένων παρέχουν αναγνώσματα που αντιπροσωπεύονται από ASVs και OTUs, των οποίων οι αλληλουχίες τους φέρουν πια βάσιμες ταξινομικές πληροφορίες σχετικά με την ταυτοποίηση βακτηριακής προέλευσης. Είναι εμφανές ότι η κάθε επεξεργαστική ροή έχει παράξει διαφορετικά αποτελέσματα ως προς τον αριθμό διατηρητέων αναγνωσμάτων σε σχέση με τον αντίστοιχο των πρωτογενών δεδομένων καθώς και ως προς τον αριθμό διατηρητέων ASVs και OTUs (**Σχήμα 3.7**). Σε ότι αφορά στον αριθμό των βακτηριακών ταξινομικών μονάδων (Taxa) στα διάφορα ταξινομικά επίπεδα που φέρουν τα σύνολα δεδομένων, δεν παρατηρούνται μεγάλες διαφορές (**Πίνακας 3.22**).

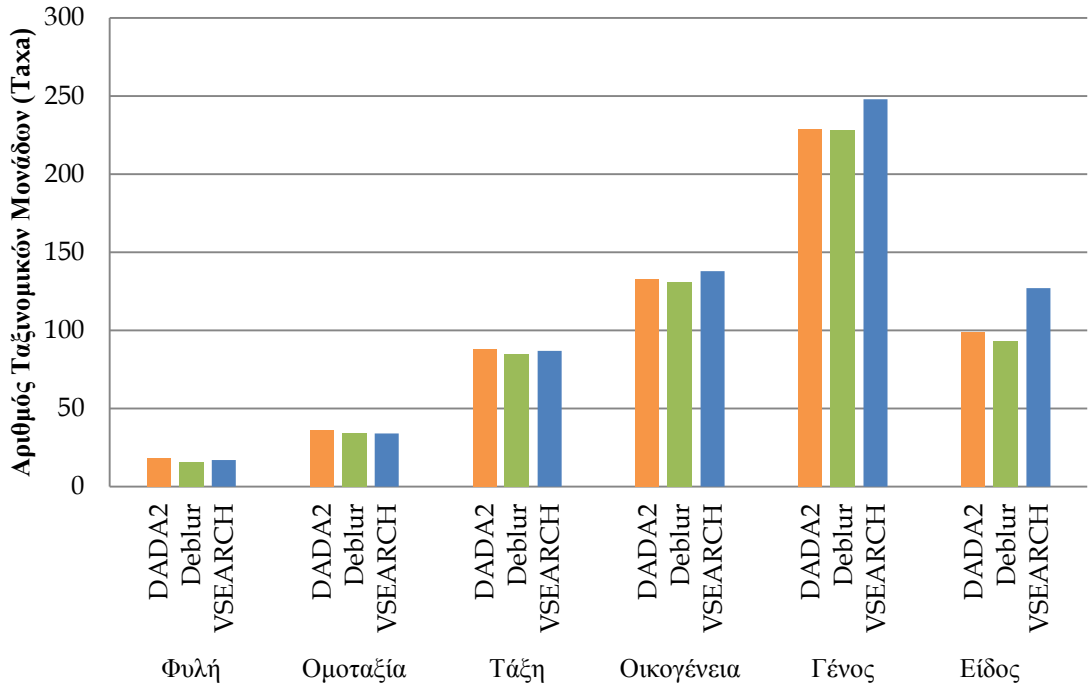


Σχήμα 3.7 Διαγράμματα ράβδων που απεικονίζουν **(A)** τον αριθμό των αρχικών και διατηρητέων αναγνωσμάτων καθώς και **(B)** τον αριθμό των παραγόμενων αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση μη στοχευόμενων αποτελεσμάτων.

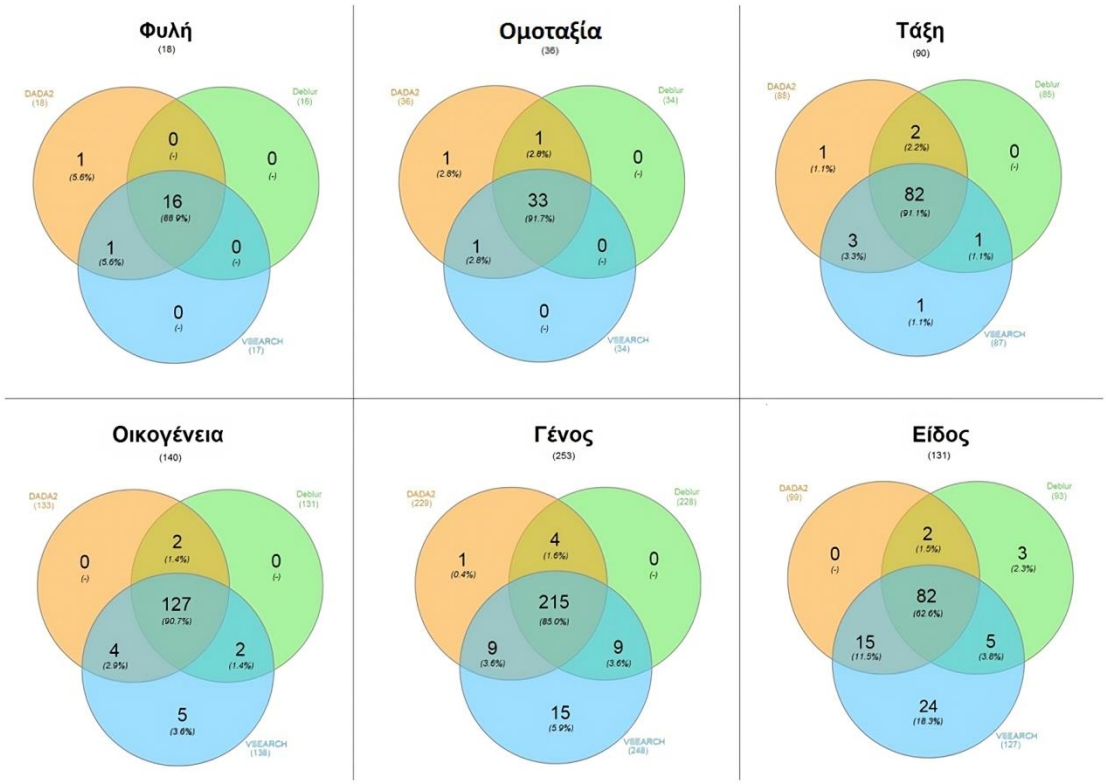
Η γενική εικόνα του διαγράμματος ράβδων του αριθμού των βακτηριακών ταξινομικών μονάδων ανά ταξινομικό επίπεδο που προέκυψε και από τις τρεις επεξεργαστικές ροές φανερώνει ένα μοτίβο (**Σχήμα 3.8**). Καθώς αυξάνεται η ευκρίνεια την ταξινομικής βακτηριακής ταυτότητας, αυξάνεται ο αριθμός ταξινομικών κατηγοριών έως και το επίπεδο γένους, ενώ εν συνεχεία παρατηρείται μειώση στο επίπεδο είδους. Η δραματική μείωση του αριθμού ταξινομήσεων σε επίπεδο είδους υποδεικνύει την αδυναμία και των τριών συνόλων δεδομένων να ταξινομηθούν σε ακριβή ταυτότητα βακτηρίου. Βέβαια, είναι σημαντικό να σημειώθεί ότι μπορεί να οφείλεται στην αδυναμία της διαδικασία της ταξινομικής ανάθεσης σε επίπεδο είδους λόγω της φύσης των δεδομένων του 16S rRNA γονιδίου, καθώς και της μη πληρότητας των βάσεων δεδομένων. Ενώ και στα τρία σύνολα παρατηρείται το ίδιο μοτίβο, συγκρίνοντας τα αποτελέσματά μεταξύ τους εντοπίζονται μερικές διαφορές. Συγκεκριμένα, ο αριθμός των βακτηριακών ταξινομήσεων των δεδομένων του Deblur είναι μικρότερος σε όλα τα ταξινομικά επίπεδα σε σχέση με τα αποτελέσματα των υπόλοιπων συνόλων. Μεταξύ των δεδομένων του DADA2 και VSEARCH, παρατηρείται μια εναλλαγή του μέγιστου αριθμού ταξινομήσεων στα διάφορα επίπεδα. Μάλιστα, ξεκινώντας από το επίπεδο φυλής μέχρι και αντίστοιχα της τάξης, ο DADA2 παρέχει ελαφρώς μεγαλύτερο αριθμό βακτηριακών ταξινομικών κατηγοριών σε σχέση με τα αποτελέσματα των άλλων επεξεργαστικών ροών, ενώ από το επίπεδο οικογένειας μέχρι και στο αντίστοιχο του είδους, τα δεδομένα του VSEARCH καταλήγουν να έχουν μεγαλύτερο αριθμό ταξινομήσεων σε σύγκριση με τα άλλα δύο σύνολα. Σε έναν βαθμό, μπορεί να χαρακτηριστεί αναμενόμενος ο αυξημένος αριθμός των ταξινομήσεων που παρέχουν τα δεδομένα του VSEARCH, δεδομένου ότι φέρουν μεγαλύτερο αριθμό OTUs. Παρόλα αυτά, δεν αντικατοπτρίζεται ανάλογα αυτή η αύξηση.

Πίνακας 3.22 Ο αριθμός των βακτηριακών ταξινομικών μονάδων (Taxa) στα διάφορα ταξινομικά επίπεδα που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση μη στοχευόμενων αποτελεσμάτων.

Αριθμός Βακτηριακών Ταξινομικών Μονάδων			
pipeline / Επίπεδο	**DADA2**	**Deblur**	**VSEARCH**
Φυλή	18	16	17
Ομοταξία	36	34	34
Τάξη	88	85	87
Οικογένεια	133	131	138
Γένος	229	228	248
Είδος	99	93	127



Σχήμα 3.8 Διάγραμμα ράβδων που απεικονίζει τον αριθμό των βακτηριακών ταξινομικών μονάδων (Taxa) στα διάφορα ταξινομικά επίπεδα που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση μη στοχευόμενων αποτελεσμάτων.



Σχήμα 3.9 Διαγράμματα Venn κοινών και μοναδικών βακτηριακών ταξινομικών κατηγοριών (taxa) στα 3 σύνολα δεδομένων που παράχθηκαν από τις επεξεργαστικές ροές των DADA2, Deblur, VSEARCH σε διαφορετικά ταξινομικά επίπεδα. Τα συνολικά taxa ανά επίπεδο εμφανίζονται κάτω από τον τίτλο του επιπέδου ταξινόμησης.

Τα διαγράμματα Venn δείχνουν μια σχετικά μεγάλη αντιστοίχιση ταξινομικών κατηγοριών από τα τρία σύνολα δεδομένων, όπου ανιχνεύθηκαν 88,9% κοινά φύλα, 91,7% κοινές ομοταξίες, 91,1% κοινές τάξεις, 90,7% κοινές οικογένειες, 85,0% κοινά γένη και 62,6% κοινά είδη (**Σχήμα 3.9**). Το μειωμένο ποσοστό κοινών ταξινομήσεων στο επίπεδο είδους υποδηλώνει την αδυναμία των επεξεργαστικών ροών να παράξουν ASVs/OTUs που να καταλήγουν στην ίδια ακριβή βακτηριακή ταυτότητα. Μπορεί επίσης να οφείλεται εξίσου στην διαδικασία της ταξινομικής ανάθεσης και στην αδυναμία της να ταξινομήσει σε επίπεδο είδους ή στη φύση της βάσης δεδομένων SILVA η οποία δεν παρέχει μεγάλο αριθμό ονομαζόμενων ταξινομικών κατηγοριών έως και επίπεδο είδους. Γενικά, παρατηρείται ένα ελαφρώς αυξημένο ποσοστό ομοιότητας των αποτελεσμάτων του DADA2 με αυτά των άλλων δύο συνόλων από το επίπεδο φυλής έως και το αντίστοιχο της τάξης. Από το επίπεδο της οικογένειας έως και αυτό του είδους, φανερώνεται ένα αυξημένο ποσοστό ομοιότητας των αποτελεσμάτων του VSEARCH με τα υπόλοιπα δεδομένα. Επιπλέον, ο VSEARCH παρουσιάζει τον μεγαλύτερο αριθμό μοναδικών ταξινομικών κατηγοριών στα περισσότερα ταξινομικά επίπεδα, σε αντίθεση με τον Deblur στον οποίο παρατηρείται πολύ μικρός αριθμός μοναδικών ταξινομήσεων.

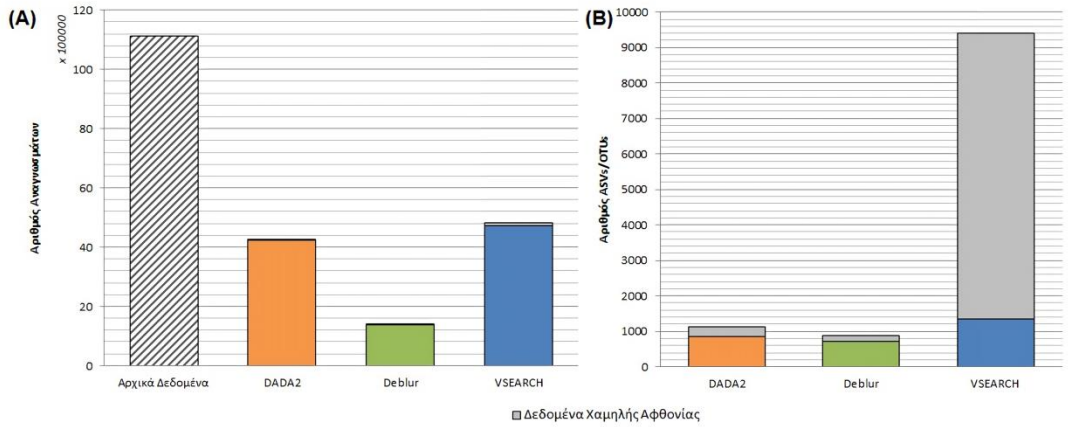
3.10 Φιλτράρισμα Χαμηλής Σχετικής Αφθονίας Taxa

Στην προηγούμενη υποενότητα, στα σύνολα δεδομένων συμπεριλαμβανόταν ένας σημαντικός αριθμός μοναδικών ή πολύ χαμηλής σχετικής αφθονίας ASVs και OTUs. Αυτά τα στοιχεία, ενώ ταξινομήθηκαν, ενδέχεται να μην αντιπροσωπεύουν την πραγματική βιολογική ποικιλότητα, αλλά να έχουν προκύψει από σφάλματα της ενίσχυσης PCR, της αλληλούχισης ή και σε κάποιες περιπτώσεις λόγω επιμόλυνσης. Για τον λόγο αυτό, τα σύνολα δεδομένων υποβλήθηκαν σε διαδικασία φιλτραρίσματος ξεχωριστά, κατά την οποία αφαιρέθηκαν οι ταξινομικές κατηγορίες που αντιπροσωπεύονταν με λιγότερη από 0,002% σχετική αφθονία στα δεδομένα. Έτσι, με την αποφυγή συμπερίληψης αυτών των αλληλουχιών στις περαιτέρω αναλύσεις, ελαχιστοποιείται το ρίσκο της πρόσδεσης εσφαλμένα αυξημένης ποικιλομορφίας στα δείγματα. Συνοπτικά, η αφαίρεση των χαμηλής σχετικής αφθονίας δεδομένων έφερε ως αποτέλεσμα την εξομάλυνση των δεδομένων μεταξύ τους όσον αφορά στον αριθμό των ταξινομημένων ASVs και OTUs καθώς και τον αριθμό ταξινομικών κατηγοριών.

Πιο αναλυτικά, η μεγαλύτερη μεταβολή και στα τρία σύνολα δεδομένων εντοπίζεται στον αριθμό ταξινομημένων ASVs και OTUs κατά την αφαίρεση των δεδομένων χαμηλής αφθονίας (**Πίνακας 3.23**, **Σχήμα 3.10**). Μάλιστα, από τα δεδομένα που έχουν προκύψει από το DADA2, εντοπίστηκε και αφαιρέθηκε περίπου το 25% των παραγόμενων ταξινομημένων ASVs, στα δεδομένα του Deblur αντίστοιχα το 18% αυτών, ενώ στην περίπτωση του VSEARCH προέκυψε πάνω από το 85% των OTUs να αποτελείται από μοναδικά και χαμηλής αφθονίας στοιχεία. Τα ιστογράμματα των ASVs και OTUs που παρέχουν μια δεδομένη συχνότητα στα σύνολα δεδομένων φανερώνουν την απουσία μοναδικών και πολύ χαμηλής αφθονίας στοιχείων (**Παράρτημα 38** (B)). Επιπλέον, παρατηρείται αμελητέα μεταβολή του συνολικού αριθμού διατηρητέων ταξινομημένων αναγνωσμάτων και στα τρία σύνολα δεδομένων (**Πίνακας 3.23**), ενώ αντίστοιχη μεταβολή εντοπίζεται και στον αριθμό αναγνωσμάτων ανά δείγμα (**Παράρτημα 38** (A)).

Πίνακας 3.23 Γενική περιγραφή αριθμού ταξινομημένων αναγνωσμάτων στο σύνολο των δειγμάτων, ο αριθμός ταξινομημένων αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) και η γενική περιγραφή των συχνοτήτων τους που προέκυψαν από τις επεξεργαστικές ροές DADA2, Deblur και VSEARCH μετά το φιλτράρισμα χαμηλής αφθονίας ταξινομικών κατηγοριών. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

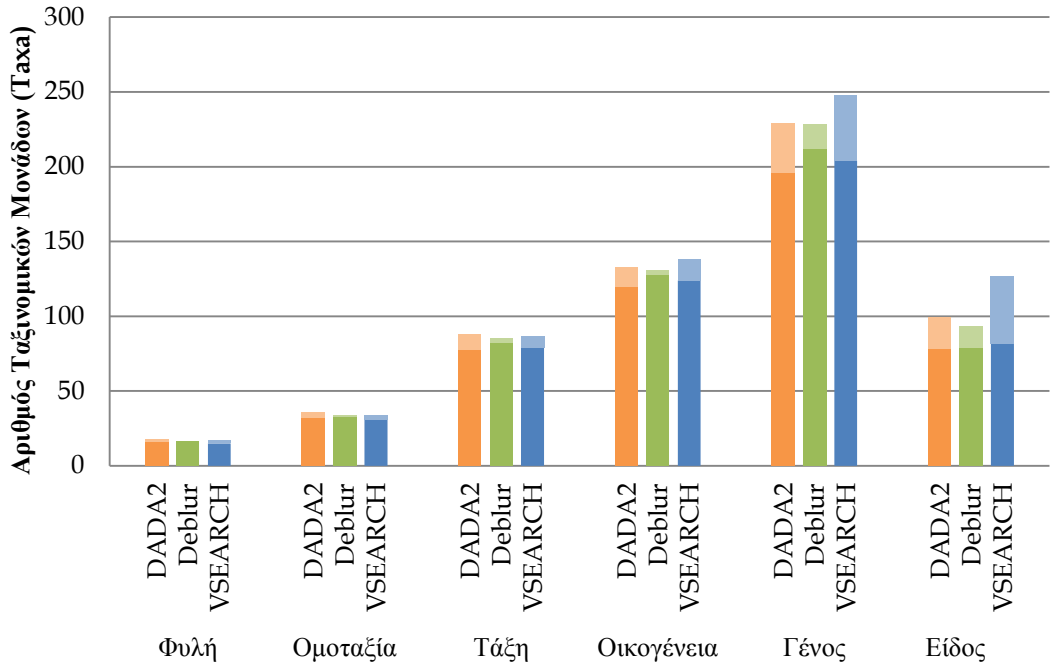
	Αριθμός Φιλτραρισμένων Ταξινομημένων Αναγνωσμάτων		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
Ελάχιστος	35610	12073	39000
Μέσος Όρος	98071	32212	110038
Μέγιστος	358992	108116	396257
Συνολικός	4217049 ***(37.83%)***	1385130 ***(12.42%)***	4731631 ***(42.44%)***
	Αριθμός Φιλτραρισμένων Ταξινομημένων ASVs/OTUs		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
	851	711	1356
	Συχνότητα Φιλτραρισμένων Ταξινομημένων ASVs/OTUs		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
Ελάχιστος	98	28	96
Μέσος Όρος	4955	1948	3489
Μέγιστος	618921	266201	462980



Σχήμα 3.10 Διαγράμματα ράβδων που απεικονίζουν **(A)** τον αριθμό των αρχικών και την επίδραση των διατηρητέων αναγνωσμάτων καθώς και **(B)** την επίδραση του αριθμού των παραγόμενων αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά το φιλτράρισμα χαμηλής αφθονίας ταξινομικών κατηγοριών.

Σε ότι αφορά τις βακτηριακές ταξινομικές κατηγορίες, το διάγραμμα ράβδων παρουσιάζει την τάση τα δεδομένα να παρέχουν σχεδόν ίδιο αριθμό ταξινομήσεων μετά την αφαίρση των χαμηλής αφθονίας στοιχείων (**Σχήμα 3.11**), ενώ τα διαγράμματα Venn φανερώνουν την αύξηση του ποσοστού όμοιων ταξινομικών κατηγοριών μεταξύ των τριών συνόλων δεδομένων (**Σχήμα 3.12**). Συγκεκριμένα, ενώ το μοτίβο του αριθμού ταξινομικών κατηγοριών ανά επίπεδο παραμένει ίδιο με αυτό πριν το φιλτράρισμα χαμηλής αφθονίας ταξινομήσεων, παρατηρείται μείωση των στοιχείων αυτών και στα τρία σύνολα δεδομένων. Η μεγαλύτερη μεταβολή του αριθμού ταξινομήσεων εντοπίζεται στις ταξινομήσεις του επίπεδου γένους και είδους, με το σύνολο δεδομένων του VSEARCH να παρουσιάζει την μεγαλύτερη αντίστοιχη απώλεια (18% στο επίπεδο γένους και 35% για επίπεδο είδους). Το σύνολο ταξινομήσεων του DADA2 παρουσιάζει την δεύτερη μεγαλύτερη απώλεια

ταξινομικών κατηγοριών στα αντίστοιχα ταξινομικά επίπεδα, (14% στο επίπεδο γένους και 21% για επίπεδο είδους). Το σύνολο δεδομένων του Deblur είναι αυτό που διατήρησε τον μεγαλύτερο αριθμό ταξινομικών κατηγοριών μετά την αφαίρεση χαμηλής αφθονίας στοιχείων σχεδόν σε όλα τα ταξινομικά επίπεδα σε σχέση με τα άλλα δύο σύνολα, και κατά επέκταση έχει το μικρότερο ποσοστό απώλειας ταξινομήσεων (7% στο επίπεδο γένους και 15% για επίπεδο είδους). Παρόλα αυτά, είναι σημαντικό να σημειωθεί ότι η επίπτωση της αφαίρεσης των χαμηλής σχετικής αφθονίας δεδομένων έφερε ως αποτέλεσμα την μείωση της διαφοράς του αριθμού ταξινομικών κατηγοριών μεταξύ των συνόλων αποτελεσμάτων των διαφορετικών επεξεργαστικών ροών (**Πίνακας 3.24**).



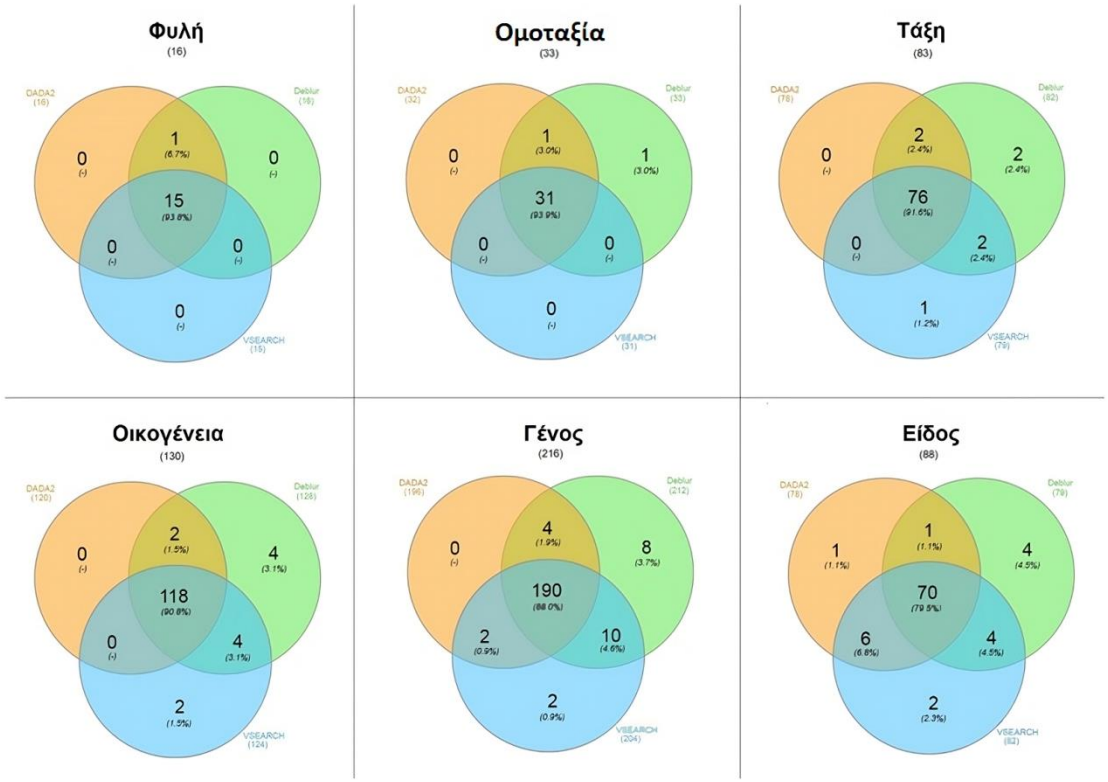
Σχήμα 3.11 Διάγραμμα ράβδων που απεικονίζει την επίδραση του αριθμού των βακτηριακών ταξινομικών μονάδων (Taxa) στα διάφορα ταξινομικά επίπεδα που προέκυψαν από τις επεξεργαστικές ροές DADA2, Deblur και VSEARCH μετά το φιλτράρισμα χαμηλής αφθονίας αποτελέσματα. Οι σκουρόχρωμες ράβδοι δείχνουν τα στοιχεία που απομένουν μετά τα φίλτρα αφθονίας και οι αντίστοιχες ανοιχτόχρωμες τα στοιχεία που φιλτραρίστηκαν.

Πίνακας 3.24 Ο αριθμός των βακτηριακών ταξινομικών μονάδων (Taxa) στα διάφορα ταξινομικά επίπεδα που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά το φιλτράρισμα χαμηλής αφθονίας ταξινομικών κατηγοριών.

	Αριθμός Φιλτραρισμένων Ταξινομικών Μονάδων		
pipeline / Επίπεδο	**DADA2**	**Deblur**	**VSEARCH**
Φυλή	16	16	15
Ομοταξία	32	33	31
Τάξη	78	82	79
Οικογένεια	120	128	124
Γένος	196	212	204
Είδος	78	79	82

Εμβαθύνοντας στο περιεχόμενο των βακτηριακών ταξινομικών κατηγοριών που παρέχουν τα τρία σύνολα δεδομένων, παρατηρείται αύξηση της αντιστοίχισης από τα σύνολα

αυτά στα διαφορετικά ταξινομικά επίπεδα κατά την αφαίρεση χαμηλής σχετικής αφθονίας ταξινομήσεων. Μάλιστα, ανιχνεύθηκαν 93,8% κοινά φύλα, 93,9% κοινές ομοταξίες, 91,6% κοινές τάξεις, 90,8% κοινές οικογένειες, 88,0% κοινά γένη και 79,5% κοινά είδη. Γενικά, εντοπίζεται σχεδόν εξολοκλήρου ταύτιση των αποτελεσμάτων του DADA2 με αυτά των άλλων δύο συνόλων σε όλα τα ταξινομικά επίπεδα. Τα αποτελέσματα του VSEARCH παρουσιάζουν επίσης σημαντικό ποσοστό ομοιότητας με τα αποτελέσματα των υπολοίπων συνόλων δεδομένων, με την μεγαλύτερη ομοιότητα να εντοπίζεται με τα δεδομένα του Deblur. Τέλος, σε αντίθεση με τα δεδομένα πριν εφαρμοστεί φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομήσεων, ο Deblur κατέληξε να παρέχει τον μεγαλύτερο αριθμό μοναδικών ταξινομικών κατηγοριών στα περισσότερα ταξινομικά επίπεδα, κάτι που ήταν αναμενόμενο δεδομένου ότι παρέχει τον μεγαλύτερο αριθμό ταξινομήσεων γενικά.



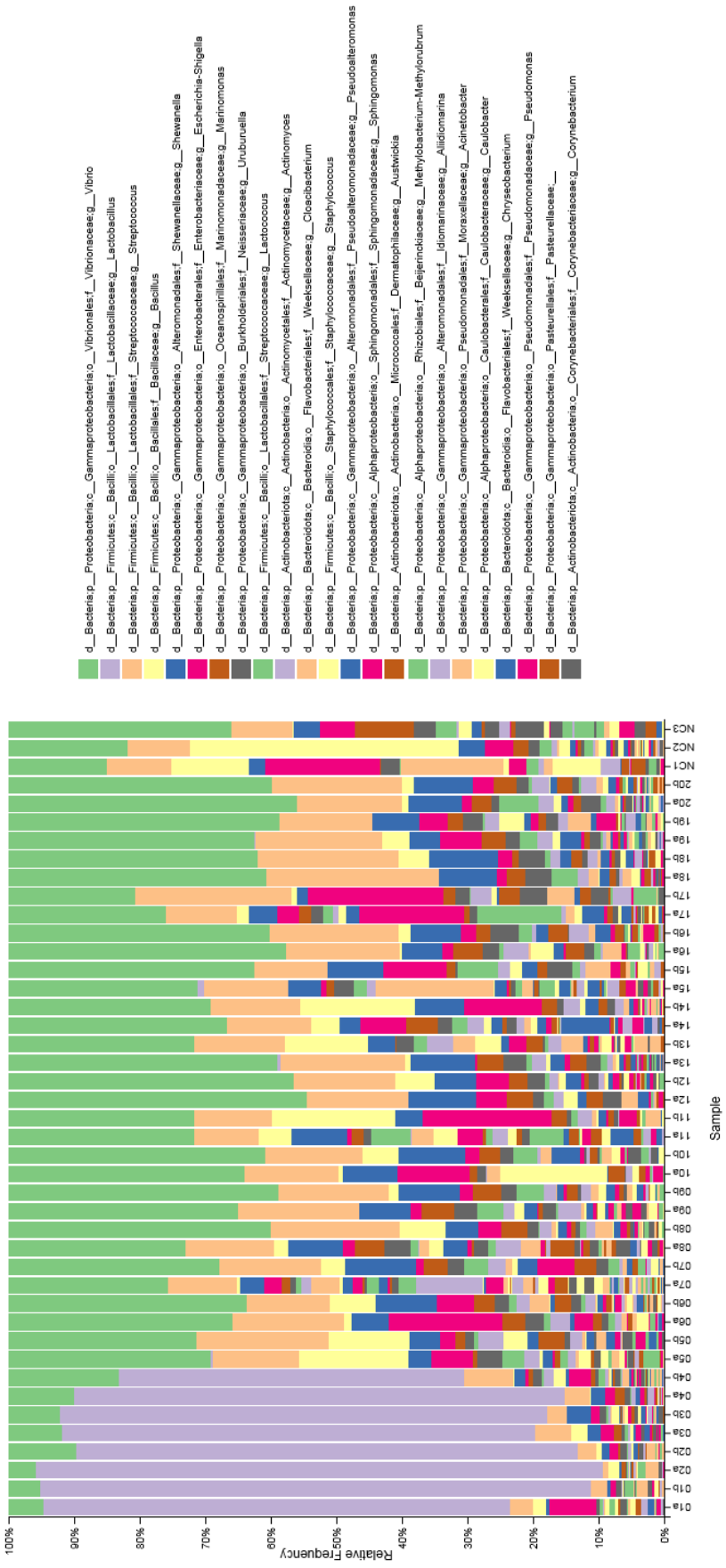
Σχήμα 3.12 Διαγράμματα Venn κοινών και μοναδικών βακτηριακών ταξινομικών κατηγοριών (taxa) στα 3 σύνολα δεδομένων που παράχθηκαν από τις επεξεργαστικές ροές των DADA2, Deblur, VSEARCH σε διαφορετικά ταξινομικά επίπεδα μετά το φιλτράρισμα χαμηλής αφθονίας taxa. Τα συνολικά taxa ανά επίπεδο εμφανίζονται κάτω από τον τίτλο του επιπέδου ταξινόμησης.

3.11 Αφαίρεση Πιθανής Επιμόλυνσης

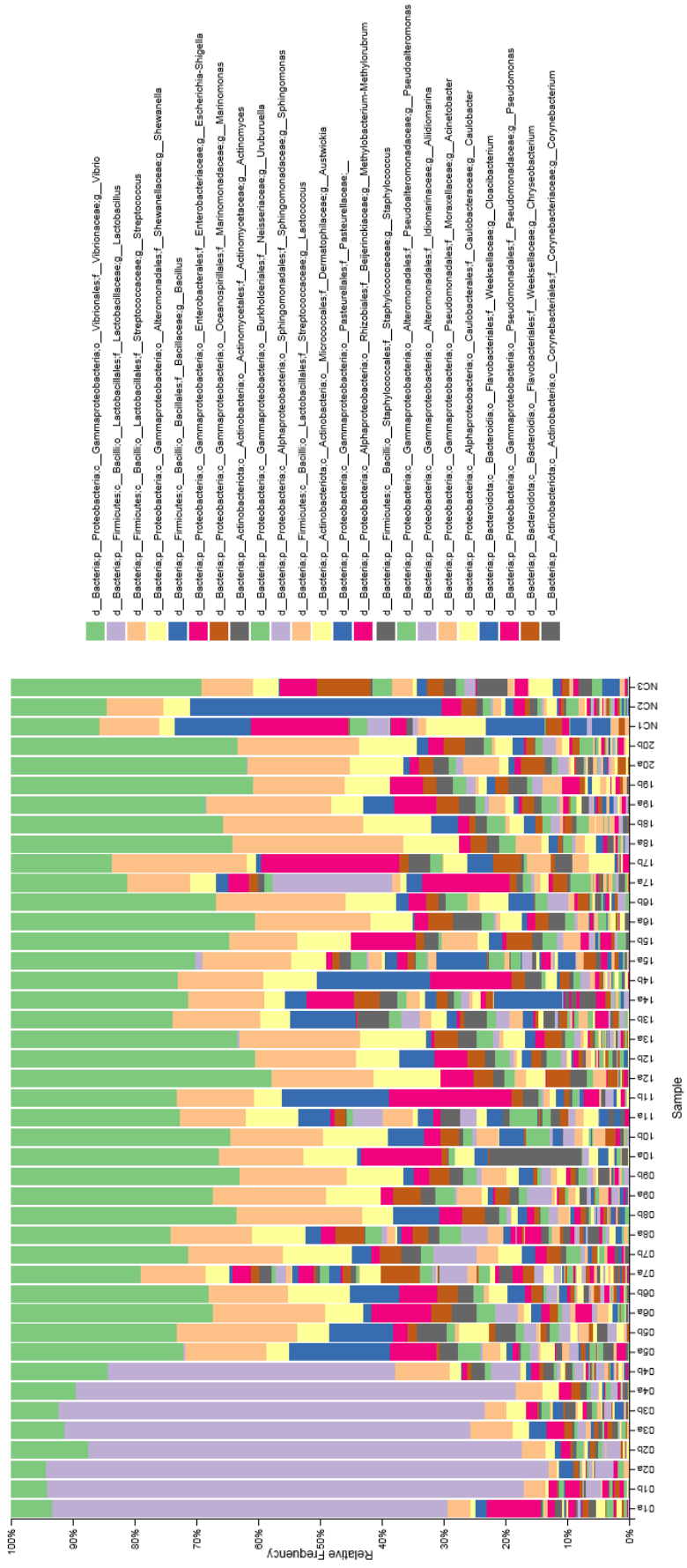
Ολοκληρώνοντας μια σειρά από έναν επίμονο ποιοτικό έλεγχο στα δεδομένα αλληλούχισης και συγκρίνοντας την απόδοση των ταξινομικών αναθέσεων των τριών διαφορετικών συνόλων δεδομένων, δίνεται πια η δυνατότητα της εις βάθος μελέτης της ταξινομικής σύνθεσης των βιολογικών δειγμάτων. Όμως, πριν την εξαγωγή συμπερασμάτων για το ταξινομικό προφίλ των βιολογικών δειγμάτων, είναι αναγκαίο ένα ακόμα στάδιο ποιοτικού ελέγχου των δεδομένων. Ο έλεγχος αυτός δεν αφορά την ίδια την τεχνολογία αλληλούχισης, όπως τον εντοπισμό και την αφαίρεση αναγνωσμάτων που παρέχουν χαμηλή βαθμολογία ποιότητας ή που να χαρακτηρίζονται χιμαιρικά, αλλά αφορά την εύρεση και διαχείριση εξωγενούς επιμόλυνσης στα βιολογικά δείγματα που ενδεχομένως προέκυψε από

σφάλματα λήψης ή συντήρησης αυτών. Είναι σημαντικό να αφαιρεθούν τυχόν επιμολύνσεις που μπορούν να έχουν προκύψει, διότι εκτός του προβλήματος της εξαγωγής εσφαλμένης ταξινομικής σύνθεσης των βιολογικών δειγμάτων, υπάρχει και κίνδυνος εσφαλμένης αυξημένης ποικιλομορφίας στις περαιτέρω στατιστικές αναλύσεις.

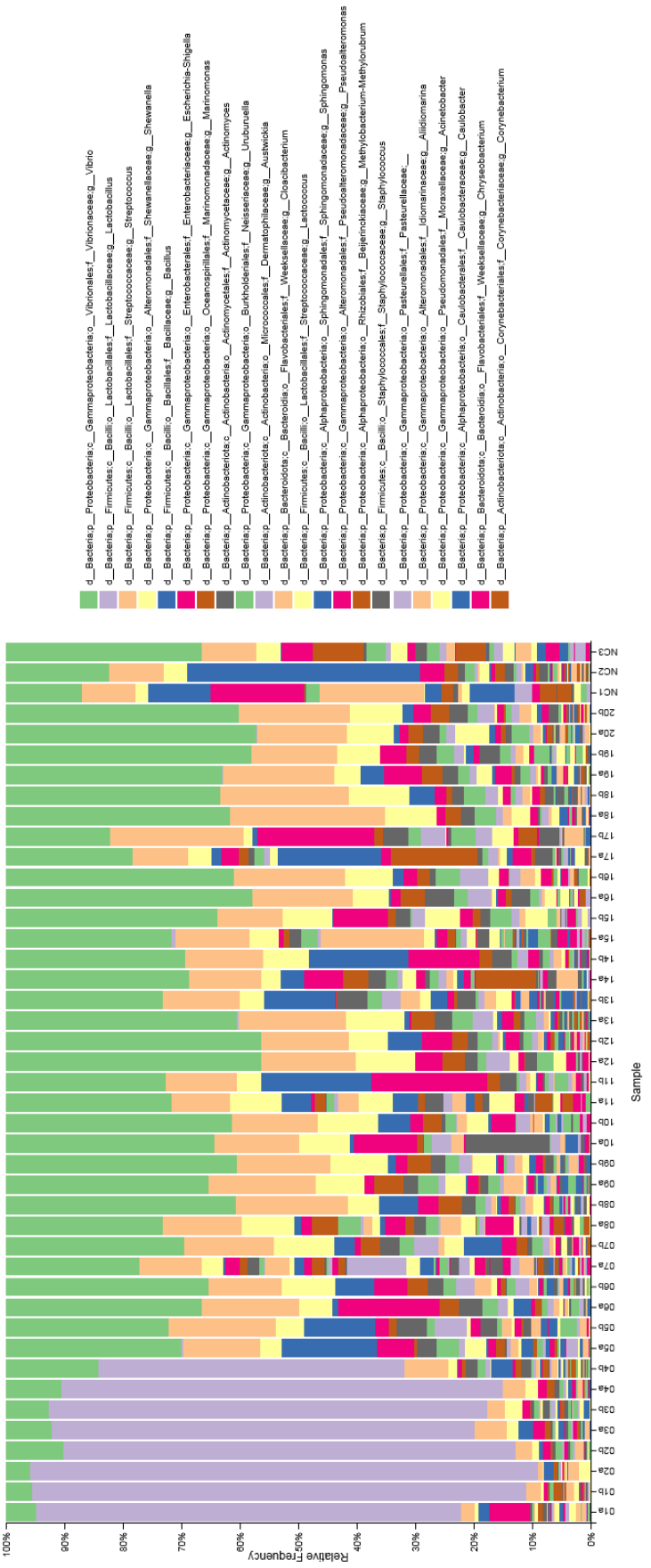
Λόγω του ότι τα δεδομένα αλληλούχισης προέρχονται κυρίως από βιολογικά δείγματα άγνωστης ταξινομικής σύνθεσης και η πλατφόρμα QIIME2 δεν παρέχει εργαλείο εντοπισμού επιμολύνσεων, ένα κριτήριο εντοπισμού ταξινομημένης αλληλουχίας που ενδεχομένως έχει προκληθεί λόγω επιμόλυνσης είναι η αφύσικα πολύ υψηλή σχετική αφθονία που παρουσιάζεται σε συγκεκριμένα δείγματα. Με μια γρήγορη επισκόπηση της βακτηριακής σύνθεσης των δειγμάτων σε επίπεδο φυλής, παρατηρείται και στα τρία σύνολα δεδομένων σημαντικά αυξημένη σχετική αφθονία σε *Firmicutes* σε 8 δείγματα αίματος σε σχέση με τα υπόλοιπα δείγματα, των οποίων η προέλευσή τους είναι από 4 ασθενείς (patient 1, 2, 3 και 4) στα δύο διαφορετικά χρονικά σημεία δειγματοληψίας (before and after treatment) (**Παράρτημα 46, Παράρτημα 47, Παράρτημα 48**). Αυξάνοντας την ταξινομική ευκρίνεια σε επίπεδο γένους, τα διαγράμματα ράβδων σχετικής αφθονίας της βακτηριακής ταξινομικής σύνθεσης των δειγμάτων παρουσιάζουν στα ίδια δείγματα αυξημένη σχετική αφθονία σε *Lactobacillus*, μία βακτηριακή ταξινομική κατηγορία που ανήκει στην φυλή των *Firmicutes* (**Σχήμα 3.13**, **Σχήμα 3.14**, **Σχήμα 3.15**)



Σχήμα 3.13 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επίπεδο γένους του συνόλου δεδομένων DADA2.



Σχήμα 3.14 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο γενους του συνόλου δεδομένων Deblur.



Σχήμα 3.15 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επίπεδο γένους του συνόλου δεδομένων VSEARCH.

Συγκεκριμένα, και στα τρία σύνολα δεδομένων που έχουν προκύψει από τις διαφορετικές επεξεργαστικές ροές εντοπίζεται σημαντικά αυξημένη σχετική αφθονία σε *Lactobacillus* στα προαναφερόμενα δείγματα, με το ποσοστό να κυμαίνεται από 46% έως και 87% ανάλογα το δείγμα και το σύνολο δεδομένων που εξετάζεται. Επίσης, σε όλο τον όγκο των δεδομένων, η συγκεκριμένη ταξινομική κατηγορία παρουσιάζεται επιπλέον σε πέντε δείγματα αίματος και ένα αντίστοιχο ελέγχου (NTC3), με το σύνολο δεδομένων που έχει προκύψει από την εφαρμογή του VSEARCH να παρουσιάζεται σε ένα ακόμα δείγμα ελέγχου (NTC2). Τα δείγματα αίματος αφορούν πέντε ασθενείς (patient 5, 13, 15, 19 και 20), των οποίων η δειγματοληψία πραγματοποιήθηκε πριν την αντιψυχωσική αγωγή (t1, before treatment). Όμως, στο καθένα από αυτά τα δείγματα, η συγκεκριμένη ταξινομική κατηγορία εμφανίζεται με σχετική αφθονία που δεν ξεπερνάει το 1%, με την εξαίρεση να αποτελεί στο δείγμα αίματος του ασθενή 15 του συνόλου δεδομένων του Deblur, το οποίο εμφανίζεται με 1,1%.

Η έντονα αυξημένη σχετική αφθονία σε *Lactobacillus* στα οκτώ πρώτα δείγματα, καθώς και η σημαντικά χαμηλή σχετική αφθονία αυτής της βακτηριακής ταξινόμησης στα υπόλοιπα δείγματα, υποδεικνύουν σε μεγάλο βαθμό ότι τα αναγνώσματα που κατέληξαν να αντιπροσωπεύουν την περιοχή V3-V4 του 16S rRNA γονιδίου των βακτηρίων γένους *Lactobacillus* στα συγκεκριμένα δεδομένα αλληλούχισης έχουν προκύψει λόγω επιμόλυνσης. Η βιβλιογραφική επισκόπηση σε ότι αφορά την σύνδεση της συγκεκριμένης βακτηριακής ταξινομικής κατηγορίας και σε προβλήματα επιμόλυνσης δειγμάτων ενισχύει ακόμα περισσότερο την υπόθεση ότι στα συγκεκριμένα δεδομένα τα βακτήρια *Lactobacillus* αποτελούν αποτέλεσμα επιμόλυνσης. Συγκεκριμένα, έχουν αναφερθεί περιπτώσεις δεδομένων αλληλούχισης και καλλιέργειας δειγμάτων ελέγχου, βιολογικών χαμηλού μικροβιακού φορτίου καθώς και συγκεκριμένα σε δείγματα αίματος, στις οποίες έχει εντοπιστεί και χαρακτηριστεί η κατηγορία των *Lactobacillus* ως αποτέλεσμα επιμόλυνσης στα δείγματα που ενδεχομένως πραγματοποιήθηκε κατά την διαχείριση και επεξεργασία αυτών (Karstens et al., 2019). Συνεπώς, η λογική πράξη για την διαχείριση αυτής της κατάστασης στα παρόντα δεδομένα είναι είτε η μη συμπερίληψη των δειγμάτων με την έντονη παρουσία ως προς την αφθονία των *Lactobacillus* σε περαιτέρω αναλυτικά στάδια, είτε η ολική αφαίρεση των αναγνωσμάτων που αντιπροσωπεύουν αυτή την ταξινομική κατηγορία από όλα τα σύνολα δεδομένων και στην συνέχεια η συμπερίληψη όλων των δειγμάτων στις μετέπειτα αναλύσεις. Εφόσον ο αριθμός των δειγμάτων των αρχικών δεδομένων που χρησιμοποιήθηκαν στην παρούσα Διπλωματική Εργασία είναι ήδη σημαντικά μικρός για την ανάλυση μικροβιώματος, δεν υφίσταται η επιλογή αφαίρεσης 8 δειγμάτων αίματος από τα 40 συνολικά. Δεδομένου ότι τα δείγματα στην παρούσα εργασία θεωρούνται πολύτιμα, εξετάζεται η παρουσία των αναγνωσμάτων που αντιπροσωπεύουν τα *Lactobacillus* από όλα τα σύνολα δεδομένων.

Πιο συγκεκριμένα, παρατηρείται ότι τα ASVs/OTUs που ταξινομήθηκαν σε αυτήν την βακτηριακή ταξινομική κατηγορία, τα οποία είναι 34 ASVs (3,9%) για τον DADA2, 8 ASVs (1,1%) για τον Deblur και 117 OTUs (8,6%) για τον VSEARCH, δεν έχουν ταξινομηθεί περεταίρω σε επίπεδο είδους. Επιπρόσθετα, τα οκτώ δείγματα που φέρουν πολύ μεγάλη σχετική αφθονία σε *Lactobacillus* στο περιεχόμενό τους έχουν προκύψει να παρέχουν μεγαλύτερο αριθμό αναγνωσμάτων σε σχέση με τα υπόλοιπα δείγματα (**Παράρτημα 37**). Ειδικά το δείγμα αίματος πριν την αντιψυχωτική αγωγή του ασθενή 1 (sample1a) παρουσιάζει σε όλα τα σύνολα δεδομένων και σε όλες τις ροές επεξεργασίας σε πολύ μεγαλύτερο βαθμό αυξημένο αριθμό αναγνωσμάτων. Δεδομένου αυτού, η αφαίρεση μεγάλου όγκου αναγνωσμάτων από αυτά τα δείγματα θα φέρει σαν αποτέλεσμα την εξομάλυνση της

διακύμανση του αριθμού αναγνωσμάτων ανά δείγμα, κατά πάσα πιθανότητα. Επίσης, η αφαίρεση αυτών των αναγνωσμάτων από τα δείγματα στα οποία παρουσιάζονται σε πολύ χαμηλή σχετική αφθονία όχι μόνο δεν θα επιφέρει σημαντική αλλαγή στο ταξινομικό προφίλ αυτών το δειγμάτων, αλλά πιθανόν θα βελτιώσει την ταξινομική εικόνα, διότι χαρακτηρίζονται ως αναγνώσματα χαμηλής σχετικής αφθονίας.

Λαμβάνοντας υπόψιν όλα τα προαναφερόμενα, αποφασίζεται να αφαιρεθούν και από τα τρία σύνολα δεδομένων τα ASVs και τα OTUs που ταξινομήθηκαν σε *Lactobacillus*, και κατά επέκταση τα αναγνώσματα που αντιπροσωπεύονται από αυτά προτού γίνει οποιαδήποτε εις βάθος ανάλυση ποικιλομορφίας των δειγμάτων. Τα αποτελέσματα που προέκυψαν μετά από αυτή την ενέργεια φανερώνουν παρόμοιες αλλαγές μεταξύ των τριών συνόλων δεδομένων, με την σημαντικότερη να αποτελεί η βελτίωση της διακύμανσης του αριθμού αναγνωσμάτων ανά δείγμα (**Σχήμα 3.16** (A)). Επιπλέον, η παραδοχή ότι η παρουσίαση των *Lactobacillus* αποτελεί αποτέλεσμα επιμόλυνσης σε έναν βαθμό επιβεβαιώνεται από την συνολική ταξινομική εικόνα των δειγμάτων μετά την αφαίρεσή τους από τα δεδομένα (**Σχήμα 3.18**, **Σχήμα 3.19**, **Σχήμα 3.20**).

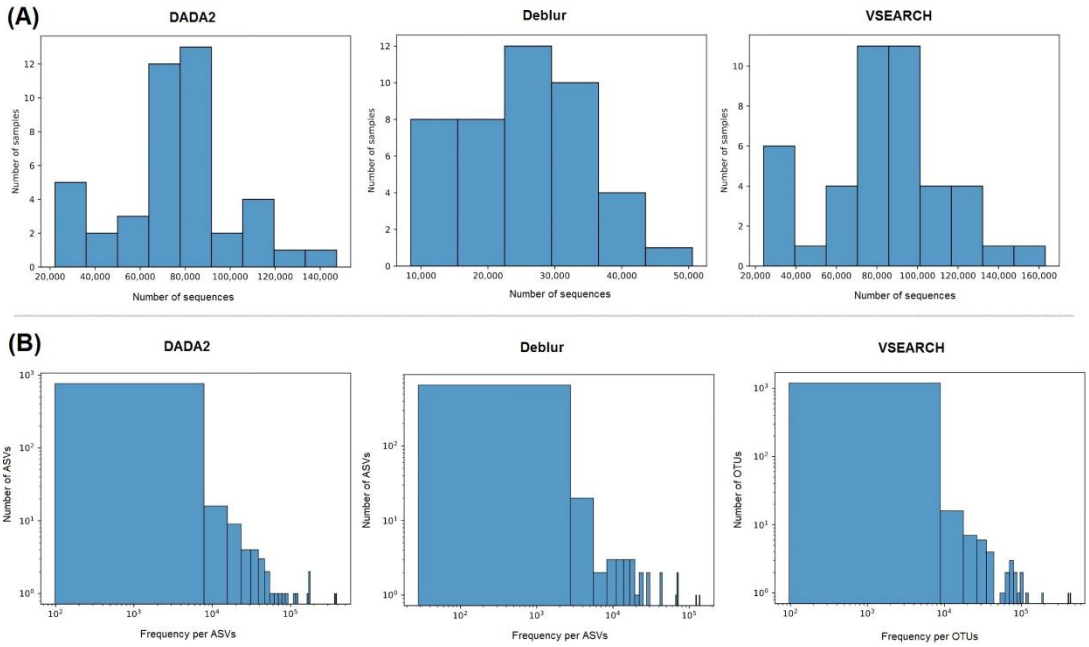
Πιο αναλυτικά, μετά την αφαίρεση των *Lactobacillus*, ο προσεγγιστικά ποσοστιαίος όγκος αναγνωσμάτων που απορρίφθηκε για τον DADA2 είναι 23,6%, για τον Deblur 19,3% και για τον VSEARCH 23,5% αντίστοιχα. Έτσι, ο ποσοστιαίος αριθμός τελικών διατηρητέων αναγνωσμάτων για το κάθε σύνολο δεδομένων σε σχέση με τον αρχικό αριθμό αναγνωσμάτων προκύπτει για τον DADA2 ίσο με 28,90%, για τον Deblur 10,02% και για τον VSEARCH 32,47% αντίστοιχα (**Πίνακας 3.25**, **Σχήμα 3.17**). Αναμενόμενη είναι η μείωση των συχνοτήτων που παρέχουν τα ASVs και OTUs, δεδομένου ότι αφαιρέθηκαν από τα δεδομένα τα αντίστοιχα που έφεραν μεγάλο όγκο αναγνωσμάτων.

Τα ιστογράμματα δειγμάτων που παρέχουν ένα δεδομένο αριθμό ταξινομημένων αναγνωσμάτων δείχνουν την μείωση της διακύμανσης του αριθμού αναγνωσμάτων ανά δείγμα (**(A)**). Μάλιστα, πριν την αφαίρεση των *Lactobacillus*, ο μεγαλύτερος αριθμός αναγνωσμάτων δείγματος ήταν περίπου δεκαπλάσιος από τον ελάχιστο αντίστοιχα και στα τρία σύνολα δεδομένων, ενώ μετά την αφαίρεση, η αναλογία μέγιστου και ελάχιστου αριθμού αναγνωσμάτων προέκυψε περίπου 6,5 στα αποτελέσματα. Βέβαια, είναι σημαντικό να σημειωθεί ότι στο σύνολο δεδομένων του Deblur κατέληξαν τρία δείγματα να φέρουν λιγότερο από 10000 αναγνώσματα, με το ελάχιστο να είναι ίσο με 8459 αναγνώσματα (**Παράρτημα 37**). Δεδομένου ότι τα δείγματα αυτά δεν απέχουν πολύ από το ελάχιστο επιτρεπόμενο αριθμό αναγνωσμάτων, θεωρείται αμελητέα αυτή η επίπτωση.

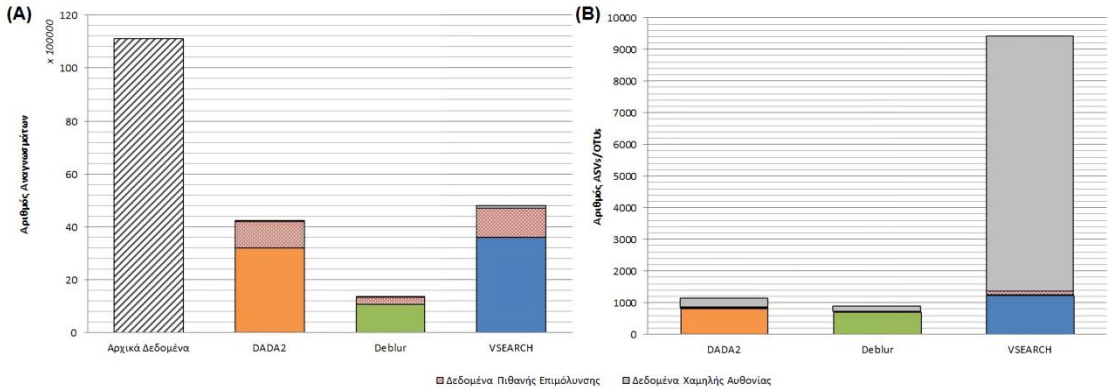
Τέλος, με μια γρήγορη επισκόπηση της τελικής βακτηριακής σύνθεσης των δειγμάτων σε επίπεδο φυλής μετά την αφαίρεση της επιμόλυνσης λόγω των *Lactobacillus*, παρατηρείται και στα τρία σύνολα δεδομένων μία ομοιομορφία στο περιεχόμενο καθώς και στις σχετικές αφθονίες στα δείγματα αίματος (**Σχήμα 3.18**, **Σχήμα 3.19**, **Σχήμα 3.20**). Τα δείγματα που προέρχονται από το ίδιο είδος περιβάλλοντος θα πρέπει να έχουν παρόμοια ταξινομικά προφίλ επειδή εκτίθενται σε παρόμοιους οικολογικούς και περιβαλλοντικούς παράγοντες. Συνεπώς, εφόσον η αφαίρεση των *Lactobacillus* από τα δεδομένα έφερε την ταξινομική ομοιομορφία ως προς την ανίχνευση και την ποσοτικοποίηση της σχετικής αφθονίας των δειγμάτων, επιβεβαιώνεται ότι η αρχική παρουσία των *Lactobacillus* με ακραία πολύ μεγάλη σχετική αφθονία στα 8 δείγματα αίματος οφείλετε πιθανότατα λόγω επιμόλυνσης.

Πίνακας 3.25 Γενική περιγραφή αριθμού ταξινομημένων αναγνωσμάτων των συνολικών δειγμάτων, ο αριθμός ταξινομημένων αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) και η γενική περιγραφή των συχνοτήτων τους που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση πιθανής επιμόλυνσης. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Τελικών Ταξινομημένων Αναγνωσμάτων		
pipeline	DADA2	Deblur	VSEARCH
Ελάχιστος	22157	8459	24059
Μέσος Όρος	74917	25987	84176
Μέγιστος	147271	50563	163046
Συνολικός	3221426 ***(28.90%)***	1117421 ***(10.02%)***	3619584 ***(32.47%)***
	Αριθμός Τελικών Ταξινομημένων ASVs/OTUs		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
	817	703	1239
	Συχνότητα Τελικών Ταξινομημένων ASVs/OTUs		
pipeline	**DADA2**	**Deblur**	**VSEARCH**
Ελάχιστος	98	28	96
Μέσος Όρος	3943	1590	2921
Μέγιστος	388062	137696	439181



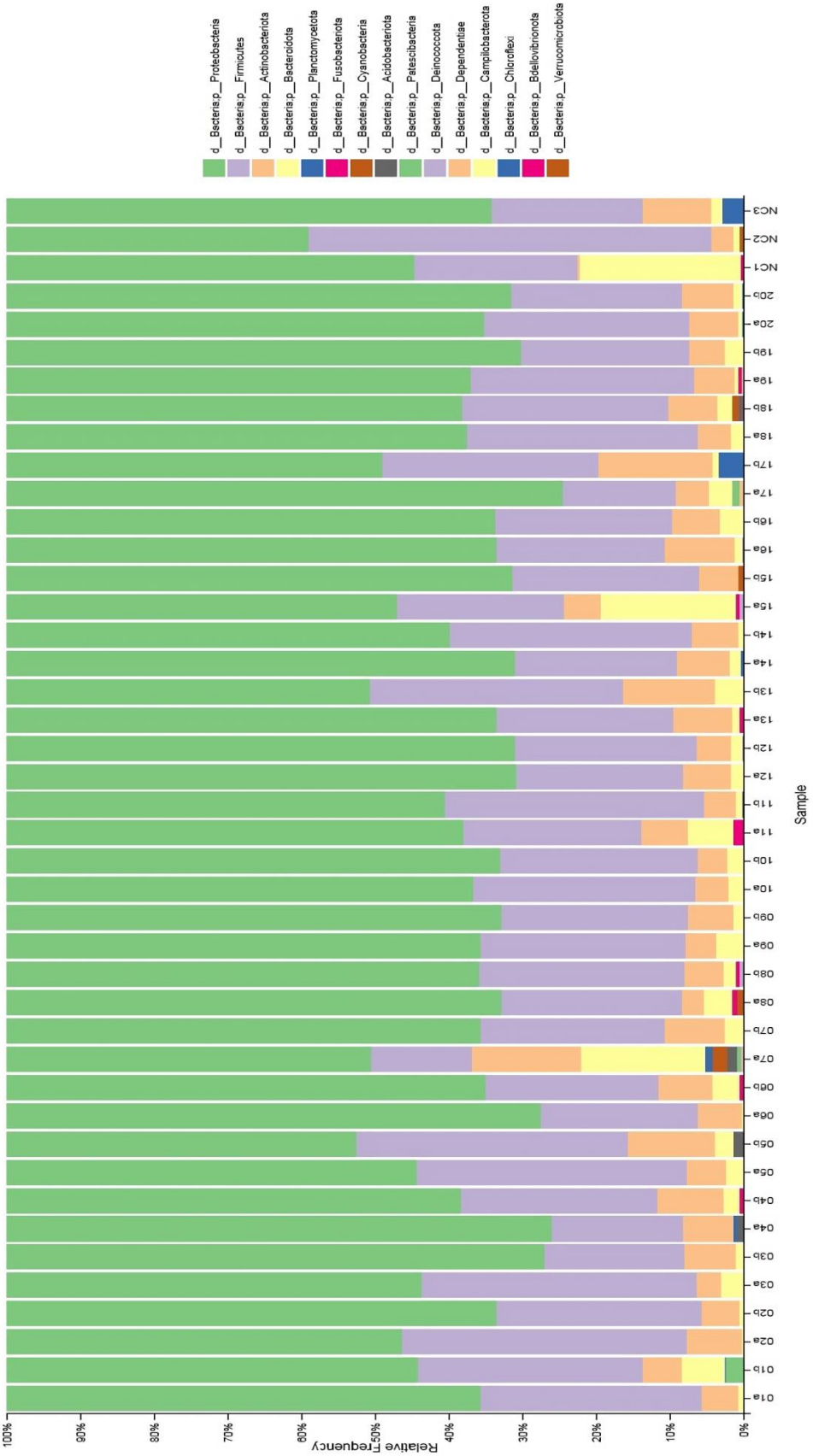
Σχήμα 3.16 (A) Ιστογράμματα δειγμάτων που παρέχουν ένα δεδομένο αριθμό ταξινομημένων αναγνωσμάτων και **(B)** ιστογράμματα αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που παρέχουν μία δεδομένη συχνότητα στο σύνολο των αποτελεσμάτων που προέκυψαν αντίστοιχα από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση δεδομένων πιθανής επιμόλυνσης.



Σχήμα 3.17 Διαγράμματα ράβδων που απεικονίζουν **(A)** τον αριθμό των αρχικών και την επίδραση των διατηρητέων αναγνωσμάτων καθώς και **(B)** την επίδραση του αριθμού των παραγόμενων αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά το φιλτράρισμα δεδομένων χαμηλής αφθονίας και πιθανής επιμόλυνσης.

Σχήμα 3.18 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο φυλης του συνόλου δεδομένων DADA2 μετά την αφαίρεση της πιθανής επιμόλυνσης.

Σχήμα 3.19 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο φυλης του συνόλου δεδομένων Deblur μετά την αφαίρεση της πιθανής επιμόλυνσης

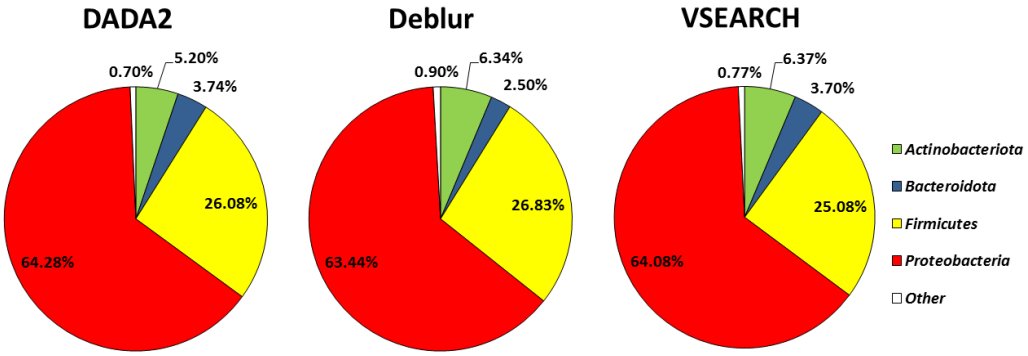


Σχήμα 3.20 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο φυλης του συνόλου δεδομένων VSEARCH μετά την αφαίρεση της πιθανής επιμόλυνσης

3.12 Ταξινομικό Προφίλ Ανθρώπινου Αίματος στη Σχιζοφρένεια

Σε αυτό το στάδιο ανάλυσης, ο ποιοτικός έλεγχος των δεδομένων θεωρείται ολοκληρωμένος και κατ' επέκταση δίνεται η δυνατότητα εξαγωγής συμπερασμάτων όσον αφορά στην βιολογία των δειγμάτων της παρούσας εργασίας. Ένας τρόπος με τον οποίο επιτυγχάνεται η εξαγωγή συμπερασμάτων είναι η εξέταση του ταξινομικού προφίλ των δειγμάτων, που στην συγκεκριμένη περίπτωση αποτελούν τα δείγματα αίματος ανθρώπων που παρουσίασαν το πρώτο ψυχωσικό επεισόδιο και διαγνώστηκαν από σχιζοφρένεια. Μεταξύ των τριών συνόλων δεδομένων δεν παρουσιάζονται σημαντικές διαφορές ως προς την ανίχνευση και την ποσοτικοποίηση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επίπεδο φυλής (**Σχήμα 3.21**).

Συγκεκριμένα, όπως και στα δείγματα αίματος από υγιής ανθρώπους, έτσι και στα αντίστοιχα από ανθρώπους που πάσχουν από σχιζοφρένεια φαίνεται να κυριαρχούνται σε επίπεδο φύλων από *Proteobacteria*, όπου οι σχετικές αφθονίες προέκυψαν για τον DADA2 να είναι ίση με 64,28% ±1,14%, για τον Deblur ίση με 63,24% ±4,97% και για τον VSEARCH ίση με 64,08% ±1,20%. Στην συνέχεια, τα δείγματα κυριαρχούνται από *Firmicutes* (DADA2: 26,08% ±0,49%, Deblur: 26,83% ±2,04% και VSEARCH: 25,08% ±0,45%) και ακολουθούν τα *Actinobacteria* (DADA2: 5,20% ±0,16%, Deblur: 6,64% ±0,79% και VSEARCH: 6,37% ±0,20%) και τα *Bacteroidetes* (DADA2: 3,74% ±0,30%, Deblur: 2,50% ±0,64% και VSEARCH: 3,70% ±0,29%).



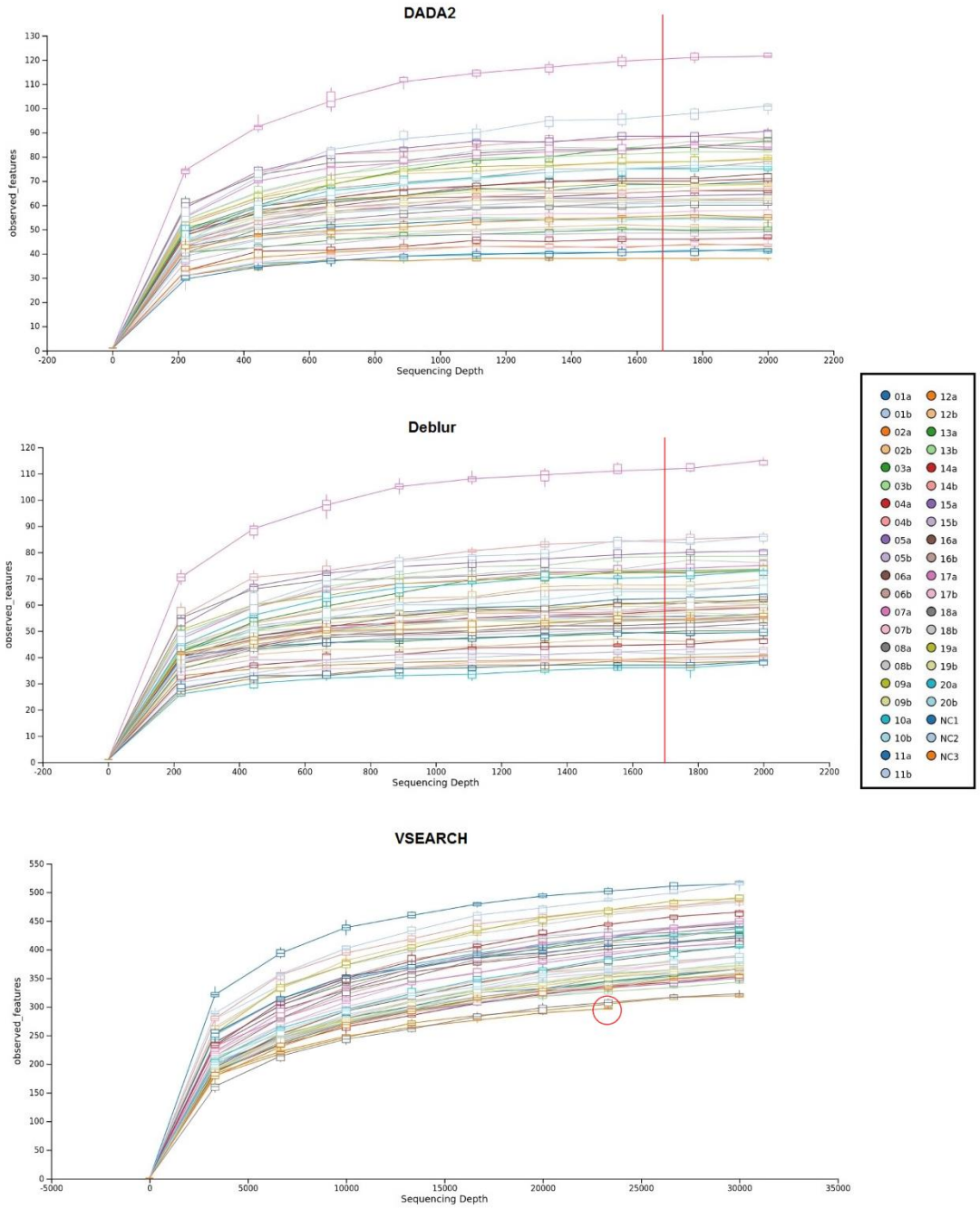
Σχήμα 3.21 Διαγράμματα πίτας που αντιπροσωπεύουν τις μέσες σχετικές αναλογίες των βακτηριακών φύλων στο αίμα σχιζοφρενών οι οποίοι παρουσίασαν το πρώτο ψυχωτικό επεισόδιο (t_1) που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH .

3.13 Καμπύλες Αραίωσης

Από την φυλογενετικές αναλύσεις και τους προσδιορισμούς των δεικτών Shannon των δεδομένων, κατασκευάστηκαν για το κάθε σύνολο δεδομένων καμπύλες αραίωσης με σκοπό την αποτίμηση βάθους ανάλυσης των δειγμάτων και κατ' επέκταση την περαιτέρω στατιστική ανάλυση της άλφα ποικιλομορφίας τους. Είναι πολύ σημαντικό να επιλεχθεί ένα βάθος δειγματοληψίας αναγνωσμάτων για κάθε σύνολο δεδομένων, το οποίο να καλύπτει όσο τον δυνατόν περισσότερο την ποικιλομορφία των βιολογικών δειγμάτων, να μπορούν να συμπεριληφθούν σε περαιτέρω αναλύσεις όλα τα δείγματα και ταυτόχρονα το κάθε δείγμα να παρέχει αριθμό αναγνωσμάτων τουλάχιστον ίσο με 10000.

Από τα γραφήματα (**Σχήμα 3.22**), παρατηρείται ότι στον DADA2 και στον Deblur οι καμπύλες φτάνουν σε ξεκάθαρο πλατό στο διάστημα βάθους αλληλούχισης 1600-1800 αναγνώσματα. Αυτό υποδηλώνει ότι η συλλογή πρόσθετων αναγνωσμάτων πέρα από αυτό το βάθος δειγματοληψίας δεν είναι πιθανό να οδηγήσουν στην παρατήρηση πρόσθετων πληροφοριών σε σχέση με το μικροβιακό περιεχόμενο. Όμως, στην περίπτωση του VSEARCH, οι καμπύλες δεν φαίνεται να πιάνουν πλατό. Μάλιστα, παρατηρείται να έχουν μικρή αλλά συνεχή άνοδο ακόμα και μετά την τιμή βάθους αλληλούχισης 25000, ξεπερνώντας έτσι το δείγμα με τον ελάχιστο αριθμό αναγνωσμάτων..

Λαμβάνοντας υπόψη τις προϋποθέσεις επιλογής βάθους αλληλούχισης, επιλέγεται και στα τρία σύνολα δεδομένων ο αριθμός αναγνωσμάτων που παρέχει το δείγμα με τον μικρότερο αριθμό αναγνωσμάτων. Συνεπώς, για τον DADA2 το βάθος είναι ίσο με d_{DADA2}=22157, για τον Deblur είναι d_{Deblur}=8459 και για τον VSEARCH είναι $d_{VSEARCH}$=24059 αντίστοιχα.



Σχήμα 3.22 Καμπύλες αραίωσης που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH που απεικονίζουν τον αριθμό των ταξινομικών ειδών που παρέχει το κάθε δείγμα έναντι της τιμής του βάθους αλληλούχισης. Παρατηρείται ότι στον DADA2 και στον Deblur οι καμπύλες φτάνουν σε ξεκάθαρο πλατό στο διάστημα βάθους αλληλούχισης 1600-1800 (κόκκινη γραμμή), ενώ στον VSEARCH παρουσιάζεται να έχουν μικρή αλλά συνεχή άνοδο ακόμα και μετά την τιμή βάθους αλληλούχισης 25000 ξεπερνώντας έτσι το δείγμα με τον ελάχιστο αριθμό αναγνωσμάτων (κόκκινος κύκλος).

3.14 Άλφα Ποικιλομορφία

Επιλέγοντας τον όγκο πληροφορίας που παρέχεται από τα προαναφερόμενα βάθη αλληλούχισης για κάθε σύνολο δεδομένων, παρέχεται η δυνατότητα διερεύνησης της μικροβιακής ποικιλομορφίας των δειγμάτων αίματος ασθενών με σχιζοφρένεια που παρουσίασαν το πρώτο ψυχωσικό επεισόδιο, των δειγμάτων αίματος των ίδιων ασθενών μετά από ένα μήνα αντιψυχωσικής θεραπείας, την μεταξύ τους σύγκριση καθώς και την σύγκριση

αυτών με τα δείγματα ελέγχου. Σκοπός αυτών των διερευνήσεων είναι αρχικά να μελετηθεί η διαφορά ως προς την ποικιλομορφία των δειγμάτων αίματος και των δειγμάτων ελέγχου και στην συνέχεια να μελετηθούν τυχόν αλλαγές στην ποικιλομορφία των δειγμάτων αίματος στα δύο διαφορετικά χρονικά σημεία. Αρχικά, για τον σκοπό της σύγκρισης της ποικιλομορφίας των τριών βασικών ομάδων δειγμάτων, κατασκευάζονται θηκογράμματα για την αναπαράσταση και σύγκριση της άλφα ποικιλομορφίας χρησιμοποιώντας τους δείκτες εντροπίας του Shannon που προέκυψαν και από τα τρία σύνολα δεδομένων (**Παράρτημα 49**). Οι τρείς βασικές ομάδες δειγμάτων είναι τα δείγματα αίματος πριν την αντιψυχωσική θεραπεία (before_treatment), τα δείγματα αίματος μετά από έναν μήνα αντιψυχωσικής θεραπείας (after_treatment) και τα δείγματα αρνητικού ελέγχου (controls).

Η πρώτη επισκόπηση των γραφημάτων φανερώνει ενδιαφέρον μοτίβα στην ποικιλομορφία των ομάδων δειγμάτων μεταξύ των τριών διαφορετικών συνόλων δεδομένων (**Σχήμα 3.23**). Ένα βασικό χαρακτηριστικό διαφοροποίησης μεταξύ αυτών των συνόλων είναι ότι τα αποτελέσματα του VSEARCH έχουν προκύψει συλλογικά να έχουν υψηλότερες τιμές εντροπίας Shannon σε σχέση με τα αποτελέσματα των μεθόδων αποθορυβοποίησης. Όμως, αυτή η παρατήρηση είναι αναμενόμενη, διότι οι προαναφερόμενες καμπύλες αραίωσης υποδεικνύουν το ίδιο φαινόμενο, δηλαδή της γενικής αυξημένης ποικιλομορφίας σε όλα τα δείγματα. Παρατηρώντας τις ποικιλομορφίες που παρέχουν οι βασικές ομάδες δειγμάτων στα διαφορετικά σύνολα δεδομένων, εντοπίζονται πολύ βασικές ομοιότητες και διαφορές.

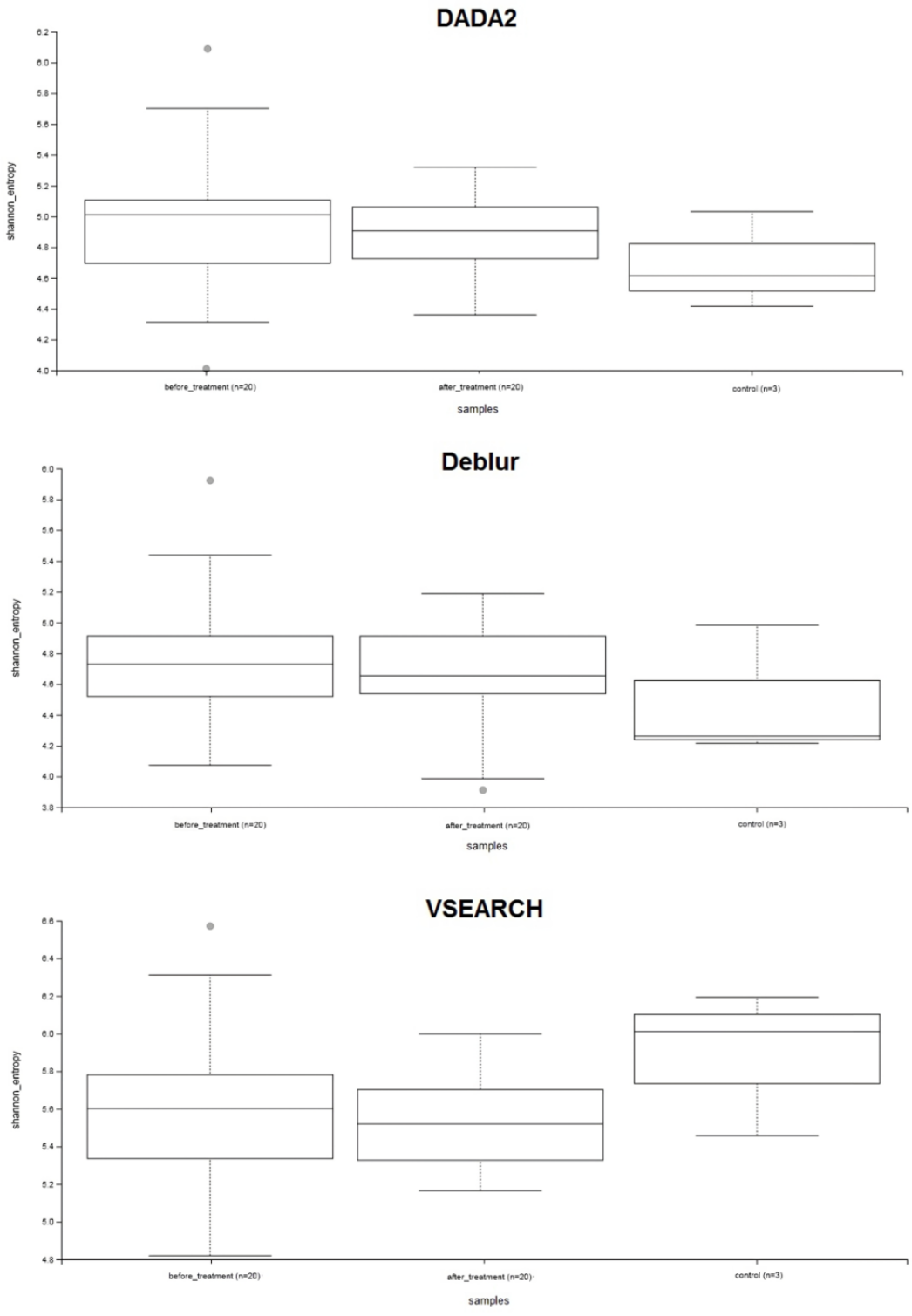
Πιο αναλυτικά, και στα τρία γραφήματα παρουσιάζεται ελαφρώς υψηλότερη διάμεση τιμή ποικιλομορφίας Shannon στην ομάδα δειγμάτων αίματος πριν την θεραπεία σε σχέση με την αντίστοιχη τιμή της ομάδας δειγμάτων αίματος μετά από ένα μήνα αντιψυχωσικής αγωγής. Αυτή η παρατήρηση υποδηλώνει μεγαλύτερο πλούτο και ομοιομορφία του μικροβιακού πληθυσμού στα δείγματα αίματος ασθενών με σχιζοφρένεια πριν την χορήγηση αντιψυχωσικών φαρμάκων. Βέβαια, είναι σημαντικό να σημειωθεί ότι και στα τρία σύνολα δεδομένων τα δείγματα αίματος «προ-θεραπείας» παρουσιάζουν μεγαλύτερη διακύμανση μεταξύ τους ως προς την ποικιλομορφία τους σε σχέση με τα δείγματα αίματος μετά την θεραπεία (**Σχήμα 3.23**).

Σε αντίθεση με τα δείγματα αίματος, τα δείγματα ελέγχου παρουσιάζουν διαφορετική ποικιλομορφία μεταξύ των μεθόδων αποθορυβοποίησης και ομαδοποίησης (**Σχήμα 3.23**). Μάλιστα, τα σύνολα δεδομένων του DADA2 και του Deblur τα δείγματα ελέγχου εμφανίζουν χαμηλότερη διάμεση τιμή εντροπίας Shannon από τις δύο διαφορετικές ομάδες δειγμάτων αίματος, ενώ στην περίπτωση του συνόλου δεδομένων του VSEARCH, τα δείγματα ελέγχου παρουσίασαν μεγαλύτερη ποικιλομορφία σε σχέση με τα δείγματα αίματος. Ένας λόγος για τον οποίο μπορεί να οφείλεται η αυξημένη ποικιλομορφία των δειγμάτων ελέγχου στα αποτελέσματα ομαδοποίησης σε OTUs είναι η τιμή του βάθος αλληλούχισης που έχει επιλεχθεί για τον προσδιορισμό της ποικιλομορφίας των δειγμάτων. Εφόσον οι καμπύλες αραίωσης δείχνουν ότι το προεπιλεγμένο βάθος αλληλούχισης δεν έχει καλύψει πλήρως την ποικιλομορφία των δειγμάτων, ενδέχεται να μην έχει καλυφθεί η ποικιλομορφία των δειγμάτων αίματος επαρκώς σε αντίθεση με τα δείγματα ελέγχου. Όμως, είναι σημαντικό να αναφερθεί ότι η επιλογή του συγκεκριμένου βάθους αλληλούχισης πραγματοποιήθηκε με γνώμονα την συμπερίληψη όλων των δειγμάτων που παρέχονται από τα πρωτογενή δεδομένα στην ανάλυση ποικιλομορφίας τους. Για την αξιολόγηση αυτών των παρατηρήσεων στις βασικές ομάδες δειγμάτων, σε κάθε σύνολο δεδομένων ξεχωριστά πραγματοποιήθηκε στατιστικός έλεγχος Kruskal-Wallis για την σύγκριση των τιμών άλφα ποικιλομορφίας

μεταξύ των δύο ομάδων δειγμάτων αίματος (before_treatment vs after_treatment), μεταξύ των ομάδων δειγμάτων αίματος προ-θεραπείας και ελέγχου (before_treatment vs controls) και τέλος μεταξύ των ομάδων δειγμάτων αίματος μετά την θεραπεία και ελέγχου (after_treatment vs controls). Και στα τρία σύνολα δεδομένων που προέκυψαν από τις τρεις διαφορετικές ροές επεξεργασίας δεν παρατηρήθηκε στατιστικά σημαντική διαφορά στην άλφα ποικιλομορφία μεταξύ των προαναφερόμενων ομάδων δειγμάτων ($p>0.05$) (**Πίνακας 3.26**). Ωστόσο, είναι σημαντικό να αναγνωριστεί ότι η αποτυχία εντοπισμού στατιστικά σημαντικών διαφορών μεταξύ των ομάδων δειγμάτων αίματος και ελέγχου μπορεί να έχει προκύψει από το μικρό μέγεθος δειγμάτων ελέγχου.

Πίνακας 3.26 Αποτελέσματα στατιστικού ελέγχου Kruskal-Wallis για την σύγκριση της άλφα ποικιλομορφίας (εντροπία Shannon) μεταξύ διαφορετικών ομάδων δειγμάτων που προέκυψε από τα σύνολα δεδομένων DADA2, Deblur και VSEARCH.

	DADA2	**Deblur**	**VSEARCH**
before_treatment (n=20) vs after_treatment (n=20)	p=0.913837	p=0.626328	p=0.766046
before_treatment (n=20) vs controls (n=3)	p=0.465209	p=0.411314	p=0.235333
after_treatment (n=20) vs controls (n=3)	p=0.235333	p=0.36131	p=0.082837



Σχήμα 3.23 Θηκογράμματα άλφα ποικιλομορφίας χρησιμοποιώντας δείκτες εντροπίας Shannon που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH, για τις τρεις βασικές ομάδες δειγμάτων, τα δείγματα αίματος πριν την αντιψυχωσική θεραπεία (before_treatment), τα δείγματα αίματος μέτα από έναν μήνα αντιψυχωσικής θεραπείας (after_treatment) και τα δείγματα ελέγχου (controls). Ο στατιστικός έλεγχος Kruskal-Wallis μεταξύ των διαφορετικών ομάδων δειγμάτων δεν παρουσίασε σημαντική στατιστική διαφορά ως προς την άλφα ποικιλομορφία τους και στα τρία σύνολα δεδομένων.

4 Συζήτηση Αποτελεσμάτων

Η παρούσα διπλωματική εργασία αφορά την διερεύνηση του μικροβιώματος του αίματος και την συσχέτιση του ρόλου αυτού με την σχιζοφρένεια αξιοποιώντας και αναλύοντας δεδομένα αλληλούχισης του 16S rRNA γονιδίου στη διαδεδομένη πλατφόρμα QIIME2. Η ανάλυση επιτυχάνθηκε χρησιμοποιώντας δύο διαφορετικά εργαλεία αποθορυβοποίησης αναγνωσμάτων, τον DADA2 και τον Deblur, και ένα εργαλείο ομαδοποίησης αντίστοιχα, τον VSEARCH, που είναι διαθέσιμα για εφαρμογή στο QIIME2, οδηγώντας στην κατασκευή τριών ξεχωριστών επεξεργαστικών ροών, με σκοπό την σύγκριση των αποτελεσμάτων τους όσο αφορά στην βιολογική ερμηνεία των δεδομένων ανάλογα την μέθοδο παρασκευής ASVs ή OTUs. Η κάθε επεξεργαστική ροή περιλαμβάνει τα απαραίτητα στάδια διαχείρισης και επεξεργασίας δεδομένων αλληλούχισης του 16S rRNA γονιδίου στη σειρά που το καθένα απαιτεί να εφαρμοστεί ανάλογα τα εργαλεία που χρησιμοποιούνται, συμπεριλαμβανομένου της αποκοπής μη-βιολογικών εκκινητών από τα paired-end αναγνώσματα, της συγχώνευσης αυτών, το ποιοτικό φιλτράρισμα αυτών, την κατασκευή αντιπροσωπευτικών αμπλικονίων (ASVs/OTUs), την αφαίρεση χιμαιρικών αλληλουχιών, την ταξινόμηση των ASVs/OTUs, τον καθαρισμό των δεδομένων από μη-στοχευμένα στοιχεία, χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών και επιμολύνσεων, την αποτίμηση του βάθους αλληλούχισης και τέλος τον προσδιορισμό των δεικτών άλφα ποικιλομορφίας.

Για την αντικειμενική σύγκριση των επεξεργαστικών ροών, αρχικά εξετάστηκε η επιρροή βασικών παραμέτρων που απαιτούνται να οριστούν από τον χρήστη στο στάδιο της προεπεξεργασίας των τριών ροών επεξεργασίας στα δεδομένα αλληλούχισης έως και το στάδιο της ταξινομικής ανάθεσης των παραγόμενων ASVs/OTUs και επιχειρήθηκε η βελτιστοποίηση των τιμών τους λαμβάνοντας υπόψιν την βέλτιστη αποκόμιση όγκου δεδομένων και βιολογικής πληροφορίας, το επίπεδο αξιοπιστίας των ίδιων των τιμών, καθώς και το γεγονός ότι οι τελικώς επιλεγμένες τιμές και των τριών επεξεργαστικών ροών να είναι όσο τον δυνατόν γίνεται παρόμοιοι και συγκρίσιμοι. Οι βασικοί παράμετροι που εξετάστηκαν αφορούν τις προδιαγραφές επικαλυπτόμενης περιοχής για την συγχώνευση των paired-end αναγνωσμάτων, τον προκαθορισμό της ποιότητας που απαιτούνται να παρέχουν τα αναγνώσματα, είτε της μορφής paired-end είτε των συγχωνευμένων, και το μήκος που πρέπει έχουν τα paired-end ή συγχωνευμένα αναγνώσματα, όπου αυτό απαιτείται και εφαρμόζεται. Στην συνέχεια, τα επιλεγμένα σύνολα δεδομένων ελέγχονται για τυχόν σφάλματα και επιμολύνσεις και εξετάζονται εις βάθος οι ταξινομικές πληροφορίες που παρέχουν. Τέλος, πραγματοποιείται ανάλυση της ποικιλομορφίας στο κάθε σύνολο δεδομένων, κατασκευάζοντας καμπύλες αραίωσης, επιλέγοντας το βάθος αλληλούχισης για το κάθε σύνολο δεδομένων ξεχωριστά και προσδιορίζοντας αντίστοιχα τους δείκτες άλφα ποικιλομορφίας Shannon. Αυτό το στάδιο ολοκληρώνεται με τον στατιστικό έλεγχο μεταξύ των δεικτών ποικιλομορφίας που προέκυψαν για τα δείγματα αίματος που συλλέχθηκαν από τους ασθενείς όταν εμφάνισαν το πρώτο ψυχωσικό επεισόδιο, τα δείγματα αίματος των ίδιων ασθενών μετά από ένα μήνα αντιψυχωσικής θεραπείας και τα δείγματα αρνητικού ελέγχου.

Η συγκριτική ανάλυση των ροών επεξεργασίας που βασίζονται στην παραγωγή ASVs και OTUs αποκάλυψε μια σειρά διαφορών αλλά και ομοιοτήτων στις ταξινομικές πληροφορίες, στις καμπύλες αραίωσης καθώς και στις μετρήσεις άλφα ποικιλομορφίας, οριοθετώντας έτσι τις κοινές και μοναδικές αποτυπώσεις αυτών των βιοπληροφορικών προσεγγίσεων, που μπορεί να έχουν βαθιές συνέπειες για στην κατανόηση της συμμετοχής του μικροβιώματος του αίματος στη σχιζοφρένεια και παραπέμπουν σε μια επανεξέταση των

πρακτικών τυποποίησης βιοπληροφορικής ανάλυσης στο πεδίο. Στον λαβύρινθο αποτελεσμάτων που προέκυψαν από τη σύγκριση βιοπληροφορικών επεξεργαστικών ροών βασισμένων στην αποθορυβοποίηση σε ASVs και ομαδοποίηση σε OTUs των αναγνωσμάτων, αποκαλύπτονται κρίσιμες γνώσεις για το βακτηριακό φορτίο που βρίσκεται στην κυκλοφορία του αίματος ατόμων με σχιζοφρένεια, γεφυρώνοντας το χάσμα τις περιορισμένης έρευνας με αυτή την θεματολογία. Στο συγκεκριμένο κεφάλαιο, αποκρυπτογραφόνται και συγκρίνονται τα ευρήματα της παρούσας διπλωματικής εργασίας με την υπάρχουσα βιβλιογραφία, εμβαθύνοντας στο δίλημμα μεταξύ των μεθόδων ομαδοποίησης και αποθορυβοποίησης, ρίχνοντας φως στην περιπλοκότητα των μεθόδων που διαμορφώνουν την ερμηνεία των δεδομένων αλληλούχισης αμπλικονίων και αναγνωρίζοντας τους περιορισμούς της ανάλυσης.

4.1 Πειραματισμός Τιμών Παραμέτρων και Συγκριτική Ανάλυση

Η ποιότητα των πρωτογενών δεδομένων αλληλούχισης καθορίζουν τον τρόπο επεξεργασίας αυτών καθώς έχουν την δύναμη να επηρεάσουν τα τελικά αποτελέσματα ως προς την βιολογική ερμηνεία των δειγμάτων αίματος. Στην παρούσα διπλωματική εργασία, τα παραγόμενα δεδομένα από την πλατφόρμα Illumina MiSeqPE250 εισήχθηκαν εύκολα στο περιβάλλον της πλατφόρμας QIIME2 και εκ πρώτης όψεως παρουσιάζουν ικανοποιητικά χαρακτηριστικά. Ο όγκος και η ποιότητα των αναγνωσμάτων που παράχθηκαν από την πλατφόρμα αλληλούχισης φανερώνει την πολύ μεγάλη απόδοση που παρέχει η τεχνολογία της Illumina, με συνολικό αριθμό αναγνωσμάτων στο ύψος των $11 \cdot 10^6$ και με βαθμολογίες ποιότητας PHRED ανά βάση από 10 και πάνω (**Πίνακας 3.1**, **Σχήμα 3.2**). Αυτό σημαίνει ότι οι βάσεις αναγνωσμάτων στα διαθέσιμα δεδομένα αλληλούχισης συνοδεύονται από μέγιστο 0,1% ποσοστό σφάλματος υποκατάστασης. Αυτό το γεγονός χαρακτηρίζεται αναμενόμενο για τα δεδομένα που προέρχονται από τις πλατφόρμες Illumina, οι οποίες είναι γνωστές για το πλεονέκτημα να παρέχουν χαμηλό ποσοστό σφαλμάτων το οποίο τις καθιστά στις πιο αξιόπιστες τεχνολογίες αλληλούχισης ανά βάση στην αγορά (Hu et al., 2021). Η πλειοψηφία των paired-end αναγνωσμάτων παρέχουν μήκος 251 bp (**Πίνακας 3.2**), που είναι ισοδύναμο με το μέγιστο μήκος που είναι ικανή να αλληλουχίσει η πλατφόρμα, η οποία παράγει paired-end 2 x 250 bp αναγνώσματα και η προθήκη μίας τελικής βάσης στα μικρο-αναγνώσματα οφείλετε στην τεχνολογίας της (Slatko et al., 2018).

Ο μόνος προβληματισμός που προκύπτει από τα πρωτογενή δεδομένα αλληλούχισης έγκειται στην έντονη διακύμανση του αριθμού αναγνωσμάτων που παρέχουν τα δείγματα, και ειδικά για το δείγμα αίματος του ασθενή 1 (sample1a) που φέρει το μέγιστο αριθμό αναγνωσμάτων και η τιμή του είναι πάνω από το διπλάσιο αριθμό αναγνωσμάτων που παρέχει το αμέσως επόμενο δείγμα (**Σχήμα 3.1**). Όπως είναι ήδη γνωστό, είναι είναι δύσκολο να συλλεχθεί ακριβώς ο ίδιος αριθμός αναγνωσμάτων σε κάθε δείγμα. Αυτό το φαινόμενο μπορεί να οφείλεται είτε στους παράγοντες που σχετίζονται με την προετοιμασία των δειγμάτων, είτε στην πλατφόρμα αλληλούχισης που χρησιμοποιείται ή λόγω των τεχνικών δυσκολιών στη φόρτωση των ίδιων μοριακών ποσοτήτων των βιβλιοθηκών αλληλούχισης στο όργανο ή ακόμα και της εγγενής μεταβλητότητας στα βιολογικά δείγματα (Gloor et al., 2017). Παρόλα αυτά, αυτή η έντονη διακύμανση αποτελεί μια κατάσταση που πρέπει να αντιμετωπιστεί μέσω της βιοπληροφορικής επεξεργασίας των δεδομένων, έτσι ώστε οι σχετικές αφθονίες των ταξινομικών κατηγοριών και οι ποικιλομορφίες των δειγμάτων να είναι όσο τον δυνατόν αντικειμενικές. Ενώ η κανονικοποίηση των δεδομένων μέσω της

αραίωσης των δειγμάτων εξυπηρετεί σε μεγάλο βαθμό στην αντιμετώπιση αυτής της κατάστασης, στην παρούσα διπλωματική εργασία πραγματοποιήθηκε η σχετική εξομάλυνση του αριθμού αναγνωσμάτων που φέρουν τα δείγματα και στις τρεις ροές επεξεργασίας μέσω της αφαίρεσης των αναγνωσμάτων που ταξινομήθηκαν σε *Lactobacillus*, τα οποία χαρακτηρίστηκαν ως αποτέλεσμα επιμόλυνσης (**Σχήμα 3.16**).

Επιστρέφοντας στο θέμα της ποιότητας των αναγνωσμάτων, ενώ η πλατφόρμα αδιαμφισβήτητα κατάφερε να δώσει πολύ καλής ποιότητας πρωτογενή δεδομένα αλληλούχισης, το ποσοστό 0.1% πιθανότητας να έχει προκύψει μια βάση λόγω σφάλματος είναι αρκετά σημαντικό και επιβλαβές στην έκταση των $11 \bullet 10^6$ αναγνωσμάτων καθώς μπορεί να προσθέσει αναξιόπιστες και δυνητικά τυχαίες αλληλουχίες στο σύνολο δεδομένων και να οδηγήσει στην εσφαλμένη ερμηνεία αυτών (Del Fabbro et al., 2013). Η ποιότητα των paired-end αναγνωσμάτων φαίνεται να είναι μειωμένη περισσότερο στις περιοχές των άκρων αυτών και κυρίως στο τελικό δεξί άκρο (**Σχήμα 3.2**). Γενικά, είναι αναμενόμενη η παρατήρηση των βαθμολογιών βάσεων στα άκρα των αναγνωσμάτων που τείνουν να παρέχουν χαμηλότερη βαθμολογία ποιότητας σε σχέση με το κεντρικό του κομμάτι βάση βιβλιογραφίας (Tan et al., 2019). Συνεπώς, η διαχείριση αυτών αποτελεί απαραίτητο βήμα για την διασφάλιση της αξιοπιστίας των αναγνωσμάτων, και κατ' επέκταση των έγκυρων μεταγενέστερων αποτελεσμάτων (Bokulich et al., 2013). Στην παρούσα ανάλυση, οι βιοπληροφορικές τεχνικές βελτίωσης ποιότητας των αναγνωσμάτων συμπεριλάμβαναν την μερική αποκοπή ενός ή και των δύο άκρων τους, την συγχώνευση των paired-end αναγνωσμάτων και την αξιολόγηση της γενικής ποιότητας τους για την επιλογή διατήρησης ή απόρριψής τους, οι οποίες συνέβαλαν αποδεδειγμένα σε ότι αφορά την αντιμετώπιση χαμηλής ποιότητας βάσεων.

Πιο συγκεκριμένα, η αποκοπή των αρχικών αλληλουχιών των paired-end αναγνωσμάτων, δηλαδή του αριστερού άκρου αυτών, που εφαρμόστηκε στην αρχή και των τριών ροών επεξεργασίας η οποία αποσκοπούσε στην απαλλαγή των μη-βιολογικών εκκινητών από τα αναγνώσματα, έφερε σαν αποτέλεσμα την αφαίρεση των αρχικών αλληλουχιών που είχαν χαμηλότερη ποιότητα σε σχέση με το κεντρικό κομμάτι των αναγνωσμάτων (**Σχήμα 3.3**). Επίσης, η μερική αποκοπή των τελικών άκρων των paired-end αναγνωσμάτων βελτιώνει σημαντικά τα μεταγενέστερα αποτελέσματα ως προς την αξιοπιστία τους βάση βιβλιογραφίας (Mohsen et al., 2019). Στην περίπτωση της επεξεργαστικής ροής του DADA2, αν και δεν προσφέρεται η δυνατότητα επισκόπησης της ποιότητας των αναγνωσμάτων από την πλατφόρμα QIIME2, μια καλή ένδειξη βελτίωσης της ποιότητας ως προς το περιεχόμενο αυτών κατά την αποκοπή του τελικού δεξιού άκρου τους είναι η αποκόμιση ASVs που παρέχουν πιο αντιπροσωπευτικό μήκος με το προσεγγιστικό αντίστοιχο της περιοχής V3-V4 του 16S rRNA γονιδίου (**Πίνακας 3.9**). Η συγχώνευση των paired-end αναγνωσμάτων που εφαρμόστηκε στις ροές επεξεργασίας των Deblur και VSEARCH φανερώνει την βελτίωση της ποιότητας αναγνωσμάτων που επιφέρει αυτό το στάδιο επεξεργασίας μέσω της υψηλής βαθμολογίας ποιότητας που παρέχουν οι βάσεις της επικαλυπτόμενης περιοχής (**Σχήμα 3.4**). Το φαινόμενο αυτό οφείλεται στο γεγονός ότι οι τιμές ποιότητας των βάσεων της επικαλυπτόμενης περιοχής υποβάλλονται σε επεξεργασία κατά την συγχώνευση των paired-end αναγνωσμάτων, κατά την οποία προσδιορίζονται και προβλέπονται βελτιωμένες τελικές τιμές ποιότητας λόγω επιτυχής αντιστοιχίας (Zhang et al., 2014).

Το ποιοτικό φιλτράρισμα, το οποίο εκτελέστηκε στις επεξεργαστικές ροές των Deblur και VSEARCH με βάση την ελάχιστη βαθμολογία ποιότητας που απαιτείται να

παρέχουν οι βάσεις των αναγνωσμάτων, έφερε επίσης σαν αποτέλεσμα την βελτίωση της ποιότητας των δεδομένων. Σε κάθε παραμετρικό σενάριο που έχει παρθεί, η βαθμολογία ποιότητας ανά βάση έχει αυξηθεί, ειδικά στις βάσεις που βρίσκονται γύρω από την επικαλυπτόμενη περιοχή (**Σχήμα 3.5**), οι οποίες προ-συγχώνευσης ήταν αυτές που βρισκόντουσαν στο τελικό άκρο των paired-end αναγνωσμάτων και που παρουσίαζαν χαμηλότερη ποιότητα σε σχέση με τις υπόλοιπες βάσεις στο κεντρικό κομμάτι των αναγνωσμάτων (**Σχήμα 3.3**). Μάλιστα, παρουσιάζεται πιο έντονη η αυξηση της βαθμολογίας ποιότητας των βάσεων καθώς αυξάνεται η τιμή του Q_{min}, αλλά η αύξηση της αξιοπιστίας των δεδομένων συνοδεύεται με ένα σοβαρό κόστος, το οποίο θα σχολιαστεί σε επόμενο στάδιο. Δυστυχώς, στην επεξεργαστική ροή του DADA2 δεν υπήρχε η δυνατότητα επισκόπησης της βαθμολογίας ποιότητας των βάσεων που φέρουν τα αναγνώσματα ύστερα του ποιοτικού ελέγχου των δεδομένων, διότι η πλατφόρμα QIIME2 διαθέτει το συγκεκριμένο εργαλείο αποθορυβοποίσης υπο την μορφή ενός pipeline που εμπεριέχει όλα τα βήματα προεπεξεργασίας μετατροπής των μοναδικών αναγνωσμάτων σε ASVs που συνοδεύονται από μια τιμή συχνότητας.

Για την ανάλυση δεδομένων αλληλούχισης του 16S rRNA γονιδίου που προέρχονται από βιολογικά δείγματα άγνωστου μικροβιακού περιεχομένου και φορτίου, εκ των οποίων επιθυμείται η διερεύνηση της όσο τον δυνατόν πιο αντιπροσωπευτικής βακτηριακής σύνθεσης και ποικιλομορφίας, είναι αναγκαία η προσαρμογή των τιμών παραμέτρων στο στάδιο προεπεξεργασίας. Αυτό οφείλεται στο γεγονός ότι οι παράγοντες που αφορούν την περικοπή των αναγνωσμάτων, τις προδιαγραφές συγχώνευσης αυτών καθώς και την ποιότητα που απαιτείται να παρέχουν, μπορούν να επηρεάσουν σημαντικά την ερμηνεία των τελικών αποτελεσμάτων της ανάλυσης. Στην περίπτωση της αποκοπής και αφαίρεσης των τελικών αλληλουχιών από τα paired-end αναγνώσματα πριν την συγχώνευσή τους, ένα επεξεργαστικό στάδιο που προσφέρεται σαν επιλογή στο πρόσθετο του DADA2 στο QIIME2, ενώ μπορεί να βελτιωθούν σημαντικά τα μεταγενέστερα αποτελέσματα ως προς την αξιοπιστίας τους (Mohsen et al., 2019), ωστόσο υπάρχει περίπτωση αφαίρεσης των βάσεων της επικαλυπτόμενης περιοχής ενός μεγάλου όγκου αυτών και κατά επέκταση να επιφέρει την αποτυχία της συγχώνευσης αυτών. Αντίστοιχα, στην περίπτωση της περικοπής των συγχωνευμένων αναγνωσμάτων σε ένα συγκεκριμένο μήκος, μια παρέμβαση που απαιτείται κατά την εφαρμογή της αποθορυβοποίησης του Deblur, το σημείο αποκοπής πρέπει να επιλεχθεί με προσοχή, διότι ενώ η διατήρηση του όσο τον δυνατόν μεγαλύτερου μήκους αναγνώσματος ισοδυναμεί με την αποκόμιση περισσότερων ταξινομικών και φυλογενετικών πληροφοριών στις περαιτέρω αναλύσεις (Del Fabbro et al., 2013), τα συγχωνευμένα αναγνώσματα που φέρουν μικρότερο μήκος από αυτό που ορίζεται απορρίπτονται αυτομάτως, το οποίο δυνητικά θα μπορούσε να οδηγήσει στην μεγάλη απώλεια όγκου χρήσιμων δεδομένων.

Στην επεξεργαστική ροή του DADA2, η διαδικασία της αποθορυβοποίησης είναι ενσωματωμένη σε ένα pipeline ο οποίος περιλαμβάνει εξίσου την επιλογή εάν αυτή είναι επιθυμητή της περικοπής των paired-end αναγνωσμάτων. Στην παρούσα ανάλυση επιλέχθηκαν 2 διαφορετικά παραμετρικά σενάρια που σχετίζονται με την αποκοπή των αναγνωσμάτων, όπου στην μία περίπτωση επιλέγεται να μην εφαρμοστεί περαιτέρω αποκοπή (no trim), ενώ στην δεύτερη επιλέγεται η αποκοπή των εμπρόσθιων αναγνωσμάτων στην $220^{η}$ βάση (f: trim@=220) και για τα ανάστροφα αντίστοιχα στην $225^{η}$ βάση (r: trim@=225) και η αφαίρεση των αλληλουχιών που εμπεριέχονται στο τελικό δεξιό άκρο αυτών. Τα αποτελέσματα που προέκυψαν μέχρι και το στάδιο της ταξινομικής ανάθεσης φανερώνουν

την ελαφρώς μεγαλύτερη αξιοπιστία που παρέχουν τα σύνολα δεδομένων, στα οποία έχουν επιβληθεί αποκοπή των τελικών άκρων των paired-end αναγνωσμάτων.

Πιο αναλυτικά, ο αριθμός των αναγνωσμάτων μετά από την εφαρμογή του πρώτου βήματος επεξεργασίας του DADA2 παρουσιάζεται μειωμένος στα σύνολα δεδομένων που έχει υποβληθεί περικοπή στα μικρο-αναγνώσματα σε σχέση με αυτά της μη-αποκοπής αντίστοιχα (**Παράρτημα 11**). Αυτό οφείλεται στο ότι τα paired-end αναγνώσματα που έφεραν μικρότερο μήκος από το σημείο αποκοπής που εφαρμόστηκε απορρίφθηκαν σε αυτό το στάδιο, αλλά δεδομένου του μικρού όγκου αυτών, με μέγιστο ποσοστό περίπου 6%, δεν χαρακτηρίζεται δυσμενής αυτή η απώλεια δεδομένων. Μάλιστα, καθώς μειώνεται το μέγιστο ποσοστό αναμενόμενων σφαλμάτων που επιτρέπεται να παρέχουν τα paired-end αναγνώσματα στα μεταγενέστερα στάδια επεξεργασίας, μειώνεται η διαφορά των ποσοστών διατηρητέων αναγνωσμάτων μεταξύ των συνόλων δεδομένων, διότι όσο πιο αυστηρό είναι το ποιοτικό φιλτράρισμα τόσο μεγαλύτερος όγκος paired-end αναγνωσμάτων που δεν έχουν υποβληθεί περικοπή απορρίπτεται. Αυτό συμβαίνει λόγω του ότι τα μικρο-αναγνώσματα φέρουν περισσότερες βάσεις χαμηλής ποιότητας στο τελικό δεξί άκρο σε σύγκριση με αυτά από τα οποία έχουν αφαιρεθεί οι τελικές βάσεις. Μετά την ολοκλήρωση της αποθορυβοποίησης, δεν εντοπίζεται καμία ιδιαίτερη επιρροή που να σχετίζεται με την επιλογή της περικοπής ή μη των paired-end αναγνωσμάτων (**Παράρτημα 12**). Με την ολοκλήρωση των επεξεργασιών που εκτελούνται από τον DADA2, που συμπεριλαμβάνει την αφαίρεση χιμαιρικών αλληλουχιών και την συγχώνευση των paired-end αναγνωσμάτων, παρατηρείται η απόρριψη μεγαλύτερου ποσοστού αναγνωσμάτων στα σύνολα δεδομένων που δεν είχε υποβληθεί αποκοπή των τελικών άκρων στα μικρο-αναγνώσματα (**Πίνακας 3.7**). Εστιάζοντας και στα σύνολα δεδομένων που εφαρμόστηκε e.e.max=0.5, παρατηρείται ανάκτηση ελαφρώς μεγαλύτερου ποσοστού διατηρητέων συγχωνευμένων αναγνωσμάτων όταν εφαρμόζεται περικοπή στα τελικά άκρα των paired-end αναγνωσμάτων. Λογικά, αυτό το φαινόμενο οφείλεται στην διαδικασία της αποθορυβοποίησης του DADA2, κατά την οποία πραγματοποιούνται διορθώσεις στις βάσεις των τελικών άκρων των μικρο-αναγνωσμάτων που παρέχουν χαμηλή βαθμολογία ποιότητας λόγω του ότι το παραμετρικό μοντέλο του αλγόριθμου τις θεωρεί αποτέλεσμα σφαλμάτων. Αυτή η διαδικασία διόρθωσης υπάρχει πιθανότητα να επέφερε σε ένα σημαντικό βαθμό την παραλλαγή των αλληλουχιών στο τελικό άκρο των paired-end αναγνωσμάτων και κατά επέκταση της επικαλυπτόμενης περιοχής με αποτέλεσμα να φέρει την αποτυχία στοίχισης αυτών κατά την συγχώνευση τους. Δεδομένου ότι και τα αποτελέσματα της ταξινομικής ανάθεσης παρουσιάζουν αμελητέες διαφορές στον αριθμό αναγνωσμάτων που κατάφεραν τα ταξινομηθούν μεταξύ των συνόλων που επιβλήθηκε η περικοπή των paired-end αναγνωσμάτων και αυτών που δεν επιβλήθηκε αντίστοιχα (**Πίνακας 3.15**), η επιλογή των σημείων αποκοπής στα paired-end αναγνώσματα δεν επηρέασε αρνητικά την διαδικασία συγχώνευσης διότι δεν αφαιρέθηκαν πολλές αλληλουχίες της επικαλυπτόμενης περιοχής.

Μεταβαίνοντας την προσοχή στα χαρακτηριστικά που φέρουν τα παραγόμενα αντιπροσωπευτικά αναγνώσματα του DADA2, παρατηρείται ότι τα σύνολα δεδομένων χωρίς αποκοπή των paired-end αναγνωσμάτων φέρουν διπλάσια ποσότητα ASVs από τα σύνολα που επιβλήθηκε περικοπή των τελικών του άκρων (**Πίνακας 3.8**). Ενώ αυτό θα μπορούσε να σημαίνει περισσότερη ευκρίνεια στην μετέπειτα ταξινομική ανάθεση, το μήκος του 50% των ASVs που φέρουν τα σύνολα δεδομένων no-trim είναι μικρότερο από 250 bp (**Πίνακας 3.9**), και συνεπώς δεν αντικατοπτρίζουν καθόλου το αντίστοιχο της στοχευόμενης περιοχής V3-V4 του 16S rRNA γονιδίου που μελετάται στην παρούσα μελέτη (Vargas-Albores et al., 2017). Επιπλέον, το γεγονός ότι παρουσιάζονται περισσότερα χαμηλής αφθονίας ASVs στα σύνολα

δεδομένων χωρίς εφαρμογή περικοπής (**Παράρτημα 15**), ενισχύεται η θεωρία ότι μια μεγάλη ποσότητα αυτών είναι αποτέλεσμα σφαλμάτων που ο DADA2 δεν κατάφερε να απορρίψει. Κατά την ταξινομική ανάθεση, ενώ και στα δύο παραμετρικά σενάρια περικοπής ένα μεγάλο ποσοστό ASVs δεν κατάφερε να ταξινομηθεί, τα σύνολα δεδομένων με αποκοπή των paired-end αναγνωσμάτων έφεραν μεγαλύτερη ποσοστιαία ανάκτηση ταξινομημένων ASVs (**Πίνακας 3.14**) και λιγότερη ποσότητα χαμηλής αφθονίας ASVs σε σχέση με τα σύνολα που δεν εφαρμόστηκε αποκοπή (**Παράρτημα 29**). Ακόμα και που τα σύνολα χωρίς περικοπή paired-end αναγνωσμάτων κατέληξαν με 30% παραπάνω σε αριθμό ταξινομημένων ASVs, ο αριθμός των ταξινομικών κατηγοριών που προέκυψε από αυτά τα σύνολα δεν ξεπερνάει το ίδιο αναλογικά με τον αντίστοιχο αριθμό που φέρουν τα σύνολα που εφαρμόστηκε περικοπή (**Πίνακας 3.16**). Πρακτικά, όλα τα παραπάνω υποδεικνύουν ότι τα σύνολα δεδομένων που εφαρμόστηκε αποκοπή των τελικών άκρων των paired-end αναγνωσμάτων ήταν πιο αποδοτικά σε σχέση με τα σύνολα που δεν εφαρμόστηκε καθόλου περικοπή στα άκρα τους.

Στην ροή επεξεργασίας του Deblur, στην διαδικασία της αποθορυβοποίησης ήταν αναγκαία η περικοπή των αναγνωσμάτων σε ίσο μήκος για να εκτελεστεί με επιτυχία ο προσδιορισμός των αποστάσεων Hamming. Στην παρούσα εργασία επιλέχθηκαν δύο σημεία αποκοπής των συγχωνευμένων αναγνωσμάτων και η αφαίρεση του τελικού δεξιού άκρου, όπου στην μία περίπτωση το μήκος των αναγνωσμάτων ορίστηκε ίσο με 250 bp και στην δεύτερη αντίστοιχα ίσο με 380 bp. Η επιλογή αυτών των τιμών μήκους έχει προκύψει από μήκος των αναγνωσμάτων που φέρει το πλήθος μετά την διαδικασία του φιλτραρίσματος χαμηλής ποιότητας αυτών και από την προσπάθεια συμπερίληψης όσο τον δυνατόν μεγαλύτερου όγκου δεδομένων στην διαδικασία αποθορυβοποίησης (**Πίνακας 3.6**). Ενώ η ανάκτηση μεγαλύτερου αριθμού διατηρητέων αναγνωσμάτων (**Πίνακας 3.10**) και παραγόμενων ASVs (**Πίνακας 3.11**) στα σύνολα δεδομένων αποκοπής στην 250^{η} βάση σε σχέση με αυτά της αποκοπής στην 380^{η} βάση είναι αναμενόμενη, λόγω του ότι λαμβάνει μέρος μεγαλύτερος αριθμός αναγνωσμάτων στην διαδικασία αποθορυβοποίησης, παρατηρείται ότι τα ASVs μήκους 250 bp φέρουν μικρότερο αριθμό ταξινομικών κατηγόριων σε σχέση με τα αντίστοιχα μήκους 380 bp (**Πίνακας 3.17**). Το γεγονός ότι με μικρότερο αριθμό ASVs ανακτήθηκε περισσότερη ταξινομική πληροφορία σημαίνει ότι η διατήρηση μεγαλύτερου μήκους αναγνωσμάτων φέρει πιο αποδοτικά ASVs και μεγαλύτερη ευκρίνεια.

Όσον αφορά στις προδιαγραφές της συγχώνευσης των paired-end αναγνωσμάτων, μία βασική παράμετρος από την οποία εξαρτάται η επιρροή των δεδομένων είναι το ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής που πρέπει να παρέχουν τα αναγνώσματα για να ενωθούν αποτελεσματικά και αξιόπιστα. Το μήκος της επικαλυπτόμενης περιοχής βασίζεται στην απόδοση του οργάνου αλληλούχισης καθώς και το προσεγγιστικό μήκος της στοχευόμενης περιοχής του 16S rRNA γονιδίου. Δεδομένου ότι η πλατφόρμα Illumina δημιούργησε ζεύγη αναγνωσμάτων με μήκος 2x250 bp και το μήκος της στοχευόμενης περιοχής έχει προσδιοριστεί να είναι κατά μέσο όρο ίσο με 458 bp με εύρος 433-483 bp (Vargas-Albores et al., 2017), το μήκος της επικαλυπτόμενης περιοχής των δεδομένων της παρούσας ανάλυσης εκτιμήθηκε να είναι κατά μέσο όρο ίσο με 42 bp με εύρος 17-67 bp. Ο ορισμός μεγάλου απαιτούμενου μήκους επικαλυπτόμενης περιοχής μπορεί να οδηγήσει στην αποτυχία στοίχισης των αλληλουχιών της επικαλυπτόμενης περιοχής των paired-end αναγνωσμάτων που προέρχονται από το μεγάλο εύρος μήκους αμπλικονίων και συνεπώς παρέχουν μικρό μήκος επικαλυπτόμενης περιοχής. Από την άλλη μεριά, όταν απαιτείται η επικαλυπτόμενη περιοχή να παρέχει μικρό μήκος, ειδικά στην περίπτωση που δεν έχουν αφαιρεθεί τα χαμηλής ποιότητας τελικά άκρα των paired-end αναγνωσμάτων, υπάρχει πιθανότητα να συγχωνευθούν paired-end αναγνώσματα που δεν προέρχονται από το ίδιο

αμπλικόνιο, που σημαίνει ότι στοίχισή τους θα ήταν τυχαία. Για την διερεύνηση της επιρροής αυτής της παραμέτρου και για την προσπάθεια στοίχισης όσο τον δυνατόν περισσότερων paired-end αναγνωσμάτων που προέρχονται από αμπλικόνια εκ των οποίων το απόλυτο μήκος τους είναι άγνωστο, επιλέχθηκαν τρία ίδια παραμετρικά σενάρια συγχώνευσης και για τις τρεις ροές επεξεργασίας της παρούσας μελέτης. Οι τρεις τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής που ορίστηκαν είναι ίσες με $Overlap_{min}$ = 10, 20 και 30. Η επιρροή αυτής της παραμέτρου και στις τρεις ροές επεξεργασίας παρουσιάστηκε, αν όχι μηδενική, το πολύ αμελητέα.

Συγκεκριμένα, σε κάθε παραμετρικό σενάριο της επεξεργαστικής ροής του DADA2 παρατηρείται η ελαφρώς μείωση του αριθμού διατηρητέων αναγνωσμάτων (**Πίνακας 3.7**) και των παραγόμενων ASVs (**Πίνακας 3.8**) κατά την ολοκλήρωση της επεξεργασίας του DADA2. Επιπλέον, τα μήκη των παραγόμενων ASVs δεν παρουσιάζουν ιδιαίτερες παραλλαγές μεταξύ των διάφορων τιμών μήκους επικαλυπτόμενης περιοχής (**Πίνακας 3.9**). Τα αποτελέσματα της ταξινομικής ανάθεσης των ASVs του DADA2 παρουσιάζουν εξίσου το ίδιο μοτίβο. Ο αριθμός των ταξινομημένων ASVs (**Πίνακας 3.14**), το ποσοστό διατηρητέων ταξινομημένων αναγνωσμάτων (**Πίνακας 3.15**) καθώς και ο αριθμός ταξινομικών κατηγοριών (**Πίνακας 3.16**) παρουσιάζουν αμελητέες διαφορές μεταξύ των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής. Η μόνη αξιοσημείωτη παρατήρηση που φανερώνεται σε όλα τα προαναφερόμενα αποτελέσματα είναι η ελαφρώς μείωση των μετρήσεων καθώς αυξάνεται το απαιτούμενο μήκος επικαλυπτόμενης περιοχής. Αυτό το φαινόμενο λογικά οφείλεται στην αποτυχία συγχώνευσης paired-end αναγνωσμάτων των οποίων το μήκος της επικαλυπτόμενης περιοχής είναι μικρότερη από 20 bp ή 30bp, κάτι που είναι αναμενόμενο δεδομένου του εύρους του μήκους της στοχευόμενης περιοχής V3-V4 του 16S rRNA γονιδίου. Παρόλα αυτά, το γεγονός ότι είναι αμελητέες οι διαφορές των αποτελεσμάτων υποδηλώνεται ότι η συντριπτική πλειοψηφία των paired-end αναγνωσμάτων των δεδομένων της παρούσας ανάλυσης παρέχουν μήκος επικαλυπτόμενης περιοχής τουλάχιστον ίσο με 30 bp.

Αυτό το συμπέρασμα επιβεβαιώνεται και από τα αποτελέσματα των επεξεργαστικών ροών του Deblur και VSEARCH. Η συγχώνευση των paired-end αναγνωσμάτων, η οποία είναι από τις πρώτες εκτελέσεις των επεξεργαστικών ροών, φανερώνει εξίσου ότι οι διάφορες τιμές του ελάχιστου μήκους επικαλυπτόμενης περιοχής επηρεάζουν ελάχιστα τον αριθμό συγχωνευμένων αναγνωσμάτων που ανακτούνται (**Πίνακας 3.3**). Επιπλέον, ο μετασχηματισμός προς αύξηση της βαθμολογίας ποιότητας των βάσεων της επικαλυπτόμενης περιοχής ρίχνει φώς στο μήκος αυτής της συντριπτικής πλειοψηφίας των αναγνωσμάτων, που είναι προσεγγιστικά ίσο με 40 bp (**Σχήμα 3.4**). Στα επόμενα επεξεργαστικά στάδια των ροών του Deblur και VSEARCH παρουσιάζεται ακόμα πιο έντονη η αποχή της επιρροής των διάφορων τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής σε ότι αφορά στον αριθμό διατηρητέων φιλτραρισμένων αναγνωσμάτων (**Πίνακας 3.5**), στον αριθμό διατηρητέων αναγνωσμάτων μετά την διαδικασία αποθορυβοποίησης του Deblur (**Πίνακας 3.10**) και ομαδοποίησης του VSEARCH (**Πίνακας 3.12**), στον αριθμό παραγόμενων ASVs (**Πίνακας 3.11**) και OTUs (**Πίνακας 3.13**) και τέλος στον αριθμό ταξινομικών μονάδων που προέκυψε κατά την ταξινομική ανάθεση των δεδομένων του Deblur (**Πίνακας 3.17**) και του VSEARCH (**Πίνακας 3.18**).

Μία εξίσου ενδιαφέρουσα παρατήρηση αποτελεί το γεγονός ότι στα τελικά αποτελέσματα του Deblur και VSEARCH παρατηρείται μηδενική διαφορά στις μετρήσεις μεταξύ των διάφορων τιμών $Overlap_{min}$, σε αντίθεση με τα αποτελέσματα του DADA2 που

παρουσιάζουν τουλάχιστον μία αμελητέα μείωση των μετρήσεων καθώς αυξάνεται το ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής. Αυτή η παρατήρηση μπορεί να οφείλεται στην σειρά που εκτελέστηκε το στάδιο την συγχώνευσης των paired-end αναγνωσμάτων, όπου στην περίπτωση του DADA2 πραγματοποιήθηκε μετά το ποιοτικό φιλτράρισμα και την αποθορυβοποίηση των αναγνωσμάτων ενώ στην περίπτωση των Deblur και VSEARCH εκτελέστηκε πριν την διαδικασία ποιοτικού φιλτραρίσματος και αποθορυβοποίησης και ομαδοποίησης αντίστοιχα. Βέβαια, ανεξαρτήτως σειράς εκτέλεσης και παραμετρικού σεναρίου, η διαδικασία συγχώνευσης επιφέρει την μείωση του ποσοστού ανάκτησης διατηρητέων συγχωνευμένων αναγνωσμάτων. Αυτή η επίπτωση είναι αναμενόμενη για κάθε επεξεργαστική ροή. Στην περίπτωση αυτής των Deblur και VSEARCH, η συγχώνευση των paired-end αναγνωσμάτων έφερε την ανάκτηση περίπου του 80% διατηρητέων αναγνωσμάτων σε σχέση με τα πρωτογενή δεδομένα, που σημαίνει ότι περίπου το 20% των paired-end αναγνωσμάτων δεν παρείχε τις απαραίτητες προδιαγραφές για την στοίχισή τους (**Πίνακας 3.3**). Σε ένα βαθμό, το ποσοστό θεωρείται αναμενόμενο, αφού τα δεδομένα αλληλούχισης σε αυτό το στάδιο δεν έχουν υποστεί καμία ουσιαστική επεξεργασία και ποιοτικό έλεγχο με αποτέλεσμα να συμπεριλαμβάνονται βάσεις κακής ποιότητας και σφάλματα αλληλούχισης στην επικαλυπτόμενης περιοχή που εμπόδισαν την διαδικασία συγχώνευσης για το 20% των αναγνωσμάτων.

Στην περίπτωση της επεξεργαστικής ροής του DADA2 ενώ θεωρητικά θα έπρεπε μετά την διαδικασία αποθορυβοποίησης να υπήρχε πλήρης στοίχιση των διατηρητέων αναγνωσμάτων προ-συγχώνευσης με βάση την λειτουργία του αλγορίθμου (Callahan et al., 2016), παρατηρείται στα αποτελέσματα της παρούσας ανάλυσης ένα ποσοστό προσεγγιστικά της τάξεως του 2% με 10% των paired-end αναγνωσμάτων να απορρίπτεται μετά την ολοκλήρωση του pipeline DADA2 (**Παράρτημα 12**, **Πίνακας 3.7**). Ωστόσο, τα αποτελέσματα τις ολοκλήρωσης του DADA2 περιλαμβάνουν και τον εντοπισμό και αφαίρεση χιμαιρικών αλληλουχιών, οπότε ενδεχομένως το ποσοστό απορριπτόμενων αναγνωσμάτων που προέκυψε μπορεί να οφείλεται σε αυτήν την διαδικασία. Παρόλα αυτά, αν παρθεί η παραδοχή ότι είναι και αποτέλεσμα αποτυχίας συγχώνευσης των paired-end αναγνωσμάτων, χαρακτηρίζεται αμελητέο το ποσοστό που απορρίφτηκε, ενώ καθώς μειώνεται το ποσοστό αναμενόμενων σφαλμάτων που επιτρέπεται να παρέχουν τα αναγνώσματα, παρατηρείται η τάση της πλήρης συγχώνευσης των διατηρητέων αναγνωσμάτων. Αυτό σημαίνει ότι όσο αυξάνεται η αξιοπιστία των paired-end αναγνωσμάτων, η πιθανότητα να συγχωνευθούν αυτά είναι μεγαλύτερη.

Η τελευταία και πιο σημαντική παράμετρος από την οποία εξαρτάται σημαντικά η τελική εξαγωγή δεδομένων είναι αυτή που σχετίζεται με το ποιοτικό φιλτράρισμα. Το ποιοτικό φιλτράρισμα με βάση παραμέτρους όπως το μέγιστο ποσοστό αναμενόμενων σφαλμάτων που επιτρέπεται να έχει το κάθε ανάγνωσμα (e.e.$_{max}$) (Edgar & Flyvbjerg, 2015) ή η ελάχιστη βαθμολογία ποιότητας Phred που απαιτείται να παρέχουν οι βάσεις των αναγνωσμάτων (Q_{min}) (Bokulich et al., 2013), παίζει σημαντικό ρόλο στην αφαίρεση αναγνωσμάτων χαμηλής ποιότητας, σφαλμάτων αλληλούχισης και τεχνουργημάτων και μπορεί να επηρεάσει σημαντικά την ποιότητα και την αξιοπιστία των δεδομένων αλληλούχισης και των μεταγενέστερων αποτελεσμάτων, συμπεριλαμβανομένου της ταξινομικής ανάθεσης και της ανάλυσης ποικιλομορφίας. Αυτό το στάδιο βιοπληροφορικής επεξεργασίας επιλέγεται να εκτελεστεί πριν την διαδικασία της αποθορυβοποίησης ή ομαδοποίησης αναγνωσμάτων για την μείωση του υπολογιστικού φόρτου και την βελτίωση της απόδοσης και αποτελεσματικότητας αυτών των σημαντικών βιοπληροφορικών διαδικασιών. Ωστόσο, το εξαιρετικά αυστηρό ποιοτικό φιλτράρισμα μπορεί να οδηγήσει στην

απώλεια αναγνωσμάτων που παρέχουν γνήσια βιολογική πληροφορία, ενώ οι υπερβολικά ελαστικές προδιαγραφές ποιοτικού ελέγχου μπορεί να επιτρέψει την συμπερίληψη εσφαλμένων αλληλουχιών στο υπόλοιπο κομμάτι της επεξεργαστικής ροής (Bokulich et al., 2013). Στις επεξεργαστικές ροές των Deblur και VSEARCH εφαρμόστηκε ποιοτικό φιλτράρισμα βασισμένο στην ελάχιστη βαθμολογία ποιότητας Phred, ενώ στην επεξεργαστική ροή του DADA2 ο ποιοτικό έλεγχος εξαρτήθηκε από το μέγιστο ποσοστό αναμενόμενων σφαλμάτων. Όπως έχει ήδη αναφερθεί, η διαδικασία του ποιοτικού φιλτραρίσματος εκτελέστηκε με διαφορετική σειρά στην επεξεργαστική ροή του DADA2 σε σχέση με τις υπόλοιπες δύο, παρόλα αυτά σε κάθε περίπτωση φαίνεται πως ο ορισμός των τιμών των προαναφερόμενων παραμέτρων επηρεάζουν την εικόνα των δεδομένων.

Πιο αναλυτικά, στην περίπτωση των επεξεργαστικών ροών Deblur και VSEARCH η διαδικασία ελέγχου και αφαίρεσης αναγνωσμάτων χαμηλής ποιότητας εκτελέστηκε χρησιμοποιώντας την μέθοδο q-score (Bokulich et al., 2013), η οποία βασίζεται στην ελάχιστη τιμή βαθμολογίας Q_{min}. Γενικά, είναι κοινώς αποδεκτό ότι οι τιμές βαθμολογίας ποιότητας $Q \geq 20$ χαρακτηρίζουν τις βάσεις επαρκώς αξιόπιστες για την διατήρησή τους (Pfeifer, 2017). Για την αξιολόγηση της επιρροής αυτής της παραμέτρου στα διαθέσιμα δεδομένα αλληλούχισης της παρούσας εργασίας, εφαρμόστηκαν τρεις διαφορετικές τιμές, όπου Q_{min}=20, 22 και 26. Ανεξαρτήτως της τιμής που επιλέχθηκε, η βελτίωση της ποιότητας των αναγνωσμάτων παρουσιάζεται έντονα μετά την εφαρμογή του ποιοτικού φιλτραρίσματος (**Σχήμα 3.5**). Μάλιστα, η βελτίωση παρατηρείται ακόμα πιο έντονη στα σύνολα δεδομένων με Q_{min}=26 σε σύγκριση με αυτών των Q_{min}=20 και 22, τα οποία παρουσιάζουν ίδια χαρακτηριστικά ποιότητας. Η ομοιότητα των συνόλων δεδομένων του Q_{min}=20 και 22 εντοπίστηκε σε όλα τα αποτελέσματα του τμήματος των επεξεργαστικών ροών μέχρι και την ταξινομική ανάθεση, ενώ οι παραλλαγές των αντίστοιχων αποτελεσμάτων παρουσιάστηκαν στα σύνολα δεδομένων με Q_{min}=26. Αρχικά, από το στάδιο της εφαρμογής του ποιοτικού φιλτραρίσματος, παρατηρήθηκε η απόρριψη ενός ποσοστού ύψους 16% των συγχωνευμένων αναγνωσμάτων για Q_{min}=20 και 22, ενώ για Q_{min}=26 το ποσοστό αυτό προέκυψε ίσο με περίπου 40%. Το μοτίβο των μετρήσεων των αποτελεσμάτων μεταξύ των συνόλων δεδομένων με Q_{min}=26 και των συνόλων με Q_{min}=20 και 22 παρουσιάστηκε ίδιο μέχρι και στον αριθμό τον ταξινομικών κατηγοριών που προέκυψαν από την ταξινομική ανάθεση των ASVs για τον Deblur και των OTUs του VSEARCH. Στα σύνολα δεδομένων με Q_{min}=26 διατηρήθηκαν λιγότερα αναγνώσματα με μήκος πάνω από 250 bp (**Πίνακας 3.6**), ανακτήθηκε μικρότερο ποσοστό διατηρητέων αναγνωσμάτων μετά την διαδικασία αποθορυβοποίησης (**Πίνακας 3.10**) και ομαδοποίησης (**Πίνακας 3.12**) και αντίστοιχα της ταξινομικής ανάθεσης, κάτι το οποίο επέφερε τον σημαντικά μειωμένο αριθμό αναγνωσμάτων σε κάποια δείγματα, παράχθηκε μικρότερος αριθμός παραγόμενων και ταξινομημένων ASVs και OTUs και προσδιορίστηκε μικρότερος αριθμός ταξινομικών κατηγοριών (**Πίνακας 3.17**, **Πίνακας 3.18**) σε σχέση με τα σύνολα δεδομένα με Q_{min}=20 και 22.

Στην περίπτωση της επεξεργαστική ροής του DADA2, το ποιοτικό φιλτράρισμα βασίζεται στην τιμή του μέγιστου ποσοστού αναμενόμενων σφαλμάτων που επιτρέπεται να παρέχουν τα αναγνώσματα. Γενικά, οι τιμές e.e.≤ 3% που προκύπτουν από τον προσδιορισμό της ποιότητας του αναγνώσματος θεωρούνται καλές και, κατά συνέπεια, εξασφαλίζουν την αξιοπιστία του αναγνώσματος (Edgar & Flyvbjerg, 2015). Για την αξιολόγηση της επιρροής αυτής της παραμέτρου στα διαθέσιμα δεδομένα αλληλούχισης της παρούσας εργασίας, εφαρμόστηκαν τρεις διαφορετικές τιμές, όπου $e.e._{max}$= 0.5, 1.5 και 2.5. Τα αποτελέσματα αυτής της επεξεργαστικής ροής παρουσιάστηκαν εξίσου με το ίδιο μοτίβο όπως και στις

άλλες δύο επεξεργαστικές ροές, με την μόνη διαφορά ότι οι μετρήσεις των διατηρητέων αναγνωσμάτων προέκυψαν πιο αναλογικές και γραμμικές με τις τιμές $e.e._{max}$ που επιλέχθηκαν. Συγκεκριμένα, παρατηρείται μια σταδιακή και διακριτή μείωση του ποσοστού διατηρητέων paired-end αναγνωσμάτων ύστερα του ποιοτικού φιλτραρίσματος (**Παράρτημα 11**), της διαδικασία αποθορυβοποίησης (**Παράρτημα 12**), της ολοκλήρωσης του DADA2 (**Πίνακας 3.7**) και της ταξινομικής ανάθεσης. Παρόλα αυτά, μεταξύ των συνόλων δεδομένων με $e.e._{max}$= 0.5 και αυτών με $e.e._{max}$=1.5 και 2.5, εντοπίστηκε η ίδια συμπεριφορά των μετρήσεων των παραγόμενων/ταξινομημένων ASVs και ταξινομικών κατηγοριών που παρατηρήθηκε και στις επεξεργαστικές ροές των Deblur και VSEARCH. Για τιμή $e.e._{max}$= 0.5 ανακτάται μικρότερος αριθμός παραγόμενων και ταξινομημένων ASVs καθώς και των ταξινομικών κατηγοριών σε σχέση με την επιλογή $e.e._{max}$=1.5 ή 2.5, όπου τα σύνολα αυτών παρουσιάζουν αμελητέες διαφορές σε αυτές τις μετρήσεις. Ως εκ τούτου, η συντήρηση πολύ αξιόπιστων αναγνωσμάτων έφερε σαν αποτέλεσμα την έντονη μείωση όγκου αναγνωσμάτων, ταξινομικών πληροφοριών και δυνητικά των πληροφοριών ως προς την ποικιλομορφία των δειγμάτων, καθώς κάποια από αυτά στην περίπτωση του DADA2 και Deblur είχαν μικρότερο αριθμό αναγνωσμάτων από αυτόν που συνιστάται προκειμένου να είναι ικανοποιητική περιγραφή της βακτηριακής σύστασής τους (Bukin et al., 2019).

Αποτελεί δεδομένο το γεγονός ότι η επιλογή ενός συγκεκριμένου βιοπληροφορικού εργαλείου καθώς και ο τρόπος με τον οποίο εφαρμόζεται αυτό μπορούν να επηρεάσουν σημαντικά το περιεχόμενο των δεδομένων αλληλούχισης (Allaband et al., 2019). Συνεπώς, η αξιοπιστία της συγκριτικής διερεύνησης μεταξύ των επεξεργαστικών ροών που βασίζονται στην παραγωγή ASVs και OTUs εξαρτάται σε ένα μεγάλο βαθμό από την επιλογή τιμών παραμέτρων βασικών επεξεργαστικών σταδίων στο πλαίσιο της διαχείρισης paired-end αναγνωσμάτων. Αναγνωρίζοντας την πιθανή εισαγωγή μεροληψιών σε αυτό το πρώιμο στάδιο της συγκριτικής ανάλυσης, και γενικά της ανάλυσης δεδομένων αλληλούχισης του 16S rRNA γονιδίου, επιχειρήθηκε η προσπάθεια, όσο αυτή ήταν εφικτή, να επιλεχτούν παρόμοια παραμετρικά σενάρια για κάθε επεξεργαστική ροή. Ο πειραματισμός στις τιμές των βασικών παραμέτρων, ο οποίος έχει εστιαστεί για ένα μεγάλο μέρος της παρούσας διπλωματικής εργασίας, καθώς και τα διάφορα βιοπληροφορικά εργαλεία βασιζόμενα σε μεθόδους αποθορυβοποίησης και ομαδοποίησης αναγνωσμάτων έφεραν στην επιφάνεια χρήσιμες πληροφορίες σε ότι αφορά το τρόπο που συμπεριφέρονται τα δεδομένα ως απόκριση αυτών των προ-επεξεργαστικών σταδίων. Μία από τις πιο χρήσιμες πληροφορίες αποτελεί ο τρόπος με τον οποίο παρουσιάζονται οι παραλλαγές των δεδομένων ανάλογα τις τιμές παραμέτρων που επιλέγονται σε ότι αφορά την συγχώνευση paired-end αναγνωσμάτων, την περικοπή paired-end ή συγχωνευμένων αναγνωσμάτων και το ποιοτικό φιλτράρισμα. Ο πειραματισμός αυτών των τιμών επέτρεψε την διάκριση της επιρροής των παραμέτρων και της αντίστοιχης των μεθόδων αποθορυβοποίησης και ομαδοποίησης στα δεδομένα αλληλούχισης. Με αυτόν τον τρόπο, διασφαλίστηκε η αξιοπιστία της σύγκρισης των μεθόδων αποθορυβοποίησης και ομαδοποίησης και η πιστότητα των μεταγενέστερων αποτελεσμάτων.

Για την σχολαστική εμβάθυνση του θέματος της επιλογής των βέλτιστων τιμών παραμέτρων, αρχικά πρέπει να γίνει πιο ξεκάθαρος και κατανοητός ο προβληματισμός που προέκυψε κατά την διεκπεραίωση της παρούσας διπλωματικής εργασίας στα πλαίσια αυτού του θέματος. Η κατασκευή του πρώιμου σταδίου της κάθε επεξεργαστικής ροής εξαρτήθηκε από την αναγκαία τεχνική και ποιοτική διαχείριση των αναγνωσμάτων έτσι ώστε να εκτελεστεί με επιτυχία και αξιοπιστία η αποθορυβοποίηση ή ομαδοποίηση αυτών. Λόγω αυτού, παρατηρήθηκε ότι κάποια από τα βασικά προεπεξεργαστικά στάδια για την η εφαρμογή του DADA2, Deblur και VSEARCH στην πλατφόρμα QIIME2 ήταν διαφορετικά

και κάποια ήταν κοινά μεταξύ τους. Δύο από τα κοινά επεξεργαστικά βήματα που χρειαζόταν να εκτελεστεί και στις τρεις επεξεργαστικές ροές ήταν η συγχώνευση των paired-end αναγνωσμάτων και το ποιοτικό φιλτράρισμα. Ενώ αυτά τα βήματα θεωρητικά έχουν κοινό στόχο, διαπιστώθηκε ότι πρακτικά αυτές οι διαδικασίες εκτός του ότι εκτελούνται με διαφορετική σειρά, το ποιοτικό φιλτράρισμα εκτελείται με διαφορετικό αλγοριθμικό τρόπο μεταξύ των επεξεργαστικών ροών του DADA2 και των Deblur και VSEARCH.

Συγκεκριμένα, για την αποθορυβοποίηση σε ASVs χρησιμοποιώντας το pipeline Deblur και για την ομαδοποίηση σε OTUs χρησιμοποιώντας το pipeline VSEARCH τα δεδομένα εισόδου έπρεπε να έχουν επεξεργαστεί ως προς την συγχώνευση των paired-end αναγνωσμάτων και το φιλτράρισμα χαμηλής ποιότητας αναγνωσμάτων χρησιμοποιώντας αλλά βιοπληροφορικά εργαλεία. Η πλατφόρμα QIIME2 παρείχε ένα pipeline για την κάθε διαδικασία, όπου ο ένας εκτελούσε την συγχώνευση των paired-end αναγνωσμάτων με βάση το ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής ($Overlap_{min}$) και ο άλλος σχετιζόταν με την διατήρηση αναγνωσμάτων των οποίων οι βάσεις παρείχαν ελάχιστη βαθμολογία ποιότητας Phred (Q_{min}). Δεδομένου ότι το pipeline που αφορά το ποιοτικό φιλτράρισμα δεχόταν δεδομένα εισόδου όπου το περιεχόμενό τους αποτελούνταν από single-end αναγνώσματα, ήταν αναγκαίο να συγχωνευτούν τα paired-end αναγνώσματα πριν το ποιοτικό έλεγχο, διότι τα συγχωνευμένα αναγνώσματα μπορούσαν να αναγνωριστούν από το pipeline ως single-end. Στην περίπτωση του pipeline DADA2, τα δεδομένα εισόδου δεν χρειάστηκαν να επεξεργαστούν προηγουμένως από άλλο pipeline σε ότι αφορά την συγχώνευση των αναγνωσμάτων και το ποιοτικό φιλτράρισμα διότι αυτές οι διαδικασίες συμπεριλαμβάνονταν στο ίδιο pipeline. Όμως στον DADA2, για να εκτελεστεί η διαδικασία αποθορυβοποίησης αποδοτικά, η σειρά με την οποία εφαρμόστηκαν αυτές οι επεξεργασίες είναι πρώτα το φιλτράρισμα χαμηλής ποιότητας paired-end αναγνωσμάτων με βάση το μέγιστο επιτρεπτό ποσοστό αναμενόμενων σφαλμάτων που παρείχαν ($e.e._{max}$), ύστερα ακολούθησε η διαδικασία αποθορυβοποίησης, και μετά τα φιλτραρισμένα και αποθορυβοποιημένα paired-end αναγνώσματα επιβλήθηκαν στην διαδικασία συγχώνευσης με βάση το ελάχιστο απαιτούμενο μήκος επικαλυπτόμενης περιοχής ($Overlap_{min}$), που είναι η ίδια λογική με αυτήν του pipeline που χρησιμοποιήθηκε για τις επεξεργαστικές ροές των Deblur και VSEARCH. Ενώ δεν υπήρχε περιθώριο πειραματισμού σε ότι έχει να κάνει με την σειρά με την οποία πρέπει να εκτελεστούν οι διαδικασίες ποιοτικού ελέγχου και συγχώνευσης paired-end αναγνωσμάτων σε κάθε επεξεργαστική ροή, υπήρχε η δυνατότητα επιλογής διάφορων τιμών των παραμέτρων $Overlap_{min}$, Q_{min} και $e.e._{max}$.

Η παράμετρος που αφορά την διαδικασία συγχώνευσης είναι ακριβώς η ίδια για την επεξεργαστική ροή του Deblur και του VSEARCH, ενώ στην επεξεργαστική ροή του DADA2 η αντίστοιχη παράμετρος δεν είναι ακριβώς ίδια, αλλά πάρθηκε με μεγάλη αξιοπιστία η παραδοχή να είναι παρόμοιας λογικής. Συνεπώς, ορίστηκαν τρεις διαφορετικές τιμές $Overlap_{min}$ οι οποίες ήταν ίδιες σε κάθε επεξεργαστική ροή. Τα αποτελέσματα που προέκυψαν από τις διάφορες τιμές $Overlap_{min}$ σε κάθε επεξεργαστική ροή επιβεβαιώνει το γεγονός ότι όχι μόνο δεν επιφέρουν καμία αλλαγή στα σύνολα δεδομένων, αλλά ότι διακρίνεται μια αντιστοίχιση αυτών των τιμών που επιλέχθηκαν.

Η διαδικασία του ποιοτικού φιλτραρίσματος αποδείχθηκε πιο περίπλοκη ως προς την επιλογή των αντίστοιχων τιμών Q_{min} και $e.e._{max}$ καθώς ο τρόπος που αξιολογούν ποιοτικά τα αναγνώσματα είναι σημαντικά διαφορετικός μεταξύ τους. Παρόλα αυτά, ο ορισμός του $e.e._{max}$ βασίζεται σχεδόν εξολοκλήρου από τις βαθμολογίες Q που παρέχουν οι βάσεις των αναγνωσμάτων. Για την προσπάθεια επιλογής παρόμοιων παραμετρικών σεναρίων για κάθε

επεξεργαστική ροή, πάρθηκε η παραδοχή ότι οι τιμές $e.e._{max}$= 2.5, 1.5 και 0.5 αντιστοιχούν στις ελάχιστες τιμές βαθμολογίας ποιότητας των βάσεων που θα πρέπει να παρέχουν κατά προσέγγιση τα αναγνώσματα: Q_{min}= 20 ή 22 ή 26, σύμφωνα με την προαναφερόμενη εξίσωση και ορισμό του $e.e._{max}$. Σε έναν μεγάλο βαθμό η παραδοχή αυτή επιβεβαιώθηκε μέσω των χαρακτηριστικών παραλλαγών των αριθμών παραγόμενων και ταξινομημένων ASVs και OTUs και των αριθμών ταξινομικών κατηγοριών που προέκυψαν από κάθε επεξεργαστική ροή.

Η επιλογή περικοπής των τελικών άκρων των paired-end αναγνωσμάτων ήταν εφικτή στην παρούσα μελέτη μόνο στην επεξεργαστική ροή του DADA2, η οποία ήταν ενσωματωμένη στο pipeline του DADA2. Αυτή η διαδικασία δεν υποστηριζόταν από άλλον διαθέσιμο pipeline στην πλατφόρμα QIIME2. Δεδομένου ότι η επιλογή αποκοπής στον DADA2 δεν ήταν υποχρεωτική, επιλέχθηκαν δύο παραμετρικά σενάρια, αυτής της μη-περικοπής και αυτής της αφαίρεσης των τελικών βάσεων σε ένα συγκεκριμένο μήκος των paired-end αγωνισμάτων για την διερεύνηση της επίπτωσης αυτών των επιλογών. Μία ακόμα παρέμβαση που αφορά στην περικοπή αναγνωσμάτων πραγματοποιήθηκε στην επεξεργαστική ροή του Deblur, όπου στο pipeline του Deblur έπρεπε να επιλεχθεί ένα συγκεκριμένο μήκος των συγχωνευμένων αναγνωσμάτων προκειμένου να πραγματοποιηθεί η διαδικασία αποθορυβοποίησης. Για την διερεύνηση της επιρροής της αποκοπής των συγχωνευμένων αναγνωσμάτων σε διαφορετικά μήκη, επιλέχθηκαν δύο παραμετρικά σενάρια, όπου στην μία περίπτωση έγινε η προσπάθεια συμπερίληψης όσο τον δυνατόν περισσότερου όγκου δεδομένων στην διαδικασία αποθορυβοποίησης, και στην δεύτερη περίπτωση επιλέχθηκε να απορριφθεί ένας σημαντικός όγκος δεδομένων για την διατήρηση μεγαλύτερου μήκους συγχωνευμένων αναγνωσμάτων στις μετέπειτα αναλύσεις. Στην επεξεργαστική ροή του VSEARCH δεν υπήρχε η δυνατότητα περικοπής paired-end ή συγχωνευμένων αναγνωσμάτων, διότι δεν ήταν διαθέσιμο κάποιο ανεξάρτητο από αυτό του DADA2 και Deblur pipeline που να εκτελούσε αυτές τις διαδικασίες. Συνεπώς, η επεξεργαστική ροή του VSEARCH κατέληξε να έχει μόνο ένα παραμετρικό σενάριο σε αυτό το πλαίσιο, δηλαδή την μη περικοπή των αναγνωσμάτων σε συγκεκριμένο σημείο.

Παρατηρώντας την επιρροή που έχουν τα διάφορα παραμετρικά σενάρια στα δεδομένα αλληλούχισης μέχρι και το σημείο της ταξινομικής ανάθεσης των ASVs και OTUs, και εντοπίζοντας διάφορα παρόμοια μοτίβα μεταξύ των τριών επεξεργαστικών ροών, φανερώθηκαν πιο ξεκάθαρα οι διαφορές την επίπτωσης των μεθόδων αποθορυβοποίησης και ομαδοποίησης στα δεδομένα αλληλούχισης στο πρώιμο στάδιο επεξεργασίας ανάλυσης του 16S rRNA γονιδίου. Βέβαια, δεδομένου ότι στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν δύο διαφορετικά εργαλεία αποθορυβοποίησης, φανερώθηκαν εξίσου διαφορές μεταξύ των επεξεργαστικών ροών του DADA2 και του Deblur. Η πρώτη εικόνα των αποτελεσμάτων παρουσίασε το γεγονός ότι η εφαρμογή της αποθορυβοποίησης με Deblur συνοδεύτηκε με την απώλεια μεγάλου όγκου αναγνωσμάτων σε σχέση με τον αρχικό όγκο πρωτογενών δεδομένων, ανεξαρτήτως παραμετρικών σεναρίων. Αντίθετα, η διατήρηση αναγνωσμάτων στις ροές επεξεργασίας των DADA2 και VSEARCH φάνηκε να επηρεάζεται περισσότερο από το πόσο αυστηρός ήταν ο ποιοτικός έλεγχος, με την μόνη εξαίρεση να αποτελούν τα σύνολα δεδομένων με Q_{min}=20 και Q_{min}=22 της επεξεργαστικής ροής του VSEARCH, στην οποία, όπως έχει ήδη αναφερθεί παραπάνω, δεν παρουσιάζονται σημαντικές αλλαγές. Σε έναν βαθμό, είναι αρκετά λογικό το φαινόμενο απώλειας μεγάλης ποσότητας αναγνωσμάτων στην περίπτωση του Deblur, διότι έστω ότι υπάρχει η παραδοχή ότι η μέση βαθμολογία ποιότητας των βάσεων είναι 20, και άρα το μέσο ποσοστό σφάλματος είναι 0,01%, τότε για μήκος αναγνωσμάτων 250 bp ή 380 bp το 1-(1-0,01)^250=91,9% ή

97,8% των αναγνωσμάτων αντίστοιχα αναμένεται να περιέχει τουλάχιστον ένα σφάλμα υποκατάστασης βάσης. Υποθέτοντας ότι ισχύει αυτή η πιθανότητα και ότι η αποθορυβοποίηση του Deblur, που δεν έχει καμία διορθωτική παρέμβαση, εκτελεστεί επιτυχώς, τα δείγματα που έχουν περίπου κατά μέσο όρο $16x10^4$ (**Πίνακας 3.5**) αναγνώσματα αναμένεται μετά την εφαρμογή του pipeline να παρέχουν 13000 ή 3500 περίπου διατηρητέα αναγνώσματα αντίστοιχα. Ενδεχομενώς, με αυτόν τον τρόπο εξηγείται και το μειωμένο ποσοστό διατηρητέων αναγνωσμάτων στην περίπτωση αποκοπής της $380^{ης}$ βάσης σε σχέση με αυτό της $250^{ης}$ βάσης. Γενικά, η τάση της αποθορυβοποίησης με Deblur να διατηρεί πολύ λιγότερο αριθμό αναγνωσμάτων σε σχέση με τον DADA2 καθώς και με διάφορα εργαλεία που εφαρμόζουν την μέθοδο ομαδοποίησης έχει αναδειχθεί στο παρελθόν (Prodan et al., 2020), όποτε η μεγάλη απώλεια όγκου δεδομένων σε αυτήν την περίπτωση είναι αναμενόμενη.

Μία αξιοσημείωτη παρατήρηση αποτέλεσε ο σημαντικός όγκος των παραγόμενων ASVs της επεξεργαστικής ροής του DADA2 που δεν κατάφερε να ταξινομηθεί σε καμία ταξινομική κατηγορία της βάσης δεδομένων SILVA κατά την ταξινομική ανάθεση. Μάλιστα, περίπου το 50% αυτών κατέληξε να μην ταξινομηθεί και τα μη-ταξινομημένα ASVs εντοπίστηκαν μόνο σε δείγματα αίματος, ενώ στην επεξεργαστική ροή του Deblur παρουσιάστηκε πλήρης ταξινόμηση των ASVs. Στην περίπτωση του VSEARCH κατέληξε να μην ταξινομηθούν τρία μοναδικά OTUs, όπου στο βάθος των 6500 με 9500 παραγόμενων OTUs, θεωρείται αμελητέα η απώλεια και κατά επέκταση παρουσιάστηκε και σε αυτήν την επεξεργαστική ροή πλήρης ταξινόμηση των επεξεργασμένων δεδομένων. Ο λόγος για τον οποίο παρουσιάζεται αυτό το φαινόμενο κατά πάσα πιθανότητα οφείλεται στην λειτουργία των Deblur και VSEARCH, στην οποία χρησιμοποιείται μία βάση δεδομένων ως βοηθητική. Συγκεκριμένα, στην περίπτωση του Deblur, η διαδικασία αποθορυβοποίησης περιλαμβάνει και τον εντοπισμό χιμαιρικών αλληλουχιών και τεχνουργημάτων με την χρήση της βάσης δεδομένων Greengenes με σκοπό να επικυρωθεί ότι η προέλευση των αλληλουχιών των παραγόμενων ASVs είναι από την αλληλούχιση του 16S rRNA γονιδίου. Αντίστοιχα, η μέθοδος ομαδοποίησης κλειστής αναφοράς βασίζεται εξ ολοκλήρου στην χρήση μίας βάσης δεδομένων, όπου στην προκειμένη περίπτωση είναι η SILVA, με σκοπό την παραγωγή OTUs με βάση το ποσοστό ομοιότητας των άγνωστης βακτηριακής προέλευσης αναγνωσμάτων με τις αλληλουχίες γνωστής και ταξινομημένης βακτηριακής προέλευσης. Δεδομένου ότι κατά την αποθορυβοποίηση με DADA2 δεν πραγματοποιήθηκε καμία χρήση μίας βάσης δεδομένων, ήταν αναμενόμενο να παρουσιαστούν μη-ταξινομημένα ASVs. Ενδιαφέρουσα παρατήρηση αποτέλεσε το γεγονός ότι αυτά τα ASVs εντοπίστηκαν μόνο στα δείγματα αίματος, που μπορεί να σημαίνει ότι οι αλληλουχίες αυτών των ASVs να προέρχονται από βακτήρια των οποίων το 16S rRNA γονίδιό τους δεν έχει ακόμα κατοχυρωθεί στην βάση δεδομένων. Λόγω του ότι γενικά οι βάσεις δεδομένων έχουν κατασκευαστεί βασιζόμενα από δεδομένα καλλιεργήσιμων βακτηρίων και ευρήματα μελετών κυρίως του εντερικού μικροβιώματος, υπάρχει πιθανότητα να μην ταξινομηθούν αλληλουχίες σε βιολογικά δείγματα, όπως αυτά του αίματος, που δεν έχουν μελετηθεί σχολαστικά ως προς την βακτηριακή τους σύνθεση. Βέβαια, ένας άλλος λόγος που εντοπίστηκαν μη-ταξινομημένα ASVs στα δείγματα αίματος και όχι σε δείγματα ελέγχου είναι η πιθανότητα να αλληλουχήθηκαν θραύσματα γονιδίων του ξενιστή, το οποίο οφείλεται στην ενίσχυσή τους κατά την εφαρμογή PCR και προετοιμασίας των βιολογικών δειγμάτων.

Επιπρόσθετα, ένα βασικό χαρακτηριστικό της μεθόδου ομαδοποίησης με VSEARCH που φανερώθηκε στην επιλογή τιμών βασικών παραμέτρων των επεξεργαστικών ροών ήταν η παραγωγή πολύ μεγαλύτερου αριθμού αντιπροσωπευτικών αναγνωσμάτων, στην προκειμένη

περίπτωση OTUs, σε σχέση με τις μεθόδους αποθορυβοποίησης ανεξαρτήτως παραμετρικού σεναρίου. Μάλιστα, στα διαθέσιμα δεδομένα της παρούσας διπλωματικής εργασίας, ο VSEARCH κατέληξε να παράγει τουλάχιστον τετραπλάσια ποσότητα ταξινομημένων OTUs σε σχέση με τον αριθμό των ταξινομημένων ASVs του DADA2 και Deblur, όπου στην περίπτωση αυτού προέκυψε ο μικρότερος αριθμός παραγόμενων ταξινομημένων ASVs ανεξαρτήτως παραμετρικού σεναρίου. Αυτό το μοτίβο έχει αναφερθεί και σε άλλες συγκριτικές αναλύσεις των συγκεκριμένων εργαλείων (Nearing et al., 2018), και λογικά λόγω αυτού του υπεραυξημένου όγκου OTUs προκύπτει ο αυξημένος βακτηριακός πλούτος σε σύγκριση με τις μεθόδους αποθορυβοποίησης (**Σχήμα 1.21**), ένα φαινόμενου που παρουσιάστηκε και στην παρούσα μελέτη σχολιάζεται και παρακάτω.

Ολοκληρώνοντας το θέμα των τιμών παραμέτρων, όλος αυτός ο επίπονος πειραματισμός, ο οποίος προτείνεται ανεπιφύλακτα να εκτελείται όχι μόνο στις συγκριτικές αναλύσεις αλλά και σε οποιαδήποτε ανάλυση δεδομένων αλληλούχισης του 16S rRNA γονιδίου που προέρχονται από βιολογικά δείγματα άγνωστης βακτηριακής σύνθεσης, φανέρωσε την ευαισθησία των δεδομένων αλληλούχισης στα διάφορα βιοπληροφορικά εργαλεία και στον τρόπο χρήση τους. Επιπλέον, συνέβαλε σημαντικά στην επιλογή των τιμών βασικών παραμέτρων για κάθε επεξεργαστική ροή και επέτρεψε την διεκπεραίωση της εις βάθους συγκριτικής ανάλυσης των μεθόδων αποθορυβοποίησης με DADA2 και Deblur και ομαδοποίησης με VSEARCH της παρούσας μελέτης με όσον δυνατόν αντικειμενικά κριτήρια. Το γεγονός ότι η διαδικασία της επιλογής των βέλτιστων τιμών παραμέτρων έλαβε μέρος έως το στάδιο της ταξινομικής ανάλυσης σε κάθε επεξεργαστική ροή οφείλεται στο ότι είναι σημείο που επιτρέπει να εκτιμηθεί η απόδοση των βημάτων επεξεργασίας ως προς την διατήρηση όσο τον δυνατόν μεγαλύτερο όγκο πρωτογενών δεδομένων και στην αποκόμιση βέλτιστης ταξινομημένης πληροφορίας. Επιπλέον, η επιλογή των τελικών παραμετρικών σεναρίων και η ανάλυση ενός συνόλου δεδομένων από κάθε επεξεργαστική ροή διευκόλυνε την διαχείριση και την εις βάθος συγκριτική ανάλυση των μετέπειτα αποτελεσμάτων, δεδομένου ότι από το κάθε σύνολο προέκυψε στην συνέχεια πολύ μεγάλος όγκος δεδομένων και πληροφοριών.

Στην κάθε επεξεργαστική ροή με τα τελικώς επιλεγμένα σύνολα δεδομένων, εντοπίστηκαν ASVs και OTUs που ανατέθηκαν στην βάση δεδομένων σε ταξινομική κατηγορία η οποία ούτε χαρακτηρίζεται βακτηριακής προέλευσης, καθώς εμπεριέχονταν αρχαία, μιτοχόνδρια, χλωροπλάστες και ευκαριώτα, αλλά ούτε προσφέρει ευκρινή βακτηριακή ονομασία. Παραδείγματος χάρη κάποια ASVs και OTUs δεν ταξινομήθηκαν σε τουλάχιστον επίπεδο φυλής ή ταξινομήθηκαν σε κατηγορία «μη καλλιεργημένου». Λόγω της φύσης του 16S rRNA γονιδίου, είναι πιθανό να ενισχυθούν, αλληλουχιθούν και ταξινομηθούν αλληλουχίες που προέρχονται από μιτοχόνδρια, χλωροπλάστες και ευκαρυωτικά κύτταρα (de la Cuesta-Zuluaga & Escobar, 2016). Επιπλέον, ενώ επί του παρόντος η βάση δεδομένων SILVA είναι η πιο ολοκληρωμένη, έχει αναφερθεί ότι ένα σημαντικό μειονέκτημα αυτής είναι οι κατοχυρωμένες ταξινομικές κατηγορίες που εμπίπτουν στην κατηγορία του «μη καλλιεργημένου» (uncultured), που ουσιαστικά αποτελούν αλληλουχίες για τις οποίες δεν έχει προσδιοριστεί η ακριβής βακτηριακή προέλευση τους (Trego et al., 2022). Αν και ο όγκος αυτών των στοιχεία στην παρούσα εργασία προέκυψε να είναι πολύ μικρός και στις τρεις επεξεργαστικές ροές (**Πίνακας 3.19**, **Πίνακας 3.21**), είναι σημαντικό να αφαιρεθούν διότι χαρακτηρίζονται αποτέλεσμα σφάλματος στην ανάλυση της περιοχής V3-V4 του 16S rRNA γονιδίου.

Η αδυναμία της ταξινόμησης σε επίπεδο είδους των αλληλουχιών του 16S rRNA γονιδίου (Gao et al., 2017) καθώς και η απουσία της επιμέλειας της ταξινόμησης σε επίπεδο είδους στη βάση δεδομένων SILVA (Trego et al., 2022) έγιναν αντιληπτές στα αποτελέσματα των ταξινομικών κατηγοριών (**Σχήμα 3.8**). Ο αριθμός βακτηριακών ταξινομικών μονάδων στα διάφορα ταξινομικά επίπεδα που προέκυψε και από τις τρεις επεξεργαστικές ροές φανέρωσε, ανεξαρτήτως μεθόδου αποθορυβοποίησης η ομαδοποίησης, την αδυναμία των ASVs και OTUs να ταξινομηθούν σε επίπεδο είδους. Μάλιστα, αυτή η αδυναμία αποδείχθηκε και στην διερεύνηση των κοινών ταξινομικών κατηγοριών σε επίπεδο είδους των τριών επεξεργαστικών ροών (**Σχήμα 3.9**), η οποία φανέρωσε το μειωμένο ποσοστό κοινών ειδών και έτσι υποδεικνύεται και πάλι η αδυναμία των ASVs και OTUs στην ταξινόμηση σε επίπεδο είδους.

Στην συγκριτική ανάλυση των μεθόδων αποθορυβοποίησης και ομαδοποίησης, η επεξεργαστική ροή του VSEARCH κατέληξε να παράγει πολύ μεγαλύτερο αριθμό OTUs σε σχέση με τον αριθμό ASVs στις επεξεργαστικές ροές των DADA2 και Deblur (**Σχήμα 3.7**), τα οποία έφεραν μεγαλύτερο αριθμό μοναδικών ταξινομικών κατηγοριών κυρίως σε επίπεδο γένους και είδους (**Σχήμα 3.8**), όχι βέβαια με την ίδια αναλογία. Προκειμένου να μειωθεί ο όγκος στοιχείων που χαρακτηρίζονται αποτέλεσμα επιμόλυνσης ή σφαλμάτων, επιβλήθηκε φιλτράρισμα ταξινομικών κατηγοριών και στα τρία σύνολα δεδομένων ξεχωριστά που παρουσιαζόταν με πολύ χαμηλή σχετική αφθονία σε αυτά. Αυτό το φίλτρο επηρέασε πολύ έντονα το σύνολο δεδομένων του VSEARCH, καθώς αποδείχθηκε ότι το 85% των παραγόμενων OTUs παρείχαν σχετική αφθονία σίγουρα λιγότερη από 0,002% (**Σχήμα 3.10** Β). Παρόλα αυτά, ύστερα από αυτή τη διαδικασία φιλτραρίσματος και με την απώλεια ενός πολύ μικρού όγκου αναγνωσμάτων (**Σχήμα 3.10** Α), παρατηρήθηκε η αύξηση της ομοιότητας των ταξινομικών πληροφοριών σε τουλάχιστον ύψους 80% μεταξύ και των τριών συνόλων δεδομένων (**Σχήμα 3.12**), ένα φαινόμενο που έχει αναφερθεί και σε άλλη έρευνα (García-López et al., 2021), και η εξομάλυνση της διακύμανσης του αριθμού των ASVs και OTUs και φυσικά των ταξινομικών κατηγοριών (**Σχήμα 3.11**). Με αυτόν τον τρόπο, ο DADA2 κατέληξε να έχει προσδιορίσει τον μικρότερο αριθμό βακτηριακών ειδών στα δεδομένα αλληλούχισης της παρούσας ανάλυσης, ο οποίος ίσως μπορεί να χαρακτηριστικεί πιο ακριβής όπως υποστηρίζεται και από την βιβλιογραφία (Nearing et al., 2018; Prodan et al., 2020).

Οι ταξινομικές συνθέσεις των δειγμάτων πριν και μετά την αφαίρεση του όγκου δεδομένων που αντιπροσωπευόταν από την επιμόλυνση *Lactobacillus* κατέληξαν να μην έχουν καμία διαφορά μεταξύ αποτελεσμάτων των τριών επεξεργαστικών ροών. Συγκεκριμένα, οι σχετικές αφθονίες των βακτηριακών ταξινομικών κατηγοριών ανά δείγμα σε επίπεδο φυλής (**Παράρτημα 46**, **Παράρτημα 47**, **Παράρτημα 48**) και γένους (**Σχήμα 3.13, Σχήμα 3.14, Σχήμα 3.15**) πριν το φιλτράρισμα των *Lactobacillus*, καθώς και οι σχετικές αφθονίες των βακτηριακών ταξινομικών κατηγοριών ανά δείγμα σε επίπεδο φυλής ύστερα του φιλτραρίσματος (**Σχήμα 3.18, Σχήμα 3.19, Σχήμα 3.20**) παρουσίασαν πολύ παρόμοια εικόνα, παρόλο που τα σύνολα παρείχαν διαφορετικό αριθμό αναγνωσμάτων και ASVs/OTUs. Βέβαια, η αφαίρεση της επιμόλυνσης επέφερε αναλογική επίπτωση στα αποτελέσματα της κάθε επεξεργαστικής ροής, όπως η ποσοστιαία απώλεια των αναγνωσμάτων (**Σχήμα 3.17**) και η εξομάλυνση της διακύμανσης του αριθμού αναγνωσμάτων ανά δείγμα (**Σχήμα 3.16**). Επιπρόσθετα, οι μέσες σχετικές αναλογίες των βακτηριακών φυλών στο αίμα σχιζοφρενών που παρουσίασαν το πρώτο ψυχωσικό επεισόδιο οι οποίες προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH δεν ανέδειξαν σημαντικές διαφορές ως προς την ανίχνευση και την σχετική ποσοτικοποίηση των βακτηριακών ταξινομικών κατηγοριών σε επίπεδο φυλής (**Σχήμα 3.21**).

Αντίθετα, οι καμπύλες αραίωσης αποτύπωσαν ξεκάθαρα την διαφοροποίηση μεταξύ των μεθόδων αποθορυβοποίησης και ομαδοποίησης. Οι καμπύλες αυτές χρησιμοποιούνται για τον προσδιορισμό του ιδανικού βάθους αλληλούχισης ώστε να καλυφθεί όλη η ποικιλομορφία όλων των δειγμάτων ταυτόχρονα (Regueira-Iglesias et al., 2023). Ενώ η κάλυψη της ποικιλομορφίας των δειγμάτων από την επεξεργαστική ροή του DADA2 και Deblur προέκυψε σε βάθος αλληλούχισης ύψους 1800 αναγνωσμάτων, στα δεδομένα του VSEARCH καθως παρουσιάστηκε η συνεχής άνοδος του αριθμού των ταξινομικών ειδών που παρέχει το κάθε δείγμα έναντι της τιμής του βάθους αλληλούχισης (**Σχήμα 3.22**), δεν ήταν αδύνατο να προσδιοριστεί το ιδανικό βάθος αλληλούχισης. Επιπλέον, οι μέσες τιμές της εντροπίας Shannon άλφα ποικιλομορφίας του συνόλου δεδομένων VSEARCH προσδιορίστηκαν αυξημένες σε σχέση με τις αντίστοιχες των μεθόδων αποθορυβοποίησης, ειδικά οι τιμές των δειγμάτων αρνητικού ελέγχου (**Σχήμα 3.23**). Παρόλα αυτά, τα αποτελέσματα του στατιστικού ελέγχου των διαφοροποιήσεων ή μη των ποικιλομορφιών μεταξύ των δειγμάτων αίματος και αρνητικού ελέγχου είχαν κοινή γραμμή μεταξύ των διαφορετικών επεξεργαστικών ροών.

4.2 DADA2, Deblur ή VSEARCH;

Η συγκριτική ανάλυση των μεθόδων αποθορυβοποίησης και ομαδοποίησης της παρούσας διπλωματικής εργασίας κατέληξε να έρχεται σε συμφωνία με τις προϋπάρχουσες έρευνες συγκριτικής ανάλυσης (García-López et al., 2021; Glassman & Martiny, 2018; Joos et al., 2020; Kerrigan & D'Hondt, 2022; Moossavi et al., 2020; Nearing et al., 2018; Prodan et al., 2020). Όπως και τα ευρήματα αυτών, έτσι και στην παρούσα μελέτη οι μέθοδοι παραγωγής ASVs και OTUs παρουσίασαν παρόμοια βιολογικά αποτελέσματα ως προς την ταξινομική ανίχνευση και σύνθεση των δεδομένων αλληλούχισης που προήλθαν από φυσικά βιολογικά δείγματα. Ταυτόχρονα, φανερώθηκε το κύριο χαρακτηριστικό των προσεγγίσεων που βασίζονται στην ομαδοποίηση των αναγνωσμάτων σε OTUs, το οποίο είναι ο αυξημένος βακτηριακός πλούτος σε σύγκριση με τις μεθόδους αποθορυβοποίησης. Δεδομένου ότι αυτό το χαρακτηριστικό εμπεριέχει τον κίνδυνο της ανακριβής ανίχνευσης της άλφα ποικιλομορφίας των βιολογικών δειγμάτων, όπως επίσης ότι βασίζεται πολύ στην ποιότητα και πληρότητα της βάσης δεδομένων και ότι οι προσεγγίσεις παραγωγής ASVs μπορούν να παρέχουν ένα σημαντικό πλεονέκτημα στην ακριβέστερη ταυτοποίηση μικροοργανισμών (Caruso et al., 2019; Needham et al., 2017; Prodan et al., 2020; Xue et al., 2018), όπως και η επιστημονική κοινότητα, έτσι και η παρούσα διπλωματική εργασία υποστηρίζει την χρήση μεθόδων αποθορυβοποίησης και παραγωγής ASVs έναντι της χρήση μεθόδων ομαδοποίησης και παραγωγής OTUs, όπως τον VSEARCH, για την ανάλυση πρωτογενών δεδομένων αλληλούχισης του 16S rRNA γονιδίου.

Μάλιστα, η μέθοδος αποθορυβοποίησης του DADA2 φάνηκε να έχει πολύ καλύτερη απόδοση και ως προς τον τρόπο διαχείρισης δεδομένων αλληλούχισης και ως προς την αξιοπιστία του σε σχέση με την αντίστοιχη μέθοδο του Deblur. Αρχικά, σε αντίθεση με του Deblur, η επεξεργαστική ροή του DADA2 εκτελεί την διαδικασία καθαρά υπολογιστικά και δεν χρησιμοποιείται καμία βάση δεδομένων, ακόμα και για τον εντοπισμό των χιμαιρικών αλληλουχιών. Αυτό σημαίνει ότι είναι απαλλαγμένο από την ποιότητα και πληρότητα των βάσεων δεδομένων και κατά επέκταση την μεροληψία που μπορεί να προσδέσουν αυτά στην προεπεξεργασία των δεδομένων αλληλούχισης. Επίσης, η εφαρμογή του DADA2 δεν βασίζεται σε συγκεκριμένο μήκος αναγνωσμάτων, κάτι που απαιτείται να οριστεί για την

αποθορυβοποίηση του Deblur. Επιπλέον, η αποθορυβοποίηση με DADA2 εξετάζονται τα δεδομένα αλληλούχισης όλων των δειγμάτων ταυτόχρονα επιτρέποντας την πιο αντικειμενική αξιολόγηση αυτών, σε αντίθεση με τον Deblur ο οποίος εστιάζει την διαδικασία μόνο ανά δείγμα. Φυσικά, η επιλογή της περικοπής των τελικών άκρων των paired-end, η οποία είναι πολύ υποσχόμενη διαδικασία για την βελτίωση της απόδοσης στην παραγωγής ASV, προσφέρεται από το QIIME2 μόνο από την χρήση του pipeline DADA2. Σε συνδυασμό με την διευκόλυνση που παρέχεται από την πλατφόρμα QIIME2 ως προς την χρήση ενός ενιαίου pipeline που συμπεριλαμβάνει την εκτέλεση της περικοπής paired-end αναγνωσμάτων, του φιλτραρίσματος χαμηλής ποιότητας αυτών, της αποθορυβοποίησης, του εντοπισμού και της αφαίρεσης χιμαιρικών αλληλουχιών και της συγχώνευσης των paired-end αναγνωσμάτων, καθιστά την εφαρμογή του DADA2 μέσω της πλατφόρμας QIIME2 πολύ πιο εύκολη στην χρήση (user-friendly). Το γεγονός ότι ο DADA2 κατάφερε να διατηρήσει μεγαλύτερο όγκο πρωτογενών δεδομένων στο τέλος της επεξεργαστικής ροής σε σχέση με τον Deblur, ο οποίος μάλιστα κατέληξε να διατηρεί λιγότερα από 10000 αναγνώσματα σε μερικά δείγματα, ευνοεί ακόμα περισσότερα την προτίμηση της χρήσης του DADA2. Τέλος, ενώ η επεξεργαστική ροή του Deblur κατάφερε να ανιχνεύσει μεγαλύτερο αριθμό ταξινομικών μονάδων, όπως έχει προαναφερθεί, ο μειωμένος αριθμός βακτηριακών γενών και ειδών που προέκυψε από τον DADA2 υπάρχει μεγάλη περίπτωση να είναι πιο ακριβής και κατά επέκταση πιο αξιόπιστος.

4.3 Συσχέτιση Μικροβιώματος του Αίματος και Σχιζοφρένειας

Τα ευρήματα της παρούσας ανάλυσης που σχετίζονται με το μικροβίωμα του αίματος σχιζοφρενών προέκυψαν πολύ ενδιαφέρον και έριξαν λίγο φως στο κενό αυτού του ερευνητικού κλινικού πεδίου. Στην παρούσα διπλωματική εργασία τα δείγματα αίματος από άτομα που παρουσίασαν το πρώτο ψυχωσικό επεισόδιο και διαγνώστηκαν με σχιζοφρένεια προέκυψε να κυριαρχούνται από *Proteobacteria* (**Σχήμα 3.21**), όπως και στο υγιές αίμα (Païssé et al., 2016; Velmurugan et al., 2020). Ωστόσο, παρατηρήθηκε πολύ αυξημένη η σχετική αφθονία των *Firmicutes* σε σχέση με την αντίστοιχη των δειγμάτων αίματος από υγιή άτομα, ενώ τα *Actinobacteria* και τα *Bacteroidetes* προέκυψαν σε παρόμοια επίπεδα. Η αυξημένη παρουσία των *Firmicutes* έχει αναφερθεί και στο εντερικό μικροβίωμα ατόμων που πάσχουν από σχιζοφρένεια (Szeligowski et al., 2020), γεγονός ότι διασταυρώνεται και τα αποτελέσματα της παρούσας ανάλυσης. Δεδομένου ότι ένα από τα πιο χαρακτηριστικά φαινόμενα της δυσβίωσης του άξονα εντέρου-εγκεφάλου αποτελεί η βακτηριακή μετατόπιση της μικροχλωρίδας από το έντερο στο κυκλοφορικό σύστημα (Munawar et al., 2021), γεγονός που έχει αποδειχθεί ότι συμβαίνει σε άτομα που πάσχουν σχιζοφρένεια (Severance et al., 2013), υπάρχει πιθανότητα η αυξημένη σχετική αφθονία των *Firmicutes* να οφείλεται από αυτό το φαινόμενο.

Έτσι, η παρούσα μελέτη οδηγείται σε μια απλή αλλά βασική ερώτηση. Τελικά, υπάρχει μικροβίωμα του αίματος, και αν ναι, πως είναι δυνατόν να ενισχυθεί η ύπαρξή του μέσω της παρούσας διπλωματικής εργασίας; Βάση βιβλιογραφίας, ένας τρόπος εξακρίβωσης της ύπαρξης μικροβιώματος στο αίμα είναι η σύγκριση του ταξινομικής σύνθεσης και τις ποικιλομορφίας των δειγμάτων αίματος με τα δείγματα αρνητικού ελέγχου (Dinakaran et al., 2014; Lauder et al., 2016; Païssé et al., 2016; Traykova et al., 2017). Στην παρούσα μελέτη, στα αποτελέσματα και των τριών επεξεργαστικών ροών της ταξινομικής σύνθεσης σε επίπεδο φυλής (**Σχήμα 3.18, Σχήμα 3.19, Σχήμα 3.20**) και των δεικτών άλφα ποικιλομορφίας (**Σχήμα 3.23**) δεν παρουσιάζουν σημαντικές διαφορές μεταξύ των δειγμάτων αίματος και των

δειγμάτων αρνητικού ελέγχου. Αυτό υποδεικνύει το πόσο επιρρεπές είναι τα δείγματα αίματος στις επιμολύνσεις (de la Cuesta-Zuluaga & Escobar, 2016). Άλλωστε, είναι γνωστό ότι οτιδήποτε χρησιμοποιείται κατά την πειραματική διαδικασία της αλληλούχισης νέας γενιάς δεν είναι ποτέ πραγματικά στείρο, καθώς έχουν ανιχνευθεί πάνω από 90 διαφορετικά μικρόβια σε επίπεδο γένους μόνο και μόνο σε αντιδραστήρια απομόνωσης DNA και προετοιμασίας δειγμάτων προς αλληλούχιση (Lauder et al., 2016; Salter et al., 2014). Βέβαια, ενδιαφέρον αποτελεί η παρατήρηση ότι στην επεξεργαστική ροή του DADA2 παρουσιάστηκε η μόνη διαφοροποίηση των δειγμάτων αίματος και των δειγμάτων αρνητικού ελέγχου. Συγκεκριμένα, κατά την ταξινομική ανάθεση των αποτελεσμάτων αποθορυβοποίησης του DADA2 προέκυψε η ανικανότητα να ταξινομηθεί ένας όγκος δεδομένων, όπου ανάλογα το παραμετρικό σενάριο είναι της τάξεως του 2% με 14% σε σχέση με τα πρωτογενή δεδομένα, στην βάση δεδομένων SILVA (**Πίνακας 3.15**). Αυτός ο μη-ανατεθειμένος όγκος δεδομένων εντοπίστηκε μόνο στα δείγματα αίματος (**Παράρτημα 5**, **Παράρτημα 6**, **Παράρτημα 7**, **Παράρτημα 8**, **Παράρτημα 9**, **Παράρτημα 10**, **Παράρτημα 25**, **Παράρτημα 26**). Ενώ αυτό το φαινόμενο μπορεί να οφείλεται στην σφαλμένη αλληλούχιση ανθρώπινου γονιδίου, υπάρχει σοβαρή πιθανότητα να οφείλεται στην μη πληρότητα της βάσης δεδομένων SILVA και στην κάλυψη της βακτηριακής σύνθεσης του αίματος, και συνεπώς η μη διαφοροποίηση των δειγμάτων αίματος και του ελέγχου στην παρούσα μελέτη να βασίαζεται σε αυτό το γεγονός. Εν κατακλείδι, η ύπαρξη του μικροβιώματος του ανθρώπινου αίματος παραμένει ακόμα υπό διερεύνηση λόγω των προκλήσεων που περιστρέφονται γύρω από δύο θέματα, δηλαδή τον υψηλό κίνδυνο μικροβιακής επιμόλυνσης σε δείγματα χαμηλής βιομάζας και την απροσδιόριστη βιωσιμότητα των μικροβίων του αίματος που βασίζεται σε μεθόδους ανεξαρτήτως καλλιέργειας (Cheng et al., 2023).

Η ευαισθησία των δειγμάτων αίματος στις εξωγενής επιμολύνσεις παρουσιάστηκε στην παρούσα εργασία και με έναν ακόμα τρόπο. Στο αίμα 4 ασθενών πριν και μετά την χορήγηση αντιψυχωσικών φαρμάκων εντοπίστηκε σε μεγάλη σχετική αφθονία η ταξινομική κατηγορία γένους *Lactobacillus* στα αποτελέσματα και των τριών επεξεργαστικών ροών (**Σχήμα 3.13, Σχήμα 3.14, Σχήμα 3.15**). Δεδομένου της σημαντικά αυξημένης σχετικής αφθονίας, τον μεγάλο αριθμό αναγνωσμάτων που παρείχαν τα συγκεκριμένα δείγματα και το γεγονός ότι οδήγησε στην εξομάλυνση της ταξινομικής σύνθεσης των αυτών των δειγμάτων με τα υπόλοιπα δείγματα αίματος κατά την αφαίρεση του από τα δεδομένα, στην παρούσα μελέτη η συγκεκριμένη ταξινομική κατηγορία χαρακτηρίστηκε αποτέλεσμα επιμόλυνσης. Πράγματι, υπάρχουν έρευνες του υποστηρίζουν την παρουσία των *Lactobacillus* ως αποτέλεσμα επιμόλυνσης συγκεκριμένα σε δείγματα αίματος, και το γεγονός ότι κυριαρχούν στη ταξινομική σύνθεση των δειγμάτων αίματος που κατηγοριοποιούνται σε δείγματα χαμηλής μικροβιακής βιομάζας, ενισχύεται η θεωρία ότι τα *Lactobacillus* στην παρούσα εργασία χαρακτηρίζονται μολυσματικά στα δεδομένα (Karstens et al., 2019; Kullar et al., 2023). Παρόλα αυτά, είναι σημαντικό να σημειωθεί ότι ένα από τα σταθερά ευρήματα είναι μια σημαντική αύξηση των *Lactobacillius* στο μικροβίωμα του στοματοφάρυγγα και στο εντερικό μικροβίωμα ασθενών με σχιζοφρένεια και σε άτομα με αυξημένο κίνδυνο σχιζοφρένειας, η οποία συσχετίστηκε ακόμη και με τη σοβαρότητα των συμπτωμάτων (Szeligowski et al., 2020). Δεδομένου ότι ένα από τα πιο χαρακτηριστικά φαινόμενα της δυσβίωσης του άξονα εντέρου-εγκεφάλου αποτελεί η βακτηριακή μετατόπιση της μικροχλωρίδας από το έντερο στο κυκλοφορικό σύστημα, υπάρχει η πιθανότητα η παρουσία των *Lactobacillius* στα δείγματα αίματος να είναι αποτέλεσμα της αυξημένης διαπερατότητας του φραγμού του εντέρου, ένα φαινόμενο που έχει συσχετιστεί με την σχιζοφρένεια (Munawar et al., 2021). Ως εκ τούτου, είναι σημαντικό να διερευνηθεί περισσότερο ο ρόλος

των *Lactobacillius* στην σχιζοφρένεια, καθώς παρότι στην παρούσα διπλωματική έχει χαρακτηριστεί ως επιμόλυνση, η παρουσία των βακτηρίων του γένους *Lactobacillus*.

5 Συμπεράσματα

Η διεκπεραίωση της παρούσας διπλωματικής εργασίας επιβεβαίωσε την πολυπλοκότητα της ανάλυσης paired-end δεδομένων αλληλούχισης του 16S rRNA γονιδίου, όπου στην προκειμένη περίπτωση ήταν η διερεύνηση των ταξινομικών πληροφοριών της υπερμεταβλητής περιοχής V3-V4, τα οποία έχουν προκύψει από πλατφόρμα αλληλούχισης Illumina. Ακόμα και με την χρήση της πλατφόρμας QIIME2, η οποία αποσκοπεί στην διευκόλυνση της χρήσης βιοπληροφορικών εργαλείων χωρίς την ανάγκη κατάρτισης γνώσεων δύσκολής γλώσσας προγραμματισμού, υπάρχει πλήθος διαφορετικών επιλογών και συνδυασμών βιοπληροφορικών εργαλείων που μπορούν να εφαρμοστούν και από τα οποία προκύπτουν διαφορετικές επεξεργαστικές ροές. Η επιλογή μεθόδων αποθορυβοποίησης ή ομαδοποίησης των αναγνωσμάτων, από τις οποίες βασίζεται κατά κόρων ο τρόπος κατασκευής της επεξεργαστικής ροής του προεπεξεργαστικού σταδίου, μπορεί εν δυνάμει να έχει αντίκτυπο στα τελικά αποτελέσματα, και κατά επέκταση στην βιολογική ερμηνεία των δεδομένων. Επιπλέον, όχι μόνο η επιλογή των βιοπληροφορικών εργαλείων, αλλά και ο τρόπος με τον οποίο χρησιμοποιούνται μπορεί να επηρεάσει τα αποτελέσματα. Αποτελεί αντίστοιχα σημαντική και η επιλογή των τιμών παραμέτρων βασικών επεξεργαστικών σταδίων, καθώς μικρές διαφοροποιήσεις μπορούν να επιφέρουν σημαντικές αλλαγές στην κατάσταση των δεδομένων.

Στα πλαίσια της παρούσας μελέτης, για την εφαρμογή των μεθόδων αποθορυβοποίησης του DADA2 και Deblur και ομαδοποίησης κλειστής αναφοράς του VSEARCH με την βάση δεδομένων SILVA έπρεπε να κατασκευαστεί μια επεξεργαστική ροή για το κάθε εργαλείο, κατά την οποία τα πρωτογενή δεδομένα επεξεργάστηκαν ανάλογα με τις ανάγκες και απαιτήσεις της κάθε μεθόδου, καθώς και με την σωστή σειρά. Οι κοινές απαιτούμενες παρεμβάσεις πριν την παραγωγή ASVs/OTUs στα συσχετιζόμενα πρωτογενή δεδομένα ήταν η αφαίρεση των μη-βιολογικών εκκινητών στις αρχικές αλληλουχίες των paired-end αναγνωσμάτων, η συγχώνευση και το ποιοτικό φιλτράρισμα αυτών, με τη μόνη διαφορά στην περίπτωση του DADA2 που η συγχώνευση και το φιλτράρισμα με βάση τη ποιότητα εκτελέστηκαν με διαφορετική σειρά. Επιπλέον, υπήρχε η δυνατότητα περικοπής των τελικών αλληλουχιών από τα paired-end αναγνώσματα στην περίπτωση του DADA2, ενώ πριν την αποθορυβοποίηση του Deblur ήταν αναγκαία η αποκοπή των συγχωνευμένων αναγνωσμάτων σε ένα συγκεκριμένο μήκος. Εφόσον τα πρωτογενή δεδομένα προέρχονταν κυρίως από δείγματα ανθρώπινου αίματος άγνωστου βακτηριακού περιεχόμενου, ήταν αναγκαίο να ληφθεί υπόψη τον αντίκτυπο που θα μπορούσε να έχει σε αυτά η περικοπή ή μη των paired-end αναγνωσμάτων, οι προδιαγραφές συγχώνευσης αυτών ως προς την επικαλυπτόμενη περιοχή, η απαιτούμενη ελάχιστη ποιότητα των αναγνωσμάτων στα τελικά αποτελέσματα και το σημείο αποκοπής των συγχωνευμένων αναγνωσμάτων.

Ενώ οι τεχνικές όπως η μερική περικοπή των αρχικών ή τελικών αλληλουχιών, η συγχώνευση και το φιλτράρισμα χαμηλής ποιότητας των paired-end ή συγχωνευμένων αναγνωσμάτων βελτίωσαν την ποιότητα των πρωτογενών δεδομένων, η προσαρμογή των παραμέτρων προεπεξεργασίας έπαιξε σημαντικό ρόλο στην παρούσα Διπλωματική Εργασία. Η προσπάθεια επιλογής παρόμοιων παραμετρικών σεναρίων για κάθε επεξεργαστική ροή στο στάδιο του πειραματισμού, όπου αυτή ήταν εφικτή, και η εξέταση των αποτελεσμάτων μέχρι και το στάδιο της ταξινομικής ανάθεσης των αναγνωσμάτων, επέτρεψε την αξιολόγηση της επιρροής των παραμετρικών συνθηκών στα δεδομένα, τον διαχωρισμό της επιρροής αυτών και των μεθόδων αποθορυβοποίησης και ομαδοποίησης και φυσικά την επιλογή παρόμοιων τελικών παραμετρικών τιμών και συνθηκών για κάθε επεξεργαστική ροή για την εις βάθος

ταξινομική ανάλυση και διερεύνηση ποικιλομορφίας των δειγμάτων. Με αυτό τον τρόπο, πραγματοποιήθηκε με επιτυχία ο πρωταρχικός σκοπός που αφορά την συγκριτική ανάλυση των μεθόδων παραγωγής ASVs/OTUs με διαφάνεια και αξιοπιστία.

Τα αποτελέσματα έδειξαν ότι οι διάφορες προδιαγραφές συγχώνευσης σε κάθε επεξεργαστική ροή, δηλαδή οι διάφορες τιμές ελάχιστου απαιτούμενου μήκος επικαλυπτόμενης περιοχής ($Overlap_{min}$=10, 20, 30), ανεξαρτήτως σειράς εκτέλεσης αυτής της διαδικασίας, δεν επηρέασαν τα δεδομένα αλληλούχισης και τις ταξινομικές πληροφορίες, εφόσον αναδείχθηκε ότι το μήκος επικαλυπτόμενης περιοχής της συντριπτικής πλειοψηφίας των αμπλικονίων είναι ίσο κατά προσέγγιση με 40 bp. Η περικοπή των τελικών αλληλουχιών των paired-end αναγνωσμάτων στην επεξεργαστική ροή του DADA2 φανέρωσε την παραγωγή ελαφρώς υψηλότερης αξιοπιστίας ASVs σε αντίθεση με την επιλογή μη αποκοπής αυτών. Στην περίπτωση του Deblur, η διατήρηση μεγαλύτερου μήκους αναγνωσμάτων αύξησε την απόδοση των παραγόμενων ASVs στην ταξινομική ανάθεση. Όσον αφορά στην διαδικασία του ποιοτικού φιλτραρίσματος, στην περίπτωση των επεξεργαστικών ροών του Deblur και VSEARCH τα συγχωνευμένα αναγνώσματα αξιολογήθηκαν με βάση την ελάχιστη βαθμολογία ποιότητας βάσης (Q_{min}=20, 22, 26), ενώ στην περίπτωση του DADA2 η αξιολόγηση των paired-end αναγνωσμάτων πραγματοποιήθηκε με βάση το μέγιστο ποσοστό αναμενόμενων σφαλμάτων που παρέχουν αυτά σαν σύνολο ($e.e._{max}$= 2.5, 1.5, 0.5). Η επιλογή των συγκεκριμένων τιμών βασίστηκε στην παραδοχή ότι με βάση αυτές τα αναγνώσματα αξιολογούνται με παρόμοια κριτήρια, η οποία επιβεβαιώθηκε από τις παραλλαγές των αριθμών παραγόμενων και ταξινομημένων ASVs και OTUs και των αριθμών ταξινομικών κατηγοριών που προέκυψαν από κάθε επεξεργαστική ροή. Η κύρια διάκριση στη διαχείριση της ποιότητας των αναγνωσμάτων προέκυψε να έγκειται στον αριθμό διατηρητέων αυτών μέχρι και την ταξινομική ανάθεση, όπου κατά την μείωση του $e.e._{max}$ παρατηρήθηκε μια σχετικά γραμμική μείωση του όγκου διατηρητέων αναγνωσμάτων, ενώ στην περίπτωση της αύξησης του Q_{min} αναδείχθηκε η μηδενική επιρροή στον όγκο αυτό μεταξύ των τιμών Q_{min}=20 και 22. Ωστόσο, ανεξαρτήτως της μεθόδου ποιοτικού φιλτραρίσματος, η συγκράτηση αναγνωσμάτων υψηλής ποιότητας μείωσε σημαντικά τον όγκο των διατηρητέων αναγνωσμάτων, τις ταξινομικές πληροφορίες και ενδεχομένως τις πληροφορίες σχετικά με την ποικιλομορφία δειγμάτων.

Ο πειραματισμός των παραμετρικών τιμών και συνθηκών μέχρι την ταξινομική ανάθεση των ASVs και των OTUs αποκάλυψε διακριτά μοτίβα, τονίζοντας τις διαφορές στον αντίκτυπο των μεθόδων του DADA2, Deblur και VSEARCH στα αρχικά στάδια της ανάλυσης του 16S rRNA γονιδίου. Η αποθορυβοποίηση με Deblur οδήγησε στην σημαντική απώλεια αναγνωσμάτων ανεξαρτήτως παραμετρικών σεναρίων σε σύγκριση με την αποθορυβοποίηση DADA2 και ομαδοποίηση VSEARCH, στις οποίες παρουσιάστηκε ο διατηρητέος όγκος αναγνωσμάτων να επηρεάζεται από το στάδιο ποιοτικού φιλτραρίσματος. Από την αποθορυβοποίηση με DADA2 προέκυψε ένας σημαντικός όγκος αναγνωσμάτων, αποκλειστικά από τα δείγματα αίματος, που δεν κατάφερε να ταξινομηθεί, υποδεικνύοντας ότι οι αλληλουχίες των μη-ανατεθειμένων ASVs μπορεί να προέρχονται από βακτήρια που δεν έχουν ακόμη καταγραφεί στη βάση δεδομένων SILVA ή, εναλλακτικά, από θραύσματα ανθρώπινου γονιδίου που έχουν ενισχυθεί κατά την προετοιμασία του δείγματος. Αντίθετα, το Deblur και το VSEARCH έδειξαν πλήρη ταξινόμηση, πιθανότατα λόγω της εξάρτησής τους από τις βάσεις δεδομένων κατά την αποθορυβοποίηση και την ομαδοποίηση. Σε κάθε παραμετρικό σενάριο, η ομαδοποίηση VSEARCH παρήγαγε έναν σημαντικά υψηλότερο αριθμό OTUs, περίπου τέσσερις φορές μεγαλύτερο από τον αριθμό ταξινομημένων ASVs που δημιουργήθηκαν από το DADA2 και το Deblur.

Η επιλογή των τελικών τιμών παραμέτρων του προεπεξεργαστικού σταδίου έλαβε υπόψη τον όγκο διατηρητέων ταξινομημένων αναγνωσμάτων, ASVs/OTUs και τον αντίστοιχο όγκο προκυπτόντων ταξινομικών κατηγοριών. Οι βέλτιστες παράμετροι στόχευσαν στην μεγαλύτερη αναπαράσταση της βάσης δεδομένων με το λιγότερο δυνατό αριθμό ASVs/OTUs, με το μέγιστο αριθμό αξιόπιστων διατηρητέων αναγνωσμάτων και παράλληλα στην επιλογή παρόμοιων παραμετρικών σεναρίων μεταξύ των τριών διαφορετικών επεξεργαστικών ροών. Λόγω της αμελητέας επιρροής των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής, επιλέχθηκε και για της τρεις επεξεργαστικές ροές για την διαδικασία της συγχώνευσης $Overlap_{min}$=10. Παρόλο που εντοπίστηκαν σοβαρές ενδείξεις παραγωγής μεγαλύτερης απόδοσης ASVs στα σύνολα δεδομένων του DADA2 που εφαρμόστηκε περικοπή των τελικών αλληλουχιών των paired-end αναγνωσμάτων, τα αποτελέσματα χωρίς αποκοπή παρουσίασαν μεγαλύτερο αριθμό διατηρητέων αναγνωσμάτων, ASVs και ταξινομικές κατηγορίες αντίστοιχα. Η επιλογή της με αποκοπής των paired-end αναγνωσμάτων επέτρεψε εξίσου την διατήρηση της συνέπειας με τις άλλες δύο επεξεργαστικές ροές, στις οποίες δεν ήταν διαθέσιμη αυτή η επιλογή. Για τον Deblur, εκτός από την απόδοση μεγαλύτερης σαφήνειας στην βιολογική προέλευση των παραγόμενων ASVs, η επιλογή του μήκους αναγνωσμάτων ίσο με 380 bp έφερε σαν αποτέλεσμα να φέρουν και οι τρεις επεξεργαστικές ροές κατά την πλειοψηφία τους παρόμοιο μήκος αναγνωσμάτων. Ως προς τις προδιαγραφές του ποιοτικού ελέγχου, για τον DADA2 επιλέχθηκε το σύνολο δεδομένων με $e.e._{max}$=1,5 και για τον Deblur και VSEARCH αντίστοιχα με Q_{min}=22.

Η εις βάθος συγκριτική ταξινομική ανάλυση των τελικών επιλεγμένων συνόλων δεδομένων της κάθε μεθόδου αποθορυβοποίησης και ομαδοποίησης φανέρωσε την αδυναμία των αλληλουχιών του 16S rRNA στην ταξινομική ανάθεση και την έλλειψη επιμέλειας της βάση δεδομένων SILVA, τα οποία οδήγησαν στην αξιοσημείωτη αποτυχία ταξινόμησης των ASVs και των OTUs σε επίπεδο είδους. Επιπλέον, η διερεύνηση κοινών ταξινομικών κατηγοριών μεταξύ των τριών συνόλων ανέδειξε εξίσου την αναξιοπιστία των ASVs και των OTUs στην ταξινόμηση σε επίπεδο είδους. Το VSEARCH παρήγαγε μεγαλύτερο αριθμό OTUs από αριθμό ASVs των DADA2 και Deblur, από τα οποία προέκυψε μεγαλύτερος αριθμός μοναδικών ταξινομικών κατηγοριών εξίσου. Το φιλτράρισμα ταξινομικών κατηγοριών με σχετική αφθονία λιγότερη από 0,002% στο κάθε σύνολο δεδομένων ξεχωριστά, μείωσε τα στοιχεία χαμηλής αφθονίας που θεωρητικά είναι αποτέλεσμα σφαλμάτων και επιμολύνσεων, ειδικά στο σύνολο δεδομένων του VSEARCH, και οδήγησε σε αυξημένη ταξινομική ομοιότητα μεταξύ όλων των συνόλων δεδομένων. Ο Deblur κατέληξε να φέρει τον μεγαλύτερο αριθμό ταξινομικών κατηγοριών, σε αντίθεση με τον DADA2 που εντόπισε τα λιγότερα είδη, που όμως μπορεί να παρέχει πιο ακριβή αποτελέσματα, καθώς πρόκειται για ανάλυση δειγμάτων χαμηλής μικροβιακής μάζας. Οι ταξινομικές συνθέσεις των δειγμάτων πριν και μετά την αφαίρεση της επιμόλυνσης *Lactobacillus* δεν έδειξαν σημαντικές διαφορές μεταξύ των τριών ροών επεξεργασίας. Παρά τους ποικίλους αριθμούς αναγνωσμάτων και ASVs/OTUs, τα αποτελέσματα σε ταξινομικά επίπεδα φυλής και γένους παρέμειναν παρόμοια. Η αφαίρεση επιμόλυνσης είχε αναλογικές επιπτώσεις σε κάθε ροή, επιφέροντας την απώλεια αναγνωσμάτων και εξομάλυνση της διακύμανσης του αριθμού αναγνωσμάτων ανά δείγμα. Μεταξύ των DADA2, Deblur και VSEARCH δεν παρουσιάστηκαν παραλλαγές στις μέσες βακτηριακές αναλογίες σε επίπεδο φυλής στο αίμα των σχιζοφρενών που παρουσίασαν το πρώτο ψυχωσικό επεισόδιο.

Οι καμπύλες αραίωσης τόνισαν τη διάκριση μεταξύ των μεθόδων αποθορυβοποίησης και ομαδοποίησης. Ο DADA2 και ο Deblur πέτυχαν την κάλυψη της ποικιλομορφίας των

δειγμάτων σε βάθος αλληλούχισης ύψους 1800 αναγνωσμάτων, ενώ ο VSEARCH παρουσίασε συνεχή αύξηση των παρατηρούμενων ταξινομικών κατηγοριών συναρτήσει του βάθος αλληλούχισης. Οι μέσες τιμές της άλφα ποικιλομορφίας Shannon στο VSEARCH ήταν υψηλότερες από τις μεθόδους αποθορυβοποίησης. Ωστόσο, οι στατιστικοί έλεγχοι έδειξαν σταθερά μοτίβα στην διερεύνηση διαφοροποιήσεων της ποικιλομορφίας μεταξύ των δειγμάτων αίματος και αρνητικού ελέγχου και στις τρείς επεξεργαστικές ροές. Ως εκ τούτου, όπως και προηγούμενες έρευνες, έτσι και η παρούσα μελέτη υποστήριξε την χρήση μεθόδων αποθορυβοποίησης έναντι των μεθόδων ομαδοποίησης για την ανάλυση δεδομένων αλληλούχισης του 16S rRNA γονιδίου. Ενώ οι μέθοδοι παραγωγής ASVs και OTUs κατάφεραν να φέρουν παρόμοια βιολογικά αποτελέσματα στην ταξινομική ανίχνευση φυσικών δειγμάτων, φανερώθηκε ότι οι προσεγγίσεις ομαδοποίησης καταλήγουν σε αυξημένο πλούτο βακτηρίων σε σύγκριση με τις μεθόδους αποθορυβοποίησης. Μάλιστα, ο DADA2 προέκυψε πιο αποδοτικός από τον Deblur όσον αφορά την υπολογιστική απόδοση, την ανεξαρτησία από τις βάσεις δεδομένων και την ευελιξία στο μήκος αναγνωσμάτων.

Εν κατακλείδι, το μικροβίωμα του αίματος σχιζοφρενών που βίωσαν το πρώτο ψυχωσικό επεισόδιο προέκυψε να κυριαρχείται από *Proteobacteria*, με αύξηση των *Firmicutes* σε σύγκριση με υγιή άτομα βάσει βιβλιογραφίας. Ωστόσο, είναι πολύ σημαντικό να αναφερθεί ότι η παρούσα εργασία αντιμετώπισε σοβαρές προκλήσεις που σχετίζονται με την επιμόλυνση των δειγμάτων. Η συμπερίληψη δειγμάτων αρνητικού ελέγχου καθώς και ο εντοπισμός ακραίας μεγάλης σχετικής αφθονίας των *Lactobacillus* σε σημαντικό αριθμό δειγμάτων αίματος υπέδειξαν την ευαισθησία των δειγμάτων αίματος σε εξωγενείς επιμολύνσεις.

6 Μελλοντικές Προτάσεις

Οι μελλοντικές προτάσεις για έρευνα που βασίζονται στην τρέχουσα Διπλωματική Εργασία, οι οποίες επιδιώκουν όχι μόνο στην βελτίωση μεθοδολογιών ανάλυσης του 16S rRNA γονιδίου αλλά και στην συμβολή μίας πιο ολοκληρωμένης κατανόησης της περίπλοκης δυναμικής των μικροβιακών κοινοτήτων του ανθρώπινου αίματος, ιδιαίτερα στο πλαίσιο της σχιζοφρένειας, μπορούν να διακριθούν σε δύο βασικούς άξονες. Η μία κατεύθυνση αφορά τη βελτίωση της συγκριτικής ανάλυσης και η άλλη αντίστοιχα την περαιτέρω εξερεύνηση των δεδομένων αλληλούχισης του 16S rRNA γονιδίου που προέρχονται από δείγματα αίματος ατόμων διαγνωσμένοι με σχιζοφρένεια.

Η επιρροή των παραμέτρων στο στάδιο της προεπεξεργασίας απαιτεί ιδιαίτερη προσοχή σε μελλοντικές έρευνες σχετικές με την ανάλυση του 16S rRNA γονιδίου. Μία εξειδικευμένη διερεύνηση στον αντίκτυπο των τιμών παραμέτρων ποιοτικού φιλτραρίσματος, όπως του $e.e._{max}$ και του Q_{min}, καθώς και η στρατηγική απόφαση για τη συγχώνευση των paired-end αναγνωσμάτων πριν ή μετά το ποιοτικό φιλτράρισμα, υπόσχεται μια πιο αναλυτική κατανόηση της συμπεριφοράς των δεδομένων και των τελικών αποτελεσμάτων. Η συμπερίληψη μεθοδολογιών *de novo* ομαδοποίησης ή ανοιχτής αναφοράς, που παρέχεται από το βιοπληροφορικό εργαλείο VSEARCH (Rognes et al., 2016), θα μπορούσε να συμβάλει σημαντικά σε μία συγκριτική ανάλυση των μεθόδων αποθορυβοποίησης και ομαδοποίησης. Επιπλέον, η αξιοποίηση διαφορετικών βάσεων δεδομένων για την ομαδοποίηση αναγνωσμάτων σε OTUs ή για την διαδικασία της ταξινομικής ανάθεσης, ο προσδιορισμός και η σύγκριση διαφόρων δεικτών ποικιλομορφίας, όπως αυτών της β-ποικιλομορφίας, θα μπορούσαν να αυξήσουν την εγκυρότητα της παρούσας συγκριτικής ανάλυσης. Επιπλέον, η ενσωμάτωση εργαλείων βιοπληροφορικής που αναπτύχθηκαν έξω από το περιβάλλον QIIME2 και η εις βάθος εξερεύνηση διαφορετικών μεθόδων αποθορυβοποίησης, συμπεριλαμβανομένης της εφαρμογής του UNOISE (Nearing et al., 2018), εξασφαλίζει σίγουρα τον εμπλουτισμό του συγκριτικού πεδίου. Φυσικά, τα αποτελέσματα της παρούσας ανάλυσης για να είναι επιστημονικά αποδεδειγμένα, θα πρέπει να είναι και επαναλήψιμα. Συνεπώς, μια όμοια ανάλυση είναι απαραίτητο να πραγματοποιηθεί. Η χρήση διαφορετικών αρχικών πρωτογενών δεδομένων αλληλούχισης της περιοχής V3-V4 του 16S rRNA γονιδίου, τα οποία ιδανικά προέρχονται από δείγματα αίματος με παρόμοιες πειραματικές συνθήκες, και η εφαρμογή των ίδιων επεξεργαστικών ροών που κατασκευάστηκαν στην τρέχων μελέτη με αντίστοιχες παραμέτρους θα επιτρέψουν τον έλεγχο της επίδρασης των μεθόδων αποθορυβοποίησης και ομαδοποίησης και την διερεύνηση του αντίκτυπου των επεξεργαστικών ροών και σε άλλα δεδομένα.

Μια στοχευμένη προσπάθεια για τη βελτίωση της ανάλυσης των πρωτογενών δεδομένων της παρούσας εργασίας είναι επιτακτική. Η χρήση πιο εξειδικευμένων βιοπληροφορικών εργαλείων για τον εντοπισμό και την αφαίρεση επιμολύνσεων που η πλατφόρμα QIIME2 δεν παρέχει, όπως το decontam, διασφαλίζει μια πιο ολοκληρωμένη προσέγγιση ανάλυσης δειγμάτων αίματος. Επιπρόσθετα, η ενσωμάτωση βιοπληροφορικών εργαλείων πρόβλεψης λειτουργικού προφίλ, όπως το PICRUSt, το οποίο υποστηρίζεται και από το QIIME2, μπορεί να ρίξει φως στις πιθανές λειτουργικές συνέπειες των βακτηριακών παραλλαγών. Ολοκληρώνοντας, η συμπερίληψη δεδομένων αλληλούχισης από δείγματα υγιούς αίματος και συσχετιζόμενα με τα διαθέσιμα δεδομένα της παρούσας Διπλωματικής Εργασίας ενισχύει την έρευνα συσχέτισης μικροβιώματος αίματος και σχιζοφρένειας, προσφέροντας μια βάση για την αξιολόγηση των αποκλίσεων στις μικροβιακές αλληλεπιδράσεις. Αυτή η προσέγγιση επιτρέπει μια πιο ολοκληρωμένη εξερεύνηση των

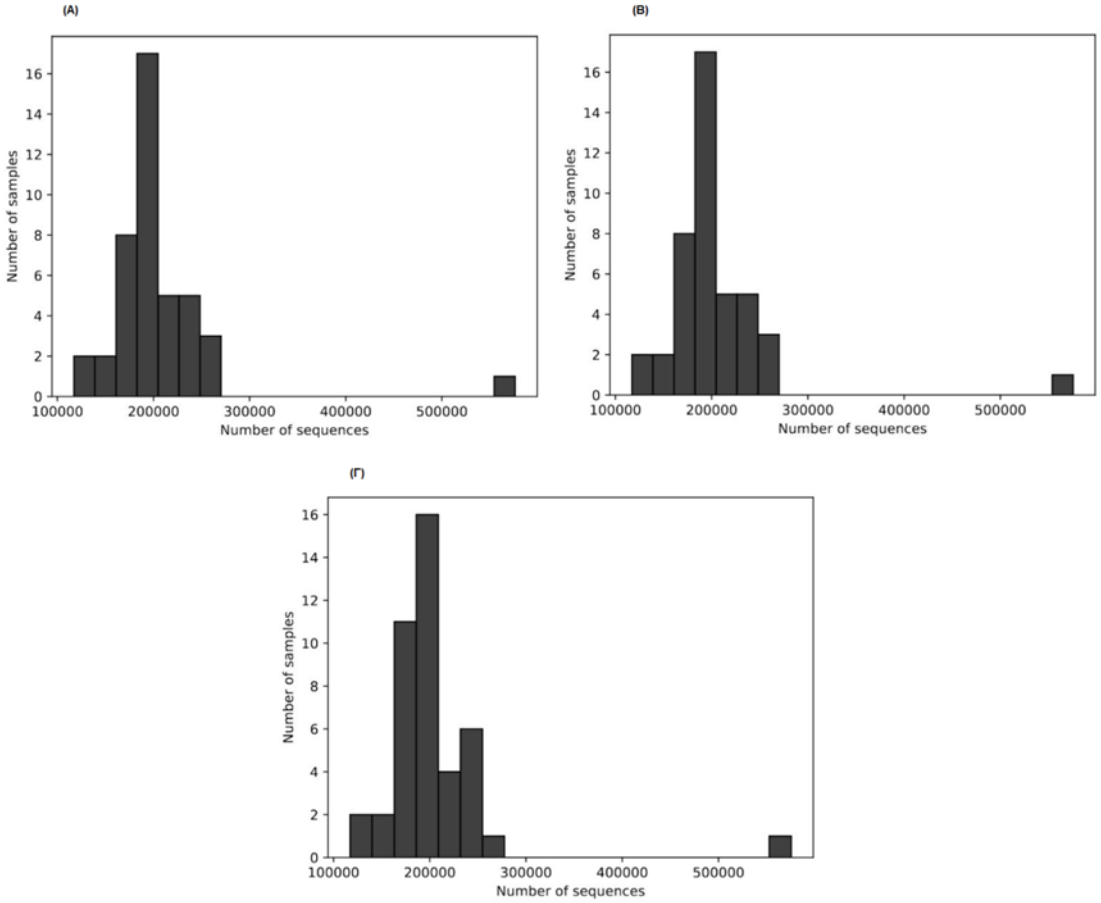
αποστάσεων β-ποικιλομορφίας, με τα δείγματα αίματος από υγιή άτομα να παίρνουν το ρόλο της αναφοράς, παρέχοντας πληροφορίες για πιθανές δυσρυθμίσεις ή μοναδικά μικροβιακά μοτίβα που σχετίζονται με την πάθηση της σχιζοφρένειας.

Παράρτημα

Παράρτημα 1. Πίνακας με τον αριθμό paired-end πρωτογενών αναγνωσμάτων και αριθμό διατηρητέων συγχωνευμένων αναγνωσμάτων ανά δείγμα μετά την εφαρμογή συγχώνευσης στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής.

Αρ. Δείγματος	Πρωτογενή		Συγχώνευση		
	Εμπρόσθια	Ανάστροφα	$Overlap_{min}$=10	$Overlap_{min}$=20	$Overlap_{min}$=30
sample1a	635568	635568	576207	575660	575630
sample1b	286442	286442	250569	250493	250314
sample2a	287471	287471	260449	260427	260215
sample2b	278024	278024	251040	251014	250175
sample3a	286627	286627	245955	245882	245154
sample3b	238137	238137	202209	202039	201905
sample4a	256150	256150	219519	219411	219390
sample4b	269149	269149	242879	242866	242535
sample5a	276445	276445	205560	205252	204729
sample5b	244451	244451	180322	180266	179854
sample6a	252395	252395	190879	190695	190323
sample6b	264515	264515	203063	202965	202884
sample7a	254884	254884	199179	199169	199092
sample7b	247215	247215	188612	188475	188133
sample8a	258191	258191	196090	195337	194001
sample8b	287333	287333	224991	224729	223511
sample9a	241789	241789	183216	183066	182846
sample9b	237894	237894	185854	185690	185554
sample10a	182016	182016	122730	122666	121802
sample10b	250520	250520	180755	180211	179495
sample11a	267362	267362	187088	186983	186274
sample11b	297998	297998	240826	240548	240387
sample12a	213449	213449	156336	156258	155683
sample12b	285872	285872	204151	203993	203633
sample13a	288582	288582	228804	228713	228480
sample13b	228547	228547	173324	173172	172885
sample14a	246166	246166	193575	193374	192314
sample14b	278361	278361	220780	220664	220254
sample15a	236775	236775	182957	182899	181939
sample15b	252457	252457	169803	169629	168858
sample16a	260691	260691	200840	200749	200493
sample16b	246481	246481	172297	172207	171587
sample17a	271749	271749	208384	199169	207990
sample17b	236239	236239	165498	165411	165272
sample18a	209518	209518	164715	164425	163840
sample18b	199792	199792	157498	157356	156842
sample19a	222217	222217	171975	171908	171549

sample19b	252215	252215	197023	196666	196296
sample20a	243703	243703	188703	188610	187800
sample20b	263337	263337	200103	199955	199589
negativecontrol1	221505	221505	202406	202406	202406
negativecontrol2	263038	263038	242976	242976	242970
negativecontrol3	126881	126881	117272	117271	117271

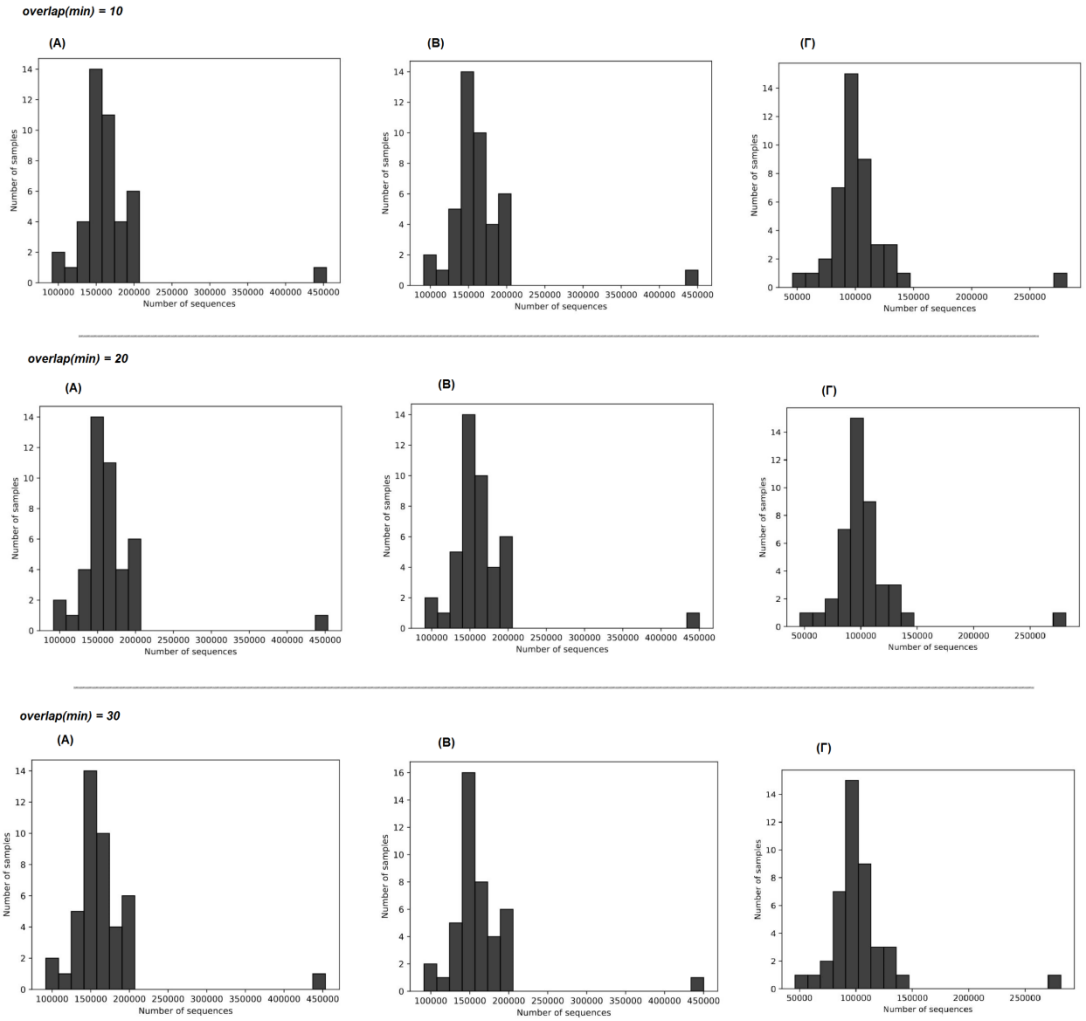


Παράρτημα 2. Ιστογράμματα συχνότητας δειγμάτων που παρέχουν μία δεδομένη τιμή αριθμού συγχωνευμένων αναγνωσμάτων στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, όπου **(Α)** $Overlap_{min}$=10, **(Β)** $Overlap_{min}$=20 και **(Γ)** $Overlap_{min}$=30.

Παράρτημα 3. Πίνακας με τον αριθμό διατηρητέων συγχωνευμένων και φιλτραρισμένων αναγνωσμάτων ανά δείγμα στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας μετά την εφαρμογή ποιοτικού φιλτραρίσματος.

Αρ. Δείγματος	**Ποιοτικό Φιλτράρισμα**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Q_{min}=20	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**
sample1a	454022	450639	281793	453580	450198	281778	453559	450177	281778
sample1b	196585	195118	122938	196523	195056	122929	196403	194937	122929
sample2a	204146	202570	128279	204135	202559	128279	203976	202400	128264
sample2b	196621	195096	122352	196604	195079	122352	196061	194537	122341
sample3a	201678	200526	137842	201618	200466	137841	201093	199943	137751
sample3b	161209	160209	101373	161098	160098	101373	160999	159999	101360
sample4a	178734	177537	120047	178658	177462	120047	178644	177448	120047

sample4b	193653	192387	107488	193647	192381	107488	193426	192160	107478
sample5a	163791	162296	110511	163534	162039	110500	163213	161722	110487
sample5b	144127	143011	76247	144083	142967	76245	143830	142714	76228
sample6a	151352	149963	98579	151225	149836	98568	150965	149577	98546
sample6b	157954	156551	97976	157887	156484	97971	157829	156426	97957
sample7a	157670	156281	86680	157664	156275	86680	157611	156222	86672
sample7b	147581	146228	92256	147499	146146	92255	147266	145918	92233
sample8a	154126	152738	98307	153591	152204	98204	152589	151208	97732
sample8b	172216	170720	100885	172022	170527	100874	171270	169776	100789
sample9a	148510	147181	99449	148415	147086	99449	148259	146933	99434
sample9b	148433	146960	92236	148358	146885	92236	148274	146801	92222
sample10a	106094	105294	81550	106052	105253	81546	105394	104596	81385
sample10b	148264	146937	99766	147894	146573	99720	147413	146097	99659
sample11a	158357	157099	111511	158273	157015	111505	157794	156539	111463
sample11b	194696	192875	128760	194489	192669	128736	194432	192612	128734
sample12a	115750	113901	66315	115703	113855	66311	115374	113531	66299
sample12b	172843	171428	126442	172731	171317	126438	172486	171076	126422
sample13a	181295	179706	104515	181229	179641	104514	181081	179493	104489
sample13b	143766	142686	94799	143675	142595	94780	143493	142415	94766
sample14a	158324	157004	97366	158176	156856	97345	157471	156153	97301
sample14b	178079	176610	111257	177988	176519	111257	177700	176232	111231
sample15a	147433	146321	85718	147394	146282	85718	146769	145659	85637
sample15b	144506	143469	106436	144398	143362	106433	143888	142856	106375
sample16a	164387	162961	106366	164321	162896	106365	164153	162732	106346
sample16b	139648	138630	87389	139595	138577	87389	139205	138189	87325
sample17a	164350	162732	92950	164299	162682	92947	164089	162473	92870
sample17b	133573	132491	86866	133516	132434	86863	133428	132347	86856
sample18a	132459	131200	87746	132264	131006	87741	131864	130608	87566
sample18b	126707	125554	82326	126596	125443	82294	126237	125085	82240
sample19a	141464	140200	96964	141415	140151	96959	141162	139900	96870
sample19b	160603	159250	104445	160350	158999	104416	160094	158744	104395
sample20a	152128	150698	99049	152063	150633	99045	151486	150059	98900
sample20b	164043	162527	108994	163936	162420	108991	163686	162172	108969
negativecontrol1	155648	154153	73044	155648	154153	73044	155648	154153	73044
negativecontrol2	184362	182784	102075	184362	182784	102075	184361	182783	102075
negativecontrol3	91941	91346	46067	91941	91346	46067	91941	91346	46067



Παράρτημα 4. Ιστογράμματα συχνότητας δειγμάτων που παρέχουν μία δεδομένη τιμή αριθμού συγχωνευμένων αναγνωσμάτων στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας, όπου **(Α)** Q_{min}=20, **(Β)** Q_{min}=22 και **(Γ)** Q_{min}=26.

Παράρτημα 5. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης DADA2 στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και για τιμή μέγιστου ποσοστού αναμενόμενων σφαλμάτων ίσο με $e.e._{max}$=2,5 και χωρίς αποκοπή των paired-end αναγνωσμάτων.

Αρ. Δείγματος	**DADA2 με $e.e._{max}$=2,5 και no-trim**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Ποιοτικό Φιλτράρισμα	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**
sample1a	528842	527854	485628	528842	527854	485086	528842	527854	485086
sample1b	235602	234919	213353	235602	234919	213253	235602	234919	213103
sample2a	240122	239336	218907	240122	239336	218817	240122	239336	218622
sample2b	231658	231029	214753	231658	231029	214665	231658	231029	213952
sample3a	241227	239940	216570	241227	239940	216412	241227	239940	215731
sample3b	196378	195244	176682	196378	195244	176551	196378	195244	176464
sample4a	217352	216302	187351	217352	216302	187169	217352	216302	187169
sample4b	223979	223026	203664	223979	223026	203480	223979	223026	203251
sample5a	214818	213111	172924	214818	213111	172460	214818	213111	171996
sample5b	186326	184671	158193	186326	184671	158140	186326	184671	157830

sample6a	194074	192121	157716	194074	192121	157473	194074	192121	157154
sample6b	203174	200922	170591	203174	200922	170540	203174	200922	170516
sample7a	194956	192839	158724	194956	192839	158724	194956	192839	158670
sample7b	189012	187124	161107	189012	187124	160911	189012	187124	160696
sample8a	199073	197247	168529	199073	197247	167798	199073	197247	166552
sample8b	221865	219938	196331	221865	219938	196054	221865	219938	195015
sample9a	193542	191471	165780	193542	191471	165615	193542	191471	165433
sample9b	190981	188960	163181	190981	188960	162902	190981	188960	162808
sample10a	146469	143030	110515	146469	143030	110350	146469	143030	109548
sample10b	197978	195738	152405	197978	195738	151639	197978	195738	151103
sample11a	212405	209600	170814	212405	209600	170540	212405	209600	169912
sample11b	243982	242130	208426	243982	242130	207980	243982	242130	207924
sample12a	162405	159822	140531	162405	159822	140458	162405	159822	140166
sample12b	236440	232046	171188	236440	232046	170713	236440	232046	170481
sample13a	231576	229418	196814	231576	229418	196711	231576	229418	196588
sample13b	181697	179837	149424	181697	179837	149267	181697	179837	149063
sample14a	196024	193455	160845	196024	193455	160615	196024	193455	159746
sample14b	223206	221302	186045	223206	221302	185968	223206	221302	185604
sample15a	186800	184202	147910	186800	184202	147752	186800	184202	146910
sample15b	196813	194069	143212	196813	194069	142957	196813	194069	142410
sample16a	209773	206958	157390	209773	206958	157322	209773	206958	157167
sample16b	190164	187783	144278	190164	187783	144208	190164	187783	143740
sample17a	206616	204497	173834	206616	204497	173774	206616	204497	173524
sample17b	178470	177115	130470	178470	177115	130378	178470	177115	130346
sample18a	170131	167712	138059	170131	167712	137790	170131	167712	137290
sample18b	161445	159405	133434	161445	159405	133179	161445	159405	132692
sample19a	182133	179212	139060	182133	179212	138950	182133	179212	138619
sample19b	204449	202110	174504	204449	202110	174019	204449	202110	173746
sample20a	194747	191740	159404	194747	191740	159236	194747	191740	158434
sample20b	213768	209956	165166	213768	209956	164922	213768	209956	164591
negativecontrol1	179521	179363	163899	179521	179363	163832	179521	179363	163832
negativecontrol2	215964	215212	193386	215964	215212	193386	215964	215212	193386
negativecontrol3	106050	105788	101103	106050	105788	101083	106050	105788	101083

Παράρτημα 6. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης DADA2 στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και για τιμή μέγιστου ποσοστού αναμενόμενων σφαλμάτων ίσο με $e.e._{max}$=1,5 και χωρίς αποκοπή των paired-end αναγνωσμάτων.

	DADA2 με $e.e._{max}$=1,5 και no-trim								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
Αρ. Δείγματος	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**
sample1a	413588	413128	383417	413588	413128	382922	413588	413128	382922
sample1b	179731	179326	166107	179731	179326	165995	179731	179326	165896
sample2a	187989	187575	173964	187989	187575	173890	187989	187575	173728

sample2b	181809	181435	171076	181809	181435	170996	181809	181435	170550
sample3a	189829	189109	173782	189829	189109	173609	189829	189109	173105
sample3b	149101	148375	137378	149101	148375	137298	149101	148375	137229
sample4a	170361	169821	147486	170361	169821	147358	170361	169821	147358
sample4b	174254	173646	161297	174254	173646	161166	174254	173646	160959
sample5a	155238	154145	128334	155238	154145	127988	155238	154145	127703
sample5b	128409	127378	111141	128409	127378	111102	128409	127378	110897
sample6a	139614	138628	115415	139614	138628	115213	139614	138628	114985
sample6b	143810	142303	122171	143810	142303	122136	143810	142303	122117
sample7a	137852	136599	113750	137852	136599	113750	137852	136599	113719
sample7b	135077	134040	117021	135077	134040	116934	135077	134040	116790
sample8a	142398	141558	124083	142398	141558	123536	142398	141558	122551
sample8b	154455	153229	137459	154455	153229	137273	154455	153229	136674
sample9a	135480	134408	117269	135480	134408	117147	135480	134408	117024
sample9b	135054	133893	116995	135054	133893	116826	135054	133893	116766
sample10a	108379	106518	85485	108379	106518	85352	108379	106518	84728
sample10b	142776	141546	112863	142776	141546	112301	142776	141546	111926
sample11a	154828	153194	128983	154828	153194	128801	154828	153194	128363
sample11b	180582	179585	156429	180582	179585	156117	180582	179585	156061
sample12a	104199	102854	92509	104199	102854	92469	104199	102854	92343
sample12b	179964	177289	133204	179964	177289	132799	179964	177289	132613
sample13a	166866	165401	143972	166866	165401	143921	166866	165401	143833
sample13b	130713	129526	109678	130713	129526	109530	130713	129526	109394
sample14a	143933	142406	120354	143933	142406	120184	143933	142406	119574
sample14b	161545	160427	137165	161545	160427	137099	161545	160427	136838
sample15a	131755	130201	106000	131755	130201	105881	131755	130201	105361
sample15b	142555	141075	109142	142555	141075	108985	142555	141075	108607
sample16a	153930	152083	118336	153930	152083	118292	153930	152083	118219
sample16b	125704	124477	96355	125704	124477	96291	125704	124477	95947
sample17a	146732	145237	126168	146732	145237	126125	146732	145237	125931
sample17b	126263	125652	94700	126263	125652	94641	126263	125652	94621
sample18a	127112	125820	105327	127112	125820	105079	127112	125820	104707
sample18b	118265	117013	99521	118265	117013	99342	118265	117013	98985
sample19a	137808	135976	107204	137808	135976	107119	137808	135976	106839
sample19b	152957	151521	134334	152957	151521	133970	152957	151521	133758
sample20a	144211	142300	121281	144211	142300	121141	144211	142300	120563
sample20b	160381	158106	125791	160381	158106	125581	160381	158106	125360
negativecontrol1	134130	133395	123790	134130	133395	123790	134130	133395	123790
negativecontrol2	163940	163301	147573	163940	163301	147573	163940	163301	147573
negativecontrol3	82277	82084	78642	82277	82084	78642	82277	82084	78642

Παράρτημα 7. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης DADA2 στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και για τιμή μέγιστου ποσοστού αναμενόμενων σφαλμάτων ίσο με $e.e._{max}$=0,5 και χωρίς αποκοπή των paired-end αναγνωσμάτων.

Αρ. Δείγματος	DADA2 με $e.e._{max}$=0,5 και no-trim								
	$Overlap_{min}$=10			$Overlap_{min}$=20			$Overlap_{min}$=30		
	Ποιοτικό Φιλτράρισμα	Αποθορυβοποίηση	Συγχώνευση	Ποιοτικό Φιλτράρισμα	Αποθορυβοποίηση	Συγχώνευση	Ποιοτικό Φιλτράρισμα	Αποθορυβοποίηση	Συγχώνευση
sample1a	118476	118367	113045	118476	118367	112937	118476	118367	112937
sample1b	46550	46408	45213	46550	46408	45146	46550	46408	45146
sample2a	54156	54079	53149	54156	54079	53063	54156	54079	53040
sample2b	54177	54049	52538	54177	54049	52460	54177	54049	52452
sample3a	53262	53009	50360	53262	53009	50298	53262	53009	50298
sample3b	35335	35048	33702	35335	35048	33702	35335	35048	33696
sample4a	44789	44480	40735	44789	44480	40670	44789	44480	40670
sample4b	46249	46092	44457	46249	46092	44301	46249	46092	44290
sample5a	36516	36108	31814	36516	36108	31779	36516	36108	31754
sample5b	26236	25999	23212	26236	25999	23212	26236	25999	23191
sample6a	34584	34295	30094	34584	34295	30011	34584	34295	29999
sample6b	33061	32578	28848	33061	32578	28848	33061	32578	28848
sample7a	29641	29185	24811	29641	29185	24780	29641	29185	24789
sample7b	32713	32297	29571	32713	32297	29571	32713	32297	29539
sample8a	33584	33361	30696	33584	33361	30593	33584	33361	30366
sample8b	27139	26762	25016	27139	26762	25011	27139	26762	24984
sample9a	21377	20947	19048	21377	20947	19005	21377	20947	19017
sample9b	20666	20388	18327	20666	20388	18327	20666	20388	18339
sample10a	22654	22116	19715	22654	22116	19618	22654	22116	19554
sample10b	26063	25717	21700	26063	25717	21611	26063	25717	21582
sample11a	30204	29812	26578	30204	29812	26511	30204	29812	26439
sample11b	37364	36944	34023	37364	36944	33883	37364	36944	33862
sample12a	10701	10423	9818	10701	10423	9818	10701	10423	9818
sample12b	44032	42968	32852	44032	42968	32680	44032	42968	32638
sample13a	27518	27196	23953	27518	27196	23916	27518	27196	23900
sample13b	23231	22844	19864	23231	22844	19802	23231	22844	19783
sample14a	29493	29008	25615	29493	29008	25594	29493	29008	25599
sample14b	30413	30046	27006	30413	30046	26989	30413	30046	26954
sample15a	18704	18298	15323	18704	18298	15284	18704	18298	15248
sample15b	29716	29289	26064	29716	29289	26039	29716	29289	25978
sample16a	30117	29633	24196	30117	29633	24105	30117	29633	24069
sample16b	13666	13378	10856	13666	13378	10856	13666	13378	10826
sample17a	33824	33425	31164	33824	33425	31127	33824	33425	31083
sample17b	29240	28911	22131	29240	28911	22131	29240	28911	22123
sample18a	26479	26163	23926	26479	26163	23882	26479	26163	23798
sample18b	22611	22214	19632	22611	22214	19511	22611	22214	19508
sample19a	30548	29888	25203	30548	29888	25079	30548	29888	25044

sample19b	31507	31128	28963	31507	31128	28901	31507	31128	28912
sample20a	30186	29769	26078	30186	29769	25985	30186	29769	25907
sample20b	35864	35153	29384	35864	35153	29351	35864	35153	29299
negativecontrol1	26625	26619	25710	26625	26619	25710	26625	26619	25710
negativecontrol2	39732	39606	36711	39732	39606	36711	39732	39606	36711
negativecontrol3	19627	19579	18687	19627	19579	18667	19627	19579	18667

Παράρτημα 8. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης DADA2 στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και για τιμή μέγιστου ποσοστού αναμενόμενων σφαλμάτων ίσο με $e.e._{max}$=2,5 και αποκοπή των paired-end αναγνωσμάτων.

Αρ. Δείγματος	**DADA2 με $e.e._{max}$=2,5 και with-trim**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Ποιοτικό Φιλτράρισμα	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**
sample1a	520785	519421	484123	520785	519421	481297	520785	519421	481297
sample1b	229449	228481	210937	229449	228481	209415	229449	228481	209377
sample2a	233676	232729	215814	233676	232729	214367	233676	232729	214192
sample2b	227764	226798	210995	227764	226798	209498	227764	226798	209302
sample3a	231208	229669	211043	231208	229669	209699	231208	229669	209135
sample3b	188999	187525	171519	188999	187525	170491	188999	187525	170395
sample4a	201002	199869	181334	201002	199869	180041	201002	199869	180035
sample4b	220278	219040	203171	220278	219040	202268	220278	219040	201955
sample5a	184720	182461	156754	184720	182461	155984	184720	182461	155686
sample5b	167468	165093	145085	167468	165093	144906	167468	165093	144631
sample6a	182167	180039	149856	182167	180039	149358	182167	180039	148583
sample6b	190232	187325	164090	190232	187325	163671	190232	187325	163194
sample7a	178627	175910	154275	178627	175910	153986	178627	175910	153449
sample7b	174219	171802	149156	174219	171802	148737	174219	171802	147860
sample8a	183204	180520	155913	183204	180520	154225	183204	180520	153678
sample8b	206766	204469	183920	206766	204469	183238	206766	204469	183110
sample9a	177744	175046	153782	177744	175046	153235	177744	175046	153049
sample9b	180542	178206	156665	180542	178206	155780	180542	178206	155601
sample10a	126103	122367	95521	126103	122367	94811	126103	122367	94327
sample10b	178045	175121	145179	178045	175121	144402	178045	175121	144067
sample11a	185444	182066	149361	185444	182066	148635	185444	182066	148018
sample11b	225696	222917	194327	225696	222917	192530	225696	222917	192389
sample12a	150498	146738	130752	150498	146738	129939	150498	146738	129334
sample12b	202658	198159	154692	202658	198159	153711	202658	198159	153521
sample13a	208526	206350	180747	208526	206350	180244	208526	206350	180183
sample13b	161147	159008	136440	161147	159008	136148	161147	159008	135441
sample14a	174468	171717	147088	174468	171717	146708	174468	171717	145761
sample14b	204007	201323	171557	204007	201323	170925	204007	201323	170446
sample15a	169465	166072	139529	169465	166072	139268	169465	166072	138582
sample15b	172115	169251	126992	172115	169251	126447	172115	169251	126089

sample16a	185674	182081	149349	185674	182081	149020	185674	182081	148721
sample16b	169311	166784	140582	169311	166784	140567	169311	166784	140107
sample17a	193123	190500	164449	193123	190500	164226	193123	190500	163871
sample17b	152942	150908	126553	152942	150908	126528	152942	150908	126119
sample18a	159644	156940	129372	159644	156940	128703	159644	156940	128189
sample18b	148987	146642	124400	148987	146642	123823	148987	146642	123427
sample19a	166488	163377	129850	166488	163377	129350	166488	163377	129122
sample19b	187701	185128	161160	187701	185128	160298	187701	185128	159746
sample20a	179560	176349	149413	179560	176349	148908	179560	176349	148468
sample20b	193147	189041	150623	193147	189041	150075	193147	189041	149599
negativecontrol1	177561	176985	165054	177561	176985	165133	177561	176985	164893
negativecontrol2	213939	212645	191131	213939	212645	190804	213939	212645	190804
negativecontrol3	105721	105249	99928	105721	105249	99606	105721	105249	99606

Παράρτημα 9. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης DADA2 στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και για τιμή μέγιστου ποσοστού αναμενόμενων σφαλμάτων ίσο με $e.e._{max}$=1,5 και αποκοπή των paired-end αναγνωσμάτων.

Αρ. Δείγματος	**DADA2 με $e.e._{max}$=1,5 και with-trim**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Ποιοτικό Φιλτράρισμα	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**
sample1a	410009	409510	385584	410009	409510	384374	410009	409510	384374
sample1b	176582	176128	165352	176582	176128	164608	176582	176128	164581
sample2a	184036	183578	172631	184036	183578	171980	184036	183578	171834
sample2b	180100	179614	169516	180100	179614	168756	180100	179614	168641
sample3a	182912	182017	170558	182912	182017	169811	182912	182017	169416
sample3b	144635	143909	134307	144635	143909	133751	144635	143909	133675
sample4a	158163	157590	146719	158163	157590	146118	158163	157590	146115
sample4b	172591	171868	161680	172591	171868	161138	172591	171868	160941
sample5a	133496	132513	116935	133496	132513	116614	133496	132513	116437
sample5b	115804	114784	102907	115804	114784	102813	115804	114784	102639
sample6a	131598	130627	112461	131598	130627	112221	131598	130627	111670
sample6b	135331	134021	118906	135331	134021	118712	135331	134021	118392
sample7a	126531	125280	111348	126531	125280	111198	126531	125280	110816
sample7b	124791	123796	109819	124791	123796	109685	124791	123796	109051
sample8a	132002	130971	116511	132002	130971	115279	132002	130971	114872
sample8b	145246	144120	131183	145246	144120	130988	145246	144120	130913
sample9a	125148	123940	110708	125148	123940	110562	125148	123940	110424
sample9b	129077	127806	114035	129077	127806	113610	129077	127806	113508
sample10a	93061	91171	74522	93061	91171	74033	93061	91171	73736
sample10b	128813	127476	109907	128813	127476	109373	128813	127476	109160
sample11a	135197	133495	114084	135197	133495	113756	135197	133495	113400
sample11b	167690	166266	148646	167690	166266	148161	167690	166266	148081
sample12a	97386	95724	87473	97386	95724	87105	97386	95724	86768

sample12b	153545	151205	120904	153545	151205	120518	153545	151205	120376
sample13a	150890	149716	132410	150890	149716	132172	150890	149716	132120
sample13b	116194	115053	101111	116194	115053	100954	116194	115053	100466
sample14a	128347	126927	111395	128347	126927	111164	128347	126927	110451
sample14b	148917	147646	128158	148917	147646	127922	148917	147646	127594
sample15a	120813	119195	103460	120813	119195	103356	120813	119195	102893
sample15b	124882	123510	100093	124882	123510	99829	124882	123510	99568
sample16a	136672	134925	114974	136672	134925	114789	136672	134925	114620
sample16b	113212	112019	96668	113212	112019	96624	113212	112019	96286
sample17a	137431	136045	120435	137431	136045	120281	137431	136045	120013
sample17b	108724	107957	92758	108724	107957	92792	108724	107957	92512
sample18a	119661	118275	100801	119661	118275	100330	119661	118275	99965
sample18b	109887	108597	94123	109887	108597	93811	109887	108597	93545
sample19a	126094	124390	101933	126094	124390	101612	126094	124390	101451
sample19b	140899	139647	125150	140899	139647	124575	140899	139647	124182
sample20a	133447	131696	114789	133447	131696	114465	133447	131696	114159
sample20b	144989	142981	118615	144989	142981	118320	144989	142981	117987
negativecontrol1	133878	132959	126306	133878	132959	126317	133878	132959	126198
negativecontrol2	163716	162809	147370	163716	162809	147149	163716	162809	147149
negativecontrol3	82707	82406	78698	82707	82406	78539	82707	82406	78539

Παράρτημα 10. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης DADA2 στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και για τιμή μέγιστου ποσοστού αναμενόμενων σφαλμάτων ίσο με $e.e._{max}$=0,5 και αποκοπή των paired-end αναγνωσμάτων.

Αρ. Δείγματος	**DADA2 με $e.e._{max}$=0,5 και with-trim**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Ποιοτικό Φιλτράρισμα	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**	**Ποιοτικό Φιλτράρισμα**	**Αποθορυβοποίηση**	**Συγχώνευση**
sample1a	120757	120619	117480	120757	120619	117269	120757	120619	117269
sample1b	47243	47156	46761	47243	47156	46655	47243	47156	46655
sample2a	54507	54438	53677	54507	54438	53590	54507	54438	53569
sample2b	55461	55374	53914	55461	55374	53846	55461	55374	53833
sample3a	52145	51971	51073	52145	51971	50937	52145	51971	50927
sample3b	35080	34952	34272	35080	34952	34198	35080	34952	34192
sample4a	41483	41353	40535	41483	41353	40474	41483	41353	40474
sample4b	47236	47081	45927	47236	47081	45843	47236	47081	45833
sample5a	32233	31997	29832	32233	31997	29788	32233	31997	29785
sample5b	23954	23791	22348	23954	23791	22348	23954	23791	22336
sample6a	34224	34036	30860	34224	34036	30842	34224	34036	30764
sample6b	32402	32123	29809	32402	32123	29800	32402	32123	29773
sample7a	27767	27458	25984	27767	27458	25981	27767	27458	25948
sample7b	31474	31231	29585	31474	31231	29580	31474	31231	29508
sample8a	33129	32922	31362	33129	32922	31135	33129	32922	31075
sample8b	27357	27197	25957	27357	27197	25938	27357	27197	25934

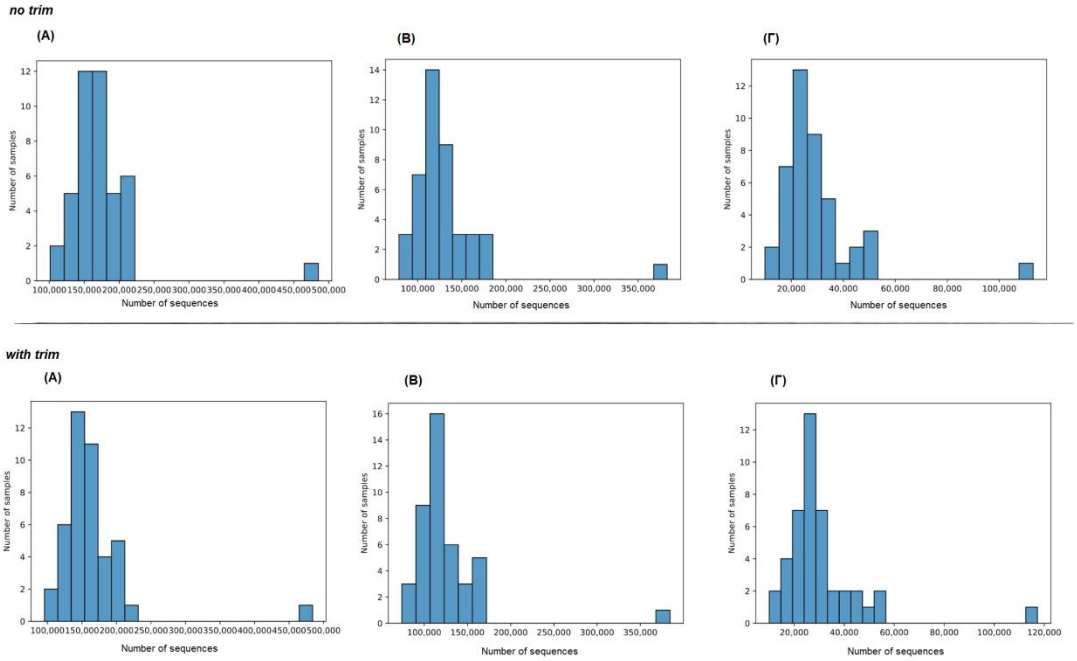
sample9a	19991	19777	18771	19991	19777	18761	19991	19777	18761
sample9b	20494	20307	19026	20494	20307	19012	20494	20307	19012
sample10a	19336	18830	17203	19336	18830	17116	19336	18830	17096
sample10b	24404	24158	22543	24404	24158	22518	24404	24158	22512
sample11a	26537	26191	24369	26537	26191	24330	26537	26191	24304
sample11b	36069	35641	33650	36069	35641	33618	36069	35641	33618
sample12a	10402	10275	10014	10402	10275	10014	10402	10275	10010
sample12b	34470	33826	29274	34470	33826	29222	34470	33826	29208
sample13a	25826	25631	23771	25826	25631	23755	25826	25631	23748
sample13b	21383	21064	19876	21383	21064	19858	21383	21064	19827
sample14a	27888	27576	25762	27888	27576	25748	27888	27576	25692
sample14b	30220	29854	27468	30220	29854	27464	30220	29854	27438
sample15a	18111	17853	16324	18111	17853	16322	18111	17853	16294
sample15b	26854	26580	24203	26854	26580	24188	26854	26580	24162
sample16a	27921	27586	24919	27921	27586	24901	27921	27586	24878
sample16b	12451	12308	11176	12451	12308	11172	12451	12308	11172
sample17a	33136	32805	31132	33136	32805	31115	33136	32805	31055
sample17b	26080	25964	24070	26080	25964	24065	26080	25964	24002
sample18a	25503	25251	23523	25503	25251	23439	25503	25251	23416
sample18b	21538	21287	19601	21538	21287	19563	21538	21287	19563
sample19a	27594	27201	24288	27594	27201	24240	27594	27201	24233
sample19b	30103	29882	28302	30103	29882	28259	30103	29882	28235
sample20a	28581	28260	25848	28581	28260	25766	28581	28260	25753
sample20b	32089	31657	28443	32089	31657	28416	32089	31657	28377
negativecontrol1	28493	28452	28116	28493	28452	28072	28493	28452	28068
negativecontrol2	41931	41742	38768	41931	41742	38717	41931	41742	38717
negativecontrol3	21329	21223	20444	21329	21223	20423	21329	21223	20423

Παράρτημα 11 Γενική περιγραφή αριθμού φιλτραρισμένων αναγνωσμάτων των συνολικών δειγμάτων που προέκυψαν από τον DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων και συνθήκες αποκοπής αναγνωσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Φιλτραρισμένων paired-end Αναγνωσμάτων					
	No trim			**With trim**		
e.e.$_{max}$	**2,5**	**1,5**	**0,5**	**2,5**	**1,5**	**0,5**
Ελάχιστος	106050	82277	10701	105721	82707	10402
Μέσος Όρος	197978	143933	30204	183204	133496	28493
Μέγιστος	528842	413588	118476	520785	410009	120757
Συνολικός	8932007 ***(80,12%)***	6575854 ***(58,99%)***	1448630 ***(12,99%)***	8270820 ***(74,19%)***	6125104 ***(54,94%)***	1395186 ***(12,51%)***

Παράρτημα 12. Γενική περιγραφή αριθμού αποθορυβοποιημένων αναγνωσμάτων στο σύνολο των δειγμάτων που προέκυψαν από τον DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων και συνθήκες αποκοπής αναγνωσμάτων. Στην παρένθεση αναγράφεται ο ποσοστιαίος αριθμός αυτού σε σχέση με τον αριθμό των πρωτογενών αναγνωσμάτων.

	Αριθμός Αποθορυβοποιημένων paired-end Αναγνωσμάτων					
	No trim			**With trim**		
e.e.$_{max}$	**2,5**	**1,5**	**0,5**	**2,5**	**1,5**	**0,5**
Ελάχιστος	105788	82084	10423	105249	82406	10275
Μέσος Όρος	195738	142303	29812	180520	132513	28260
Μέγιστος	527854	413128	118367	519421	409510	120619
Συνολικός	8847554 ***(79,36%)***	6525583 ***(58,54%)***	1433579 ***(12,86%)***	8168121 ***(73,27%)***	6074162 ***(54,49%)***	1384880 ***(12,42%)***

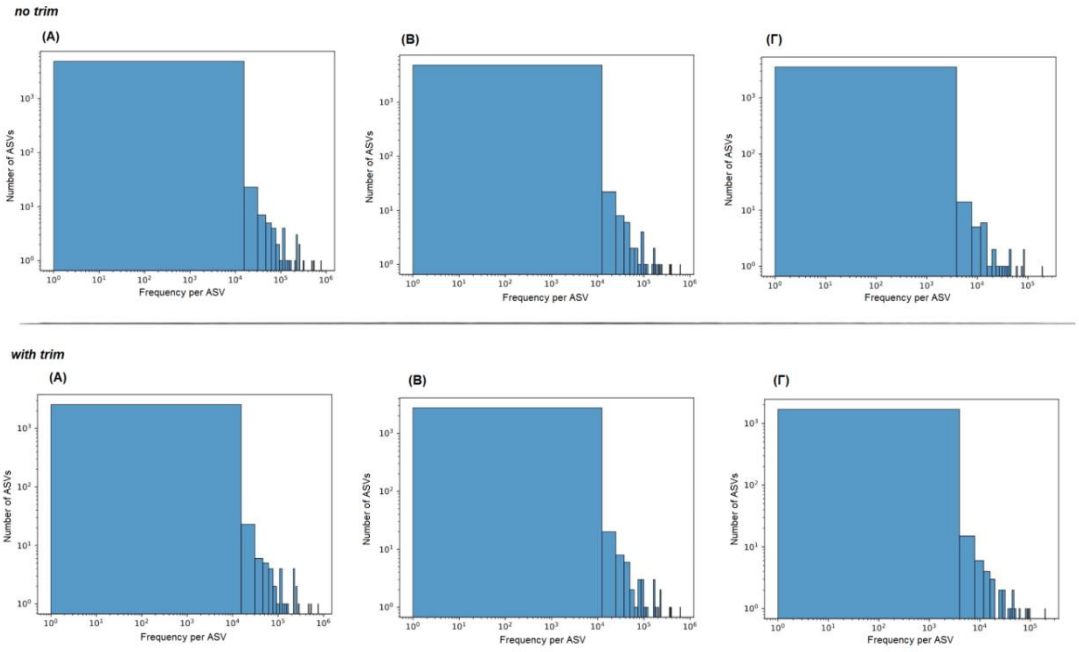


Παράρτημα 13. Ιστόγραμμα δειγμάτων που παρέχουν ένα δεδομένο αριθμό τελικών διατηρητέων αναγνωσμάτων του DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, όπου **(Α)** e.e.$_{max}$=2.5, **(Β)** e.e.$_{max}$=1.5 και **(Γ)** e.e.$_{max}$=0.5, και συνθήκες αποκοπής αναγνωσμάτων. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστης επικάλυψης δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 14. Γενική περιγραφή των συχνοτήτων που παρουσιάζουν συνολικά τα αντιπροσωπευτικά αναγνώσματα (ASVs) του DADA2 στις τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και συνθήκες αποκοπής αναγνωσμάτων.

Συχνότητα Αντιπροσωπευτικών Αναγνωσμάτων - ASVs									
	No trim								
e.e.$_{max}$	**2,5**			**1,5**			**0,5**		
Overlap$_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	1	1	1	1	1	1	1	1	1
Μέσος Όρος	1507	1532	1559	1141	1160	1181	363	367	371
Μέγιστη	775751	775751	775751	618921	618921	618921	194608	194608	194608

	With trim								
e.e.$_{max}$	**2,5**			**1,5**			**0,5**		
Overlap$_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	1	1	1	1	1	1	1	1	1
Μέσος Όρος	2491	2616	2695	1928	2026	2086	729	755	769
Μέγιστη	767045	767045	767045	617245	617245	617245	198484	198484	198484



Παράρτημα 15. Ιστόγραμμα αντιπροσωπευτικών αναγνωσμάτων (ASVs) που παρέχουν μία δεδομένη συχνότητα στον συνολικό όγκο αποτελεσμάτων του DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, όπου **(Α)** e.e.$_{max}$=2.5, **(Β)** e.e.$_{max}$=1.5 και **(Γ)** e.e.$_{max}$=0.5, και συνθήκες αποκοπής αναγνωσμάτων. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 16. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης Deblur στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, ελάχιστης βαθμολογίας ποιότητας και αποκοπή των συγχωνευμένων αναγνωσμάτων στην 250^{η} βάση.

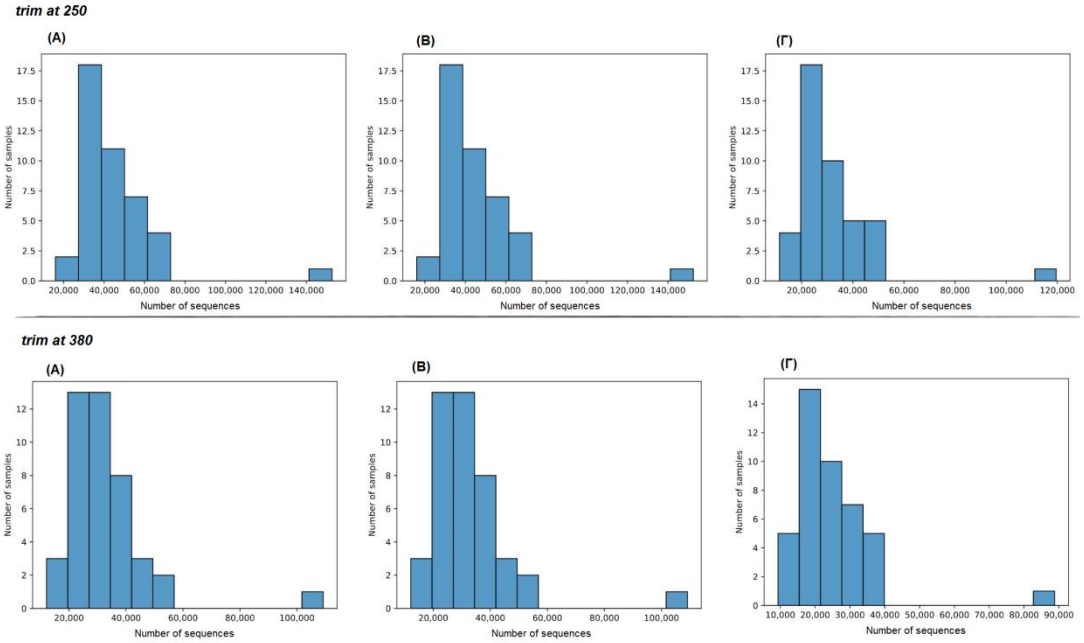
Αρ. Δείγματος	**Deblur με trim@=250**								
	Overlap$_{min}$=10			**Overlap$_{min}$=20**			**Overlap$_{min}$=30**		
	Q$_{min}$=20	**Q$_{min}$=22**	**Q$_{min}$=26**	**Q$_{min}$=20**	**Q$_{min}$=22**	**Q$_{min}$=26**	**Q$_{min}$=20**	**Q$_{min}$=22**	**Q$_{min}$=26**
sample1a	152680	152541	119524	152680	152541	119524	152680	152541	119524
sample1b	59361	59317	46276	59361	59317	46276	59361	59317	46276
sample2a	63830	63760	50588	63830	63760	50588	63830	63760	50588
sample2b	63403	63349	50492	63403	63349	50492	63403	63349	50492
sample3a	54581	54534	42901	54581	54534	42901	54581	54534	42901
sample3b	43847	43805	33347	43847	43805	33347	43847	43805	33347
sample4a	45965	45915	35751	45965	45915	35751	45965	45915	35751
sample4b	66478	66405	46748	66478	66405	46748	66478	66405	46748
sample5a	44527	44478	31959	44527	44478	31959	44527	44478	31959

sample5b	40641	40606	20957	40641	40606	20957	40641	40606	20957
sample6a	39084	39032	28006	39084	39032	28006	39084	39032	28006
sample6b	46229	46175	32891	46229	46175	32891	46229	46175	32891
sample7a	43718	43670	26832	43718	43670	26832	43718	43670	26832
sample7b	43423	43384	30849	43423	43384	30849	43423	43384	30849
sample8a	41963	41915	29427	41963	41915	29427	41963	41915	29427
sample8b	53035	52964	36733	53035	52964	36733	53035	52964	36733
sample9a	28910	28871	19596	28910	28871	19596	28910	28871	19596
sample9b	34252	34199	23860	34252	34199	23860	34252	34199	23860
sample10a	15933	15907	11805	15933	15907	11805	15933	15907	11805
sample10b	33401	33357	23964	33401	33357	23964	33401	33357	23964
sample11a	29857	29818	20136	29857	29818	20136	29857	29818	20136
sample11b	46655	46594	34443	46655	46594	34443	46655	46594	34443
sample12a	17156	17123	11307	17156	17123	11307	17156	17123	11307
sample12b	29761	29725	21641	29761	29725	21641	29761	29725	21641
sample13a	54202	54143	36865	54202	54143	36865	54202	54143	36865
sample13b	31637	31596	21271	31637	31596	21271	31637	31596	21271
sample14a	44376	44322	29830	44376	44322	29830	44376	44322	29830
sample14b	52578	52523	37910	52578	52523	37910	52578	52523	37910
sample15a	34621	34592	21896	34621	34592	21896	34621	34592	21896
sample15b	30849	30808	22829	30849	30808	22829	30849	30808	22829
sample16a	38569	38526	26429	38569	38526	26429	38569	38526	26429
sample16b	29576	29546	18622	29576	29546	18622	29576	29546	18622
sample17a	51120	51063	34069	51120	51063	34069	51120	51063	34069
sample17b	36390	36337	24905	36390	36337	24905	36390	36337	24905
sample18a	28619	28582	20673	28619	28582	20673	28619	28582	20673
sample18b	28405	28362	20047	28405	28362	20047	28405	28362	20047
sample19a	27943	27906	20051	27943	27906	20051	27943	27906	20051
sample19b	38471	38410	26995	38471	38410	26995	38471	38410	26995
sample20a	34239	34193	24626	34239	34193	24626	34239	34193	24626
sample20b	35054	35009	24377	35054	35009	24377	35054	35009	24377
negativecontrol1	54754	54681	36699	54754	54681	36699	54754	54681	36699
negativecontrol2	66855	66773	49696	66855	66773	49696	66855	66773	49696
negativecontrol3	38369	38328	25677	38369	38328	25677	38369	38328	25677

Παράρτημα 17. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης Deblur στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, ελάχιστης βαθμολογίας ποιότητας και αποκοπή των συγχωνευμένων αναγνωσμάτων στην 380^{η} βάση.

Αρ. Δείγματος	**Deblur με trim@=380**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Q_{min}=20	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**
sample1a	109189	109106	88782	109189	109106	88782	109189	109106	88782
sample1b	40917	40890	33193	40917	40890	33193	40917	40890	33193

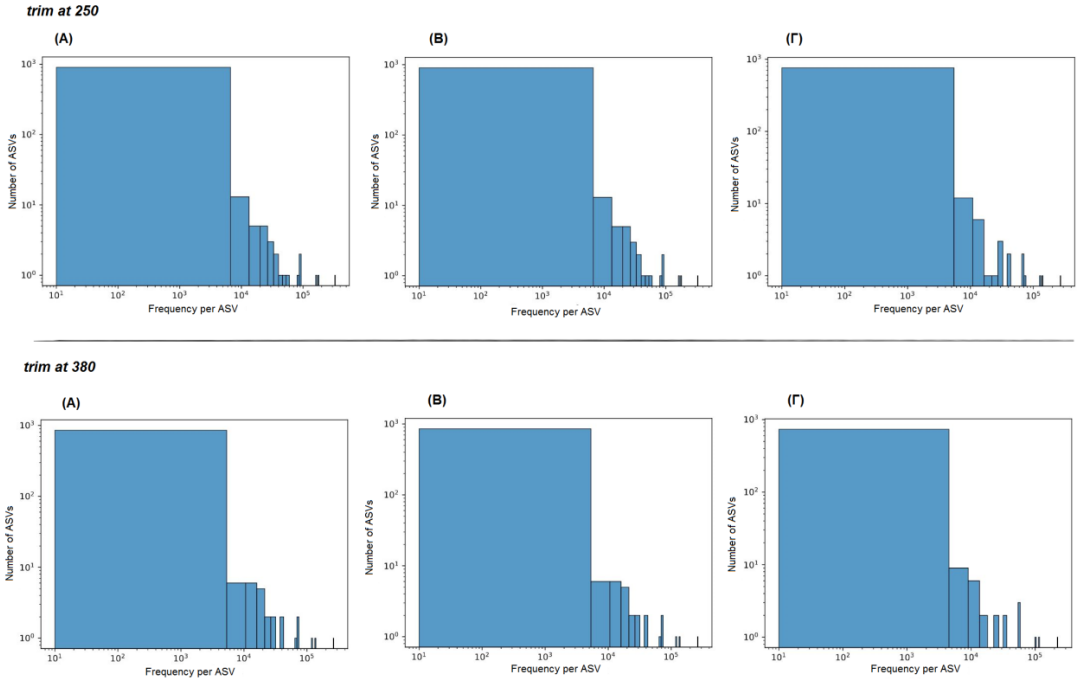
sample2a	45948	45908	37841	45948	45908	37841	45948	45908	37841
sample2b	46935	46905	38738	46935	46905	38738	46935	46905	38738
sample3a	41641	41604	33938	41641	41604	33938	41641	41604	33938
sample3b	33139	33118	26348	33139	33118	26348	33139	33118	26348
sample4a	34622	34597	27962	34622	34597	27962	34622	34597	27962
sample4b	51952	51902	38182	51952	51902	38182	51952	51902	38182
sample5a	34213	34180	25908	34213	34180	25908	34213	34180	25908
sample5b	27484	27464	16356	27484	27464	16356	27484	27464	16356
sample6a	30069	30030	22689	30069	30030	22689	30069	30030	22689
sample6b	33804	33773	25566	33804	33773	25566	33804	33773	25566
sample7a	33557	33530	22147	33557	33530	22147	33557	33530	22147
sample7b	34394	34370	25789	34394	34370	25789	34394	34370	25789
sample8a	32256	32229	23946	32256	32229	23946	32256	32229	23946
sample8b	37370	37334	27903	37370	37334	27903	37370	37334	27903
sample9a	20170	20151	14757	20170	20151	14757	20170	20151	14757
sample9b	24429	24400	18149	24429	24400	18149	24429	24400	18149
sample10a	12112	12099	9407	12112	12099	9407	12112	12099	9407
sample10b	24577	24552	18744	24577	24552	18744	24577	24552	18744
sample11a	22232	22210	15937	22232	22210	15937	22232	22210	15937
sample11b	35681	35642	27827	35681	35642	27827	35681	35642	27827
sample12a	13311	13287	9322	13311	13287	9322	13311	13287	9322
sample12b	20999	20979	16068	20999	20979	16068	20999	20979	16068
sample13a	36141	36114	26598	36141	36114	26598	36141	36114	26598
sample13b	23149	23127	16609	23149	23127	16609	23149	23127	16609
sample14a	33464	33429	23872	33464	33429	23872	33464	33429	23872
sample14b	36350	36328	27949	36350	36328	27949	36350	36328	27949
sample15a	22514	22499	15420	22514	22499	15420	22514	22499	15420
sample15b	24024	23993	18692	24024	23993	18692	24024	23993	18692
sample16a	28084	28056	20439	28084	28056	20439	28084	28056	20439
sample16b	17578	17564	12247	17578	17564	12247	17578	17564	12247
sample17a	40373	40327	28373	40373	40327	28373	40373	40327	28373
sample17b	28815	28781	21016	28815	28781	21016	28815	28781	21016
sample18a	22560	22528	16986	22560	22528	16986	22560	22528	16986
sample18b	21606	21577	15984	21606	21577	15984	21606	21577	15984
sample19a	21196	21169	15750	21196	21169	15750	21196	21169	15750
sample19b	29574	29534	21877	29574	29534	21877	29574	29534	21877
sample20a	26108	26082	19746	26108	26082	19746	26108	26082	19746
sample20b	26403	26373	19517	26403	26373	19517	26403	26373	19517
negativecontrol1	43276	43226	30546	43276	43226	30546	43276	43226	30546
negativecontrol2	50687	50635	39335	50687	50635	39335	50687	50635	39335
negativecontrol3	29515	29492	20751	29515	29492	20751	29515	29492	20751



Παράρτημα 18. Ιστόγραμμα δειγμάτων που παρέχουν ένα δεδομένο αριθμό τελικών διατηρητέων αναγνωσμάτων του Deblur στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, όπου **(Α)** Q_{min}=20, **(Β)** Q_{min}=22 και **(Γ)** Q_{min}=26, και συνθήκες αποκοπής συγχωνευμένων αναγνωσμάτων. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 19. Γενική περιγραφή των συχνοτήτων που παρουσιάζουν συνολικά τα αντιπροσωπευτικά αναγνώσματα (ASVs) του Deblur στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, μήκους επικαλυπτόμενης περιοχής και σημείων αποκοπής.

Συχνότητα Αντιπροσωπευτικών Αναγνωσμάτων - ASVs									
	trim@=250								
Q_{min}	**20**			**22**			**26**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	10	10	10	10	10	10	10	10	10
Μέσος Όρος	2016	2016	2016	2016	2016	2016	1711	1711	1711
Μέγιστη	333281	333281	333281	332984	332984	332984	272693	272693	272693
	trim@=380								
Q_{min}	**20**			**22**			**26**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	10	10	10	10	10	10	10	10	10
Μέσος Όρος	1586	1586	1586	1585	1585	1585	1387	1387	1387
Μέγιστη	266392	266392	266392	266201	266201	266201	225615	225615	225615

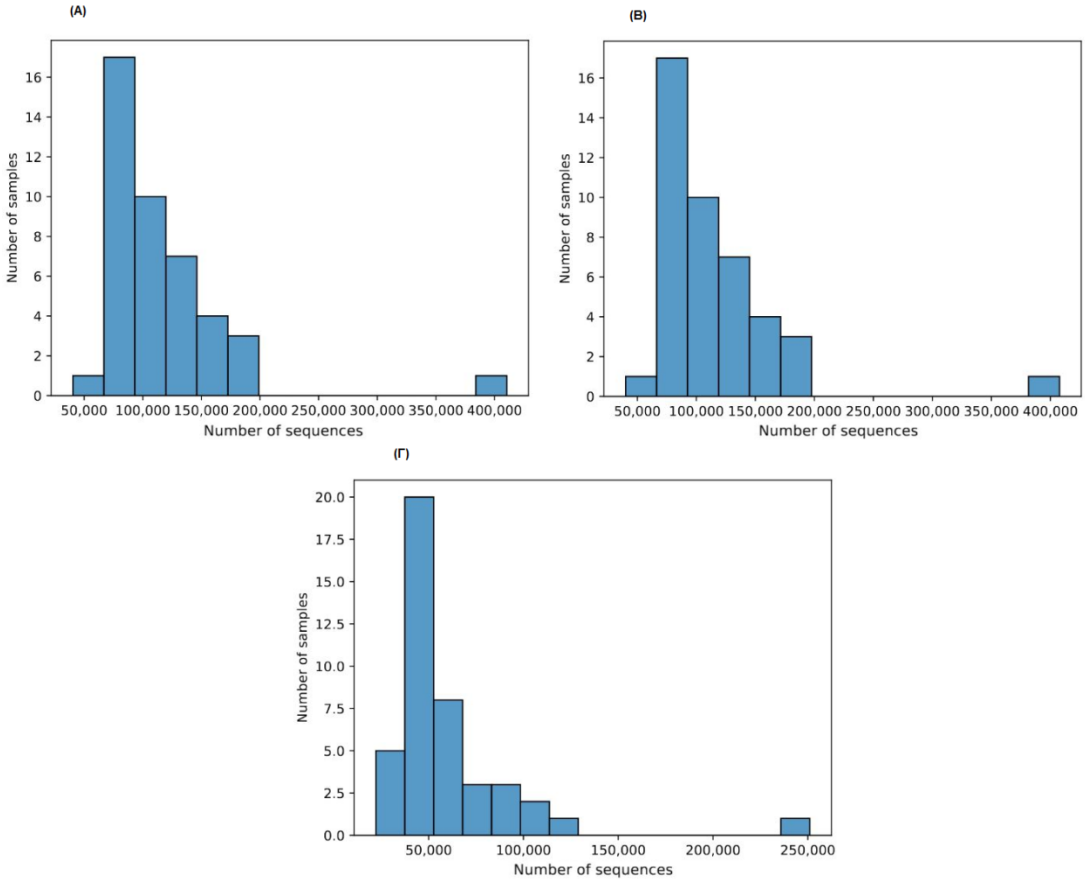


Παράρτημα 20. Ιστόγραμμα αντιπροσωπευτικών αναγνωσμάτων (ASVs) που παρέχουν μία δεδομένη συχνότητα στο σύνολο των αποτελεσμάτων του Deblur στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, όπου **(Α)** Q_{min}=20, **(Β)** Q_{min}=22 και **(Γ)** Q_{min}=26, και σημείων αποκοπής συγχωνευμένων αναγνωσμάτων. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα

Παράρτημα 21. Πίνακας με τον αριθμό διατηρητέων αναγνωσμάτων ανά δείγμα μετά της εφαρμογή του πρόσθετου αποθορυβοποίησης VSEARCH στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας.

Αρ. Δείγματος	**VSEARCH**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Q_{min}=20	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**
sample1a	410488	407800	251040	410488	407800	251040	410488	407800	251040
sample1b	171512	170421	102610	171512	170421	102610	171512	170421	102610
sample2a	185177	183958	114192	185177	183958	114192	185177	183958	114192
sample2b	178341	177133	110951	178341	177133	110951	178339	177131	110951
sample3a	152287	151500	93985	152287	151500	93985	152287	151500	93985
sample3b	130129	129409	75217	130129	129409	75217	130129	129409	75217
sample4a	134378	133512	79990	134378	133512	79990	134378	133512	79990
sample4b	176588	175586	96808	176588	175586	96808	176588	175586	96808
sample5a	98859	98050	53168	98859	98050	53168	98859	98050	53168
sample5b	72421	71904	34984	72421	71904	34984	72421	71904	34984
sample6a	92035	91204	49424	92035	91204	49424	92035	91204	49424
sample6b	104373	103561	55266	104373	103561	55266	104373	103561	55266
sample7a	108837	108001	47339	108837	108001	47339	108837	108001	47339
sample7b	99993	99176	52832	99993	99176	52832	99993	99176	52832
sample8a	98229	97352	50627	98229	97352	50627	98228	97351	50627
sample8b	124340	123446	63121	124340	123446	63121	124340	123446	63121
sample9a	81914	81208	39737	81914	81208	39737	81914	81208	39737

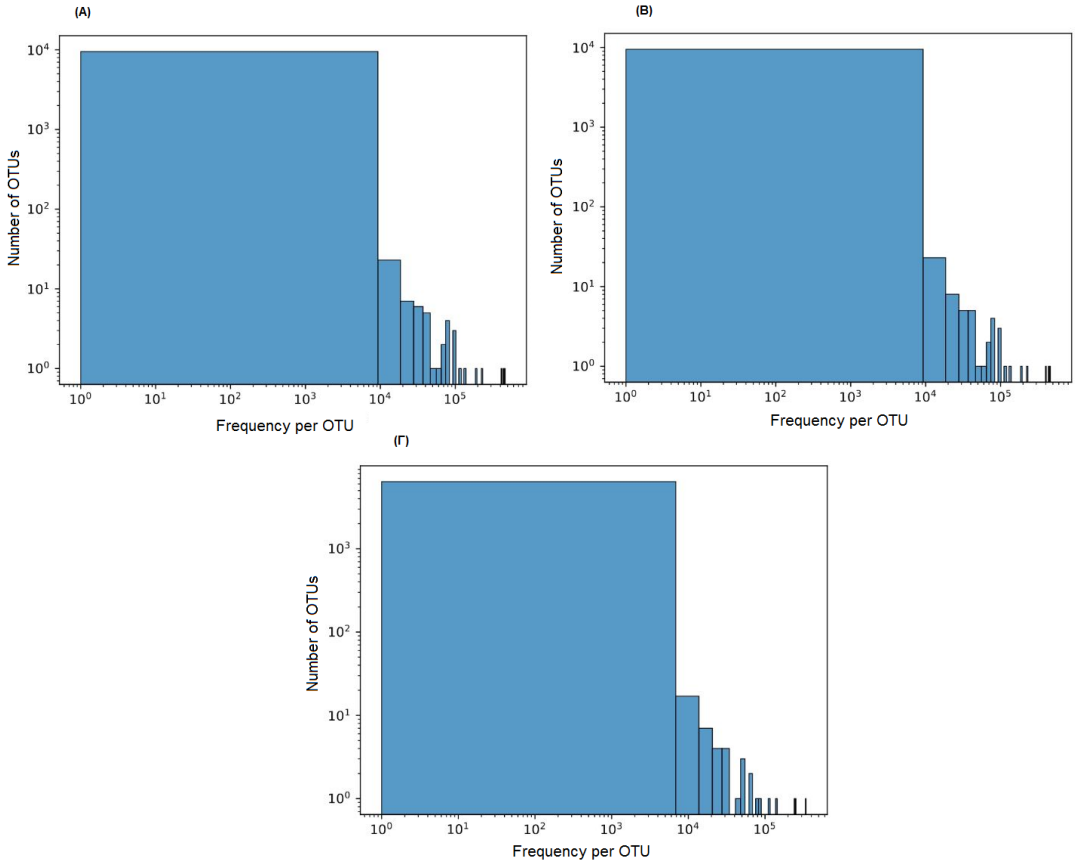
sample9b	94782	93911	47866	94782	93911	47866	94782	93911	47866
sample10a	40389	40029	22129	40389	40029	22129	40389	40029	22129
sample10b	84898	84211	44899	84898	84211	44899	84898	84211	44899
sample11a	72687	72128	36460	72687	72128	36460	72687	72128	36460
sample11b	124224	123121	67498	124224	123121	67498	124224	123121	67498
sample12a	72156	71036	29559	72156	71036	29559	72156	71036	29559
sample12b	78730	78103	42878	78730	78103	42878	78730	78103	42878
sample13a	134039	133047	67506	134039	133047	67506	134039	133047	67506
sample13b	81086	80479	39976	81086	80479	39976	81086	80479	39976
sample14a	103993	103190	52579	103993	103190	52579	103993	103190	52579
sample14b	123439	122576	67260	123439	122576	67260	123439	122576	67260
sample15a	93954	93356	42077	93954	93356	42077	93954	93356	42077
sample15b	69408	68876	39011	69408	68876	39011	69408	68876	39011
sample16a	97481	96718	49584	97481	96718	49584	97481	96718	49584
sample16b	77991	77488	34953	77991	77488	34953	77991	77488	34953
sample17a	120462	119452	58205	120462	119452	58205	120462	119452	58205
sample17b	78417	77794	39495	78417	77794	39495	78417	77794	39495
sample18a	79662	78933	41748	79662	78933	41748	79662	78933	41748
sample18b	77920	77242	40795	77920	77242	40795	77920	77242	40795
sample19a	77527	76838	41133	77527	76838	41133	77527	76838	41133
sample19b	98341	97494	51292	98341	97494	51292	98341	97494	51292
sample20a	92988	92167	49366	92988	92167	49366	92988	92167	49366
sample20b	92991	92160	48598	92991	92160	48598	92991	92160	48598
negativecontrol1	151432	150201	72560	151432	150201	72560	151432	150201	72560
negativecontrol2	168183	166991	95890	168183	166991	95890	168182	166990	95890
negativecontrol3	86598	86116	44494	86598	86116	44494	86598	86116	44494



Παράρτημα 22. Ιστόγραμμα δειγμάτων που παρέχουν ένα δεδομένο αριθμό τελικών διατηρητέων αναγνωσμάτων του VSEARCH στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας, όπου **(Α)** Q_{min}=20, **(Β)** Q_{min}=22 και **(Γ)** Q_{min}=26. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 23. Γενική περιγραφή των συχνοτήτων που παρουσιάζουν συνολικά οι λειτουργικές ταξινομικές μονάδες (OTUs) του VSEARCH στις διάφορες τιμές ελάχιστης βαθμολογίας ποιότητας και ελάχιστου μήκους επικαλυπτόμενης περιοχής.

Συχνότητα Λειτουργικών Ταξινομικών Μονάδων - OTUs									
Q_{min}	**20**			**22**			**26**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	1	1	1	1	1	1	1	1	1
Μέσος Όρος	511	511	511	508	508	508	410	410	410
Μέγιστη	464371	464371	464371	462980	462980	462980	343130	343130	343130



Παράρτημα 24. Ιστόγραμμα λειτουργικών ταξινομικών μονάδων (OTUs) που παρέχουν μία δεδομένη συχνότητα στο σύνολο των αποτελεσμάτων του VSEARCH στις διάφορες τιμές βαθμολογίας ποιότητας, όπου **(Α)** Q_{min}=20, **(Β)** Q_{min}=22 και **(Γ)** Q_{min}=26. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 25. Πίνακας με τον αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα τις ροής επεξεργασίας του DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και χωρίς αποκοπή των paired-end αναγνωσμάτων.

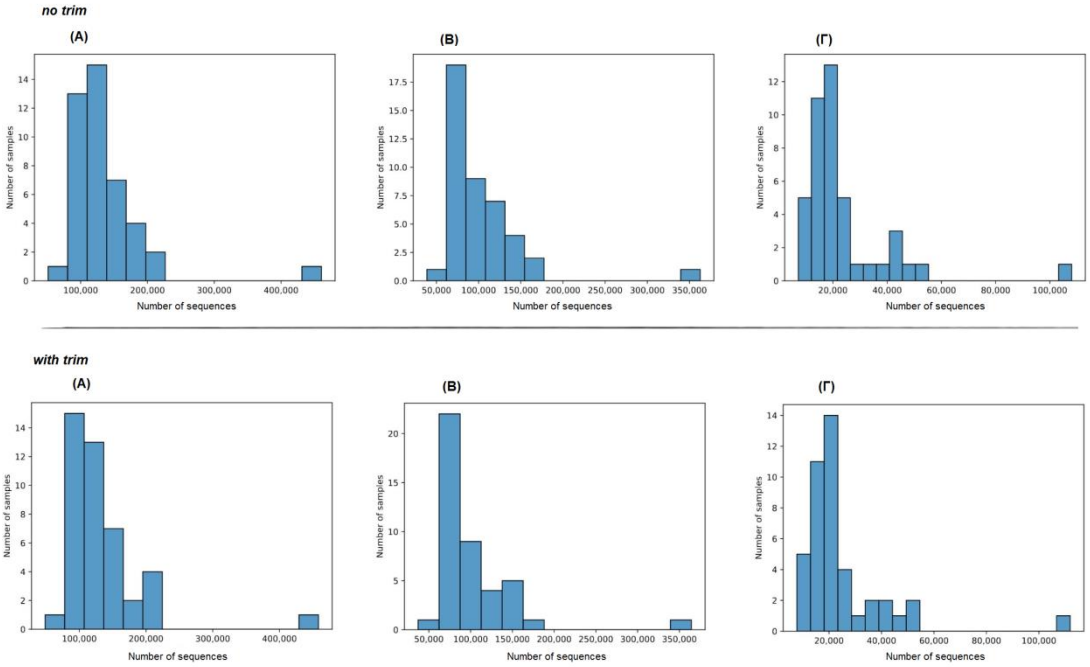
Αρ. Δείγματος	**DADA2 με no-trim**								
	e.e$_{max}$=2,5			**e.e$_{max}$=1,5**			**e.e$_{max}$=0,5**		
	Overlap$_{min}$=10	**Overlap$_{min}$=20**	**Overlap$_{min}$=30**	**Overlap$_{min}$=10**	**Overlap$_{min}$=20**	**Overlap$_{min}$=30**	**Overlap$_{min}$=10**	**Overlap$_{min}$=20**	**Overlap$_{min}$=30**
sample1a	460422	460422	460422	362722	362722	362722	108024	107968	107968
sample1b	195682	195682	195574	151894	151857	151788	41096	41042	41042
sample2a	208096	208006	207992	165082	165008	164996	50070	49984	49984
sample2b	201790	201723	201010	161044	160979	160533	50730	50652	50644
sample3a	173650	173576	173576	138744	138642	138642	41736	41677	41723
sample3b	149290	149290	149254	115468	115468	115442	28850	28850	28850
sample4a	154409	154295	154295	121862	121769	121769	34105	34040	34040
sample4b	194209	194092	193879	153781	153650	153498	42872	42716	42705
sample5a	125898	125898	125766	91242	91242	91154	22181	22181	22160
sample5b	122659	122606	122408	84092	84053	83920	16568	16568	16557
sample6a	115534	115456	115278	83592	83517	83389	20649	20574	20576
sample6b	130937	130937	130937	92440	92440	92440	21294	21294	21294

sample7a	131170	131170	131170	92119	92119	92119	19025	18994	19006
sample7b	122411	122411	122411	87749	87749	87749	21632	21632	21617
sample8a	121173	121173	121157	87155	87155	87139	20897	20853	20839
sample8b	158719	158719	158102	110657	110657	110335	20148	20148	20144
sample9a	103335	103285	103285	70639	70588	70588	10091	10048	10060
sample9b	116658	116658	116658	83183	83183	83183	13491	13491	13503
sample10a	51231	51066	50908	38630	38497	38381	8631	8534	8545
sample10b	107614	107222	107222	78746	78459	78459	15100	15017	15017
sample11a	95545	95424	95186	69880	69792	69646	13685	13615	13572
sample11b	152080	151854	151798	112226	112073	112017	23671	23551	23530
sample12a	100478	100478	100478	66157	66157	66157	7450	7450	7450
sample12b	96767	96436	96343	72796	72488	72378	16751	16584	16545
sample13a	162616	162583	162523	117246	117246	117206	19249	19215	19199
sample13b	98174	98128	98066	71179	71114	71069	13239	13182	13170
sample14a	122495	122495	122317	90836	90836	90708	19337	19337	19357
sample14b	148954	148954	148909	109406	109406	109361	22087	22039	22013
sample15a	114540	114455	114377	81684	81604	81548	11844	11805	11797
sample15b	88786	88786	88658	67051	67051	66992	15386	15341	15312
sample16a	119146	119146	119110	87780	87780	87744	17759	17652	17625
sample16b	101763	101693	101373	67760	67696	67435	7249	7249	7219
sample17a	138167	138167	138081	99148	99148	99088	23333	23300	23300
sample17b	96472	96472	96472	67981	67981	67981	15055	15055	15047
sample18a	96556	96556	96532	72430	72374	72374	15795	15729	15706
sample18b	93603	93540	93389	69416	69395	69304	13815	13703	13700
sample19a	92097	92047	91987	69131	69093	69052	15459	15341	15341
sample19b	123287	123069	123069	92673	92523	92523	19789	19727	19748
sample20a	110957	110887	110830	83672	83604	83562	17992	17889	17854
sample20b	111079	111009	110897	83622	83524	83440	18646	18608	18566
negativecontrol1	163899	163832	163832	123790	123790	123790	25710	25710	25710
negativecontrol2	193386	193386	193386	147573	147573	147573	36711	36711	36711
negativecontrol3	101103	101083	101083	78642	78642	78642	18687	18667	18667

Παράρτημα 26. Πίνακας με τον αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα τις ροής επεξεργασίας του DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενης περιοχής και με αποκοπή των paired-end αναγνωσμάτων.

Αρ. Δείγματος	**DADA2 με with-trim**								
	e.e$_{max}$=2,5			**e.e$_{max}$=1,5**			**e.e$_{max}$=0,5**		
	Overlap$_{min}$=10	**Overlap$_{min}$=20**	**Overlap$_{min}$=30**	**Overlap$_{min}$=10**	**Overlap$_{min}$=20**	**Overlap$_{min}$=30**	**Overlap$_{min}$=10**	**Overlap$_{min}$=20**	**Overlap$_{min}$=30**
sample1a	458539	456249	456249	364382	363584	363584	111922	111763	111763
sample1b	195417	193994	193994	152452	151784	151784	42574	42481	42481
sample2a	206724	205277	205263	165228	164548	164536	51301	51214	51214
sample2b	199745	198268	198239	160648	159903	159884	52213	52145	52141
sample3a	172004	170835	170835	138631	138041	138041	42941	42845	42845
sample3b	148003	147104	147068	115252	114783	114756	29889	29815	29815

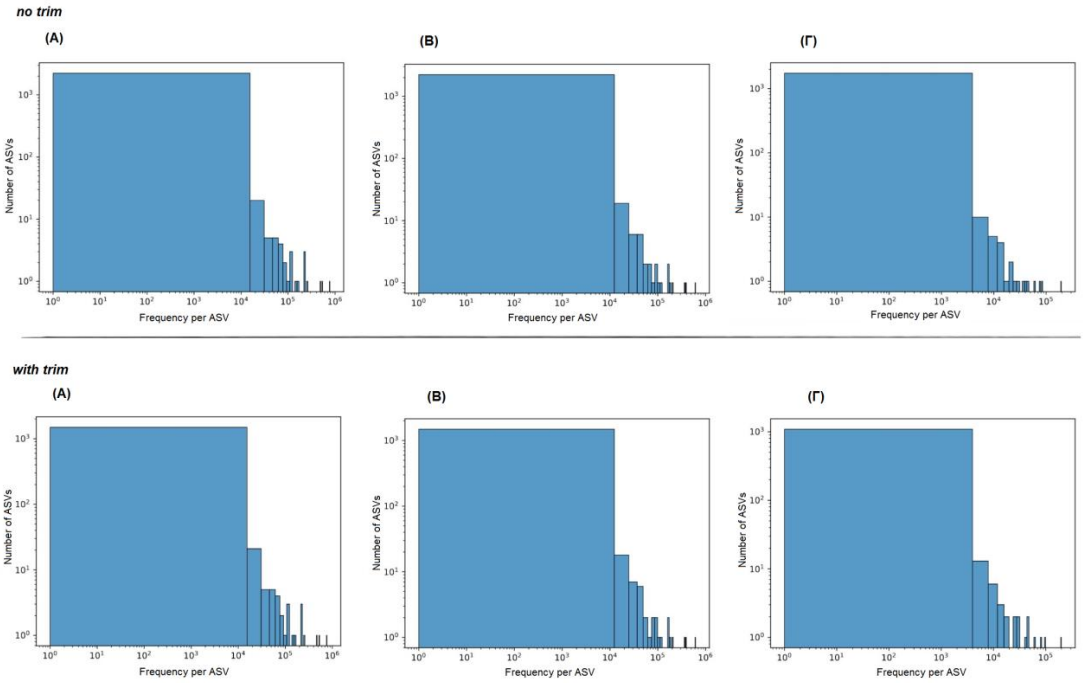
sample4a	152324	151143	151143	121231	120666	120666	34760	34699	34699
sample4b	195267	194364	194129	155188	154646	154498	44667	44583	44573
sample5a	120276	119851	119836	87899	87843	87833	22448	22434	22431
sample5b	118629	118461	118261	82646	82561	82424	17125	17125	17113
sample6a	112489	112158	111983	82467	82319	82192	21826	21818	21802
sample6b	128607	128196	128189	92277	92090	92084	22938	22929	22929
sample7a	128763	128486	128509	91737	91598	91598	20060	20060	20060
sample7b	120101	119727	119727	87226	87127	87127	23110	23110	23110
sample8a	113830	113673	113673	82818	82760	82760	21480	21480	21480
sample8b	152794	152384	152384	108359	108329	108329	21797	21784	21784
sample9a	100455	99931	99931	70025	69791	69791	11004	10994	10994
sample9b	115436	114845	114855	83646	83360	83369	14887	14873	14873
sample10a	48782	48482	48332	37061	36882	36772	8510	8490	8478
sample10b	103568	103181	103181	77090	76854	76854	16134	16116	16116
sample11a	91754	91244	91244	68118	67942	67942	13881	13865	13865
sample11b	149958	148161	148159	111544	111059	111057	25147	25115	25115
sample12a	94429	93688	93688	62912	62584	62584	7740	7740	7740
sample12b	92143	91235	91283	69820	69389	69415	16441	16403	16403
sample13a	157782	157331	157342	115467	115239	115247	20803	20787	20787
sample13b	96531	96350	96313	71057	70958	70929	14582	14569	14569
sample14a	119408	119156	119125	89575	89436	89411	21049	21049	21049
sample14b	145950	145362	145362	108761	108563	108563	23903	23899	23899
sample15a	107083	107027	106847	79448	79412	79307	12828	12826	12826
sample15b	82809	82613	82613	63015	62987	62987	15597	15593	15587
sample16a	112808	112554	112554	84323	84185	84185	18650	18632	18632
sample16b	99313	99298	99006	67563	67519	67284	8248	8244	8244
sample17a	135759	135678	135593	98343	98294	98233	24730	24718	24703
sample17b	92993	92846	92846	66576	66501	66501	15764	15759	15759
sample18a	93203	92957	92957	70609	70490	70490	16189	16173	16173
sample18b	90627	90174	90043	67694	67475	67402	14468	14430	14430
sample19a	89054	88778	88719	67767	67642	67600	15818	15802	15802
sample19b	117696	117121	117121	89343	89005	89005	20223	20180	20180
sample20a	108732	108392	108367	82347	82144	82122	18733	18691	18688
sample20b	105668	105304	105222	80980	80806	80753	19311	19284	19278
negativecontrol1	165054	165133	164893	126306	126317	126198	28116	28072	28068
negativecontrol2	191131	190804	190804	147370	147149	147149	38768	38717	38717
negativecontrol3	99928	99606	99606	78698	78539	78539	20444	20423	20423



Παράρτημα 27. Ιστόγραμμα δειγμάτων που παρέχουν ένα δεδομένο αριθμό ταξινομημένων αναγνωσμάτων του DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, όπου **(Α)** $e.e._{max}$=2.5, **(Β)** $e.e._{max}$=1.5 και **(Γ)** $e.e._{max}$=0.5, και συνθήκες αποκοπής paired-end αναγνωσμάτων. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστου μήκους επικαλυπτόμενης περιοχής δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 28. Γενική περιγραφή των συχνοτήτων που παρουσιάζουν συνολικά τα ταξινομημένα αντιπροσωπευτικά αναγνώσματα (ASVs) του DADA2 στις διαφορές τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, ελάχιστου μήκους επικαλυπτόμενςη περιοχής και συνθήκες αποκοπής paired-end αναγνωσμάτων.

Συχνότητα Ταξινομημένων Αντιπροσωπευτικών Αναγνωσμάτων - ASVs									
	No trim								
$e.e._{max}$	**2,5**			**1,5**			**0,5**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	1	1	1	1	1	1	1	1	1
Μέσος Όρος	2565	2594	2628	1922	1943	1969	572	578	585
Μέγιστη	775751	775751	775751	618921	618921	618921	194608	194608	194608
	With trim								
$e.e._{max}$	**2,5**			**1,5**			**0,5**		
$Overlap_{min}$	**10**	**20**	**30**	**10**	**20**	**30**	**10**	**20**	**30**
Ελάχιστη	1	1	1	1	1	1	1	1	1
Μέσος Όρος	3698	3928	3987	2817	2991	3032	945	982	989
Μέγιστη	767045	767045	767045	617245	617245	617245	198484	198484	198484



Παράρτημα 29. Ιστογράμματα ταξινομημένων αντιπροσωπευτικών αναγνωσμάτων (ASVs) που παρέχουν μία δεδομένη συχνότητα στον συνολικό όγκο αποτελεσμάτων ταξινόμησης του DADA2 στις διάφορες τιμές μέγιστου ποσοστού αναμενόμενων σφαλμάτων, όπου **(Α)** $e.e._{max}$=2.5, **(Β)** $e.e._{max}$=1.5 και **(Γ)** $e.e._{max}$=0.5, και συνθήκες αποκοπής αναγνωσμάτων. Σημειώνεται ότι μεταξύ των διαφορετικών τιμών ελάχιστης επικάλυψης δεν παρατηρείται διαφορά στα γραφήματα και δεν παρουσιάζονται επιπρόσθετα.

Παράρτημα 30. Πίνακας με τον αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα της ροής επεξεργασίας του Deblur στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, ελάχιστης βαθμολογίας ποιότητας και αποκοπή των συγχωνευμένων αναγνωσμάτων στην 250η βάση.

Αρ. Δείγματος	**Deblur με trim@=250**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Q_{min}=20	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**
sample1a	152680	152541	119524	152680	152541	119524	152680	152541	119524
sample1b	59361	59317	46276	59361	59317	46276	59361	59317	46276
sample2a	63830	63760	50588	63830	63760	50588	63830	63760	50588
sample2b	63403	63349	50492	63403	63349	50492	63403	63349	50492
sample3a	54581	54534	42901	54581	54534	42901	54581	54534	42901
sample3b	43847	43805	33347	43847	43805	33347	43847	43805	33347
sample4a	45965	45915	35751	45965	45915	35751	45965	45915	35751
sample4b	66478	66405	46748	66478	66405	46748	66478	66405	46748
sample5a	44527	44478	31959	44527	44478	31959	44527	44478	31959
sample5b	40641	40606	20957	40641	40606	20957	40641	40606	20957
sample6a	39084	39032	28006	39084	39032	28006	39084	39032	28006
sample6b	46229	46175	32891	46229	46175	32891	46229	46175	32891
sample7a	43718	43670	26832	43718	43670	26832	43718	43670	26832
sample7b	43423	43384	30849	43423	43384	30849	43423	43384	30849
sample8a	41963	41915	29427	41963	41915	29427	41963	41915	29427
sample8b	53035	52964	36733	53035	52964	36733	53035	52964	36733

sample9a	28910	28871	19596	28910	28871	19596	28910	28871	19596
sample9b	34252	34199	23860	34252	34199	23860	34252	34199	23860
sample10a	15933	15907	11805	15933	15907	11805	15933	15907	11805
sample10b	33401	33357	23964	33401	33357	23964	33401	33357	23964
sample11a	29857	29818	20136	29857	29818	20136	29857	29818	20136
sample11b	46655	46594	34443	46655	46594	34443	46655	46594	34443
sample12a	17156	17123	11307	17156	17123	11307	17156	17123	11307
sample12b	29761	29725	21641	29761	29725	21641	29761	29725	21641
sample13a	54202	54143	36865	54202	54143	36865	54202	54143	36865
sample13b	31637	31596	21271	31637	31596	21271	31637	31596	21271
sample14a	44376	44322	29830	44376	44322	29830	44376	44322	29830
sample14b	52578	52523	37910	52578	52523	37910	52578	52523	37910
sample15a	34621	34592	21896	34621	34592	21896	34621	34592	21896
sample15b	30849	30808	22829	30849	30808	22829	30849	30808	22829
sample16a	38569	38526	26429	38569	38526	26429	38569	38526	26429
sample16b	29576	29546	18622	29576	29546	18622	29576	29546	18622
sample17a	51120	51063	34069	51120	51063	34069	51120	51063	34069
sample17b	36390	36337	24905	36390	36337	24905	36390	36337	24905
sample18a	28619	28582	20673	28619	28582	20673	28619	28582	20673
sample18b	28405	28362	20047	28405	28362	20047	28405	28362	20047
sample19a	27943	27906	20051	27943	27906	20051	27943	27906	20051
sample19b	38471	38410	26995	38471	38410	26995	38471	38410	26995
sample20a	34239	34193	24626	34239	34193	24626	34239	34193	24626
sample20b	35054	35009	24377	35054	35009	24377	35054	35009	24377
negativecontrol1	54754	54681	36699	54754	54681	36699	54754	54681	36699
negativecontrol2	66855	66773	49696	66855	66773	49696	66855	66773	49696
negativecontrol3	38369	38328	25677	38369	38328	25677	38369	38328	25677

Παράρτημα 31. Πίνακας με τον αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα της ροής επεξεργασίας του Deblur στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής, ελάχιστης βαθμολογίας ποιότητας και αποκοπή των συγχωνευμένων αναγνωσμάτων στην 380^{η} βάση.

Αρ. Δείγματος	**Deblur με trim@=380**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Q_{min}=20	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**
sample1a	109189	109106	88782	109189	109106	88782	109189	109106	88782
sample1b	40917	40890	33193	40917	40890	33193	40917	40890	33193
sample2a	45948	45908	37841	45948	45908	37841	45948	45908	37841
sample2b	46935	46905	38738	46935	46905	38738	46935	46905	38738
sample3a	41641	41604	33938	41641	41604	33938	41641	41604	33938
sample3b	33139	33118	26348	33139	33118	26348	33139	33118	26348
sample4a	34622	34597	27962	34622	34597	27962	34622	34597	27962
sample4b	51952	51902	38182	51952	51902	38182	51952	51902	38182
sample5a	34213	34180	25908	34213	34180	25908	34213	34180	25908
sample5b	27484	27464	16356	27484	27464	16356	27484	27464	16356

sample6a	30069	30030	22689	30069	30030	22689	30069	30030	22689
sample6b	33804	33773	25566	33804	33773	25566	33804	33773	25566
sample7a	33557	33530	22147	33557	33530	22147	33557	33530	22147
sample7b	34394	34370	25789	34394	34370	25789	34394	34370	25789
sample8a	32256	32229	23946	32256	32229	23946	32256	32229	23946
sample8b	37370	37334	27903	37370	37334	27903	37370	37334	27903
sample9a	20170	20151	14757	20170	20151	14757	20170	20151	14757
sample9b	24429	24400	18149	24429	24400	18149	24429	24400	18149
sample10a	12112	12099	9407	12112	12099	9407	12112	12099	9407
sample10b	24577	24552	18744	24577	24552	18744	24577	24552	18744
sample11a	22232	22210	15937	22232	22210	15937	22232	22210	15937
sample11b	35681	35642	27827	35681	35642	27827	35681	35642	27827
sample12a	13311	13287	9322	13311	13287	9322	13311	13287	9322
sample12b	20999	20979	16068	20999	20979	16068	20999	20979	16068
sample13a	36141	36114	26598	36141	36114	26598	36141	36114	26598
sample13b	23149	23127	16609	23149	23127	16609	23149	23127	16609
sample14a	33464	33429	23872	33464	33429	23872	33464	33429	23872
sample14b	36350	36328	27949	36350	36328	27949	36350	36328	27949
sample15a	22514	22499	15420	22514	22499	15420	22514	22499	15420
sample15b	24024	23993	18692	24024	23993	18692	24024	23993	18692
sample16a	28084	28056	20439	28084	28056	20439	28084	28056	20439
sample16b	17578	17564	12247	17578	17564	12247	17578	17564	12247
sample17a	40373	40327	28373	40373	40327	28373	40373	40327	28373
sample17b	28815	28781	21016	28815	28781	21016	28815	28781	21016
sample18a	22560	22528	16986	22560	22528	16986	22560	22528	16986
sample18b	21606	21577	15984	21606	21577	15984	21606	21577	15984
sample19a	21196	21169	15750	21196	21169	15750	21196	21169	15750
sample19b	29574	29534	21877	29574	29534	21877	29574	29534	21877
sample20a	26108	26082	19746	26108	26082	19746	26108	26082	19746
sample20b	26403	26373	19517	26403	26373	19517	26403	26373	19517
negativecontrol1	43276	43226	30546	43276	43226	30546	43276	43226	30546
negativecontrol2	50687	50635	39335	50687	50635	39335	50687	50635	39335
negativecontrol3	29515	29492	20751	29515	29492	20751	29515	29492	20751

Παράρτημα 32. Πίνακας με τον αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα της ροής επεξεργασίας του VSEARCH στις διάφορες τιμές ελάχιστου μήκους επικαλυπτόμενης περιοχής και ελάχιστης βαθμολογίας ποιότητας.

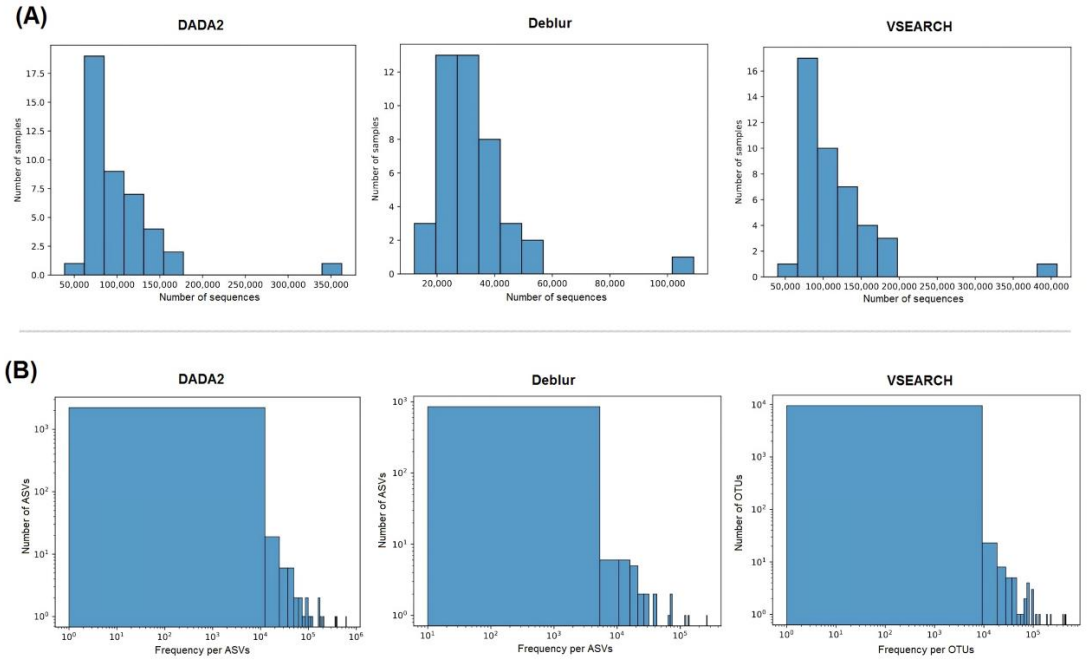
Αρ. Δείγματος	**VSEARCH**								
	$Overlap_{min}$=10			**$Overlap_{min}$=20**			**$Overlap_{min}$=30**		
	Q_{min}=20	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**	**Q_{min}=20**	**Q_{min}=22**	**Q_{min}=26**
sample1a	410488	407800	251040	410488	407800	251040	410488	407800	251040
sample1b	171512	170421	102610	171512	170421	102610	171512	170421	102610
sample2a	185177	183958	114192	185177	183958	114192	185177	183958	114192
sample2b	178341	177133	110951	178341	177133	110951	178339	177131	110951

sample3a	152287	151500	93985	152287	151500	93985	152287	151500	93985
sample3b	130129	129409	75217	130129	129409	75217	130129	129409	75217
sample4a	134378	133512	79990	134378	133512	79990	134378	133512	79990
sample4b	176588	175586	96808	176588	175586	96808	176588	175586	96808
sample5a	98859	98050	53168	98859	98050	53168	98859	98050	53168
sample5b	72421	71904	34984	72421	71904	34984	72421	71904	34984
sample6a	92035	91204	49424	92035	91204	49424	92035	91204	49424
sample6b	104373	103561	55266	104373	103561	55266	104373	103561	55266
sample7a	108837	108001	47339	108837	108001	47339	108837	108001	47339
sample7b	99993	99176	52832	99993	99176	52832	99993	99176	52832
sample8a	98229	97352	50627	98229	97352	50627	98228	97351	50627
sample8b	124340	123446	63121	124340	123446	63121	124340	123446	63121
sample9a	81914	81208	39737	81914	81208	39737	81914	81208	39737
sample9b	94782	93911	47866	94782	93911	47866	94782	93911	47866
sample10a	40389	40029	22129	40389	40029	22129	40389	40029	22129
sample10b	84898	84211	44899	84898	84211	44899	84898	84211	44899
sample11a	72687	72128	36460	72687	72128	36460	72687	72128	36460
sample11b	124224	123121	67498	124224	123121	67498	124224	123121	67498
sample12a	72156	71036	29559	72156	71036	29559	72156	71036	29559
sample12b	78730	78102	42878	78729	78102	42877	78729	78102	42877
sample13a	134039	133047	67506	134039	133047	67506	134039	133047	67506
sample13b	81086	80479	39976	81086	80479	39976	81086	80479	39976
sample14a	103993	103190	52579	103993	103190	52579	103993	103190	52579
sample14b	123439	122575	67260	123438	122575	67259	123438	122575	67259
sample15a	93954	93356	42077	93954	93356	42077	93954	93356	42077
sample15b	69408	68876	39011	69408	68876	39011	69408	68876	39011
sample16a	97481	96718	49584	97481	96718	49584	97481	96718	49584
sample16b	77991	77488	34953	77991	77488	34953	77991	77488	34953
sample17a	120462	119452	58205	120462	119452	58205	120462	119452	58205
sample17b	78417	77794	39495	78417	77794	39495	78417	77794	39495
sample18a	79662	78933	41748	79662	78933	41748	79662	78933	41748
sample18b	77920	77242	40795	77920	77242	40795	77920	77242	40795
sample19a	77527	76838	41133	77527	76838	41133	77527	76838	41133
sample19b	98341	97494	51292	98341	97494	51292	98341	97494	51292
sample20a	92988	92167	49366	92988	92167	49366	92988	92167	49366
sample20b	92991	92160	48598	92991	92160	48598	92991	92160	48598
negativecontrol1	151432	150201	72560	151432	150201	72560	151432	150201	72560
negativecontrol2	168183	166991	95890	168183	166991	95890	168182	166990	95890
negativecontrol3	86598	86115	44494	86597	86115	44493	86597	86115	44493

Παράρτημα 33. Πίνακας με τον αρχικό αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα για τα τρία κυρίαρχα σύνολα δεδομένων των επεξεργαστικών ροών των DADA2, Deblur και VSEARCH καθώς και τον αντίστοιχο αριθμό αναγνωσμάτων μετά το φιλτράρισμα μη-στοχευόμενων στοιχείων.

Αρ. Δείγματος	Αρχικά Κυριάρχα Σύνολα Δεδομένων			Φιλτράρισμα μη-στοχευόμενων στοιχείων		
	DADA2	Deblur	VSEARCH	DADA2	Deblur	VSEARCH
sample1a	362722	109106	407800	359280	108177	404121
sample1b	151894	40890	170421	151291	40890	170421
sample2a	165082	45908	183958	164139	45705	183183
sample2b	161044	46905	177133	158116	46280	174870
sample3a	138744	41604	151500	136223	41439	150961
sample3b	115468	33118	129409	114016	32981	128826
sample4a	121862	34597	133512	120429	34547	133296
sample4b	153781	51902	175586	144842	48178	162760
sample5a	91242	34180	98050	87114	34180	98049
sample5b	84092	27464	71904	56733	21351	65368
sample6a	83592	30030	91204	80466	30030	91204
sample6b	92440	33773	103561	89619	33773	103561
sample7a	92119	33530	108001	89952	33525	107980
sample7b	87749	34370	99176	84970	34277	98851
sample8a	87155	32229	97352	85802	32229	97352
sample8b	110657	37334	123446	107264	37315	123365
sample9a	70639	20151	81208	69184	20151	81208
sample9b	83183	24400	93911	81356	24400	93911
sample10a	38630	12099	40029	35796	12099	40029
sample10b	78746	24552	84211	75024	24552	84211
sample11a	69880	22210	72128	64754	22210	72127
sample11b	112226	35642	123121	109151	35642	123121
sample12a	66157	13287	71036	62837	13287	71036
sample12b	72796	20979	78102	67332	20979	78102
sample13a	117246	36114	133047	114337	36047	132735
sample13b	71179	23127	80479	67387	22966	79836
sample14a	90836	33429	103190	88445	33361	102926
sample14b	109406	36328	122575	107156	36249	122179
sample15a	81684	22499	93356	78716	22012	91116
sample15b	67051	23993	68876	61479	23993	68876
sample16a	87780	28056	96718	84122	27887	95997
sample16b	67760	17564	77488	64940	17564	77488
sample17a	99148	40327	119452	96308	40327	119452
sample17b	67981	28781	77794	65946	28663	77394
sample18a	72430	22528	78933	70681	22528	78933
sample18b	69416	21577	77242	67419	21577	77242
sample19a	69131	21169	76838	66224	21169	76837

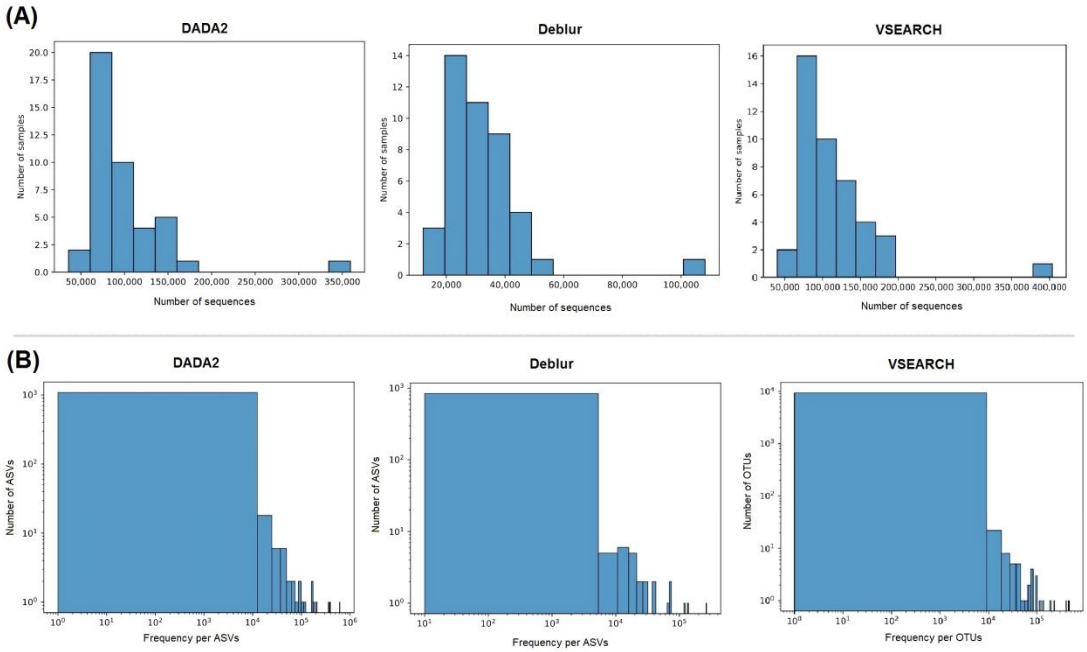
sample19b	92673	29534	97494	89955	29445	97218
sample20a	83672	26082	92167	79627	26022	91958
sample20b	83622	26373	92160	79847	26373	92160
negativecontrol1	123790	43226	150201	123755	43226	150201
negativecontrol2	147573	50635	166991	147539	50625	166952
negativecontrol3	78642	29492	86115	78642	29492	86115



Παράρτημα 34. (A) Ιστογράμματα δειγμάτων που παρέχουν ένα δεδομένο αριθμό ταξινομημένων αναγνωσμάτων και **(B)** ιστογράμματα αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που παρέχουν μία δεδομένη συχνότητα στο σύνολο των αποτελεσμάτων που προέκυψαν αντίστοιχα από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH.

Παράρτημα 35. Γενική περιγραφή των συχνοτήτων που παρουσιάζουν συνολικά τα ταξινομημένα αντιπροσωπευτικά αναγνώσματα (ASVs) και οι λειτουργικές ταξινομικές μονάδες (OTUs) που προέκυψαν από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH.

Συχνότητα Ταξινομημένων ASVs/OTUs			
pipeline	**DADA2**	**Deblur**	**VSEARCH**
Ελάχιστος	1	10	1
Μέσος Όρος	1922	1585	508
Μέγιστος	618921	266201	462980

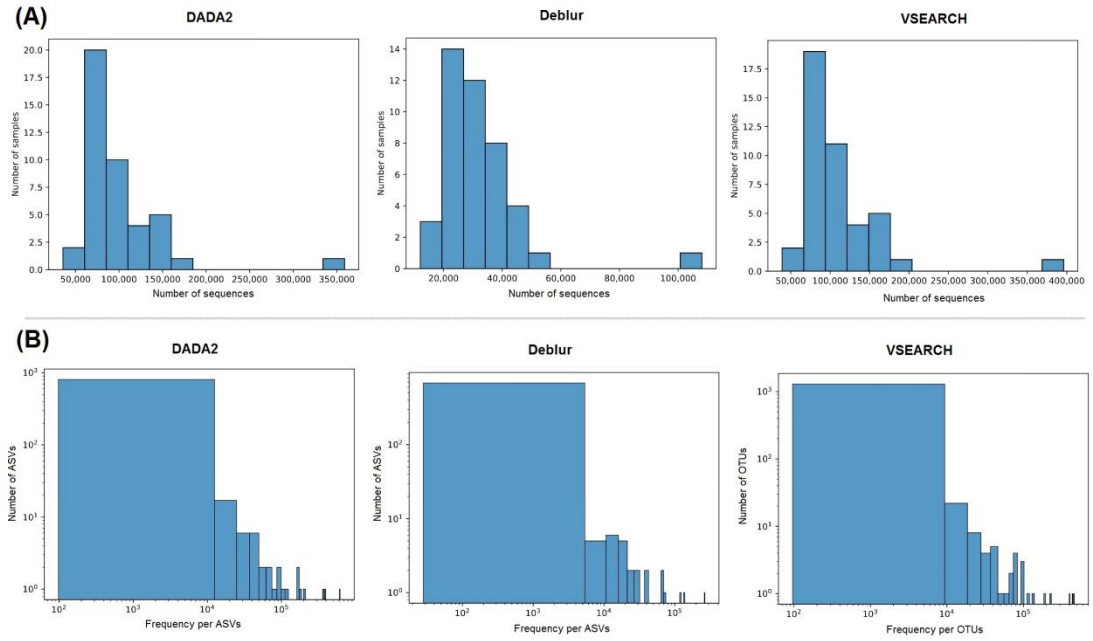


Παράρτημα 36. (A) Ιστογράμματα δειγμάτων που παρέχουν ένα δεδομένο αριθμό ταξινομημένων αναγνωσμάτων και **(B)** ιστογράμματα αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που παρέχουν μία δεδομένη συχνότητα στο σύνολο των αποτελεσμάτων που προέκυψαν αντίστοιχα από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά την αφαίρεση μη στοχευόμενων αποτελεσμάτων.

Παράρτημα 37. Πίνακας με τον αριθμό διατηρητέων ταξινομημένων αναγνωσμάτων ανά δείγμα για τα τρία κυρίαρχα σύνολα δεδομένων των επεξεργαστικών ροών των DADA2, Deblur και VSEARCH μετά την εφαρμογή φιλτραρίσματος χαμηλής αφθονίας ταξινομικών κατηγοριών καθώς και μετά την αφαίρεση πιθανής επιμόλυνσης από τα δεδομένα.

Αρ. Δείγματος	Φιλτράρισμα χαμηλής αφθονίας taxa			Αφαίρεση πιθανής Επιμολυνσης		
	DADA2	Deblur	VSEARCH	DADA2	Deblur	VSEARCH
sample1a	358992	108116	396257	103185	38910	108129
sample1b	151161	40884	168763	24215	9352	26244
sample2a	163998	45655	182111	22157	8459	24059
sample2b	157972	46262	173763	37097	13773	39302
sample3a	135708	41352	149811	37774	14194	41535
sample3b	113816	32972	127644	29240	10247	31996
sample4a	120279	34475	132452	30239	9982	32546
sample4b	144124	48057	159414	68321	25725	76007
sample5a	87089	34170	95937	86816	34056	95686
sample5b	56527	21314	63447	56527	21314	63447
sample6a	80189	29974	89633	80189	29974	89633
sample6b	89146	33711	102408	89146	33711	102408
sample7a	89227	33357	102486	89227	33357	102486
sample7b	84772	34208	97611	84772	34208	97611
sample8a	85534	32160	92780	85534	32160	92780
sample8b	107132	37240	120204	107132	37240	120204
sample9a	69183	20114	79217	69183	20114	79217
sample9b	81147	24312	92327	81147	24312	92327

sample10a	35610	12073	39000	35610	12073	39000
sample10b	75024	24535	83028	75024	24535	83028
sample11a	64549	22185	69546	64549	22185	69546
sample11b	108946	35589	120737	108946	35589	120737
sample12a	62753	13250	70247	62753	13250	70247
sample12b	66434	20873	76071	66434	20873	76071
sample13a	114035	35952	130736	113395	35801	130118
sample13b	66956	22885	77543	66956	22885	77543
sample14a	88148	33289	100935	88148	33289	100935
sample14b	106632	36137	119749	106632	36137	119749
sample15a	77919	21880	88267	77269	21639	87690
sample15b	61479	23966	67767	61479	23966	67767
sample16a	83987	27864	94269	83987	27864	94269
sample16b	64898	17538	76498	64898	17538	76498
sample17a	96105	40241	116326	96105	40241	116326
sample17b	65691	28617	75668	65691	28617	75668
sample18a	70618	22514	77539	70618	22514	77539
sample18b	67234	21524	76404	67234	21524	76404
sample19a	65901	21096	74978	65722	21039	74837
sample19b	89886	29410	95246	89886	29410	95246
sample20a	79581	25984	90304	79540	25972	90266
sample20b	79317	26216	90642	79317	26216	90642
negativecontrol1	123648	43194	146747	123648	43194	146747
negativecontrol2	147271	50563	163047	147271	50563	163046
negativecontrol3	78431	29422	84072	78413	29419	84048



Παράρτημα 38. (A) Ιστογράμματα δειγμάτων που παρέχουν ένα δεδομένο αριθμό ταξινομημένων αναγνωσμάτων και **(B)** ιστογράμματα αντιπροσωπευτικών αναγνωσμάτων (ASVs) και λειτουργικών ταξινομικών μονάδων (OTUs) που παρέχουν μία δεδομένη συχνότητα στο σύνολο των

αποτελεσμάτων που προέκυψαν αντίστοιχα από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά το φιλτράρισμα χαμηλής αφθονίας ταξινομικών κατηγοριών.

Παράρτημα 39. Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο φυλής των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH πριν και μετά το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

Φυλή	
Πρίν το φιλτράρισμα	**Μετά το φριλτράρισμα**
[DADA2] και [Deblur] και [VSEARCH]	**[DADA2] και [Deblur] και [VSEARCH]**
Proteobacteria	*Proteobacteria*
Actinobacteriota	*Actinobacteriota*
Firmicutes	*Acidobacteriota*
Acidobacteriota	*Firmicutes*
Deinococcota	*Deinococcota*
Bacteroidota	*Bacteroidota*
Patescibacteria	*Patescibacteria*
Fusobacteriota	*Fusobacteriota*
Dependentiae	*Dependentiae*
Cyanobacteria	*Cyanobacteria*
SAR324 clade (Marine group B)	*Planctomycetota*
Chloroflexi	*Verrucomicrobiota*
Planctomycetota	*Chloroflexi*
Verrucomicrobiota	*Bdellovibrionota*
Bdellovibrionota	*Campilobacterota*
Campilobacterota	**[DADA2] και [Deblur]**
[DADA2] και [VSEARCH]	*SAR324 clade (Marine group B)*
Sumerlaeota	
[DADA2]	
Spirochaetota	

Παράρτημα 40. Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο ομοταξίας των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH πριν και μετά το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

Ομοταξία	
Πριν το φιλτράρισμα	**Μετά το φιλτράρισμα**
[DADA2] και [Deblur] και [VSEARCH]	**[DADA2] και [Deblur] και [VSEARCH]**
Gammaproteobacteria	*Gammaproteobacteria*
Actinobacteria	*Actinobacteria*
Bacilli	*Vicinamibacteria*
Vicinamibacteria	*Alphaproteobacteria*
Alphaproteobacteria	*Bacilli*
Deinococci	*Deinococci*
Bacteroidia	*Bacteroidia*
Gracilibacteria	*Gracilibacteria*
Clostridia	*Clostridia*
Thermoanaerobacteria	*Thermoanaerobacteria*

Fusobacteriia	*Fusobacteriia*
Coriobacteriia	*Coriobacteriia*
Babeliae	*Babeliae*
Cyanobacteriia	*Cyanobacteriia*
Saccharimonadia	*Negativicutes*
Negativicutes	*Vampirivibrionia*
Chloroflexia	*Phycisphaerae*
Vampirivibrionia	*Thermoleophilia*
Blastocatellia	*Verrucomicrobiae*
Phycisphaerae	*Planctomycetes*
Thermoleophilia	*Rubrobacteria*
Verrucomicrobiae	*KD4-96*
Planctomycetes	*Saccharimonadia*
Rubrobacteria	*Bdellovibrionia*
KD4-96	*Desulfitobacteriia*
Acidimicrobiia	*TK10*
Bdellovibrionia	*Acidobacteriae*
Desulfitobacteriia	*Incertae Sedis*
TK10	*Thermaerobacteria*
Acidobacteriae	*Campylobacteria*
Incertae Sedis	*Blastocatellia*
Thermaerobacteria	**[DADA2] και [Deblur]**
Campylobacteria	*Acidimicrobiia*
[DADA2] και [Deblur]	**[Deblur]**
Symbiobacteriia	*Chloroflexia*
[DADA2] και [VSEARCH]	
Sumerlaeia	
[DADA2]	
Ktedonobacteria	

Παράρτημα 41. Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο τάξης των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH πριν και μετά το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

Τάξη			
Πριν το φιλτράρισμα		**Μετά το φιλτράρισμα**	
[DADA2] και [Deblur] και [VSEARCH]		**[DADA2] και [Deblur] και [VSEARCH]**	
Gammaproteobacteria Incertae Sedis	*Paenibacillales*	*Gammaproteobacteria Incertae Sedis*	*Paenibacillales*
Propionibacteriales	*Phormidesmiales*	*Propionibacteriales*	*Phormidesmiales*
Lactobacillales	*Diplorickettsiales*	*Vicinamibacterales*	*Azospirillales*
Vicinamibacterales	*Azospirillales*	*Sphingomonadales*	*Ferrovibrionales*
Sphingomonadales	*Ferrovibrionales*	*Rhizobiales*	*Kineosporiales*
Rhizobiales	*Kineosporiales*	*Lactobacillales*	*Thermoanaerobacterales*
Alteromonadales	*Saccharimonadales*	*Deinococcales*	*Lachnospirales*
Enterobacterales	*Thermoanaerobacterales*	*Flavobacteriales*	*Sphingobacteriales*

Pseudomonadales	*Lachnospirales*	*Absconditabacteriales (SR1)*	*Veillonellales-Selenomonadales*
Micrococcales	*Sphingobacteriales*	*Burkholderiales*	*Aeromonadales*
Deinococcales	*Veillonellales-Selenomonadales*	*Staphylococcales*	*Frankiales*
Flavobacteriales	*Aeromonadales*	*Acholeplasmatales*	*Acidaminococcales*
Absconditabacteriales (SR1)	*Frankiales*	*Enterobacterales*	*Obscuribacterales*
Thermincolales	*Acidaminococcales*	*Oceanospirillales*	*Phycisphaerales*
Burkholderiales	*Thermomicrobiales*	*Clostridiales*	*Thermincolales*
Staphylococcales	*Obscuribacterales*	*Vibrionales*	*Pedosphaerales*
Acholeplasmatales	*Blastocatellales*	*Pasteurellales*	*Brevibacillales*
Oceanospirillales	*Brevibacillales*	*Corynebacteriales*	*Rhodospirillales*
Clostridiales	*Phycisphaerales*	*Bacillales*	*Pirellulales*
Vibrionales	*Pedosphaerales*	*Micromonosporales*	*Caulobacterales*
Pasteurellales	*Oxyphotobacteria Incertae Sedis*	*Caldicellulosiruptorales*	*Micavibrionales*
Corynebacteriales	*Rhodospirillales*	*Cytophagales*	*Chitinophagales*
Bacillales	*Pirellulales*	*Alteromonadales*	*Cellvibrionales*
Micromonosporales	*Micavibrionales*	*Micrococcales*	*Rubrobacterales*
Caldicellulosiruptorales	*Chitinophagales*	*Pseudonocardiales*	*Gaiellales*
Caulobacterales	*Cellvibrionales*	*Rhodobacterales*	*Solirubrobacterales*
Cytophagales	*Rubrobacterales*	*Xanthomonadales*	*Saccharimonadales*
Pseudonocardiales	*Gaiellales*	*Pseudomonadales*	*Streptomycetales*
Rhodobacterales	*Solirubrobacterales*	*Tistrellales*	*Bdellovibrionales*
Xanthomonadales	*Streptomycetales*	*Fusobacteriales*	*Desulfitobacteriales*
Tistrellales	*Microtrichales*	*Exiguobacterales*	*Bryobacterales*
Bifidobacteriales	*Thermoactinomycetales*	*Thermales*	*Planctomycetales*
Fusobacteriales	*Bdellovibrionales*	*Coriobacteriales*	*DTU014*
Exiguobacterales	*Desulfitobacteriales*	*Peptostreptococcales-Tissierellales*	*Cyanobacteriales*
Thermales	*Acetobacterales*	*Legionellales*	*Bifidobacteriales*
Coriobacteriales	*Bryobacterales*	*Babeliales*	*Thermaerobacterales*
Peptostreptococcales-Tissierellales	*Planctomycetales*	*Actinomycetales*	*Campylobacterales*
Legionellales	*DTU014*	*Bacteroidales*	*Blastocatellales*
Babeliales	*Cyanobacteriales*	**[DADA2] και [Deblur]**	
Actinomycetales	*Thermaerobacterales*	*Microtrichales*	*Erysipelotrichales*
Bacteroidales	*Campylobacterales*	**[Deblur] και [VSEARCH]**	
[DADA2] και [Deblur]		*Acetobacterales*	*Isosphaerales*
Symbiobacteriales	*Erysipelotrichales*	**[Deblur]**	
[DADA2] και [VSEARCH]		*Oxyphotobacteria Incertae Sedis*	*Thermomicrobiales*
Sumerlaeales	*Oscillospirales*	**[VSEARCH]**	
Alphaproteobacteria Incertae Sedis		*Diplorickettsiales*	
[Deblur] και [VSEARCH]			
Isosphaerales			
[DADA2]			
C0119			
[VSEARCH]			

Clostridia UCG-014		

Παράρτημα 42 Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο οικογένειας των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH πριν και μετά το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

Οικογένεια			
Πριν το φιλτράρισμα		**Μετά το φιλτράρισμα**	
[DADA2] και [Deblur] και [VSEARCH]		**[DADA2] και [Deblur] και [VSEARCH]**	
Nocardioidaceae	*Bacteroidaceae*	*Nocardioidaceae*	*Halomonadaceae*
Sphingomonadaceae	*Vicinamibacteraceae*	*Sphingomonadaceae*	*Kineosporiaceae*
Xanthobacteraceae	*Halomonadaceae*	*Xanthobacteraceae*	*Gemellaceae*
Shewanellaceae	*Kineosporiaceae*	*Lactobacillaceae*	*Family III*
Lactobacillaceae	*Mycobacteriaceae*	*Deinococcaceae*	*Beijerinckiaceae*
Enterobacteriaceae	*Gemellaceae*	*Crocinitomicaceae*	*Lachnospiraceae*
Pseudomonadaceae	*Family III*	*Oxalobacteraceae*	*env.OPS 17*
Promicromonosporaceae	*Beijerinckiaceae*	*Staphylococcaceae*	*Intrasporangiaceae*
Deinococcaceae	*Lachnospiraceae*	*Acholeplasmataceae*	*Nocardiaceae*
Crocinitomicaceae	*env.OPS 17*	*Comamonadaceae*	*Promicromonosporaceae*
Thermincolaceae	*Intrasporangiaceae*	*Enterobacteriaceae*	*Veillonellaceae*
Oxalobacteraceae	*Nocardiaceae*	*Marinomonadaceae*	*Aeromonadaceae*
Staphylococcaceae	*Rhizobiales Incertae Sedis*	*Clostridiaceae*	*Hafniaceae*
Acholeplasmataceae	*Veillonellaceae*	*Vibrionaceae*	*Dermatophilaceae*
Enterococcaceae	*Aeromonadaceae*	*Pasteurellaceae*	*Geodermatophilaceae*
Comamonadaceae	*Hafniaceae*	*Corynebacteriaceae*	*Acidaminococcaceae*
Marinomonadaceae	*Dermatophilaceae*	*Bacillaceae*	*Pseudoalteromonadaceae*
Morganellaceae	*Sphingobacteriaceae*	*Morganellaceae*	*Vagococcaceae*
Clostridiaceae	*Geodermatophilaceae*	*Micromonosporaceae*	*Obscuribacteraceae*
Vibrionaceae	*Acidaminococcaceae*	*Enterococcaceae*	*Aerococcaceae*
Pasteurellaceae	*Carnobacteriaceae*	*Caldicellulosiraptoraceae*	*Phycisphaeraceae*
Corynebacteriaceae	*Pseudoalteromonadaceae*	*Hymenobacteraceae*	*Thermincolaceae*
Bacillaceae	*Vagococcaceae*	*Idiomarinaceae*	*Propionibacteriaceae*
Micromonosporaceae	*JG30-KF-CM45*	*Micrococcaceae*	*Mycobacteriaceae*
Caldicellulosiraptoraceae	*Obscuribacteraceae*	*Cellulomonadaceae*	*Listeriaceae*
Caulobacteraceae	*Aerococcaceae*	*Pseudonocardiaceae*	*Dermabacteraceae*
Streptococcaceae	*Blastocatellaceae*	*Weeksellaceae*	*Pedosphaeraceae*
Moraxellaceae	*Burkholderiaceae*	*Rhodobacteraceae*	*Brevibacillaceae*
Hymenobacteraceae	*Brevibacillaceae*	*Xanthomonadaceae*	*Cytophagaceae*
Idiomarinaceae	*Phycisphaeraceae*	*Pseudomonadaceae*	*Pirellulaceae*
Micrococcaceae	*Propionibacteriaceae*	*Neisseriaceae*	*Caulobacteraceae*
Cellulomonadaceae	*Listeriaceae*	*Moraxellaceae*	*Dietziaceae*
Pseudonocardiaceae	*Dermabacteraceae*	*Geminicoccaceae*	*Chitinophagaceae*
Weeksellaceae	*Pedosphaeraceae*	*Fusobacteriaceae*	*Cellvibrionaceae*
Rhodobacteraceae	*Devosiaceae*	*Exiguobacteraceae*	*Burkholderiaceae*
Microbacteriaceae	*Cytophagaceae*	*Thermaceae*	*Rubrobacteriaceae*

Xanthomonadaceae	*Sporolactobacillaceae*	*Coriobacteriaceae*	*Selenomonadaceae*
Neisseriaceae	*Pirellulaceae*	*Shewanellaceae*	*Sporolactobacillaceae*
Geminicoccaceae	*Dietziaceae*	*Rhodocyclaceae*	*Solirubrobacteraceae*
Bifidobacteriaceae	*Chitinophagaceae*	*Legionellaceae*	*Saccharimonadaceae*
Fusobacteriaceae	*Cellvibrionaceae*	*Streptococcaceae*	*Streptomycetaceae*
Exiguobacteraceae	*Brevibacteriaceae*	*Babeliaceae*	*Microbacteriaceae*
Thermaceae	*Rubrobacteriaceae*	*Actinomycetaceae*	*Devosiaceae*
Coriobacteriaceae	*Selenomonadaceae*	*Porphyromonadaceae*	*Alteromonadaceae*
Rhodocyclaceae	*Solirubrobacteraceae*	*Pleomorphomonadaceae*	*Hydrogenophilaceae*
Legionellaceae	*Saccharimonadaceae*	*Paenibacillaceae*	*Bdellovibrionaceae*
Babeliaceae	*Streptomycetaceae*	*Leptotrichiaceae*	*Trueperaceae*
Actinomycetaceae	*Sporomusaceae*	*Nodosilineaceae*	*Desulfitobacteriaceae*
Porphyromonadaceae	*Ilumatobacteraceae*	*Prevotellaceae*	*Brevibacteriaceae*
Pleomorphomonadaceae	*Thermoactinomycetaceae*	*Flavobacteriaceae*	*Bryobacteraceae*
Paenibacillaceae	*Alteromonadaceae*	*Rhizobiaceae*	*Spirosomaceae*
Leptotrichiaceae	*Hydrogenophilaceae*	*Planococcaceae*	*Bifidobacteriaceae*
Nodosilineaceae	*Bdellovibrionaceae*	*Azospirillaceae*	*Rhodanobacteraceae*
Prevotellaceae	*Trueperaceae*	*Peptostreptococcaceae*	*Thermaerobacteraceae*
Flavobacteriaceae	*Desulfitobacteriaceae*	*Alcaligenaceae*	*Chroococcidiopsaceae*
Rhizobiaceae	*Acetobacteraceae*	*Yersiniaceae*	*Campylobacteraceae*
Planococcaceae	*Bryobacteraceae*	*Ferrovibrionaceae*	*Blastocatellaceae*
Diplorickettsiaceae	*Spirosomaceae*	*Bacteroidaceae*	*Carnobacteriaceae*
Azospirillaceae	*Rhodanobacteraceae*	*Vicinamibacteraceae*	
Peptostreptococcaceae	*Thermaerobacteraceae*	**[DADA2] and [Deblur]**	
Alcaligenaceae	*Chroococcidiopsaceae*	*Tannerellaceae*	*Erysipelatoclostridiaceae*
Yersiniaceae	*Campylobacteraceae*	**[Deblur] and [VSEARCH]**	
Ferrovibrionaceae		*Arcobacteraceae*	*Isosphaeraceae*
[DADA2] και [Deblur]		*Acetobacteraceae*	*Rhizobiales Incertae Sedis*
Tannerellaceae	*Erysipelatoclostridiaceae*	**[Deblur]**	
[DADA2] και [VSEARCH]		*Ilumatobacteraceae*	*JG30-KF-CM45*
Sumerlaeaceae	*Ethanoligenenaceae*	*Sporomusaceae*	
Leuconostocaceae		**[VSEARCH]**	
[Deblur] και [VSEARCH]		*Diplorickettsiaceae*	*Sphingobacteriaceae*
Arcobacteraceae	*Isosphaeraceae*		
[VSEARCH]			
Microtrichaceae	*Microscillaceae*		
Hyphomicrobiaceae	*Ruminococcaceae*		
Nitrosomonadaceae			

Παράρτημα 43 Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο γένους των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH πριν το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

Γένος			
Πριν το φιλτράρισμα			
[DADA2] και [Deblur] και [VSEARCH]			
Acidibacter	*Janthinobacterium*	*Aeribacillus*	*Streptomyces*
Nocardioides	*Legionella*	*Blastococcus*	*Pelosinus*
Sphingomonas	*Streptococcus*	*Carnobacterium*	*Conexibacter*
Bradyrhizobium	*Candidatus_Babela*	*Enterobacter*	*Erythrobacter*
Shewanella	*Adhaeribacter*	*Neisseria*	*Klebsiella*
Lactobacillus	*Actinomyces*	*Pseudoalteromonas*	*Ferruginibacter*
Escherichia-Shigella	*Novosphingobium*	*Peptostreptococcus*	*Vitreoscilla*
Pseudomonas	*Porphyromonas*	*Amaricoccus*	*Filifactor*
Promicromonospora	*Pleomorphomonas*	*Azospirillum*	*Thermoactinomyces*
Deinococcus	*Paenibacillus*	*Vagococcus*	*Rheinheimera*
Brumimicrobium	*Leptotrichia*	*Methylobacterium-Methylorubrum*	*Ornithinibacillus*
Thermincola	*Nodosilinea PCC-7104*	*Rhodococcus*	*Microvirga*
Massilia	*Alloprevotella*	*Candidatus Obscuribacter*	*Achromobacter*
Staphylococcus	*Flavobacterium*	*Psychrobacter*	*Dechloromonas*
Acholeplasma	*Tepidimonas*	*Abiotrophia*	*Tepidiphilus*
Enterococcus	*Prevotella*	*Cupriavidus*	*Bdellovibrio*
Hydrogenophaga	*Shimwellia*	*Brevibacillus*	*Pseudoxanthomonas*
Marinomonas	*Rickettsiella*	*Moraxella*	*Truepera*
Morganella	*Peptoniphilus*	*SM1A02*	*Desulfosporosinus*
Clostridium sensu stricto 1	*Caldibacillus*	*Bergeyella*	*Frederiksenia*
Vibrio	*Skermanella*	*Diaphorobacter*	*Craurococcus-Caldovatus*
Aggregatibacter	*Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium*	*Thermobacillus*	*Solirubrobacter*
Corynebacterium	*Peptoclostridium*	*Moellerella*	*Fenollaria*
Bacillus	*Paenalcaligenes*	*Brochothrix*	*Kingella*
Micromonospora	*Yersinia*	*Ensifer*	*Saccharopolyspora*
Caldicellulosiruptor	*Ferrovibrio*	*Salinispora*	*Pseudochrobactrum*
Brevundimonas	*Rothia*	*Enhydrobacter*	*Eubacterium yurii group*
Lactococcus	*Lawsonella*	*Brachybacterium*	*Stomatobaculum*
Acinetobacter	*Myroides*	*Proteus*	*Joostella*
Hymenobacter	*Serratia*	*CENA359*	*Bryobacter*
Aliidiomarina	*Thermomonas*	*Caulobacter*	*Lacihabitans*
Kocuria	*Glutamicibacter*	*Devosia*	*Luteimonas*
Actinotalea	*Bacteroides*	*Planococcus*	*CL500-29 marine group*
Actinophytocola	*Lysinibacillus*	*Clostridium sensu stricto 12*	*Cutibacterium*
Chryseobacterium	*Halomonas*	*Siphonobacter*	*Aerococcus*

Paracoccus	*Mycobacterium*	*Microbacterium*	*Clostridium sensu stricto 5*
Anoxybacillus	*Haemophilus*	*Vulcaniibacterium*	*Brachymonas*
Empedobacter	*Gemella*	*Sporolactobacillus*	*Bosea*
Stenotrophomonas	*hermoanaerobacterium*	*Pir4 lineage*	*Lysobacter*
Comamonas	*Anaerostipes*	*Dietzia*	*Chiayiivirga*
Actinomycetospora	*Actinoalloteichus*	*Gallicola*	*Schlegelella*
Salinicoccus	*Gordonia*	*Micrococcus*	*Undibacterium*
Uruburuella	*Phreatobacter*	*Cellvibrio*	*Thermaerobacter*
Candidatus_Alysiosphaera	*Isoptericola*	*Ralstonia*	*Cnuella*
Bifidobacterium	*Veillonella*	*Brevibacterium*	*Chroococcidiopsis SAG 2023*
Fusobacterium	*Pseudonocardia*	*Rubrobacter*	*Mesonia*
Geobacillus	*Aeromonas*	*Eremococcus*	*Campylobacter*
Exiguobacterium	*Hafnia-Obesumbacterium*	*Lautropia*	*Elizabethkingia*
Thermus	*Idiomarina*	*Flaviflexus*	*Granulicatella*
Collinsella	*Altererythrobacter*	*Marmoricola*	*Croceicoccus*
Cloacibacterium	*Haematobacter*	*Sphingobium*	*Ruminococcus gauvreauii group*
Anaerococcus	*Austwickia*	*TM7a*	*Shinella*
Methyloversatilis	*Pedobacter*	*Jeotgalicoccus*	*Finegoldia*
Rubellimicrobium	*Coenonia*	*Quadrisphaera*	
[DADA2] και [Deblur]			
Caldinitratiruptor	*Mangrovibacter*	*Testudinibacter*	*Erysipelatoclostridium*
[DADA2] και [VSEARCH]			
Ancylobacter	*Gelidibacter*	*Blastomonas*	*JGI 0001001-H03*
Sumerlaea	*Acuticoccus*	*Capnocytophaga*	*Ethanoligenens*
Leuconostoc			
[Deblur] and [VSEARCH]			
Pelomonas	*Tundrisphaera*	*Aliterella*	*Delftia*
Pseudarcobacter	*Desemzia*	*Nitratireductor*	*Phenylobacterium*
Acidovorax			
[DADA2]			
Limnobacter			
[VSEARCH]			
Arthrobacter	*Aeromicrobium*	*Puia*	*IS-44*
Sediminibacterium	*Selenomonas*	*IMCC26207*	*Parvimonas*
Zoogloea	*Hydrogenophilus*	*Hyphomicrobium*	*Ruminococcus*
Roseateles	*Ochrobactrum*	*Meiothermus*	

Παράρτημα 44 Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο γένους των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH μετά το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

Γένος			
Μετά το φιλτράρισμα			
[DADA2] and [Deblur] and [VSEARCH]			
Acidibacter	*Streptococcus*	*Blastococcus*	*Quadrisphaera*

Nocardioides	*Candidatus Babela*	*Actinomycetospora*	*Streptomyces*
Sphingomonas	*Adhaeribacter*	*Pseudonocardia*	*Microbacterium*
Bradyrhizobium	*Actinomyces*	*Enterobacter*	*Erythrobacter*
Lactobacillus	*Novosphingobium*	*Neisseria*	*Klebsiella*
Deinococcus	*Porphyromonas*	*Pseudoalteromonas*	*Devosia*
Brumimicrobium	*Pleomorphomonas*	*Peptostreptococcus*	*Ferruginibacter*
Massilia	*Paenibacillus*	*Amaricoccus*	*Vitreoscilla*
Staphylococcus	*Leptotrichia*	*Comamonas*	*Rheinheimera*
Acholeplasma	*Nodosilinea PCC-7104*	*Vagococcus*	*Ornithinibacillus*
Hydrogenophaga	*Alloprevotella*	*Methylobacterium-Methylorubrum*	*Microvirga*
Marinomonas	*Rubellimicrobium*	*Rhodococcus*	*Achromobacter*
Clostridium sensu stricto 1	*Flavobacterium*	*Candidatus Obscuribacter*	*Dechloromonas*
Vibrio	*Tepidimonas*	*Psychrobacter*	*Tepidiphilus*
Aggregatibacter	*Prevotella*	*Abiotrophia*	*Bdellovibrio*
Corynebacterium	*Shimwellia*	*SM1A02*	*Truepera*
Bacillus	*Caldibacillus*	*Bergeyella*	*Desulfosporosinus*
Morganella	*Skermanella*	*Diaphorobacter*	*Frederiksenia*
Micromonospora	*Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium*	*Altererythrobacter*	*Brevibacterium*
Enterococcus	*Peptoclostridium*	*Thermincola*	*Solirubrobacter*
Caldicellulosiruptor	*Paenalcaligenes*	*Thermobacillus*	*Pseudoxanthomonas*
Hymenobacter	*Yersinia*	*Mycobacterium*	*Kingella*
Aliidiomarina	*Ferrovibrio*	*Brochothrix*	*Eubacterium yurii group*
Kocuria	*Rothia*	*Salinispora*	*Stomatobaculum*
Actinotalea	*Lactococcus*	*Enhydrobacter*	*Joostella*
Actinophytocola	*Lawsonella*	*Brachybacterium*	*Bryobacter*
Chryseobacterium	*Myroides*	*Proteus*	*Lacihabitans*
Paracoccus	*Serratia*	*Planococcus*	*Cutibacterium*
Anoxybacillus	*Thermomonas*	*Brevibacillus*	*Caulobacter*
Empedobacter	*Glutamicibacter*	*Clostridium sensu stricto 12*	*Clostridium sensu stricto 5*
Stenotrophomonas	*Bacteroides*	*Siphonobacter*	*Peptoniphilus*
Pseudomonas	*Lysinibacillus*	*Vulcaniibacterium*	*Bosea*
Salinicoccus	*Halomonas*	*Pir4 lineage*	*Bifidobacterium*
Uruburuella	*Haemophilus*	*Brevundimonas*	*Chiayiivirga*
Acinetobacter	*Gemella*	*Dietzia*	*Promicromonospora*
Candidatus Alysiosphaera	*Thermoanaerobacterium*	*Gallicola*	*Schlegelella*
Fusobacterium	*Anaerostipes*	*Micrococcus*	*Undibacterium*
Geobacillus	*Actinoalloteichus*	*Cellvibrio*	*Thermaerobacter*
Exiguobacterium	*Gordonia*	*Ralstonia*	*Cnuella*
Thermus	*Isoptericola*	*Rubrobacter*	*Chroococcidiopsis SAG 2023*
Escherichia-Shigella	*Veillonella*	*Eremococcus*	*Luteimonas*
Collinsella	*Aeromonas*	*Lautropia*	*Campylobacter*
Cloacibacterium	*Hafnia-Obesumbacterium*	*Flaviflexus*	*Croceicoccus*
Shewanella	*Idiomarina*	*Sporolactobacillus*	*Ruminococcus gauvreauii group*

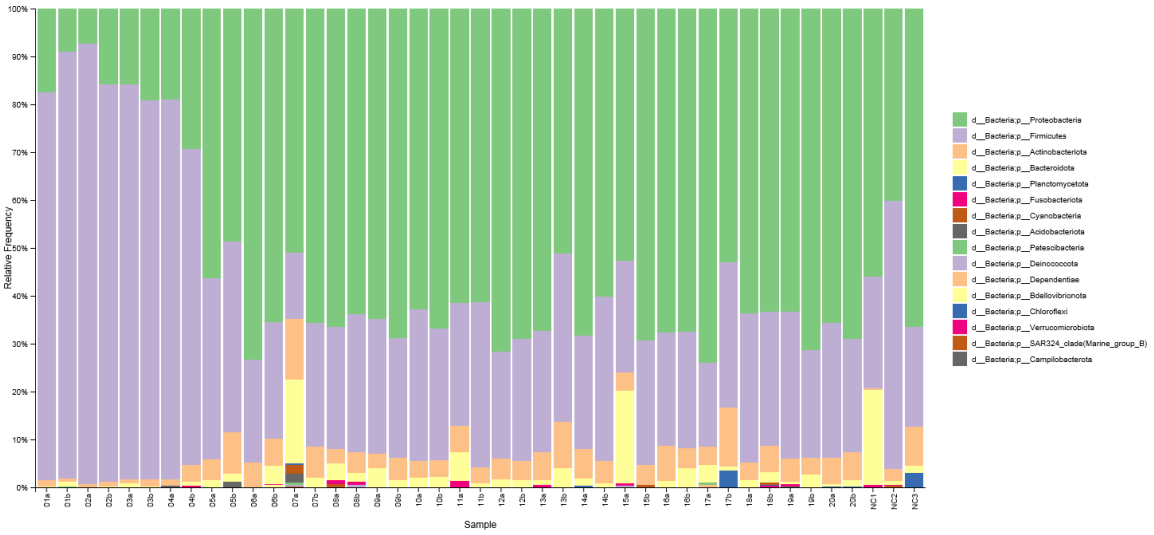
Anaerococcus	*Haematobacter*	*Marmoricola*	*Granulicatella*
Methyloversatilis	*Austwickia*	*Sphingobium*	*Shinella*
Janthinobacterium	*Coenonia*	*TM7a*	*Jeotgalicoccus*
Legionella	*Aeribacillus*		
[DADA2] και [Deblur]			
Ensifer	*Mangrovibacter*	*Testudinibacter*	*Erysipelatoclostridium*
[DADA2] και [VSEARCH]			
Ancylobacter	*JGI 0001001-H03*		
[Deblur] και [VSEARCH]			
Pelomonas	*Craurococcus-Caldovatus*	*Aliterella*	*Saccharopolyspora*
Pseudarcobacter	*Tundrisphaera*	*Phreatobacter*	*Carnobacterium*
Acidovorax	*Desemzia*		
[Deblur]			
Conexibacter	*CENA359*	*CL500-29 marine group*	*Pelosinus*
Moellerella	*Lysobacter*	*Nitratireductor*	*Phenylobacterium*
[VSEARCH]			
Rickettsiella	*Pedobacter*		

Παράρτημα 45 Πίνακας με τις κοινές και μοναδικές ταξινομηκές κατηγορίες σε επίπεδο είδους των τριών συνόλων δεδομένων από τις επεξεργαστικές ροές των DADA2, Deblur και VSEARCH πριν και μετά το φιλτράρισμα χαμηλής σχετικής αφθονίας ταξινομικών κατηγοριών.

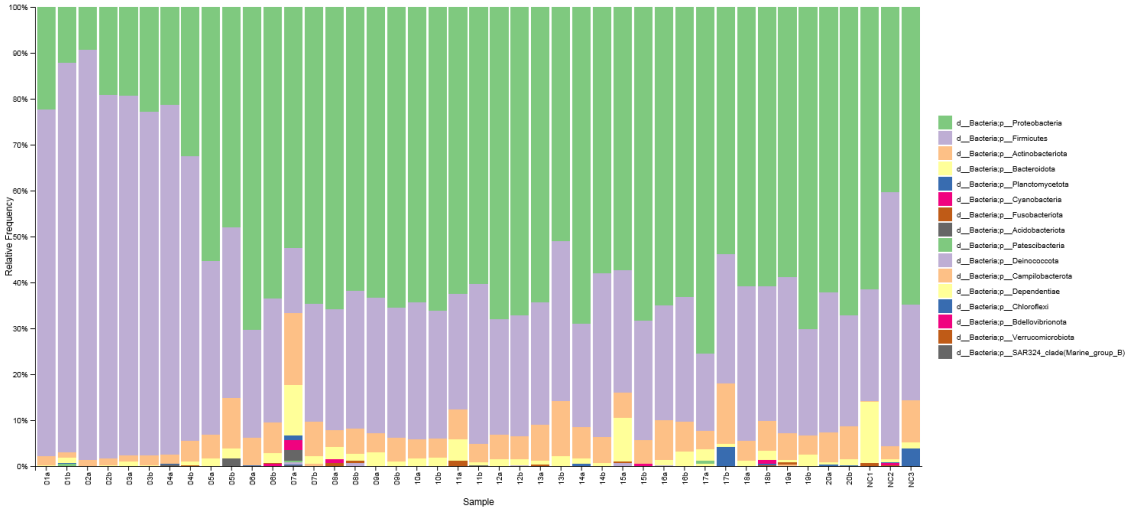
Είδος			
Πριν το φιλτράρισμα		**Μετά το φιλτράρισμα**	
[DADA2] και [Deblur] και [VSEARCH]		**[DADA2] και [Deblur] και [VSEARCH]**	
Deinococcus aerolatus	*Corynebacterium propinquum*	*Deinococcus aerolatus*	*Acinetobacter baumannii*
Staphylococcus lentus	*Nocardioides furvisabuli*	*Staphylococcus lentus*	*Clostridium magnum*
Caldicellulosiruptor acetigenus	*Peptoniphilus duerdenii*	*Caldicellulosiruptor acetigenus*	*Psychrobacter alimentarius*
Corynebacterium appendicis	*Clostridium magnum*	*Corynebacterium appendicis*	*Acinetobacter guillouiae*
Lactococcus piscium	*Psychrobacter alimentarius*	*Hymenobacter tibetensis*	*Bacillus circulans*
Acinetobacter baumannii	*Sporolactobacillus laevolacticus*	*Aliidiomarina shirensis*	*Prevotella copri*
Hymenobacter tibetensis	*Acinetobacter guillouiae*	*Acinetobacter ursingii*	*Schaalia odontolytica*
Aliidiomarina shirensis	*Bacillus circulans*	*Staphylococcus equorum*	*Ralstonia mannitolilytica*
Acinetobacter ursingii	*Prevotella copri*	*Fusobacterium periodonticum*	*Hymenobacter rutilus*
Staphylococcus equorum	*Ralstonia mannitolilytica*	*Thermus scotoductus*	*Enterococcus cecorum*
Bifidobacterium pullorum	*Hymenobacter rutilus*	*Collinsella tanakaei*	*Rubrobacter xylanophilus*
Fusobacterium periodonticum	*Enterococcus cecorum*	*Legionella pneumophila*	*Nocardioides lentus*
Thermus scotoductus	*Rubrobacter xylanophilus*	*Candidatus Babela*	*Sporolactobacillus laevolacticus*
Collinsella tanakaei	*Nocardioides lentus*	*Streptococcus parauberis*	*Haemophilus pittmaniae*
Legionella pneumophila	*Corynebacterium durum*	*Boudabousia marimammalium*	*Brevundimonas olei*
Candidatus Babela	*Parageobacillus thermoglucosidasius*	*Porphyromonas pasteri*	*Methylobacterium variabile*
Streptococcus parauberis	*Haemophilus pittmaniae*	*Paenibacillus validus*	*Acinetobacter radioresistens*
Boudabousia marimammalium	*Brevundimonas olei*	*Vibrio metschnikovii*	*Morganella psychrotolerans*
Porphyromonas pasteri	*Methylobacterium variabile*	*Prevotella veroralis*	*Ornithinibacillus contaminans*
Schaalia odontolytica	*Acinetobacter radioresistens*	*Flavobacterium dankookense*	*Brevibacterium ravenspurgense*

Paenibacillus validus	*Filifactor alocis*	*Shewanella baltica*	*Solirubrobacter taibaiensis*
Vibrio metschnikovii	*Morganella psychrotolerans*	*Ferrovibrio denitrificans*	*Pseudomonas guguanensis*
Lactobacillus fermentum	*Ornithinibacillus contaminans*	*Streptococcus dysgalactiae*	*Acinetobacter schindleri*
Prevotella veroralis	*Hymenobacter qilianensis*	*Myroides odoratus*	*Kocuria palustris*
Flavobacterium dankookense	*Brevibacterium ravenspurgense*	*Leptotrichia buccalis*	*Joostella marina*
Shewanella baltica	*Solirubrobacter taibaiensis*	*Porphyromonas asaccharolytica*	*Staphylococcus pettenkoferi*
Ferrovibrio denitrificans	*Pseudomonas guguanensis*	*Lactococcus piscium*	*bacterium 28W232*
Streptococcus dysgalactiae	*Acinetobacter schindleri*	*Vibrio diazotrophicus*	*Methylobacterium komagatae*
Myroides odoratus	*Kocuria palustris*	*Porphyromonas gingivalis*	*Thermaerobacter subterraneus*
Leptotrichia buccalis	*Joostella marina*	*Anaerostipes caccae*	*Bifissio spartinae*
Porphyromonas asaccharolytica	*Staphylococcus pettenkoferi*	*Corynebacterium aurimucosum*	*Campylobacter concisus*
Vibrio diazotrophicus	*bacterium 28W232*	*Austwickia chelonae*	*Stenotrophomonas rhizophila*
Porphyromonas gingivalis	*Cutibacterium granulosum*	*Streptococcus salivarius*	*Myroides phaeus*
Deinococcus proteolyticus	*Methylobacterium komagatae*	*Coenonia anatina*	*Corynebacterium kroppenstedtii*
Anaerostipes caccae	*Thermaerobacter subterraneus*	*Corynebacterium propinquum*	*TM7 phylum*
Prevotella paludivivens	*Bifissio spartinae*	**[DADA2] και [Deblur]**	
Corynebacterium aurimucosum	*Ilumomonas thalassia*	*Testudinibacter aquarius*	
Austwickia chelonae	*Campylobacter concisus*	**[DADA2] και [VSEARCH]**	
Streptococcus salivarius	*Stenotrophomonas rhizophila*	*Acinetobacter bohemicus*	*Corynebacterium amycolatum*
Coenonia anatina	*Myroides phaeus*	*Lactococcus raffinolactis*	*Bacillus nealsonii*
Moraxella atlantae	*Corynebacterium kroppenstedtii*	*Prevotella paludivivens*	*Neisseria perflava*
[DADA2] και [Deblur]		**[Deblur] και [VSEARCH]**	
Caldinitratiruptor microaerophilus	*Testudinibacter aquarius*	*Deinococcus antarcticus*	*Paenibacillus glucanolyticus*
[DADA2] και [VSEARCH]		*Paenibacillus urinalis*	*Carnobacterium maltaromaticum*
Comamonas granuli	*Bacillus nealsonii*	**[DADA2]**	
Acinetobacter bohemicus	*Porphyromonas catoniae*	*Streptococcus parasanguinis*	
Lactococcus raffinolactis	*Streptococcus parasanguinis*	**[Deblur]**	
Lactobacillus melliventris	*Lactobacillus iners*	*Comamonas denitrificans*	*Pseudomonas geniculata*
Leuconostoc carnosum	*Aggregatibacter aphrophilus*	*Peptoniphilus duerdenii*	*Thermoanaerobacterium thermosaccharolyticum*
Prevotella nanceiensis	*Neisseria perflava*	**[VSEARCH]**	
Corynebacterium amycolatum	*Capnocytophaga sputigena*	*Desemzia incerta*	*Lactobacillus iners*
bacterium BW1PhC42			
[Deblur] και [VSEARCH]			
Deinococcus antarcticus	*Carnobacterium maltaromaticum*		
Paenibacillus urinalis	*Pseudomonas geniculata*		
Paenibacillus glucanolyticus			
[Deblur]			
Comamonas denitrificans	*Thermoanaerobacterium thermosaccharolyticum*		
Acinetobacter beijerinckii			
[VSEARCH]			
Clostridium indolis	*Enterococcus faecium*		
Shewanella profunda	*Rubellimicrobium aerolatum*		
Anaerococcus nagyae	*Vibrio ordalii*		

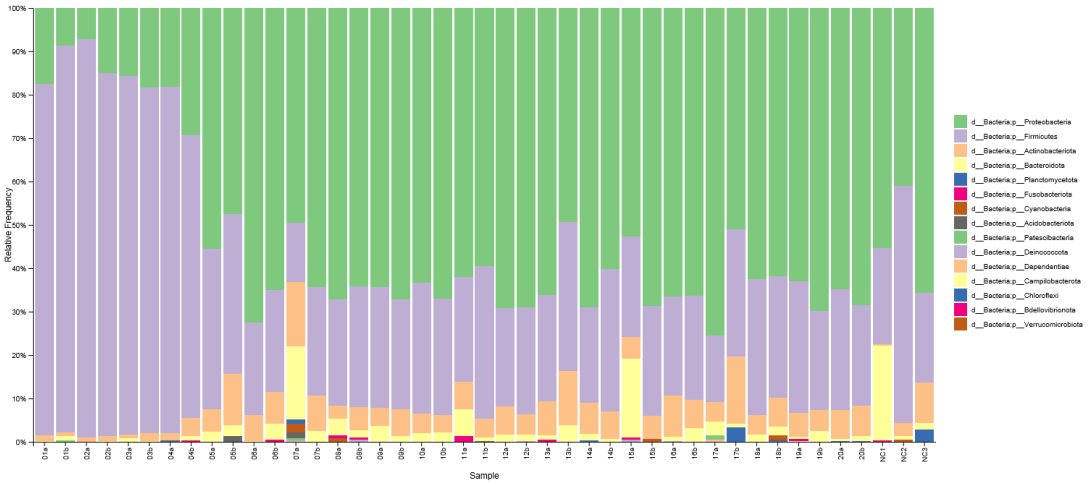
Prevotella bivia	*Ruminococcus bicirculans*		
bacterium AF13	*Kingella oralis*		
TM7 phylum	*Aliidiomarina iranensis*		
Lepisosteus oculatus	*Tepidimonas fonticaldi*		
Clostridium hiranonis	*Corynebacterium matruchotii*		
Streptococcus vestibularis	*Peptoniphilus lacrimalis*		
Anaerococcus octavius	*Corynebacterium testudinoris*		
Clostridium putrefaciens	*bacterium CYCU-0258*		
bacterium RFB	*Acinetobacter septicus*		



Παράρτημα 46 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο φυλης του συνόλου δεδομένων DADA2.



Παράρτημα 47 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο φυλης του συνόλου δεδομένων Deblur.



Παράρτημα 48 Διάγραμμα ραβδών για την απεικόνιση της σχετικής αφθονίας των βακτηριακών ταξινομικών κατηγοριών σε επιπεδο φυλης του συνόλου δεδομένων VSEARCH.

Παράρτημα 49. Αναλυτικές μετρήσεις εντροπίας Shannon άλφα ποικιλομορφίας για την κατασκευή των θηκογραμμάτων που προέκυψαν από τα σύνολα δεδομένων DADA2, Deblur και VSEARCH, για τις τρεις βασικές ομάδες δειγμάτων, τα δείγματα αίματος πριν την αντιψυχωσική θεραπεία (before_treatment), τα δείγματα αίματος μετά από έναν μήνα αντιψυχωσικής θεραπείας (after_treatment) και τα δείγματα ελέγχου (controls).

	Ομάδες δειγμάτων:	**before_treatment (n=20)**	**after_treatment (n=20)**	**controls (n=3)**
DADA2	Διάμεσος	5,01	4,91	4,61
	Ενδοτεταρτημοριακό εύρος (IQR)	0,82	0,36	0,31
	Ελάχιστη	4,32	4,47	4,41
	Μέγιστη	5,70	5,32	5,04
	Αριθμός ακραίων τιμών	2	-	-
Deblur	Διάμεσος	4,73	4,66	4,28
	Ενδοτεταρτημοριακό εύρος (IQR)	0,39	0,36	0,37
	Ελάχιστη	4,09	3,99	4,21
	Μέγιστη	5,43	5,19	5,00
	Αριθμός ακραίων τιμών	1	1	-
VSEARCH	Διάμεσος	5,60	5,51	6,02
	Ενδοτεταρτημοριακό εύρος (IQR)	0,46	0,39	0,38
	Ελάχιστη	4,82	5,15	5,57
	Μέγιστη	6,30	6,00	6,20
	Αριθμός ακραίων τιμών	1	-	-

Βιβλιογραφία

Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors a unique microbiome. *Science Translational Medicine*, *6*(237), 237ra65. https://doi.org/10.1126/scitranslmed.3008599

Aggarwal, N., Kitano, S., Puah, G. R. Y., Kittelmann, S., Hwang, I. Y., & Chang, M. W. (2023). Microbiome and Human Health: Current Understanding, Engineering, and Enabling Technologies. *Chemical Reviews*, *123*(1), 31–72. https://doi.org/10.1021/acs.chemrev.2c00431

Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J. J., Tripathi, A., Brenner, D. A., Loomba, R., Smarr, L., Sandborn, W. J., Schnabl, B., Dorrestein, P., Zarrinpar, A., & Knight, R. (2019). Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clinical Gastroenterology and Hepatology : The Official Clinical Practice Journal of the American Gastroenterological Association*, *17*(2), 218–230. https://doi.org/10.1016/j.cgh.2018.09.017

Almeida, O. G. G., & De Martinis, E. C. P. (2019). Bioinformatics tools to assess metagenomic data for applied microbiology. *Applied Microbiology and Biotechnology*, *103*(1), 69–82. https://doi.org/10.1007/s00253-018-9464-9

Alves, L. de F., Westmann, C. A., Lovate, G. L., de Siqueira, G. M. V., Borelli, T. C., & Guazzaroni, M.-E. (2018). Metagenomic Approaches for Understanding New Concepts in Microbial Science. *International Journal of Genomics*, *2018*, 2312987. https://doi.org/10.1155/2018/2312987

Amar, J., Lange, C., Payros, G., Garret, C., Chabo, C., Lantieri, O., Courtney, M., Marre, M., Charles, M. A., Balkau, B., & Burcelin, R. (2013). Blood microbiota dysbiosis is associated with the onset of cardiovascular events in a large general population: The D.E.S.I.R. study. *PloS One*, *8*(1), e54461. https://doi.org/10.1371/journal.pone.0054461

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed.* (pp. xliv, 947). American Psychiatric Publishing, Inc. https://doi.org/10.1176/appi.books.9780890425596

Ames, N. J., Ranucci, A., Moriyama, B., & Wallen, G. R. (2017). The Human Microbiome and Understanding the 16S rRNA Gene in Translational Nursing Science. *Nursing Research*, *66*(2), 184–197. https://doi.org/10.1097/NNR.0000000000000212

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*, *2*(2). https://doi.org/10.1128/mSystems.00191-16

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data* [Java]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, *22*(1), 177. https://doi.org/10.1186/s12859-021-04115-6

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, *71*(12), 7724–7736. https://doi.org/10.1128/AEM.71.12.7724-7736.2005

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., & Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, *3*, e1029. https://doi.org/10.7717/peerj.1029

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PloS One*, *7*(10), e46679. https://doi.org/10.1371/journal.pone.0046679

Azad, M. B., Konya, T., Maughan, H., Guttman, D. S., Field, C. J., Chari, R. S., Sears, M. R., Becker, A. B., Scott, J. A., & Kozyrskyj, A. L. (2013). Gut microbiota of healthy Canadian infants: Profiles by mode of delivery and infant diet at 4 months. *CMAJ : Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, *185*(5), 385–394. https://doi.org/10.1503/cmaj.121189

Bagdasarian, N., Rao, K., & Malani, P. N. (2015). Diagnosis and treatment of Clostridium difficile in adults: A systematic review. *JAMA*, *313*(4), 398–408. https://doi.org/10.1001/jama.2014.17103

Bahrani-Mougeot, F. K., Paster, B. J., Coleman, S., Ashar, J., Barbuto, S., & Lockhart, P. B. (2008). Diverse and novel oral bacterial species in blood following dental procedures. *Journal of Clinical Microbiology*, *46*(6), 2129–2132. https://doi.org/10.1128/JCM.02004-07

Baquero, F., & Nombela, C. (2012). The microbiome as a human organ. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *18 Suppl 4*, 2–4. https://doi.org/10.1111/j.1469-0691.2012.03916.x

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., … Schloter, M. (2020). Microbiome definition re-visited: Old concepts and new challenges. *Microbiome*, *8*(1), 103. https://doi.org/10.1186/s40168-020-00875-0

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, *22*(1), 178–193. https://doi.org/10.1093/bib/bbz155

Bhat, A. H., Prabhu, P., & Balakrishnan, K. (2019). A critical analysis of state-of-the-art metagenomics OTU clustering algorithms. *Journal of Biosciences*, *44*. https://api.semanticscholar.org/CorpusID:207913831

Bhavesh Tiwarekar, Kiran Kirdat, Shivaji Sathe, Xavier Foissac, & Amit Yadav. (2023). Chimera alert! The threat of chimeric sequences causing inaccurate taxonomic classification of phytoplasma strains. *bioRxiv*, 2023.04.10.535501. https://doi.org/10.1101/2023.04.10.535501

Biasucci, G., Benenati, B., Morelli, L., Bessi, E., & Boehm, G. (2008). Cesarean delivery may affect the early biodiversity of intestinal bacteria. *The Journal of Nutrition*, *138*(9), 1796S-1800S. https://doi.org/10.1093/jn/138.9.1796S

Bleidorn, C. (2016). Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, *14*(1), 1–8. https://doi.org/10.1080/14772000.2015.1099575

Blevins, S. M., & Bronze, M. S. (2010). Robert Koch and the "golden age" of bacteriology. *International Journal of Infectious Diseases : IJID : Official Publication of the International Society for Infectious Diseases*, *14*(9), e744-751. https://doi.org/10.1016/j.ijid.2009.12.003

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A., & Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, *10*(1), 57–59. https://doi.org/10.1038/nmeth.2276

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Branton, W. G., Ellestad, K. K., Maingat, F., Wheatley, B. M., Rud, E., Warren, R. L., Holt, R. A., Surette, M. G., & Power, C. (2013). Brain microbial populations in HIV/AIDS: α-proteobacteria predominate independent of host immune status. *PloS One*, *8*(1), e54673. https://doi.org/10.1371/journal.pone.0054673

Bukin, Y. S., Galachyants, Y. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *6*, 190007. https://doi.org/10.1038/sdata.2019.7

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Calle, M. L. (2019). Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, *17*(1), e6. https://doi.org/10.5808/GI.2019.17.1.e6

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cao, Q., Sun, X., Rajesh, K., Chalasani, N., Gelow, K., Katz, B., Shah, V. H., Sanyal, A. J., & Smirnova, E. (2020). Effects of Rare Microbiome Taxa Filtering on Statistical Analysis. *Frontiers in Microbiology*, *11*, 607325. https://doi.org/10.3389/fmicb.2020.607325

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Caruso, V., Song, X., Asquith, M., & Karstens, L. (2019). Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems*, *4*(1). https://doi.org/10.1128/mSystems.00163-18

Castillo, D. J., Rifkin, R. F., Cowan, D. A., & Potgieter, M. (2019). The Healthy Human Blood Microbiome: Fact or Fiction? *Frontiers in Cellular and Infection Microbiology*, *9*. https://www.frontiersin.org/articles/10.3389/fcimb.2019.00148

Cernava, T., Rybakova, D., Buscot, F., Clavel, T., McHardy, A. C., Meyer, F., Meyer, F., Overmann, J., Stecher, B., Sessitsch, A., Schloter, M., & Berg, G. (2022). Metadata harmonization-Standards are the key for a better usage of omics data for integrative microbiome analysis. *Environmental Microbiome*, *17*(1), 33. https://doi.org/10.1186/s40793-022-00425-1

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, *11*(4), 265–270. JSTOR.

Check Hayden, E. (2009). Genome sequencing: The third generation. *Nature*, *457*(7231), 768–769. https://doi.org/10.1038/news.2009.86

Chen, R., Wang, Z., Chen, J., & Qiao, G.-X. (2015). Avoidance and Potential Remedy Solutions of Chimeras in Reconstructing the Phylogeny of Aphids Using the 16S rRNA Gene of Buchnera: A Case in Lachninae (Hemiptera). *International Journal of Molecular Sciences*, *16*(9), 20152–20167. https://doi.org/10.3390/ijms160920152

Cheng, H. S., Tan, S. P., Wong, D. M. K., Koo, W. L. Y., Wong, S. H., & Tan, N. S. (2023). The Blood Microbiome and Health: Current Evidence, Controversies, and Challenges. *International Journal of Molecular Sciences*, *24*(6). https://doi.org/10.3390/ijms24065633

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nature Reviews. Genetics*, *20*(6), 341–355. https://doi.org/10.1038/s41576-019-0113-7

Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Flater, J., Tiedje, J. M., Hofmockel, K. S., Gelder, B., & Howe, A. (2017). Strategies to improve reference databases for soil microbiomes. *The ISME Journal*, *11*(4), 829–834. https://doi.org/10.1038/ismej.2016.168

Church, D. L., Cerutti, L., Gürtler, A., Griener, T., Zelazny, A., & Emler, S. (2020). Performance and Application of 16S rRNA Gene Cycle Sequencing for Routine Identification of Bacteria in the Clinical Microbiology Laboratory. *Clinical Microbiology Reviews*, *33*(4). https://doi.org/10.1128/CMR.00053-19

Clapp, M., Aurora, N., Herrera, L., Bhatia, M., Wilen, E., & Wakefield, S. (2017). Gut microbiota's effect on mental health: The gut-brain axis. *Clinics and Practice*, *7*(4), 987. https://doi.org/10.4081/cp.2017.987

Cloutier, M., Aigbogun, M. S., Guerin, A., Nitulescu, R., Ramanakumar, A. V., Kamat, S. A., DeLucia, M., Duffy, R., Legacy, S. N., Henderson, C., Francois, C., & Wu, E. (2016). The Economic Burden of Schizophrenia in the United States in 2013. *The Journal of Clinical Psychiatry*, *77*(6), 764–771. https://doi.org/10.4088/JCP.15m10278

Cogen, A. L., Nizet, V., & Gallo, R. L. (2008). Skin microbiota: A source of disease or defence? *The British Journal of Dermatology*, *158*(3), 442–455. https://doi.org/10.1111/j.1365-2133.2008.08437.x

Coughlan, L. M., Cotter, P. D., Hill, C., & Alvarez-Ordóñez, A. (2015). Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. *Frontiers in Microbiology*, *6*, 672. https://doi.org/10.3389/fmicb.2015.00672

Cryan, J. F., O'Riordan, K. J., Cowan, C. S. M., Sandhu, K. V., Bastiaanssen, T. F. S., Boehme, M., Codagnone, M. G., Cussotto, S., Fulling, C., Golubeva, A. V., Guzzetta, K. E., Jaggar, M., Long-Smith, C. M., Lyte, J. M., Martin, J. A., Molinero-Perez, A., Moloney, G., Morelli, E., Morillas, E., … Dinan, T. G. (2019). The Microbiota-Gut-Brain Axis. *Physiological Reviews*, *99*(4), 1877–2013. https://doi.org/10.1152/physrev.00018.2018

Dacey, D. P., & Chain, F. J. J. (2021). Concatenation of paired-end reads improves taxonomic classification of amplicons for profiling microbial communities. *BMC Bioinformatics*, *22*(1), 493. https://doi.org/10.1186/s12859-021-04410-2

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. https://doi.org/10.1186/s40168-018-0605-2

de la Cuesta-Zuluaga, J., & Escobar, J. S. (2016). Considerations For Optimizing Microbiome Analysis Using a Marker Gene. *Frontiers in Nutrition*, *3*. https://www.frontiersin.org/articles/10.3389/fnut.2016.00026

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS One*, *8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

Dinakaran, V., Rathinavel, A., Pushpanathan, M., Sivakumar, R., Gunasekaran, P., & Rajendhran, J. (2014). Elevated levels of circulating DNA in cardiovascular disease patients: Metagenomic profiling of microbiome in the circulation. *PloS One*, *9*(8), e105221. https://doi.org/10.1371/journal.pone.0105221

Domingue, G. J., & Schlegel, J. U. (1977). Novel bacterial structures in human blood: Cultural isolation. *Infection and Immunity*, *15*(2), 621–627. https://doi.org/10.1128/iai.15.2.621-627.1977

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, *107*(26), 11971–11975. https://doi.org/10.1073/pnas.1002601107

Dominguez-Bello, M. G., Godoy-Vitorino, F., Knight, R., & Blaser, M. J. (2019). Role of the microbiome in human development. *Gut*, *68*(6), 1108. https://doi.org/10.1136/gutjnl-2018-317503

Doré, J., Simrén, M., Buttle, L., & Guarner, F. (2013). Hot topics in gut microbiota. *United European Gastroenterology Journal*, *1*(5), 311–318. https://doi.org/10.1177/2050640613502477

Dos Santos, H. R. M., Argolo, C. S., Argôlo-Filho, R. C., & Loguercio, L. L. (2019). A 16S rDNA PCR-based theoretical to actual delta approach on culturable mock communities revealed severe losses of diversity information. *BMC Microbiology*, *19*(1), 74. https://doi.org/10.1186/s12866-019-1446-2

Dueholm, M. K. D., Nierychlo, M., Andersen, K. S., Rudkjøbing, V., Knutsson, S., Arriaga, S., Bakke, R., Boon, N., Bux, F., Christensson, M., Chua, A. S. M., Curtis, T. P., Cytryn, E., Erijman, L., Etchebehere, C., Fatta-Kassinos, D., Frigon, D., Garcia-Chaves, M. C., Gu, A. Z., … MiDAS Global Consortium. (2022). Author Correction: MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nature Communications*, *13*(1), 4017. https://doi.org/10.1038/s41467-022-31423-z

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics (Oxford, England)*, *31*(21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)*, *27*(16), 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, *27*(2), 105–117. https://doi.org/10.1016/j.tim.2018.11.003

Erzin, G., Pries, L.-K., Dimitrakopoulos, S., Ralli, I., Xenaki, L.-A., Soldatos, R. – F., Vlachos, I., Selakovic, M., Foteli, S., Kosteletos, I., Nianiakas, N., Mantonakis, L., Rizos, E., Kollias, K., Van Os, J., Guloksuz, S., & Stefanis, N. (2023). Association between exposome score for schizophrenia and functioning in first-episode psychosis: Results from the Athens first-episode psychosis research study. *Psychological Medicine*, *53*(6), 2609–2618. Cambridge Core. https://doi.org/10.1017/S0033291721004542

Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., Dillon, M. R., Bolyen, E., Caporaso, J. G., & Knight, R. (2020). QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols in Bioinformatics*, *70*(1), e100. https://doi.org/10.1002/cpbi.100

Fadeev, E., Cardozo-Mino, M. G., Rapp, J. Z., Bienhold, C., Salter, I., Salman-Carvalho, V., Molari, M., Tegetmeyer, H. E., Buttigieg, P. L., & Boetius, A. (2021). Comparison of Two 16S rRNA Primers (V3-V4 and V4-V5) for Studies of Arctic Microbial Communities. *Frontiers in Microbiology*, *12*, 637526. https://doi.org/10.3389/fmicb.2021.637526

Forner, L., Larsen, T., Kilian, M., & Holmstrup, P. (2006). Incidence of bacteremia after chewing, tooth brushing and scaling in individuals with periodontal inflammation. *Journal of Clinical Periodontology*, *33*(6), 401–407. https://doi.org/10.1111/j.1600-051X.2006.00924.x

Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P. J., Soen, Y., & Shental, N. (2018). Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome*, *6*(1), 17. https://doi.org/10.1186/s40168-017-0396-x

Funkhouser, L. J., & Bordenstein, S. R. (2013). Mom knows best: The universality of maternal microbial transmission. *PLoS Biology*, *11*(8), e1001631. https://doi.org/10.1371/journal.pbio.1001631

Galloway-Peña, J., & Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digestive Diseases and Sciences*, *65*(3), 674–685. https://doi.org/10.1007/s10620-020-06091-y

Gao, X., Lin, H., Revanna, K., & Dong, Q. (2017). A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*, *18*(1), 247. https://doi.org/10.1186/s12859-017-1670-4

García-López, R., Cornejo-Granados, F., Lopez-Zavala, A. A., Cota-Huízar, A., Sotelo-Mundo, R. R., Gómez-Gil, B., & Ochoa-Leyva, A. (2021). OTUs and ASVs Produce

Comparable Taxonomic and Diversity from Shrimp Microbiota 16S Profiles Using Tailored Abundance Filters. *Genes*, *12*(4). https://doi.org/10.3390/genes12040564

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Gebrayel, P., Nicco, C., Al Khodor, S., Bilinski, J., Caselli, E., Comelli, E. M., Egert, M., Giaroni, C., Karpinski, T. M., Loniewski, I., Mulak, A., Reygner, J., Samczuk, P., Serino, M., Sikora, M., Terranegra, A., Ufnal, M., Villeger, R., Pichon, C., … Edeas, M. (2022). Microbiota medicine: Towards clinical revolution. *Journal of Translational Medicine*, *20*(1), 111. https://doi.org/10.1186/s12967-022-03296-9

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392–400. https://doi.org/10.1038/nm.4517

Glassman, S. I., & Martiny, J. B. H. (2018). Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere*, *3*(4). https://doi.org/10.1128/mSphere.00148-18

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Gołębiewski, M., & Tretyn, A. (2020). Generating amplicon reads for microbial community assessment with next-generation sequencing. *Journal of Applied Microbiology*, *128*(2), 330–354. https://doi.org/10.1111/jam.14380

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Gupta, P. K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology*, *26*(11), 602–611. https://doi.org/10.1016/j.tibtech.2008.07.003

GUZE, S. B. (1995). Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV). *American Journal of Psychiatry*, *152*(8), 1228–1228. https://doi.org/10.1176/ajp.152.8.1228

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., Petrosino, J. F., Knight, R., & Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, *21*(3), 494–504. https://doi.org/10.1101/gr.112730.110

Hajj, J., Blaine, N., Salavaci, J., & Jacoby, D. (2018). The "Centrality of Sepsis": A Review on Incidence, Mortality, and Cost of Care. *Healthcare (Basel, Switzerland)*, *6*(3). https://doi.org/10.3390/healthcare6030090

Hancock, A. M., Witonsky, D. B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J., Coop, G., & Di Rienzo, A. (2010). Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, *107*(supplement_2), 8924–8930. https://doi.org/10.1073/pnas.0914625107

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics*, *16*(1), 169. https://doi.org/10.1186/s12859-015-0611-3

Hjorthøj, C., Stürup, A. E., McGrath, J. J., & Nordentoft, M. (2017). Years of potential life lost and life expectancy in schizophrenia: A systematic review and meta-analysis. *The Lancet. Psychiatry*, *4*(4), 295–301. https://doi.org/10.1016/S2215-0366(17)30078-0

Hornung, B. V. H., Zwittink, R. D., & Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology*, *95*(5). https://doi.org/10.1093/femsec/fiz045

Hosseini, S. M., Pratas, D., & Pinho, A. J. (2016). A Survey on Data Compression Methods for Biological Sequences. *Inf.*, *7*, 56.

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Huber, T., Faulkner, G., & Hugenholtz, P. (2004). Bellerophon: A program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics (Oxford, England)*, *20*(14), 2317–2319. https://doi.org/10.1093/bioinformatics/bth226

Hussing, C., Kampmann, M.-L., Mogensen, H. S., Børsting, C., & Morling, N. (2018). Quantification of massively parallel sequencing libraries—A comparative study of eight methods. *Scientific Reports*, *8*(1), 1110. https://doi.org/10.1038/s41598-018-19574-w

Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., Taylor, J., & Nekrutenko, A. (2020). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*, *48*(W1), W395–W402. https://doi.org/10.1093/nar/gkaa434

Jiménez, E., Fernández, L., Marín, M. L., Martín, R., Odriozola, J. M., Nueno-Palop, C., Narbad, A., Olivares, M., Xaus, J., & Rodríguez, J. M. (2005). Isolation of Commensal Bacteria from Umbilical Cord Blood of Healthy Neonates Born by Cesarean Section. *Current Microbiology*, *51*(4), 270–274. https://doi.org/10.1007/s00284-005-0020-3

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. https://doi.org/10.1038/s41467-019-13036-1

Joos, L., Beirinckx, S., Haegeman, A., Debode, J., Vandecasteele, B., Baeyen, S., Goormachtig, S., Clement, L., & De Tender, C. (2020). Daring to be differential: Metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. *BMC Genomics*, *21*(1), 733. https://doi.org/10.1186/s12864-020-07126-4

Jünemann, S., Kleinbölting, N., Jaenicke, S., Henke, C., Hassa, J., Nelkner, J., Stolze, Y., Albaum, S. P., Schlüter, A., Goesmann, A., Sczyrba, A., & Stoye, J. (2017). Bioinformatics for NGS-based metagenomics and the application to biogas research. *Journal of Biotechnology*, *261*, 10–23. https://doi.org/10.1016/j.jbiotec.2017.08.012

Kadri, K. (2019). Polymerase Chain Reaction (PCR): Principle and Applications. In M. L. Nagpal, O.-M. Boldura, C. Baltă, & S. Enany (Eds.), *Synthetic Biology*. IntechOpen. https://doi.org/10.5772/intechopen.86491

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe, A. J., Braun, J., & McWeeney, S. (2019). Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems*, *4*(4). https://doi.org/10.1128/mSystems.00290-19

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kerrigan, Z., & D'Hondt, S. (2022). Patterns of Relative Bacterial Richness and Community Composition in Seawater and Marine Sediment Are Robust for Both Operational Taxonomic Units and Amplicon Sequence Variants. *Frontiers in Microbiology*, *13*, 796758. https://doi.org/10.3389/fmicb.2022.796758

Kostopoulos, D. (2015). Taxonomy & Systematics. In *LIFE EVOLUTION: CHORDATA* (1st ed., p. 51). Kallipos, Open Academic Editions. http://hdl.handle.net/11419/1912

Kowarsky, M., Camunas-Soler, J., Kertesz, M., De Vlaminck, I., Koh, W., Pan, W., Martin, L., Neff, N. F., Okamoto, J., Wong, R. J., Kharbanda, S., El-Sayed, Y., Blumenfeld, Y., Stevenson, D. K., Shaw, G. M., Wolfe, N. D., & Quake, S. R. (2017). Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(36), 9623–9628. https://doi.org/10.1073/pnas.1707009114

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Kullar, R., Goldstein, E. J. C., Johnson, S., & McFarland, L. V. (2023). Lactobacillus Bacteremia and Probiotics: A Review. *Microorganisms*, *11*(4). https://doi.org/10.3390/microorganisms11040896

Lakhtakia, R. (2014). The Legacy of Robert Koch: Surmise, search, substantiate. *Sultan Qaboos University Medical Journal*, *14*(1), e37-41. https://doi.org/10.12816/0003334

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, *82*(20), 6955–6959. https://doi.org/10.1073/pnas.82.20.6955

Lane, N. (2015). The unseen world: Reflections on Leeuwenhoek (1677) "Concerning little animals". *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1666). https://doi.org/10.1098/rstb.2014.0344

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. https://doi.org/10.1038/nbt.2676

Lauder, A. P., Roche, A. M., Sherrill-Mix, S., Bailey, A., Laughlin, A. L., Bittinger, K., Leite, R., Elovitz, M. A., Parry, S., & Bushman, F. D. (2016). Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome*, *4*(1), 29. https://doi.org/10.1186/s40168-016-0172-3

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239

Li, Q., Wang, C., Tang, C., Zhao, X., He, Q., & Li, J. (2018). Identification and Characterization of Blood and Neutrophil-Associated Microbiomes in Patients with Severe Acute Pancreatitis Using Next-Generation Sequencing. *Frontiers in Cellular and Infection Microbiology*, *8*. https://www.frontiersin.org/articles/10.3389/fcimb.2018.00005

Li, Y., & Chen, L. (2014). Big biological data: Challenges and opportunities. *Genomics, Proteomics & Bioinformatics*, *12*(5), 187–189. https://doi.org/10.1016/j.gpb.2014.10.001

Liu, Y.-X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell*, *12*(5), 315–330. https://doi.org/10.1007/s13238-020-00724-8

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, *8*(1), 51. https://doi.org/10.1186/s13073-016-0307-y

Louca, S., Doebeli, M., & Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, *6*(1), 41. https://doi.org/10.1186/s40168-018-0420-9

Madigan, M., Sattley, W., Aiyer, J., Stahl, D., & Buckley, D. (2021). *Brock Biology of Microorganisms, Global Edition*. Pearson Deutschland. https://elibrary.pearson.de/book/99.150005/9781292405063

Malysa, G., Hernaez, M., Ochoa, I., Rao, M., Ganesan, K., & Weissman, T. (2015). QVZ: lossy compression of quality values. *Bioinformatics (Oxford, England)*, *31*(19), 3122–3129. https://doi.org/10.1093/bioinformatics/btv330

Manzo, V. E., & Bhatt, A. S. (2015). The human microbiome in hematopoiesis and hematologic disorders. *Blood*, *126*(3), 311–318. https://doi.org/10.1182/blood-2015-04-574392

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Marini, F., Linke, J., & Binder, H. (2020). ideal: An R/Bioconductor package for interactive differential expression analysis. *BMC Bioinformatics*, *21*(1), 565. https://doi.org/10.1186/s12859-020-03819-5

Martel, J., Wu, C.-Y., Huang, P.-R., Cheng, W.-Y., & Young, J. D. (2017). Pleomorphic bacteria-like structures in human blood represent non-living membrane vesicles and protein particles. *Scientific Reports*, *7*(1), 10650. https://doi.org/10.1038/s41598-017-10479-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., & Droit, A. (2022). Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. *Frontiers in Microbiology*, *13*. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.811495

Matsuo, Y., Komiya, S., Yasumizu, Y., Yasuoka, Y., Mizushima, K., Takagi, T., Kryukov, K., Fukuda, A., Morimoto, Y., Naito, Y., Okada, H., Bono, H., Nakagawa, S., & Hirota, K. (2021). Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing confers species-level resolution. *BMC Microbiology*, *21*(1), 35. https://doi.org/10.1186/s12866-021-02094-5

Mayer, E. A., Knight, R., Mazmanian, S. K., Cryan, J. F., & Tillisch, K. (2014). Gut microbes and the brain: Paradigm shift in neuroscience. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(46), 15490–15496. https://doi.org/10.1523/JNEUROSCI.3299-14.2014

McCutcheon, R. A., Reis Marques, T., & Howes, O. D. (2020). Schizophrenia-An Overview. *JAMA Psychiatry*, *77*(2), 201–210. https://doi.org/10.1001/jamapsychiatry.2019.3360

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618. https://doi.org/10.1038/ismej.2011.139

McKnight, P. E., & Najab, J. (2010). Mann-Whitney U Test. In *The Corsini Encyclopedia of Psychology* (pp. 1–1). https://doi.org/10.1002/9780470479216.corpsy0524

McLaughlin, R. W., Vali, H., Lau, P. C. K., Palfree, R. G. E., De Ciccio, A., Sirois, M., Ahmad, D., Villemur, R., Desrosiers, M., & Chan, E. C. S. (2002). Are there naturally occurring pleomorphic bacteria in the blood of healthy humans? *Journal of Clinical Microbiology*, *40*(12), 4771–4775. https://doi.org/10.1128/JCM.40.12.4771-4775.2002

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Mirzaei, M. (2020). *Science and Engineering In Silico*. https://api.semanticscholar.org/CorpusID:229079725

Mitchell, A. J., Gray, W. D., Schroeder, M., Yi, H., Taylor, J. V., Dillard, R. S., Ke, Z., Wright, E. R., Stephens, D., Roback, J. D., & Searles, C. D. (2016). Pleomorphic Structures in Human Blood Are Red Blood Cell-Derived Microparticles, Not Bacteria. *PloS One*, *11*(10), e0163582. https://doi.org/10.1371/journal.pone.0163582

Mitra, S., Stärk, M., & Huson, D. H. (2011). Analysis of 16S rRNA environmental sequences using MEGAN. *BMC Genomics*, *12*(3), S17. https://doi.org/10.1186/1471-2164-12-S3-S17

Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N., Pusey, A. E., Peeters, M., Hahn, B. H., & Ochman, H. (2016). Cospeciation of gut microbiota with hominids. *Science (New York, N.Y.)*, *353*(6297), 380–382. https://doi.org/10.1126/science.aaf3951

Mohsen, A., Park, J., Chen, Y.-A., Kawashima, H., & Mizuguchi, K. (2019). Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks. *BMC Bioinformatics*, *20*(1), 581. https://doi.org/10.1186/s12859-019-3187-5

Mokhtari, E. B., & Ridenhour, B. J. (2022). Filtering ASVs/OTUs via mutual information-based microbiome network analysis. *BMC Bioinformatics*, *23*(1), 380. https://doi.org/10.1186/s12859-022-04919-0

Moossavi, S., Atakora, F., Fehr, K., & Khafipour, E. (2020). Biological observations in microbiota analysis are robust to the choice of 16S rRNA gene sequencing processing algorithm: Case study on human milk microbiota. *BMC Microbiology*, *20*(1), 290. https://doi.org/10.1186/s12866-020-01949-7

Morgan, X. C., & Huttenhower, C. (2012). Chapter 12: Human microbiome analysis. *PLoS Computational Biology*, *8*(12), e1002808. https://doi.org/10.1371/journal.pcbi.1002808

Moriyama, K., Ando, C., Tashiro, K., Kuhara, S., Okamura, S., Nakano, S., Takagi, Y., Miki, T., Nakashima, Y., & Hirakawa, H. (2008). Polymerase chain reaction detection of

bacterial 16S rRNA gene in human blood. *Microbiology and Immunology*, *52*(7), 375–382. https://doi.org/10.1111/j.1348-0421.2008.00048.x

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K. E., Venter, J. C., & Telenti, A. (2017). The blood DNA virome in 8,000 humans. *PLoS Pathogens*, *13*(3), e1006292. https://doi.org/10.1371/journal.ppat.1006292

Munawar, N., Ahsan, K., Muhammad, K., Ahmad, A., Anwar, M. A., Shah, I., Al Ameri, A. K., & Al Mughairbi, F. (2021). Hidden Role of Gut Microbiome Dysbiosis in Schizophrenia: Antipsychotics or Psychobiotics as Therapeutics? *International Journal of Molecular Sciences*, *22*(14). https://doi.org/10.3390/ijms22147671

Murray, P., & Witebsky, F. G. (2014). *The Clinician and the Microbiology Laboratory*. *1*, 233–265. https://doi.org/10.1016/B978-1-4557-4801-3.00016-3

Mysara, M., Saeys, Y., Leys, N., Raes, J., & Monsieurs, P. (2015). CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Applied and Environmental Microbiology*, *81*(5), 1573–1584. https://doi.org/10.1128/AEM.02896-14

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364. https://doi.org/10.7717/peerj.5364

Needham, D. M., Sachdeva, R., & Fuhrman, J. A. (2017). Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME Journal*, *11*(7), 1614–1629. https://doi.org/10.1038/ismej.2017.29

Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., Meltser, A., Douglas, G. M., Kamer, I., Gopalakrishnan, V., Dadosh, T., Levin-Zaidman, S., Avnet, S., Atlan, T., Cooper, Z. A., … Straussman, R. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science (New York, N.Y.)*, *368*(6494), 973–980. https://doi.org/10.1126/science.aay9189

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nikkari, S., McLaughlin, I. J., Bi, W., Dodge, D. E., & Relman, D. A. (2001). Does blood of healthy subjects contain bacterial ribosomal DNA? *Journal of Clinical Microbiology*, *39*(5), 1956–1959. https://doi.org/10.1128/JCM.39.5.1956-1959.2001

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H., & Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology*, *8*(11), e1000546. https://doi.org/10.1371/journal.pbio.1000546

Olde Loohuis, L. M., Mangul, S., Ori, A. P. S., Jospin, G., Koslicki, D., Yang, H. T., Wu, T., Boks, M. P., Lomen-Hoerth, C., Wiedau-Pazos, M., Cantor, R. M., de Vos, W. M., Kahn, R. S., Eskin, E., & Ophoff, R. A. (2018). Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. *Translational Psychiatry*, *8*(1), 96. https://doi.org/10.1038/s41398-018-0107-9

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., & Iliopoulos, I. (2015). Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, *9*, 75–88. https://doi.org/10.4137/BBI.S12462

Païssé, S., Valle, C., Servant, F., Courtney, M., Burcelin, R., Amar, J., & Lelouvier, B. (2016). Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion*, *56*(5), 1138–1147. https://doi.org/10.1111/trf.13477

Pan, A. Y. (2021). Statistical analysis of microbiome data: The challenge of sparsity. *Current Opinion in Endocrine and Metabolic Research*, *19*, 35–40. https://doi.org/10.1016/j.coemr.2021.05.005

Panaiotov, S., Filevski, G., Equestre, M., Nikolova, E. R., & Kalfin, R. (2018). Cultural Isolation and Characteristics of the Blood Microbiome of Healthy Individuals. *Ai Magazine*, *08*, 406–421.

Panaiotov, S., Hodzhev, Y., Tsafarova, B., Tolchkov, V., & Kalfin, R. (2021). Culturable and Non-Culturable Blood Microbiota of Healthy Individuals. *Microorganisms*, *9*(7). https://doi.org/10.3390/microorganisms9071464

Pandey, V., Nutter, R. C., & Prediger, E. (2008). Applied Biosystems SOLiD™ System: Ligation-Based Sequencing. In *Next Generation Genome Sequencing* (pp. 29–42). https://doi.org/10.1002/9783527625130.ch3

Pascal, M., Perez-Gordo, M., Caballero, T., Escribese, M. M., Lopez Longo, M. N., Luengo, O., Manso, L., Matheu, V., Seoane, E., Zamorano, M., Labrador, M., & Mayorga, C. (2018). Microbiome and Allergic Diseases. *Frontiers in Immunology*, *9*, 1584. https://doi.org/10.3389/fimmu.2018.01584

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Peeters, J., Thas, O., Shkedy, Z., Kodalci, L., Musisi, C., Owokotomo, O. E., Dyczko, A., Hamad, I., Vangronsveld, J., Kleinewietfeld, M., Thijs, S., & Aerts, J. (2021). Exploring the Microbiome Analysis and Visualization Landscape. *Frontiers in Bioinformatics*, *1*. https://www.frontiersin.org/articles/10.3389/fbinf.2021.774631

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., van den Brandt, P. A., & Stobberingh, E. E. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics*, *118*(2), 511–521. https://doi.org/10.1542/peds.2005-2824

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, *118*(2), 111–124. https://doi.org/10.1038/hdy.2016.102

Potgieter, M., Bester, J., Kell, D. B., & Pretorius, E. (2015). The dormant blood microbiome in chronic, inflammatory diseases. *FEMS Microbiology Reviews*, *39*(4), 567–591. https://doi.org/10.1093/femsre/fuv013

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PloS One*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PloS One*, *15*(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

Qian, X.-B., Chen, T., Xu, Y.-P., Chen, L., Sun, F.-X., Lu, M.-P., & Liu, Y.-X. (2020). A guide to human microbiome research: Study design, sample collection, and

bioinformatics analysis. *Chinese Medical Journal*, *133*(15), 1844–1855. https://doi.org/10.1097/CM9.0000000000000871

Qiu, J., Zhou, H., Jing, Y., & Dong, C. (2019). Association between blood microbiome and type 2 diabetes mellitus: A nested case-control study. *Journal of Clinical Laboratory Analysis*, *33*(4), e22842. https://doi.org/10.1002/jcla.22842

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

Reay, W. R., Kiltschewskij, D. J., Geaghan, M. P., Atkins, J. R., Carr, V. J., Green, M. J., & Cairns, M. J. (2022). Genetic estimates of correlation and causality between blood-based biomarkers and psychiatric disorders. *Science Advances*, *8*(14), eabj8969. https://doi.org/10.1126/sciadv.abj8969

Regueira-Iglesias, A., Balsa-Castro, C., Blanco-Pintos, T., & Tomás, I. (2023). Critical review of 16S rRNA gene sequencing workflow in microbiome studies: From primer selection to advanced data analysis. *Molecular Oral Microbiology*, *38*(5), 347–399. https://doi.org/10.1111/omi.12434

Reitmeier, S., Hitch, T. C. A., Treichel, N., Fikas, N., Hausmann, B., Ramer-Tait, A. E., Neuhaus, K., Berry, D., Haller, D., Lagkouvardos, I., & Clavel, T. (2021). Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Communications*, *1*(1), 31. https://doi.org/10.1038/s43705-021-00033-z

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rohwer, R. R., Hamilton, J. J., Newton, R. J., & McMahon, K. D. (2018). TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere*, *3*(5). https://doi.org/10.1128/mSphere.00327-18

Romano-Keeler, J., & Weitkamp, J.-H. (2015). Maternal influences on fetal microbial colonization and immune development. *Pediatric Research*, *77*(1), 189–195. https://doi.org/10.1038/pr.2014.163

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. https://doi.org/10.1038/nature10242

Saha, S., Johnson, J., Pal, S., Weinstock, G. M., & Rajasekaran, S. (2019). MSC: a metagenomic sequence classification algorithm. *Bioinformatics (Oxford, England)*, *35*(17), 2932–2940. https://doi.org/10.1093/bioinformatics/bty1071

Saladin, K. S. (2011). *Human Anatomy* (p. 520). McGraw-Hill.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*, 87. https://doi.org/10.1186/s12915-014-0087-z

Sambo, F., Finotello, F., Lavezzo, E., Baruzzo, G., Masi, G., Peta, E., Falda, M., Toppo, S., Barzon, L., & Di Camillo, B. (2018). Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics*, *19*(1), 343. https://doi.org/10.1186/s12859-018-2360-6

Sato, J., Kanazawa, A., Ikeda, F., Yoshihara, T., Goto, H., Abe, H., Komiya, K., Kawaguchi, M., Shimizu, T., Ogihara, T., Tamura, Y., Sakurai, Y., Yamamoto, R., Mita, T., Fujitani, Y., Fukuda, H., Nomoto, K., Takahashi, T., Asahara, T., … Watada, H. (2014). Gut dysbiosis and detection of "live gut bacteria" in blood of Japanese patients with type 2 diabetes. *Diabetes Care*, *37*(8), 2343–2350. https://doi.org/10.2337/dc13-2817

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Sekse, C., Holst-Jensen, A., Dobrindt, U., Johannessen, G. S., Li, W., Spilsberg, B., & Shi, J. (2017). High Throughput Sequencing for Detection of Foodborne Pathogens. *Frontiers in Microbiology*, *8*, 2029. https://doi.org/10.3389/fmicb.2017.02029

Severance, E. G., Gressitt, K. L., Stallings, C. R., Origoni, A. E., Khushalani, S., Leweke, F. M., Dickerson, F. B., & Yolken, R. H. (2013). Discordant patterns of bacterial translocation markers and implications for innate immune imbalances in schizophrenia. *Schizophrenia Research*, *148*(1–3), 130–137. https://doi.org/10.1016/j.schres.2013.05.018

Shehadul Islam, M., Aryasomayajula, A., & Selvaganapathy, P. R. (2017). A Review on Macroscale and Microscale Cell Lysis Methods. *Micromachines*, *8*(3), 83. https://doi.org/10.3390/mi8030083

Shulman, S. T., Friedmann, H. C., & Sims, R. H. (2007). Theodor Escherich: The First Pediatric Infectious Diseases Physician? *Clinical Infectious Diseases*, *45*(8), 1025–1029. https://doi.org/10.1086/521946

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science : A Publication of the Protein Society*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290

SIMPSON, E. H. (1949). Measurement of Diversity. *Nature*, *163*(4148), 688–688. https://doi.org/10.1038/163688a0

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Socała, K., Doboszewska, U., Szopa, A., Serefko, A., Włodarczyk, M., Zielińska, A., Poleszak, E., Fichna, J., & Wlaź, P. (2021). The role of microbiota-gut-brain axis in neuropsychiatric and neurological disorders. *Pharmacological Research*, *172*, 105840. https://doi.org/10.1016/j.phrs.2021.105840

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sudo, N., Chida, Y., Aiba, Y., Sonoda, J., Oyama, N., Yu, X.-N., Kubo, C., & Koga, Y. (2004). Postnatal microbial colonization programs the hypothalamic-pituitary-adrenal system for stress response in mice. *The Journal of Physiology*, *558*(Pt 1), 263–275. https://doi.org/10.1113/jphysiol.2004.063388

Swanson, H. I. (2015). Drug Metabolism by the Host and Gut Microbiota: A Partnership or Rivalry? *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *43*(10), 1499–1504. https://doi.org/10.1124/dmd.115.065714

Szeligowski, T., Yun, A. L., Lennox, B. R., & Burnet, P. W. J. (2020). The Gut Microbiome and Schizophrenia: The Current State of the Field and Clinical Applications. *Frontiers in Psychiatry*, *11*, 156. https://doi.org/10.3389/fpsyt.2020.00156

Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, *9*(1), 2856. https://doi.org/10.1038/s41598-019-39076-7

Tedeschi, G. G., Amici, D., & Paparelli, M. (1969). Incorporation of nucleosides and amino-acids in human erythrocyte suspensions: Possible relation with a diffuse infection of mycoplasms or bacteria in the L form. *Nature*, *222*(5200), 1285–1286. https://doi.org/10.1038/2221285a0

Teng, F., Darveekaran Nair, S. S., Zhu, P., Li, S., Huang, S., Li, X., Xu, J., & Yang, F. (2018). Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Scientific Reports*, *8*(1), 16321. https://doi.org/10.1038/s41598-018-34294-x

Thanukos, A. (2009). A Name by Any Other Tree. *Evolution: Education and Outreach*, *2*(2), 303–309. https://doi.org/10.1007/s12052-009-0122-7

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics—A guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3

Traykova, D., Schneider, B., Chojkier, M., & Buck, M. (2017). Blood Microbiome Quantity and the Hyperdynamic Circulation in Decompensated Cirrhotic Patients. *PloS One*, *12*(2), e0169310. https://doi.org/10.1371/journal.pone.0169310

Trego, A., Keating, C., Nzeteu, C., Graham, A., O'Flaherty, V., & Ijaz, U. Z. (2022). Beyond Basic Diversity Estimates-Analytical Tools for Mechanistic Interpretations of Amplicon Sequencing Data. *Microorganisms*, *10*(10). https://doi.org/10.3390/microorganisms10101961

Urbaniak, C., Cummins, J., Brackstone, M., Macklaim, J. M., Gloor, G. B., Baban, C. K., Scott, L., O'Hanlon, D. M., Burton, J. P., Francis, K. P., Tangney, M., & Reid, G. (2014). Microbiota of human breast tissue. *Applied and Environmental Microbiology*, *80*(10), 3007–3014. https://doi.org/10.1128/AEM.00242-14

Valdes, A. M., Walter, J., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ (Clinical Research Ed.)*, *361*, k2179. https://doi.org/10.1136/bmj.k2179

Vargas-Albores, F., Ortiz-Suárez, L. E., Villalpando-Canchola, E., & Martínez-Porchas, M. (2017). Size-variable zone in V3 region of 16S rRNA. *RNA Biology*, *14*(11), 1514–1521. https://doi.org/10.1080/15476286.2017.1317912

Velmurugan, G., Dinakaran, V., Rajendhran, J., & Swaminathan, K. (2020). Blood Microbiota and Circulating Microbial Metabolites in Diabetes and Cardiovascular Disease. *Trends in Endocrinology and Metabolism: TEM*, *31*(11), 835–847. https://doi.org/10.1016/j.tem.2020.01.013

Wagh, V. V., Vyas, P., Agrawal, S., Pachpor, T. A., Paralikar, V., & Khare, S. P. (2021). Peripheral Blood-Based Gene Expression Studies in Schizophrenia: A Systematic Review. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.736483

Wakita, Y., Shimomura, Y., Kitada, Y., Yamamoto, H., Ohashi, Y., & Matsumoto, M. (2018). Taxonomic classification for microbiome analysis, which correlates well with the metabolite milieu of the gut. *BMC Microbiology*, *18*(1), 188. https://doi.org/10.1186/s12866-018-1311-8

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Wei, Z.-G., Zhang, X.-D., Cao, M., Liu, F., Qian, Y., & Zhang, S.-W. (2021). Comparison of Methods for Picking the Operational Taxonomic Units From Amplicon Sequences. *Frontiers in Microbiology*, *12*, 644012. https://doi.org/10.3389/fmicb.2021.644012

Weinroth, M. D., Belk, A. D., Dean, C., Noyes, N., Dittoe, D. K., Rothrock, M. J., Ricke, S. C., Myer, P. R., Henniger, M. T., Ramírez, G. A., Oakley, B. B., Summers, K. L., Miles, A. M., Ault-Seay, T. B., Yu, Z., Metcalf, J. L., & Wells, J. E. (2022). Considerations and best practices in animal science 16S ribosomal RNA gene sequencing microbiome studies. *Journal of Animal Science*, *100*(2). https://doi.org/10.1093/jas/skab346

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, *3*, e1487. https://doi.org/10.7717/peerj.1487

Whittle, E., Leonard, M. O., Harrison, R., Gant, T. W., & Tonge, D. P. (2018). Multi-Method Characterization of the Human Circulating Microbiome. *Frontiers in Microbiology*, *9*, 3266. https://doi.org/10.3389/fmicb.2018.03266

Wilkins, O. G., Capitanchik, C., Luscombe, N. M., & Ule, J. (2021). Ultraplex: A rapid, flexible, all-in-one fastq demultiplexer. *Wellcome Open Research*, *6*, 141. https://doi.org/10.12688/wellcomeopenres.16791.1

Wolfe, A. J., Toh, E., Shibata, N., Rong, R., Kenton, K., Fitzgerald, M., Mueller, E. R., Schreckenberger, P., Dong, Q., Nelson, D. E., & Brubaker, L. (2012). Evidence of uncultivated bacteria in the adult female bladder. *Journal of Clinical Microbiology*, *50*(4), 1376–1383. https://doi.org/10.1128/JCM.05852-11

World Health Organization. (1992). The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines. *Classification Internationale Des Troubles Mentaux et Des Troubles Du Comportement : Descriptions Cliniques et Directives Pour Le Diagnostic*. WHO IRIS. https://iris.who.int/handle/10665/37958

Xenaki, L. A., Kollias, C. T., Stefanatou, P., Ralli, I., Soldatos, R.-F., Dimitrakopoulos, S., Hatzimanolis, A., Triantafyllou, T.-F., Kosteletos, I., Vlachos, I. I., Selakovic, M., Foteli, S., Mantonakis, L., Ermiliou, V., Voulgaraki, M., Psarra, E., Gülöksüz, S., van Os, J., & Stefanis, N. C. (2020). Organization framework and preliminary findings from the Athens First-Episode Psychosis Research Study. *Early Intervention in Psychiatry*, *14*(3), 343–355. https://doi.org/10.1111/eip.12865

Xenaki, L. A., Stefanatou, P., Ralli, E., Hatzimanolis, A., Dimitrakopoulos, S., Soldatos, R. F., Vlachos, I. I., Selakovic, M., Foteli, S., Kosteletos, I., Nianiakas, N., Ntigridaki, A., Triantafyllou, T.-F., Voulgaraki, M., Mantonakis, L., Tsapas, A., Bozikas, V. P., Kollias, K., & Stefanis, N. C. (2022). The relationship between early symptom severity, improvement and remission in first episode psychosis with jumping to conclusions. *Schizophrenia Research*, *240*, 24–30. https://doi.org/10.1016/j.schres.2021.11.039

Xia, Y., Sun, J., & Chen, D.-G. (2018). Community Diversity Measures and Calculations. In Y. Xia, J. Sun, & D.-G. Chen (Eds.), *Statistical Analysis of Microbiome Data with R* (pp. 167–190). Springer Singapore. https://doi.org/10.1007/978-981-13-1534-3_6

Xue, Z., Kable, M. E., & Marco, M. L. (2018). Impact of DNA Sequencing and Analysis Methods on 16S rRNA Gene Bacterial Community Analysis of Dairy Products. *mSphere*, *3*(5). https://doi.org/10.1128/mSphere.00410-18

Yang, B., Wang, Y., & Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*, *17*, 135. https://doi.org/10.1186/s12859-016-0992-y

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*(5), 303–314. https://doi.org/10.1038/nrg3186

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zheng, P., Zeng, B., Liu, M., Chen, J., Pan, J., Han, Y., Liu, Y., Cheng, K., Zhou, C., Wang, H., Zhou, X., Gui, S., Perry, S. W., Wong, M.-L., Licinio, J., Wei, H., & Xie, P. (2019). The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice. *Science Advances*, *5*(2), eaau8317. https://doi.org/10.1126/sciadv.aau8317

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, *18*(1), 186. https://doi.org/10.1186/s13059-017-1319-7